PHYLOGENETIC CONFLICTS: DISTINGUISHING GENE FLOW FROM INCOMPLETE LINEAGE SORTING

Nicolas Galtier
Institut des Sciences de l'Evolution
Univ Montpellier - CNRS - IRD
nicolas.galtier@umontpellier.fr

ABSTRACT

Gene flow and incomplete lineage sorting are two distinct sources of phylogenetic conflict, *i.e.*, gene trees that differ in topology from each other and from the species tree. Distinguishing between the two processes is a key objective of current evolutionary genomics. This is most often pursued via the so-called ABBA-BABA type of method, which relies on a prediction of symmetry of gene tree discordance made by the incomplete lineage sorting hypothesis. Gene flow, however, need not be asymmetric, and the ABBA-BABA hypothesis-testing approach is not a proper way of measuring gene flow prevalence. We introduce Aphid, a maximum-likelihood method aimed at quantifying the sources of phylogenetic conflict via topology and branch length analysis of three-species gene trees. Aphid draws information from the fact that gene trees affected by gene flow tend to have shorter branches, and gene trees affected by incomplete lineage sorting longer branches, than the average gene tree. Accounting for the among-loci variance in mutation rate and gene flow time, Aphid returns estimates of the speciation times and ancestral effective population size, and a posterior assessment of the contribution of gene flow and incomplete lineage sorting to the conflict. Simulations suggest that Aphid is reasonably robust to a wide range of conditions. An analysis of exon trees in mammals illustrates the potential of the approach. Notably, 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.

Keywords speciation · coalescence · ABBA-BABA · effective population size · apes

1 Introduction

Phylogenetic conflicts occur when distinct genes support distinct trees. Conflicts often result from errors, when a reconstructed tree does not reflect the true history of the considered locus. Hidden paralogy, alignment problems and substitution model mis-specification are, among others, well-documented sources of discrepancy between gene trees and species trees [Scornavacca et al., 2020]. These problems are presumably minimized when closely related species are analyzed. In this case, it is likely that a majority of observed conflicts reflect genuine differences among genealogies. Two biological processes are potentially involved: gene flow (GF) and incomplete lineage sorting (ILS).

GF, or introgression, happens when a piece of DNA is transferred from one species to another. If the immigrant allele spreads and reaches fixation, then the previously accumulated divergence between the donor and recipient species is erased, and the two species get to be genealogically related at this locus, regardless of the speciation history. Loci experiencing GF are therefore likely to have genealogies that differ from each other, and from loci unaffected by GF. GF has been documented in a wide range of taxa [Green et al., 2010, Abby et al., 2012, Fontaine et al., 2015, Meyer et al., 2017, Ropars et al., 2018, Edelman et al., 2019, Glémin et al., 2019, Zhang et al., 2021, Suvorov et al., 2022a], and in eukaryotes typically happens via hybridization between closely related species, when reproductive isolation is not yet complete.

ILS is an independent source of phylogenetic conflict that happens when within-species polymorphism lasts longer than the time between two successive events of speciation. Consider for instance that, at one particular locus, three alleles, X_1, X_2 and X_3, segregate in an ancestral species. Imagine that the ancestral species quickly splits twice, and that all three alleles are transmitted to the newly formed species A, B and C. Now A, B and C start diverging, and each of them randomly keeps one of X_1, X_2 or X_3 by drift. The phylogeny at this locus will reflect the pre-existing genealogical relationships between X_1, X_2 and X_3 in the ancestral species, not the sequence of speciation events giving rise to A, B and C. The expected prevalence of ILS depends on the time between speciation events and on the effective population size [Pamilo and Nei, 1988]. ILS has been documented in a number of taxa, including great apes [Mallet et al., 2016, Dutheil et al., 2009, Hobolth et al., 2011, Guerzoni and McLysaght, 2016, Mendes and Hahn, 2016, Meleshko et al., 2021, Rivas-González et al., 2023].

GF and ILS are two distinctive processes, each responsible for a fraction of the genuine variation between gene trees. A natural question to ask is the relative contribution of ILS and GF to the observed conflict, which raises the issue of how to distinguish between the two processes. The literature on this subject is rich and diverse [Hibbins and Hahn, 2022]. To model phylogenies with ILS is relatively straightforward. The Multi-Species Coalescent model assumes that loci have independent coalescence histories within the framework of a shared species tree, divergence dates, and effective population size. Inference methods based on this model have been developed and applied to numerous data sets [Rannala and Yang, 2017], including a modified version also accounting for the auto-correlation of genealogies along chromosomal segments [Dutheil et al., 2009]. Adding GF to the picture is a more challenging task. Recent developments have been made in this direction, both theoretical and in terms of inference tools [Long and Kubatko, 2018, Wen and Nakhleh, 2018, Glémin et al., 2019, Flouri et al., 2020]. Some of these methods in principle fulfil the objective of jointly estimating the prevalence of ILS and GF. These methods, however, are computationally demanding, and/or require specific hypotheses about which lineages have exchanged genes. They have not been applied to a large number of data sets so far.

Rather, the empirical literature is dominated by a hypothesis-testing approach, in which the existence of GF is demonstrated via rejection of an ILS-only scenario. This is based on a prediction made by the ILS model, which is that, if $((A, B), C)$ is the true species tree, the alternative $((A, C), B)$ and $((B, C), A)$ trees should be observed in roughly equal frequencies. GF, in contrast, can be asymmetric, *i.e.*, there could be more gene flow between A and C than between B and C, for instance, leading to a larger number of $((A, C), B)$ than $((B, C), A)$ gene trees. A significant imbalance between the discordant topologies therefore demonstrates the existence of GF. The popular ABBA-BABA test [Green et al., 2010, Durand et al., 2011] and its variations [Martin et al., 2015, Pease and Hahn, 2015, Blischak et al., 2018] implement this idea. These methods are efficient and have routinely been applied to a large number of data sets.

It should be noted that, although perfectly sound for detecting the existence of GF, ABBA-BABA-like approaches are not suitable for quantifying its prevalence. This is because GF does not need to be asymmetric, but only its asymmetric component is detected by ABBA-BABA. If GF between C and A (on one hand) and between C and B (on the other hand) occurred at a similar rate, then ABBA-BABA tests would simply have no power to reject the ILS hypothesis [Vanderpool et al., 2020]. Yet, even in the case of symmetric gene flow, the ILS and GF hypothesis can in principle be distinguished since they make distinctive predictions regarding coalescence times. Indeed, under GF the time to the

most recent common ancestor of the interacting lineages is reduced compared to non-GF scenarios, whereas ILS, in contrast, entails relatively old coalescences. This means that besides inbalance in discordant topology counts, one can potentially gain information on the prevalence of GF and ILS by examining gene tree branch lengths. Although present in the literature [Holder et al., 2001, Joly et al., 2009, J and MR, 2013, Meyer et al., 2017, Suvorov et al., 2022b], this idea has rarely been used as a basis for of method distinguishing GF from ILS. A notable exception is the recently developed QuIBL method [Edelman et al., 2019], which focuses on the distribution of the length of the internal branch in three-species rooted gene trees. The ILS hypothesis predicts an exponential distribution for this variable regardless of tree topology, whereas in the case of GF, gene trees departing the species tree topology should be associated with an increased mean and variance of the internal branch length.

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. Aphid basically asks whether discordant gene trees tend to be taller, indicating ILS, or shorter, indicating GF, than the majority gene trees. Modeling a set of three-species rooted gene trees as a mixture of no-event, ILS and GF scenarios, Aphid considers both terminal and internal branch lengths and accounts for differences in mutation rate among loci. Below we introduce the details of the Aphid method, assess its performances using simulations, and apply it to eight triplets of recently diverged species of mammals. We show that, unlike ABBA-BABA tests, Aphid can reveal the existence of symmetric GF during species divergence, for instance in the human/chimpanzee/gorilla case.

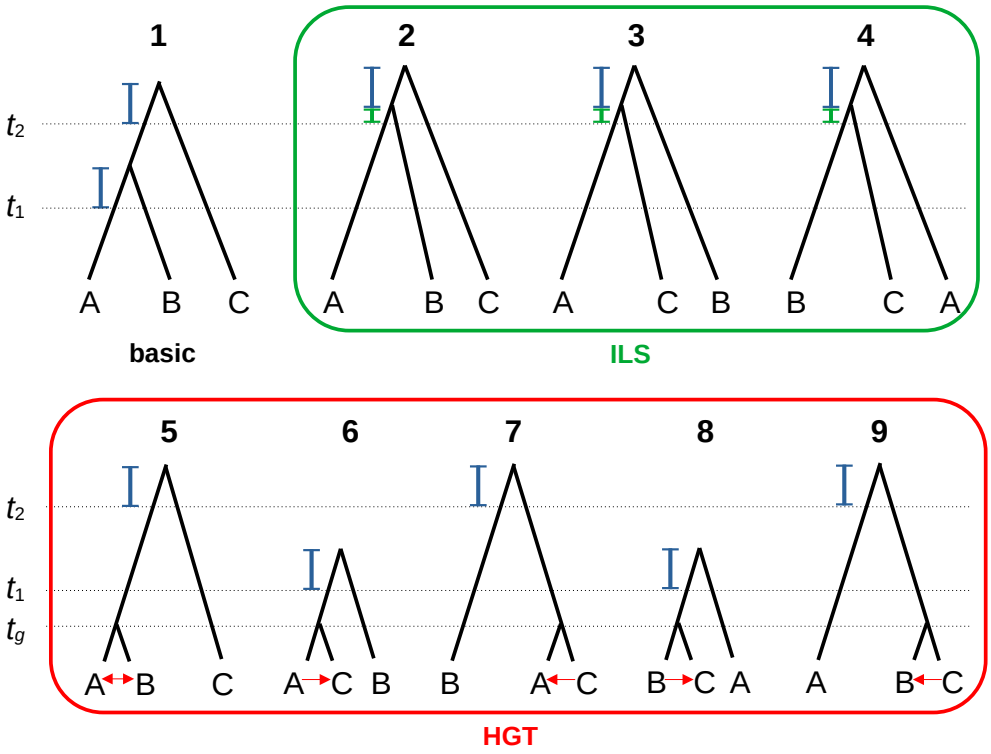

Figure 1: Typical scenarios involving incomplete lineage sorting (green), gene flow (red) or no event. Branch lengths depend on divergence times t_1 and t_2 and effective population size N_e. The vertical blue and green bars are of length $2N_e$ and $2N_e/3$, respectively. Observed gene trees are modelled as a mixture of such scenarios. See Supplementary Figure S1 and Supplementary Table 1 for a detailed version.

2 The method

2.1 Mixture of genealogies

Consider a triplet of species $((A, B), C)$ such that the $(A, B)|C$ split occurred t_2 generations ago and the $A|B$ split occurred t_1 generations ago, and assume a constant population size of N_e diploid individuals across the whole tree. Suppose that we sample exactly one haplome per species and reconstruct the genealogical history (gene tree) of a number of genomic segments (loci). We model this set of gene trees as a mixture of scenarios from three categories: no-event, ILS, and GF. The no-event category corresponds to the scenario in which neither GF nor ILS has been involved (Fig.1, scenario S_1). Under this scenario the tree topology must be $((A, B), C)$, and coalescence times are assumed to equal $t_1 + 2N_e$ and $t_2 + 2N_e$ for the $A|B$ and $(A, B)|C$ split, respectively, $2N_e$ being the average coalescence time between two lineages in a panmictic population. The ILS category corresponds to three scenarios in which the coalescence between the A and B lineages is older than t_2. In this case the three possible topologies can be generated, hence scenario S_2, S_3 and S_4 in Fig.1. Coalescence times are assumed to equal $t_2 + 2N_e/3$ and $t_2 + 2N_e/3 + 2N_e$, respectively, $2N_e/3$ being the average time of the first coalescence among three lineages in a panmictic population. Finally, the GF category corresponds to scenarios in which an event of gene flow has occurred. Fig.1 displays five such scenarios, depending on which lineage is the donor and which the recipient. Scenario S_5 involves an event of gene flow between the A and B lineages. This does not affect the topology, but leads to a younger $A|B$ coalescence time compared to scenario 1. Scenario S_6 and S_7 involve a transfer from A to C and C to A, respectively, which both result in an $((A, C), B)$ topology while also modifying the expected coalescence times and branch lengths. Scenario S_8 and S_9 similarly model gene flow between B and C.

The five GF scenarios shown in Fig.1 are such that an event of GF happened at one specific time in the past, this time being shared by all loci experiencing GF. Our method actually relaxes this assumption by assuming a discrete distribution, instead of fixed values, of GF time. The user is asked to choose a finite number, n_c, of GF times, expressed as fractions of t_1. The oldest of these GF times is assigned probability p_a (for "ancient"), while each of the other GF times has probability $(1 - p_a)/(n_c - 1)$. In this study, the default setting was $n_c = 2$ and it was assumed that GF events could occur at time t_1 or $t_1/2$. We therefore considered ten GF scenarios, as explicitly described in Supplementary Figure S1 and Supplementary Table 1.

Our model assigns prior probabilities to these scenarios. GF between A and B, A and C, and B and C are assumed to occur with probability p_{AB}, p_{AC} and p_{BC}, respectively. GF is here assumed to have the same probability to occur from A to C than from C to A, and from B to C than from C to B - events of GF from A to B and from B to A lead to the very same pattern and are indistinguishable anyway. The p_{AB}, p_{AC} and p_{BC} parameters, combined with p_a, define the prior probabilities of the ten GF scenarios, as detailed in Supplementary Table 1. In the absence of gene flow, ILS will occur if the A and B lineages have not yet coalesced at time t_2, which happens with probability $p_{ILS} = e^{-(t_2-t_1)/2N_e}$ [Pamilo and Nei, 1988]. Each of the three ILS scenarios therefore has probability $(1 - p_{AB} - p_{AC} - p_{BC})p_{ILS}/3$. The no-event scenario, finally, has probability $(1 - p_{AB} - p_{AC} - p_{BC})(1 - p_{ILS})$. This parametrization entails a constraint on the sum of GF probabilities, $p_{AB} + p_{AC} + p_{BC}$, which must be lower than one.

2.2 Mutation rate

The above description expresses time in generation units, whereas what we can measure from the data is sequence divergence, which for a neutral locus depends on divergence time and mutation rate. Calling μ the locus mutation rate, the t_1, t_2 and N_e parameters are re-scaled as:

$\tau_1 = t_1\mu$

$\tau_2 = t_2\mu$

$\theta = 4N_e\mu$

Crucially, the method accounts for the variation in mutation rate among loci. This is required if one aims at interpreting differences in branch length among gene trees in terms of GF or ILS. We introduce a locus-specific factor, α_i, which is the relative mutation rate of locus i and multiplies the predicted branch lengths for this locus. Information on α_i's will be obtained by using additional species besides A, B and C (see below). Supplementary Table 1 expresses branch lengths under the 14 scenarios as functions of the re-scaled parameters.

2.3 Parameter estimation

The model includes seven parameters controlling the distribution of gene tree topology and branch lengths (divergence times τ_1 and τ_2, population mutation rate θ, gene flow probabilities p_{AB}, p_{AC}, p_{BC}, p_a) and one relative mutation rate per locus, α_i.

The whole set of parameters is not identifiable from three-species gene trees only since the predicted branch lengths are expressed as products of α_i and τ_1, τ_2, or θ (Supplementary Table 1). An external source of information on the α_i's is therefore needed. Aphid requires that every input gene tree includes at least one species from a pre-specified outgroup clade, in addition to A, B and C. The relative mutation rate at locus i, α_i, is estimated as $l_i/\bar{l}$, where l_i is the mean distance from root to tip in the ith gene tree, and $\bar{l}$ is the average of l_i's across gene trees. This relies on the assumptions that the mutation rate is constant in time and among lineages, *i.e.*, that sequences evolve in a clock-wise manner.

The other seven parameters are estimated in the maximum likelihood framework. The probability of the ith gene tree given the parameters is written as:

$$P(G_i) = \sum_k P(S_k)P(G_i|S_k) \tag{1}$$

where G_i is the ith gene tree. The prior probabilities of the various scenarios, $P(S_k)$, are functions of the parameters as defined above and in Supplementary Table 1. The probability of a gene tree given a scenario, $P(G_i|S_k)$ equals zero if the topology of G_i does not match S_k, and is written as:

$$P(G_i|S_k) = \mathcal{P}(a_i; \alpha_i\bar{a}[k])\mathcal{P}(b_i; \alpha_i\bar{b}[k])\mathcal{P}(c_i; \alpha_i\bar{c}[k])\mathcal{P}(d_i; \alpha_i\bar{d}[k])) \tag{2}$$

if the topology of G_i matches S_k. Here, a_i, b_i, c_i and d_i are the terminal branch lengths of the A, B and C lineages and internal branch in G_i, respectively, $\bar{a}[k]$, $\bar{b}[k]$, $\bar{c}[k]$ and $\bar{d}[k]$ are the expected branch lengths under scenario S_k (Supplementary Table 1), and $\mathcal{P}(j; \lambda) = \lambda^j e^{-\lambda}/j!$ is the probability mass function of the Poisson distribution. Note that branch lengths are here defined as the total, not per site, number of sequence changes having occurred between two nodes. The method can easily account for unresolved (*i.e.*, star-like) gene trees: the probability of an unresolved gene tree given a scenario is given by equation (2) irrespective of scenario topology. Finally, the likelihood of the whole data set is obtained by multiplying the likelihood across gene trees, here assumed to be independent. The likelihood was maximized using an adapted version of the Newton-Raphson algorithm, which relies on differentiation of the likelihood function with respect to the parameters.

2.4 Gene tree annotation, GF/ILS contribution, imbalance

Having estimated the parameters, one can calculate the posterior probability of each of the scenarios for any particular gene tree using an empirical Bayesian approach:

$$P(S_k|G_i) = P(G_i, S_k)/\sum_k P(G_i, S_k) \tag{3}$$

The posterior probability for a gene tree to be affected by ILS can be calculated as:

$$P(\text{ILS}|G_i) = \sum_{S_k \in \mathcal{S}_{ILS}} P(S_k|G_i) \tag{4}$$

and similarly for GF:

$$P(\text{GF}|G_i) = \sum_{S_k \in \mathcal{S}_{GF}} P(S_k|G_i) \tag{5}$$

where $\mathcal{S}_{ILS}$ and $\mathcal{S}_{GF}$ are the sets of ILS and GF scenarios, respectively.

The estimated contributions of ILS and GF to the phylogenetic conflict in the whole data set are expressed as:

$$P_{\text{ILS}} = \sum_{T_i \neq ((\text{A,B}),\text{C})} P(\text{ILS}|G_i)/n \tag{6}$$

and

$$P_{\text{GF}} = \sum_{T_i \neq ((\text{A,B}),\text{C})} P(\text{GF}|G_i)/n \tag{7}$$

where T_i is the topology of gene tree G_i and n is the number of analyzed gene trees. Equation 6 and 7 sum posterior probabilities across all gene trees having a discordant topology, either $((A, C), B)$ or $((B, C), A)$. Separating these two terms, we define the discordant topology imbalance associated with ILS as

$$I_{\text{ILS}} = \frac{\max(P_{\text{ILS}}[1], P_{\text{ILS}}[2])}{P_{\text{ILS}}} \tag{8}$$

where

$$P_{\text{ILS}}[1] = \sum_{T_i = ((\text{A,C}),\text{B})} P(\text{ILS}|G_i)/n \tag{9}$$

$$P_{\text{ILS}}[2] = \sum_{T_i = ((\text{B,C}),\text{A})} P(\text{ILS}|G_i)/n \tag{10}$$

and similarly for GF:

$$I_{\text{GF}} = \frac{\max(P_{\text{GF}}[1], P_{\text{GF}}[2])}{P_{\text{GF}}} \tag{11}$$

$$P_{\text{GF}}[1] = \sum_{T_i = ((\text{A,C}),\text{B})} P(\text{GF}|G_i)/n \tag{12}$$

$$P_{\text{GF}}[2] = \sum_{T_i = ((\text{B,C}),\text{A})} P(\text{GF}|G_i)/n \tag{13}$$

I_{ILS} and I_{GF} vary between 0.5 (no imbalance) and 1 (maximal imbalance).

2.5 Confidence interval, hypothesis testing

95% confidence intervals around maximum likelihood parameter estimates were defined as intervals including all values of the considered parameter such that the log-likelihood dropped off by less than 1.92 units. Likelihood ratio tests were conducted to assess the significance of the detected GF. To this end, the likelihood was maximized under the constraint that p_{AC} and/or p_{BC} equalled zero. Twice the difference in log-likelihood between the unconstrained and constrained models was compared to a χ^2 distribution with one (if a single parameter was set to zero) or two (if two parameters were set to zero) degrees of freedom. This procedure allows testing the significance of gene flow between A and C, of gene flow between B and C, or of gene flow generally.

2.6 Aphid: requirements, data filtering, running time

The Aphid program takes as input a set of rooted gene trees with branch lengths, a focal triplet of species and a set of outgroup species. Because branch lengths are usually expressed in per site number of changes, sequence length must also be input for each gene tree. Aphid only analyzes gene trees in which the three focal species are present and monophyletic and at least one outgroup species is present. When the considered gene tree contained several outgroup species, we here required these to be monophyletic. In addition, Aphid filters gene trees based on a criterion of clock-likeness. Specifically, the average distance from root to tips is calculated separately for the focal species, on one hand, and the outgroup, on the other hand, and the ratio between these two distances must be lower than a user-defined threshold, which was here set to 2. Aphid considers a gene tree as topologically unresolved when the internal branch length is below a user-defined threshold, which was here set to 0.5 mutations. Gene trees in Aphid are supposed to describe the coalescence history of unlinked genomic segments, which implies assuming that no recombination has happened within a segment, and that recombination between segments is high. The segment size and between-segment distance satisfying these conditions are expected to vary among species depending on the effective population size and recombination landscape. Aphid typically takes less than a minute for analyzing data sets of thousands of gene trees on a laptop. In this study we used Aphid version 0.10.

3 Simulations

Coalescent simulations were conducted under the multi-species coalescent model with gene flow in order to assess the performance of the Aphid method. We considered an ((A,B),C) ultrametric species tree with divergence time t_1 between A and B and t_2 between (A, B) and C, and population size N_e. Migration was assumed to occur at constant rate m. Migration occurred between the three A, B and C species when time was lower than t_1 (rate m for each of the three pairs of species), and between the ancestral (A, B) and C species when time was between t_1 and t_2, both ways. This implies that the simulations departed the Aphid scenarios in several respects. For instance, very recent ($t < t_1/2$) and ancient ($t > t_1$) gene flow times were possible, as well as multiple events of gene flow.

We used a sample size of one gene per species and simulated 100 data sets, each consisting of 2000 independently-generated gene trees. An outgroup distant from A, B or C by 2 million generations was added to every gene tree. Gene trees were assumed to correspond to genomic segments of length following a log-uniform distribution between 100 and 1000 bp. The mutation rate was assumed to vary by a factor of nine among genomic segments, with a mean value of 2.10^{-8} per generation per bp. The t_1, t_2, N_e and m parameters differed among the 100 simulated data sets as they were drawn from uniform distributions: t_1 took values in [0.5; 1.5] million generations, $t_2 - t_1$ in [0.25; 0.75] million generations, N_e in [50,000; 150,000] and m in [10^{-8}, 10^{-7}] per generation. The percentage of simulated gene trees with a discordant ((A,C),B) or ((B,C),A) topology) varied from 6% to 50% across data sets, with a mean of 26%.

Fig.2A and B compare the estimated and simulated values of divergence parameters τ_1 and τ_2 and population mutation rate θ. There was a tendency to slightly underestimate τ_1 and τ_2 and slightly overestimate θ. This might result from the occurrence, in some simulations, of events of GF at times older than t_1. Such ancient GF events are not allowed in any

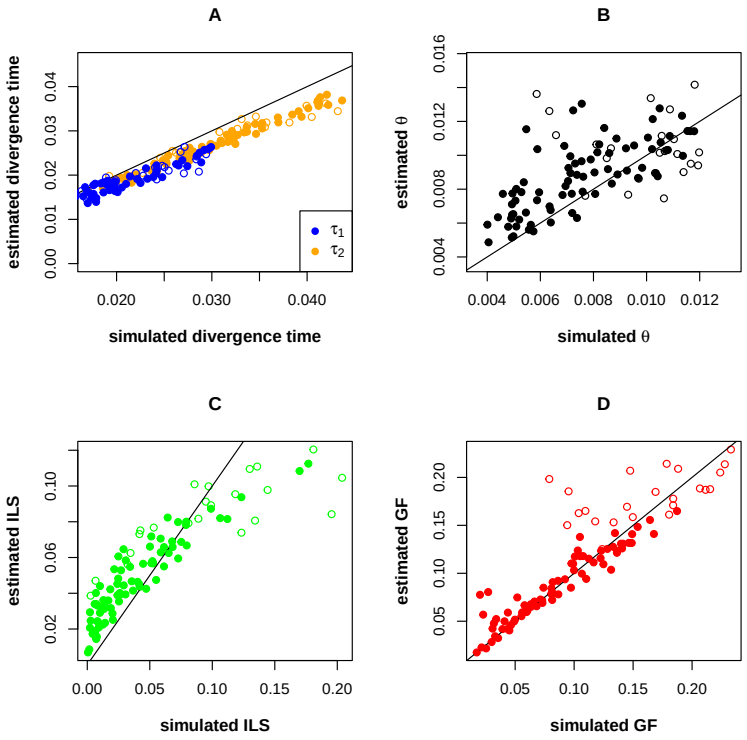

Figure 2: Simulation-based assessment of the performances of Aphid. A. Estimated divergence parameters τ_1 and τ_2. B. Estimated θ. C. Estimated ILS prevalence in phylogenetic conflicts. D. Estimated GF prevalence in phylogenetic conflict. True values are on the X-axis. Closed dots correspond to simulated data sets in which discordant topologies amounted 35% of the gene trees or less.

of the scenarios considered in Aphid (Fig.1). This appears to be mitigated by young estimated divergence times and a high estimated θ, such that the discordant gene trees due to ancient GF are accommodated by ILS scenarios. Please note, however, that a similar bias was reported with a related method irrespective of gene flow [Dutheil et al., 2009], so the explanation might be more complex than that. Fig.2C and D compare the estimated and simulated proportion of discordant gene trees resulting from ILS and GF, respectively. The prevalence of ILS (equation 6) and GF (equation 7) were reasonably well estimated under a wide range of conditions, particularly when the percentage of discordant gene trees was below 35% (closed dots). Spearman's correlation coefficient between true and predicted prevalence was 0.91 for ILS, 0.93 for GF. ILS tended to be slightly overestimated when its real effect was low, and underestimated when it was high.

We tested the QuIBL program [Edelman et al., 2019] on these simulated data sets using the default options, with the exception of the numsteps parameter, which was set to 50 in agreement with the program documentation. QuIBL did not perform particularly well here, the estimated ILS and GF prevalence being only weakly correlated with the simulated ones (Supplementary Fig. 2A): Spearman's correlation coefficient between true and predicted prevalence was 0.42 for ILS, 0.50 for GF. QuIBL assumes a constant mutation rate across loci, whereas the mutation varied substantially in our simulations. We therefore conducted a second series of simulations with a constant mutation rate. The correlation between the simulated and estimated GF/ILS contributions was improved, while not reaching the accuracy of Aphid (Supplementary Fig. 2B).

We evaluated the performance of Aphid in annotating the status of a specific gene tree. We calculated, for each locus of each simulated data set, the posterior probability of being affected by ILS (equation (4)) or by GF (equation (5)),

and defined the posterior probability of no-event as one minus the sum of these two probabilities. If one of the three posterior probabilities was above 0.90 then the corresponding locus was annotated accordingly. If none was above 0.90 then no annotation was made. The percentage of annotated gene trees varied greatly among simulated data sets, from 7% to 79%, the average across data sets being 32%. We focused on the annotation of simulated gene trees with a discordant $((A, C), B)$ or $((B, C), A)$ topology. We found that 95% of the discordant gene trees predicted to result from GF actually experienced an event of GF in the simulations; only 5% of these were false GF positives, *i.e.*, discordant gene trees having actually experienced ILS. The specificity was much lower as far as ILS-predicted discordant gene trees were concerned: only 57% of these were gene trees having truly experienced ILS in the simulations, the other 43% being false GF negatives. We suggest that this result might be at least in part explained by the existence in the simulations of ancient events of GF, occurring earlier than t_1. The resulting gene trees have relatively long terminal branches, and are not properly taken into account by the Aphid GF scenarios.

Overall these simulations suggest that (i) Aphid estimates the relative contribution of ILS and GF to phylogenetic conflicts with reasonably good accuracy across a wide range of conditions, provided the overall conflict prevalence is not too high, and (ii) Aphid only annotates a limited fraction of genes (with the sequence length considered here), while calling a small number of false GF positives.

4 Data analysis

We applied the Aphid program to eight mammalian data sets downloaded from the Orthomam v10 database [Scornavacca et al., 2019]. These data sets, depicted in Supplementary Table 2, correspond to triplets of species having diverged between 1.5 and 12 Mya according to the TimeTree of Life (`http://www.timetree.org/`). We used gene trees reconstructed from alignment of exons of length 400 bp or above - exons were previously identified as appropriate gene tree reconstruction units in mammals [Scornavacca and Galtier, 2017]. There were 1926 to 3417 eligible exon trees per data set after filtering for taxon sampling and clock-likeness (see above 2.6). In all eight data sets a majority of the gene trees agreed with the Orthomam reference species tree topology (`https://orthomam.mbb.cnrs.fr/`). The percentage, among resolved gene trees, of discordant topologies varied from 5% to 42% among data sets, and the average sequence divergence varied from 0.06% to 1.4%. We here specifically focus on the three primate data sets, Macaca, Papionini, and Homininae (Fig.3), which illustrate the diversity of patterns we obtained. The results for the other five data sets are given in Supplementary Table 2.

Dot size in Fig.3 is proportional to the estimated contribution of ILS (green, equation 8) or GF (red, equation 11) to the conflict. The two processes contributed a substantial fraction of the conflict in the three data sets. In all three cases a model assuming no GF (parameters p_{AC} and p_{BC} fixed to zero) lead to a significant drop in log-likelihood and was rejected with a *p*-value below 10^{-10} by a likelihood ratio test. The Y-axis in Fig.3 is the estimated imbalance between the two discordant topologies associated with the detected effects. As far as ILS was concerned, the imbalance was low in the three data sets, as expected. The estimated GF was associated with strong discordant topology imbalance in Macaca. In this data set, Aphid detected significant GF between the two indonesian species *Macaca fascicularis* and *M. nemestrina*, but much less GF between *M. mulatta* and *M. nemestrina*, consistent with published ABBA-BABA-based analyses [Vanderpool et al., 2020, Song et al., 2021]. This is essence says that in Macaca, the $((fascicularis, nemestrina), mulatta)$ and $((mulatta, nemestrina), fascicularis)$ topologies appear in roughly equal numbers as far as tall discordant gene trees are concerned, whereas the former topology is highly dominant among short discordant gene trees. Note that the signal captured by ABBA-BABA is averaged across all the discordant gene trees, irrespective of branch lengths. Black squares in Fig.3 show the magnitude of this merged imbalance, which in Macaca was much weaker than the GF-specific imbalance. By separating the signal from tall *vs.* short gene trees, Aphid appears to provide a more informed description of the situation.

Interestingly, in the other two data sets, Aphid detected substantial gene flow despite low discordant topology imbalance - a signal necessarily missed by ABBA-BABA. This was notably the case in the Homininae data set, *i.e.*, the ((Hu-

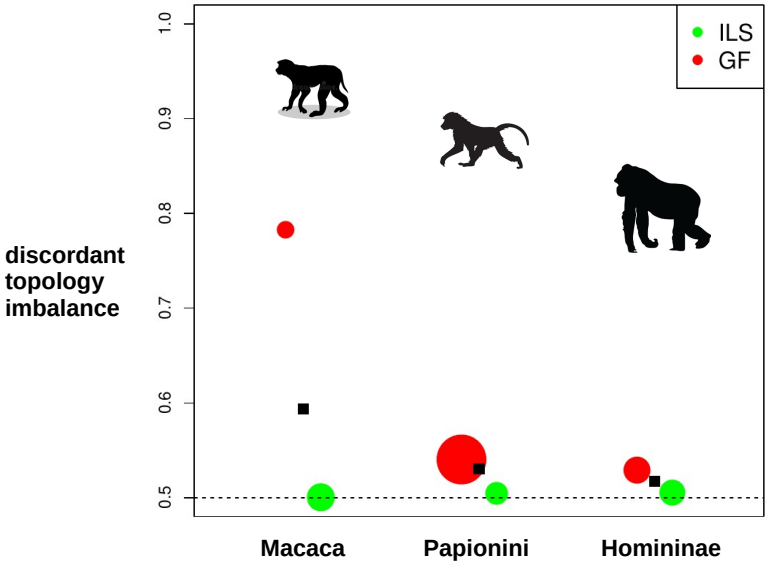

Figure 3: Aphid analysis of three primate data sets. Dot size is proportional to the estimated contribution of ILS (green) or GF (red) to the phylogenetic conflict. Y-axis: imbalance between the contributions of the ((A,C),B) and ((B,C),A) topologies to the detected effect. Black squares: overall imbalance, merging the ILS and GF signals.

man,Chimpanzee),Gorilla) triplet. This triplet has been thoroughly studied during the last decades, with the conflict most often interpreted in terms of ILS [Hobolth et al., 2007, Dutheil et al., 2009, Rivas-González et al., 2023]. Aphid rather suggests that roughly half of the discordant gene trees are shorter than expected under ILS, and more likely to result from chimpanzee <-> gorilla and human <-> gorilla gene flow. The estimated p_a parameter equalled 0.95 for the Homininae data set, meaning that most of the detected gene flow is inferred to have happened shortly after the human|chimpanzee split.

Using whole-genome data and empirical estimates of the mutation rate and generation time, Rivas-González et al. [2023] dated the average coalescence time between human and chimpanzee at 10Mya (their Fig. 1A). The average sequence divergence between human and chimpanzee in our data set is $6.85\,10^{-3}$ per bp. Combining these two numbers, we get an estimated neutral mutation rate of $3.43\,10^{-4}$ per bp per My for our data. Dividing the estimated τ_1 and τ_2 by the mutation rate yields an estimated human|chimpanzee speciation time of 8.4My (95% confidence interval: [8.0; 8.8]) and an estimated human|gorilla speciation time of 11.4My (95% confidence interval: [11.1; 11.7]). Assuming a generation time of 20 years for apes, we get a per bp per generation neutral mutation rate of $6.85\,10^{-9}$. Dividing our estimated θ by four times this figure yields an estimated effective population size of 60,200 (95% confidence interval: [54,800-65,500]). Rather using a human-gorilla coalescence calibration, dated at 12.4 Mya by [Rivas-González et al., 2023], provides younger speciation time estimates (human|chimpanzee: 7.7 [7.3-8.1]; human|gorilla: 10.4 [10.1-10.7]) and a smaller estimated effective populatiopn size (55,000 [50,100-59,800]). Assuming that the phylogenetic conflict entirely results from ILS, Rivas-González et al. [2023] rather reported estimated speciation times of 6 Mya (human|chimpanzee) and 8 Mya (human|gorilla), and ancestral effective population sizes of 177,000 (human-chimpanzee ancestor) and 107,000 (human-gorilla ancestor).

5 Discussion

Aphid models a set of gene trees as a mixture of typical scenarios, and returns an estimate of the contribution of the various (categories of) scenarios to the data. The analysis of simulated and real data suggested that Aphid is reasonably robust to a wide range of conditions and has the potential of revealing gene flow despite balanced discordant topology counts, thus extracting more information from the data than the simpler ABBA-BABA.

Likelihood calculation in Aphid is approximate in that it does not account for the variance in coalescence time among loci. The model assumes that gene trees associated with a given scenario only differ from each other due to the stochasticity of mutation, neglecting the variance due to drift. There are two connected reasons why we made this approximation. 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]. Secondly, the calculation and maximization of the approximated likelihood are achieved efficiently, since the likelihood can be analytically differentiated with respect to parameters. The Aphid method typically takes a few seconds to analyse a data set of several thousand gene trees, offering possibilities such as data re-sampling for instance. Our simulations suggest that inferences are reasonably robust despite the approximation. Yet, one should keep in mind that Aphid is a heuristic and does not rely on a proper population genetic model.

As Aphid aims to interpret branch lengths in terms of coalescence times, properly modeling the among loci variation in mutation rate is essential. The strategy retained here consists in (1) including an outgroup, (2) removing gene trees departing the molecular clock hypothesis, (3) estimating a locus relative mutation rate as the average root-to-tip distance. Although plausible this procedure could probably be improved. Step (2) above, in particular, implied discarding up to 40% of gene trees in our analysis of mammalian data sets. Instead, one might explicitly model the among lineage rate variation and estimate the triplet mutation rate even for gene trees departing the molecular clock assumption [Thorne et al., 1998, Lartillot et al., 2009]. Another potential issue with the above procedure is that the average root-to-tip distance depends not only on mutation rate but also on the age of the root node, which is expected to differ among loci due to the variance of the coalescent, and possibly ancient GF. For this reason we recommend using a relatively distant outgroup, such that the standard deviation in root age is small compared to root age. It should be noted that the uncertainty in locus-specific mutation rate estimation was not taken into account in the confidence intervals we calculated here.

Aphid has a lot in common with the CoalHMM method [Dutheil et al., 2009], which also uses discrete categories of scenarios with branch lengths equal to theoretical expectations. CoalHMM does not analyse supposedly independent gene trees but models the topological variation across a genome - a more ambitious objective. By explicitly modeling recombination, CoalHMM uniquely offers the opportunity of comprehensively analyzing genome-wide data without the need of identifying unlinked, non-recombining genomic windows. Another major difference is that CoalHMM does not consider GF, and rather assumes that the topological conflict entirely results from ILS - not an insignificant assumption. The ABBA-BABA-based literature has reported a significant imbalance between discordant topologies in many taxa, suggesting that GF is not a generally negligible source of phylogenetic conflict.

Aphid also has similarities with the maximum-likelihood method proposed by Yang [2010]. Like Aphid, this method analyses supposedly independent three-species gene trees and models the among loci variation in mutation rate. Unlike Aphid and CoalHMM, it numerically integrates across all possible coalescence times, thus providing an exact likelihood calculation, while also accounting for multiple mutations at the same site. A major difference is that the method introduced by Yang [2010] models GF between the two most recently diverged species - A and B according to our notations - whereas Aphid mainly focuses on GF between A and C or B and C. Detecting GF between A and B is a difficult problem since this does not entail any topological change. Yang [2010] proposes a test of the null hypothesis of an absence of GF, drawing information from the extra variance in $A|B$ coalescence time expected in case of GF. A similar analysis is in principle possible in Aphid by testing the significance of the p_{AB} parameter. Unlike Yang's

method Aphid was not specifically designed for this goal, and we did not assess its performance with this respect in the current study.

The published method maybe most similar in spirit to Aphid is QuIBL [Edelman et al., 2019]. Like Aphid, QuIBL aims at partitioning the discordant gene trees into GF and non-GF categories via a mixture model. Unlike Aphid, QuIBL is an exact method in that it properly accounts for the among loci variance in coalescence time. This, however, comes at the cost of only focusing on a single branch, namely the internal branch (d according to our notations), the distribution of which can be analytically predicted conditionally on the topology. QuIBL did not perform particularly well in the simulations we conducted here. This is in part explained by the fact that the mutation rate varied substantially across loci in our simulations, whereas QuIBL assumes a constant mutation rate (Supplementary Fig. 2). It should also be noted that we simulated relatively short sequences - 100-1000 bp, to be compared to windows of 5kb in Edelman et al. [2019]. Because it analyzes only part of the available information, QuIBL likely requires more data than Aphid to achieve the same accuracy. One, however, should not conclude from this study that Aphid is generally superior to QuIBL. Further work is needed to characterize the applicability conditions for both methods.

Like QuIBL, Aphid assumes a constant N_e throughout the whole species tree, whereas CoalHMM and Yang's method allow the (AB) ancestral N_e to differ from the (ABC) ancestral N_e. Simulations indicate that the CoalHMM method is able to correctly estimate the two distinct ancestral N_e's when there is no GF [Dutheil et al., 2009]. Intuitively, it seems plausible that, in the absence of GF, the minimal sequence divergence between two genomes informs on the speciation time, and its variance on the ancestral N_e. In models accounting for GF, such as Aphid and QuIBL, the variance in coalescence time does not only depend on the ancestral N_e but also on the time and amount of GF, perhaps raising identifiability issues. To assume a constant N_e when GF is modelled, even if biologically less realistic, might therefore appear safer - a formal argument on this subject would be welcome, though.

Besides the interpretation of the phylogenetic conflict, explicitly modeling gene flow is likely to also make a difference in estimating parameters, particularly speciation times and ancestral effective population sizes. To assume no GF when there is some is expected to result in underestimated speciation times - due to recent GF-induced coalescences - and overestimated ancestral N_e - due to GF-induced discordances [Leaché et al., 2014]. Indeed, in apes our speciation time estimates are older, and our ancestral N_e smaller, than estimates obtained by the CoalHMM method [Rivas-González et al., 2023], which only considers ILS as a potential source of conflict. The Aphid speciation time estimates are consistent with a recent review of the palaeontological data in apes [Almécija et al., 2021], in which the human|chimpanzee split was dated at 7.5 Mya and the human|gorilla split at 11 Mya. The Aphid ancestral N_e estimate is close to current estimated N_e in humans, chimpanzee, bonobo and gorilla (20,000-60,000 according to Supplementary Table S2 in Rivas-González et al. [2023]).

6 Conclusion

Aphid is an efficient three-species method quantifying the contribution of gene flow and incomplete lineage sorting to the phylogenetic conflict, while providing gene flow-aware estimates of speciation times and ancestral effective population size.

7 Acknowledgments

The author thanks Iago Bonicci and Khalid Belkhir from the Montpellier Bioinformatics & Biodiversity platform, Julien Dutheil, Marjolaine Rousselle, Céline Scornavacca, and Carole Smadja for comments, advice and help.

8 Supplementary Material, Data accessibility

Aphid and documentation are available from `https://gitlab.mbb.cnrs.fr/ibonnici/aphid`. The supplementary figures, supplementary tables, data sets and scripts used in this study are available from `https://doi.org/10.48579/PRO/F6KEDM`.

References

C Scornavacca, F Delsuc, and N Galtier. *Phylogenetics in the genomic era*. Open Access Book, 2020. URL `https://hal.science/hal-02535070`.

RE Green, J Krause, AW Briggs, T Maricic, U Stenzel, M Kircher, N Patterson, H Li, W Zhai, MH Fritz, NF Hansen, EY Durand, AS Malaspinas, JD Jensen, T Marques-Bonet, C Alkan, K Prüfer, M Meyer, HA Burbano, JM Good, R Schultz, A Aximu-Petri, A Butthof, B Höber, B Höffner, M Siegemund, A Weihmann, C Nusbaum, ES Lander, C Russ, N Novod, J Affourtit, M Egholm, C Verna, P Rudan, D Brajkovic, Ž Kucan, I Gušic, VB Doronichev, LV Golovanova, C Lalueza-Fox, M de la Rasilla, J Fortea, A Rosas, RW Schmitz, PLF Johnson, EE Eichler, D Falush, E Birney, JC Mullikin, M Slatkin, R Nielsen, J Kelso, M Lachmann, D Reich, and S Pääbo. A draft sequence of the neandertal genome. *Science*, 328(5979), 2010. doi:10.1126/science.

SS Abby, E Tannier, M Gouy, and V Daubin. Lateral gene transfer as a support for the tree of life. *Proc Natl Acad Sci U S A*, 109(13), 2012. doi:10.1073/pnas.1116871109.

MC Fontaine, JB Pease, A Steele, RM Waterhouse, DE Neafsey, IV Sharakhov, X Jiang, AB Hall, F Catteruccia, E Kakani, SN Mitchell, YC Wu, HA Smith, RR Love, MK Lawniczak, MA Slotman, SJ Emrich, MW Hahn, and NJ Besansky. Mosquito genomics. extensive introgression in a malaria vector species complex revealed by phylogenomics. *Science*, 347(6217), 2015. doi:10.1126/science.1258524.

BS Meyer, M Matschiner, and W Salzburger. Disentangling incomplete lineage sorting and introgression to refine species-tree estimates for lake tanganyika cichlid fishes. *Syst Biol*, 66(4), 2017. doi:10.1093/sysbio/syw069.

J Ropars, C Maufrais, D Diogo, M Marcet-Houben, A Perin, N Sertour, K Mosca, E Permal, G Laval, C Bouchier, L Ma, K Schwartz, K Voelz, RC May, J Poulain, C Battail, P Wincker, AM Borman, A Chowdhary, S Fan, SH Kim, P Le, Pape, O Romeo, JH Shin, T Gabaldon, G Sherlock, ME Bougnoux, and C d'Enfert. Gene flow contributes to diversification of the major fungal pathogen candida albicans. *Nat Commun*, 9(1), 2018. doi:10.1038/s41467-018-04787-4.

NB Edelman, PB Frandsen, M Miyagi, B Clavijo, J Davey, RB Dikow, G García-Accinelli, SM Van, Belleghem, N Patterson, DE Neafsey, R Challis, S Kumar, GRP Moreira, C Salazar, M Chouteau, BA Counterman, R Papa, M Blaxter, RD Reed, KK Dasmahapatra, M Kronforst, M Joron, CD Jiggins, WO McMillan, F Di, Palma, AJ Blumberg, J Wakeley, D Jaffe, and J Mallet. Genomic architecture and introgression shape a butterfly radiation. *Science*, 366(6465), 2019. doi:10.1126/science.aaw2090.

S Glémin, C Scornavacca, J Dainat, C Burgarella, V Viader, M Ardisson, G Sarah, S Santoni, J David, and V Ranwez. Pervasive hybridizations in the history of wheat relatives. *Sci Adv*, 5(5), 2019. doi:10.1126/sciadv.aav9188.

D Zhang, FE Rheindt, H She, Y Cheng, G Song, C Jia, Y Qu, P Alström, and F Lei. Most genomic loci misrepresent the phylogeny of an avian radiation because of ancient gene flow. *Syst Biol*, 70(5), 2021. doi:10.1093/sysbio/syab024.

A Suvorov, C Scornavacca, MS Fujimoto, P Bodily, M Clement, KA Crandall, MF Whiting, DR Schrider, and SM Bybee. Deep ancestral introgression shapes evolutionary history of dragonflies and damselflies. *Syst Biol*, 71(3), 2022a. doi:10.1093/sysbio/syab063.

P Pamilo and M Nei. Relationships between gene trees and species trees. *Mol Biol Evol*, 5(5), 1988. doi:10.1093/oxfordjournals.molbev.a040517.

J Mallet, N Besansky, and MW Hahn. How reticulated are species? *Bioessays*, 38(2), 2016. doi:10.1002/bies.201500149.

JY Dutheil, G Ganapathy, A Hobolth, T Mailund, MK Uyenoyama, and MH Schierup. Ancestral population genomics: the coalescent hidden markov model approach. *Genetics*, 183(1), 2009. doi:10.1534/genetics.109.103010.

A Hobolth, JY Dutheil, J Hawks, MH Schierup, and T Mailund. Incomplete lineage sorting patterns among human, chimpanzee, and orangutan suggest recent orangutan speciation and widespread selection. *Genome Res*, 21(3), 2011. doi:10.1101/gr.114751.110.

D Guerzoni and A McLysaght. De novo genes arise at a slow but steady rate along the primate lineage and have been subject to incomplete lineage sorting. *Genome Biol Evol*, 8(4), 2016. doi:10.1093/gbe/evw074.

FK Mendes and MW Hahn. Gene tree discordance causes apparent substitution rate variation. *Syst Biol*, 65(4), 2016. doi:10.1093/sysbio/syw018.

O Meleshko, MD Martin, TS Korneliussen, C Schröck, P Lamkowski, J Schmutz, A Healey, BT Piatkowski, AJ Shaw, DJ Weston, KI Flatberg, P Szövényi, K Hassel, and HK Stenøien. Extensive genome-wide phylogenetic discordance is due to incomplete lineage sorting and not ongoing introgression in a rapidly radiated bryophyte genus. *Mol Biol Evol*, 38(7), 2021. doi:10.1093/molbev/msab063.

I Rivas-González, M Rousselle, F Li, L Zhou, JY Dutheil, K Munch, Y Shao, D Wu, MH Schierup, and G Zhang. Pervasive incomplete lineage sorting illuminates speciation and selection in primates. *Science*, 380(6648), 2023. doi:10.1126/science.abn4409.

MS Hibbins and MW Hahn. Phylogenomic approaches to detecting and characterizing introgression. *Genetics*, 220(2), 2022. doi:10.1093/genetics/iyab173.

B Rannala and Z Yang. Efficient bayesian species tree inference under the multispecies coalescent. *Syst Biol*, 66(5), 2017. doi:10.1093/sysbio/syw119.

C Long and L Kubatko. The effect of gene flow on coalescent-based species-tree inference. *Syst Biol*, 67(5), 2018. doi:10.1093/sysbio/syy020.

D Wen and L Nakhleh. Coestimating reticulate phylogenies and gene trees from multilocus sequence data. *Syst Biol*, 67 (3), 2018. doi:10.1093/sysbio/syx085.

T Flouri, X Jiao, B Rannala, and Z Yang. A bayesian implementation of the multispecies coalescent model with introgression for phylogenomic analysis. *Mol Biol Evol*, 37(4), 2020. doi:10.1093/molbev/msz296.

EY Durand, N Patterson, D Reich, and M Slatkin. Testing for ancient admixture between closely related populations. *Mol Biol Evol*, 28(8), 2011. doi:10.1093/molbev/msr048.

SH Martin, JW Davey, and CD Jiggins. Evaluating the use of abba-baba statistics to locate introgressed loci. *Mol Biol Evol*, 32(1), 2015. doi:10.1093/molbev/msu269.

JB Pease and MW Hahn. Detection and polarization of introgression in a five-taxon phylogeny. *Syst Biol*, 64(4), 2015. doi:10.1093/sysbio/syv023.

PD Blischak, J Chifman, AD Wolfe, and LS Kubatko. Hyde: A python package for genome-scale hybridization detection. *Syst Biol*, 67(5), 2018. doi:10.1093/sysbio/syy023.

D Vanderpool, BQ Minh, R Lanfear, D Hughes, S Murali, RA Harris, M Raveendran, DM Muzny, MS Hibbins, RJ Williamson, RA Gibbs, KC Worley, J Rogers, and MW Hahn. Primate phylogenomics uncovers multiple rapid radiations and ancient interspecific introgression. *PLoS Biol*, 18(12), 2020. doi:10.1371/journal.pbio.3000954.

MT Holder, JA Anderson, and AK Holloway. Difficulties in detecting hybridization. *Syst Biol*, 50(6), 2001. doi:10.1080/106351501753462911.

S Joly, PA McLenachan, and PJ Lockhart. A statistical approach for distinguishing hybridization and incomplete lineage sorting. *Am Nat*, 174(2), 2009. doi:10.1086/600082.

Smith J and Kronforst MR. Do heliconius butterfly species exchange mimicry alleles? *Biology Letters*, 9(4), 2013. doi:10.1098/rsbl.2013.0503.

A Suvorov, BY Kim, J Wang, EE Armstrong, D Peede, ERR D'Agostino, DK Price, P Waddell, M Lang, V Courtier-Orgogozo, JR David, D Petrov, DR Matute, DR Schrider, and AA Comeault. Widespread introgression across a phylogeny of 155 drosophila genomes. *Curr Biol*, 32(1), 2022b. doi:10.1016/j.cub.2021.10.052.

C Scornavacca, K Belkhir, J Lopez, R Dernat, F Delsuc, EJP Douzery, and V Ranwez. Orthomam v10: Scaling-up orthologous coding sequence and exon alignments with more than one hundred mammalian genomes. *Mol Biol Evol*, 36(4), 2019. doi:10.1093/molbev/msz015.

C Scornavacca and N Galtier. Incomplete lineage sorting in mammalian phylogenomics. *Syst Biol*, 66(1), 2017. doi:10.1093/sysbio/syw082.

Y Song, C Jiang, KH Li, J Li, H Qiu, M Price, ZX Fan, and J Li. Genome-wide analysis reveals signatures of complex introgressive gene flow in macaques (genus macaca). *Zool Res*, 42(4), 2021. doi:10.24272/j.issn.2095-8137.2021.038.

A Hobolth, OF Christensen, T Mailund, and MH Schierup. Genomic relationships and speciation times of human, chimpanzee, and gorilla inferred from a coalescent hidden markov model. *PLoS Genet*, 3(2), 2007. doi:10.1371/journal.pgen.0030007.

Z Yang. Likelihood and bayes estimation of ancestral population sizes in hominoids using data from multiple loci. *Genetics*, 162(4), 2002. doi:10.1093/genetics/162.4.1811.

JL Thorne, H Kishino, and IS Painter. Estimating the rate of evolution of the rate of molecular evolution. *Mol Biol Evol*, 15(12), 1998. doi:10.1093/oxfordjournals.molbev.a025892.

N Lartillot, T Lepage, and S Blanquart. Phylobayes 3: a bayesian software package for phylogenetic reconstruction and molecular dating. *Bioinformatics*, 25(17), 2009. doi:10.1093/bioinformatics/btp368.

Z Yang. A likelihood ratio test of speciation with gene flow using genomic sequence data. *Genome Biol Evol*, 2, 2010. doi:10.1093/gbe/evq011.

AD Leaché, RB Harris, B Rannala, and Z Yang. The influence of gene flow on species tree estimation: a simulation study. *Syst Biol*, 63(1), 2014. doi:10.1093/sysbio/syt049.

S Almécija, AS Hammond, NE Thompson, KD Pugh, S Moyà-Solà, and DM Alba. Fossil apes and human evolution. *Science*, 372(6542), 2021. doi:10.1126/science.abb4363.