

Review of Phylogenetic Conflicts manuscript by Galtier

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.
- 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.

Additional comments and suggestions:

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

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?

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

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

Equations 8 and 11: Why do the equations for discordant topology

imbalance 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.

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?

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

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

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

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

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.

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.

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?

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

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.