

MS: MBE-22-0920

Title: Inferring the direction of between-species gene flow using genomic sequence data

11 February 2023

Dear Editor,

Thanks for giving us a chance to revise and resubmit the ms. I include a copy of the email correspondence concerning our appeal.

We have now revised the ms. to improve its readability. We have mostly followed our plans in our appeal letter. Here is a summary of our major changes. Our point-to-point responses follow.

We have rewritten Introduction to better explain the biological significance of our results and the objectives of the paper. We have added some 'orientational' text to explain our objectives or to summarize the results at the beginning of several sections.

We followed reviewer 1's suggestion and moved the coalescent-time densities into Appendix A (p.17). We added the distributions under the B model at the end of the appendix, and used them to explain the unidentifiability issues under the B model when there is only one sequence per species (p.3). We encountered such unidentifiability issues in our analysis of *Heliconius* genomic data from the Z chromosome (p.11).

We have three main chunks of results, and we have rewritten the text to highlight them.

First, we have developed a theory to "predict" parameter estimates under model O with the introgression direction is misspecified. We added a new table (table 1) to summarize our arguments, which are described in the section "Asymptotic analysis and best-fitting parameter values" (p.3-4). The summary table is later referred to when we discuss the simulation results. Inclusion of the table has allowed us to shorten the text and we think it has also helped to improve clarity of the ms. as well.

Second, we characterized the information content concerning the introgression probability when the introgression direction is correctly specified, in eq. 3. We have rewritten the section "Information content for estimating introgression probability under the true model" on p.5. We used a coin tossing analogy to explain the theory and hope this is clearer than before.

Third, we characterized the behavior of the Bayesian test of introgression when the introgression direction is misspecified. We have restructured the section "Bayesian test of introgression: power and false positives" (p.6-7) discusses both the implications of the best-fitting parameter values to the asymptotic behavior of test and the simulation results. (These were in two sections in the submitted version.)

In response to reviewer 2's suggestion of examining more genomic data, we have now analyzed the coding and noncoding data from all 21 chromosomes (whereas the submitted version used chromosome 1 only). We described and discussed the results for chromosome 1 in detail and then commented on the results for other chromosomes. See p.10-11.

I include a pdf with changes in response to review comments highlighted in red, to make it easier to read this cover letter. As mentioned above, we have made many changes, so those summarized above are not highlighted in the file.

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

Thanks for your time in handling our ms.

Best wishes,

ziheng yang

on behalf of the authors.

From: Molecular Biology and Evolution <onbehalf@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, dealing with a technical problem in population genetics. Please see our summary of our main results above.

In addition to the first reviewer's comments, based on my own reading I found lack of explanation. That 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, $\theta = 4N\mu$ 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 identifiability (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 Thawornawattana 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 $\hat{\theta}_X(O) < \theta_X(I)$, $\hat{\theta}_Y(O) > \theta_Y(I)$ and $\hat{\phi}_X > \phi_Y$. Furthermore, we expect $\hat{\phi}_X$ to be larger in case b than in case a. Note that when $\theta_Y(O)$ and $\theta_X(I)$ are fixed with $\theta_Y(O) > \theta_X(I)$ (or when $\hat{\theta}_Y(O)$ is similar in the two cases, table S1), the smaller $\Delta\tau$ of case b (than in case a) means a larger $\hat{\phi}_X$ according to eq. 4. We have $\phi_X^* \approx 0.27$ 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).

As summarized above in our comments to the Editor, we have made major restructuring and rewritten multiple sections to improve the presentation and readability of our paper. We hope that the paper is now comprehensible to any serious reader who is familiar with statistical inference and the coalescent model.

We hope that our real data analysis can be understood by empirical biologists who may wish to use our program to analyze their own genomic data.

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 ϕ).

We have restructured and edited our paper to improve its readability. We moved the equations into an appendix. Nevertheless, we have many more important results than summarized by the reviewer here. Please see our summary in our comments to the Editor.

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, so that it was very confusing to have a new gamma parameter when we continued to work on bpp. We then changed to ϕ : this sounds like f and looks like 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 $\tau_1 \rightarrow 0$ (and rescaling M).

We did not find a relevant equation in Lohse and Frantz (2014). We have cited Yu et al. (2014). The coalescent time densities are special cases of the gene-tree density under the model.

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 should add 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 constrained 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 $\rho \sim \theta$, it seems important to consider the bias induced by recombination in this study.

Yes, the reviewer is right that the MSci model studied in our paper assumes no recombination among sites of the same locus. We have added a citation to a recent paper (Zhu et al., 2022), which discussed the issue extensively and conducted simulations to examine the impact of within-locus recombination on estimation of parameters in MSci (p.3). 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.

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 ϕ 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 ϕ 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 $(4942 + 5341 \text{ loci}) * 500 \text{ bp}$. There is ample opportunity to create replicates from the real data by making additional subsets of the genome, and it will be interesting to see how 95% HPD CIs for ϕ and other parameters for real data match up with those of simulated data.

Yes, estimates of 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 *Heliconius* datasets has more than 2000 loci, so the estimates are fairly precise.

In response to the reviewer's suggestion, we have now analyzed coding and noncoding data from all chromosomes from the *Heliconius*. The results are summarized in SI figure S7, tables S5 and S6. Please see also our comments to the Editor.

I think a critical assumption (in all of these MSC models) is that there is no recombination within loci, which will in practice rarely be true and difficult to know. While the statement " t_X is largely determined by the smallest coalescent time t_{ab} " is true, in practice, 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 have now included a citation to Zhu et al. (2022), which examined the impact of within-locus recombination on various inference problems under MSci (p.3). That paper also included an extensive discussion of the issue.

This is mentioned by reviewer #1 as well. Please see also our response above.

Other comments:

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

We followed the suggestion and deleted "between-species".

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.

Changed as suggested.

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.

Deleted as suggested.

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".

Changed as suggested.

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.

Changed as suggested (p.12, top).

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.

This is now lost during editing.

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 realisation of the model). Please reword this.

We changed "rate" to "magnitude".

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

Fixed.

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

This is an insightful comment. Changed as suggested.

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.

Inferring or selecting the model of introgression is currently a challenging task, and also an active research area. We have added a short paragraph at the end of Discussion to comment on the options available to an empirical biologist (p.12).

141: "introgression" misspelled

Fixed.

Fig. 1: Consider including τ_R and τ_X in the schematic.

Changed as suggested.

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

We have changed n to N to be consistent.

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

The reviewer is right, but this is so well-known that we do not think we need to mention it, especially given that we are short of space. It is also well-known in the population genetics community that the coalescent provides surprisingly good approximations to the fisher-wright process even when the population size is fairly small (in the order of a few hundreds for example).

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.

We have rewritten the figure legend to improve the clarity. We added explanatory text for each of the four cases. We now mention that the vertical dotted lines indicate discontinuity points at τ_X and τ_R .

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

Changed as suggested.

305-312: There are two sentences that are partly redundant and should be merged. One starting with "The true distribution $f_l(t_{bb})$ is discontinuous at τ_X and τ_R ", the other one with "The true distribution $f_l(t_{bb})$ is discontinuous at τ_R "

Changed as suggested.

smallest coalescent time between sequences from the two species (t_{ab}): in practise very noisy.

354: I think (τ_X, τ_Y) should read (τ_X, τ_R) .

Fixed.

363: Correct "...depending as..."

Fixed.

416: I cannot follow the conjecture: "As $\phi_{X*} > 0$ according to our analysis, model O is a 'less wrong' model than model O". 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."

This section is about Bayesian asymptotics, and 'less wrong' means 'smaller K-L divergence from the true model'. We have added "judged by the Kullback-Leibler divergence" (p.6).

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 $\phi_{A \rightarrow B}$ is used, but before this was called ϕ_Y . (I do find the former clearer, but either way the use should be consistent)

We have made changes to 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"

The text is lost during editing.

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.)

This part of the text is now rewritten. The two factors are
(i) the number of *B* sequences reaching *Y* and
(ii) the ease with which one can tell the parental path taken by each *B* sequence at *Y*.
Please see p.5.

559: "if P_x is greater, or if the branch length $2/\theta(\tau_R - \tau_X)$ is greater" The "or" does not make any sense here because given the definition of P_x 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 now say "or in other words ..."

567 onward: ϕ_Y in the "short tree" vs "long tree" scenario appears to be affected antagonistically by parameters τ_Y and τ_B . 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 θ_B by τ_B (since $\tau_B = 0$). The number of coalescent events in population B or the number of B sequences reaching the hybridisation node Y has a distribution given by n_B , the number of sequences sampled from B, and $2\tau_Y/\theta_B$, the age of the species in coalescent units. This is actually analytically tractable even though the equations are complex. It is true that the effects of τ_Y and θ_B are in opposite directions, but the process is well-characterized. We have now cited Tavaré (1984) and Wakeley (2009) (p.5), and also included a new figure (figure S4).

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

Fixed.

589: "If we use the same population size θ_B 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. θ_B is the same in cases a&b, therefore the number of sequences reaching Y should NOT be the same.

We could not understand this text either. It appears to be text inserted by mistake. It is now deleted. Thanks to the reviewer.

600: Consider clarifying: "...have a higher chance of coalescing with other sequences in population X, BEFORE τ_R ."

Yes, we changed this to "a higher chance of coalescence during (τ_X, τ_R) ."

627: "...were smaller by half."

This text is lost during editing.

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 improve if the true species divergence time is given.

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

Fixed.

Fig. 5: The trees seem very deep. At the introgression time 1.5θ 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.

If we treat the tips as different species, the trees we used are not very deep. If we treat them as populations of the same species, the trees are indeed deep. There are huge variations among biological systems, and scenarios we evaluated are not atypical.

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

This is now lost during editing.

1277: and --> into

Fixed.