**Associate Editor**

Figure 4 is not cited in the text. Figures 2 and 3 could be reformatted to black and white if authors wish. For Figure 2, authors could use the species symbols as they did in Figure 1 (square for A. flavescens and circle for A. formosa). For Figure 3 populations could be given a number for example on the x axis that could match the populations list of Table 1.

* Figure 4 is now cited in the text.
* Figures 2 and 3 have been reformatted to black and white according to the above suggestions. Note that Table 1 has slightly changed; there is an additional column giving a number identifier which figure 3 references, an additional column reporting the mean hybrid index for each group, and an additional row for *A. flavescens* var. *miniana* type specimens (these are not among the sampled populations but this makes the labelling of Figure 3 more consistent.)

**Reviewer 1**

One of the main concerns in this manuscript is the analytical procedure used to calculate the hybrid index. I believe the linear discriminant analysis (LDA) is somehow appropriate (though it still needs justification of why authors decided to use LDA); however, the description of how it was performed is not well described and makes it difficult to assess accurately. Authors missed to describe the variables included in the analysis, the metric used (i.e., whether the variables were standardized or not), did all variables were significant or most influential? How many were used or discarded? Were any cases misclassified - how many? Etc. Thus, it is hard to follow and understand what exactly the authors did. I think this part needs to be re-written, and all the stats outcome provided for a better evaluation.

* We have added a short description with citation as justification of the use of LDA as a method of constructing a hybrid index.
* In the updated version of the manuscript we have focused only on the continuous traits and left out the previously mentioned categorical trait, for two reasons. First, this trait technically does not meet the assumptions of the LDA model, namely normality, and second, it contributed least to the discriminant function. Furthermore, the results remain highly consistent and we believe a simpler model is sufficient and preferable in this situation. Due to this the figures have been updated based on the new analysis.
* A table of trait means for each species and loadings onto the discriminant axis have been provided as supplementary material (Table S1). We have also included as supplementary material a diagram of the floral measurements on a pressed flower (Figure S1). Note that this same floral diagram appeared in our previous publication Groh et al (2019). We believe it is a valuable visual aid to help the reader better understand our data collection process.
* We manually checked that out implementation of LDA produces identical results whether variables are scaled prior to running the LDA analysis or not. The only difference is in the values of trait loadings onto the discriminant axis. As the scaling of traits depends on the particular sample variance of traits in a given data set, we avoid scaling the variables prior to running LDA, and prefer to report the loadings of raw trait values onto the discriminant axis as these could potentially be used by taxonomists.
* We have added the classification accuracy of the discriminant function in the text.

Another major concern about the stats part in this paper is the use of PCA in addition to LDA. These two methods aim for different things, and I don’t see a clear justification for adding PCA to the analysis. I think the incorporation of the PCA brings more data manipulation (statistically speaking), which may cause divergences in the final results and perhaps wrong conclusions. I would suggest looking for alternatives to deal with the variation in flower size; for example, using flower diameter/height ratio that may solve the problem of species-specific trait combinations rather than PCA scores (See Page 6, lines 126-131).

* We agree that this approach may not have been adequately explained, however PCA is a robust and commonly used method for quantifying allometry using multiple traits. Our approach to size correction with PCA is directly analogous to a univariate regression of a focal trait against size, the only difference being that in this case both the focal trait (the hybrid index constructed from LDA) and size (PC2) are multivariate. We believe our interpretation of PC2 as a size variable is sound, given that all continuous traits load with the same sign onto this axis, meaning that it captures the variation resulting from joint increase in all traits.
* Regarding the reviewer’s suggestion of the use of a ratio, we would argue that this may be less informative, as a given ratio only represents relative dimensions of two traits, and thus may not be sensitive to size at all if the two traits grow at similar rates throughout development.
* We would also like to point out that the effect of size on our analyses is overall relatively slight, and thus our procedure is unlikely to lead to any incorrect conclusions. As a robustness check, we have run all analyses without the size-correction procedure and confirmed that all results are all highly similar. Nonetheless, we report results from analyses using the size-correction procedure as we believe this practice should be encouraged in morphometric studies.

Concerning the result section, I think the authors need to check the written part in the result section. Authors did a good writing job for the most part; however, some parts might be misplaced from the discussion. I believe authors should concentrate on describing what they found from the data they have collected, avoiding inferences and citations.

* We agree with the reviewer’s comment that some material from the results section is more suited to the discussion section. We have edited these two sections to achieve this.

Title: Aquilega? It is written in italics.

* We have edited the title.

Introduction: Perhaps including a brief definition/description of what hybrid zone and secondary contact are, might be useful for not well-familiarized readers with this terminology.

* We have added a brief definition of a hybrid zone in the beginning of the introduction.

P. 4 l. 69: please check the sentence “… taxa, and there is a natural bias for collectors to collect plants in flowering condition” seems not right.

* We have edited this sentence - it should now be clearer.

P. 4 l. 88: Replace “We combine this …” for “We combine this herbarium data…”

* done

P. 5 and 6 l. 108-111: Please clarify how herbarium sheet labels were reassigned coordinates? Moreover, provide a table summarizing all the floral measurements? Inline 111, I didn’t find any supplemental data

- We have indicated that we used the most accurate information available on the herbarium sheet labels. We have now included a table of floral measurements as supplementary material. Data files (and scripts) are also now included in a Github repo which is referenced at the end of the text.

P. 6 l. 126-137: See general comment at the beginning.   
  
P. 9 l. 179-183: these sentences are confusing and too long, a bit shorter may help. Also, when you say the sigmoid function transformed size-adjusted discriminant scores, means double transformation? Since they were already extracted from the PCA, where they were transformed already?

- We shortened these sentences. The meaning should be clear now; we first predict. discriminant scores, then include a regression-based size correction by using a size value derived from PCA, then re-scale the values by a sigmoid function to generate a hybrid index value in this case.

P. 10 l. 209: “hybrid index values are strongly predicted…” I don’t see any r-square and p-values to confirm that statement.

- We added r-squared values and P values for these regressions.   
  
P. 10 and 11 l. 217- 229: no citations in the result section. Also, this paragraph fits better in the discussion.

- We agree that this section was too verbose for the results section. We have stripped it down to a simpler reporting of the findings.

P. 12 l. 258-259: you said, “… we visually observed some morphological affinity to A. flavescens” which ones?

- By this statement we meant that several *A. flavescens*-like individuals were observed in this population, but we did not record any morphometric data from this population. We adjusted the sentence for this to be clearer.

P. 13 l. 262-267: fits better in the discussion

- We agree with the reviewer’s point. This material has been moved to the discussion under the heading “A note on taxonomic practice.”

P. 14 l. 284-290: fits better in the discussion

* This material has been moved to the discussion under the heading “Ecological and breeding system considerations”

P. 16 l. 335-336: how frequent visits by bees mean pollinator-isolation absence or weak? Please clarify

-We have worded this sentence more carefully. The meaning should now be clearer.

Reviewer 2

This manuscript represents an extension of the recent paper published in 2018 by Groh et al. to look more completely at the entire known range of hybrids of A. flavescens and A. formosa. In doing so, much of the methodology is not explained in this manuscript requiring the reader to have access to Groh et al., which may or may not be appropriate for publication in Botany. I personally find it difficult to follow and discern what was actually done as most details are not available. The same is true for the data. I see no reference to the data as a link (previous paper used a Github repo) or as an associated file. This would definitely help the reader to better understand what data is being used for which analyses as sometimes the sample numbers do not seem to add up.

* As described above, we have updated the manuscript to include more detail about the data collection and analysis. Regarding the reviewer’s point about sample numbers, we have more information in the text to hopefully clarify this, as well as including the data files themselves and analysis scripts in a Github repo that is referenced in the text.

The use of herbarium specimens to enhance a data set of morphological measurements is promising. I struggle a little bit to believe that all measurements could be made for every specimen as once pressed the three dimensional shape of the inflorescence is compromised and access to all measurements may be challenging. If trying to make the measurements from an image of the herbarium sheet, this is even more challenging. Also, I see no mention of the implications of drying of the inflorescence on measurements.

* The reviewer’s concern is valid. There were many specimens were not included in our analyses because not all traits could be measured, but our final data set includes only those for which we were confident in our measurements. We also note that our measurements don’t describe three-dimensional aspects of the flower. For example, corolla width (diameter) will of course give a different measure on a three-dimensional flower, but as the flowers are always flattened in our data set, this measurement does not depend on the three-dimensional structure of the flower. Most of the fully-developed flowers of *Aquilegia* are pressed in roughly the same conformation as the flower in Fig S1. While our measurements are not without error, we assume this measurement error is not biased in any way with respect to identification status, so this should not alter our conclusions here. The most common issue was for sepals to be somewhat crinkled (and therefore having a shorter width), but the sepal dimensions little to the species discrimination, so the noise introduced by this does not strongly influence our results or conclusions.
* Regarding digital specimens, we feel confident that at the resolution of our measurements we were able to accurately measure these traits from digitized specimens (most digitized specimens were scanned at an adequate pixel density).
* For specimens pressed correctly, the dimensions of the flower change little (personal experience), and this should not be biased with respect to taxonomic identity. In any case, we do not claim our measurements reflect measurements made on non-pressed flowers. We have added a brief disclosure statement of this in the methods section.

Reviewer 3  
  
This is nice case study of hybridization and introgression patterns within the genus Aquilegia. I have only some comments, particularly on the Material and Methods section:  
  
(1) Lines 98-100: The terms “floral phenotype” and “specimen phenotype” require explanation, they are not self-explanative.

* We have rewritten these sentences. The meaning should now be clearer.

(2) Lines 109-110: morphological characters measured should be enumerated more explicitly, reference to Groh et al. (2018) is not sufficient (albeit the paper is freely available online), moreover it is not clear whether one categorical character that is mentioned (petal laminae cleft, according to Groh et al., 2018) is binary or multistate one (only binary character would be acceptable for subsequent analyses).

- We now include Figure S1 and Table S1 which give more detail on the floral characters used.

(3) Lines 116-118: why exactly 40% of the data points with the least minimum geographic distance to a record of the alternative species in the data set where filtered out? Is this number purely arbitrary are there is some particular reason for that?

-We have now noted in the methods section that this value was chosen based on a value that resulted in nearly complete separation of the two groups without sacrificing too much data in the training set.

(4) Lines 160-169: I wonder which criteria were set up for the identification of introgressed and pure populations on both species as mentioned in this paragraph (and perhaps elsewhere in the paper). Identification of pure populations is the most critical step in all studies of hybrids.

- We edited the text to note that the two introgressed populations of *A. flavescens* were brought to our attention by taxonomists at the UBC herbarium. These regions were also identified as likely harboring introgressed individuals in our biogeographic analysis. The intermediacy in color and floral form is in line with these facts in suggesting introgression. We have also now noted in the text that the additional “pure” *A. formosa* population was sampled in a region with no records of *A. flavescens,* and that the population suspected of introgression was sampled in a region harboring a known hybrid population. Thus, our assessment of “pure” vs. introgressed is in each case in line with biogeographic data, and this is corroborated by our hybrid index values (Figure 3B).

The paper Groh et al. (2018) was based on both morphological and molecular (microsatellites and cpDNA sequences), was there any possibility to use those molecular data also in this paper?

- While addition of molecular data would make for a highly interesting study, we believe the manuscript stands alone in its emphasis on leveraging phenotype data alone from herbarium specimens.

The extension of the Discussion section beyond the genus Aquilegia would be highly appreciated. There are numerous studies dealing with the hybridization and some patterns that are common with other genera or conclusions that are more generally valid should be mentioned.

-We have included further discussion of similar patterns of introgression in other genera in the discussion under the section “Models of introgression”.