Host-symbiont-gene phylogenetic reconciliation

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

[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*

[*] *To whom correspondence should be addressed: eric.tannier@inria.fr*

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 coevolution of two different levels, 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 coevolution 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 host-symbiont 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.6782794.

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 [18].

Yet the essence of living systems lies in dependence: constraint, cooperation or conflict [48]. Symbiotic micro-organisms coevolve with animals or plants [50]. 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 coevolving entities. Gene/species systems have been studied by phylogenetic reconciliation, accounting for events of gene duplication, horizontal gene transfer and gene loss (DTL model) [14, 38, 56, 9, 32]. The same model can be applied with little to no modification to symbiont/host [11, 47, 13], protein domain/gene coevolution [42, 51], or biogeography [28, 45, 46]. DTL models have also been used to reconstruct genome histories [16], detect highways of lateral gene transfers in bacteria, archaea or eukaryota [7], assess the relative role of duplication and gene transfer in the evolution of genomes [49], infer ancient symbiotic relationships [5], reconstruct histories of gene fusion and fission [15], model endosymbiotic gene transfer [4].

A limitation of reconciliation methods is their separate application on molecular studies on one side (gene/species coevolution), and ecological studies on the other (host/symbiont coevolution). 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 coevolution. In support of this claim, a number of empirical studies already rely on host symbiont histories when proposing horizontal gene transfers between symbionts [41, 39, 27, 37], when often, only symbiont gene/species comparisons do not provide enough statistical support for them [59, 43].

Three level reconciliations have been introduced by Stolzer et al. [51] 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 [25] 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 [25] and propose different heuristics and problem variants [26, 24]. A probabilistic model without transfers has been proposed by Muhammad, Sennblad, and Lagergren [35]. 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, but 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 coevolving 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 [55, 54]. 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 [23], 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 [27] 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 coevolution between all three scales.

2 2-level reconciliation, definitions and preliminaries

We denote by G, S, H respectively the gene tree, symbiont (or species) tree and host tree. Given a tree T, $|T|$ is the number of nodes of T.

We briefly describe in this section a two-level DTL reconciliation model, based on the undated version of the "ODT" model as implemented in "ALE undated" [54]. It is a birth and death like model generating a rooted phylogenetic tree G inside S, with speciation at all speciation nodes of S, and duplications, transfers and loss specific to G along the branches of S. We thus have three rates for duplication, transfer and loss events, concerning the evolution of genes inside their species. A gene tree can originate in any branch of the species tree with a uniform prior.

The input of 2-level reconciliation inference is one gene tree, and one species tree, with a many to one matching of the leaves. Both trees are assumed undated, binary

and rooted.

We call reconciliation scenario a list of events of kinds D,T,L, or S for each internal gene tree node, that can be the result of the birth and death process (the events happening on a branch are assigned to its child node). These lists transcribe into a mapping of the gene tree nodes to the species tree nodes it evolves in. We note $R_{G,S}$ the set of all possible reconciliation scenarios between G and S.

We denote by p^S, p^D, p^T, p^L the probabilities for a gene evolving on a species tree branch during the process to undergo the S,D,T,L events, with $p^S + p^D + p^T + p^L = 1$. These probabilities are constant along the species tree and for all gene families. When confusion is possible we add a S index for symbiont/gene reconciliation, and H index for host/symbiont one. When a transfer occurs, the receiver branch is chosen according to a uniform probability, avoiding ancestor branches of the donor. This avoids impossible transfers but is not sufficient to guarantee that the overall scenario is time feasible. The likelihood of a scenario r, $P(r|S)$ is the product of the probabilities of all events.

Summing over all possible scenarios for one gene tree and species tree we obtain the likelihood of the gene tree $P(G|S)$. We do not have to enumerate all scenarios to compute that sum, because we can compute this likelihood using dynamic programming, considering matching all couples of gene and species sub-trees, starting from the leaves, and enumerating all possible events to get each match. This in return enables us to sample scenarios according to their likelihood, or finding the most likely scenario, by backtracking through the table constructed.

We will call such a reconciliation of a gene tree and symbiont tree, "2-level" reconciliation, in comparison with the host/symbiont/gene 3-level reconciliation that we introduce in the following section.

3 3-level reconciliation, likelihood estimation and scenario inference

3.1 Elements of the probabilistic model

We use an undated framework similar to the one implemented in ALE undated [54] presented in previous section. A rooted binary host phylogenetic tree H is given, without branch lengths. We do not include the generation parameters in our model and consider instead the host tree as a parameter. We model the evolution of one or multiple symbiont trees S with the DTL model [54] (see previous section) adding the possibility for a symbiont to live temporarily in an unknown host (this feature is described in more detail in Section 3.8).

We then model the evolution of genes in the symbiont trees with duplication, loss and intra horizontal transfer, meaning that horizontal transfer is possible only between symbiont branches that are present in the same host branch. We thus have six rates in our model, three for the duplication, transfer and loss between host and symbiont, and three additional ones for duplication, intra-transfer and loss events concerning genes inside their species (note that both p_H^S and p_S^S the speciation probabilities for the host/symbiont and symbiont/gene reconciliations are not free parameters as in both

cases $p^S = 1 - p^D - p^T - p^L$). An illustration of the realization of such a model, as

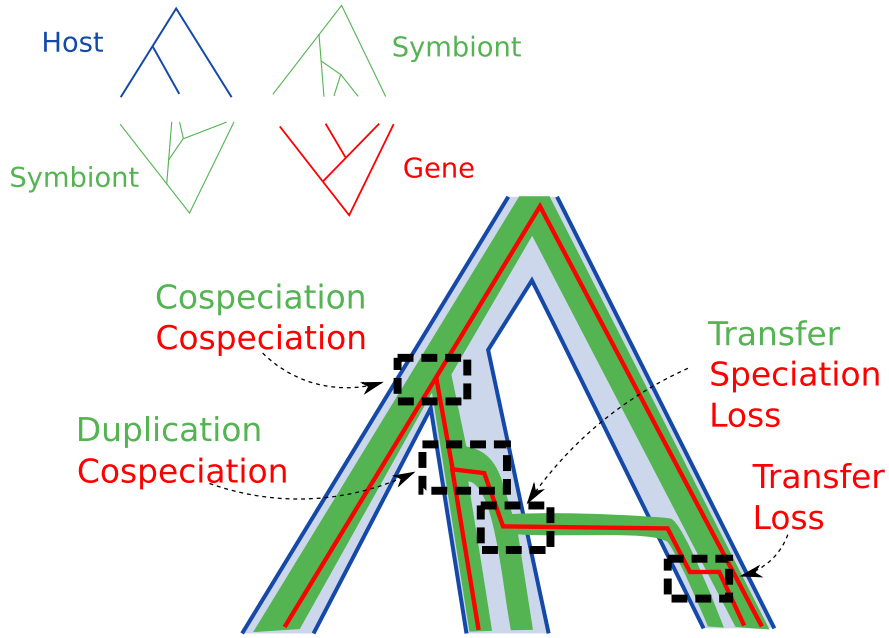


Figure 1: An example of a 3-Level reconciliation input (top left, with three trees and associations between the leaves of two couples of trees) and a possible reconciliation scenario for this input. Events of the host/symbiont co-evolution are written in red, while events of the symbiont/gene reconciliation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

well as the input for the inference part, is given in Figure 1.

This model can be immediately used for simulations, but we chose to use an external simulator for our tests [23], to minimize the similarities in models and implementation between simulation and inference.

3.2 Monte Carlo approximation of the likelihood

The inference consists in computing the probability that G have been generated by the model: $P(G|S, H)$, from a set of input trees (one rooted host tree H, one or several rooted symbiont trees S, one or several unrooted gene trees G), and given the DTL probabilities for the two reconciliation levels. It is then possible to estimate the evolutionary rates and sample among reconciliation scenarios. We also propose a way to propose one likely symbiont tree if only the host tree and some gene trees are given. All given trees are supposed to be binary, and branch lengths are not taken into account.

Because a similar computation in a parsimonious framework is NP-hard [25], it is probably not possible to exactly and quickly compute $P(G|S, H)$. We thus apply an approximation technique based on sampling reconciliations. The probability of a gene tree can indeed be decomposed by summing over all possible host/symbiont

reconciliation scenarios $r_{S,H}$:

$$P(G|S,H) = \sum_{r_{S,H} \in R_{S,H}} P(G|S,H,r_{S,H}) P(r_{S,H}|S,H) \tag{1}$$

The number of reconciliations in this sum is at least exponential in the size of the input (and even the number of scenarios maximizing $P(r_{S,H}|S,H)$ can be exponential [13]), so we use a Monte Carlo approach to estimate it, sampling a reasonable number N of symbiont/host reconciliations:

$$P(G|S,H) \simeq \frac{1}{N} \sum_{n=1}^{N} P(G|S,H,r_n) \tag{2}$$

where r_n is sampled in the set $R_{S,H}$ of all reconciliations according to its likelihood $P(r_n|S,H)$.

3.3 Reconciliation inference and ghost lineages

Sampling reconciliations in $R_{S,H}$ can be done with the dynamic programming algorithm implemented in "ALE undated" and is a two-level reconciliation problem [56].

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. It consists in checking, during the dynamic programming process, for all gene transfer possibilities, if the donor symbiont i and receiver one j 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.

In our model gene transfer can only occur between two symbiont species inside a same host. However 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 unsampled and gene transfers are often "from the dead" [52, 19, 61]. In consequence in the model a transfer can occur from a donor that is now extinct. This transfer is traced back to an ancestor of this extinct donor that is not in the same host than the receiver. See in Figure 2 how an "inter" transfer between i and j (on the left) can be modelled (on the right) by an extinct sister lineage to symbiont i, that switched host, and transferred a gene to j while being in the same host. As the sister lineage goes extinct, in the inferred gene history 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:

$$P_S^T(i \to j) = \sum_{e \in H_i, h \in H_j} P_H^T(e \to h) \tag{3}$$

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

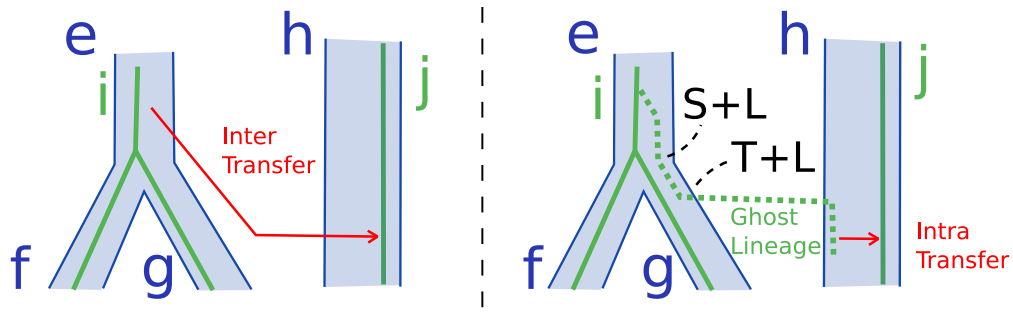


Figure 2: Computation of inter transfer rate from intra transfer rate and ghost species: the left inter transfer can be modeled by multiple scenarios without inter gene transfer but implying ghost symbiont lineages, such as the one on the right that implies first a speciation and a loss, and then a transfer and a loss before the extinction of the species.

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 $P_e = P(e \to h)$ 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 [33], or figure 2: we denote by f, g the children of a host e.

$$\begin{cases} P_e &= \frac{1}{|S_h|} p_S^T \text{ if } e = h \\ P_e &= p_H^S (P_f E_g + P_g E_f) + 2 p_H^D P_e E_e + \sum_{k \in H} \frac{p_H^T}{|H|} P_k E_e \end{cases} \tag{4}$$

This equation has a self dependency due to the Transfer/Loss event, which is already accounted for in reconciliation methods [20, 55]. We forbid successions of several Transfer/Loss events to break this self dependency and solve this equation.

3.4 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 the one that maximises the marginal likelihood [60]. That is, at each step of the backtracking of the dynamic programming procedure we select the maximum likelihood position.

This approach is similar to the one of Stolzer et al. [51], but in a probabilistic setting, using marginal likelihood, and with a way of computing the inter transfer probabilities

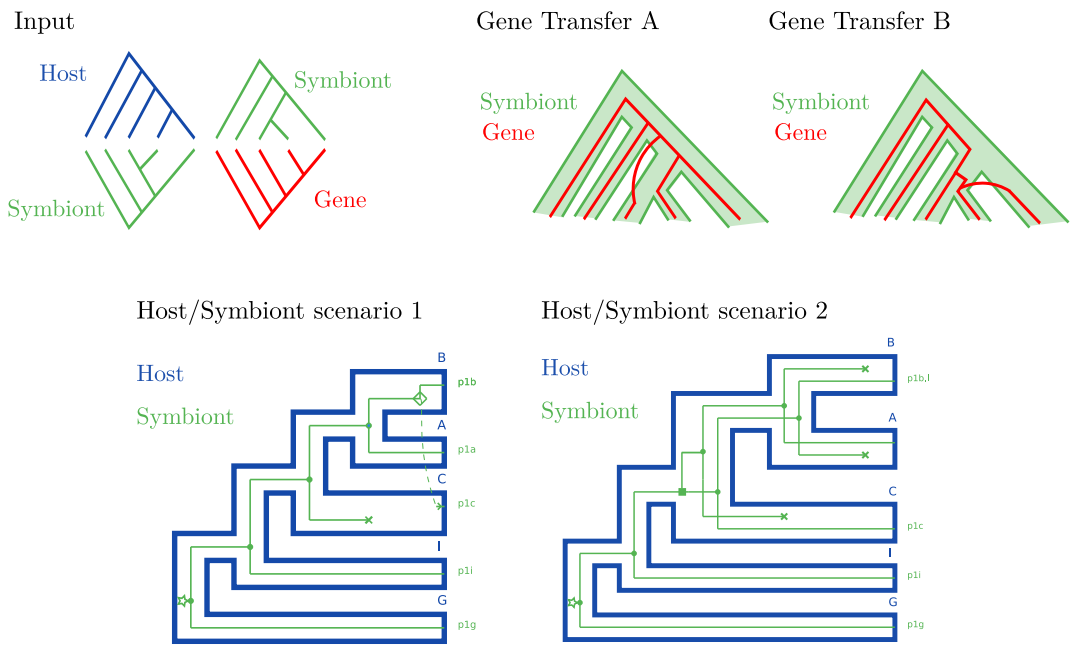


Figure 3: An example of input where the Sequential heuristic is less robust than the Monte Carlo one. We compare the support for two gene transfer scenarios, scenario A and B. There are two main possible host/symbiont reconciliation scenarios, scenario 1 and scenario 2. In scenario 1, gene transfer A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

from the host/symbiont and symbiont/gene DTL reconciliation parameters instead of using an additional parameter.

We study the differences between the Sequential heuristic, the Monte Carlo sampling, and a classic 2-level approach, in order to measure the importance of taking into account the dependency between host symbiont and symbiont gene reconciliations. The description of the 3-level model is also important from a theoretical point of view to get a better understanding of what we compute in the different heuristic inference methods. Writing the equations also enables us to see the different parameters that can be inferred, using a maximum likelihood framework for instance. An example is the possibility to infer a host/symbiont reconciliation aware of the genes.

The faster Sequential heuristic may not be as robust as the Monte Carlo one. Li and Bansal in [26] 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.

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

Table 1: Comparison of the support for the two gene transfer scenarios in the example presented in figure 3.

3.5 Time complexity and tractability

DTL reconciliation methods use a dynamic programming approach to compute the probability of the coevolution of two trees (and all of their subtrees) [10]. Then backtracking produces reconciliation scenarios. In a gene/species reconciliation, if all transfers have the same probability, *i.e.* this probability is independent from the donor-receiver couple, DTL reconciliation can be computed in quadratic time [6].

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

The first part of our algorithm, is to reconcile host and symbiont, a classic 2-level reconciliation, with quadratic complexity $O(hs)$. In our implementation sampling a host symbiont reconciliation scenario is then done in cubic complexity (as we extend the transfer sum) however it might be possible to also get a quadratic complexity here, though as we do not consider all couples, it is in general faster than the forward computation.

Then we compute the gene transfer probabilities between all couple of symbiont nodes, this is done with a dynamic programming similar to the one for reconciliation in $O(hs)$, and a final sum (equation 3) over all hosts of the considered symbionts in $O(h^2s^2)$, which in the reasonable case where the number of symbiont nodes per host nodes (in the reconciliation scenario) is below a constant k, we get $O(h^2k^2 + hs)$ for this part.

Finally we can compute the host aware gene/symbiont reconciliation. The difference with classic 2-level reconciliation is that transfer rates depend on the donor-receiver couple. In consequence we cannot use the efficient computation trick used for uniform rates. 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.

For the Monte Carlo approach, we repeat all steps except the initial host/symbiont reconciliation, leading to a total complexity of $O(N(hs + h^2k^2 + s^2g))$ where h, s, g are the size of the host, symbiont and gene trees, k is a bound on the number of symbiont 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. 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.

3.6 Symbiont tree inference

In case the symbiont tree is unknown, we devised an option to infer the symbiont tree by amalgamation [12, 55] 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[20, 55].

3.7 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 [56], 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.

3.8 Free living relatives of symbionts

In the course of their evolutionary history, some symbiont may live outside a host, or within an unknown host.

This is particularly important for us because we invoke unknown hosts in the case of inter host horizontal gene transfers (section 3.3). In order to consider these cases we added the possibility for a symbiont to be "free living", meaning associated to no host.

We did that by adding the symbiont tree as a possible host tree, and matching the symbiont leaves with no host to themselves. In that way, we see transfer between free living as less likely than when a common host is known. The utility of this model

addition is visible in the *Cinara* aphids example developed in the Results section (see Fig. 6).

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 [17]. The scenarios can be visualised using Thirdkind[1] [40], 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 [23]. 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 [22], 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

[1] https://crates.io/crates/thirdkind

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 each symbiont/gene coevolution. 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 [36], but with the inference of ancestral characters [58], the number of transfers [53], the ability to infer better trees [8], or the ability to map the correct event type to each gene node [22]. 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.

$$\text{Precision} = \frac{\sum_{T \in L_{t,sim}} P_{obs}(T)}{\sum_{T \in L_{t,obs}} P_{obs}(T)} \text{ and Recall } = \frac{\sum_{T \in L_{t,sim}} P_{obs}(T)}{\sum_{T \in L_{t,sim}} 1} \tag{5}$$

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.

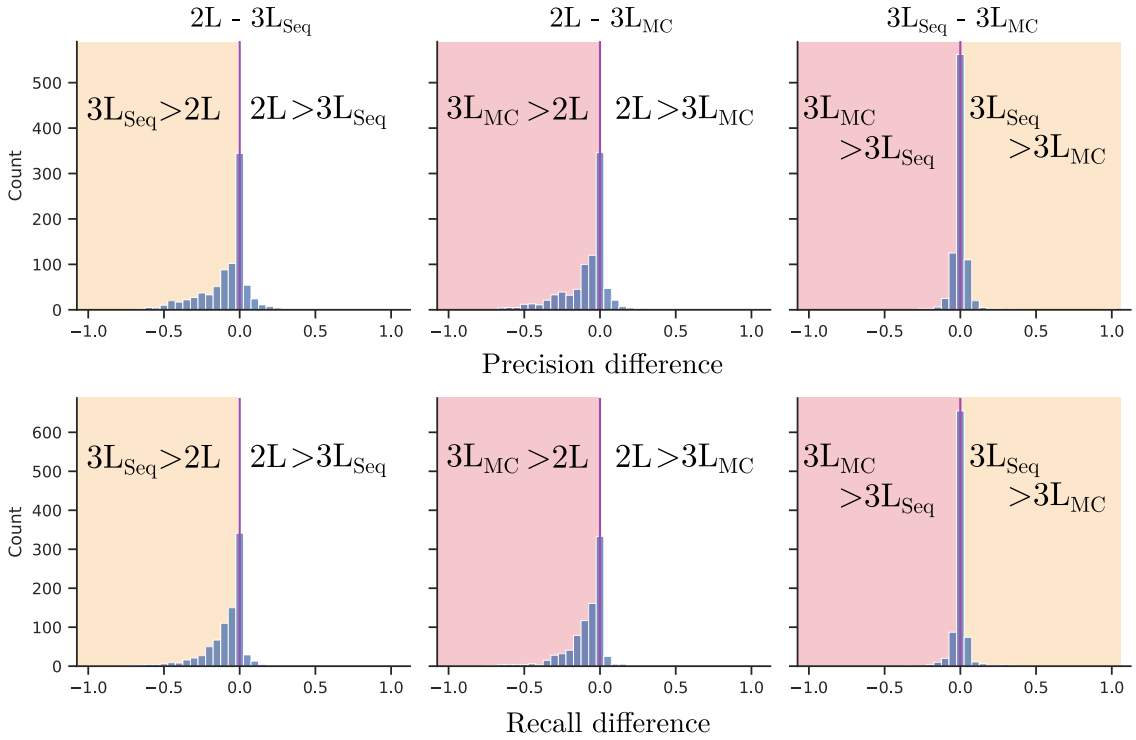


Figure 4: Distribution of differences of precision and recall on the inference of horizontal gene transfers for all combinations of two approaches: 2-level (2L), 3-level with the Monte Carlo heuristic (3L$_{MC}$) and 3-level with the Sequential heuristic (3L$_{Seq}$), centered on 0, and for all 874 gene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

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 1.0 model with no 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.

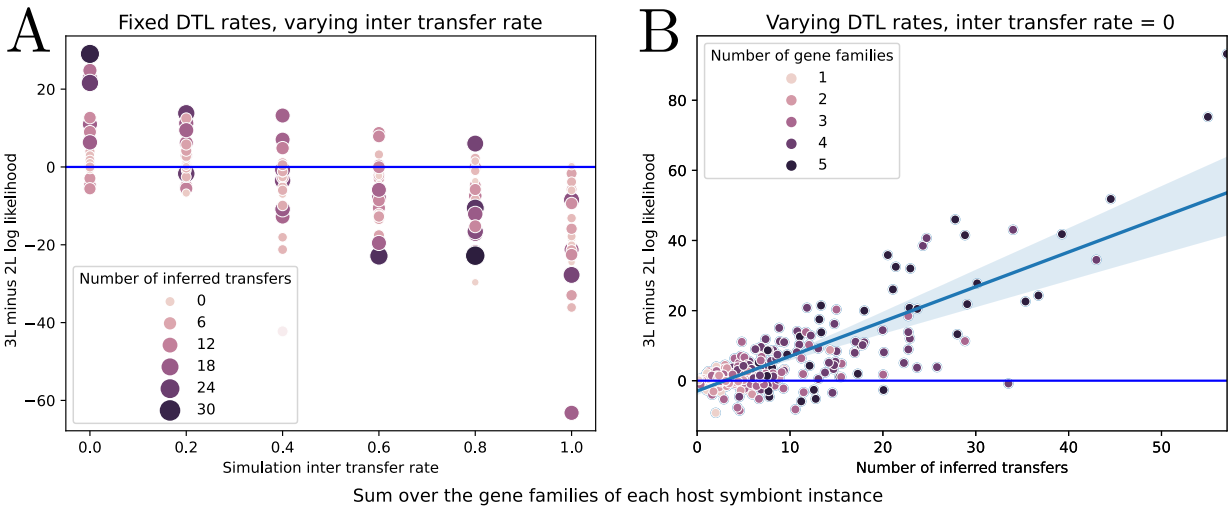


Figure 5: A test of host symbiont co-evolution. We measure the difference of likelihood between the 3-level model and the 2-level model, using the estimation of these likelihoods provided by our "2-level" and "3-level Sequential" heuristics, in order to differentiate instances where gene trees are generated 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 [27] 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. [27]. *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 [31], to complement the *Cinara* aphids symbionts phylogeny proposed in [27]. 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.

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. [27]. 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 [1]. Bacterial strains have been divided in different populations corresponding to geographical areas (Africa 1, Africa 2, Asia 2, East Asia, North East Africa, Europe) [57, 29].

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 [2, 21]. A phylogenetic tree was built based on the concatenation of universal-unicopy genes (322 gene families), and

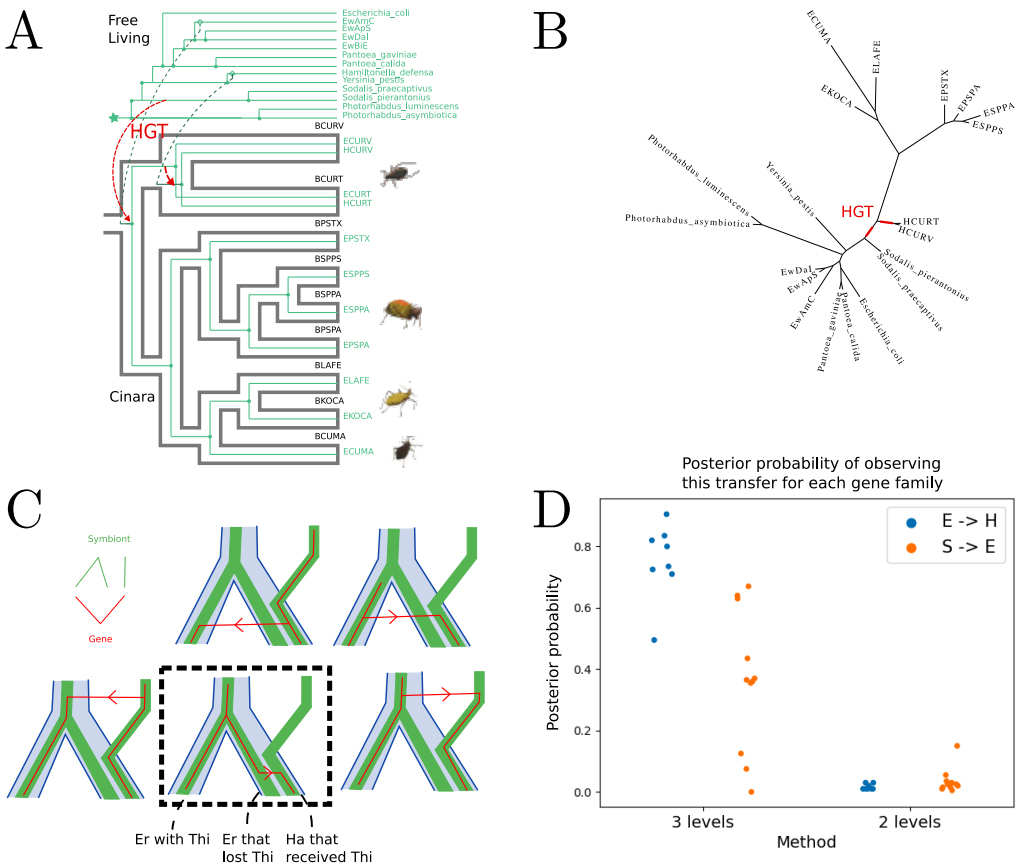


Figure 6: The evolution of *Cinara* and their enterobacteria symbionts. (A) The coevolutionary scenario identified by Manzano-Marín et al. [27]. The reconciliation of the hosts (*Cinara* aphids) and symbionts (bacteria) are depicted along with the position of the horizontal gene transfers (in red). (B) 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.

a sample of 113 strains representing the diversity of H. pylori in the old world (excluding strains from the Americas) was obtained using Treemmer [30]. 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 [57, 29]. 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 3.6). 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 reconciliation frameworks [44].

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.

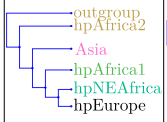
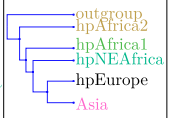
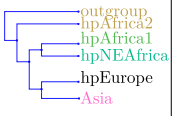
Figure 7 (B) is an illustration of a reconciliation scenario for the maximum likelihood host tree with Thirdkind [40]. We see the host tree and the amalgamated strain tree reconciled (I events are represented as transfers from a parent node to one of its child). On top of these two embedded trees red 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 [59] 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.

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	-653×10^3	-658×10^3	-658×10^3	-676×10^3

B

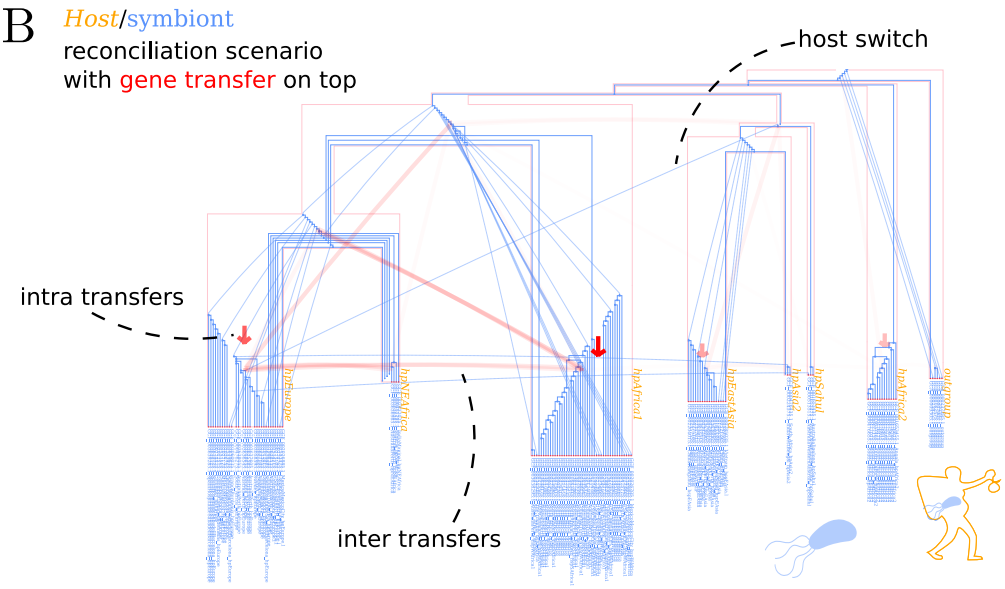


Figure 7: Co-evolution of human populations and *Helicobacter pylori*. (A) Log likelihood of the different population trees. (B) The representation with Third-Kind [40] of one possible reconciliation scenario of *Helicobacter pylori* strain tree and the population tree maximizing likelihood. Aggregated gene transfers are depicted on top of the DTL reconciliation, with the opacity corresponding to the number of time the transfers were seen across the 1034 gene families.

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 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 [34].

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

A comparison of the inference method to similar ones [51, 25, 35] 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. [51] 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 coevolving entities can be studied with it: species/gene/protein domain as it was done in previous studies [51, 24, 35], 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 multilevel co-evolution, could arise with the more and more numerous studied biological systems that fit into this multi-scale coevolution framework, notably with an increasing interest for hologenomics [3].

6 Acknowledgements

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

7 Funding

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

References

[1] Mark Achtman. "How old are bacterial pathogens?" In: *Proceedings of the Royal Society B: Biological Sciences* 283.1836 (Aug. 17, 2016), p. 20160990. DOI: 10.1098/rspb.2016.0990. URL: https://royalsocietypublishing.org/doi/full/10.1098/rspb.2016.0990 (visited on 09/12/2019).

[2] Mark Achtman, Takeshi Azuma, Douglas E. Berg, Yoshiyuki Ito, Giovanna Morelli, Zhi-Jun Pan, Sebastian Suerbaum, Stuart A. Thompson, Arie Van Der Ende, and Leen-Jan Van Doorn. "Recombination and clonal groupings within Helicobacter pylori from different geographical regions". In: *Molecular Microbiology* 32.3 (1999), pp. 459–470. DOI: https://doi.org/10.1046/j.1365-2958.1999.01382.x. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1046/j.1365-2958.1999.01382.x. URL: https://onlinelibrary.wiley.com/doi/abs/10.1046/j.1365-2958.1999.01382.x.

[3] Antton Alberdi, Sandra B. Andersen, Morten T. Limborg, Robert R. Dunn, and M. Thomas P. Gilbert. "Disentangling host–microbiota complexity through hologenomics". In: *Nature Reviews Genetics* 23.5 (May 2022). Number: 5 Publisher: Nature Publishing Group, pp. 281–297. ISSN: 1471-0064. DOI: 10.1038/s41576-021-00421-0. URL: https://www.nature.com/articles/s41576-021-00421-0 (visited on 06/01/2022).

[4] Yoann Anselmetti, Nadia El-Mabrouk, Manuel Lafond, and Aïda Ouangraoua. "Gene tree and species tree reconciliation with endosymbiotic gene transfer". In: *Bioinformatics* 37 (Supplement_1 July 1, 2021), pp. i120–i132. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btab328. URL: https://doi.org/10.1093/bioinformatics/btab328 (visited on 10/20/2021).

[5] Marc Bailly-Bechet, Patricia Martins-Simões, Gergely J. Szöllősi, Gladys Mialdea, Marie-France Sagot, and Sylvain Charlat. "How Long Does Wolbachia Remain on Board?" In: *Molecular Biology and Evolution* 34.5 (May 1, 2017), pp. 1183–1193. ISSN: 0737-4038. DOI: 10.1093/molbev/msx073. URL: https://academic.oup.com/mbe/article/34/5/1183/2992913 (visited on 12/02/2019).

[6] Mukul S. Bansal, Eric J. Alm, and Manolis Kellis. "Efficient algorithms for the reconciliation problem with gene duplication, horizontal transfer and loss". In: *Bioinformatics* 28.12 (June 15, 2012). Publisher: Oxford Academic, pp. i283–i291. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/bts225. URL: https://academic.oup.com/bioinformatics/article/28/12/i283/269262 (visited on 06/01/2020).

[7] Mukul S. Bansal, Guy Banay, J. Peter Gogarten, and Ron Shamir. "Detecting Highways of Horizontal Gene Transfer". In: *Journal of Computational Biology* 18.9 (Sept. 1, 2011), pp. 1087–1114. DOI: 10.1089/cmb.2011.0066. URL: https://www.liebertpub.com/doi/abs/10.1089/cmb.2011.0066 (visited on 02/20/2020).

[8] Mukul S. Bansal, Yi-Chieh Wu, Eric J. Alm, and Manolis Kellis. "Improved gene tree error correction in the presence of horizontal gene transfer". In: *Bioinformatics* 31.8 (Apr. 15, 2015), pp. 1211–1218. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btu806. URL: https://academic.oup.com/bioinformatics/article/31/8/1211/212713 (visited on 11/18/2019).

[9] Bastien Boussau and Celine Scornavacca. "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, 2020, 3.2:1–3.2:23. URL: https://hal.archives-ouvertes.fr/hal-02535529 (visited on 08/20/2020).

[10] Michael Charleston. "Jungles: a new solution to the host/parasite phylogeny reconciliation problem". In: *Mathematical Biosciences* 149.2 (May 1998), pp. 191–223. ISSN: 0025-5564. DOI: 10.1016/s0025-5564(97)10012-8.

[11] Michael Charleston and Ran Libeskind-Hadas. "Event-Based Cophylogenetic Comparative Analysis". In: *Modern Phylogenetic Comparative Methods and Their Application in Evolutionary Biology*. Springer Berlin Heidelberg, 2014, pp. 465–480. DOI: 10.1007/978-3-662-43550-2_20. URL: https://doi.org/10.1007/978-3-662-43550-2_20.

[12] Lawrence A. David and Eric J. Alm. "Rapid evolutionary innovation during an Archaean genetic expansion". In: *Nature* 469.7328 (Jan. 2011), pp. 93–96. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/nature09649. URL: http://www.nature.com/articles/nature09649 (visited on 02/19/2020).

[13] Beatrice Donati, Christian Baudet, Blerina Sinaimeri, Pierluigi Crescenzi, and Marie-France Sagot. "EUCALYPT: efficient tree reconciliation enumerator". In: *Algorithms for Molecular Biology* 10.1 (Jan. 23, 2015), p. 3. ISSN: 1748-7188. DOI: 10.1186/s13015-014-0031-3. URL: https://doi.org/10.1186/s13015-014-0031-3 (visited on 02/18/2020).

[14] Jean-Philippe Doyon, Vincent Ranwez, Vincent Daubin, and Vincent Berry. "Models, algorithms and programs for phylogeny reconciliation". In: *Briefings in Bioinformatics* 12.5 (Sept. 2011), pp. 392–400. ISSN: 1477-4054. DOI: 10.1093/bib/bbr045.

[15] Wandrille Duchemin, Yoann Anselmetti, Murray Patterson, Yann Ponty, Sèverine Bérard, Cedric Chauve, Celine Scornavacca, Vincent Daubin, and Eric Tannier. "DeCoSTAR: Reconstructing the Ancestral Organization of Genes or Genomes Using Reconciled Phylogenies". In: *Genome Biology and Evolution* 9.5 (May 1, 2017), pp. 1312–1319. DOI: 10.1093/gbe/evx069. URL: https://academic.oup.com/gbe/article/9/5/1312/3200386 (visited on 09/13/2019).

[16] Wandrille Duchemin, Vincent Daubin, and Eric Tannier. "Reconstruction of an ancestral Yersinia pestisgenome and comparison with an ancient sequence". In: *BMC Genomics* 16.10 (Oct. 2, 2015), S9. ISSN: 1471-2164. DOI: 10.1186/1471-2164-16-S10-S9. URL: https://doi.org/10.1186/1471-2164-16-S10-S9 (visited on 01/05/2022).

[17] Wandrille Duchemin, Guillaume Gence, Anne-Muriel Arigon Chifolleau, Lars Arvestad, Mukul S. Bansal, Vincent Berry, Bastien Boussau, François Chevenet, Nicolas Comte, Adrián A. Davín, Christophe Dessimoz, David Dylus, Damir Hasic, Diego Mallo, Rémi Planel, David Posada, Celine Scornavacca, Gergely Szöllősi, Louxin Zhang, Éric Tannier, and Vincent Daubin. "RecPhyloXML: a format for reconciled gene trees". In: *Bioinformatics* 34.21 (Nov. 1, 2018), pp. 3646–3652. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/bty389. URL: https://academic.oup.com/bioinformatics/article/34/21/3646/4995844 (visited on 09/12/2019).

[18] Joseph Felsenstein. *Inferring Phylogenies*. Oxford, New York: Oxford University Press, Sept. 4, 2003. 580 pp. ISBN: 978-0-87893-177-4.

[19] Gregory P. Fournier, Jinling Huang, and J. Peter Gogarten. "Horizontal gene transfer from extinct and extant lineages: biological innovation and the coral of life". In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 364.1527 (Aug. 12, 2009), pp. 2229–2239. ISSN: 0962-8436. DOI: 10.1098/rstb.2009.0033. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2873001/ (visited on 09/21/2021).

[20] Edwin Jacox, Cedric Chauve, Gergely J. Szöllősi, Yann Ponty, and Celine Scornavacca. "ecceTERA: comprehensive gene tree-species tree reconciliation using parsimony". In: *Bioinformatics (Oxford, England)* 32.13 (2016), pp. 2056–2058. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btw105.

[21] Keith A. Jolley, James E. Bray, and Martin C. J. Maiden. "Open-access bacterial population genomics: BIGSdb software, the PubMLST.org website and their applications". In: *Wellcome Open Research* 3 (2018), p. 124. ISSN: 2398-502X. DOI: 10.12688/wellcomeopenres.14826.1.

[22] Misagh Kordi, Soumya Kundu, and Mukul S. Bansal. "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, Sept. 4, 2019, pp. 514–523. ISBN: 978-1-4503-6666-3. DOI: 10.1145/3307339.3342168. URL: https://doi.org/10.1145/3307339.3342168 (visited on 02/18/2020).

[23] Soumya Kundu and Mukul S. Bansal. "SaGePhy: an improved phylogenetic simulation framework for gene and subgene evolution". In: *Bioinformatics* 35.18 (Sept. 15, 2019), pp. 3496–3498. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btz081. URL: https://academic.oup.com/bioinformatics/article/35/18/3496/5305633 (visited on 09/25/2019).

[24] Lei Li and Mukul S. Bansal. "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, 2018, pp. 386–397. ISBN: 978-1-4503-5794-4. DOI: 10.1145/3233547.3233603. URL: http://doi.acm.org/10.1145/3233547.3233603 (visited on 09/25/2019).

[25] Lei Li and Mukul S. Bansal. "An Integrated Reconciliation Framework for Domain, Gene, and Species Level Evolution". In: *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 16.1 (Jan. 2019), pp. 63–76. DOI: 10.1109/TCBB.2018.2846253.

[26] Lei Li and Mukul S. Bansal. "Simultaneous Multi-Domain-Multi-Gene Reconciliation Under the Domain-Gene-Species Reconciliation Model". In: *Bioinformatics Research and Applications*. Ed. by Zhipeng Cai, Pavel Skums, and Min Li. Lecture Notes in Computer Science. Springer International Publishing, 2019, pp. 73–86. ISBN: 978-3-030-20242-2.

[27] Alejandro Manzano-Marín, Armelle Coeur D'acier, Anne-Laure Clamens, Céline Orvain, Corinne Cruaud, Valérie Barbe, and Emmanuelle Jousselin. "Serial horizontal transfer of vitamin-biosynthetic genes enables the establishment of new nutritional symbionts in aphids' di-symbiotic systems". In: *The ISME Journal* (Oct. 17, 2019), pp. 1–15. ISSN: 1751-7370. DOI: 10.1038/s41396-019-0533-6. URL: https://www.nature.com/articles/s41396-019-0533-6 (visited on 11/05/2019).

[28] Andrés Martínez-Aquino. "Phylogenetic framework for coevolutionary studies: a compass for exploring jungles of tangled trees". In: *Current Zoology* 62.4 (Aug. 2016), pp. 393–403. ISSN: 1674-5507. DOI: 10.1093/cz/zow018. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5804275/ (visited on 09/13/2019).

[29] Francis Mégraud, Philippe Lehours, and Filippa F. Vale. "The history of Helicobacter pylori: from phylogeography to paleomicrobiology". In: *Clinical Microbiology and Infection* 22.11 (Nov. 1, 2016), pp. 922–927. ISSN: 1198-743X. DOI: 10.1016/j.cmi.2016.07.013. URL: https://www.sciencedirect.com/science/article/pii/S1198743X16302373 (visited on 09/21/2021).

[30] Fabrizio Menardo, Chloé Loiseau, Daniela Brites, Mireia Coscolla, Sebastian M. Gygli, Liliana K. Rutaihwa, Andrej Trauner, Christian Beisel, Sonia Borrell, and Sebastien Gagneux. "Treemmer: a tool to reduce large phylogenetic datasets with minimal loss of diversity". In: *BMC Bioinformatics* 19.1 (May 2, 2018), p. 164. ISSN: 1471-2105. DOI: 10.1186/s12859-018-2164-8. URL: https://doi.org/10.1186/s12859-018-2164-8 (visited on 01/10/2022).

[31] Kerrin Mendler, Han Chen, Donovan H Parks, Briallen Lobb, Laura A Hug, and Andrew C Doxey. "AnnoTree: visualization and exploration of a functionally annotated microbial tree of life". In: *Nucleic Acids Research* 47.9 (May 21, 2019), pp. 4442–4448. ISSN: 0305-1048. DOI: 10.1093/nar/gkz246. URL: https://doi.org/10.1093/nar/gkz246 (visited on 09/20/2021).

[32] Hugo Menet, Vincent Daubin, and Eric Tannier. "Phylogenetic reconciliation". June 2021. URL: https://hal.archives-ouvertes.fr/hal-03258402 (visited on 10/20/2021).

[33] Benoit Morel, Alexey M. Kozlov, Alexandros Stamatakis, and Gergely J. Szöllősi. *GeneRax: A tool for species tree-aware maximum likelihood based gene family tree inference under gene duplication, transfer, and loss.* preprint. Bioinformatics, Sept. 26, 2019. DOI: 10.1101/779066. URL: http://biorxiv.org/lookup/doi/10.1101/779066 (visited on 03/10/2020).

[34] Benoit Morel, Paul Schade, Sarah Lutteropp, Tom A. Williams, Gergely J. Szöllősi, and Alexandros Stamatakis. *SpeciesRax: A tool for maximum likelihood species tree inference from gene family trees under duplication, transfer, and loss.* Company: Cold Spring Harbor Laboratory Distributor: Cold Spring Harbor Laboratory Label: Cold Spring Harbor Laboratory Section: New Results Type: article. Mar. 29, 2021, p. 2021.03.29.437460. DOI: 10.1101/2021.03.29.437460. URL: https://www.biorxiv.org/content/10.1101/2021.03.29.437460v1 (visited on 01/06/2022).

[35] Sayyed Auwn Muhammad, Bengt Sennblad, and Jens Lagergren. "Species tree-aware simultaneous reconstruction of gene and domain evolution". In: *bioRxiv* (June 2, 2018), p. 336453. DOI: 10.1101/336453. URL: https://www.biorxiv.org/content/10.1101/336453v1 (visited on 09/25/2019).

[36] Agnieszka Mykowiecka, Anna Muszewska, and Paweł Górecki. "Inferring time-consistent and well-supported horizontal gene transfers". In: 2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). IEEE Computer Society, Dec. 1, 2018, pp. 79–83. ISBN: 978-1-5386-5488-0. DOI: 10.1109/BIBM.2018.8621558. URL: https://www.computer.org/csdl/proceedings-article/bibm/2018/08621558/17D45Xi9rWz (visited on 10/20/2021).

[37] Atsushi Nakabachi, Reiko Ueoka, Kenshiro Oshima, Roberta Teta, Alfonso Mangoni, Mihaela Gurgui, Neil J. Oldham, Gerhild van Echten-Deckert, Keiko Okamura, Kohei Yamamoto, Hiromitsu Inoue, Moriya Ohkuma, Yuichi Hongoh, Shin-ya Miyagishima, Masahira Hattori, Jörn Piel, and Takema Fukatsu. "Defensive Bacteriome Symbiont with a Drastically Reduced Genome". In: *Current Biology* 23.15 (Aug. 5, 2013), pp. 1478–1484. ISSN: 0960-9822. DOI: 10.1016/j.cub.2013.06.027. URL: https://www.cell.com/current-biology/abstract/S0960-9822(13)00752-5 (visited on 11/06/2019).

[38] Luay Nakhleh. "Computational approaches to species phylogeny inference and gene tree reconciliation". In: *Trends in ecology and evolution* 28 (2013), pp. 719–728.

[39] Naruo Nikoh, Takahiro Hosokawa, Minoru Moriyama, Kenshiro Oshima, Masahira Hattori, and Takema Fukatsu. "Evolutionary origin of insect–Wolbachia nutritional mutualism". In: *Proceedings of the National Academy of Sciences of the United States of America* 111.28 (July 15, 2014), pp. 10257–10262. ISSN: 0027-8424. DOI: 10.1073/pnas.1409284111. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4104916/ (visited on 11/06/2019).

[40] Simon Penel, Hugo Menet, Théo Tricou, Vincent Daubin, and Eric Tannier. "Thirdkind: displaying phylogenetic encounters beyond 2-level reconciliation". In: *Bioinformatics (Oxford, England)* (Feb. 9, 2022), btac062. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btac062.

[41] Thomas Penz, Stephan Schmitz-Esser, Suzanne E. Kelly, Bodil N. Cass, Anneliese Müller, Tanja Woyke, Stephanie A. Malfatti, Martha S. Hunter, and Matthias Horn. "Comparative Genomics Suggests an Independent Origin of Cytoplasmic Incompatibility in Cardinium hertigii". In: *PLOS Genetics* 8.10 (Oct. 25, 2012), e1003012. ISSN: 1553-7404. DOI: 10.1371/journal.pgen.1003012. URL: https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1003012 (visited on 11/06/2019).

[42] Matthew D. Rasmussen and Manolis Kellis. "Unified modeling of gene duplication, loss, and coalescence using a locus tree". In: *Genome Research* 22.4 (Apr. 1, 2012), pp. 755–765. ISSN: 1088-9051. DOI: 10.1101/gr.123901.111. URL: http://genome.cshlp.org/cgi/doi/10.1101/gr.123901.111 (visited on 10/23/2019).

[43] Matt Ravenhall, Nives Škunca, Florent Lassalle, and Christophe Dessimoz. "Inferring Horizontal Gene Transfer". In: *PLOS Computational Biology* 11.5 (May 28, 2015), e1004095. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1004095. URL: https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1004095 (visited on 02/19/2020).

[44] Richard H Ree, Brian R Moore, Campbell O Webb, and Michael J Donoghue. "A likelihood framework for inferring the evolution of geographic range on phylogenetic trees". In: (2005), p. 13.

[45] Richard H. Ree and Stephen A. Smith. "Maximum Likelihood Inference of Geographic Range Evolution by Dispersal, Local Extinction, and Cladogenesis". In: *Systematic Biology* 57.1 (Feb. 1, 2008), pp. 4–14. ISSN: 1063-5157. DOI: 10.1080/10635150701883881. URL: https://academic.oup.com/sysbio/article/57/1/4/1703014 (visited on 10/02/2019).

[46] Fredrik Ronquist. "Dispersal-Vicariance Analysis: A New Approach to the Quantification of Historical Biogeography". In: *Systematic Biology* 46.1 (Mar. 1, 1997), pp. 195–203. ISSN: 1063-5157. DOI: 10.1093/sysbio/46.1.195. URL: https://academic.oup.com/sysbio/article/46/1/195/1685511 (visited on 10/03/2019).

[47] Santi Santichaivekin, Qing Yang, Jingyi Liu, Ross Mawhorter, Justin Jiang, Trenton Wesley, Yi-Chieh Wu, and Ran Libeskind-Hadas. "eMPRess: a systematic cophylogeny reconciliation tool". In: *Bioinformatics* (btaa978 Nov. 20, 2020). ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btaa978. URL: https://doi.org/10.1093/bioinformatics/btaa978 (visited on 06/07/2021).

[48] Jan Sapp. *Evolution by association*. Oxford University Press, 1994.

[49] Joel Sjöstrand, Ali Tofigh, Vincent Daubin, Lars Arvestad, Bengt Sennblad, and Jens Lagergren. "A Bayesian Method for Analyzing Lateral Gene Transfer". In: *Systematic Biology* 63.3 (May 1, 2014), pp. 409–420. ISSN: 1063-5157. DOI: 10.1093/sysbio/syu007. URL: https://doi.org/10.1093/sysbio/syu007 (visited on 01/06/2022).

[50] Justin L. Sonnenburg and Erica D. Sonnenburg. "Vulnerability of the industrialized microbiota". In: *Science* 366.6464 (2019), eaaw9255. DOI: 10.1126/science.aaw9255. eprint: https://www.science.org/doi/pdf/10.1126/science.aaw9255. URL: https://www.science.org/doi/abs/10.1126/science.aaw9255.

[51] Maureen Stolzer, Katherine Siewert, Han Lai, Minli Xu, and Dannie Durand. "Event inference in multidomain families with phylogenetic reconciliation". In: *BMC Bioinformatics* 16.14 (Oct. 2, 2015), S8. ISSN: 1471-2105. DOI: 10.1186/1471-2105-16-S14-S8. URL: https://doi.org/10.1186/1471-2105-16-S14-S8 (visited on 09/25/2019).

[52] Gergely J. Szöllosi, Eric Tannier, Nicolas Lartillot, and Vincent Daubin. "Lateral gene transfer from the dead". In: *Systematic Biology* 62.3 (May 1, 2013), pp. 386–397. ISSN: 1076-836X. DOI: 10.1093/sysbio/syt003.

[53] Gergely J. Szöllősi, Bastien Boussau, Sophie S. Abby, Eric Tannier, and Vincent Daubin. "Phylogenetic modeling of lateral gene transfer reconstructs the pattern and relative timing of speciations". In: *Proceedings of the National Academy of Sciences* 109.43 (Oct. 23, 2012). Publisher: National Academy of Sciences Section: Biological Sciences, pp. 17513–17518. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.1202997109. URL: https://www.pnas.org/content/109/43/17513 (visited on 06/05/2020).

[54] Gergely J. Szöllősi, Adrián Arellano Davín, Eric Tannier, Vincent Daubin, and Bastien Boussau. "Genome-scale phylogenetic analysis finds extensive gene transfer among fungi". In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 370.1678 (Sept. 26, 2015). Publisher: Royal Society, p. 20140335. DOI: 10.1098/rstb.2014.0335. URL: https://royalsocietypublishing.org/doi/full/10.1098/rstb.2014.0335 (visited on 10/20/2021).

[55] Gergely J. Szöllősi, Wojciech Rosikiewicz, Bastien Boussau, Eric Tannier, and Vincent Daubin. "Efficient Exploration of the Space of Reconciled Gene Trees". In: *Systematic Biology* 62.6 (Nov. 2013), pp. 901–912. ISSN: 1063-5157. DOI: 10.1093/sysbio/syt054. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3797637/ (visited on 02/18/2020).

[56] Gergely J. Szöllősi, Eric Tannier, Vincent Daubin, and Bastien Boussau. "The Inference of Gene Trees with Species Trees". In: *Systematic Biology* 64.1 (Jan. 2015), e42–e62. ISSN: 1063-5157. DOI: 10.1093/sysbio/syu048. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4265139/ (visited on 12/02/2019).

[57] Langgeng Agung Waskito and Yoshio Yamaoka. "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, 2019, pp. 1–16. ISBN: 978-3-030-21916-1. DOI: 10.1007/5584_2019_356. URL: https://doi.org/10.1007/5584_2019_356 (visited on 09/21/2021).

[58] Nicolas Wieseke, Tom Hartmann, Matthias Bernt, and Martin Middendorf. "Cophylogenetic Reconciliation with ILP". In: *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 12.6 (Nov. 1, 2015), pp. 1227–1235. ISSN: 1545-5963. DOI: 10.1109/TCBB.2015.2430336. URL: http://ieeexplore.ieee.org/document/7103015/ (visited on 10/28/2019).

[59] Bhagya K. Wijayawardena, Dennis J. Minchella, and J. Andrew DeWoody. "Hosts, parasites, and horizontal gene transfer". In: *Trends in Parasitology* 29.7 (July 1, 2013), pp. 329–338. ISSN: 1471-4922, 1471-5007. DOI: 10.1016/j.pt.2013.05.001. URL: https://www.cell.com/trends/parasitology/abstract/S1471-4922(13)00073-1 (visited on 11/05/2019).

[60] Ziheng Yang. *Computational molecular evolution*. Oxford series in ecology and evolution. Oxford: Oxford University press, 2006. ISBN: 978-0-19-856699-1 978-0-19-856702-8.

[61] Olga Zhaxybayeva and J. Peter Gogarten. "Cladogenesis, coalescence and the evolution of the three domains of life". In: *Trends in genetics: TIG* 20.4 (Apr. 2004), pp. 182–187. ISSN: 0168-9525. DOI: 10.1016/j.tig.2004.02.004.