Review of Manuscript MBE-23-0129.R1 by Thawornwattana *et al.* "Inferring the direction of introgression using genomic sequence data"

Simon Aeschbacher

21 May 2023

Summary

Thawornwattana and coauthors have revised their manuscript on inferring the direction of pulse-introgression between two populations according to the comments of three reviewers. I have read the revised manuscript and the point-by-point answers by the authors to the reviewers' comments.

The authors have dealt with most of my comments to a satisfactory extent. However, I still think that the Discussion needs substantial editorial revision to improve the reading flow, accommodate the majority of the MBE readers, and include explicit a mentioning of limitations and assumptions of the approach. I still disagree with the author's use of nomenclature and abbreviations for some models, and I noticed a few comments that the authors did not seem to have addressed as requested. For a detailed list of my comments, see below ("Re. authors' response to major concerns") and ("Re. authors' response to minor comments").

As part of their revisions (major change iii according to the authors's response), and as far as I can judge in response to main comment 3) by reviewer 2, the authors have now introduced a major subsection "Unidentifiability of introgression models" right at the beginning of the Discussion. I found this decision unfortunate for two reasons. First, this new subsection is far too long and too technical. The reviewers and the editor seemed to agree on a strong need to make this already technical manuscript understandable to a majority of the MBE readers. Instead, the authors now dive into classifying different types of incompatibilities and devote far too much attention to a topic that is subordinate to the main story.

Second, I was surprised that the authors placed this subsection at the beginning of the Discussion, where it is guaranteed to deter many MBE readers and thus prevent them from reaching the concise summary and an interpretation of the main results that one would typically expect at the beginning of an Discussion. In short, I strongly request i) a reduction to a minimally complex discussion of the problem of non-identifiability, and ii) that this part be moved away from the beginning of the Discussion to a place after the summary of the main theoretical results, but still before the discussion of the *Heliconius* results.

Finally, I spotted a few more minor issues, which I list at the very end of my review.

Re. authors' response to major concerns

1. Resolved.

2. Original comment: The Discussion seems to be structured in a somewhat unconventional way – the results of the application to empirical data are discussed before the main results obtained from the theoretical analyses – and the transitions between the subsections are not yet worked out well. The Discussion also misses i) a concise summary of the main results early on, ii) an account of the assumptions and limitations of the framework and approach used, and iii) a perspective on future research. Last, but not least, the Discussion ends flat in my view. I recommend that the authors end on a somewhat more detailed and specific statement of the implications that the study has. > *Authors' response: We have now added a few sentences in the last subsection to summarize our theoretical results and also edited the ending paragraph to have a more positive tone.* » The authors only partially addressed my concern. The authors added a more positive ending and inserted a summary of the theoretical results. I specifically asked for i) a concise summary of the main results early on (not at the end of the Discussion), ii) an account of the assumptions and limitations (still missing), and iii) a perspective of future research (vague to missing). My overall concern that the Discussion is structured in an unconventional way has become even stronger now that the authors inserted at the beginning of their Discussion a very technical subsection "Unidentifiability of introgression models". As explained in more detail below, I consider this new subsection problematic for two reasons, and I think the authors should now really make an effort to revise their Discussion so that it is well suited for the generic readership of MBE. The Discussion now starts very technically instead of restating the purpose of the study and providing a summary of main results and their implications.
3. The acronyms used by the authors to denote the demographic scenarios are confusing and inconsistent with previous literature that introduced the scenarios. Specifically, it is unclear to me why the "MSC-with-introgression" model should be abbreviated as "MSci", given that "MSC" (and not "MSc") is the abbreviation for "Multi Species Coalescent" apparently recognised by the authors. Second, Yu et al. (2012) already introduced the acronym "MSNC" for the exact same model as the one the authors consider here. In short, I recommend that the authors use "MSNC" instead of "MSci" for better congruence with the previous literature and to aim for a parsimonious use of terms in the literature. > *When preparing Flouri et al. (2020), we made an effort to follow the convention in the existing literature, but the problem was that many terminologies already existed at the time. In the end we chose to use the terms that appeared most sensible to us. Degnan (2018) encourages model specifications that explicitly identify the biological factors being considered, and we have followed his suggestion. We used "multispecies coalescent with introgression" or MSci, following the terminology "MSC + hybridisation" of Degnan (2018, p.3) and "coalescent with hybridization" (Blischak et al. 2018). These terms emphasize the two biological processes in the model: coalescent and introgression. We did not like the term 'network' as it had been used to describe a variety of processes, some not being even biological. Degnan (2018) and Solis-Lemus et al. (2017) have discussed the confusions caused by the use of the term "network", such as the distinction between "implicit networks" and "explicit networks". Similarly, we used the term "introgression probability" as in Long (1991) and Martin and Jiggins (2018). This parameter has been referred to as the 'inheritance probability' (Yu et al. 2014) and 'heritability' (Solis- Lemus and Ane 2016). These are most unfortunate: "inheritance probability" concerns the transfer of genetic material from parents to offspring in pedigree analysis, whereas "heritability" is a fundamental concept in quantitative genetics. In short, we did not invent new terms but made a choice among existing terms.* » I do not consider the debate about the terminology related to the abbreviation "MSci" settled. First, "MSci" is a very unintuitive abbreviation for either "multispecies coalescent with introgression" as well as "coalescent with hybridisation". Degnan's (2018) suggestion to use model specifications that explicitly identify the biological factors being considered sounds

good in principle, but if we were serious about this suggestion, we wound need to abandon the term "coalescence". What I am arguing for here is to stick to a term that has been established (multi-species network coalescent) and to use an abbreviation that is intuitive (MSNC). I consider the arguments related to the use of "network" distracting and I did not refer to "introgression probability" at all. I suppose the disagreement between the authors and me will persist, and I thus suggest the editor takes a final decision.

4. Resolved.
5. Resolved.
6. In the application to the *Heliconius* data, and as the authors briefly mention (p. 11, l.11–15), estimates of some parameters (including θ_C, θ_M, τ_s, and τ_c) appear as outliers for a subset of chromosomes (chromosomes 5, 10, 13, 15, 19, and/or 20, depending on the parameter). It also seems as if these estimates are in most cases stronger outliers for the noncoding data as compared for the respective coding data. I found it particularly striking that τ_c . I missed an attempt at explaining this variation, and an effort to check if the outliers might be due to a technical artefact or if they reflect a biological signal. Related to this point, did the authors observe variation in introgression rate between chromosomes? Was this variation consistent with previous results on adaptive introgression (wing pattern loci) and reduced effective gene flow? > *We can confidently rule out technical or numerical errors. Each analysis was conducted using 10 independent MCMC runs, and the pattern is seen in multiple models. We have added the following text (p.10, left column): "A likely explanation is that some individuals are partially inbred, with large variations in heterozygosity across chromosomes."* » I appreciate the authors' effort to rule out technical or numerical errors as an explanation for variation in parameter estimates between *Heliconius* chromosomes. However, besides a technical error, there is the possibility that the observed variation reflects some important biological variation. I sill miss any attempt by the authors to appreciate this possibility and to discuss potential mechanisms. It is known that there are two "categories" of chromosomes in *Heliconius*, short ones with higher per-base pair recombination rates, and long ones with lower per-base pair recombination rates. Variation in recombination rate could explain why "effective" neutral parameters such as the effective population size and the effective introgression probability vary truly among chromosomes. This variation might be picked up by BPP, and I think it can only strengthen the manuscript if the authors discuss this possibility.

Re. authors' response to minor comments

C: comment; **Q:** question; **S:** suggestion; **R:** request.

I will only mention the minor comments to which I would like to add something to the author's response. I kept page and line numbers to refer to the original text for consistency, even if the respective positions may now have been moved due to other revisions.

Introduction

- [. . .]
- [p.1R, l.42–47] **R:** Use "MSNC" instead of "MSci" throughout the manuscript, and simplify the sentence here accordingly. > *Authors' response: "MSci" has been used in many papers now, so we prefer to keep it for consistency.* » I do not think that repeated use of a "bad" abbreviation fixes the issue of a bad abbreviation. As mentioned above, I think it is the responsibility of the editor to make a final decision.

- [p.1R, l.50–54] **R:** Use "IM model" instead of "MSC-M model" throughout the manuscript, do not introduce "MSC-M" here, and simplify the sentence here accordingly. > *We have considered this suggestion and decided to keep both terminologies, as both IM and MSC-M are commonly used. Please see our comment above about terminology.* » I did not check the literature extensively to search for "MSC-M", but I have been in the field for long enough to have encountered "IM" (for "isolation with migration") many times. I think I have clearly stated my opinion and I again appeal to the editor to make a final decision.
- [. . .]
- [p.3L, l.30–31] **S:** Insert ". . . and constant population sizes" after ". . . no gene flow". > *Added as suggested.* » I could not find the change. If I did not miss it, I ask the authors to add the change.
- [. . .]

Discussion

- [. . .]
- [p.12L, l.11] **C/S:** I wonder why the authors first discuss the Heliconius results rather than the outcome of their original work. It would seem preferable to swap the order of content, i.e. start with a discussion of the theoretical results and then follow with the Heliconius data. This later order would be consistent with the order of content in the Results, and it would avoid a sharp break here in the Discussion. > *Please see our response to major comment #4.* » My original minor comment was related to major comment 2, which as described above is still not satisfactorily addressed.
- [. . .]

Figures

- Fig. 1
  - [l.20, caption] **S:** I suggest to state that the direction of the arrows indicates the direction of introgression *forward* in time. > *Added as suggested.* » The authors inserted the sentence "The arrow points to introgression direction in the real world (forward in time)." I find "in the real world" unnecessary and suggest the authors write "The arrow points to introgression direction in forward in time." for maximum clarity and brevity.

Tables

- Table 2
  - **S:** I suggest to report the Bayes factors from thermodynamic integration on the logarithmic scale (e.g. report 1087.1 instead of $e^{1087.1}$. I further suggest to horizontally align the numerical values in a given column by the decimal point. > *We prefer the non-log scale to make the values comparable to those from another approach (Savage-Dickey density ratio).* » I the non-log scale is so difficult to read that it is also difficult to compare the values to the Savage-Dickey density ratios. Also, the authors did not address my point about the alignment of the numerical values. Both pints ultimately concern the journal's typesetting rules, so I consider my contribution as a reviewer done.

Additional minor issues spotted after the first round of review

C: comment; **Q:** question; **S:** suggestion; **R:** request.

In the following, I refer to page and line numbers in the revised manuscript (the one with changes by the authors marked in red).

Introduction

- [l.3] **R**: Replace "While it has . . . " by "While gene flow has . . . ".

Results

- [l.154] **S:** Insert "the" after "in terms of".
- [l.914] **R:** Replace "we analyzed . . . " by "We analyzed . . . ".
- [l.988] **C/R:** I dislike the authors's use of "tower" to describe the posterior surface here and at other places in the new subsection "Unidentifiability of introgression models" in the Discussion. I request that the authors use "peak" or "mode", as either of these two terms will be much more readily understood by the readers of MBE.

Discussion

- [l.1054] **Q:** I did not understand why the authors introduce another abbreviation (BDI) for the model of bidirectional introgression that they previously introduced as model B.
- [l.1098–1092] **C/R:** I do not think that the sentence "No systematic studies have examined the frequency of unidirectional versus bidirectional gene flow given that two species are involved in introgression or hybridization" given the existing previous work on single-pulse admixture between sister taxa. See example recent works by Hahn and Hibbins.
- [1119–1121] **C:** The statement "Most introgressed alleles are 1120 expected to be purged in the recipient species because of incompatibilities with the host genomic background." seems to unilateral to me. We simply do not yet know – and there is ample disagreement about – the extent to which genetic incompatibilities between genes and genes, or between genes and the genetic background are responsible for variation in effective migration rate along the genome. An alternative explanation is that immigrating alleles are locally maladapted to the recipient species' environment. I thus suggest that the authors rephrase their sentence to a more balanced statement.
- [l.1177-1186] **C/R** I disagree that the (parameter) estimates are in general consistent among chromosomes and between coding and noncoding data. There seems to be strong differences between chromosomes and between coding and noncoding data. I also do not see how these differences would be explained by the fact that some individuals were from inbred lines. This fact may explain differences among individuals or populations, as it is a genome-wide effect on average. However, variation in estimates *along* the genome require alternative explanations, including variation in recombination rate and/or the strength of different types of selection. As mentioned above, I miss these biological mechanisms as potential explanations for the within-genome variation in parameter estimates that the authors found.
- [l.1196] **R:** Replace ". . . comparison had the . . . " by ". . . comparison have the . . . ".
- [l.1200] **R:** Replace "excludes . . . " by "exclude . . . " (CIs seems to be plural).
- [l.1203] **S:** Replace "test" by "tests".
- [l.1205–1280] **S:** To address my request above that the Discussion be restructured, I suggest that this subsection ("Inferring the direction of gene flow using genomic data") be the first one after a short introductory paragraph to the Discussion that reminds the reader of the purpose of the study and re-states the main results.
- [l.1220–1221] ***S:** Insert "our" before "Bayesian test of gene flow . . . ".

- [l.1238ff] **C:** In my view, here is where the authors could start with a small part on the issue of identifiability, the relation of their work to the literature, and the limitations of their approach.
- [l.1277] **S:** Replace "... offer ample chances ..." by "... thus there is ample room ...".
- [l.1317] **R:** Replace "consists" by "consisted".
- [l.1368] **R:** Replace "are" by "were".
- [l.1426] **R:** Replace "... the Z chromosome 21 for which ..." by "... the Z chromosome (chromosome 21), for which ...".
- [.1450] **R:** Replace "... is used to ..." by "... was used to ...".
- [l.1452–l.1454] **C:** This sentence is not yet complete. Suggestion: "We used the Bayes factor $B_{10} = \frac{M_1}{M_0}$, where ...".

Figures

- Figure 5, first line of caption: **R:** Replace "simulatie" by "simulate".

Tables

- Table 1, row (e), column 3 (Notes): **C/S:** The clause "... but there is no such deficit or in the true model I or in the data." is unclear. Perhaps replace by "... but there is no such deficit, nor in the true model I nor in the data."?
- Table 1, footnote: Insert "the" after "the behavior of". Insert "the" before "analysis of sequence data".