Inferring the direction of introgression using genomic sequence data

Sent
15-May-2023
From
EiC.MBE@gmail.com, EAssist.MBE@gmail.com
To
z.yang@ucl.ac.uk
CC
EiC.MBE@gmail.com, EAssist.MBE@gmail.com
Subject
Editorial Decision to (major) Revise MBE-23-0129
Body
15-May-2023

MS: MBE-23-0129
Title: Inferring the direction of introgression using genomic sequence data

Dear Prof. Yang,

The in-depth review of your manuscript by the editors and the peer reviewers is now complete. Based on their assessment, it is clear that your manuscript requires a substantial revision before it can be considered further for publication in MBE. Comments from the Associate Editor and external reviewers are included below.

We invite you to revise your manuscript within 60 days and submit it for further consideration. A delayed submission will be treated as a new submission. If you need an extension, please contact Tom Whitehead by e-mail (EAssist.MBE@gmail.com) before the deadline.

Please note that all articles accepted for publication will be published as Open Access and subject to payment of an Article Processing Charge (APC). Find out more on our Open Access page here:
https://academic.oup.com/mbe/pages/Open_Access

If you are planning to promote your manuscript in the press, please contact the Editorial Office (EAssist.MBE@gmail.com) as soon as possible so we can take appropriate action.

Most importantly, the revised manuscript will be subject to editorial and external reviews, and its eventual acceptance depends on the reviewers' and editors' enthusiasm. Note that manuscripts invited to be revised are accepted at a very high rate. But, manuscripts deemed to require more than one major revision are often rejected. So, you must revise it to satisfy all editorial and reviewer concerns. A re-review by original and new reviewers may raise additional concerns, so anticipate them in advance and change thoroughly.

You can create a revision by using the URL: *** PLEASE NOTE: This is a two-step process. After clicking on the link, you will be directed to a webpage to confirm. ***

*** PLEASE NOTE: This is a two-step process. After clicking on the link, you will be directed to a webpage to confirm. ***

LINK REMOVED

Following is a checklist of needed actions and files.

[#1] Number and respond to each of the comments of the reviewers and the Editor. You must enter these comments in the Authors' Response section on the website when you submit the revision.

[#2] Please follow the Author Guidelines at: (https://academic.oup.com/mbe/pages/General_Instructions In the File Upload section of the submission website, please provide the following:

[#3] Upload a file in DOC/DOCX format, RTF format, or LaTeX format containing the text, tables and figure legends. If you are submitting a LaTex/Tex file, please include accompanying .tex files (e.g., .bib, .sty, .cls,.bst).

[#4] Upload files containing high-resolution figures – this does not apply to any supplementary figures. If you have color figures, please confirm which ones are to be published in color in print. If you have color figures that are to be published color online only, please ensure the figure legends do not refer to color. Images are required as high-resolution files (600-1200 dpi for line drawings and 350 dpi for color and half-tone artwork). Most formats are acceptable, including eps, tiff, pdf, jpg, and Powerpoint.

[#5] Upload a copy of the manuscript in PDF (for “Advance Access”) that contains the complete manuscript, including text, any tables, figure legends, and figures, but excluding any supplementary material.

[#6] Upload any supplementary information files. There is no limit on supplementary files. All supplementary information, including supplementary tables and supplementary figure legends, must be uploaded separately as "supplementary files" and not placed in the manuscript. (https://academic.oup.com/mbe/pages/Supplementary_Information).

[#7] Double-check the format of citations and references to ensure that they meet MBE requirements (https://academic.oup.com/mbe/pages/General_Author_Guidelines#References). Failure to do so will result in a delay in manuscript processing.

[#8] Double-check the format for revisions. Articles, Letters, Brief Communications, Perspectives, Reviews, and Protocols all have different format requirements.
Articles must have headings in order: Introduction, Results, Discussion, and Materials and Methods. Results and Discussion may be combined. Subsections are allowed throughout.
Brief Communications must be in the resources category.
(https://academic.oup.com/mbe/pages/manuscript_types)

[#9] To ensure brevity, please make use of unlimited supplementary files to keep manuscripts below 20

pages using 1.5 line spacing for the main text (excluding references, figure legends, or tables). If the manuscript is longer, please justify the extent in response to review.

[#10] Your original files will be available to you when you upload your revised manuscript. Please delete any redundant files before completing the submission.

We look forward to receiving a revised version of the manuscript for further consideration.

Sincerely,

Board of Editors
Molecular Biology and Evolution

Associate Editor
Editors' comments to the author:
Dear Prof. Yang,

I was requested to handle your appeal to MBE after the previous version of the manuscript was rejected. I was asked to consider it as a new manuscript (I did not handle the previous submission). Three new expert reviewers agreed to review it and provided valuable comments, which I am sure you can address. I am sorry for the long time it took. It was not trivial to recruit reviewers and the paper is not a trivial reviewing task...

I think the main common theme is that for the work to have bigger impact, more explanations for readers outside the immediate field (such as myself) are needed. In other words, it is OK for the paper to be much longer, if it makes it easier to read.

I look forward to receiving your revised version.
Tal Pupko

Reviewer: 1
Comments to the Author
I worry that the paper will become too long for mbe if an extensive example is included.

Reviewer: 2
Comments to the Author
The manuscript of Thawornwattana and colleagues addresses the statistical power and impact of model misspecification on inferences to detect, date and quantify introgression (gene flow) between species. This is an interesting study on properties of multi-species coalescent with introgression, combining analytical results, simulations and application to genomic data from Heliconius butterflies, which became a speciation genetics model system to study the impact of gene flow. The authors consider general models of divergence with introgression with 2, 3 and 4 species, which approximate well scenarios that researchers might face when working with genomic data from natural populations. I enjoyed the fact that the analysis of natural populations is framed in the context of the theoretical results, illustrating the link between theory and data. The main conclusions are that population genomics data allow to detect

introgression, even when misspecifying the direction of introgression, and that it is possible to infer introgression proportions in models allowing for gene flow in both directions. Another important conclusion is that the power to detect and quantify introgression depends on the effective sizes of ancestral populations. The authors provide an explanation for their results in terms of times of coalescent in a simple two population model, which provides an intuitive and clear interpretation of their simulation and Heliconius results even for models with more populations. The methods and results are sound, and I am convinced that the results of this manuscript are relevant for a wide audience working on speciation, introgression and hybridization, including theoreticians and empiricists.

I have read the revised version of the manuscript, as well as the response to the editor and reviewers. The authors addressed most of the comments and concerns of reviewers. However, my main concern is that the manuscript is still difficult to read and potentially confusing in some parts. Although the manuscript improved, I still partially agree with the concerns of reviewer#1.

Main comments:
1) The manuscript is written in a very technical way, which I think will be difficult to follow by many readers that are interested in the general questions addressed in the manuscript. Thus, I partially agree with the comments of reviewer #1 in a previous version, and I think that the readability of the manuscript might not be appropriate for the wide audience of MBE and fits more with specific journals (not TBE for the reasons given by the authors). For instance, in the response to the editor and reviewers the authors suggest they will “We will probably include in the Discussion section a Q-A table summarizing our main results”, but this was not done in this revised version. I think that the authors should try to highlight the main results and explanations in the main text, moving several of the details into supplement, keeping in the main the key theoretical results that provide clear an intuitive explanations.

2) Model assumptions regarding effective sizes and main conclusions: The authors discuss that the main impact of model misspecification is on the estimates of the introgression probabilities (phi_Y, phi_X), and ancestral effective sizes (theta_Y and theta_X), since those are expected to differ between models and determine the fit to the distribution of coalescent times for a pair of lineages (t_aa, t_ab and t_bb). In Table 1 the authors present a summary of their main results for two population models, which seem to capture very well what happens in more complex models with three and four populations. I think it is important to note that the analysis of data under the wrong model (e.g., real data from model I analyzed with model O) will be affected by assumptions about changes in effective size. Under model O, the only way to fit the distribution of coalescent times for two lineages sampled in population B is to fit a different effective size (theta) during the time interval between the split and admixture events (tau_X to tau_R). If the model O would assume the same theta, the distribution of t_bb would not not be possible to fit assuming a constant effective size. I think this is a general point that should be addressed, as results indicate that expansions or collapses in Ne could be due to misspecification of the model.

3) Identifiability of introgression probability parameters in model with bidirectional gene flow (model B) for two populations. The simulation results indicate that under model B it is possible to estimate very well asymmetries in introgression. However, in the main text (page 3, lines 35-44, second column) the authors concluded that under the conditions used in simulations (theta_X=theta_Y) those parameters are not identifiable. Is this related with the MCMC mixture, with priors, or with the choice of introgression proportion? Indeed, all analyses were done using a single value of introgression of 0.2. Could that explain

why in this case the admixture in both directions is identifiable? I think it would be important to further extend on the ability to infer introgression in both directions, as it is also one of the conclusions from the analysis of Heliconius dataset.

Minor comments:

Reference to other approaches to derive the distributions of coalescent times: I agree with a comment of reviewer#1 in the previous version regarding the lack of papers in the topic from other authors on general models of isolation with migration and secondary contact (e.g., Lohse and Frantz 2014 Genetics, Costa and Wilkinson-Herbots 2021 TPB and references therein). Although those studies focused on deriving probabilities of the number of pairwise differences assuming the infinite sites model, they also derive the distribution of coalescent times (Lohse and Frantz 2014 using generating functions, and Costa and Wilkinson-Herbots 2021 using Markov chain theory). I think that the authors should also point the readers to those other approaches to find the distribution of coalescent times of a pair of lineages sampled either within or between populations.

Line 55, page 3, left column: Consider indicating explicitly that t_aa, t_ab and t_bb correspond to a sample size of 2, e.g., "(…) we examine the coalescent times of a sample size of 2 lineages (t_aa, t_ab, t_bb) as important summaries (…)".

Line 10, page 3, right column: The result of non-identifiability is only valid if all the populations have the same effective size. It is important to mention that explicitly.

Lines 15-21, page 5, left column: This sentence was unclear to me, and I think more explanation about the identifiability of phi_Y under model B would be required (main comment 3). Do you mean that the older tau_X the less lineages reach node Y, and hence the harder it is to estimate phi_Y? Why would that indicate that there are no differences in information content between model I and B? The rationale for this sentence is later expanded in the manuscript, but in a different context, and not to explain the identifiability of phi_Y and phi_X under model B. Is this because the identifiability of theta_X and theta_Y parameters (related to t_aa and t_bb) allow to correctly distinguish phi_Y from a phi_X of zero?

Line 15-21, page 5, left column. At several places of the manuscript you use "the ease with which one can tell the parental path taken by each sequence at the hybridization node." I think this sentence is vague and confusing, and it would be better to clarify its meaning the first time it is mentioned. I think it would be clearer if you frame it in terms of coalescent probabilities, as this reflects the probability that lineages from B migrate into A and coalesce during the time interval from tau_X to tau_R (or between tau_X and tau_S for models with 3 species).

Lines 22-25, page 5, left column: Could it be that you reach identifiable phi_Y and phi_X in model B, and similar to those estimated in model I for phi_Y, due to MCMC mixing issues in model B, such that only a region of the parameter space is explored? The results in Figure 3 indicate that that is unlikely, as there are no major differences between the posteriors obtained for the 100 datasets.

Lines 15-16, page 6, left column. Typo in table number? The information in Table 3 does not refer to the

simulations.

Lines 58-60, page 12, left column. The authors mention that their approach is better than other methods, as several parameters are estimated besides the migration rates. However, many of the previous studies mentioned use methods (e.g., IMa2) that also estimate effective sizes and times of divergence. Thus, for me, this sentence might be misleading, leading readers to think that the other methods do not estimate effective sizes and times of divergence. What the authors wrote is valid for methods based on tree frequencies (e.g. Martin et al 2019), but not for other (e.g. Lohse et al 2016, Kronforst et al. 2013).

Line 28, page 13, left column: For model d, the relative times are defined in terms of theta_1 and not theta_0, as the text indicates.

Lines 50-55, page 13, left column: What is the effect of choice of priors? Why using prior distributions with mean close to the true value? When analysing data from natural populations this is unlikely to be the case (i.e., a user might specify a prior with a mean far from the value with higher posterior probability).

Line 6, page 14, right column: I think there is a typo and where it is written P(H1|X)~1% it should be P(H0|X)~1%. This is because a Bayes factor B10=100 means that the posterior probability of P(H1|X) is 100 times larger than the posterior of P(H0|X), i.e. P(H1|X)=100*P(H0|X). Thus, if P(H1|X) is larger than 0.99 then P(H0|X)~0.01.

Line 47, page 17, left column: I think it would increase clarity to explain the mid expression in A2, if you mention that coalescent between tau_X and tau_R is only possible if both lineages do not migrate and stay in pop B (with probability (1-phi)^2), or if both lineages migrate to pop A (with probability phi^2).

Reviewer: 3
Comments to the Author
Please find my review in the attached PDF document.
Attached Files
49187000-File000005-1231476716.docx