Review of the paper Host-symbiont-gene phylogenetic reconciliation, by Hugo Menet, Alexia Nguyen Trung, Vincent Daubin and Eric Tannier.

Overall comment:

The paper considers the co-evolution of hosts, symbionts and genes within symbionts for which it proposes a three-level probabilistic model of evolution. An important contribution is the proposal of two methods to estimate reconciliations and evolutionary event rates in this probabilistic framework. The authors also propose a method to infer the symbiont phylogeny through amalgamation form gene trees and a host tree and a test to check whether considering three entangled levels is worthwhile. Reconciliation inference methods are evaluated both from simulated and real data.

Overall, this paper presents an invaluable step forward on three level reconciliations, a framework jointly considering, e.g., a host tree, a symbiont tree and gene trees (or, e.g., a species tree, a gene tree and a domain tree) linked together through evolutionary events. This problem has been previously considered in a parsimony setting by Stolzer et al, then by Li and Bansal, and in a probabilistic setting by Muhammad et al. The latter focus on inferring gene and domain trees under a Duplication-Loss (DL) model, while the focus of the current paper is on inferring reconciliations, event rates and places of these events in the reconciliations, in a probabilistic model allowing not only duplications and losses but also transfers (DTL model). The practical evaluation is convincing. In particular (but not only!) a documented transfer is only inferred if taking into account the three-level picture. This shows that the model and sampling process presented here are not only elegant but also needed in practice. Moreover reported complexities, confirmed by reported running times, show that the method is computationally efficient.

General comments/questions to authors:

Current models such as the one presented here allow gene transfers between symbionts. There are a number of documented cases where gene transfers occur between hosts and symbionts. Allowing such transfers in models would necessitate to trace also species genes (not only symbiont genes), but modelling this might not be as hard as adding a fourth level to the reconciliation model, as genes in hosts and symbionts are in essence evolving from the same ancestral genes… To what extent could your model be extended to integrate this kind of events?

You consider *undated* trees and somewhere you remark that indeed this might lead to time inconsistent scenarios sometimes. Though in the experimental part of the paper you never come back to this point. Could you state the percentage of unfeasible scenarios you get in the sampling process both on simulated and biological data? I think this is a point that deserves some stats in the experiments.
Also, from an unfeasible (time inconsistent) scenario, do you have a method that brings to a feasible scenario? Would that be possible without loosing too much in likelihood? Please say something about that in the paper (or that this point remains to be investigated).

It might seem the Monte Carlo approach you propose is from the importance sampling family. Could you state that explicitly (and give reference) or indicate in which aspect it differs from this family of other MC methods?

Detailed comments:

I lacked time to make several reading passes over the paper and to read in detail some recent related works, which might explain the relatively large number of comments / questions I mention below. But overall, I'm confident with the fact that the paper is a useful addition to the known theory and practice in the reconciliation field.

P1, L16: « *hosts, symbionts and <u>symbiont</u> genes* » would be more precise than « *and their genes* ». There are other places in the paper were you might be more explicit

P1, L20: please indicate that you first fix the H/S reconciliation before sampling S/G reconciliations.
P2, L79: « *it is possible to jointly handle three nested levels in a single computational model …* »: this sentence is overly long, please break it in two.

P5, Fig1 is rather useful.
P5, L169: « *The inference consists in* » -> « *begins with* » + state that this is done for all gene trees independently (as G can also sometimes denote a set of gene trees).
P5,L169-171: how do you estimate the D,T,L rates for the S/H reconciliation?
P5, L172: « it is then possible to estimate the evolutionary rates »: based on the sampled scenarios?
P5, L173: « among reconciliation scenarios »: of both H/S and S/G?
P5, L175: it is a bit surprising that a probabilist approach does not take branch length into account, usually, this is a plus of such methods over parsimony ones. For instance the probability of a non-speciation event would intuitively be considered smaller for short branches.
P5+6: in sections 3.2+3.5 there is an important Figure missing somewhere here that would clearly depict the steps of the inference process, with their respective input and how they coordinate together. The text in its current state is not enough for me to be sure of the whole multi-step inference process. I'm sure readers would consider this a useful addition.
P6, L196, « In our model, gene transfer » (add a comma there)
P6, L204: « *while being in the same target (?) host* » (currently there is some ambiguity)
P7, Figure2: the clarity of this figure should be improved.
P7,L220: « *among the |S_h| symbiont branches present in h* » (might be worth precising).
P7, équation (4): the last sum is over $k\in H$ ancestor branches of e?
P8, Figure 3: in the middle picture on the top row: is that a transfer from the dead?
P9, Table 1: the legend is not self contained. State what is 2-level in particular (G/S without being aware of the H/S co-evolution?)
P9, section 3.5: please number the inference steps and refer to these numbers whenever appropriate. Here again the lack of a figure depicting the inference process makes it hard to parse the text.
P9,L276: « *Finally we can compute the host aware gene/symbiont reconciliation* »: not coming back to put into question the h/s reconciliation? Why not then? And if not, this seems like a contradiction with what is stated on page 12.
P9,L278: « *on the donor-receiver symbiont couple »?*
P9,L282: « we repeat all steps except the initial host/symbiont reconciliation » -> « *we repeat steps 2 and 3*»?
P10, L293: running times seem quite reasonable given the model sophistication, good! Is that a result of the fact that you do not entangle together the three levels but rather consider the first two levels, then the second and third level with only the knowledge for the second level of whether several symbionts belong to the same host or not (ie you mostly consider partial information from the h/s reconciliation, maybe because only this partial information is relevant to the s/g reconciliation? Could you elaborate a bit in the paper on this.
P10,L307: « *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* »: ok, the inference process might loose a lot of its potential accuracy doing that?
P12, L366: does this mean that gene trees are rather similar to one another? Or gene trees are close to symbiont trees? Could you give distance measures between input trees? (dRF or dMAST for instance)
P12,L378-380: here R(S,H) is sampled which seems to be a contradiction with what is said in section 3.5
P17, L511-512: it might be a good place to recall that the program is available on GitHub, this was only briefly mentioned in the abstract.
P19,L549-550: this is quite honest from you to recognise this.

Concerning the **code repository**: beware that some comments are not in English, for instance in main.py.