Host-symbiont-gene phylogenetic reconciliation

Hugo Menet, Alexia Nguyen Trung, Vincent Daubin, Eric Tannier

Peer Community Journal

Section: Evolutionary Biology

RESEARCH ARTICLE

Published
2023-05-22

Cite as
Hugo Menet, Alexia Nguyen Trung, Vincent Daubin and Eric Tannier (2023) *Host-symbiont-gene phylogenetic reconciliation* , Peer Community Journal, 3: e47.

Correspondence
vincent.daubin@cnrs.fr
eric.tannier@inria.fr

Peer-review
Peer reviewed and recommended by PCI Evolutionary Biology, https://doi.org/10.24072/pci.evolbiol.100593

Host-symbiont-gene phylogenetic reconciliation

Hugo Menet[1], Alexia Nguyen Trung[1], Vincent Daubin[1], and Eric Tannier[1,2]

Volume **3** (2023), article e47

https://doi.org/10.24072/pcjournal.273

Abstract

Motivation Biological systems are made of entities organized at different scales (*e.g.* macro-organisms, symbionts, genes…) which evolve in interaction. These interactions range from independence or conflict to cooperation and coevolution, which results in them having a common history. The evolution of such systems is approached by phylogenetic reconciliation, which describes the common patterns of diversification between two different levels, *e.g.* genes and species, or hosts and symbionts for example. The limit to two levels hides the multi-level inter-dependencies that characterize complex systems. **Results** We present a probabilistic model of evolution of three nested levels of organization which can account for the codivergence of hosts, symbionts and their genes. This model allows gene transfer as well as host switch, gene duplication as well as symbiont diversification inside a host, gene or symbiont loss. It handles the possibility of ghost lineages as well as temporary free-living symbionts. Given three phylogenetic trees, we devise a Monte Carlo algorithm which samples evolutionary scenarios of symbionts and genes according to an approximation of their likelihood in the model. We evaluate the capacity of our method on simulated data, notably its capacity to infer horizontal gene transfers, and its ability to detect hostsymbiont co-evolution by comparing host/symbiont/gene and symbiont/gene models based on their estimated likelihoods. Then we show in a aphid enterobacter system that some reliable transfers detected by our method, are invisible to classic 2-level reconciliation. We finally evaluate different hypotheses on human population histories in the light of their coevolving *Helicobacter pylori* symbionts, reconciled together with their genes. **Availability** Implementation is available on GitHub https://github.com/hmenet/TALE. Data are available on Zenodo https://doi.org/10.5281/zenodo.7667342.

[1]Univ Lyon, Université Lyon 1, CNRS, Laboratoire de Biométrie et Biologie Évolutive UMR5558, F-69622 Villeurbanne, France, [2]Inria, Centre de recherche de Lyon, 69622 Villeurbanne, France

Contents

1. Introduction

The toolbox of evolutionary biology largely relies on the assumption of statistical independence of biological objects at any level of organization: organisms from different species are isolated from a biological system based on their genomes, genomes are cut into independent genes, and inside genes, nucleotides are evolving independently from each other (Felsenstein, 2003).

Yet the essence of living systems lies in dependence: constraint, cooperation or conflict (Sapp, 1994). Symbiotic micro-organisms coevolve with animals or plants (JL Sonnenburg and ED Sonnenburg, 2019). The ensemble they form is gathered under the holobiont concept. It allows to see genes as entities not only following their own interest, not only participating to the functioning of the genome they are hosted by, but also participating to, and probably evolving with, a larger biological system.

A powerful tool to study these inter-dependencies is phylogenetic reconciliation: an ensemble of models and methods explaining the differences and similarities between phylogenies of two entities diversifying concommitantly in evolution. Gene/species systems have been studied by phylogenetic reconciliation, accounting for events of gene duplication, horizontal gene transfer and gene loss (DTL model) (Boussau and Scornavacca, 2020; Doyon et al., 2011; Menet et al., 2022; Nakhleh, 2013; Szöllősi et al., 2015b). The same model can be applied with little to no modification to symbiont/host (Charleston and Libeskind-Hadas, 2014; Donati et al., 2015; Santichaivekin et al., 2020), protein domain/gene cophylogeny (Rasmussen and Kellis, 2012; Stolzer et al., 2015), and biogeography has been imagined as one possible level (Martínez-Aquino, 2016;

Ree and Smith, 2008; Ronquist, 1997). DTL models have also been used to reconstruct genome histories (Duchemin et al., 2015), detect highways of lateral gene transfers in bacteria, archaea or eukaryota (Bansal et al., 2011), assess the relative role of duplication and gene transfer in the evolution of genomes (Sjöstrand et al., 2014), infer ancient symbiotic relationships (Bailly-Bechet et al., 2017), reconstruct histories of gene fusion and fission (Duchemin et al., 2017), model endosymbiotic gene transfer (Anselmetti et al., 2021).

A limitation of reconciliation methods is their separate application on molecular studies on one side (gene/species cophylogeny), and ecological studies on the other (host/symbiont cophylogeny). The striking methodological unity of the two (the same DTL model is applied on both the molecular and ecological systems) and the growing interest for multi-level systems integrating molecular and ecological inter-dependencies (e.g. the holobiont concept) calls for a unique model for host, symbiont, gene cophylogeny. In support of this claim, a number of empirical studies already rely on host symbiont histories when proposing horizontal gene transfers between symbionts (Manzano-Marín et al., 2019; Nakabachi et al., 2013; Nikoh et al., 2014; Penz et al., 2012), when often, only symbiont gene/species comparisons do not provide enough statistical support for them (Ravenhall et al., 2015; Wijayawardena et al., 2013).

Three level reconciliations have been introduced by Stolzer et al. (2015) and applied to protein domain, gene and species. They describe two embedded DTL models and an inference method by parsimony. The inference method first reconciles genes and species trees in a DTL model. Then, knowing which genes are present in which species, it reconciles the protein domains with the genes. This defines two kinds of horizontal protein domain transfers between genes, depending on whether the genes are in the same species (which we will call "intra" transfer) or not ("inter" transfer), with a different cost for those two events. Further efforts in this direction have been published by Li and Bansal (2019a) with a duplication/loss model between gene and species and a DTL model, forbidding inter species transfers, between protein domains and genes. They show NP-hardness of inferring the most parsimonious couple of nested reconciliations (Li and Bansal, 2019a) and propose different heuristics and problem variants (Li and Bansal, 2018, 2019b). A probabilistic model without transfers has been proposed by Muhammad et al. (2018). It aims at inferring dated gene trees from protein domain alignments using Markov Chain Monte Carlo. These attempts prove that it is possible to jointly handle three nested levels in a single computational model. However none of them can yet handle host/symbiont/gene systems in a statistical framework because of specific limitations of each of them: parsimony framework, no transfer or no inter-host transfer, no joint inference between levels of organization, no explicit handling of absent lineages.

We propose a probabilistic model that describes the evolution of three nested entities at three different scales, adapted to a host/symbiont/gene system. In our model a symbiont tree is generated by a DTL model inside the host, with a possibility of evolving temporarily outside the host phylogeny. A gene is generated by a DTL model inside the symbiont, where gene transfer is more likely between symbionts that share a common host ("intra" transfer) than for those that do not ("inter" transfer).

Based on this model we propose an inference method extending the two-level reconciliation "ALE" software (Szöllősi et al., 2015a, 2013). It takes three trees as input, constructs joint scenarios and estimates event rates and likelihoods according to the model. Our implementation also features the possibility to infer a symbiont species tree if only the host tree and several symbiont gene trees are given as input. In addition a comparison of the likelihood of two-level and three-level reconciliations can be used as a test for multi-scale coevolution.

We report a benchmark test of the inference method on simulated data, using an external simulator (Kundu and Bansal, 2019), showing that under the hypothesis that gene transfers are more likely between symbionts of a same host, the three-level reconciliation represent a significant gain compared to the two-level one in terms of the capacity to retrieve the symbiont donors and receivers of horizontal gene transfers.

We use the inference method to identify horizontal gene transfers between *Cinara* aphid symbionts that are detected by expertise (Manzano-Marín et al., 2019) but missed by two-level reconciliations.

Finally we show on genes of *Helicobacter pylori* from human populations how likelihood computations can be used to compare different hypotheses on the diversification of a host, given the genes of its symbionts, taking into account the evolutionary dependencies between all three scales.

2. 2-level reconciliation, definitions and preliminaries

Because we base our model on the two-level DTL reconciliation model implemented in "ALE undated" (Morel et al., 2020; Szöllősi et al., 2015a), together with the inference methods, this section is devoted to their brief description.

2.1. Model and parameters

We denote by G, S respectively a set of gene trees and a species tree, and $\delta_D^S, \delta_S^T, \delta_S^L$ is the set of rates at which a gene evolving in a branch of S undergo the D,T,L (speciation, duplication, transfer, loss) events. These rates are constant along the species tree and for all gene trees.

The model is generates a rooted phylogenetic tree G, given S and the rates, according to a birth and death like model. A gene tree can originate in any branch of the species tree with a uniform prior. Speciation occurs at all nodes of S, while duplications, transfers and loss can occur along the branches of S with the given rates.

When a transfer occurs, the receiver branch is chosen according to a uniform probability, avoiding ancestor branches of the donor. This avoids certain impossible transfers but is not sufficient to guarantee that the overall scenario is time feasible. Indeed, two transfers might be incompatible with respect to time (Davín et al., 2018).

2.2. Inference

The core of the inference method consists in computing the probability $P_{\theta_S}(G|S)$ of generating G given S and $\theta_S = (p_S^S, p_S^D, p_S^T, p_S^L)$, the probabilities of S,D,T,L events, proportional to the rates and satisfying $p^S + p^D + p^T + p^L = 1$. That is, $P_{\theta_S}(G|S)$ is the likelihood of G, S and θ_S. S is assumed binary and rooted. G is binary but can be rooted or not. A mapping of the leaves of G to the leaves of S is needed (the species in which each extant gene is found).

We call *reconciliation scenario* a list of events of kind D, T, L, or S associated to each internal gene tree node, that can be the result of the birth and death process. These lists transcribe into a mapping of the nodes of G to the nodes of S. We note $R_{G,S}$ the set of all possible reconciliation scenarios by which G can be produced from S. The likelihood of a scenario $r \in R_{G,S}$, $P_{\theta_S}(r|S)$ is the product of the probabilities of all events. Thus we have

$$P_{\theta_S}(G|S) = \sum_{r \in R_{G,S}} P_{\theta_S}(r|S)$$

We do not need to fully enumerate all scenarios to compute this sum. Indeed, a dynamic programming scheme along S and G allows us to sum over scenarios individually on each branch of S and ensures tractability. The dynamic programming scheme consists first in a "forward step" traversing the nodes of G and S in post-order: a node is examined only if its children have been examined before.

If e, u are nodes of S and G, f and g are descendants of e, v and w are descendants of u (if any of these do not exist the corresponding terms must be dropped), and $P_{e,u} = P_{\theta_S}(e, u)$ is the probability of generating the subtree of G rooted at u in the subtree of S rooted at e, then:

$$
\begin{aligned}
P_{e,u} =& p^S \left(P_{g,v}P_{f,w} + P_{g,w}P_{f,v} + E_f P_{g,u} + P_{f,u}E_g\right) \\
&+ p^D \left(P_{e,v}P_{e,w} + 2P_{e,u}E_e\right) \\
&+ \left(\frac{1}{|S|}\sum_h p^T P_{h,w}\right) P_{e,v} + \left(\frac{1}{|S|}\sum_h p^T P_{h,v}\right) P_{e,w} \\
&+ \left(\frac{1}{|S|}\sum_h p^T E_h\right) P_{e,u} + \left(\frac{1}{|S|}\sum_h p^T P_{u,h}\right) E_e,
\end{aligned} \tag{1}
$$

where E_e is the probability that a gene on branch e of S goes extinct:

$$E_e = p^L + p^S E_f E_g + p^D E_e E_e + \left(\frac{1}{|S|}\sum_h p^T E_h\right) E_e.$$

The sum of probabilities at the root of G, for all node of S, gives $P_{\theta_S}(G|S)$.

A "backward step" then traverses the nodes in the reverse order. It allows one to sample the scenarios based on their probability, or to select the scenario that maximises the marginal likelihood (Yang, 2006): this means, at each step of the backtracking procedure we select the scenario with maximum likelihood.

Note that this is different from finding the most likely reconciliation scenario. It is possible to find it by a similar procedure, storing the maximum probability in the forward step instead of the sum of the probabilities, and computing the scenario realizing this maximum in the backward step, as in a parsimony algorithm. We did not use this possibility, sticking to the ALE principle.

2.3. About time consistency

Simulated scenarios according to the model, and inferred reconciliations do not need to be *time consistent*: a set of transfers might indicate histories that are not feasible on a timeline. This is a known drawback of undated models (Davín et al., 2018). There have been attempts to investigate this aspects in several directions. For example, Eucalypt (Donati et al., 2015) or Notung (Stolzer et al., 2015) propose to infer only time feasible scenarios, without any guarantee that such a scenario exists or is can be found in reasonable computing time. Producing a time feasible scenario is NP-complete. Moreover, inferring only time feasible scenarios for one gene family or one symbiont, depending on the biological context, does not guarantee that the combination of scenarios from several gene families or several symbiont will be time consistent: a set of transfers from different genes might not be consistent.

Producing time consistent scenarios with several gene families or symbiont goes back to producing a dated tree (Chauve et al., 2017).

On the other hand, measuring the degree of inconsistency can make this hindsight a strength: for example, one can compare scenarios in relation to this consistency, with the assumption that the more consistent the scenario, the more realistic it is. We used this to compare 2-level and 3-level scenarios in the evaluation of the method by simulations.

3. 3-level reconciliation, likelihood estimation and scenario inference

3.1. Elements of the probabilistic model

The 3-level model is based on two nested 2-level models based on the one presented in the previous section, with the following extensions and restrictions.

3.1.1. Host/Symbiont. A *host* tree H is unique, given, rooted. Inside H, a *symbiont* tree S is generated with the DTL 2-level model, with parameters $\delta_H^D, \delta_H^L, \delta_H^T$, adding the possibility for a symbiont to live temporarily in an unknown host.

Indeed, in the course of their evolutionary history, some symbionts may live outside a host, or within an unknown host. This is a general interesting feature, and is particularly important for us because we invoke unknown hosts in the inference process in the case of inter host horizontal gene transfers (section 3.3).

The utility of this model addition is visible in the *Cinara* aphids example developed in the Results section (see Fig. 6).

3.1.2. Host/Symbiont/Gene. The evolution of a gene tree G inside H/S reconciliations also follows an adaptation of the DTL model. G is generated in one or several symbiont trees with duplication, loss and intra horizontal transfer, with rates $\delta_S^D, \delta_S^L, \delta_S^T$. "Intra" means that horizontal transfer is possible only between symbiont branches (from the same symbiont tree or not) that are present in the same host branch (as the trees S are generated in H).

Note that the gene tree G can refer to a family of genes found in symbionts as well as the host. In the latter case, to remain generic, we simply assume that S can be a copy of H, reconciled with H with only speciations. That is, the host genes are contained in a specific compartment and can transfer to a symbiont, and be transferred from a symbiont.

An illustration of the realization of such a model is given in Figure 1.

This model can be immediately used for simulations, but we chose to use an external simulator for our tests (Kundu and Bansal, 2019). Though this does not allow an identifiability study, which we postpone to a future work, it controls some of the effects of similarities in models and implementation between simulation and inference, providing more difficult instances for testing.

3.2. Monte Carlo approximation of the likelihood

Like in the previous section, the inference consists in estimating the parameters (trees and evolutionary rates) and sampling reconciliation scenarios. We consider as input a single rooted binary tree H, one or several rooted or unrooted binary symbiont trees $S = \{S_i\}$, and one or several rooted or unrooted binary gene trees $G = \{G_i\}$. Both parameter estimation and sampling are accomplished through a calculation of the probability $P(G|S, H)$ that gene trees G have been generated by the model, given H, S, and given the DTL probabilities for the two reconciliation levels $\theta_S = (p_S^S, p_S^D, p_S^T, p_S^L)$ and $\theta_H = (p_H^S, p_H^D, p_H^T, p_H^L)$ derived from the rates: the DTL probabilities are proportional to the rates and the sum of all three probabilities is 1.

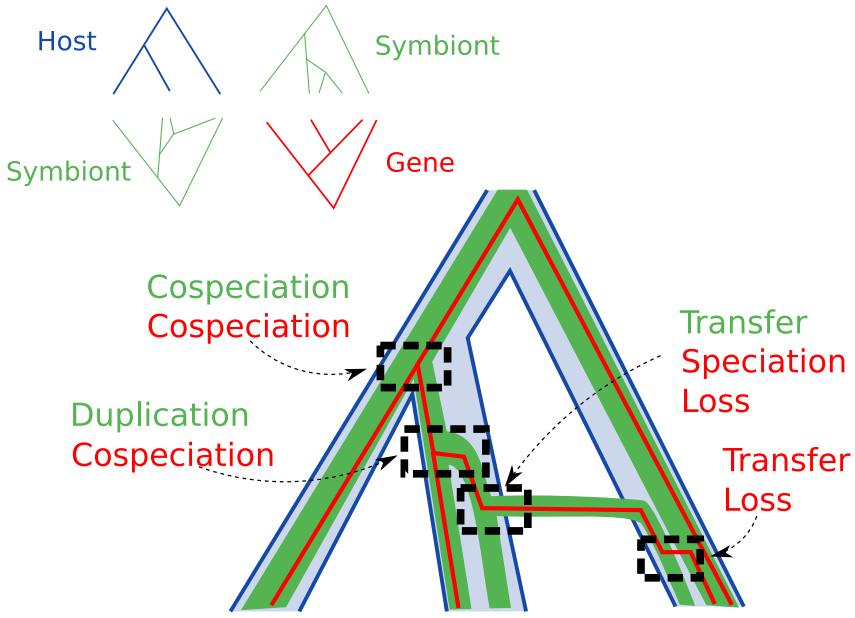


Figure 1 – An example of a 3-Level reconciliation input (top left, with three trees and associations between the leaves of two couples of trees) and a possible reconciliation scenario for this input. Events of the host/symbiont co-evolution are written in red, while events of the symbiont/gene reconciliation are written in green.

The probability $P(G|S,H)$ can be decomposed by summing over all possible host/symbiont reconciliation scenarios $r_{S,H} \in R_{S,H}$:

$$P(G|S,H) = \sum_{r_{S,H} \in R_{S,H}} P(G|S,H,r_{S,H}) P(r_{S,H}|S,H). \quad (2)$$

The number of reconciliations in this sum is at least exponential in the size of the input (even the number of scenarios maximizing $P(r_{S,H}|S,H)$ can be exponential (Donati et al., 2015)). The similar computation in a parsimonious framework is NP-hard (Li and Bansal, 2019a), so it is probably not possible to exactly and quickly compute $P(G|S,H)$.

So we apply a Monte Carlo approximation technique. The goal is to sample a reasonable number N of symbiont/host reconciliations and approximate $P(G|S,H)$:

$$P(G|S,H) \simeq \frac{1}{N} \sum_{n=1}^{N} P(G|S,H,r_n) \quad (3)$$

where r_n is sampled in the set $R_{S,H}$ of all reconciliations according to its likelihood $P(r_n|S,H)$. In consequence the term in equation 3 approximates the term in equation 2 according to the Monte Carlo principle.

3.3. Reconciliation inference

The computation of $P(G|S,H)$, as well as sampling reconciliations in $R_{S,H}$, is done by successive steps of dynamic programming as shown in Algorithm 1. Steps 2 and 8 are the exact executions of the algorithm ALE (Szöllősi et al., 2013), with the additional possibility that a symbiont is free living. Free living symbiont are handled by adding a copy of the symbiont tree as an additional host tree. Indeed the reconciliation algorithm can accommodate multiple host trees on separate sets of leaves. Symbiont leaves with no host are matched to themselves instead of a host. In that way, we hypothesize that transfer between free living is less likely than when a common host is known.

Given $r_n \in R_{S,H}$, the probability $P(G|S,H,r_n)$ can be computed with an adaptation of the same dynamic programming algorithm (step 15 of Algorithm 1). The only modification is that during the dynamic programming process, for all gene transfer possibilities, it is checked if the donor and receiver symbiont share a host in r_n. If they do, then it is an "intra" transfer and the transfer has the probability defined by the transfer rate.

3.4. Inter species transfer through ghost species

Transfer between two symbionts in different hosts is possible through ghost species. Indeed it is always reasonable to assume that a major part of species are extinct or not sampled and gene transfers are often "from the dead" (Fournier et al., 2009; Szöllosi et al., 2013; Tricou et al., 2022; Zhaxybayeva and Gogarten, 2004).

In consequence, a transfer can have occured from a donor that is now extinct. Figure 2 shows how an "inter" transfer between symbionts i and j (on the left) can occur, even if it is not explicitly modeled, through a sister lineage to i, that switched host and transferred a gene to j (on the right). The sister lineage then goes extinct, which explains that the gene looks like it is transferred from i to j.

We denote by $P_S^T(i \to j)$ the probability for a gene present in symbiont i to undergo a horizontal transfer to symbiont j, and $P_H^T(e \to h)$ the probability for a gene present in a symbiont associated to host e to transfer to a symbiont associated to host h. Let H_i (H_j) be the set of host branches that contain symbiont i (resp. j). We go from P_H^T to P_S^T by summing over all possibles hosts h of the receiver symbiont j and all hosts e of the donor symbiont i:

$$P_S^T(i \to j) = \sum_{e \in H_i, h \in H_j} P_H^T(e \to h) \quad (4)$$

Algorithm 1 The Monte Carlo inference algorithm

Input. H is a rooted binary tree, θ_H the probabilities of symbiont events, $S = \{S_i\}$ is a set of symbiont trees, and θ_S the probabilities of gene events, $G = \{G_i\}$ is a set of gene trees.

(1) **for** all symbiont tree in S_i **do**
(2) Compute $P(S_i|H)$ with the "extended" 2-level DTL model.
(3) (2-level dynamic programming allowing free living symbionts H, S forward)
(4) **end for**
(5) Repeat steps 1-4 (5 times, parameterizable) to optimize θ_H.
(6) **for** n in $[0 \dots N]$ ($N = 100$, number of samples, parameterizable) **do**
(7) **for** all S_i **do**
(8) Sample a reconciliation $r_{n,i}$ with probability proportional to $P_{\theta_H}(r_{n,i}|H, S_i)$
(9) (2-level dynamic programming H, S backward)
(10) **end for**
(11) Construct $r_n = \cup_i r_{n,i}$ a reconciliation of the set S with H.
(12) **end for**
(13) **for** all r_n **do**
(14) **for** all gene tree G_i **do**
(15) Compute $P(G_i|S, H, r_n)$
(16) (3-level dynamic programming $G|S, H, r_n$ forward)
(17) Sample scenarios of reconciliation between G_i and S knowing r_n.
(18) (3-level dynamic programming $G|S, H, r_n$ backward)
(19) **end for**
(20) Compute $P(G|S, H, r_n) = \prod P(G_i|S, H, r_n)$
(21) **end for**
(22) Approximate $P(G|S, H)$ by $\frac{1}{N} \sum_{n=1}^{N} P(G|S, H, r_n)$
(23) Repeat steps 13-22 (5 times, parameterizable) to estimate θ_S

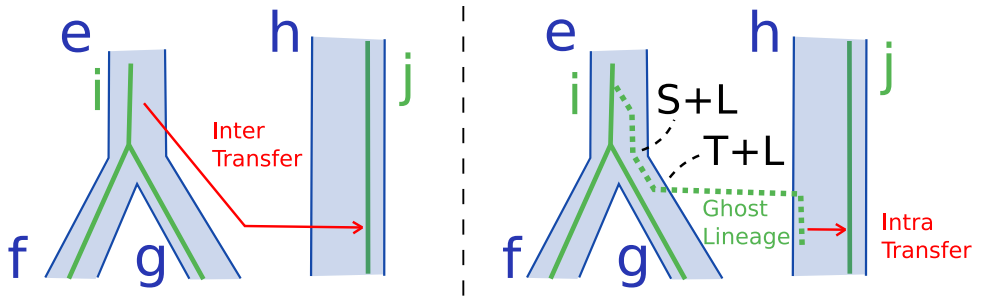


Figure 2 – A gene transfer between two symbiont lineages that are in different host lineages (inter transfer) is explained with intra-transfers and ghost lineages. Left part shows the host phylogeny (blue pipes), the reconciled symbiont phylogeny (green lines) and a gene transfer (in red) from lineage i to lineage j, while i is in host lineage e and j is in host lineage h. This direct inter transfer is forbidden by the model. Right part shows a mechanism allowed by the model that has the exact same result, and the way to compute the associated probability. First the symbiont lineage i undergoes a speciation and a loss (S+L), and then a transfer and a loss (T+L) before the extinction of the symbiont (or its absence in the taxon sampling) inside j. Now the gene transfer (in red) is an intra transfer, as it is transfered between two symbionts inside h.

At fixed h we rewrite with $P_e = P^T(e \to h)$. Recall p_S^T are the probability of horizontal transfer in the symbiont/gene reconciliation, and $p_H^S, p_H^D, p_H^T, p_H^L$ the probabilities of speciation, duplication, transfer and loss in the host/symbiont reconciliation. Let E_e be the probability of extinction, that is, the probability that a gene is present in a branch e of the host tree and absent from all the leaves. Let $|S_h|$ be the number of symbiont branches matched to host h in the host/symbiont reconciliation scenario. The initial case in our inductive definition of $P_e = P^T(e \to h)$ is the case

where $e = h$, so when the donor symbiont is in one of the receiver symbiont host, in that case the probability to transfer to that one symbiont of h, is uniform among the $|S_h|$ symbionts present in h. Then, for the induction, we rewrite the undated reconciliation equations, to progress a symbiont in the host tree from any host e to host h of the receiver symbiont and such that the symbiont species we invoke then goes extinct. The notations are similar to those used in the undated ALE description in (Morel et al., 2020), Section 2 or figure 2: we denote by f, g the children of a host e, and by $|H|$ the number of nodes in $|H|$.

$$\begin{cases} P_e &= \frac{1}{|S_h|} p_S^T \text{ if } e = h \\ P_e &= p_H^S(P_f E_g + P_g E_f) + 2p_H^D P_e E_e + \sum_{k \in H} \frac{p_H^T}{|H|} P_k E_e \end{cases} \tag{5}$$

Note that the last sum in the equation is limited to the k that are not ancestors of e, as in ALE. This equation has a self dependency due to the Transfer/Loss event, which is already accounted for in reconciliation methods (Jacox et al., 2016; Szöllősi et al., 2013). We forbid successions of several Transfer/Loss events to break this self dependency and solve this equation.

3.5. Sequential and 2-level estimation of the likelihood

Because the Monte Carlo approach can be computationally heavy, we devised an alternative "Sequential" heuristic. Instead of sampling scenarios randomly like in the Monte Carlo, we select only one of them, maximizing the marginal likelihood (Yang, 2006). That is, at each step of the backtracking of the dynamic programming procedure we select the event maximizing the probability in the sum of Equation (1). In other words, we decompose $P_{(\theta_S,\theta_H)}(G|S,H)$ into

$$P_{(\theta_S,\theta_H)}(G|S,H) \simeq P_{\theta_S}(G|S,H,\hat{r}_{S,H}) P_{\theta_H}(\hat{r}_{S,H}|S,H), \tag{6}$$

where $\hat{r}_{S,H}$ is the reconciliation scenario maximizing the marginal likelihood. Note that is can be different from taking the most likely scenario, which is also a possible strategy, consisting in changing the Equation (1) from a sum to a max, and backtracking in this alternative dynamic programming table. So this variant consists in removing the "for" loop of step 6 of Algorithm 1 and replacing step 8 by a systematic choice of a maximum instead of choosing in the sum of Equation 1 with probabilities proportional to the term values.

This approach is similar to the one of Stolzer et al. (2015). The differences are, apart from using a probabilistic setting, that we use marginal likelihood, and that we compute the inter transfer probabilities from the host/symbiont and symbiont/gene DTL reconciliation parameters instead of using an additional parameter (described in the previous sections).

The faster Sequential heuristic may not be as robust as the Monte Carlo one. Li and Bansal (2019b) present an example where the sequential approach cannot propose a solution at all, in a parsimony model where inter horizontal gene transfer are forbidden. In figure 3 we present another illustration, with this time an emphasis on the "not continuous" aspect of the Sequential heuristic in regard to the host and symbiont reconciliation events rates.

A small change in the transfer rate of the host and symbiont makes a big difference for the gene and symbiont reconciliation with the Sequential heuristic, but a small one for the Monte Carlo one, see the results in table 1.

3.6. Time complexity and tractability

We denote h, s, g the number of nodes of the host, symbiont, and gene trees respectively.

It has been demonstrated that 2-level parsimony DTL reconciliations can be computed in quadratic time (Bansal et al., 2012) if all transfers have the same probability.

In our implementation sampling one host symbiont reconciliation scenario (line 8 in Algorithm 1) is done in cubic time $O(hs^2)$ complexity because we parse the transfer sum from equation 1.

Computing the gene transfer probabilities between all couples of symbiont nodes (section 3.3) is done with a dynamic programming similar to the one for reconciliation in $O(hs)$, presented in equation 5. A final sum (equation 4) over all hosts of the considered symbionts in $O(h^2s^2)$, in

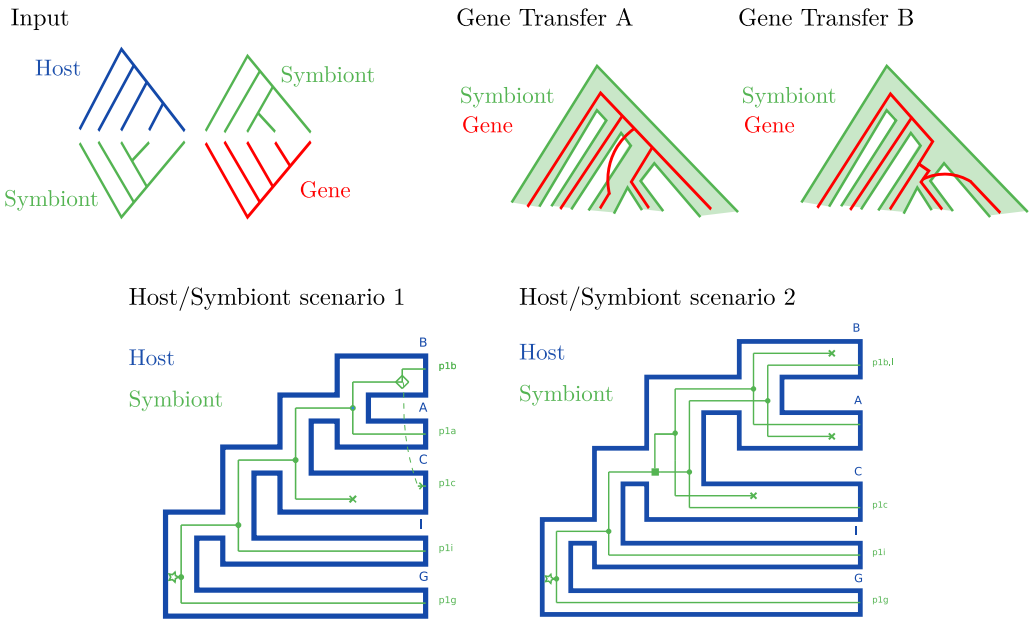


Figure 3 – An example of input where the Sequential heuristic is less robust than the Monte Carlo one. We compare the support for two gene transfer scenarios, scenario A and B. There are two main possible host/symbiont reconciliation scenarios, scenario 1 and scenario 2. In scenario 1, gene transfer A is more likely, and in scenario 2, gene transfer B is more likely (both gene transfers involve ghost species, whatever the scenario. The support for both gene transfers for the Sequential and Monte Carlo heuristics are presented in table 1.

Table 1 – Comparison of the support for the two gene transfer scenarios in the example presented in figure 3. Column 1 contains the method: Monte Carlo (Section 3.2), Sequential (Section 3.4) and 2-level (which consists in reconciling G with S without information form H). Then columns 2 and 3 contain the support of gene transfers, respectively A and B (in reference to Figure 3), according to reconciliation scenario 1 or 2, obtained with different transfer probabilities.

Heuristic	Gene transfer A	Gene transfer B
Host Symbiont rates **T 0.006** D 0.1 L 0.1		
Monte Carlo	0.43	0.27
Sequential	0.90	< 0.05
2-level	0.18	0.21
Host Symbiont rates **T 0.005** D 0.1 L 0.1		
Monte Carlo	0.35	0.33
Sequential	< 0.05	0.49
2-level	0.19	0.23

the reasonable case where the number of symbiont nodes per host nodes (in the reconciliation scenario) is below a constant k, yields $O(h^2k^2 + hs)$ for this part.

Finally the host aware gene/symbiont reconciliation (line 15) differs with classic 2-level reconciliation in that transfer rates depend on the donor-receiver couple. In consequence we cannot use the efficient computation trick used for uniform rates (Bansal et al., 2012; Szöllősi et al., 2013), that enable to compute equation 1 without computing for each couple of gene and symbiont subtrees the transfer sum. Here for each couple of gene and symbiont subtrees, we must explicitly consider transfers toward all symbiont nodes, yielding a cubic complexity of $O(s^2g)$ for host aware symbiont/gene reconciliation.

This leads to a total complexity of $O(N(hs + h^2k^2 + s^2g))$ where k is a bound on the number of symbionts per host in the sampled reconciliations (s in the worst case), and N is the number of samples in the Monte Carlo approach.

The datasets presented give a good idea of the size of the data we can consider with this new method. We here give the computation time for the Sequential heuristic. Computation on the Cinara aphid dataset, with a size of 25 leaves for the symbiont tree, 9 leaves for the host, and 13 gene families takes about 3 minutes on a single laptop core, including the rate estimation steps. This is a dataset on which it would be possible to use the Monte Carlo approach. The pylori dataset is larger, the symbiont has 119 leaves, the host 7 leaves, and there are 1034 gene families, of which 322 have 119 leaves. Reconciliation, with fixed rates (without rate estimation) took just under a day using 8 cores.

3.7. Symbiont tree inference

In case the symbiont tree is unknown, we devised an option to infer the symbiont tree by amalgamation (David and Alm, 2011; Szöllősi et al., 2013) of universal unicopy gene trees, guided by the host tree.

Clade prior probabilities are computed from universal unicopy gene trees, and dynamic programming is used to compute the likelihood. A symbiont tree is sampled in the backtracking phase at the same time as the host/symbiont reconciliation scenario.

This amalgamation is also implemented for the symbiont/gene part, to account for gene tree being unrooted, and to be able to include uncertainty in gene tree topology, just like in 2-level reconciliations(Jacox et al., 2016; Szöllősi et al., 2013).

3.8. Rates estimation and likelihood comparison

In our model, the data is the gene trees, and the free parameters are the three DTL probabilities of the symbiont/gene reconciliation. We consider the host/symbiont DTL parameters as fixed, *i.e.* estimated without knowing the data. This makes it possible to compare, based on the likelihood, our approach and a 2-level one (symbiont/gene reconciliation, unaware of the host), because they have the same free parameters, and because they both define a probability distribution on the same space of gene trees associated to the symbiont tree.

In practice we estimate the host/symbiont DTL parameters, as done in ALE (Szöllősi et al., 2015b), with an expectation maximization method, and then fix these parameters. Then we run the Monte Carlo or sequential approach multiple times to estimate rates for the symbiont/gene reconciliation with the same expectation maximisation method.

3.9. Output format and solution visualization

Our implementation can output a sample of full scenarios, both for symbiont/genes and the corresponding host/symbiont reconciliations. The scenarios are given in RecPhyloXML, a common standard for reconciliation output endorsed by a significant part of the gene/species reconciliation community (Duchemin et al., 2018). The scenarios can be visualised using Thirdkind `https://crates.io/crates/thirdkind` (Penel et al., 2022), a reconciliation viewer that handles 3-level reconciliations. We also output event frequencies based on the reconciliation scenario sampling. Indeed we sample a number (100 by default) of symbiont/gene reconciliations and observe the frequency of each event in these replicas, thus obtaining an estimate of the posterior probability of events. It is this result that we use to evaluate the ability of our method to infer specific events, such as receptors and donors of horizontal symbiont transfers, which we compare to simulated scenarios or previously proposed scenarios on aphids Cinara.

4. Experimental results

4.1. Simulated datasets

4.1.1. Description of the simulation process. Our probabilistic model can be used for simulation, however in order to test our method, we chose to use an exterior simulation framework. We

used the available software Sagephy developed by Kundu and Bansal (2019). Sagephy generates three embedded trees and allows replacing transfers on top of additive ones. We used the parameters proposed by the same team in another article (Kordi et al., 2019), as representative of small (D 0.133, T 0.266, L 0.266), medium (D 0.3, T 0.6, L 0.6) and high (D 0.6, T 1.2, L 1.2) transfer rates, without replacing transfers. The software enables to specify an inter transfer rate, corresponding to the probability for a gene transfer to be between symbionts hosted by different hosts ("inter" transfer). When a horizontal transfer is chosen during generation of the gene tree (inside a symbiont tree and knowing a host/symbiont reconciliation), the transfer is chosen to be an inter host one with the inter transfer rate. So an inter transfer rate of 0 corresponds to our inference model of only intra transfer, and of 1 corresponds to a case where transfers are only between symbionts in separate hosts.

We constructed two simulated datasets, one with a combination of the different rates for the DTL parameters, and one with only medium rates but with different inter transfer rates. For the first dataset, we used all 9 combinations of small, medium and high rates for the symbiont generation and the gene generation, with only intra host gene transfer (*i.e.* an inter transfer rate of zero). For the second dataset, we used only medium rates for both symbiont and genes generation, but we used 6 inter transfer rates going from 0 to 1.

For both datasets, and for each set of rates, we generated 50 instances consisting of 1 host tree with 100 leaves, 1 symbiont tree and 5 gene trees, each generated in the pruned version of the other trees (branches that do not reach present are pruned before the generation of the next tree). We then selected host leaves with a probability of 0.08 to simulate unexhaustive sampling, resulting in host trees with an average size of 8 leaves. We thus simulate extinct lineages, and even with a simulation inter transfer rate of 0, some gene transfers will be inter. This ended up to 399 instances for the first dataset and 226 instances for the second one, and at least 29 instances of 5 genes for each set of parameters.

We compared the results from three approaches. (1) The "2-level" heuristic which is a 2-level reconciliation between the gene and symbiont trees, ignorant of the host tree. (2) The "Sequential" heuristic, which consists in computing the most likely host/symbiont DTL reconciliation and doing the symbiont/gene reconciliation, given that host/symbiont reconciliation. (3) The full 3-level "Monte Carlo" method, summing the results of the gene reconciliations over 50 sampled host/symbiont reconciliation scenarios. We let our approaches estimate evolutionary rates.

We measured first the capacity of the three methods to infer the correct symbiont donor and recipient of gene transfers (with precision and recall), and second, the likelihood they attribute to symbiont/gene cophylogeny. Identifying the exact donor and recipient of simulated transfers is usually considered a hard task for reconciliation algorithms. Usually reconciliation studies are not evaluated with this strong criterion (Mykowiecka et al., 2018), but with the inference of ancestral characters (Wieseke et al., 2015), the number of transfers (Szöllősi et al., 2012), the ability to infer better trees (Bansal et al., 2015), or the ability to map the correct event type to each gene node (Kordi et al., 2019). We chose to look at the capacity to infer specific transfers because we feel that it is in this task that our model has the capacity to show its utility. It can infer more precise gene transfers because transfers are constrained by additional elements compared to other methods.

Our probabilistic reconciliation approaches output estimates of the posterior probabilities of evolutionary events, so we used these probabilities as weights for our precision and recall definition in Figure 4 for the detection of horizontal gene transfer donor and receiver symbionts. Denoting by $L_{t,sim}$ the list of simulated transfers and $L_{t,obs}$ the list of observed transfers, and $P_{obs}(T)$ the estimation of our approach for the probability of transfer T.

$$\text{Precision} = \frac{\sum_{T \in L_{t,sim}} P_{obs}(T)}{\sum_{T \in L_{t,obs}} P_{obs}(T)} \text{ and Recall } = \frac{\sum_{T \in L_{t,sim}} P_{obs}(T)}{\sum_{T \in L_{t,sim}} 1} \tag{7}$$

4.1.2. The 3-level method infers more true transfers than the 2-level method. Overall the Monte Carlo and sequential approaches give similar results on these simulated datasets, and better results (in particular for recall and to a lesser extent for precision) than the 2-level approach

(Figure 4). In most cases, the faster Sequential heuristic can advantageously replace the Monte Carlo one because they have the same recall and precision. In a few case, that might be the more interesting ones, the Monte Carlo has a slight advantage, and though it is more computationally costly, it is also theoretically more robust.

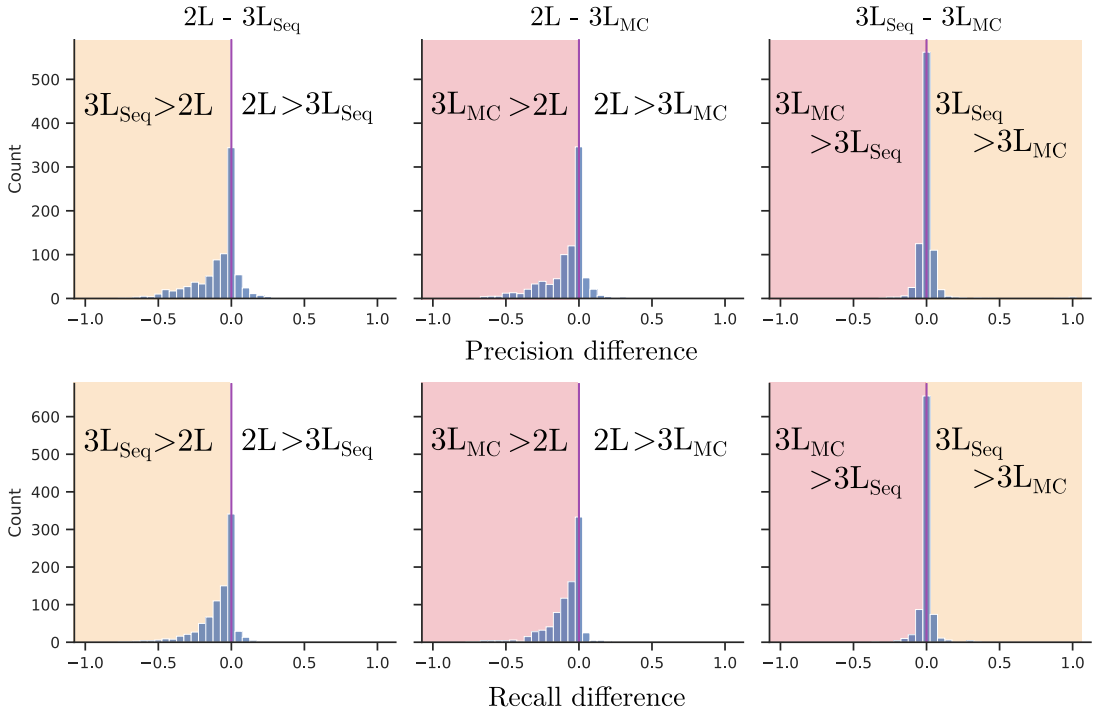


Figure 4 – Distribution of differences of precision and recall on the inference of horizontal gene transfers for all combinations of two approaches: 2-level (2L), 3-level with the Monte Carlo heuristic ($3L_{MC}$) and 3-level with the Sequential heuristic ($3L_{Seq}$), centered on 0, and for all 874 gene families of the 3-level simulation, with no inter host gene transfer, that undergo at least one transfer.

In addition we measured the time consistency of reconciliation scenarios in the 2-level and 3-level inferences. Indeed, we have already remarked that we work in an undated framework, and in consequence transfers might be incompatible (Davín et al., 2018). For each simulation condition we listed all inferred transfers and checked compatibility. For 2-level reconciliations, 35% of the conditions lead to time incompatibilities, this same measure dropping to 15% if 3-level reconciliations were performed.

4.1.3. A host-symbiont co-evolution test. The reconciliation likelihood difference between 3-level inference and 2-level inference is a marker of host-symbiont co-evolution. Indeed, Figure 5 (A) shows that when the simulation model is less dependent from the host phylogeny (inter transfer rates of 0.6, 0.8 and 1.0), the likelihood difference between the 2-level and 3-level inference methods are mostly in favor of the 2-level. It happens for almost all instances in the simulation dataset with with no intra transfers (inter transfer rates of 1.0), the farthest one from the model behind our heuristic that privileges intra transfers. For all these instances a preference for 2-level reconciliation (according to the likelihood) is more likely when few transfers are inferred (we sum over 1 to 5 gene families generated for each host and symbiont instance). This is a sign of the precision of the method to not classify 2-level instances as 3-level ones.

In a model with only intra transfers (inter transfer rate of 0), we have a very good recall for the detection of the 3-level model, almost all only intra transfer instances are classified as 3-level as they should be. A more detailed exam of this recall is presented in Figure 5 (B) with the first simulated dataset, with only intra transfer and varying DTL parameters.

Figure 5 (B) shows the likelihood difference when only intra transfers occur in the simulations. We see that when the number of transfers is higher, the likelihood difference better reflects the

mode of simulation. In practice a way to increase the number of transfers is to increase the number of gene families considered.

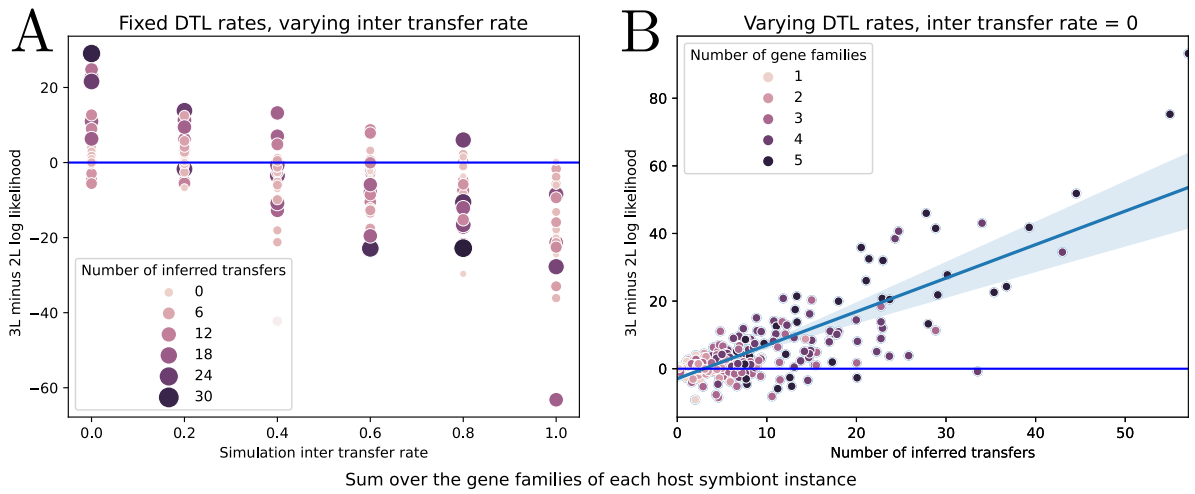


Figure 5 – A test of host symbiont co-evolution. We measure the difference of likelihood between the 3-level model and the 2-level model, using the estimation of these likelihoods provided by our "2-level" and "3-level Sequential" heuristics, in order to differentiate instances where gene trees are generated in a 3-level host/symbiont/gene model or in a 2-level symbiont/gene model. Each instance is composed of a host tree, a symbiont tree, and 1 to 5 gene families. For one instance we sum the differences over all gene families. (A) Sensitivity of the likelihood difference to the value of the inter host gene transfer probability in Sagephy. As expected, the more an inter transfer rate is probable (independent from the host phylogeny), the less we detect host-symbiont co-evolution with the likelihood difference measure. Colors indicate the number of inferred transfers. (B) Sensitivity of the likelihood difference to the number of inferred transfers (dataset with only intra transfers). Colors depict the number of gene families considered in the host and symbiont instance. Because transfers carry the co-evolution signal, the sensitivity of the method increases with the number of transfers, which are higher if we increase the number of gene families.

4.2. Precise identification of a gene transfer in enterobacteria symbiotic of *Cinara* aphids

A recent study on Cinara aphids enterobacteria systems (Manzano-Marín et al., 2019) identified one host switch and two horizontal gene transfers, one intra-host from *Erwinina* to *Hamiltonella* and one inter-host from *Sodalis* to *Erwinia*. The genes transferred (thi) and some others (bioa,d,b) were first inherited through gene transfers, probably from *Sodalis* related symbionts. Moreover, those genes transferred are part of functions to complement the lack in the sap-feeding host nutrition. It seems that a new endosymbiont acquires the genes of another one to sustain the host. This exemplifies a case where a symbiont gene can co-evolve with the symbiont host, more than with the symbiont itself. We reproduced this scenario in Figure 6 (A), and a representative gene tree witnessing the transfers is reproduced in Figure 6 (B).

Gene trees including *Cinara* endosymbionts and other enterobacteria species were available from the supplementary material made available by Manzano-Marín et al. (2019). *Cinara* and their endosymbionts phylogenies show exact correspondences on the studied period. We kept all enterobacteria associated to a *Cinara* aphid (of *Erwinia* and *Hamiltonella* genus), and chose a representative subset of the other enterobacteria present in the gene trees, notably containing *Sodalis* species, closest identified parent to one of the transferred genes, and other *Erwinia* and *Hamiltonella* genus species. We used the phylogeny proposed in Annotree for these species (Mendler et al., 2019), to complement the *Cinara* aphids symbionts phylogeny proposed in (Manzano-Marín et al., 2019). We used our 3-level reconciliation on the host tree and symbiont tree, using the possibility of our method to take into account these "free living" bacteria. As the host and symbiont (apart from the free living) are identical, we used the sequential heuristic.

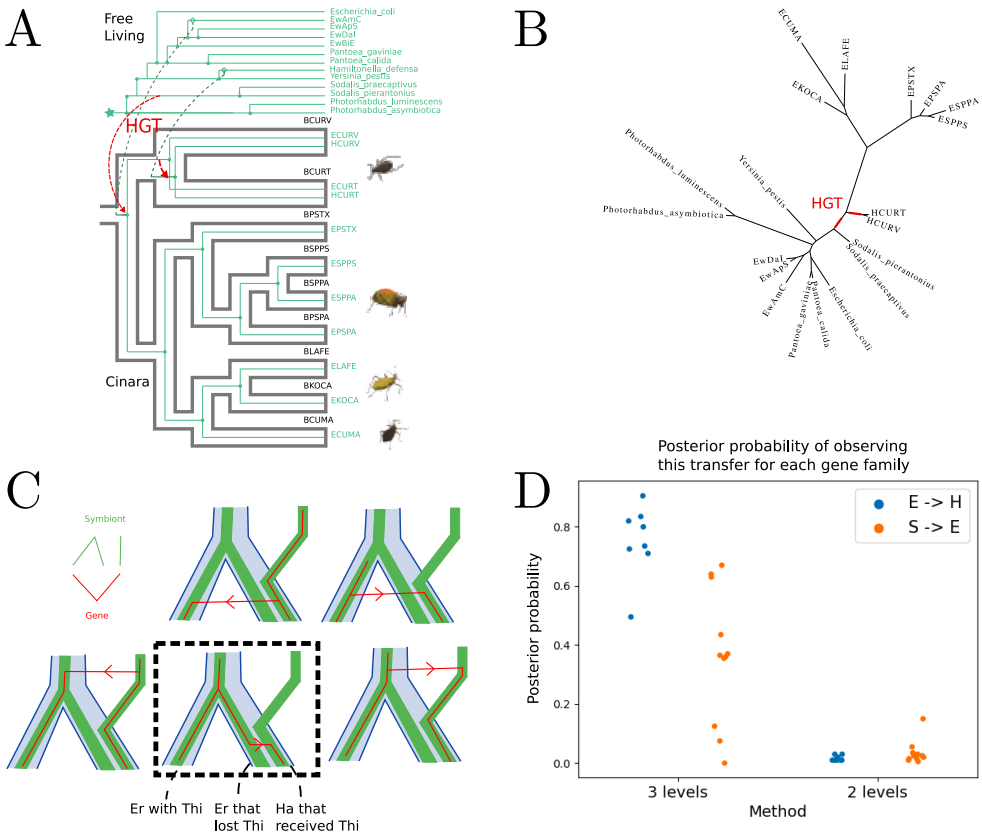


Figure 6 – The evolution of *Cinara* and their enterobacteria symbionts. (A) The evolutionary scenario identified by Manzano-Marín et al. (2019). The reconciliation of the hosts (*Cinara* aphids) and symbionts (bacteria) are depicted along with the position of the horizontal gene transfers (in red). (B) Phylogenetic tree of one gene with the position of the two transfers. (C) Theoretical explanation of the difference between the results of the 2-level and 3-level reconciliation methods. The two top reconciliations are a bit more likely in a 2-level framework, as they require a single transfer while the bottom ones require a transfer and a loss, but one of the bottom one (with the dotted square) is better in a 3-level model, as it allows an intra-host transfer. (D) Support (a posteriori probability of the transfer, computed from its observed frequency in the reconciliation sample) for the identified HGTs, from *Erwinia* to *Hamiltonella*, and from *Sodalis* to *Erwinia*, for 3-level and 2-level reconciliations.

We tested the capacity of the 3-level method compared to a 2-level one to detect the gene transfers identified by Manzano-Marín et al. (2019). The intra transfer from *Erwinina* to *Hamiltonella* is retrieved in around 80 percent of the scenarios sampled by the 3-level method, and both are better retrieved than in the method that does not take the host into account (Figure 6 (D)). A theoretical explanation using a toy example is given in Figure 6 (C). An alternative transfer, in the other direction, from *Hamiltonella* to *Erwinia* is slightly more likely but the configuration of the host evolution supports the intra transfer.

This exemplifies how multi-scale dependencies can only be captured by 3-level models.

4.3. *Helicobacter pylori* genes as documents for human migrations

Helicobacter pylori is a bacterial symbiont of a significant proportion of humans, which has been supposed to be a marker of human migrations across the Earth (Achtman, 2016). Bacterial strains have been divided in different populations corresponding to geographical areas (Africa 1, Africa 2, Asia 2, East Asia, North East Africa, Europe) (Mégraud et al., 2016; Waskito and Yamaoka, 2019).

The supposed coevolving complex made by humans, bacterial symbiont and their genes makes it an accessible system for the host/symbiont/gene reconciliation method. In particular gene transfers should be more probable between *Helicobacter* strains if they are hosted by a same human population.

We collected available current strains of *H. pylori* from the NCBI which have a genetic population assigned by MLST allelic profile (Achtman et al., 1999; Jolley et al., 2018). A phylogenetic tree was built based on the concatenation of universal-unicopy genes (322 gene families), and a sample of 113 strains representing the diversity of H. pylori in the old world (excluding strains from the Americas) was obtained using Treemmer (Menardo et al., 2018). Then, 6 non pylori strains were added (*H. hepaticus*, *H. acinonychis*, *H. canadensis*, *H felis*, *H. bizzozeronii*, *H. cetorum*), as external groups.

In this study we considered the 1034 gene families, including 322 universal unicopy families, that displayed strains from the external groups and from at least 3 continents.

We then considered four different population trees (host trees) containing the geographical areas as leaves, coherent with the scientific literature (Mégraud et al., 2016; Waskito and Yamaoka, 2019). 322 universal unicopy gene trees were used, and the strain (symbiont) tree was amalgamated from gene trees with the population trees as a guide (see subsection 3.7). As strains were much more numerous than populations, and subject to a more complex diversification than DTL events, we allowed an additional event, named I, that consists in a duplication followed by a speciation and loss of one of the copies, with a specific rate, inferior to the combination of these three events. This event allows a strain to be present in a population and one of its descendants, and is used as one of the default events in biogeography frameworks (Ree et al., 2005).

We then applied our sequential approach and compared the likelihood of the gene/strains aware of the host reconciliation to compare the population trees. The results are depicted in Figure 7 (A). The likelihood of the systems according to the population tree is reported, divided into two components: the likelihood of the population/strain comparison, and the likelihood of the gene/strain aware of the population comparison. The population tree on the left column is the most likely given the model, the method and the used data. Assessing the robustness of the result would require a sensibility study which is out of the scope of this contribution.

Figure 7 (B) is an illustration of a reconciliation scenario for the maximum likelihood host tree with Thirdkind (Penel et al., 2022). We see the host tree and the amalgamated strain tree reconciled (I events are represented as transfers from a parent node to one of its child). On top of these two embedded trees red lines represent the aggregation of gene transfers depending on the host of the donor and receiver strains. The opacity of the transfer lines are proportional to the number of times a certain kind of transfer is observed across the 1034 gene families in one sampled scenario.

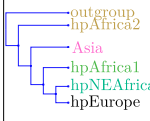
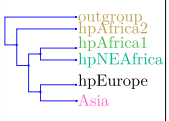
5. Discussion

In a review on horizontal gene transfer in host symbiont systems (Wijayawardena et al., 2013) the authors highlight the need of plurality of evidence to robustly assess the existence of transfers. Evidence can be of multiple types, gene trees, donor receiver ecology, or host symbiont association. We provide a framework were these multiple evidence can be gathered, and the proof of concept that it can work, on *Cinara* aphids and their enterobacteria.

Our method uses a probabilistic framework that enables rate estimation, tree inference, tree comparison and model comparison. We also introduced a method to compute the inter transfer rate from the intra transfer one and the modeling of ghost lineages in the host symbiont reconciliation. We introduced a Monte Carlo approach that enables to estimate event probabilities and likelihood, by sampling through multiple host symbiont scenarios in a double DTL model. Implementation is available on GitHub https://github.com/hmenet/TALE.

While our intuition is that the Monte Carlo approach is more robust than the sequential one, notably in cases where gene events happen around uncertain host symbiont reconciliation nodes, our evaluation on simulated data did not show a big difference in most cases. We think that in biological data, we can expect more interaction between the events of the host symbiont

A

Pylori putative population tree				
Symbiont tree	Amalgamation of the 322 universal unicopy gene trees			
Host/symbiont Log likelihood	- 308	- 320	- 322	- 320
Gene trees	1034 gene trees			
Symbiont/gene Log likelihood	-653x10^3	-658x10^3	-658x10^3	-676x10^3

B

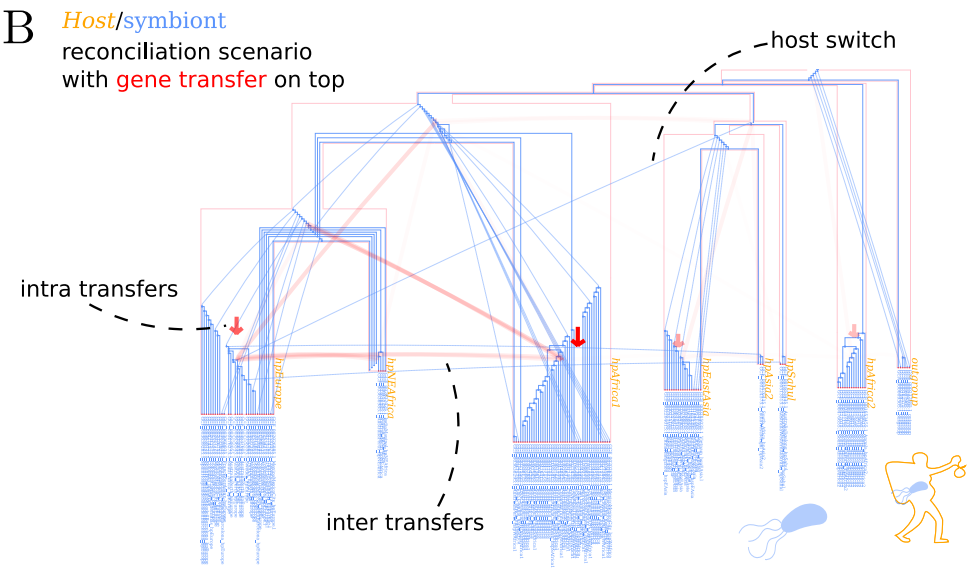


Figure 7 – Co-evolution of human populations and *Helicobacter pylori*. (A) Log likelihood of the different population trees. (B) The representation with ThirdKind (Penel et al., 2022) of one possible reconciliation scenario of *Helicobacter pylori* strain tree and the population tree maximizing likelihood. Aggregated gene transfers are depicted on top of the DTL reconciliation, with the opacity corresponding to the number of times the transfers were seen across the 1034 gene families.

reconciliation and the ones of the gene symbiont one, which are independent in our simulation. Developing new simulation frameworks that can model such dependencies, for instance by increasing the loss rates when multiple genes or symbionts are present, or using a functional approach to the evolution of genes, could be important to the understanding of these multi-level models.

The ability of our inference methods to be used for model comparison seems promising. We saw that with an increasing number of gene families we could increase our confidence in the answer. However the different gene families must contain a part of independent information, as is the case in the simulation where all families dependence are completely in the host and symbiont trees. For instance in the *Cinara* aphids dataset, the genes considered are mostly similar, and do not really make the number of independent transfers increase, and with only one intra transfer, that necessitates an additional loss to occur, the 2-level model displays a better likelihood than the 3-level. If more independent transfers were present, we can suppose that some of them might not necessitate such a loss and the test would favor a host symbiont co-evolution.

All these features deserve further tests to know their domain of validity and to draw biological conclusions. In particular, the inference of the symbiont tree, with the use of amalgamation,

from an input distribution of universal unicopy gene tree would deserve to be tested against other standard methods as concatenate or species tree reconstruction with 2-level reconciliation model as it is implemented in SpeciesRax (Morel et al., 2022).

An interesting future direction in this line would be to construct, instead of a symbiont tree, compartment trees, which would depict the evolution of inter-dependent genes that are not necessarily in the same species.

A comparison of the inference method to similar ones (Li and Bansal, 2019a; Muhammad et al., 2018; Stolzer et al., 2015) could also be undertaken. However in an host/symbiont/gene framework, horizontal transfer in the host/symbiont reconciliation are crucial, and only the model of Stolzer et al. (2015) takes these events into account. Moreover the sequential heuristic is simply a rewriting of this model in a probabilistic framework.

More generally, the model is not bound to host/symbiont/gene systems, but any set of three nested inter-dependent entities can be studied with it: species/gene/protein domain as it was done in previous studies (Li and Bansal, 2018; Muhammad et al., 2018; Stolzer et al., 2015), or geography/species/gene, and so on. As the scales of biological observation are probably infinite, so are the combination of three nested scales.

Examples presented in this article show the possibilities of the method, but still derive no biologically significant breakthrough. However the necessity of such a method, detecting multi-level co-evolution, could arise with the more and more numerous studied biological systems that fit into this multi-scale cophylogeny framework, notably with an increasing interest for hologenomics (Alberdi et al., 2022).

Acknowledgements

This work was performed using the computing facilities of the CC LBBE/PRABI.

Preprint version 2 of this article has been peer-reviewed and recommended by Peer Community In Evolutionary Biology (https://doi.org/10.24072/pci.evolbiol.100593) (Jousselin, 2023).

Fundings

This work was supported by the French National Research Agency (Grant ANR-19-CE45-0010 Evoluthon).

Conflict of interest disclosure

The authors declare that they comply with the PCI rule of having no financial conflicts of interest in relation to the content of the article.

The authors declare the following non-financial conflict of interest: Vincent Daubin and Eric Tannier are both recommenders of PCI Evolutionary Biology.

Data, script, code, and supplementary information availability

Data are available online on Zenodo: https://doi.org/10.5281/zenodo.7667342 (Menet et al., 2023). Script and codes are available online on Github : https://github.com/hmenet/TALE;

References

Achtman M (2016). *How old are bacterial pathogens? Proceedings of the Royal Society B: Biological Sciences* **283**, 20160990. https://doi.org/10.1098/rspb.2016.0990.

Achtman M, Azuma T, Berg DE, Ito Y, Morelli G, Pan ZJ, Suerbaum S, Thompson SA, Van Der Ende A, Van Doorn LJ (1999). *Recombination and clonal groupings within Helicobacter pylori from different geographical regions*. *Molecular Microbiology* **32**, 459–470. https://doi.org/https://doi.org/10.1046/j.1365-2958.1999.01382.x.

Alberdi A, Andersen SB, Limborg MT, Dunn RR, Gilbert MTP (2022). *Disentangling host–microbiota complexity through hologenomics*. *Nature Reviews Genetics* **23**. Number: 5 Publisher: Nature Publishing Group, 281–297. https://doi.org/10.1038/s41576-021-00421-0.

Anselmetti Y, El-Mabrouk N, Lafond M, Ouangraoua A (2021). *Gene tree and species tree reconciliation with endosymbiotic gene transfer*. *Bioinformatics* **37** (Supplement_1), i120–i132. https://doi.org/10.1093/bioinformatics/btab328.

Bailly-Bechet M, Martins-Simões P, Szöllősi GJ, Mialdea G, Sagot MF, Charlat S (2017). *How Long Does Wolbachia Remain on Board? Molecular Biology and Evolution* **34**, 1183–1193. https://doi.org/10.1093/molbev/msx073.

Bansal MS, Alm EJ, Kellis M (2012). *Efficient algorithms for the reconciliation problem with gene duplication, horizontal transfer and loss*. *Bioinformatics* **28**. Publisher: Oxford Academic, i283–i291. https://doi.org/10.1093/bioinformatics/bts225.

Bansal MS, Banay G, Gogarten JP, Shamir R (2011). *Detecting Highways of Horizontal Gene Transfer*. *Journal of Computational Biology* **18**, 1087–1114. https://doi.org/10.1089/cmb.2011.0066.

Bansal MS, Wu YC, Alm EJ, Kellis M (2015). *Improved gene tree error correction in the presence of horizontal gene transfer*. *Bioinformatics* **31**, 1211–1218. https://doi.org/10.1093/bioinformatics/btu806.

Boussau B, Scornavacca C (2020). *Reconciling Gene trees with Species Trees*. In: *Phylogenetics in the Genomic Era*. Ed. by Celine Scornavacca, Frédéric Delsuc, and Nicolas Galtier. No commercial publisher | Authors open access book, 3.2:1–3.2:23. URL: hal.science/hal-02535529.

Charleston M, Libeskind-Hadas R (2014). *Event-Based Cophylogenetic Comparative Analysis*. In: *Modern Phylogenetic Comparative Methods and Their Application in Evolutionary Biology*. Springer Berlin Heidelberg, pp. 465–480. https://doi.org/10.1007/978-3-662-43550-2_20. URL: https://doi.org/10.1007/978-3-662-43550-2_20.

Chauve C, Rafiey A, Davín AA, Scornavacca C, Veber P, Boussau B, Szöllősi Gj, Daubin V, Tannier E (2017). *MaxTiC: Fast ranking of a phylogenetic tree by Maximum Time Consistency with lateral gene transfers*. *bioRxiv*, 127548. https://doi.org/10.1101/127548.

David LA, Alm EJ (2011). *Rapid evolutionary innovation during an Archaean genetic expansion*. *Nature* **469**, 93–96. https://doi.org/10.1038/nature09649.

Davín AA, Tannier E, Williams TA, Boussau B, Daubin V, Szöllősi GJ (2018). *Gene transfers can date the tree of life*. *Nature Ecology & Evolution* **2**. Number: 5 Publisher: Nature Publishing Group, 904–909. https://doi.org/10.1038/s41559-018-0525-3.

Donati B, Baudet C, Sinaimeri B, Crescenzi P, Sagot MF (2015). *EUCALYPT: efficient tree reconciliation enumerator*. *Algorithms for Molecular Biology* **10**, 3. https://doi.org/10.1186/s13015-014-0031-3.

Doyon JP, Ranwez V, Daubin V, Berry V (2011). *Models, algorithms and programs for phylogeny reconciliation*. *Briefings in Bioinformatics* **12**, 392–400. https://doi.org/10.1093/bib/bbr045.

Duchemin W, Anselmetti Y, Patterson M, Ponty Y, Bérard S, Chauve C, Scornavacca C, Daubin V, Tannier E (2017). *DeCoSTAR: Reconstructing the Ancestral Organization of Genes or Genomes Using Reconciled Phylogenies*. *Genome Biology and Evolution* **9**, 1312–1319. https://doi.org/10.1093/gbe/evx069.

Duchemin W, Daubin V, Tannier E (2015). *Reconstruction of an ancestral Yersinia pestis genome and comparison with an ancient sequence*. *BMC Genomics* **16**, S9. https://doi.org/10.1186/1471-2164-16-S10-S9.

Duchemin W, Gence G, Arigon Chifolleau AM, Arvestad L, Bansal MS, Berry V, Boussau B, Chevenet F, Comte N, Davín AA, Dessimoz C, Dylus D, Hasic D, Mallo D, Planel R, Posada D, Scornavacca C, Szöllősi G, Zhang L, Tannier É, et al. (2018). *RecPhyloXML: a format for reconciled gene trees*. *Bioinformatics* **34**, 3646–3652. https://doi.org/10.1093/bioinformatics/bty389.

Felsenstein J (2003). *Inferring Phylogenies*. Oxford, New York: Oxford University Press. 580 pp.

Fournier GP, Huang J, Gogarten JP (2009). *Horizontal gene transfer from extinct and extant lineages: biological innovation and the coral of life*. *Philosophical Transactions of the Royal Society B: Biological Sciences* **364**, 2229–2239. https://doi.org/10.1098/rstb.2009.0033.

Jacox E, Chauve C, Szöllősi GJ, Ponty Y, Scornavacca C (2016). *ecceTERA: comprehensive gene tree-species tree reconciliation using parsimony*. *Bioinformatics (Oxford, England)* **32**, 2056–2058. https://doi.org/10.1093/bioinformatics/btw105.

Jolley KA, Bray JE, Maiden MCJ (2018). *Open-access bacterial population genomics: BIGSdb software, the PubMLST.org website and their applications*. *Wellcome Open Research* **3**, 124. https://doi.org/10.12688/wellcomeopenres.14826.1.

Jousselin E (2023). *Reconciling molecular evolution and evolutionary ecology studies: a phylogenetic reconciliation method for gene-symbiont-host systems*. Peer Community in Evolutionary Biology. https://doi.org/10.24072/pci.evolbiol.100593.

Kordi M, Kundu S, Bansal MS (2019). *On Inferring Additive and Replacing Horizontal Gene Transfers Through Phylogenetic Reconciliation*. In: *Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*. BCB '19. Niagara Falls, NY, USA: Association for Computing Machinery, pp. 514–523. https://doi.org/10.1145/3307339.3342168.

Kundu S, Bansal MS (2019). *SaGePhy: an improved phylogenetic simulation framework for gene and subgene evolution*. *Bioinformatics* **35**, 3496–3498. https://doi.org/10.1093/bioinformatics/btz081.

Li L, Bansal MS (2018). *An Integer Linear Programming Solution for the Domain-Gene-Species Reconciliation Problem*. In: *Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*. BCB '18. event-place: Washington, DC, USA. New York, NY, USA: ACM, pp. 386–397. https://doi.org/10.1145/3233547.3233603.

Li L, Bansal MS (2019a). *An Integrated Reconciliation Framework for Domain, Gene, and Species Level Evolution*. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* **16**, 63–76. https://doi.org/10.1109/TCBB.2018.2846253.

Li L, Bansal MS (2019b). *Simultaneous Multi-Domain-Multi-Gene Reconciliation Under the Domain-Gene-Species Reconciliation Model*. In: *Bioinformatics Research and Applications*. Lecture Notes in Computer Science. Springer International Publishing, pp. 73–86. https://doi.org/10.1007/978-3-030-20242-2_7.

Manzano-Marín A, Coeur d'acier A, Clamens AL, Orvain C, Cruaud C, Barbe V, Jousselin E (2019). *Serial horizontal transfer of vitamin-biosynthetic genes enables the establishment of new nutritional symbionts in aphids' di-symbiotic systems*. *The ISME Journal*, 1–15. https://doi.org/10.1038/s41396-019-0533-6.

Martínez-Aquino A (2016). *Phylogenetic framework for coevolutionary studies: a compass for exploring jungles of tangled trees*. *Current Zoology* **62**, 393–403. https://doi.org/10.1093/cz/zow018.

Mégraud F, Lehours P, Vale FF (2016). *The history of Helicobacter pylori: from phylogeography to paleomicrobiology*. *Clinical Microbiology and Infection* **22**, 922–927. https://doi.org/10.1016/j.cmi.2016.07.013.

Menardo F, Loiseau C, Brites D, Coscolla M, Gygli SM, Rutaihwa LK, Trauner A, Beisel C, Borrell S, Gagneux S (2018). *Treemmer: a tool to reduce large phylogenetic datasets with minimal loss of diversity*. *BMC Bioinformatics* **19**, 164. https://doi.org/10.1186/s12859-018-2164-8.

Mendler K, Chen H, Parks DH, Lobb B, Hug LA, Doxey AC (2019). *AnnoTree: visualization and exploration of a functionally annotated microbial tree of life*. *Nucleic Acids Research* **47**, 4442–4448. https://doi.org/10.1093/nar/gkz246.

Menet H, Daubin V, Tannier E (2022). *Phylogenetic reconciliation*. *PLoS Comput Biol* **18**, e1010621. https://doi.org/10.1371/journal.pcbi.1010621.

Menet H, Nguyen Trung A, Daubin V, Tannier E (2023). *Host symbiont gene reconciliation supplementary material [Data set]*. Zenodo. https://doi.org/10.5281/zenodo.7667342.

Morel B, Kozlov AM, Stamatakis A, Szöllősi GJ (2020). *GeneRax: A Tool for Species-Tree-Aware Maximum Likelihood-Based Gene Family Tree Inference under Gene Duplication, Transfer, and Loss*. *Molecular Biology and Evolution* **37**, 2763–2774. https://doi.org/10.1093/molbev/msaa141.

Morel B, Schade P, Lutteropp S, Williams TA, Szöllősi GJ, Stamatakis A (2022). *SpeciesRax: A Tool for Maximum Likelihood Species Tree Inference from Gene Family Trees under Duplication,*

Transfer, and Loss. *Molecular Biology and Evolution* **39**, msab365. https://doi.org/10.1093/molbev/msab365.

Muhammad SA, Sennblad B, Lagergren J (2018). *Species tree-aware simultaneous reconstruction of gene and domain evolution*. *bioRxiv*, 336453. https://doi.org/10.1101/336453.

Mykowiecka A, Muszewska A, Górecki P (2018). *Inferring time-consistent and well-supported horizontal gene transfers*. In: 2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). IEEE Computer Society, pp. 79–83. https://doi.org/10.1109/BIBM.2018.8621558.

Nakabachi A, Ueoka R, Oshima K, Teta R, Mangoni A, Gurgui M, Oldham NJ, Echten-Deckert G, Okamura K, Yamamoto K, Inoue H, Ohkuma M, Hongoh Y, Miyagishima Sy, Hattori M, Piel J, Fukatsu T (2013). *Defensive Bacteriome Symbiont with a Drastically Reduced Genome*. *Current Biology* **23**, 1478–1484. https://doi.org/10.1016/j.cub.2013.06.027.

Nakhleh L (2013). *Computational approaches to species phylogeny inference and gene tree reconciliation*. *Trends in ecology and evolution* **28**, 719–728.

Nikoh N, Hosokawa T, Moriyama M, Oshima K, Hattori M, Fukatsu T (2014). *Evolutionary origin of insect–Wolbachia nutritional mutualism*. *Proceedings of the National Academy of Sciences of the United States of America* **111**, 10257–10262. https://doi.org/10.1073/pnas.1409284111.

Penel S, Menet H, Tricou T, Daubin V, Tannier E (2022). *Thirdkind: displaying phylogenetic encounters beyond 2-level reconciliation*. *Bioinformatics (Oxford, England)*, btac062. https://doi.org/10.1093/bioinformatics/btac062.

Penz T, Schmitz-Esser S, Kelly SE, Cass BN, Müller A, Woyke T, Malfatti SA, Hunter MS, Horn M (2012). *Comparative Genomics Suggests an Independent Origin of Cytoplasmic Incompatibility in Cardinium hertigii*. *PLOS Genetics* **8**, e1003012. https://doi.org/10.1371/journal.pgen.1003012.

Rasmussen MD, Kellis M (2012). *Unified modeling of gene duplication, loss, and coalescence using a locus tree*. *Genome Research* **22**, 755–765. https://doi.org/10.1101/gr.123901.111.

Ravenhall M, Škunca N, Lassalle F, Dessimoz C (2015). *Inferring Horizontal Gene Transfer*. *PLOS Computational Biology* **11**, e1004095. https://doi.org/10.1371/journal.pcbi.1004095.

Ree RH, Moore BR, Webb CO, Donoghue MJ (2005). *A likelihood framework for inferring the evolution of geographic range on phylogenetic trees*. *Evolution* **59**, 2299–2311. https://doi.org/10.1111/j.0014-3820.2005.tb00940.x.

Ree RH, Smith SA (2008). *Maximum Likelihood Inference of Geographic Range Evolution by Dispersal, Local Extinction, and Cladogenesis*. *Systematic Biology* **57**, 4–14. https://doi.org/10.1080/10635150701883881.

Ronquist F (1997). *Dispersal-Vicariance Analysis: A New Approach to the Quantification of Historical Biogeography*. *Systematic Biology* **46**, 195–203. https://doi.org/10.1093/sysbio/46.1.195.

Santichaivekin S, Yang Q, Liu J, Mawhorter R, Jiang J, Wesley T, Wu YC, Libeskind-Hadas R (2020). *eMPRess: a systematic cophylogeny reconciliation tool*. *Bioinformatics* (btaa978). https://doi.org/10.1093/bioinformatics/btaa978.

Sapp J (1994). *Evolution by association*. Oxford University Press.

Sjöstrand J, Tofigh A, Daubin V, Arvestad L, Sennblad B, Lagergren J (2014). *A Bayesian Method for Analyzing Lateral Gene Transfer*. *Systematic Biology* **63**, 409–420. https://doi.org/10.1093/sysbio/syu007.

Sonnenburg JL, Sonnenburg ED (2019). *Vulnerability of the industrialized microbiota*. *Science* **366**, eaaw9255. https://doi.org/10.1126/science.aaw9255.

Stolzer M, Siewert K, Lai H, Xu M, Durand D (2015). *Event inference in multidomain families with phylogenetic reconciliation*. *BMC Bioinformatics* **16**, S8. https://doi.org/10.1186/1471-2105-16-S14-S8.

Szöllosi GJ, Tannier E, Lartillot N, Daubin V (2013). *Lateral gene transfer from the dead*. *Systematic Biology* **62**, 386–397. https://doi.org/10.1093/sysbio/syt003.

Szöllősi GJ, Boussau B, Abby SS, Tannier E, Daubin V (2012). *Phylogenetic modeling of lateral gene transfer reconstructs the pattern and relative timing of speciations*. *Proceedings of the National Academy of Sciences* **109**. Publisher: National Academy of Sciences Section: Biological Sciences, 17513–17518. https://doi.org/10.1073/pnas.1202997109.

Szöllősi GJ, Davín AA, Tannier E, Daubin V, Boussau B (2015a). *Genome-scale phylogenetic analysis finds extensive gene transfer among fungi*. *Philosophical Transactions of the Royal Society B: Biological Sciences* **370**. Publisher: Royal Society, 20140335. https://doi.org/10.1098/rstb.2014.0335.

Szöllősi GJ, Rosikiewicz W, Boussau B, Tannier E, Daubin V (2013). *Efficient Exploration of the Space of Reconciled Gene Trees*. *Systematic Biology* **62**, 901–912. https://doi.org/10.1093/sysbio/syt054.

Szöllősi GJ, Tannier E, Daubin V, Boussau B (2015b). *The Inference of Gene Trees with Species Trees*. *Systematic Biology* **64**, e42–e62. https://doi.org/10.1093/sysbio/syu048.

Tricou T, Tannier E, Vienne DM (2022). *Ghost lineages can invalidate or even reverse findings regarding gene flow*. *PLoS Biology* **20**, e3001776. https://doi.org/10.1371/journal.pbio.3001776.

Waskito LA, Yamaoka Y (2019). *The Story of Helicobacter pylori: Depicting Human Migrations from the Phylogeography*. In: *Helicobacter pylori in Human Diseases: Advances in Microbiology, Infectious Diseases and Public Health Volume 11*. Ed. by Shigeru Kamiya and Steffen Backert. Advances in Experimental Medicine and Biology. Cham: Springer International Publishing, pp. 1–16. https://doi.org/10.1007/5584_2019_356.

Wieseke N, Hartmann T, Bernt M, Middendorf M (2015). *Cophylogenetic Reconciliation with ILP*. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* **12**, 1227–1235. https://doi.org/10.1109/TCBB.2015.2430336.

Wijayawardena BK, Minchella DJ, DeWoody JA (2013). *Hosts, parasites, and horizontal gene transfer*. *Trends in Parasitology* **29**, 329–338. https://doi.org/10.1016/j.pt.2013.05.001.

Yang Z (2006). *Computational molecular evolution*. Oxford series in ecology and evolution. Oxford: Oxford University press.

Zhaxybayeva O, Gogarten JP (2004). *Cladogenesis, coalescence and the evolution of the three domains of life*. *Trends in genetics: TIG* **20**, 182–187. https://doi.org/10.1016/j.tig.2004.02.004.