Decision

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.

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.

My own review

In what follows, quotes from the preprint are in *italic font*; my own responses are in roman font.

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.

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.

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

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.

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 t_1. "No-event" is the case in which $x < t_2 - t_1$. Consequently, the appropriate expectation is $E[x|x < t_2 - t_1]$. 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 $z = (t_2 - t_1)/2N_e$,

$$E[x|x < t_2 - t_1] = 2N_e\big(1 - z/(e^z - 1)\big)$$

This is substantially smaller than $2N_e$ 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.

Eqns. 1–2: Why are some probabilities written as "P" and others as "$\mathcal{P}$?" In other words, why do some use a caligraphic font?

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

Eqns 6–7: These equations sum over $\{i : T_i \neq ((A, B), C)\}$, 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 T_i is ever defined.

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.)

177: Why 1.92 units of log likelihood?

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).

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.

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.

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.

283: If you're writing in LaTeX, consider using `$\leftrightarrow$`, which prints as $\leftrightarrow$.

References

[1] William S. Cleveland. *Visualizing Data*. Summit, NJ: Hobart Press, 1993.

[2] William S. Cleveland. *The Elements of Graphing Data*. 2nd. Summit, NJ: Hobart Press, 1994.

[3] Alan R. Rogers. "Legofit: estimating population history from genetic data". *BMC Bioinformatics* 20 (2019), p. 526. DOI: 10.1186/s12859-019-3154-1.

[4] Alan R. Rogers. "An efficient algorithm for estimating population history from genetic data". *Peer Community Journal* 2 (2022), e32. DOI: 10.24072/pcjournal.132.

[5] Alan R. Rogers and Ryan J. Bohlender. "Bias in estimators of archaic admixture". *Theoretical Population Biology* 100 (Mar. 2015), pp. 63–78. ISSN: 0040-5809. DOI: 10.1016/j.tpb.2014.12.006.