Yang, Ziheng

From: Molecular Biology and Evolution <onbehalfof@manuscriptcentral.com>
Sent: Monday, January 23, 2023 3:52 PM
To: Yang, Ziheng
Cc: eassist.mbe@gmail.com
Subject: Resubmission of Manuscript - MBE-22-0920

23-Jan-2023

MS: MBE-22-0920 (Article)
Title: Inferring the direction of between-species gene flow using genomic sequence data

Dear Dr. Yang,

The Editorial Board has considered your appeal of their decision to decline publication of this manuscript and has indicated that your manuscript can be resubmitted for further review.

The manuscript should be submitted as a new manuscript online. It will be subjected to initial editorial review, which is followed by in-depth review by external experts for a subset of manuscripts submitted.

You should indicate the previous manuscript number and provide detailed responses to editorial and reviewer comments, along with a statement that you have been advised to resubmit the manuscript by the editorial board. We usually try to assign the same handling editors, subject to availability; the editors will decide whether to invite further external review and whether to invite the same (subject to availability) or new reviewers.

The editors encourage you to focus on the clarity of this manuscript and its accessibility to readers who may be well-versed in the relevant statistics yet find the language of the manuscript difficult to follow. The board believes in the importance of the topic, but recognizes that this importance will not be appreciated by readers who find the manuscript too difficult to read.

Please consult the following URL for MBE Editorial policies:
https://academic.oup.com/mbe/pages/Editorial_Process

Thank you for continuing to consider Molecular Biology and Evolution as a venue for the publication of your best work.

With our best wishes,

Heather Rowe
Managing Editor
From MBE Website

Best regards,

Heather Rowe
From MBE Website

MS: MBE-22-0920
Title: Inferring the direction of between-species gene flow using genomic sequence data

6 January 2023

Dear Editor,

Thanks for your email with the reviewers' comments. We are disappointed about the editorial decision to reject our ms, and would like to appeal against the decision for reasons given below.

It appears that the decision to reject is mostly based on reviewer #1's frustration at the difficulty of reading and understanding our ms. This reviewer seems to get lost very early on. In contrast, reviewer #2 clearly understood our paper and made many insightful suggestions, which we will be able to accommodate easily. Both reviewers pointed out that our paper addresses an important but challenging problem which should interest all researchers who use genomic data to infer interspecific gene flow. For example, reviewer #1 highlights the importance of the problem "given that population genomic inference under these models [is] used to study speciation in an increasing number of taxa".

The difficulty the reviewer and associate editor had with our ms. is chiefly due to conceptual problems we outline here. Our paper is indeed challenging to read, even for the authors. We went through many rounds of writing but the material is still hard to read. We try to make predictions about the simulation results based on our understanding of the model and inference method, and we then compare the simulation results with our predictions. This is harder, for both the authors and the readers, than simply running the simulations and describing results. For this reason, understanding the text requires very careful reading, as well as a good knowledge of statistical theory and the multispecies coalescent model.

Would you give us a chance to re-submit? We hope to improve the readability and bring the results to a wider readership, taking into account reviewer comments as far as possible. Reviewer 2's suggestions are mostly excellent and we will follow them. We will follow some of reviewer 1's suggestions but there are a few that we do not agree with. We would hope that you assign our paper to an AE with statistical expertise who might find a statistically well-trained reviewer if there is concern about the soundness of our work.

We do not agree with the AE's suggestion, citing reviewer #1, that our paper lacks any new method or empirical result and is suitable for TPB. The theoretical analysis and simulations were motivated by counter-intuitive results we observed in our analysis of the genomic data in *Heliconius* butterflies. This work made us realize that our initial expectations about the problem were incorrect. We now believe we understand the model and the inference method much better. Indeed our empirical analysis of the small *Heliconius* case we use here reveals many of the same results obtained in the theoretical analyses and simulations. Our paper should be of broad interest to those who wish to use phylogenomic data to infer the history of species divergence and gene flow, and does not really fit TPB, which mostly publishes theoretical results in population genetics.

Our overarching interest here is to investigate the effects of incorrect assumptions on inference of introgression. Here are some of the main questions explored in the paper. Suppose introgression occurs from species A->B but we analyze genomic data assuming B->A introgression. (1) Do we detect introgression despite the incorrect assumed direction? (2) How does the estimated introgression probability ($\varphi_{B->A}$) compare with the true introgression probability ($\varphi_{A->B}$)? (3) How reliable are estimates of the time of introgression, as well as other parameters such as species divergence times and population sizes? (4) Do the results differ depending on whether gene flow is between sister lineages or between non-sister lineages, and whether gene flow is from a large population to a small one, or in

the opposite direction? (5) How can we infer the direction of introgression (A->B vs. B->A) and are typical genomic data informative about the direction? Perhaps our aims should have been expressed more clearly; but we provide answers to all of those questions and also develop novel strategies for deriving such answers.

We will follow the reviewers' suggestions to provide better orientation for the reader, explicitly stating the biological significance and rationale of the results. We believe that the empirical data analysis should be understandable to empirical biologists involved in analysis of genomic data, and will try to make this section more prominent in the paper. If you have other suggestions, we will be happy to consider and incorporate them.

I include a point-by-point response to the Editor's and reviewers' comments below.

Thanks for your consideration.

Best wishes,

ziheng yang

on behalf of the authors.

From: Molecular Biology and Evolution <onbehalfof@manuscriptcentral.com>
Sent: Tuesday, December 27, 2022 2:52 PM
To: Yang, Ziheng <z.yang@ucl.ac.uk>
Cc: EiC.MBE@gmail.com; EAssist.MBE@gmail.com
Subject: Editorial Decision to Reject MBE-22-0920

27-Dec-2022

MS: MBE-22-0920
Title: Inferring the direction of between-species gene flow using genomic sequence data

Dear Prof. Yang,

Thank you for submitting your manuscript to Molecular Biology and Evolution (MBE). We regret to inform you that it did not receive high enough priority for publication after an in-depth review by the editors and the peer reviewers. Specific comments from the editors and external reviewers are included below.

In general, MBE seeks to publish research, methods, and resources of broad significance in molecular evolutionary biology. Even when the external reviewers find a manuscript to be scientifically and technically sound, the ultimate priority for publication is determined based on the novelty and impact of the work presented. MBE does not publish manuscripts judged by the reviewers to contain mostly descriptive work, confirmatory results, and discoveries with a limited gene and taxonomic scope. All of these factors were considered in deciding the publication priority for your manuscript.

Thank you for considering Molecular Biology and Evolution, and please continue to consider MBE as a venue for the publication of your best work.

Sincerely,

Board of Editors
Molecular Biology and Evolution

Associate Editor
Editors' comments to the author:
Thank you very much for submitting a manuscript with interesting topics.
As the second reviewer commented, the results may be useful for many readers.
However, as the first reviewers commented, since this manuscript contained a lot of technical nature and the absence of any new method or empirical result, TPB might be a better fit.
I agree with this reviewer's comments.

We disagree with those comments. Our paper was motivated by analysis of the genomic data in *Heliconius* butterflies, and our results should interest researchers using genomic data to infer interspecific gene flow. It is not a theoretical population genetic paper. Please see our comments to the Editor above.

In addition to the first reviewer's comments, based on my own reading I found lack of explanation.
Thant is in the first paragraph of Result, line 160~164 on page 3.
The authors measured the coalescence time in unit of expected time to accumulate one mutation per site. Why is the time measured by the accumulation of mutations per site?
An advantage to use this time unit should be clarified here.

Time and rate are confounded in comparative analysis of sequence data. In the notation of our paper, θ = 4$N\mu$ is identifiable, but not the population size N (or the average coalescent waiting time of 2N generations) or the mutation rate μ individually. As a result, time or age on both the species tree and on the gene trees is measured by distance or the expected number of mutations per site. This is the case in both phylogenetics and population genetics.

We will add a sentence to clarify this.

Reviewer: 1
Comments to the Author
Thawornwattana et al use simulations (and analytic results for pairwise coalescence times) to study how likelihood inference of population models involving discrete gene flow are biased by model-mis-specification Understanding the biases (in terms of model selection and parameter estimates) and the identifyability (or not) of parameters and models is an important problem given that population genomic inference under these models used to study speciation in an increasing number of taxa. The authors also re-analyse data from Heliconius to demonstrate that likelihood inference based on a single diploid sample from each taxon agrees with previous estimates. My main comments are:

We thank the reviewer for highlighting the importance of the problems studied in our paper. We agree that more and more researchers working in the field of population genomics will have to consider problems addressed in our paper.

1) I found this manuscript barely human readable (despite the fact that I consider myself an expert in this area and really wanted to understand what Thawornawatttana et al have found). I appreciate that the topic is technical and some amount of mathematical notation and jargon is unavoidable. However, given this, it is all the more important to present things as clearly and succinctly as possible. I detect very little effort in this direction by the authors. The results section is excessively long (10 pages, single spacing!), repetitive in places and contains an unnecessary density of notation and level of detail. Picking a paragraph from the Results at random:

"Case b (same θ short tree) is similar to case a, but the divergence times (τ R , τ X) were half smaller. As in case a, we expect θ̂X(O) < θX(I) , θ̂Y(O) > θY(I) and ϕ̂X > ϕY. Furthermore, we expect ϕ̂X to be larger in case b than in case a. Note that when θY(O) and θX(I) are fixed with θY(O) > θX(I) (or when θ̂Y(O) is similar in the two cases, table S1), the smaller Δτ of case b (than in case a) means a larger ϕ̂X according to eq. 4. We have ϕX∗ ≈ 0.27 and 0.30 for cases a and b respectively (table S1). "

What reader is supposed to be able to parse text such as this, let alone make sense of it? [I have to say I rather agree with the reviewer's comment here; it is a well selected sentence, notwithstanding that it was perhaps better explained in Case a]. More importantly, it is often unclear why the information that is presented, is there, i.e. what question does the text in this (and other) results paragraph address? Reading (or rather trying to read) this MS, I was left unclear about what the main findings are (the one place where they are stated clearly is the Abstract).

We suspect that two reasons may explain the reviewer's frustration here. First our text is too terse and understanding the material requires a good knowledge of statistical inference and coalescent theory. Second the reviewer may lack the knowledge and/or patience to read our ms. carefully. We will make an effort to try to improve the readability of our ms., by providing more explanations to better orient the reader. We also request that our ms. be assigned to a different AE with more statistical expertise who might find a statistically well-trained reviewer.

A substantial rewrite/reorganisation is needed to make this manuscript readable. I suggest:
- condensing the results section to at most half of its current length

- relegating substantial parts of the Results (in particular eqn. 1-3, which are not particularly informative) and most of the figures 3&6 to an appendix/Supplement.
- finding subheadings that summarise the key findings of each result section. The current subheadings (e.g. "Performance under the true model") are rather uninformative.
- instead of exhaustively describing the results of the sensitivity analyses (for all parameters and model comparisons), the Results should highlight the main findings in a more distilled/digestible form. For example, I wonder whether it would make sense to structure the Results by the three main findings as summarized in the abstract:
i) it is easier to infer gene flow from a small population to a large one than in the opposite direction
ii) it is easier to infer gene flow from outgroup species to an ingroup than in the opposite direction.
iii) if introgression is assumed to occur in the wrong direction, the time of introgression tends to be correctly estimated
- it would help to use the standard notation of f for the admixture fraction (see Durand et al 20) (instead of phi).

We will go through our ms. to make editorial changes to make the ms. more concise and more readable. We will move the three equations to an Appendix. We will consider moving the real data analysis section to the front of the paper, as motivating examples, before the sections on theoretical analysis and computer simulation.

Nevertheless, we have many more important results than summarized by the reviewer here (see for example the questions mentioned above in our main comments to the Editor). We will probably include in the Discussion section a Q-A table summarizing our main results.

Regarding notation, we respect the reviewer's opinion. However, there are multiple commonly-used standards. Population geneticists (working on the D-statistic or variants) tend to use f, but phylogeneticists have used γ, for example, in programs HyDe, SNaQ, PhyloNet, PhyloNetworks, etc. We initially used γ when we implemented the model in bpp. However, there are many gamma and inverse-gamma models and priors in the program, such as the gamma and inverse-gamma priors on τ and θ, the gamma prior on the shape parameter of the gamma distribution of rates for loci, the gamma prior on the shape parameter of the gamma distribution of rates among sites, and so on, and having a new gamma parameter was extremely confusing when we continued to work on bpp. We then changed to a different symbol. Note that φ sounds similar to f and looks similar to γ.

2) While the authors cite most of their own work on the topic, papers by others are omitted. This includes two studies that contain more general versions of the mathematical results (eqn 1-3) described in this paper:
- Lohse and Frantz 2014 (Genetics) give the coalescence time distributions for a sample of 3 under an MSci model, these include f(t_ab), i.e. eqn 3 (denoted f(t_a) in Lohse and Frantz)
- the MSci model for a pair of populations the authors consider seems to be a special case of the generalised isolation with migration model which has been studied in depth by Costa & Wilkinson Herbots 2021 (TPB): the results for the distribution of coalescence times (1-3) can be obtained from the secondary contact model considered in Costa & Wilkinson Herbots 2021 by taking the limit of tau_1 ->0 (and rescaling M).

We did not find an equation in Lohse and Frantz (2014) that resembles equations 1-3, but we will add relevant citations.

3) what is the (biological) rationale for allowing for a change in population size that coincides exactly with the time of admixture? More importantly, is this level of model complexity relevant for the main question of the paper? Would the Results be easier to describe (without loss of information) if

simulations were limited to a simpler model with three theta parameters, i.e. theta_X = theta_A & theta_Y = theta_B?

Population size varies a lot even among closely related species, and also over time, so assuming different population sizes for different branches on the species tree adds biological realism. However in this case the MSci model was initially implemented with a separate theta for every branch on the species tree, and the simpler model mentioned by the reviewer is only recently added in the program. We do not believe the theory will be simpler under the simpler model.

4) I am assuming (but perhaps I have missed it) that the simulations do not include recombination within loci. Given that for most organisms rho ~ theta, it seems important to consider the bias induced by recombination in this study.

Yes, the MSci model studied in our paper assumes no recombination among sites of the same locus. A recent paper (Zhu et al., 2022 Mol. Ecol. 31: 2814-2829, DOI: 10.1111/mec.16433) conducted simulations to examine the impact of within-locus recombination on estimation of parameters in MSci, and found that the estimates are robust to recombination at rates up to 10x the human rate (see figure 6 in the paper). Overall other factors such as the number of loci, the number of sequences sampled per species, and the mutation rate, are far more important than the recombination rate. We will add a brief discussion of the effects of recombination on our analysis.

Reviewer: 2
Comments to the Author
In this manuscript Thawornwattana et al. investigate the effect of mis-specification of the directionality of gene flow on the inference of population parameters and introgression presence/strength in 2-4 species MSci models. They discuss theoretical expectations and compare them to inferences based on a Bayesian method implemented in the program BPP applied to simulated data under different scenarios and to a Heliconius case study.

Overall opinion
The paper addresses a relevant problem, i.e. inferring the direction of gene flow, which has proven challenging even with genomic data. Although methods exist to infer the direction of gene flow, e.g. rank-ordering of local genomic divergence estimates (Fig. S39 in Green et al. 2010) or the DFOIL method based on D-statistics (Pease & Hahn 2015), these often require a specific sampling setup and cannot accommodate a case of two sister species where introgressed haplotypes have permeated throughout the entire species. The current study achieves this by leveraging information from probability densities of coalescent times under different gene flow scenarios. The theoretical expectations match well with the results from simulated and real data.
The authors adequately explore the limits of their modelling approach, describing how the model parameters, including timing, populations sizes and strength of introgression are affected under various scenarios. Specifically, the case of a misspecified direction of introgression receives due attention. One main result is that the problem of misspecification in unidirectional models does not seem to be a big caveat, because it appears that the bidirectional model performs well to identify the direction in all test cases, even if it comes at a higher computational cost.

We thank the reviewer for an accurate summary of our results.

Major/General comment:

In the simulated data, the difference in phi between the true value and any single replicate can be considerable. To get close to the true value, an average and confidence interval of a 100 replicates are needed. Yet for the real Heliconius data, only two replicates are considered, one coding and one non-coding. Indeed, their resulting phi estimates do not match well, for essentially unknown reasons. This

may simply result from stochasticity inherent in the sequence data. Of the 347 Mbp available in a Heliconius genome, the dataset covers only ~5 Mbp, assuming (4942 + 5341 loci) * 500 bp. There is ample opportunity to create replicates form the real data by making additional subsets of the genome, and it will be interesting to see how 95% HPD CIs for phi and other parameters for real data match up with those of simulated data.

Yes, estimates of the rate of gene flow or the introgression probability often have large CIs in small or moderate-sized datasets, and to get precise estimates, thousands of loci are often needed. Ancient introgression involving ancestral species is in particular hard to infer. The Bayesian method provides the CI as a measure of confidence, so the uncertainty can be assessed even if one dataset is analysed. Each of the two Heliconius datasets (coding and noncoding) has about 5000 loci, so the estimates are fairly previse.
We will follow the reviewer's suggestion and analyze datasets from the other chromosomes and present the results in the SI.

I think a critical assumption (in all of these MSC models) is that there is no recombination within loci, which will in practise rarely be true and difficult to know. While the statement "τX is largely determined by the smallest coalescent time t_ab" is true, in practise, if loci are not non-recombining, the smallest t_ab might never be found, even with large amounts of data. I think this should be discussed, citing studies which investigate the impact of this assumption.

We will add a brief discussion of the effects of recombination. This is mentioned by reviewer #1 as well. Please see our response above.

Other comments:

Reviewer #2's minor comments are mostly excellent. We will make changes as requested. Below are responses to a few comments that need clarifications.

Title: Consider 'interspecific gene flow' or simply 'introgression' as alternative to 'between-species gene flow'.

Abstract: "We found that it is easier to infer gene flow from a small population to a large one than in the opposite direction, and easier to infer inflow (gene flow from outgroup species to an ingroup species) than outflow (gene flow from an ingroup species to an outgroup species)." - As explained in line 579 to end of paragraph, it is also easier to infer gene flow when the time between initial divergence and subsequent introgression is larger. This is not focused on as much in the paper, yet might be worth mentioning more explicitly.

Abstract: "We discuss factors that cause gene flow to be asymmetrical, including geography, behavior, and incompatibility of introgressed alleles with the host genomic background." - These factors are only briefly touched on, in the first section of the discussion, and do not constitute a main feature of the paper. Consider dropping the sentence from the abstract.

16: "Gene flow is thus intrinsically asymmetrical, being more likely in one direction than in the other." I know what the authors want to say here, but I think this is not the right way to say it. Gene flow is not "intrinsically asymmetric". Without any additional information there is intrinsic directionality, so the expectation is that gene flow is symmetric. That said, the variance around this expectation might be large, so any realisation might be more likely directional rather than symmetric. A possible alternative wording could be something like: "Given that these factors likely differ between species and that drift on introgressed material acts independently in different recipient species it is sensible to assume that gene flow is in most cases asymmetric".

Also, one point that could be added to intro or discussion when discussing asymmetry: Geographic context might be an important reason why gene flow is often not symmetric. For example, if a smaller subpopulation of one species A comes in contact with another larger census size species B (e.g. through migration) this subpopulation of A might be fully absorbed in B, making the introgression "intrinsically" (this time for real!) asymmetric.

25: “Two types of models” - as elaborated on in lines 31 and 38, these represent a difference in the mode of introgression (instantaneous pulse vs continuous gene flow, see e.g. Hibbins & Hahn (2021) already cited in this manuscript). Consider naming this concept more explicitly.

32-33 and 140-142: "rate is measured by the probability which is the proportion" I know what the authors want to say here, but rate, probability, and proportion are as the authors are surely aware different things. First there is no "rate" in a pulse-model and second a probability (in a stochastic model) is not the same as a proportion (in a realistation of the model). Please reword this.

42: First part of equation should be “MAB” with m capitalized.

46 and 1240: "\phi is an 'effective rate' that reflects the combined effects of gene flow, natural selection and genetic drift". I don't understand how \phi reflects genetic drift. First, drift on population level is explicitly accounted for in the model in the \thetas. Second, randomness due to small introgressed "sub-population" could change the variance in \phi in realisation of the underlying model, but not the expectation. Therefore, I don't understand how \phi is "effective" with respect to genetic drift. (With respect to natural selection is makes of course sense.)

This is an insightful comment. We take the reviewer’s point and will fix our text.

126: “Our results provide practical guidelines for inferring introgression from genomic sequence data.” The follow-up to this is found in the last section of the discussion. The authors chose to open this section with a discussion about model misspecification. It might help the reader here if instead, the authors would discuss the steps of inferring introgression in practical order, something like (1) Bayesian test of introgression, (2) choosing the appropriate model for BPP, (3) running BPP, (4) potential troubleshooting, including model misspecification and the option to apply the bidirectional model.

141: “introgression” misspelled

Fig. 1: Consider including tauR and tauX in the schematic.

168: “…with n sites in each sequence.” but on line 1364: “N = 500 sites”.

206f: if the underlying model is a Wright-Fisher model, these equations only holds for sufficiently large Ne. This should be clarified.

Fig. 2: The figure caption is confusing, especially the first sentence. I needed to take a good look at table S1 to understand what is going on here. The explanation needs to be clearer. First, the legend "I model"/"O model" is not clear. I first assumed that this designates the true underlying model, but this is only true for the first case. In the second case the underlying model is still the I model, it is only the inference that is based on the O model. That needs to be made clearer.

Also in the description it is confusing that two things change at the same time between black solid and red dashed: (1) The former is true, the latter inferred; (2) The latter uses the O model for inference instead of the true I model. The authors need to guide the reader more.

Furthermore, for the captions of the subfigures, e.g., "small to large" it is not immediately clear what the "to" refers to. The authors mean introgression from a small to a large population. This is explained in the text around line 300 but the figure is first referred to much earlier. When looking at the figure the reader could also interpret it as a temporal change, e.g., \theta_A>\theta_X (or other way round depending on whether they think time-backwards or forwards).

Finally, consider including annotations in the legend for the blue and green vertical dotted lines.

300: Consider clarifying "…(c) small to large theta, and (d) large to small theta". (As above for Fig. 2c and 2d.)

305-312: There are two sentences that are partly redundant and should be merged. One starting with "The true distribution fI(tbb) is discontinuous at τX and τR", the other one with "The true distribution fI(tbb) is discontinuous at τR"

smallest coalescent time between sequences from the two species (tab): in practise very noisy.

354: I think (\tau_X, \tau_Y) should read (\tau_X, \tau_R).

363: Correct "…depending as…"

416: I cannot follow the conjecture: "As ϕX∗ > 0 according to our analysis, model O is a 'less wrong' model than model 0". Do the authors mean that it is less wrong because it infers at least "some" gene flow even if it is the wrong direction? This seems to be in contrast to the statement a few lines below "rejecting the null and accepting model O may be considered a false positive error. In this paper, we use the second interpretation.".

491: remove the word "assumed"; it is misleading as one think this is assumed by the inference model, but the opposite is true.

Our text is correct here. The inference model, B (for bidirectional introgression), assumes both A->B and B->A introgression, whilst the true model I has A->B introgression only. The B->A introgression assumed in the inference model B is nonexistent.

494: here suddenly the notation \phi_A→B is used, but before this was called \phi_Y. (I do find the former clearer, but either way the use should be consistent)

535: "increasing the number of sampled sequences (nB) is less effective than increasing the number of sequences reaching node Y, which is in turn less effective than increasing the number of loci (L)" I understand what the authors are trying to say here, but I think it needs to be reformulated, because "increasing nB" is precisely a way of "increasing the number of sequences reaching node Y". Better would be: "increasing the number of sampled sequences (nB) is less effective than decreasing the pairwise coalescence probability of sequences before reaching node Y"

540: it is not immediately clear what the "first" is that this "second" refers to. (I was actually missing this second factor when reading the preceeding paragraph, so it should be made clear in the preceeding paragraph that only one factor is considered.)

559: "if Px is greater, or if the branch length 2/θ(τR−τX) is greater" The "or" does not make any sense here because given the definition of Px the two statements are precisely equivalent. I think the authors might mean "or if the amount of data L is greater".

The reviewer is correct that the two conditions are equivalent. We used “or” to mean “in other words” or “put in another way”. We will rephrase.

567 onward: phiY in the “short tree” vs “long tree” scenario appears to be affected antagonistically by parameters tauY and tauB. It could be worth exploring scenarios were these are varied in isolation, to better assess the relative effect of each.

The reviewer seems to mean thetaB by tauB (since tauB = 0). The number of coalescent events in population B or the number of B sequences reaching the hybridisation node Y has a distribution given by nB, the number of sequences sampled from B, and 2*tauY/thetaB, the age of the species in coalescent units. It is true that the effects of tauY and thetaB are in opposite directions, but the process is simple and well-characterized, given by the single-population coalescent theory.

586: “(eq. 6)” is missing the closing bracket.

589: "If we use the same population size θB in cases a&b, the number of sequences reaching Y will be the same, and the performance differences between the two cases will be even greater." I don't understand this sentence. θB is the same in cases a&b, therefore the number of sequences reaching Y should NOT be the same.

600: Consider clarifying: “…have a higher chance of coalescING with other sequences in population X, BEFORE tauR.”

627: “…were smaller by half.”

755: Does it add very little information, or none at all? Consider clarifying why.

The added information is “very little” but nonzero. Adding sequences from an outgroup species should help with the estimation of the species divergence time (the age of the root of the species tree), which in turn should help the estimation of other parameters in the model. As a thought experiment, estimation of introgression probability should be more precise if the true species divergence time is given.

808: “Different population sizes” with population in singular instead of plural.

Fig. 5: The trees seem very deep. At the introgression time 1.5\theta most lineages within A and B will already have coalesced and lineages surviving and migrating will likely coalesce between X and S. It would be good to clarify this in the text and discuss the implications.

1248: "This reasoning appears to suggest that by norm" that is a rather clumsy construction

1277: and --> into