bioRxiv preprint doi: https://doi.org/10.1101/166215; this version posted July 20, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

# A geography aware reconciliation method to investigate host/parasite cospeciation

**V. Berry[1], F. Chevenet[1&2], J-P. Doyon[1], E. Jousselin[3]**

[1] Institut de Biologie Computationnelle, LIRMM, UMR 5506 CNRS, Université de Montpellier, France

[2] MIVEGEC, CNRS 5290, IRD 224, Université de Montpellier, France

[3] **CBGP, INRA, CIRAD, IRD, Montpellier SupAgro, Univ. Montpellier,** Centre de Biologie pour la Gestion des Populations, 755 avenue du campus Agropolis CS 30016 34988 Montferrier-sur-Lez, **France**

**Corresponding authors: V. Berry vberry@lirmm.fr and E. Jousselin emmanuelle.jousselin@inra.fr**

**Running title (45 characters): Geography aware reconciliation**

**ABSTRACT (250 WORDS)**

Cospeciation studies aim at investigating whether hosts and symbionts speciate simultaneously or whether the associations diversify through host shifts. This problem is often tackled through reconciliation analyses that map the symbiont phylogeny onto the host phylogeny by mixing different types of diversification events. These reconciliations generate numerous scenarios that are difficult to interpret and not always biologically realistic. Researchers have underlined that the biogeographic histories of both hosts and symbionts influence the probability of cospeciation and host switches, but up to now no reconciliation software integrate geographic data. We present a new functionality in the *Mowgli* software that bridges this gap. The user can provide geographic information on both the host and symbiont extant and ancestral taxa. Constraints in the reconciliation algorithm have been implemented to generate biologically realistic codiversification scenarios. In addition, we updated the reconciliation viewer *SylvX* in order to visualize ancestral characters states on the phylogenetic trees. We apply our method to the fig/fig wasp association, a textbook example of cospeciation, and demonstrate that it generates cospeciation scenarios that are geographically consistent. In addition, we show that the comparison of reconciliations obtained with and without geographic constraints can help solving ambiguities in the biogeographic histories of the partners. With the development of robust methods in historical biogeography and the advent of next generation sequencing that leads to better-resolved trees, a geography aware reconciliation method represents a substantial advance that is likely to be useful to researchers studying the evolution of biotic interactions and biogeography.

**Keywords: Ancestral trait, biogeography, cophylogeny, host/parasite interactions, software, , tree visualization.**

bioRxiv preprint doi: https://doi.org/10.1101/166215; this version posted July 20, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

## 1) INTRODUCTION

Biotic interactions play a prominent role in species diversification. Interactions that result into long-term associations persisting over evolutionary time scales can sometimes lead to cospeciation, i.e. the concomitant occurrence of speciation in lineages that are ecologically associated (Brooks 1981; Page 1990, 1991). The idea that such a pattern can occur first stemmed from parasitological studies suggesting that parasite classifications reflect the phylogenetic relationships of their hosts (Fahrenholz 1913). Hafner & Nadler (1988; 1994) were the first authors to thoroughly test this assertion. They used the association between some small mammals (pocket gophers) and their lice as a model system and provided a clear demonstration that the phylogenies of the two interacting lineages were parallel. This study, spurred further research on coevolutionary diversification. The developments of specific methods that aimed at testing the congruence of the phylogenetic histories of interacting organisms have since played an important role in the study of cospeciation. It is indeed these methods that moved cospeciation studies beyond visual comparisons of phylogenetic trees and ad-hoc narratives for these visualizations. It soon became apparent that the study of the concordance between phylogenetic trees could be applied to reconciling gene trees and species trees (Page & Charleston 1997; Page & Charleston 1998) which further enhanced the interest of evolutionary biologists for methodological developments in this field.

Reviews on cospeciation methods (Brooks *et al.* 2004, Doyon, 2011 , de Vienne *et al.* 2013; Johnson & Clayton 2004; Martínez-Aquino 2016; Paterson & Banks 2001; Stevens 2004) all emphasize the diversity and the complexity of the scenarios that must be explored when testing for the congruence of speciation events in two interacting lineages. To compare host and parasite phylogenies, Brooks & McLennan (Brooks 1981; Brooks & McLennan 1991) first developed a parsimony method (the Brooks Parsimony Analysis, BPA). In this method the associations between hosts and their parasites are transformed into a matrix of host characters and the parsimony tree reconstructed from such a matrix is then compared to the host phylogeny. A decade later, Page and collaborators developed a fundamentally different method, called “tree reconciliation”, a term first coined in the work of Goodman *et al.* (1979) that compared gene and species trees. This method attempts to reconcile the phylogenetic history of the parasite with that of their hosts: the parasite phylogeny is “mapped” onto the host phylogeny (i.e. each node in the parasite tree is assigned to a node in the host phylogeny). In such a map, the diversification events of the parasites are linked to their host phylogenetic history, so that four types of events are considered: cospeciation events, host switches, sorting events and duplication events (Page 1994a; Page 1994b) (see material and method for a description of each event). When graphically displayed, reconciliation maps greatly ease our understanding of the evolution of biotic interactions.

Algorithms to optimize reconciliations are numerous. One of the first reconciliation software, *TreeMap* 2, uses an algorithm called "Jungles" (Charleston 1998) where each event is assigned a cost: the chosen reconciliations are the ones that have minimum costs. However it generates an exponential number of scenarios. Recent methods have proposed more efficient algorithms that moreover can just search for an optimal reconciliation: e.g. *Tarzan*, (Merkle & Middendorf, 2005); *Jane*, (Conow *et al.* 2010) ; *Core-PA* (Merkle *et al.*, 2010), *Mowgli*, (Doyon *et al.* 2010), *COALA* (Baudet *et al.* 2015), ecce*TERA* (Jacox *et al.* 2016), *Notung* (Stolzer *et al.* 2012). Recently, the *RASCAL* software proposed to infer suboptimal scenarios to achieve lower computing times (Drinkwater & Charleston 2016). Cospeciation is witnessed on a reconciliation map whenever a speciation node in the parasite phylogeny is mapped onto a speciation on the host phylogeny. Another requirement for demonstrating that two interacting lineages have cospeciated is to provide evidence of the temporal congruence of the cospeciation events in the host and parasite phylogenies (Page 1991). Though reconciliation algorithms do not strictly enforce the simultaneity of cospeciation events, they can enforce time

bioRxiv preprint doi: https://doi.org/10.1101/166215; this version posted July 20, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

consistency in the sequence of evolutionary events meaning that the parasite cannot switch back in time onto a host that no longer exists (*i.e.* transfers cannot occur towards a node in the host phylogeny that has already split into child species at the time of the transfer event) (Merkle *et al.* 2005; Nojgaard *et al.* 2017). This constraint is explicit in *Mowgli* (Doyon *et al.* 2011; Doyon *et al.* 2010), *ecceTera* (Jacox *et al.* 2016) and *RASCAL* (Drinkwater & Charleston 2016). Hence, reconciliation methods have greatly improved in the last decade; algorithms are now efficient and some algorithms have solved the time consistency issue that affected some of the first methodological developments in the field. However, the multiplicity of optimal evolutionary scenarios can still be daunting. It is generally difficult to sort these scenarios out a posteriori and identify the most realistic ones. Much remains to be done to improve these inferences and translate them into evolutionary scenarios that give insights into the biological factors that govern the evolution of interspecific associations.

Some key information that could significantly improve our inferences but are overlooked in codiversification methods are the geographic locations of extant and ancestral nodes. Indeed, the biogeographic histories of interacting lineages necessarily constrain their common part of evolutionary history (Alcala *et al.in press*, Martinez-Aquino *et al.* 2014; Nieberding *et al.* 2010). Obviously, cospeciation can only happen between taxa that co-occur in the same area. Host switches can also be influenced by the geographic proximity of the "sending" and the "receiving" hosts. In their seminal paper, Hafner & Adler (1988) highlight that "no case of suspected host-switching involves hosts that are geographically distinct" implying that this was evidence that their codiversification scenario was plausible. Indeed, in parasites that live their entire life cycle onto their hosts and have little or no dispersal forms (e.g. Clayton *et al.* 2004), transfers can only happen between hosts that coexist in the same locality. On the other hand, in biotic interactions where the symbionts can experience long dispersal events, transfers do not need to happen between sympatric hosts. However, they are only possible if the "source" and "destination" host geographic locations coincide with the geography of the dispersal events along the corresponding branches in the parasite phylogeny. Therefore, a more accurate mapping of cospeciation and host switch events can be obtained if the geographic locations of both hosts and parasites are known prior to conducting the reconciliation. Methods for inferring historical biogeography from phylogenetic reconstructions have improved in the last decades and generate robust scenarios (Beeravolu Reddy & Condamine 2016; Meseguer *et al.* 2015; Ree & Sanmartin 2009; Ronquist *et al.* 2011). These can serve as inputs for reconciliation analyses. In this paper we build on these advances to provide a geography-aware reconciliation method, pushing further the realism of scenarios proposed by such methods.

We first describe constraints that need to be enforced to ensure geographic consistencies in reconciliations and implement them in the *Mowgli* software that already generates time-consistent reconciliations (Doyon *et al.* 2010). We also update the reconciliation viewer, *SylvX* (Chevenet *et al.* 2016), in order to integrate and visualize annotations (e.g. geographic areas) at ancestral nodes for the host and parasite phylogenies. We then test these new developments on a mock dataset and on a 'textbook' example of cospeciation, the interaction between figs (*Ficus*) and their pollinating fig wasps (Cruaud *et al.* 2012).We show on this example that our approach can filter out scenarios that are geographically implausible and point out inconsistencies in the historical biogeography of the lineages involved in the interactions.

## 2) METHODS

### Extending *Mowgli* to account for geographic information

bioRxiv preprint doi: https://doi.org/10.1101/166215; this version posted July 20, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

In this section we first recall the reconciliation model followed by *Mowgli* (Doyon *et al.* 2010), in particular the precise set of evolutionary events it considers.

Only rooted trees are considered, whose leaf nodes (tips) are each labelled by a taxon. Internal nodes usually have two descendants, but the case where an internal node has a single child also occurs when the evolution of an ancestral lineage living a relatively long period of time is decomposed into a set of consecutive time periods called *slices* (see Fig. 3 of Supplementary Material 1). This slicing of branches is a transparent artefact that allows reconciliation methods to achieve fast computing times while still ensuring time consistency of host switches (see Doyon et al. 2010; Jacox *et al.* 2016; Libeskind-Hadas & Charleston 2009).

Let *P* and *H* denote respectively a parasite and a host tree. Reconciliation algorithms usually consider each current and ancestral host to be associated with one or several specific parasites at any time (e.g. in *Mowgli*, *TreeMap*, *Jane*). However, the identity of the host can vary over time, *e.g.* after a *host switch*. This evolutionary event is one of the four types of *events* considered in cospeciation studies:

- a host switch, also known as a host transfer (T event), happens when a parasite lineage from a source host is transferred to a destination host. The transfer of the parasite has to be time consistent, that is the two branches $(x_p,x)$ and $(x'_p,x')$ have to belong to the same time slice;

- a cospeciation (S event) happens when the speciation of a parasite shortly follows or coincides with the speciation of its host. This is considered by *Mowgli* as a joint speciation of both parasite and its host;

- a *within host speciation* also known as duplication (D event), models a speciation of a parasite *u* of *P*, where both descendant species continue to live on the host that *u* lived on. This is represented by *u* evolving along a $(x_p,x)$ branch of H and then dividing into two new lineages in $(x_p,x)$;

- a parasite loss (L event) occurs when a parasite lineage goes extinct while its host persists.

An illustration of these events can be found in Fig.1 of Supplementary Material 1.

*Mowgli* also sometimes considers combinations of events in order to speed up computations. A SL event occurs shortly after a cospeciation (S): one of the parasite child lineages is quickly lost on a child lineage of the host. A TL event occurs when a parasite *u* evolving on a branch $(x_p,x)$ is lost (L) on this branch shortly after having switched (T) to another host $(x'_p,x')$.

As explained above, accounting for geographic information can lead to more realistic codiversification scenarios. We integrate such information first by assigning a set of *areas* to each node of *P* and *H*. For an extant taxon this means that a population of the corresponding species is reported to live in each of the assigned area. Similarly, when an internal taxon is assigned to one or several areas, this means that populations of this now extinct taxon are inferred to have lived at the time in these geographical zones. The inference of ancestral areas can be conducted with different approaches and tools, *e.g.* ancestral character state reconstruction using Maximum Likelihood optimization as implemented in *R* (Revell, 2012) or the dispersal-extinction-cladogenesis (DEC) model implemented in *Lagrange* (Ree & Smith 2008).

In order to avoid biogeographic inconsistencies and propose meaningful reconciliations between the *P* and *H* trees, specific constraints have to be implemented in reconciliation algorithms. We detail below how we model these constraints in the context of the four D/T/L/S events or combinations thereof.

bioRxiv preprint doi: https://doi.org/10.1101/166215; this version posted July 20, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

First, note that areas of a node and its parent in the host or parasite tree can be different, due to dispersal and vicariance events. During the reconciliation process, the time period represented by a branch between nodes $x_p$ and $x$ of the host tree is considered to be assigned the union of the areas of $x_p$ and $x$: if a species changes area along the way from one area assigned to $x_p$ to a different one assigned to $x$, we do not know exactly when this happens, so we consider that at any time between $x_p$ and $x$, part of the population of the evolving species can live in any area proposed for $x_p$ or $x$.

First considering nodes of the trees, we denote by *area(x)* the set of geographic areas where a species $x$ is observed (tip) or inferred (internal node). If $x$ is a fictive node, in the middle of branch $(y_p,y)$ resulting from the slicing of the initial tree $H$, then *area(x)* is defined as the union of *area(y)* and $area(y_p)$.

Now considering branches $(x_p,x)$ of the $P$ and $H$ trees, let $area(x_p,x)$ be the set of areas where the species might have lived during this period of its evolution: this is the union of $areas(x_p)$ and *area(x)*. Note that each area for which exactly one of the two species $x$ and $x_p$ is present represents species migration or extinction events that have occurred along this branch. Note that only $(x_p,x)$ branches being one slice high are considered for $H$, as Mowgli operates on this level of detail.

We now detail which geographic constraints apply so that the reconciliation between a parasite tree $P$ and a host tree $T$ is geographically consistent.

- An extant parasite $u$ can be mapped onto a current host $x$, only if $area(u) \subseteq area(x)$ (Figure 1A). The reconciliation mapping cannot be computed by *Mowgli*, if this constraint is not respected.

- We allow the mapping of an ancestral parasite $u$ at a speciation node $x$ in the host tree, only if $area(u) \subseteq area(x)$ (Figure 1B). Each area where $x$ is present and $u$ is not (i.e., $area(x) \setminus area(u)$) represents a population of host $x$ that does not have $u$ as a parasite.

- If a branch $(u_p,u)$ of the parasite is going through a node $x$ of the host tree (which happens when the host speciates into two descendant hosts but the parasite sticks to only one of them), then the sets of area($x$) and area($u$ $u_p$) must have common elements (Figure 1C).

- If a branch $(u_p,u)$ of the parasite tree is mapped for all or part of it in a host branch $(x_p,x)$ (Figure 1D), then we require that $area(u_p,u)$ and $area(x_p,x)$ have a non-empty intersection, i.e. that there is at least one area where the parasite and the host were able to meet.

- If an internal node $u$ of $P$ is mapped in a branch $(x_p,x)$ of the host tree (Figure 1E) then we require that $area(u) \subseteq area(x_p,x)$.

When respecting the above constraints, *Mowgli* will propose a scenario that os geographically consistent. However, this scenario can have a higher cost than those obtained when not accounting for geographic information. This simply results from the fact that the search space contains geographically inconsistent scenarios that are possibly less costly. *Mowgli*'s extension described above, allows choosing the less costly scenario among those that are geographically consistent.

### ***SylvX* new functionalities**

We extended the *SylvX* editor in order to visualize current and ancestral geographic areas of hosts and symbionts. Pie charts can be used to display alternative areas for each node of the tree and/or the reconciliation. Area colour sets can be dynamically updated and tuned using the Hue, Saturation and Value scales. Thresholds are available to simplify views.

bioRxiv preprint doi: https://doi.org/10.1101/166215; this version posted July 20, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

### Implementation

*Mowgli* takes a "host tree" and a "parasite tree" in Newick format. A list of nodes with their geographic annotations can be given in the same file. Biogeographic inferences typically generate probability or likelihood values for each character state (area) at each node. *Mowgli* can take a single area at each node: the most likely area for instance. Alternatively, when several areas are specified in *Mowgli*, they are interpreted as a multi-state character. In this case, the geographic area of a node is considered as the combination of all listed areas and therefore any parasite living in at least one of these areas can be mapped on this node (see Supplementary Material 1).

*SylvX* takes a host tree in a Newick format with node id numbers together with a reconciliation (with symbiont tree node *id*.). The host tree (*outputspeciestree*) and reconciliations obtained with *Mowgli* can be directly imported into *SylvX*. Annotation files for the host and parasite phylogenies giving node information can be imported in a CSV format. As many annotations as needed can be added in the annotation files and it is up to the user to choose which ones to plot onto the species tree and the reconciliation map through *SylvX* interface.

In order to obtain files that are compatible between both software, we provide a *Perl* script that can be run through single command lines in order to: 1) obtain tree node identifiers that will be used by both programs and 2) transform these trees and corresponding annotations file into *Mowgli* format (see Supplementary Material 2 for a description of the full procedure to generate files and perform a complete analysis).

## 3) WORKED EXAMPLES

### Datasets

To demonstrate the method and its utility, we tested it on two datasets. We first created a mock dataset: two phylogenetic trees for a hypothetical host/parasite interaction in which extant and ancient geographic areas for each lineage are informed. The dataset was generated so that: 1) present-day geographic areas of associated taxa are consistent (i.e. hosts and associated parasites live in the same area); 2) the two phylogenies are not perfectly parallel but show some cospeciation events; 3) some geographic locations at nodes that we would like to cospeciate do not coincide in the parasite and host phylogenies. We ran *Mowgli* on this dataset successively with and without enforcing geographic constraints using in both cases the default parameters (cost 0 for a cospeciation, 1 for a loss and 1 for a transfer, 1 for duplication, not enforcing the root of the parasite tree to be mapped on the root of the species tree).

As a second dataset, we used a subset of the data from the latest phylogenetic investigation of figs (*Ficus*) and their pollinating figs wasps (Cruaud *et al.* 2012). For both partners of the association, biogeographic scenarios were available for 200 taxa. From the complete phylogenetic trees (available in http://datadryad.org, doi: 10.5061/dryad.hr620), we derived two trees of 23 taxa each, that included a couple of representative species for each *Ficus* main taxonomic subdivision. We excluded one of the fig subgenera (*Pharmacosycea*) and its associated pollinators (*Tetrapus* spp.) whose phylogenetic position is still debated. We have not tested our method on the total dataset presented in Cruaud *et al.* (2012) as some uncertainties remain concerning the root of the phylogenetic trees, which could lead to spurious interpretations. The most likely ancestral geographic areas of each node were directly derived from the biogeographic reconstructions of Cruaud *et al.* 2012 (obtained with Maximum Likelihood Optimization). We ran *Mowgli* on this dataset with and without enforcing geographic constraints (using default event costs and not enforcing the root of the parasite tree to be mapped on the root of

bioRxiv preprint doi: https://doi.org/10.1101/166215; this version posted July 20, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

the species tree), and explored how these reconstructions shed light on the biogeographic history of the association.

**Results**

Figure 2 represents the reconstruction obtained on the mock dataset. Taking geographic constraints into account (Figure 2B) precludes obtaining a cospeciation event at a node where the two associates do not live in the same area (node S1 of the host tree in Figure 2A and 2B). The transfer T1 preceding this cospeciation event in Figure 2A is also geographically inconsistent as it suggests a dispersal (the donor host lives in Asia or Africa, and the receiving host lives in America) while the parasite actually stays in Africa. The scenario obtained when enforcing geographic constraints is more costly but consistent (Figure 2B): it entails one more transfer and also one less cospeciation event but is biologically more realistic.

On the fig/fig wasp dataset (Figure 3), not accounting for geographical constraints leads to geographic inconsistency in one node (S1 in the host tree of Figure 3A) and the transfer that follows (event T7 on Figure 3A). Enforcing geographic constraints generates a reconciliation scenario that is more costly (Figure 3B, one more transfer is necessary to reconcile the two phylogenies) but coherent with the figs and the fig wasps biogeographic histories. This scenario suggests that the fig wasps independently colonized figs in the Neotropics and in the Afrotropics through two distinct host switches from Asia rather than accompanied the speciation of their hosts, as was suggested by Figure 3A (and node 29 of Figure S12 in Cruaud *et al.* 2012). The annotation of ancestral geographic areas on the reconciliation map also shows that host switches occurred both in "sympatric" settings (within the same geographic areas as broadly defined in our dataset) and allopatric settings (i.e. host switches occur between two geographically distant hosts). Overall, four switches out of seven occurred in sympatry (T1, 2, 4, 5) while the remaining three switches (T3, 6, 7) correspond to long distance dispersal events.

**4 DISCUSSION**

We provide here significant extensions for two reconciliation tools to infer co-diversification scenarios that, for the first time, can take the historical biogeographies of the associated lineages into account. The extension of the *Mowgli* software precludes geographic inconsistency during the reconciliation process. The resulting reconciliations can then be visualized and edited in the *SylvX* updated graphical interface that now integrates annotations of ancestral nodes. *Mowgli* is already one of the few time-consistent efficient methods that build optimal reconciliations. With the integration of geographic constraints in its algorithm, this tool now provides more realistic codiversification scenarios than other reconciliation methods. In some cospeciation studies, a large number of optimal reconciliations can be proposed for a combination of compared trees (Nguyen *et al.* 2013) and their interpretation can be daunting. Producing biologically realistic scenarios can diminish the complexity of these coevolutionary analyses by reducing the optimal set of scenarios to a handful of possibilities that can be further scrutinized for biological interpretation. In addition, geography-aware reconciliations can reveal whether host switches generally occur in sympatry or whether they are associated with dispersal events of the symbionts. Alternatively, comparing reconciliations can point out discrepancies in the biogeographic reconstructions of the host and symbiont lineages under investigation.

In the particular example of the fig/fig wasp interaction presented here, the geographic inconsistency revealed at one of the cospeciating nodes in the analysis ran without constraints might actually point out some ambiguity in the biogeographic history of the *Ficus* hosts. According to the

bioRxiv preprint doi: https://doi.org/10.1101/166215; this version posted July 20, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

inference conducted in Cruaud *et al.* (2012) the most likely area for the common ancestor of Neotropical figs (belonging to the *Americana* section) and Afrotropical figs (belonging to the *Galoglychia* section) is Africa while the proposed cospeciating pollinators lived in Asia (S1 Figure 3A). In order to respect geographic consistency, our geography-aware reconciliation suggests that the current association of figs wasps with *Galoglychia* in Africa and *Americana* in the Neotropics is the result of two independent switches (Figure 3B, events T6 and T7) of the pollinators from an Asian fig ancestor (the ancestor of the *Conosycea* figs). However, the biogeographic analysis of Cruaud *et al.* (2012) also suggested that the node S1 of *Ficus* could be situated in Asia (though with a smaller likelihood than the Afrotropics). "Correcting" the biogeographic inference of *Ficus* and locating node S1 in Asia would actually reconcile the host/fig and fig wasp trees as in Figure 3a without violating any geographic constraints. It would diminish the cost of the reconciliation by inducing one less transfer and entail one more cospeciation event. We will not conclude on the biogeographic history of the fig/fig wasp association as the purpose of our study is not to explore alternative scenarios for this association. It is merely to demonstrate the utility of our method in revealing inconsistency between biogeographic scenarios and a cospeciation hypothesis and therefore improve codiversification scenarios and/or biogeographic inferences

**Perspectives**

The cospeciation tools developed in this study could be applied to all interspecific interactions for which biogeographic scenarios are available for both partners. For now those are few, but fast developments in new sequencing technologies generate more accurate and more exhaustive phylogenies. Methods in historical biogeography have also improved. So we can hope that more datasets will be available in the near future and that cospeciation could be tested on more systems (Cruaud & Rasplus 2016). For instance, robust phylogenies and biogeographic scenarios are now available for groups of lice that have been model systems in coevolutionary studies (Boyd *et al.* 2017a). Once comprehensive phylogenies of the hosts are available, our method could be used to better understand the geography of host switches in this system. Geography-aware reconciliation could also be applied to explore the diversification history of the numerous parasitic wasps that are part of the microfauna exploiting figs: several lineages of parasitic wasps have been shown to partly cospeciate with their host figs (Jousselin *et al.* 2008) and biogeographic scenarios for some lineages are available (Cruaud *et al.* 2011). Other nursery pollination/mutualisms such as the interaction between *Yucca* and their pollinating moths are also good candidates for including geographic constraints into coevolutionary scenarios as some studies have questioned the respective role of geography and host-plant association in driving the diversification of Yucca moths (Althoff *et al.* 2012). Parasitoid/host insect associations (e.g. Deng *et al.* 2013; Hamerlinck *et al.* 2016; Wilson *et al.* 2012) and herbivorous insect/plant interactions (*e.g.* McLeish *et al.* 2007; Percy *et al.* 2004) showing significant cospeciation could also be studied.

Furthermore, the approach presented in this paper does not only apply to geographic information and could be extended to other biological traits. For instance in systems where the species are partitioned into different habitats (*e.g.* forest canopy species vs savannah species), geographic areas could be replaced by traits related to the ecological niches; constraints that are similar to the ones applied for geography could then be easily transferrable. Informing ancestral characters for habitats on the host and the symbiont phylogenetic trees and using "*Mowgli* with constraints" would result in constraining cospeciation and host switches to associates sharing the same ecological habitats. In a similar way, the respective climatic niches of associated organisms could also be used when parasite (or symbiont) distributions are known to be strongly constrained by thermal tolerance (see Singh *et al.* 2017 for a recent study showing that climatic conditions influences the patterns of association between

bioRxiv preprint doi: https://doi.org/10.1101/166215; this version posted July 20, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

fungi and their algal partners). In many specialized interactions that show significant cospeciation, such as host/obligate bacterial endosymbionts (*e.g.* Jousselin *et al.* 2009, Rosenblueth *et al.* 2012, Boyd *et al.* 2017b) or host/viruses associations (e.g. Ramsden *et al.* 2009; Garamszegi 2009), inferring ancestral character states for some ecological traits for the "symbiotic" lineages (the parasite) is not always straightforward. However, the evolution of these obligate associations and their maintenance are still governed by some phenotype matching between the partners. For instance in host/bacterial symbiont associations, the metabolic complementarity of the host and the symbiont (Zientz *et al.* 2004) could be reconstructed and used to constrain the reconciliations. In host/virus associations, information about the host immune system and viruses adaptations could be used (Longdon *et al.* 2014). The extension of *Mowgli* proposed here could probably be adapted to fit the biological properties of these associations

Independently of the new functionalities implemented in *Mowgli*, the concomitant update of *SylvX* allows the comparison of ancestral states for any character of the hosts and/or the symbionts. This can help interpreting reconciliations by replacing them in their biological context. One of the most useful functionalities of *SylvX* is now to be able to visualize whether host switches are associated with evolutionary transitions in character states in both the parasite and/or the host. It can therefore help understanding the biological processes that are associated with these transfers. Mapping characters of the associates throughout the reconciliation can also help investigating whether there is correlated trait evolution in host and parasites. Until now, such correlations could only be investigated on one of the associates phylogeny (*e.g.* Sorci *et al.* 2003, Jousselin *et al.* 2003, Johnson *et al.* 2005). Looking at simultaneous transitions in character states in both partners throughout a host/parasite reconciliation might help identifying co-adapted traits that constrain the association.

In conclusion, we provide here a framework that can integrate the character histories of the associates into the reconciliation process. This allows recovering more biologically realistic scenarios. It can also shed lights on character history inferences by pointing out inconsistencies between the character states of the two associates on the reconciliation map. The new developments made in *SylvX* facilitate these interpretations. The use of "constraint-aware" reconciliations is preferable to current practices that consist in elaborating ad-hoc narratives once the reconciliations are obtained and compared with the character histories of associates.

## ACKNOWLEDGMENTS

We thank A. Cruaud, JY Rasplus & coll. for sharing the fig/fig-wasp dataset. *Funding*: ANR Phylospace.

## AUTHOR CONTRIBUTION

VB, JPD and EJ designed the study. VB and JPD developed *Mowgli*, FC developed *SylvX*. VB and EJ wrote the manuscript with contributions of FC.

## SOFTWARE AVAILABILITY

*SylvX* is available on www.sylvx.org with manual and example files and can be installed on any platforms. Mowgli is available Dryad doi:10.5061/dryad.78bd0 (and will be shortly on http://www.atgc-montpellier.fr/Mowgli/) with manual and example files and can be installed on Linux and OSX (Mac) computers, *GeoRecHelper* and example files are available on Dryad doi:10.5061/dryad.78bd0.

## REFERENCES

bioRxiv preprint doi: https://doi.org/10.1101/166215; this version posted July 20, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Alcala N, Jenkins T, Christe P, Vuilleumier S Host shift and cospeciation rate estimation from co-phylogenies. *Ecology Letters, in press*.

Althoff DM, Segraves KA, Smith CI, Leebens-Mack J, Pellmyr O (2012) Geographic isolation trumps coevolution as a driver of yucca and yucca moth diversification. *Molecular Phylogenetics and Evolution* **62**, 898-906.

Baudet C, Donati B, Sinaimeri B, *et al.* (2015) Cophylogeny Reconstruction via an Approximate Bayesian Computation. *Systematic Biology* **64**, 416-431.

Beeravolu Reddy C, Condamine F (2016) An Extended Maximum Likelihood Inference of Geographic Range Evolution by Dispersal, Local Extinction and Cladogenesis. *BioRxiv*. https://doi.org/10.1101/038695

Boyd BM, Allen JM, Nguyen N-P, *et al.* (2017a) Phylogenomics using Target-Restricted Assembly Resolves Intrageneric Relationships of Parasitic Lice (Phthiraptera: Columbicola). *Systematic Biology*, syx027.

Boyd BM, Allen JM, Nguyen N-P, *et al.* (2017b) Primates, Lice and Bacteria: Speciation and Genome Evolution in the Symbionts of Hominid Lice. *Molecular Biology and Evolution* **34**, 1743-1757.

Brooks D (1981) Hennig's parastological method: a proposed solution *Systematic Zoology* **30**, 229-249.

Brooks D, McLennan DE (1991) *Phylogeny, Ecology, and Behavior*, Chicago.

Brooks DR, Dowling APG, van Veller MGP, Hoberg EP (2004) Ending a decade of deception: a valiant failure, a not so-valiant failure, and a success story. *Cladistics* **20**, 32-46.

Charleston MA (1998) Jungles: a new solution to the host/parasite phylogeny reconciliation problem. *Mathematical Biosciences* **149**, 191-223.

Chevenet F, Doyon JP, Scornavacca C, *et al.* (2016) *SylvX*: a viewer for phylogenetic tree reconciliations. *Bioinformatics* **32**, 608-613.

Clayton DH, Bush SE, Johnson KP (2004) Ecology of congruence: past meets present. *Systematic Biology* **53**, 165-173.

Conow C, Fielder D, Ovadia Y, Libeskind-Hadas R (2010) *Jane*: a new tool for the cophylogeny reconstruction problem. *Algorithms for Molecular Biology* **5**.

Cruaud A, Jabbour-Zahab R, Genson G, *et al.* (2011) Out of Australia and back again: the world-wide historical biogeography of non-pollinating fig wasps (Hymenoptera: Sycophaginae). *Journal of Biogeography* **38**, 209-225.

Cruaud A, Rasplus J-Y (2016) Testing cospeciation through large-scale cophylogenetic studies. *Current Opinion in Insect Science* **18**, 53-59.

Cruaud A, Ronsted N, Chantarasuwan B, *et al.* (2012) An Extreme Case of Plant-Insect Codiversification: Figs and Fig-Pollinating Wasps. *Systematic Biology* **61**, 1029-1047.

de Vienne DM, Refregier G, Lopez-Villavicencio M, *et al.* (2013) Cospeciation vs host-shift speciation: methods for testing, evidence from natural associations and relation to coevolution. *New Phytologist* **198**, 347-385.

Deng J, Yu F, Li H-B, *et al.* (2013) Cophylogenetic relationships between Anicetusparasitoids (Hymenoptera: Encyrtidae) and their scale insect hosts (Hemiptera: Coccidae). *BMC Evolutionary Biology* **13**, 275.

Doyon J-P, Ranwez V, Daubin V, Berry V (2011) Models, algorithms and programs for phylogeny reconciliation. *Briefings in Bioinformatics* **12**, 392-400.

Doyon JP, Scornavacca C, Gorbunov KY, *et al.* (2010) An Efficient Algorithm for Gene/Species Trees Parsimonious Reconciliation with Losses, Duplications and Transfers. In: *Comparative Genomics* (ed. Tannier E), pp. 93-108.

bioRxiv preprint doi: https://doi.org/10.1101/166215; this version posted July 20, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Drinkwater B, Charleston MA (2016) RASCAL: A Randomized Approach for Coevolutionary Analysis. *Journal of Computational Biology* **23**, 218-227.

Fahrenholz H (1913) Ectoparasiten und Abstammungslehre. *Zool. Anz.* **41**, 371-374.

Garamszegi LZ (2009) Patterns of co-speciation and host switching in primate malaria parasites. *Malaria Journal* **8**.

Goodman M, Czelusniak J, Moore GW, Romero-Herrera AE, Matsuda G (1979) Fitting the Gene Lineage into its Species Lineage, a Parsimony Strategy Illustrated by Cladograms Constructed from Globin Sequences. *Systematic Zoology* **28**, 132-163.

Hafner MS, Nadler SA (1988) Phylogenetic trees support the coevolution of parasites and their hosts. *Nature* **332**, 258-259.

Hafner MS, Sudman PD, Villablanca FX, *et al.* (1994) Disparate rates of molecular evolution in cospeciating hosts and parasites. *Science* **265**, 1087-1090.

Hamerlinck G, Hulbert D, Hood GR, Smith JJ, Forbes AA (2016) Histories of host shifts and cospeciation among free-living parasitoids of Rhagoletis flies. *Journal of Evolutionary Biology* **29**, 1766-1779.

Jacox E, Chauve C, Szollosi GJ, Ponty Y, Scornavacca C (2016) *ecceTERA*: comprehensive gene tree-species tree reconciliation using parsimony. *Bioinformatics* **32**, 2056-2058.

Johnson KP, Bush SE, Clayton DH (2005) Correlated evolution of host and parasite body size: Tests of Harrison's rule using birds and lice. *Evolution* **59**, 1744-1753.

Johnson KP, Clayton DH (2004) Untangling coevolutionary history. *Systematic Biology* **53**, 92-94.

Jousselin E, Desdevises Y, Coeur d'acier A (2009) Fine-scale cospeciation between *Brachycaudus* and *Buchnera aphidicola*: bacterial genome helps define species and evolutionary relationships in aphids. *Proceedings of the Royal Society B: Biological Sciences* **276**, 187-196.

Jousselin E, Rasplus JY, Kjellberg F (2003) Convergence and coevolution in a mutualism: Evidence from a molecular phylogeny of *Ficus*. *Evolution* **57**, 1255-1269.

Jousselin E, van Noort S, Berry V, *et al.* (2008) One Fig to Bind Them All: Host Conservatism in a Fig Wasp Community Unraveled by Cospeciation Analyses among Pollinating and Non-pollinating Fig Wasps. *Evolution* **62**, 1777-1797.

Libeskind-Hadas R, Charleston MA (2009) On the Computational Complexity of the Reticulate Cophylogeny Reconstruction Problem. *Journal of Computational Biology* **16**, 105-117.

Longdon B, Brockhurst MA, Russell CA, Welch JJ, Jiggins FM (2014) The Evolution and Genetics of Virus Host Shifts. *Plos Pathogens* **10**, e1004395.

Martínez-Aquino A (2016) Phylogenetic framework for coevolutionary studies: a compass for exploring jungles of tangled trees. *Current Zoology*.

Martinez-Aquino A, Ceccarelli FS, Eguiarte LE, Vazquez-Dominguez E, de Leon GPP (2014) Do the Historical Biogeography and Evolutionary History of the Digenean *Margotrema* spp. across Central Mexico Mirror Those of Their Freshwater Fish Hosts (Goodeinae)? *Plos One* **9**.

McLeish MJ, Crespi BJ, Chapman TW, Schwarz MP (2007) Parallel diversification of Australian gall-thrips on Acacia. *Molecular Phylogenetics and Evolution* **43**, 714-725.

Merkle D, Middendorf M (2005) Reconstruction of the cophylogenetic history of related phylogenetic trees with divergence timing information. *Theory in Biosciences* **123**, 277-299.

Meseguer AS, Lobo JM, Ree R, Beerling DJ, Sanmartin I (2015) Integrating Fossils, Phylogenies, and Niche Models into Biogeography to Reveal Ancient Evolutionary History: The Case of *Hypericum* (Hypericaceae). *Systematic Biology* **64**, 215-232.

Nguyen T-H, Ranwez V, Berry V, Scornavacca C (2013) Support Measures to Estimate the Reliability of Evolutionary Events Predicted by Reconciliation Methods. *Plos One* **8**, e73667.

bioRxiv preprint doi: https://doi.org/10.1101/166215; this version posted July 20, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Nieberding C, Jousselin E, Desdevises Y (2010) The use of co-phylogeographic patterns to predict the nature of host-parasite interactions, and vice versa. In: *The Biogeography of Host-parasite interactions* (eds. Morand S, Krasnov BR), pp. 59-70. Oxford University Press, New York.

Nøjgaard N, Geiß M, Stadler PF et al. (2017) Forbidden time travel : chracterization of Time-consistent tree reconciliation map. The LIPIcs Leibniz International Proceedings in Informatics: "Algorithms in Bioinformatics", 17th International Workshop, WABI 2017, Boston MA, USA.

Page RD (1994a) Maps between trees and cladistic analysis of historical associations among genes, organisms, and areas. *Systematic Biology* **43**, 58-77.

Page RD, Charleston MA (1997) From gene to organismal phylogeny: reconciled trees and the gene tree/species tree problem. *Molecular Phylogenetics and Evolution* **7**, 231-240.

Page RD, Charleston MA (1998) Trees within trees: phylogeny and historical associations. *Trends in Ecology and Evolution* **13**, 356-359.

Page RDM (1990) Temporal congruence and cladistic analysis of biogeography and cospeciation. *Systematic Zoology* **39**, 205-226.

Page RDM (1991) Clocks, clades, and cospeciation: comparing rates of evolution and timing of cospeciation events in host-parasite assemblages. *Systematic Zoology* **40**, 188-198.

Page RDM (1994b) Parallel phylogenies: reconstructing the history of host-parasite assemblages. *Cladistics* **10**, 155-173.

Paterson AM, Banks J (2001) Analytical approaches to measuring cospeciation of host and parasites: through a glass, darkly. *International Journal for Parasitology* **31**, 1012-1022.

Percy DM, Page RD, Cronk QCB (2004) Plant-insect interactions: double-dating associated insect and plant lineages reveals asynchronous radiations. *Systematic Biology* **53**, 120-127.

Ramsden C, Holmes EC, Charleston MA (2009) Hantavirus Evolution in Relation to Its Rodent and Insectivore Hosts: No Evidence for Codivergence. *Molecular Biology and Evolution* **26**, 143-153.

Ree RH, Sanmartin I (2009) Prospects and challenges for parametric models in historical biogeographical inference. *Journal of Biogeography* **36**, 1211-1220.

Ree RH, Smith SA (2008) Maximum likelihood inference of geographic range evolution by dispersal, local extinction, and cladogenesis. *Systematic Biology* **57**, 4-14.

Revell LJ (2012) phytools: an R package for phylogenetic comparative biology (and other things). *Methods in Ecology and Evolution* **3**, 217-223.

Ronquist F, Sanmartin I, Futuyma DJ, Shaffer HB, Simberloff D (2011) Phylogenetic methods in biogeography. *Annual Review of Ecology, Evolution, and Systematics* **42**, 441-464.

Rosenblueth M, Sayavedra L, Samano-Sanchez H, Roth A, Martinez-Romero E (2012) Evolutionary relationships of flavobacterial and enterobacterial endosymbionts with their scale insect hosts (Hemiptera: Coccoidea). *Journal of Evolutionary Biology* **25**, 2357-2368.

Singh G, Dal Grande F, Divakar PK, *et al.* (2017) Fungal–algal association patterns in lichen symbiosis linked to macroclimate. *New Phytologist* **214**, 317-329.

Sorci G, Skarstein F, Morand S, Hugot JP (2003) Correlated evolution between host immunity and parasite life histories in primates and oxyurid parasites. *Proceedings of the Royal Society B-Biological Sciences* **270**, 2481-2484.

Stevens J (2004) Computational aspects of host-parasite phylogenies. *Brief Bioinform* **5**, 339-349.

Stolzer M, Lai H, Xu ML, *et al.* (2012) Inferring duplications, losses, transfers and incomplete lineage sorting with nonbinary species trees. *Bioinformatics* **28**, I409-I415.

Wilson JS, Forister ML, Dyer LA, *et al.* (2012) Host conservatism, host shifts and diversification across three trophic levels in two Neotropical forests. *Journal of Evolutionary Biology* **25**, 532-546.

bioRxiv preprint doi: https://doi.org/10.1101/166215; this version posted July 20, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Zientz E, Dandekar T, Gross R (2004) Metabolic interdependence of obligate intracellular bacteria and their insect hosts. *Microbiology and Molecular Biology Reviews* **68**, 745-770.

bioRxiv preprint doi: https://doi.org/10.1101/166215; this version posted July 20, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

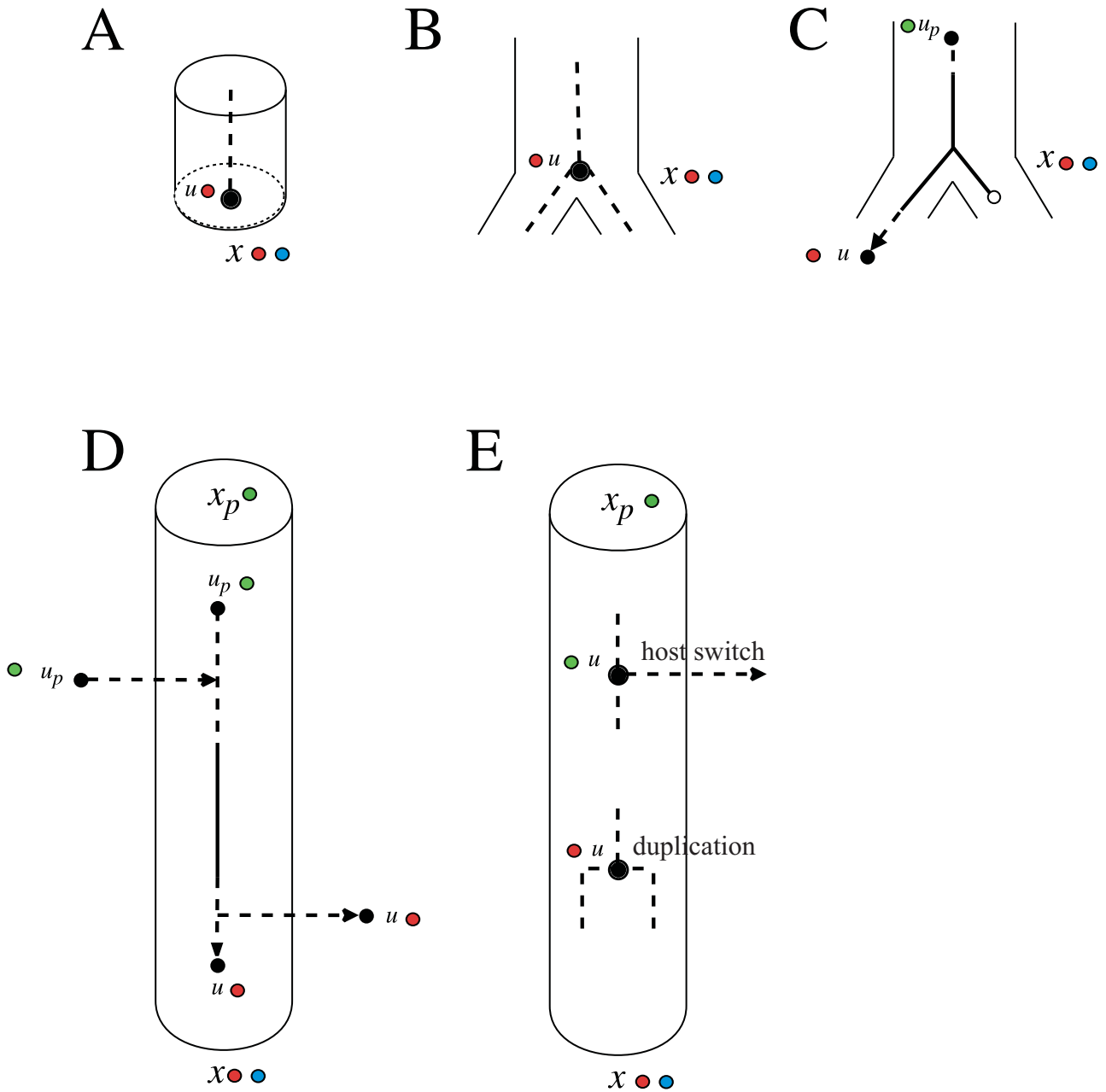
Figure 1: Description of how geographical constraints are handled by the *Mowgli* software, black dots ● represent nodes of the parasite tree, plain and dash black lines represent branches of the parasite tree, cylinders represent branches of the host tree, ellipses closing the cylinders represent nodes of the host tree; coloured dots ●,●,● represent different geographic areas of the parasite or the host. A) A parasite leaf $u$ can be mapped to a leaf $x$ of the host tree if the areas of the host contains all areas of the parasite. B) *Mowgli* accepts that a parasite node $u$ cospeciates with a host at a node $x$, if the two nodes share an area. C) To map a parasite node $u$ inside a branch $(x_p,x)$ – to represent the origin of a host switch, or a duplication of the parasite, *Mowgli* requires that $u$ lived in an area of $x_p$ or $x$ (if $u$ lived in several areas, the constraint holds for each of these areas); D) To map a branch $(u_p,u)$ of the parasite tree inside a branch $(x_p,x)$ of the host tree, *Mowgli* requires that $u_p$ or $u$ lives in an area also assigned to node $x_p$ or to node $x$. This is required whether $u_p$ is located at $x_p$, between $x$ and $x_p$, or is in another branch due to a host switch and similarly for $u$; E) A parasite on a branch $(u_p,u)$ can go through a speciation $x$ of their hosts whenever $area(u_p,u)$ inhabits some area of $x$.

Figure 2: Results of the reconciliations obtained on a simulated data set with default cost settings: A) without enforcing geographic constraints (7 coSpeciation events, 1 Tranfer, 1 Loss), B) when enforcing geographic constraints (6 S, 2 T, 1 L). The big pie charts correspond to the host ancestral geographic areas while small pie charts correspond to the symbiont ancestral geographic areas. The leaves of the species tree are also coloured according to their contemporary geographic distribution. Annotations files given for the host tree and the symbiont tree specified a single most likely area at each node.

Figure 3: Results of the reconciliations obtained on the fig/fig wasp data set with default cost settings: A) without enforcing geographic constraints (events: 17 S, 6 T, 1 L), b) when enforcing geographic constraints (events: 16 S, 7 T, 1 L). Big pie charts correspond to the *Ficus* ancestral geographic areas, small pie charts correspond to the pollinator ancestral geographic areas. The leaves of the phylogenetic trees are coloured according to contemporary geographic distribution of the associates. Annotation files given for the host tree and the symbiont tree indicate the likelihood of each geographic area at each node (i.e. each section of the pies correspond to the likelihood of different states; here it does not indicate a multistate character, i.e. the proportion of the population in each area). Annotation file for the specie tree also included *Ficus* taxonomic subdivisions and are reported in the right inside of the figure.

Supplementary Material 1: *Mowgli* manual

Supplementary Material 2: Description of the pipeline to generate trees and annotation files that can be taken as inputs for both *Mowgli* and *Sylvx*.


A
u
x
B
u
x
C
$u_p$
x
u
D
$x_p$
$u_p$
$u_p$
u
u
x
E
$x_p$
u
host switch
u
duplication
x

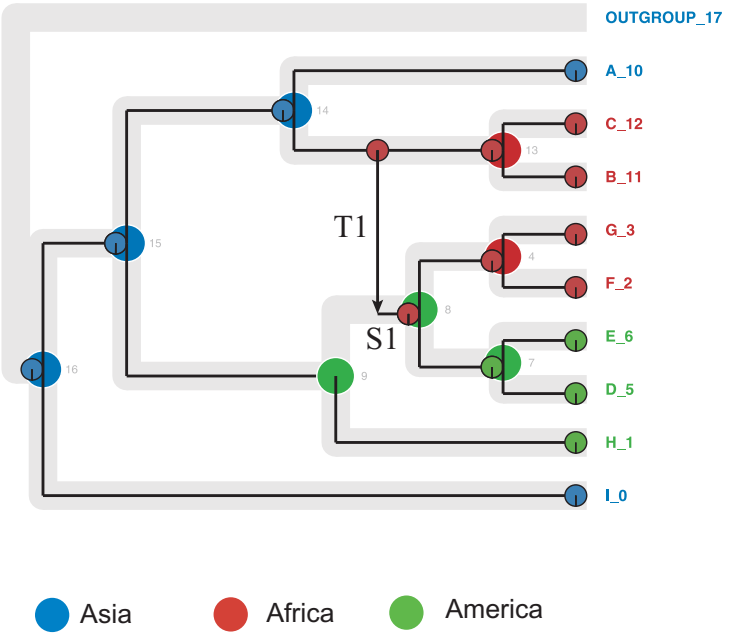



Figure 2A

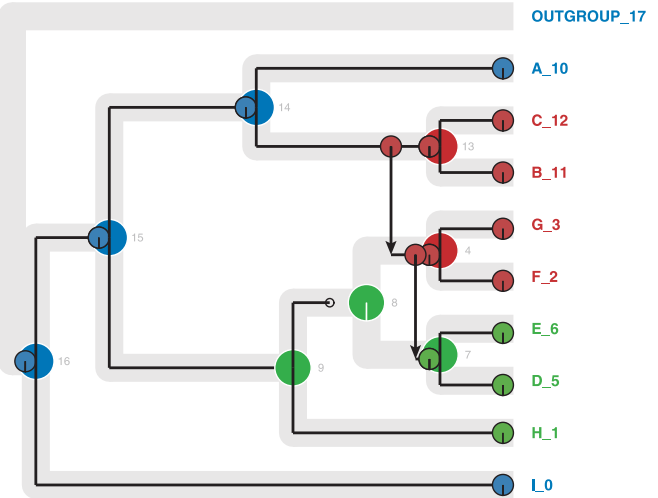



Figure 2B

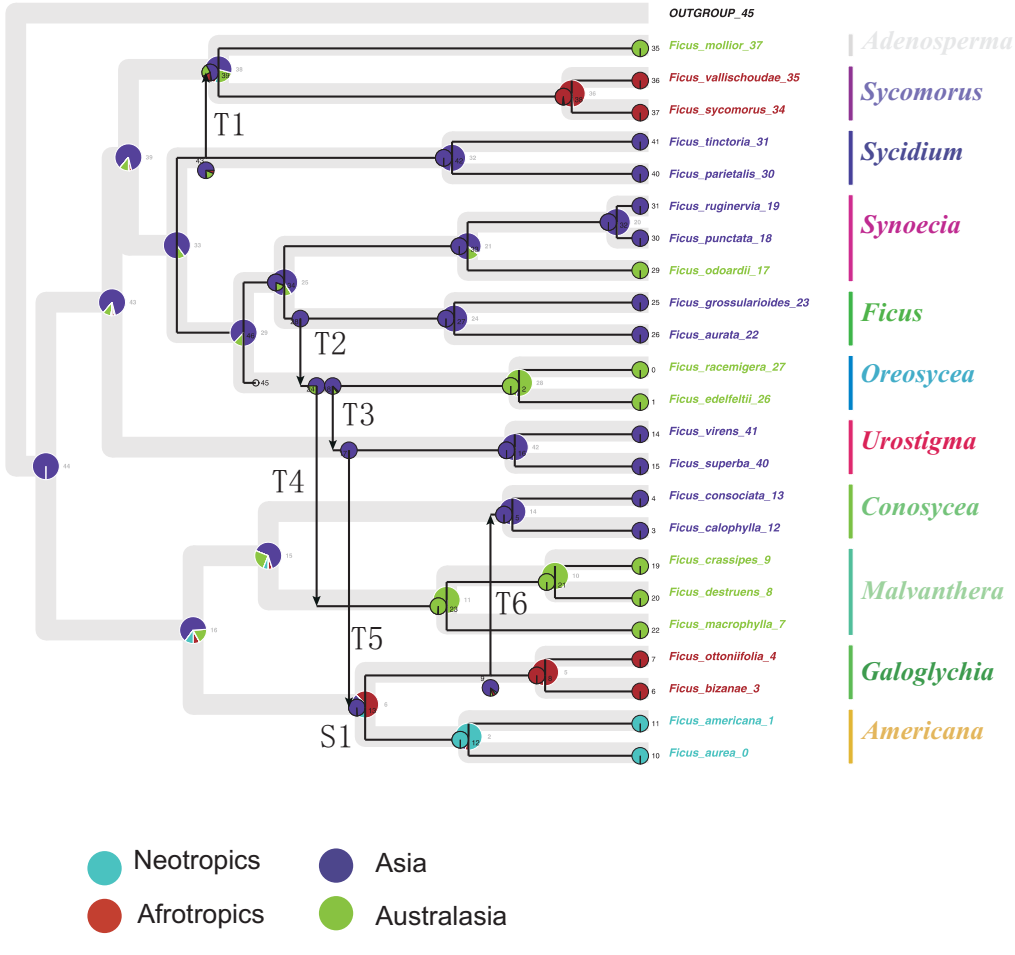



Figure 3A

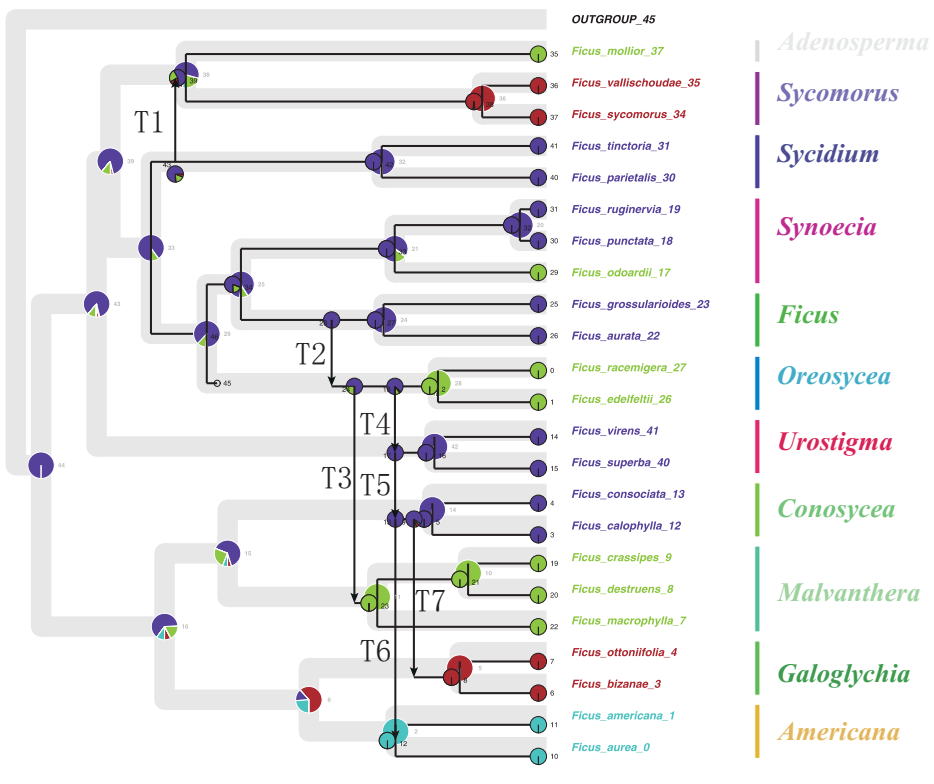



Figure 3B