A. Recommender's comments

This preprint introduces “Aphid,” a new statistical method that estimates the contributions of gene flow (GF) and incomplete lineage sorting (ILS) to phylogenetic conflict in estimated gene genealogies. All three reviews (including my own, which is below) were positive, and I anticipate a positive decision after revision.

Aphid is based on the observation that GF tends to make gene genealogies shorter, whereas ILS makes them longer. Rather than fitting the full likelihood, it models the distribution of gene genealogies as a mixture of several canonical gene genealogies in which coalescence times are set equal to their expectations under different models. This simplification makes Aphid far faster than competing methods. In addition, it deals gracefully with bidirectional gene flow—an impossibility under competing models. Because of these advantages, Aphid represents an important addition to the toolkit of evolutionary genetics.

In addition, this preprint argues that about half of the phylogenetic conflict in the human- chimp-gorilla clade results from gene flow. This is a substantive finding that is both new and important.

Thanks for the positive comments, much appreciated

The reviews make a number of helpful points. I will not try to summarize these but will merely highlight a few that seem particularly important:
1. My own review (comments on lines 97–99) points out an error in the formula for expected coalescence time in the “no event” case. It also suggests a different formula, which may eliminate the bias seen in Fig. 2c.
2. Reviewer 1 (comment 1) points out that because the mutation rate is similar to the recombination rate per nucleotide, a genomic region that is large enough to have mutations is likely also to have recombined. It is therefore a problem that Aphid ignores recombination. The simulation results suggest that this problem isn’t serious. Nonetheless, I’d like to see it discussed.
3. Reviewer 2 worries that, because the analysis uses exons only, selection may bias results. He suggests repeating the analysis using intergenic data. This seems like a good idea. If you decide against it, I hope you will justify the restriction to exons.
4. It would be useful to include some measure of goodness of fit in Aphid’s output. Poor fit could result from variation in population size (which Aphid ignores), from misspecification of the phylogeny, or from misspecification of the pattern of gene flow among subdivisions. Consequently, the user needs feedback about goodness of fit.

Thanks so much for the quality of the evaluation. You don’t often benefit from such constructive comments from 4 qualified scientists. I also received spontaneous comments on the preprint from two additional colleagues, which also were quite useful – discussed at the end of this response, section E and F. Before addressing the comments one by one, I’m here giving an overview of the revising work I have done.

There were mainly five categories of comments :
1. clarify/correct various aspects of the method and manuscript
2. discuss method assumptions and conditions of applicability
3. do more simulations
4. analyze non-coding data
5. improve the method in various ways

In short, I have followed ~all the suggestions in category 1-4 ; as far as category 5 is concerned, I mention these as opportunities for future work.

The reasons for this is that what I did first was suggestion 4, i.e., non-coding data analysis in apes. This was highly rewarding : the new analysis makes sense in many respects and, in my opinion, strengthens the claim of ancient gene flow between human, chimpanzee and gorilla, while also providing a more accurate estimation of its prevalence - which turns out to be a bit lower than suggested by exon trees. Given these new results, I decided to put more emphasis on the ape analysis in the revised version (as suggested by two reviewers), which included modifying the title and abstract. For the same reason, I thought it was essential to be convincing and thorough about the reliability and limitations of current version of Aphid, i.e., address comments in category 1-3. I propose to postpone the study of its potential improvements, which are many, entail additional developments + simulations, and will be more appropriately published in a future technical article on the program itself – especially if the method attracts some interest from the community.

Below I address the reviewers' comments, which I have (re)numbered for clarity.

0. It would be useful to include some measure of goodness of fit in Aphid's output. Poor fit could result from variation in population size (which Aphid ignores), from misspecification of the phylogeny, or from misspecification of the pattern of gene flow among subdivisions. Consequently, the user needs feedback about goodness of fit.

This is an excellent suggestion. Indeed I had in mind to prospect in this direction. I've tried something, which was to implement a degenerate model with as many scenarios as gene trees. Calling *n* the number of gene trees in a data set, this model has 5*n* parameters (4 branch lengths and one prior probability per scenario) whose ML estimates are trivially obtained (observed branch lengths and 1/*n*, respectively). Empirically the difference in log-likelihood between this model and the Aphid model seems to be a reasonnable measure of the goodness of fit. One problem, however, is that the Aphid model and the degenerate model are not nested, and we have no *a priori* knowledge of the distribution of the log-likelihood ratio under the null hypothesis that the Aphid model is true. This null distribution might be approched via simulations, but this would be computationnally demanding and increase the running time by two or three orders of magnitude. There is still the possibility of offering this as an option to the user. There also might be more clever ways of assessing the goodness of fit – happy to take suggestions. I have not modified the manuscript based on this comment, and intend to address this issue thoroughly in the future.

1. line 5–7, 58–66: "Gene flow, however, need not be asymmetric, and the ABBA-BABA hypothesis-testing approach is not a proper way of measuring gene flow prevalence". Yes and no. Patterson's D is a test statistic: it tests the hypothesis of zero gene flow. It's often misconstrued as an estimate of the level of gene flow, but that is not the case. So I think you're right that "this hypothesis-testing approach is not a proper way of measuring gene flow prevalence." However, Reich and Patterson also have a variety of other statistics (reviewed here [5]) that really are estimators. These statistics do not make the mistake of treating a test statistic as an estimator. They do however assume one-directional gene flow, as does my own method, Legofit [3, 4]. Thus, you're right to point out that this one-directional assumption is a weakness of existing methods. In summary, you make two valid points. However, they apply to different sets of methods. Your sentence (lines 5–7) and later paragraph (58–66) conflate these different sets.

Thanks for this comment. I modified the text now only mentioning the problem of symmetric vs. asymmetric gene flow, and citing more of Reich's and Patterson's papers. Abstract now says: "*Gene flow, however, need not be asymmetric, and when it is not, ABBA-BABA approaches do not properly measure the prevalence of gene flow*."

2. line 15–18: "Aphid predicts that roughly half of the human/chimpanzee/gorilla phylogenetic conflict is due to ancient gene flow. This also translates in older estimated speciation times and a smaller estimated effective population size in this group, compared to existing analyses assuming no gene flow." This is an important finding and deserves emphasis earlier in the manuscript.

I followed this suggestion, also encouraged by the analysis of non-coding data. This result is now mentioned in the manuscript title and at the end of the introduction.

3. Fig 1: For the case of HGT (do you mean GF?), the figure shows coalescent events at the time (tg) of gene flow. These events should precede tg to account for coalescence time within the donor population.

This is correct, and corresponds to major comment 1 by Reviewer 1. I modified the manuscript to clarify that *tg* in Aphid is the time of coalescence of two lineages brought together by gene flow (and see below response to Reviewer 1).

4. line 82: "We note that not only the internal but also terminal branch lengths differ in expectation under an ILS vs. a GF scenario (Fig.1). Here we introduce a new method, Aphid, aiming at capturing this information." Legofit [3, 4] also makes use of this information, provided that singleton site patterns are included in the analysis. Because legofit relies on site pattern frequencies and does not estimate gene genealogies, it does not need to assume that no recombination happens within blocks of chromosome or that there is free recombination between blocks. On the other hand, the simplifications used by Aphid make it much faster.

Thanks for this comment. Sorry I had missed the Legofit method despite its obvious relevance, reason being that it does not stem from the "ILS vs. GF" phylogenetic literature I've been focusing on. I have now included a section comparing Aphid and Legofit in the discussion, and cite the Legofit papers in the introduction.

5. line 97–99: Coalescence times are assumed to equal their expected values. However, the formula given doesn't reflect the conditioning that is implicit in the "no-event" case. Let x represent the coalescence time for the (A, B) pair of lineages, measured from an origin at time t1. "No-event" is the case in which $x < t2 - t1$. Consequently, the appropriate expectation is $E[x|x < t2 - t1]$. See Rogers and Bohlender [5, Eqn. A1, p. 71] for the appropriate formula. Translating that formula into the notation of the current paper, and setting $z = (t2 - t1)/2Ne$,

$$E[x|x < t2 - t1] = 2Ne\,(1 - z/(\exp(z) - 1))$$

This is substantially smaller than 2Ne if z is small, and this discrepancy probably introduces bias into Aphid. This may be why, in Fig. 2c, estimated ILS is much too low when true ILS is high. High ILS implies small z, and that is when we expect bias. In calculating the formula above, care is needed to avoid numerical error when z is small.

Thanks for a great suggestion, and for sharing the corrected formula. I was aware of this problem, which I did not prioritize because it is expected to mainly concern the concordant, not discordant, gene trees. Yet I take the point about a possible reduction of the bias in parameter estimation.

The corrected formula is great, however, its implementation in Aphid would create complications. This is because in this formula the expected branch length expressed in generations is not a linear function of t_1, t_2 and N_e, such that the expected branch length

expressed in mutations is no longer a function of just $\tau_1=t_1.\mu$, $\tau_2=t_2.\mu$, and $\theta=4N_e\mu$. Using this formula would imply estimating the mutation rate as an independent parameter, not just its product with the other three parameters – presumably based on a tiny amount of information, given the difficulty in distinguishing among scenarios entailing an ((A,B),C) topology.

I've tried something different, which was to constrain the younger coalescence time to be younger than t_2 in the no-event scenario Specifically, length *a* and *b* in scenario no_event were set to min(t_2, $t_1+\theta/2$), and length *d* was adjusted accordingly. This is not an exact solution like the one you're proposing, but kind of goes in the same direction. I did not detect any conspicuous improvement based on simulations. I did not modify the manuscript based on this comment.

6. Eqns. 1–2: Why are some probabilities written as “P ” and others as “P?” In other words, why do some use a caligraphic font?

The caligraphic *P* refers to the probability mass of the Poisson distribution, as indicated in the paragraph below equation 2.

7. Eqn. 1: I presume that P(Sk) is related to the material in lines 116–124, but this relationship is not clear. More detail is needed about P(Sk).

I now refer to the very paragraph in which these prior probabilities are introduced, and to Supplementary Table 1 in which formulas are made explicit.

8. Eqns 6–7: These equations sum over {i : Ti != ((A, B), C)}, which ignores gene trees that lack phylogenetic conflict. Don’t these concordant gene trees carry useful information? Why not sum over everything? I don’t think Ti is ever defined.

Concordant gene trees do carry useful information, and in particular, are presumably key in correctly estimating τ_1 and τ_2 - and in cascade all the other parameters. Equations 6 and 7 aim at addressing the very question I'm asking in this ms, which is the partitioning of the conflict into ILS vs GF. This question indeed disregards gene trees having the species tree topology.

To sum over all topologies could also be meaningful in measuring the overall prevalence of GF and ILS in the data. The reason I'm refraining to do this is that distinguishing between scenarios is a more difficult task as far as ((A,B),C) topologies are concerned – due to the existence of the "intermediate" no_event scenario. Parameter p_{AB}, in particular, has a large sampling variance (see below new simulations), so I'm a bit concerned about the reliability of these estimates. Please note that the detailed output of Aphid provides the posterior probability of every scenario for every gene tree in tabular format, making that summation easy to do.

***T_i* is defined as "the topology of gene tree G_i" immediately after equation 7.**

9. line 171: What is the rationale of the imbalance measures (Eqns. 9–13)? For GF, I suppose that imbalance must mean that gene flow is directional rather than symmetric. What does it mean for ILS? (See point 7 of Reviewer 1.)

Yes, as far as GF is concerned, the imbalance is intended to reflect the asymmetry between A ↔ C and B ↔ C gene flow, in the spirit of ABBA-BABA. As far as ILS is concerned, the Aphid model predicts no imbalance, so this calculation should be seen as a sanity check. If substantial ILS imbalance is detected, that might reflect a failure of Aphid to capture the true

history, or maybe the existence of ancient GF from ghost lineages, as now mentioned in the Discussion (l422-429, see also response to J. Joseph's comments below). I have clarified the meaning of these two indices in the revised version.

10. line 177: Why 1.92 units of log likelihood?

This results from Wilk's theorem – e.g. see section 5.1.4 in:
https://strimmerlab.github.io/publications/lecture-notes/MATH20802/likelihood-based-confidence-interval-and-likelihood-ratio.html
1.92 is half the 0.95 threshold of a chi2 distribution with one degree of freedom.

11. Fig 2c: The legend should distinguish between the horizontal axis (the fraction of ILS) and the fraction of discordant trees (coded as the filled versus open circles).

I clarified that the fraction of discordant topologies equals simulated ILS + simulated GF.

12. line 237–239: I only realized at this point that Aphid is making local estimates for relatively small chromosome segments. I had previously assumed that its goal was global estimates, which aggregate across the whole genome. It would be a good idea to distinguish between these goals in the introduction.

Actually in my analyses only a small fraction of loci did get an posterior annotation with reasonable support – even in simulations. The typical gene tree analyzed here carries too little information to be reliably categorized as ILS, GF or no_event, probably because built based on relatively short segments. Existing data on recombination rates suggest that, at least in mammals, stretches of non-recombining SNPs are not longer than the 100-1000 bp segments we used (see response to Reviewer 2 comment 0). For this reason I am not sure I should emphasize this possibility of posterior annotation of gene trees. It was still useful to be examined based on simulations since this analysis revealed some of the problems faced by Aphid for distinguishing among scenarios under certain conditions, particularly ancient gene flow.

13. Fig. 3: The text says that dot sizes are proportional to the contributions of ILS and GF. But it isn't clear what this means. Does "dot size" refer to the diameter or the area of the dots? Cleveland [1, 2] has shown that we humans are bad at decoding information that is encoded in areas. We're much better at decoding lengths. Yet when circles are used in graphs, information is usually encoded in areas. Consequently, even if you encode the information in the diameters of circles, readers are likely to assume that the area is relevant. I suggest redrawing these graphs and using the lengths of bars or lines to encode these data. Reviewer 2 makes a similar point.

I checked and found that neither the diameter nor the area of the dots I'm drawing are actually proportional to the estimated contributions – I used the arbitray cex=1+20*p formula in R, where p is the proportion to be represented. I reformulated to avoid using the word proportional, and now report the exact estimates in the figure, as suggested by Reviewer 2.

14. line 282: I would say "suggests instead," not "rather suggests." The latter has a different meaning in British English, which may be confused for the one intended here.

Modified as suggested

15. line 283: If you're writing in LATEX, consider using $\leftrightarrow$, which prints as ↔.

Done as suggested

B. Reviewer 1's comments

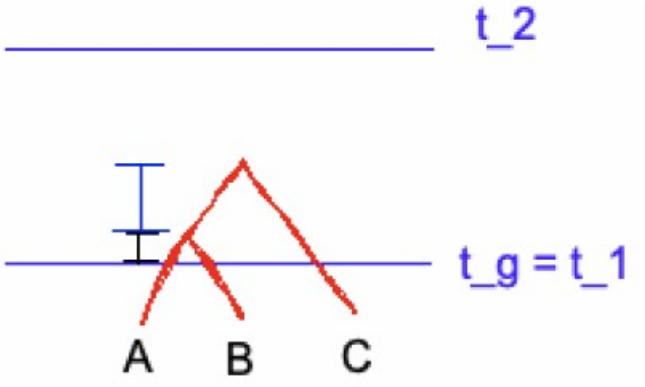
The manuscript titled "Phylogenetic conflict: distinguishing gene flow from incomplete lineage sorting" by Galtier describes a new method, called Aphid, to distinguish gene flow (GF) from incomplete lineage sorting (ILS), while estimating population parameters for the speciation process of three species. 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.

Thanks for the positive comments, much appreciated

Major comments:

1. The model, as described in section 2.1, ignores polymorphism due to gene flow. Aphid assumes that coalescent events happen exactly at times tg=t1 (or tg=t1/2), tg 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.

Thanks for a great comment. Not to mention that, in GF scenarios, coalescence time differs from gene flow time was indeed a weakness of the first version. In addition, what this comment I think highlights, is that GF and ILS are not mutually exclusive, that scenarios more complex than the ones considered by Aphid exist, and that discordances due to relatively ancient GF (occurring shortly after speciation) are likely to be very difficult to distinguish from ILS.

You suggest including additional scenarios and re-defining probabilities. That might be an option to think of – even though I am not 100% sure there really is a chance to distinguish these "GF+ILS" scenarios from simple ILS scenarios. I did not attempt to implement this in the current version of Aphid.

Rather, I did two things:
- I now clearly state that parameter t_g in Aphid is not the GF time but rather the coalescence time of two lineages brought together in the same pop by GF, which on average is 2Ne generations older than GF time. This is explained in the Methods section, and recalled when interpretating data analysis.
- I inserted a paragraph in the Discussion about the problems posed in Aphid by relatively ancient GF and the possible interaction between GF and ILS (lines 400-409), a discussion also fueled by the new simulations I performed (see below response to your comment 2).

One could think of reparametrizing Aphid, calling t_g the GF time, and recalculating branch lengths (for instance, expected length a in scenario 5 would be t_g+2N_e). This could be more natural in terms of parameter interpretation, while also requiring some adjustments.

2. I have some comments regarding simulations:

2.1. 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.

I re-conducted simulations assuming asymmetric GF, and found that Aphid performs similarly well, and is able to detect the asymmetry, as now indicated (lines 276-284).

2.2. 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.

Thanks for a pertinent suggestion. I did that and found that the bias remains, contradicting my hypothesis. The text was modified accordingly. A recent preprint (Rivas-Gonzalez et al biorxiv 546039) thoroughly addresses a similar bias affecting the CoalHMM method. I am now referring to this manuscript when discussing this problem.

Aphid performed better in terms of gene tree annotation in these simulations without ancient GF, compared to the basic scheme, indicating that ancient GF is a source of confusion, as I now mention and discuss (line 268-275, 400-409) – and this is relevant to your comment 1 above as well.

2.3. 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.

I did not exactly follow this suggestion but did something related in simulating data under the very Aphid model, following a suggestion by Z. Yang (see below). This analysis shows that Aphid performs as expected under its very assumptions (a useful sanity check of the program itself), and gives an idea of the accuracy of parameter estimation, for a given distribution of sequence length (lines 210-220).

All these scenarios might be a bit more challenging to simulate with your custom script, but can be simulated using msprime.

I rather modified my own program.

3. 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.

This is a possibility I consider exploring in the future. The problem of a limited number of gene trees was addressed in the Homininae case by analyzing genome-wide data.

Minor comments:

4. 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”.

I agree but we are here talking about the first sentence of the manuscript, which I think needs to be general and concise. We have plenty of space in the rest of the ms to define more precisely what a gene is in this context. I use the term "genomic segment" in several instances.

5. Line 45: distinctive → different.

Modified as suggested

6. Line 76: wrong format for reference J and MR 2013.

Corrected

7. 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.

Agreed. In the revised version I use "speciation time" instead of the vague "divergence time". "Divergence" is now only used twice as a general term indicating that two lineages evolve independently, and thus diverge. In addition, the term "sequence divergence" refers to mutations accumulated since the most recent common ancestor of two alleles, as you're suggesting.

8. 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.

Indeed the parameters I used limit the percentage of discordant gene trees to ~50% (58% in the revised version, where I analyse 1000 simulations, not 100). The main reasons for this choice is that this is roughly the observed level of conflict observed in the data sets I'm analyzing here.

One could think of increasing the conflict in simulated data sets, e.g. by increasing the migration rate and Ne. What happens when I do that is that we simulate a lot of gene trees affected both by GF and ILS, as you describe in major comment 1 above. This is a situation where Aphid necessarly performs poorly, in the sense that the method is based on the assumption that the two processes do not interact. Figure 3C and 3D, for instance, do not even make sense when a large fraction of gene trees experience both ILS and GF – Aphid will predict one of the two processes (presumably ILS most of the times), and I don't know if I should call this a success or a failure.

I included a discussion of this limitation (lines 400-409), in line with my accounting of your major comment 1, recommending that Aphid be used when the conflict is not too high.

9. 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.

I completely agree. Rivas-Gonzalez et al (biorxiv 546039) recently addressed this problem by adding more classes of coalescence times in CoalHMM. I am now referring to this literature when discussing this problem (lines 439-442).

10. It would be useful to have a confusion matrix for the results in lines 234-246, containing all three categories (GF, ILS, no-event).

Now added as supplementary figure 4.

11. 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?

With the new analysis of non-coding data the emphasis of the revised version now has a clear focus on primates and apes. I decided to remove the non-primate data analyses from the ms. Leaving them as is not informative, as you're suggesting; commenting on these five additional data sets would take space and distract the reader from the main messages I'm trying to pass regarding the strengths/limitations of the method and gene flow in apes.

12. Throughout the text, you use the word “we” instead of the singular “I”, even though there is only a single author.

Yes, there seems to be some debate about this issue. I happen to have done this work by myself but am not so comfortable personalizing scientists in a ms – as far as science is concerned what matters is what has been done not who did it – and the "we" appears to me as a general designation of the author(s) whoever they are. I can imagine others rather find the single-author "we" pompous. In the revised version I mostly use the passive form, and "I" instead of "we".

13. The pa 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 agree with this comment, while recalling that, as you're highlighting in major comment 1, parameter t_g in Aphid is the time of coalescnce of lineages brought together by migration – so the detected GF indeed occurred more recently than the average t_g.

14. I believe the conclusion is unnecessary, you already summarize the method in the discussion section.

The discussion of the revised version was enriched in many respects. I would like to keep this brief recap at very end if you don't mind.

C. Reviewer 2’s comments

As more species have had their genomes sequenced, it has become increasingly clear that there is widespread phylogenetic incongruence between local gene trees and consensus species trees, particularly when the time intervals between successive speciations are relatively small. These incongruences can be caused either by Incomplete Lineage Sorting (ILS) or Gene Flow (GF), and many analytical methods address one or the other, but few both jointly. This paper introduces a new, computationally efficient approach called Aphid to modelling the contribution of both ILS and GF, and with them the times of species separation and basic population genetic parameters of the ancestral species. It appears to me to be novel and effective, giving interesting results, and I recommend publication after revision.

The key step is to greatly simplify the set of possible gene trees considered for explaining the observed data, allowing an efficient maximum likelihood approach to a mixture model over this set. This is very much an intentional heuristic, which results in an enormous reduction in state space, but it appears to be remarkably effective, with good results from reasonable simulations and a demonstration of application to real data. The ideas are nice and the exposition looks sound.

The analysis with Aphid of the human/chimp/gorilla relationships is a potentially important addition to the study of hominine evolution. It suggests that approximately half their genetic discordance is due to gene flow, and consequently that the H-C and HC-G main separation dates are older and ancestral population sizes smaller, which makes sense in a number of ways. There is potential for the author or others to build on this in the future in looking at the relationships of hominine species and subspecies, and considering further the relationship to other data than can be done in this short more technical article.

Thanks for the positive comments, much appreciated

0. Like many other phylogenetic approaches, Aphid takes as input a set of supposedly independent gene trees, each built under an assumption of no internal recombination. This referee has general concerns about the no-recombination assumption. For most mammals the average recombination rate is comparable to the mutation rate (e.g. 1e-8 compared to 1.25e-8 for humans) which means that there are on average as many recombination events as mutation events in the ancestral genealogy over a stretch of genome. Given that there have to be mutations present to enable the tree to be defined, then there should also be recombinations. There are many species with smaller mutation rates and much higher recombination rates (because they have smaller chromosomes), such as most invertebrates and many plants, for which the ratio of recombination events to mutations is much higher than one, often an order of magnitude higher. It is true that recombinations are clustered at hotspots, and that neighbouring trees separated by recombinations are correlated, but I would appreciate if you could explicitly discuss the issues for Aphid around the (almost certainly wrong) assumption of no recombination in gene trees.

Thanks for an important comment. I agree, and have tried to (grossly) assess the importance of the problem, based on estimates of the mean and variance of the recombination rate in humans. My conclusion is that, although certainly not negligible, recombination is probably not of major concerned for most of the data I'm analysing. This is a difficult and a general problem, optimally addressed by CoalHMM – at the cost of heavy calculations. I've added a specific section on this subject in the discussion (section 4.3).

Major points:

1. L95-101: It took me a while to realise that it was intentional that you are only considering trees with branch lengths at the expected values, rather than all possible trees. There is a good paragraph about this at the end in the discussion, but this modelling decision should be made much clearer earlier on because it is central to your method. First you should say early on that this is a heuristic approach involving major simplification – nothing wrong with that. Then, in the section from lines 95 and following, something more like “We model this set of observed gene trees as coming from a limited mixture of characteristic trees which we call scenarios, which have fixed branch lengths set to the expected values of the branch lengths for that scenario. These fall into three categories.” Then instead of “the coalescence times are assumed to equal” something more like “we model the coalescence times to be fixed at…”. This isn’t really an assumption, because it is flagrantly false – it is a modelling decision.

Thanks for this, corrected as suggested. In the revised version, the abstract says that Aphid is an approximate ML method, the introduction says that Aphid makes a number of simplifications, and the Method section contains the specific changes you suggested.

2. L176: 95% confidence intervals. Are these calculated by re-optimising the likelihood over all the parameters other than the one being investigated for each test value of the parameter under consideration? If so then say this. If not, then this is necessary. Otherwise, if parameters are coupled, it may be that the true confidence intervals are much wider.

Yes, this is exactly what the method does, now clarified as suggested.

a. Related to this, please say for your simulations for what fraction of simulations the true value of each parameter is within the confidence interval, and discuss as appropriate.

I did not calculate confidence intervals for simulated data as this considerably enlengthens the running time. Rather, I assessed the reliability of the confidence intervals calculated by Aphid empirically, thanks to the non-coding data set newly analyzed in the revised version (lines 359-371). I found that the estimated confidence intervals do a good job of capturing the sampling variance, but recall that departure from the model assumptions is another (probably more) important source of uncertainty.

b. And you don't show the confidence intervals in SuppTable2, nor the significance test results. Please add them. Sorry that this adds lots more columns. You could perhaps transpose the table and have columns per data set and rows per feature – your choice.

Done as suggested

3. L188-191: how is the root defined for deciding whether trees are imbalanced? Do you need a super-outgroup to set the root, beyond the ones whose lengths from tip to root are being compared to those from? If so, say so. Else explain how this is done. (If you assume ultrametric behaviour to define the root then of course you underestimate root-to-tip variation.)

The exon trees I am analyzing (already present in the first version) were downloaded from the OrthoMaM data base and come with a root. The analysis of non-coding data I am newly conducting indeed implies using a super-outgroup, as you're suggesting, and as explicitly described in the corresponding section (line 330). Please note that subsection "Aphid: requirements, data filtering, running time" starts like: "The Aphid program takes as input a set of rooted gene trees with branch lengths".

4. Related to the comments above about the assumption of no recombination within the loci, I would like to see for the real data (Supp Table 2) a new column giving the fraction of 2:2 SNPs involving A,B,C and an outgroup O that is incongruent in the test regions. This is what CoalHMM uses, and is independent of the no-recombination assumption. My memory is that for Human/Chimp/Gorilla/Orang this fraction is 30%, not 26% as you have. If you see such a difference it might help motivate discussion of the consequences of the no-recombination assumption. If you see no difference, then the explanation may be due to lower Ne in exons and/or my faulty memory. In any case, if there is no difference that is a nice validation that the model is behaving reasonably.

This is not a particularly natural thing to do in this manuscript since Aphid analyses gene trees, not sequence alignments. I do analyze non-coding sequence alignments in the revised

version. The mean incongruence detected by Aphid for this data set is 28.4%. I could not find a numerical estimate in any of the founding coalHMM papers (Hobolth et al 2007, Dutheil et al 2009). I did not modify the ms based on this comment.

5. Your discussion of the real data focuses almost exclusively on macaques and hominins. I think you should at least provide a bit more overview of the other results, otherwise why do them? It looks to me that for horses and mice there is negligible evidence for either GF or ILS – please give the significance test results in the table. For the others, the GF is similar to or greater than ILS. Quite surprising to me and worth remarking on. Is there other literature on these cases?

With the new analysis of non-coding data the emphasis of the revised version now has a clear focus on primates and apes. I decided to remove the non-primate data analyses from the ms. Leaving them as is not informative, as you're suggesting; commenting on these five additional data sets would take space and distract the reader from the main messages I'm trying to pass regarding the strengths/limitations of the method and gene flow in apes.

6. For macaques Song et al. had two M. fascicularis and only one of those shows the strong gene flow signature (the one from Mauritius, where the Portugese introduced M. fascicularis several hundred years ago from SE Asia). They interpret that as meaning that the gene flow has occurred within the last 330k years, since that is their estimated divergence time. However you estimate 63% p_a which is much more ancient. I wonder about the accuracy of your p_a estimate – all the other values are above 90%. You don't discuss this in your section on simulation. Could you please address how accurately p_a is estimated on the simulation data, and comment on this discrepancy in the macaque analysis.

A good point. It is not possible to approach the reliability of p_a estimation from the simulations I conducted because these simulations are based on the multi-species coalescent (MSC), which is not the Aphid model and has no p_a parameter. Rather, gene flow in the MSC is modeled via a migration rate parameter, which here was supposed to be constant in time and across pairs of lineages (and see above response to reviewer 1, comment 2.2). In the simulations newly performed under Aphid's very model, parameter p_a was estimated with reasonable accuracy (Supplementary Figure 2). This however does not inform much on how the estimated p_a will respond to real life situations.

Importantly, please note that my usage of term "GF time" in the first version was actually not appropriate (and see Reviewer 1's major comment 1). Actually, a high value of p_a means that a large fraction of the gene trees assigned to the GF category are such that the first coalescence (thinking backwards in time) occurred at a time close to the A/B speciation time. This coalescence time if by definition more ancient than the actual time of introgression, as now clarified at various places in the ms.

As far as the macaque data set is concerned, I note that in Song et al., if the signal of gene flow from/to *nemestrina* was indeed stronger with the *fascicularis* individual from Mauritius than with the other *fascicularis* individual, it was substantial for both (see their Figure 4a and 5a). The discussion suggests that only the former is significant, but this seems to be mainly a matter of deciding on a *p*-value threshold. I do not think that the Song et al. analysis rejects the hypothesis of gene flow between *nemestrina* and *fascicularis* before the divergence between the two analyzed *fascicularis* populations.

7. I note that Song et al. inferred bi-directional gene flow in this case, which is possible in principle with careful application of D-stats or 5-taxon tests. I realise that because your method

only models symmetric bi-directional gene flow it does not selectively demonstrate bi-directional versus uni-directional gene flow. You should state this somewhere.

Included at the begining of the last section "Potential improvements"

8. L343-345: you discuss not testing p_AB. Why not? This would be simple using the same scheme as for p_AC, p_BC. Maybe you have low power for this, but that would be good to report. It would not invalidate the paper at all from my perspective, just show the limits of the approach.

I did that and found that the p_{AB}=0 hypothesis is strongly rejected.

Overall, I do not strongly trust Aphid inferences regarding p_AB. The newly provided Supp Figure 4 regarding the reliability of gene tree annotation based on simulations perhaps gives an idea why – there is heavy confusion between the various scenarios entailing a canonical ((A,B),C) topology, due to the existence of the "intermediate" no_event scenario, which is difficult to distinguish from a GF scenario with a relatively ancient t_g.

9. L357: why not distinguish N_e(AB) from N_e(ABC)? I would be interested in what happens if you add that to the model. But I realise that this is substantially more work, so I do not require this. If you tried it and it didn't work well because of indeterminism, I would appreciate a statement to that effect as again being useful to understand the limits of the model, without requiring that you present the results to demonstrate this. In my view it is much more useful to describe things that didn't work than to bury them. I hope the editor and the other referee(s) take the same view!

I leave this possibility for future developments.

Minor points:

1. L24: "These problems are presumably minimized" – this is imprecise. They are presumably less of an issue, but not as small as possible, which is the meaning of minimized. Something more like "these potential problems are presumably much reduced" or "much less of a concern"

"minimized" replaced with "mitigated"

2. L76: "J & MR" ref should be Smith and Kronfest.

corrected

3. L79: "The ILS hypothesis predicts an exponential distribution for this variable regardless of tree topology" is incorrect. This needs to be "The ILS hypothesis predicts an exponential distribution for this variable for trees discordant with the species tree", which is what the Edelman paper says.

Corrected. My intention was to express the fact that, as soon as three lineages reach t_2 (thinking backwards in time), then we have an exponential distribution for the internal branch length and equiprobable topologies, but my formulation was incorrect, thanks for pointing this out.

4. L118: "as from" in place of "than from" is better English

Corrected.

5. L122: You need to state here that you assume at most one GF event.

Note totally sure how to phrase this here. Actually, histories involving two relatively recent events of GF are expected to generate gene trees that resemble one-GF-event gene trees, so, might be appropriately captured by our mixture of scenarios. Your comment is in part covered by the new piece of discussion about complex scenarios (lines 400-409)

6. L156: you talk about star topologies here, but later (L192) you rule them out in an indirect way, by saying that you ban trees with internal branches under 0.5 mutations – since the number of mutations is discrete this means with 0 mutations, i.e. star trees. I suggest just to say at L156 that you exclude star topologies with d = 0 (and again at L192).

This is not what I do: these star trees are included, and entail a likelihood calculation in which all scenarios confer a non-zero probability to the data, as now clarified in section 2.6 (formerly line 192).

7. L265: “lead” -> “led”

Corrected

8. L269: capitalise “Indonesia”

Corrected

9. Supp Table 2: why is the asymmetry_ILS < 0.5. By definition it should be greater than 0.5, as you say in L174. Also the table header is asymmetry_ILS while the text is imbalance_ILS.

Sorry this indeed was outdated material. Supp Table 2 was entirely reworked.

10. L282,L292: you must change “neutral mutation rate” to “exon mutation rate” or even more correctly “exon accepted mutation rate”. By using exons you are clearly not considering neutral sequence.

"Exon accepted mutation rate" is what I am intending to express by "neutral mutation rate", here assuming that deleterious mutations are not accepted and favourable ones are rare. I think "neutral mutation rate" is kind of a classical term in this context so decided to keep it.

11. L308: “appears” not “appear” in “appears to exist”

That sentence was modified (see below response to Z. Yang's comment 2)

12. L317,L320: “departing from”

Corrected

13. Figure 3: I find using the magnitude of the disks for the fraction explained hard to evaluate. Fine to leave the disks, but could you add next to them the actually number in text as a percentage (e.g. 12%, 4% etc. – no need for more precision here).

Figure reworked following your advice.

14. L350: “conditional” not “conditionally”

Corrected

15. Supp Text: “do not coalesce” rather than “do not coalesced”

Corrected

16. Reference list: Maybe this is an editorial rather than an author point, but for this style (name, year) surely the references should be in alphabetical order of first author.

Reformated as suggested

D. Reviewer 3’s comments

In this manuscript, Galtier introduces a new method which can distinguish gene flow (GF) from incomplete lineage sorting (ILS). The use of ABBA-BABA approaches have become commonplace, most notably in studies of Neanderthal and Denisovan introgression. However, ABBA-BABA approaches can only detect asymmetric GF. Notably, gene tree lengths are informative when it comes to distinguishing between GF and ILS (as GF tends to yield shorter gene trees than ILS). With this in mind, Galtier devised a maximum likelihood method (Aphid) that leverages gene tree length to estimate the prevalence of GF and ILS. Overall, I found this manuscript to be well written - no easy trick given the complexity of ILS. Mathematically, everything appears to be in good shape. Notably, Aphid outperformed a similar approach (QuIBL), and the comparisons of both methods were reasonably thorough given the length of this manuscript. As such, this work is a valuable addition to the literature. However, there are two obvious improvements that would enhance an already strong manuscript.

1) First, despite Scornavacca and Galtier 2017, I'm not completely sold on the use of exon trees. For one, balancing selection could potentially bias results. The author is encouraged to re-run Aphid using intergenic data to see if ((human, chimpanzee), gorilla) divergence times match up with the outputs from exon tree data.

I followed this suggestion and analyzed non-coding sequence alignments in apes. This is a much larger data set than the exon data set. The analysis makes sense in many respects:
- the estimated prevalence of ILS is increased with non-coding compared to coding sequence data, which is consistent with a reduction in "local Ne" in coding regions due to linked selection
- the estimated prevalence of gene flow is more accurately estimated with the non-coding data set, and a bit lower than with the coding data set, although still substantial (9-10% instead of 13%)
- the X chromosome differs significantly from the autosomes, with a smaller estimated theta and a lower estimated prevalence of GF, in agreement to previous reports in apes.
- the size of the data set allowed me to perform replicated analyses via subsampling and empirically assess the reliability of the estimates, which by the way gives credit to the confidence intervals calculated by Aphid.

I have added an entire section on this new analysis, and in part reorganized the manuscript, putting more emphasis on the ape result than in the first version. Thanks so much for inciting me to do this additional effort, which in my opinion significantly improved the ms.

2) Additional benchmarking would improve the utility of Aphid. How divergent can taxa be for Aphid to still give accurate results? The inclusion of additional simulations would enable the author to clearly spell out when it would (and would not) be a good idea to use Aphid.

This comment is largely addressed in my response to Reviewer 1 above (his major comment 2), who expressed a similar request and specific suggestions. I strengthened the simulation study by moving from 100 to 1000 simulated data sets, providing detailed results regarding gene tree posterior annotation, simulating asymmetric gene flow, simulating without ancient gene flow, and simulating under the very model underlying Aphid. This helps I think characterize the potential and limitations of the method, and in particular, the problem posed by ancient gene flow, which can easily be confused with ILS. This results in a recommendation, now explicitly stated in the discussion, of refraining from using Aphid, or being careful with the interpretation, when the proportion of discordant gene tree exceeds 50-55%.

Additional comments and suggestions:

3. Line 71: It might be good to cite some classic papers of "treeness" that paved the way for the initial ABBA-BABA papers (maying something like Piazza and Cavali-Sforza 1983 or Felsenstein 1982?)

I'm not sure I identify the papers you're referring to, and feel like this is perhaps starting a bit too early given the overall flow of the introduction?

4. Figure 1: It looks like the red HGT term is a holdover from an earlier version of this manuscript I assume it should be GF instead?

Exactly; corrected.

5. Line 132: Being able to incorporate locus-specific mutation rates is a nice feature.

Thanks for this comment

6. Equation 1: The text would benefit from explicitly mentioning what the index k is in this equation.

Done, and see above response to recommender's comment 7

7. Equations 8 and 11: Why do the equations for discordant topologyimbalance associated with ILS or GF have the maximum of two values in the numerator? It might be useful for the manuscript text to mention why I_ILS and I_GF are defined the way they are.

I've added a comment below these equations explaining what I_ILS and I_GF mean and are intended to measure.

8. Line 223: This is an interesting finding that is worth exploring in more details. How do parameter estimates scale with the percentage of discordant trees? Is the 35% an arbitrary cutoff?

The relationship between parameter estimates and the percentage of discordant trees is shown in Figure 3A, B and C. I clarified that 35% is indeed an arbitrary cut off.

9. Line 175: The confidence interval approach taken here seems appropriate.

Thanks for this assessment

10. Figure 2: It would be useful to include a diagram of the simulated scenarios as a panel in this figure.

This is not an easy task since the space of simulated gene trees is continuous instead of a finite number of scenarios as in Aphid, as now clarified (lines 226-227).

11. Lines 206-213: The Parameter values used in these simulations are appropriate.

Thanks for this assessment

12. Figure 2B: It might be good to discuss why theta is harder to estimate than other parameters.

Thanks for an interesting question, which I don't find easy to answer. Indeed the ratio of standard deviation to mean was higher for θ (0.29) than for τ_1 (0.16) or τ_2 (0.12) in my main simulations (and same in simulations under Aphid's model, see line 219). I often find useful to try and put words on what kind of signal is captured from the data by the various parameters when fitting a model. In the case of τ_1, τ_2 and θ in Aphid, my intuition would be that
- we measure with very good precision the average branch lengths in, particularly, the no-event scenario, so that $\tau_1+\theta/2$ and $\tau_2+\theta/2$ should be estimated with good accuracy
- we measure with reasonnably good precision the proportion of gene trees experiencing ILS, such that $(\tau_2-\tau_1)/\theta$, which determine the probability of ILS, should be estimated with acceptable accuracy.
It's not clear for me why this should result in a higher accuracy for τ_1 and τ_2 than θ. Maybe the fact that the former appear on the numerator and the latter on the denominator of the Prob(ILS) equation? I would like to do proper research and make sure that what I'm writing above is pertinent, or makes any sense at all, before addressing this issue. I did not modify the ms based on this comment.

13. Lines 240-244: I like this assessment, but the performance here depends on time depth of when these taxa diverged. Some additional advice about when to use (and when not to use) Aphid would be good.

Agreed, advice added, see above response to your comment 2.

14. Figure 3: Despite having only a small number of elements, this figure is not clear. Should the reader read anything into the different heights of each primate silhouette? Dot size also tends to be a poor way to quantify relative magnitude (as some readers might focus on the diameter of each dot, while others focus on the area of each dot). In any case, the author is advised to rework this figure for the sake of clarity.

The figure was reworked as suggested by this and other reviewers.

15. Line 283: Given that gene trees were reconstructed from exons, how might lineage-specific selection affect this pattern? Do similar patterns arise if intergenic data are used instead?

16. Line 290: What sort of divergence time estimates (including 95%CI) arise if intergenic data is used instead of exon data?

See above response to your comment 1.

17. Lines 305-314: I appreciate the openness regarding the caveats of this approach and am okay with the use of approximate likelihood calculations.

The example input files that are included on the gitlab page are a good detail, and they are likely to facilitate the use of Aphid by other teams of researchers.

Thanks for the positive comments, much appreciated

E. ***Julien Joseph's comments***

1. The method and the manuscript do not account for introgression from extinct or unsampled (so-called "ghost") lineages.

This is a great comment. I added a paragraph discussing this issue, including references to recent, insightful research on the subject. Indeed, gene flow from ghost lineages is likely to generate discordant genealogies with long branches (if the donor lineage is distant), which might confuse Aphid. I also note that if the process is asymmetric then it might manifest itself in an estimated index of ILS-associated imbalance well above 0.5.

F. Ziheng Yang's comments

1. p.7 Simulations. I think strictly speaking, this kind of simulation should use fixed parameter values to generate replicate datasets. Then concepts like bias, variance, CI coverage are all well defined. You sampled parameter values for each replicate dataset, which is a kind of Bayesian simulation. This strategy is used quite often but I don't think it is appropriate. It leads to results that are hard to interpret: for example what kind of results in figure 2 indicate that the program is correctly written, or the method is working as expected.

Thanks for this comment. The simulations I was conducting aimed at examine the robustness of Aphid under conditions departing from its assumptions, which seems required knowing that the model underlying Aphid is largely phenomenological – in particular, to assume a discrete set of fixed-branch-lengths scenarios has no biological realism. Yet I agree that these simulations do not address the aims you're mentioning regarding program correctness and sampling variance of parameter estimates. In the revised version I included additional simulations under the very Aphid model (lines 210-220, Supplementary Figure 2), which indeed suggest that the equations are OK and the program correctly implemented.

2. p.11 "First, no closed-form expression appear to exist for the distribution of all branch lengths conditional on the topology under the multi-species coalescent [Yang 2002]"
I am not sure what you mean precisely. The density of gene tree topologies and coalescent times (branch lengths) in the case of 3 species and 3 sequences is given by takahata et al. 1995 and copied in yang 2002. The density of gene tree topologies (labelled histories) and coalescent times (branch lengths) in the general case of an arbitrary number of species and an arbitrary number of sequences is given in rannala & yang 2003. The density is easier to explain using examples but requires

excessive notations to write down.

You're right my formulation was not exact, now corrected.

3. You should perhaps mention that estimated gene trees and branch lengths involve errors and uncertainties, when the method is applied to real data.

A good point, covered at lines 502-503 of the revised version, at the end of the discussion of the effect of recombination on gene tree-based methods.

Also thanks for the many additional references you sent, most of which are cited in the revised version.