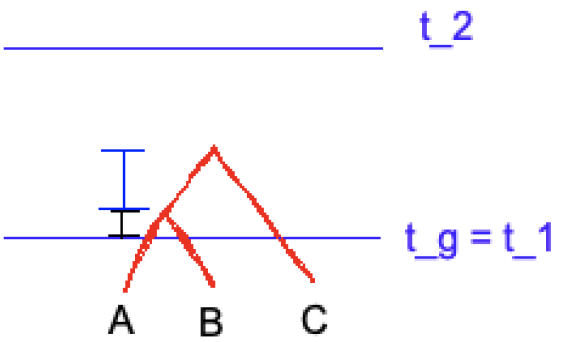
Review for “Phylogenetic conflicts: distinguishing gene flow from incomplete lineage sorting” by Galtier

The manuscript titled “Phylogenetic conflict: distinguishing gene flow from incomplete lineage sorting” by Galtier describes a new method, called Aphid, to distinguish gene flow (GF) from incomplete lineage sorting (ILS), while estimating population parameters for the speciation process of three species. The method uses gene trees and a maximum likelihood approach to fit the model, and then predicts whether gene trees are affected by ILS, GF or neither using posterior decoding. Unlike previous methods to detect GF (such as the ABBA-BABA test), Aphid is able to detect instances of multi-directional GF and GF right after the speciation event between the two most closely related species, which might not produce discordant gene trees, but rather gene trees concordant with the species tree that have short branch lengths. The author validates the model with simulations and later applies Aphid to exon trees of different mammal species.

I believe the manuscript is written very clearly, and the motivation, theoretical framework, and empirical data analysis are all easy to follow. The code is also straightforward to install, and all of the examples can be run without problems. I have some major comments, mainly with regards to some of the approximations and assumptions of the model.

Major comments:

- The model, as described in section 2.1, ignores polymorphism due to gene flow. Aphid assumes that coalescent events happen exactly at times $t_g=t_1$ (or $t_g=t_1/2$), t_g being the time when a GF event happens. However, coalescence between lineages that are co-segregating due to GF happens deeper in time than the actual hybridization event, more specifically 2Ne generations deeper on average. Moreover, due to this delay in the times for coalescence, there might be cases where the two lineages co-segregating due to GF fail to coalesce backwards in time, and thus are now co-segregating with the remaining lineage. For example, in scenario 7, if lineages A and C fail to coalesce before reaching t2, then they will be co-segregating together with B, thus creating cases of ILS due to GF. Another example is scenario 6, in which, if lineages A and C fail to coalesce before t1, then there can be cases where A and B (the scenario depicted below), or B and C (corresponding to scenario 8) find common ancestry, since all three lineages will be co-segregating. This creates an additional scenario:



Moreover, this also shifts all of the probabilities of the model. I believe these issues should be addressed by recalculating the probabilities and branch lengths, and, perhaps, it will fix some of the biases observed in fig. 2.

- I have some comments regarding simulations:
    - In the simulation procedure described in section 3, migration happens at a constant rate between all populations. However, I believe that Aphid should also be tested in the case of unidirectional GF, especially because in fig. 3 you detect discordant topology imbalance in macaca.
    - In order to test whether the bias you observe in fig. 2A is because of simulations having GF older than t1 as claimed in lines 214-218, you could simulate gene trees lacking GF between the ancestral AB species and C, and see whether you recover unbiased estimates.
    - You could also simulate in a more pulse-like manner instead of having a constant migration rate, which would be more similar to the model proposed in Aphid and might perform better.

  All these scenarios might be a bit more challenging to simulate with your custom script, but can be simulated using msprime.
- I believe that the molecular clock assumption is too strict, especially because the mutation rate for each of the lineages can vary quite a bit (see, as an example in primates, Moorjani et al. 2016 https://doi.org/10.1073/pnas.160037411). Instead of removing gene trees based on the clock-likeness, one could model the mutation rate per species separately, as you propose in the discussion section. Given that your model is very fast and efficient, I do not think that adding these parameters will influence the speed, while adding gene trees that were previously filtered out might improve the estimation of GF through posterior decoding.

Minor comments:

- Line 21: the word “genes” might be misinterpreted as only coding regions of the genome. I would use something like “trees reconstructed from different genomic locations, also known as gene trees”.
- Line 45: distinctive → different.
- Line 76: wrong format for reference J and MR 2013.
- Legend of fig. 1 (and in some other places throughout the text): “divergence times” in the manuscript refers to the speciation or split times, i.e., the times in which different species begin to be isolated. While “divergence” is often defined this way in the literature, in many other cases it refers to the average time of coalescent rather than the actual split times (see, for example, Prado-Martinez et al. 2013, https://doi.org/10.1038/nature12228), which are on average 2Ne deeper than split

times. I suggest that you explicitly define what “divergence” means in your manuscript.

- Line 213: is there a reason why you do not simulate discordant topologies over 50%? In many cases with rapid radiation (such as in birds, see Suh et al. 2015 https://doi.org/10.1371/journal.pbio.1002224), the discordant gene tree proportion can easily reach 2/3, and ILS and GF are much more difficult to distinguish in these cases. I believe that if your method performs well in such extreme cases, it can be useful to solve long-standing phylogenetic conflicts due to ILS and ancient GF.
- Line 220: the reason for the bias might also be that coalescent times in Aphid are assumed to be average coalescent times, instead of modeling the whole range of coalescent time values. Such approximation is similar to that in CoalHMM (Dutheil et al. 2009), in which the authors indicate that the similar bias that they observe in their model might be due to coalescent events being modeled as single time points.
- It would be useful to have a confusion matrix for the results in lines 234-246, containing all three categories (GF, ILS, no-event).
- Line 262: there is no other mention of the additional datasets analyzed by Aphid anywhere else in the manuscript other than this reference to supplementary table 2. Maybe mention them later in the text or include them in fig. 3?
- Throughout the text, you use the word “we” instead of the singular “I”, even though there is only a single author.
- The p_a proportions in supplementary table 2 are mostly above 0.9, suggesting that most of the GF happened around the last speciation event. Given also that most of the asymmetry indices for GF are ~0.5, the model seems to capture isolation with migration right after the speciation event for the analyzed taxa, such as the one modeled by Mailund et al. 2012 (https://doi.org/10.1371/journal.pgen.1003125), but with the additional advantage that you model GF between three species.
- I believe the conclusion is unnecessary, you already summarize the method in the discussion section.