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

Simon Aeschbacher

13 May 2023

Summary

Thawornwattana and coauthors investigate the extent to which the direction of introgression between two focal populations can be inferred from genome-wide polymorphism data. The authors derive theoretical predictions under the multi-species network coalescent with a pulse of introgression to delineate parameter combinations under which the timing and direction of gene flow can be identified. Using the existing BPP inference framework (Flouri et al. 2018) and simulated data, the authors explore the effects of model misspecification, unequal sizes of the donor and recipient populations, and of the addition of a third and fourth non-focal population as source of information. One key finding of this theoretical work is that inference under a model of bidirectional introgression is more robust to mis-inference than inference under the 'wrong' unidirectional model, albeit at a computational cost. To illustrate the inference of the timing, extent, and direction of introgression with BPP, the authors apply BPP to the butterfly sister species *Heliconius melpomene* and *H. cydno*, a well-studied model system of speciation with gene flow. In line with previous studies, the authors find that more gene flow occurred from *M. cydno* to *M. melpomene* than in the opposite direction. Overall, this study contributes significantly to our understanding of how and when genome-scale polymorphism data can inform about the extent and direction of gene flow. This is an important contribution, even though the study can only explore a moderate set of models and scenarios. I thus think the paper should ultimately be published in MBE. However, I am concerned about the length of the manuscript (redundancy of results), a conceptual mismatch between the theory and the application to the data, a lack of biological interpretation of the *Heliconius* results, and the content and organisation of the Discussion. I am confident that all of these concerns can be addressed in another round of revision.

Major concerns

1. The authors consider four multi-species coalescent models with a pulse of introgression (models I, O, and B in Fig. 1a-c) and the multi-species coalescent model without gene flow (model Ø in Fig. 1d). For models I, O and B, the authors chose a parameterisation that allows the sizes of the two populations A and B to change at the time of admixture ($\tau_X = \tau_Y$) from θ_A to θ_X and θ_B to θ_Y, respectively. In model Ø, however, the parameterisation does not allow for such a change. Therefore, model Ø is degenerate in two ways compared to the other models: there is no introgression, *and* there cannot be a change in population size. The second degeneration seem irrelevant to the theoretical part of the study, because the focus there is either i) on the effect of a misspecification of the *direction* of introgression conditioning on introgression and on

$\theta_A = \theta_X$ and $\theta_B = \theta_Y$; or ii) on a test for the mere presence of gene flow where model Ø is per definition the null model. However, in the application to the *Heliconius* data, the authors do no longer constrain the inference on $\theta_A = \theta_X$ and $\theta_B = \theta_Y$, but model Ø remains constrained. Dropping the constraints in models I, O, and B adds two degrees of freedom to the inference and may introduce additional confounding between population sizes (θs) and introgression rates (ϕs). Thus, the application reaches beyond the scope explored by the theory. I am not asking for an extension of the theoretical work, because I think this would bee too much for this paper. However, I request that the authors discuss the concern I raised in the Discussion as a potential limitation of their study.

2. 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.
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.
4. I am concerned about the length of manuscript, especially the Results section. Large parts on pages 7 to 9 elaborate on results from the three- and four-population models that are analogous to the respective results obtained under the two-population model. I suggest that a detailed account of the results for the three- and four-population model be deferred to a Supplementary Text. The main text could contain just a brief summary of the overall correspondence between the results obtained under the models with different numbers of populations. Together, the the results from the three- and four-population model in my view should not absorb more than a single page in the main text.
5. In the application to the *Heliconius* data, the differences between coding and noncoding DNA in the estimates of the mutation-scaled split times and population sizes are striking (see Figure S7). The difference is particularly large for the time to the root, τ_r, with estimates for noncoding DNA generally at least two-fold higher than estimates for coding DNA. The authors explain this difference by a reduced (effective) neutral mutation rate in coding regions due to purifying selection. Can the authors verify this explanation by comparing the differences in respective parameters between coding and noncoding DNA found in previous studies of the same two focal *Heliconius* species?
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?

Minor comments

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

Abstract

- No comments.

Introduction

- [p.1L, l.39–43] **R:** The Introduction starts with a view of gene flow that is biased towards its facilitating role in speciation. The emphasis of gene flow as a facilitator of diversification and adaptation is a comparatively recent one, and it is much less well supported than the longer established view of gene flow as hindering diversification and (local) adaptation. In short, I encourage the authors to rephrase "Gene flow between species is an important evolutionary process that can facilitate species diversification and adaptation" to "Gene flow between species is an important evolutionary process that can both facilitate [appropriate references] as well as limit [appropriate references] species diversification and adaptation".
- [p.1L, l.43] **R:** "It occurs . . . " → "Gene flow occurs . . . "
- [p.1, l.48–49] **Q:** Do all three references refer to all three factors (mate choice, ecological selection, hybrid incompatibility)? If not, please distribute the references accordingly.
- [p.1L, l.52–53] **S:** Insert comma after "species" and replace ", being . . . " → ", i.e. . . . "
- [p.1R, l.49] **S:** Omit "or γ" to avoid confusion.
- [p.1R, l.42–47] **R:** Use "MSNC" instead of "MSci" throughout the manuscript, and simplify the sentence here accordingly.
- [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.
- [p.2L, l.7] **S:** Specify the reference to Westram et al. (2022) as a reference to a review, e.g. "(see Westram et al. 2022 for a review). Otherwise, cite original papers on the concept of effective gene flow, e.g. Barton & Bengtsson (1986) Heredity, or Petry (1983) Theor. Popul. Biol.
- [p.2L, l.18] **R:** Insert ", at least in principle," after ". . . and thus", to defuse the contradiction between the statement that multilocus sequence alignments are informative about the direction of gene flow on the one hand, and the counter example given just below (lines 21–27).
- [p.2L, l.28–29] **S:** Please back this sentence with a reference.
- [p.2L, l.30–33] **S:** The authors may want to cite Hibbins & Hahn (2019) Genetics, doi: 10.1534/genetics.118.301831, and Hibbins & Hahn (2021) Genetics, doi: https://doi.org/10.1093/genetics/iyab220.
- [p.2L, l.33–34] **C:** The transition between these two paragraphs, and thus from describing background to stating the questions addressed in this paper, seems to abrupt. I suggest to work out the gap of knowledge addressed by the paper somewhat more precisely. At least, the authors should *first* mention that they will now state the questions addressed in the paper, and then state the questions – not the other way round as they currently do (cf. l.48–49).
- [p.2L, l.34] **S:** Please introduce the notation "A → B" at this point.
- [p.2L, l.50–51] **S:** Omit the parentheses surrounding "in particular, the introgression probability", and insert commas after ". . . of parameters" and after "(Flouri et al. 2020)".
- [p.2L, l.52] **R:** Insert "a" before "Bayesian".
- [p.2L, l.54] **S:** Omit "Bayesian".

- [p.2L, l.56] **R:** Insert "the" before "D statistic".
- [p.2L, l.55 to p.2R, p.10] **C/S:** I find the switch here back to providing more background and more on the gap of knowledge addressed in the paper after the statement of the questions unfortunate. I thus suggest to place this content before the statement of the questions.
- [p.2R, l.7–10] **R:** This sentence is incomplete; please formulate a full sentence.
- [p.2R, l.17–22] **C:** It seems as if the authors here provide a second list of objectives on top of the objectives (questions) stated earlier on (p.2L, l.35–48). This apparent second set of objectives is confusing. I suggest the authors be more clear about the core objectives of the paper (inference of the extent, timing, and direction of gene flow), and that they differentiate the core objectives from lower-level objectives (number of species included in the analysis, inflow from sister vs. non-sister species).
- [p.2R, l.29–31] **C:** The claim "to demonstrate the feasibility of inferring the direction of gene flow, as well as its timing and strength" by application to empirical data in the absence of information about the ground truth is problematic. I think this is not feasible. I suggest to rephrase the sentence so that the purpose of the application to the *Heliconius* data becomes a bit more modest (realistic), e.g. "..., and to demonstrate how the framework can be applied to inferring the direction of gene flow, as well as its timing and strength.".

Results

- [p.2R, l.47–50] **S:** To increase clarity: "Divergence time is defined as $\tau = T\mu$, where T is the divergence time in generations and μ is the mutation rate per site per generation." → "Throughout, we measure time in multiples of the expected waiting time to a mutation, and thus define the scaled divergence time τ as $\tau = T\mu$, where T is the divergence time in generations and μ is the mutation rate per site per generation."
- [p.3L, l.9] **R:** I find N an unfortunate choice for the number of sites, as it is easily confused with the much more common meaning of population size. Please use another symbol, e.g. "S", instead of N.
- [p.3L, l.13] **S:** I think it would be unwise to coin this assumption as the standard. Thus, please replace "We make the standard assumption of ..." → "We assume ...".
- [p.3L, l.30–31] **S:** Insert "... and constant population sizes" after "... no gene flow".
- [p.3L, l.32–38] **C/S:** This paragraph summarised background information that, according to my taste, should be moved to the Introduction (if it needs to be included at all). At this current position, the paragraph seems to divert from the story line.
- [p.3R, l.35–38] **S:** I suggest to rephrase this part as follows to increase clarity: "Similarly, introgression probabilities ϕ_X and ϕ_Y are also identifiable with only one sequence per species per locus under model B in general. However, in the special case of $\theta_X = \theta_Y$, we have ...."
- [p.4L, l.13] **S:** "... values, with $\mathbb{E}(\hat{\Theta}_\mathrm{I}) \approx \Theta_\mathrm{I}$ ..." → "... values, i.e. $\mathbb{E}(\hat{\Theta}_\mathrm{I}) \approx \Theta_\mathrm{I}$ ..."
- [p.4L, l.20] **S:** "In effect, ..." → "That is, ..."
- [p.4R, l.4] **R:** In equation (1), replace the equals (=) sign by an approximate (≈) sign.
- [p.4R, l.16] **S:** To increase clarity, replace "... compared with $\phi_X^* = 0.27$, ..." by "... compared with inferred values of $\phi_X^* = 0.27$, ...".
- [p.5L, l.53–55] **C:** I find this insertion of the coin-tossing analogy unnecessary. I would omit these three lines, and also omit "or coin-tossing setup" in line 48 just above.
- [p.5L, l.56–57] **C:** I agree that if a B sequence coalesces with an A sequence between τ_X and τ_R this means that the B sequence has taken the introgression path. However, the data will not allow to determine with certainty if this event has occurred. In particular, the event cannot be distinguished from coalescence between the A and B sequences after τ_R in the past. Therefore,

the formulation "... it will be clear that ..." is misleading, because it will *not* be clear ("clear" in the sense of certain) from the data. Coalescence of A and B between τ_X and τ_R only contributes probabilistic evidence. I suggest to rephrase this sentence accordingly.

- [p.5R, l.36] **C:** I found the remark that the Fisher information is in this case a 5 x 5 matrix and the reference to eq. A3 not very informative. I suggest the authors either drop the remark on the Fisher information, or alternatively explain i) how it follows from eq. A3 that the Fisher information is a 5 x 5 matrix, and ii) how this contributes to the approximate nature of eq. 2.
- [p.5R, l.42] **S:** "... for estimating ϕ_Y if P_X is greater, ..." → "... for estimating ϕ_Y the closer P_X is to 1, ..."
- [p.6L, l.6–15] **C:** I found the line of argumentation based on the 'factors' 1.36 and 1.61 by which $n_B - c_B$ and P_X differ between scenarios **a** and **b**, respectively, misleading. I found this because it remains unclear to what relative extents $n_B - c_B$ and P_X contribute to the information content about ϕ_Y. I appreciate that the empirical observations support the authors' conclusion, but I suggest the formulation be improved.
- [p.6L, l.44] **S:** Insert "also" before "apply", and insert a comma after "apply.
- [p.6L, l.50–52] **C:** The use of the subscripts X and Y to *theta* seems inconsistent between the text here and Table 1.
- [p.6R, l.30] **Q:** Is "Bayesian test" the name of a method? If so, quote for the first time. Otherwise, rephrase to "... *the* Bayesian [...] by Ji et al. (2022) ..." (i.e. insert "the").
- [p.6R, l.37–38] **S:** To my understanding, there are no "limiting parameter values". Rephrase "... determined by the limiting parameter values when $L \to \infty$ ..." → "... determined by the parameter values in the limit of $L \to \infty$ ...".
- [p.7L, l.36ff] I am concerned about the length of the manuscript. The parts that follow form here on three and four populations seem too long. I suggest that the authors condense these parts to a single page in the main text that summarises the main points. The more elaborate description of the results could be moved to a Supplementary Text. See my respective Major Concern above.
- [p.7L, l.42] **S:** Omit ", creating five scenarios".
- [p.8L, l.51] **C:** I found the formulation "..., on balance, the data ..." unfortunate. I suggest the authors start a new sentence with ". Overall, the data ...".
- [p.7R, l.8] ***S:** "In case where ..." → "For those cases in which ...".
- [p.7R, l.59] **R:** "incur" $rarr; "incurs"
- [p.9L, l.3–18] **C:** The results described in the paragraph by large confirm the results from the two-population model. I do not think that this level of detail is required for the main text here.
- [p.9L, l.54 to p.9R, l.54] **C:** The results described in these paragraphs by large confirm the results from the two-population model. I do not think that this level of detail is required for the main text here.
- [p.9R, l.59 to p.10L, l.5] **C/S:** Assessing the feasibility of an analysis seems meaningless to me unless there is a known truth. I suggest to rephrase this part, e.g. to "To illustrate the application of our results from the asymptotic analysis and simulations to the inference of the direction of gene flow from genomic data, we ...".
- [p.10L, l.30–34] **C/R:** This sentence is too long and suffers from inconsistency between singular and plural grammatical number ("factor" vs "show" vs "is"). Please split the sentence and fix the grammar.
- [p.11L, l.13–14] **Q:** How are these unusually large estimates for the subset of chromosomes mentioned here to be interpreted? Please elaborate on this observation.
- [p.11L, l.15–16] **C/R:** General results for all autosomes are announced here, but I could not find these results below. Please describe the general trends for the autosomes and clearly delineate

this part.

- [p.11L, l.16] **R:** When mentioning chromosome 21 here, please make clear that this is the sex chromosome.
- [p.11L, l.19] **Q/R:** What is meant by "proportionally larger"? Please increase the clarity.
- [p.11L, l.20] **Q/R:** What is being referred to by "S7"? Please clarify.
- [p.11L, l.21] **Q/S:** Do the authors mean "... reduced *effective* neutral mutation rate ..." here?
- [p.11L, l.29] **C/R:** It was unclear to me what is meant by a "local" impact of introgression. Please clarify.
- [p.11R, l.3–11] **C/R:** The authors first state an expectation about the posteriors of ϕ_M and ϕ_C, but they then do not state explicitly if this expectation was met or not in their application to the *Heliconius* data. Please add an explicit statement relating to the expectation.
- [p.11R, l.9] **C/S** "... species t is prudent ..." $rarr; "... species, we recommend .."

Discussion

- [p.11R, l.47–48] **S:** I suggest to rewrite "The *A. arabiensis* → *A. gambiae+A. coluzzii* introgression is ..." as "The introgression from *A. arabiensis* to *A. gambiae* and *A. coluzzii* has been ...".
- [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.
- [p.12L, l.59] **Q/S:** What do the authors mean by 'better' in this context? It seems difficult to qualify the parameter estimates in terms of how close they are to the biological truth. I might be missing what evidence the authors are referring to, and would thus suggest to clarify this point.
- [p.12L, l.60 to p.12R, l.1] **C/S:** The authors seem to suggest that no previous studies provided estimates of species divergence times and population sizes. According to my recollection, however, previous studies did estimate the divergence time and the scaled population sizes of the two species considered here (e.g. Lohse et al. 2016 Genetics, doi: 10.1534/genetics.115.183814, Table 2). I suggest the authors choose a more accurate phrasing.
- [p.12L, l.8–11] **C:** This part of the Discussion starts abrupt, somewhat disconnected from the preceding subsection. The problem may be reduced if the authors follow my earlier suggestion of swapping the subsections "Inferring the direction of gene flow using genomic data" and "Gene flow in *Heliconius* butterflies". However, I had some more generic concerns about the structure and content of the Discussion (see my respective Major Comment above).
- [p.12R, l.29–51] **C/S:** This paragraph seems to have been added in a previous revision of the manuscript. The paragraph appears disconnected from the rest of the Discussion, and to me it seems unfortunate to end a paper with such a technical paragraph.

Materials and Methods

- [p.13L, l.21–22] **Q:** I was wondering why, in scenarios (c) "small to large" and (d) "large to small", the authors did not only vary the size of one of the populations and its ancestral populations, but also changed τ_R and τ_X both in relative and absolute terms compared to scenarios (a) and (b).
- [p.13L, l.42–43] **S:** Rephrase "This is the so-called A00 analysis, with the model fixed (Yang 2015)." → "Note that this setting corresponds to the A00 strategy of Yang (2015)."
- [p.13L, l.48–49] **Q:** The authors chose a rate parameter of β for the priors of τ_R and θ such that the prior mean is close to the true values, with α fixed at 2. I agree that these priors are "diffuse"

as claimed by the authors, but I am concerned by the fact that the mean was approximately matched to the true. Did the authors check how robust their inference is against different choices of priors, in particular priors for which the mean is not approximately equal to the true value?

- [p.13R, l.14] **C:** I found it confusing that the authors refer to five branches on a two-species tree. The authors seem to interpret changes in population size as delineating branches, even though the populations remain the same in terms of their identity. I realise that my point is about a technical detail, but if the authors find a less confusing expression for the branch sections they refer to, it would seem preferable.

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.

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.

Appendix A

[l.1915–1924] **Q:** Were the likelihood functions derived here not already provided collectively by Wilkinson-Herbots (2008 Theor Popul Biol; 2021 Theor Popul Biol) and Costa and Wilkinson-Herbots (2017 Genetics; 2021 Theor Popul Biol)? **R:** Please make clear which results were obtained before and by whom, and which results are novel.

Figures and Tables

- [Figure S4]
  - **S:** In the caption, replace the last sentence "See legend to figure 6." by "Other details as in Figure 6.".
- [Figure S6]
  - **S:** In the caption, replace the last sentence "See legend to figure S5." by "Other details as in Figure S5.".