Version dated: February 1, 2021

Evolution of butterfly-plant networks over time, as revealed by Bayesian inference of host repertoire

Mariana P Braga[1,2], Niklas Janz[1], Sören Nylin[1], Fredrik Ronquist[3], and Michael J Landis[2]

[1] *Department of Zoology, Stockholm University, Stockholm, SE-10691, Sweden;*

[2] *Department of Biology, Washington University in St. Louis, St. Louis, MO, 63130, USA;*

[3] *Department of Bioinformatics and Genetics, Swedish Museum of Natural History, Stockholm, SE-10405, Sweden*

Corresponding author: Mariana P Braga, Department of Biology, Washington University in St. Louis, St. Louis, MO, 63130, USA; E-mail: mariana@wustl.edu

Short running title: Evolution of butterfly-plant networks

Keywords: coevolution, ecological networks, herbivorous insects, host-parasite interactions, modularity, nestedness, phylogenetics

Statement of authorship: MPB, NJ and SN designed the basis for the biological study. SN collected the data. MPB and MJL designed the statistical analyses. MPB analyzed the data, generated the figures, and wrote the first draft of the manuscript. All authors contributed to the final draft.

Data accessibility statement: No new data was used.

Abstract.—The study of herbivorous insects underpins much of the theory that concerns the evolution of species interactions. In particular, Pieridae butterflies and their host plants have served as a model system for studying evolutionary arms-races. To learn more about how the two lineages co-evolved over time, we reconstructed ecological networks and network properties using a phylogenetic model of host-repertoire evolution. In tempo and mode, host-repertoire evolution in Pieridae is slower and more conservative when compared to similar model-based estimates previously obtained for another butterfly clade, Nymphalini. Our study provides detailed insights into how host shifts, host range expansions, and recolonizations of ancestral hosts have shaped the Pieridae-angiosperm network through a phase transition from a disconnected to a connected network. Our results demonstrate the power of combining network analysis with Bayesian inference of host repertoire evolution in understanding how complex species interactions change over time.

For more than a century, evolutionary ecologists have studied the coevolutionary dynamics that result from intimate ecological interactions among species (Darwin 1877; Ehrlich and Raven 1964; Forister et al. 2012; Vienne et al. 2013). Butterflies and their host-plants are among the most studied of such systems; hence, various aspects of butterfly-plant coevolution have inspired theoretical frameworks that elucidate how interactions evolve in nature (Janz 2011). Two prominent and opposing conceptual hypotheses that explain host-associated diversification derive from empirical work in butterfly-plant systems: the escape-and-radiate hypothesis (Ehrlich and Raven 1964) and the oscillation hypothesis (Janz and Nylin 2008). The escape-and-radiate model predicts that butterflies and host-plant lineages have diversified in bursts as a result of the competitive release that follows the colonization of a brand new host. Thus, butterfly diversification would often be associated with complete host shifts, i.e. new hosts replace ancestral hosts (Fordyce 2010). In contrast, the oscillation hypothesis assumes that butterflies colonizing new hosts may retain the ability to use the ancestral host or hosts. According to this hypothesis, at any point in time, butterflies can use more hosts than they actually feed on in nature. Defining the set of hosts used by a parasite as its *host repertoire*, the oscillation hypothesis allows for a lineage to possess a realized host repertoire (analogous to realized niche) that is a subset of its fundamental host repertoire (Nylin et al. 2018). And while the fundamental host repertoire is phylogenetically conserved, the realized repertoire is less stable over evolutionary time, resulting in oscillations in the number of hosts used (i.e. host range). These oscillations in the realized host repertoire are thought to spur diversification.

In recent years, there has been a clear trend from a somewhat simplified escape-and-radiate hypothesis to more complex models of coevolution, shifting from one-to-one associations to diffuse coevolution, from tight to more loosely connected evolutionary trajectories, and from interacting species-pairs to networks of interacting species (Guimarães et al. 2011). In line with this trend, Braga et al. (2018) recently suggested that coevolving host-parasite associations in general may be characterized by processes fitting *both* the escape-and-radiate *and* the oscillation hypotheses. This was based on network and phylogenetic analyses of two butterfly families, Nymphalidae and Pieridae. Specifically, it was shown that these alternative diversification scenarios generate different structural patterns in the networks that characterize extant interactions between butterfly families and their host plants. The escape-and-radiate scenario generates network modularity (Olesen et al. 2007; Braga et al. 2018), where each module is composed of a given host taxon and closely-related butterflies that diversified after shifting to the given host. Conversely, oscillations in host range produces a specialist-generalist gradient in both trophic levels, where specialized butterflies use a subset of the host plants used by closely-related generalists. Thus, the oscillation hypothesis generates network nestedness (Bascompte et al. 2003; Braga et al. 2018).

While network analysis is a powerful tool for classifying interaction patterns predicted by alternative coevolutionary hypotheses, other methods are needed to directly identify what mechanisms generated the observed interaction patterns. In the case of host-parasite coevolution, methodological and computational constraints have so far hindered the explicit modeling of host repertoire evolution without strongly reducing the inherent complexity of the system. These constraints have been relaxed by recent developments concerning phylogenetic Bayesian inferences of evolution of discrete traits (Landis et al. 2013), allowing Braga et al. (2020) to develop a Bayesian method specifically for inferring the evolution of host repertoires. Unlike previous approaches used for reconstruction of past ecological interactions (Ferrer-Paris et al. 2013; Tsang et al. 2014; Jurado-Rivera and Petitpierre 2015; Navaud et al. 2018, e.g.), this method explicitly accounts for the possibility that a parasite may have multiple hosts and that interactions with different hosts evolve interdependently. This feature allows us to uncover the entire distribution of ancestral host ranges at any given point in time, including the "long tail" of generalists (Forister et al. 2015; Nylin et al. 2018), as well as temporal changes in host range.

In this paper, we perform a Bayesian analysis of host repertoire evolution in Pieridae butterflies using the method developed by Braga et al. (2020). Pieridae is an interesting system for this comparison because the diversification of the group was first explained solely by the escape-and-radiate hypothesis (Fordyce 2010; Edger et al. 2015), but more recent evidence suggests that these butterflies also underwent oscillations in host range (Braga et al. 2018). We represent ancestral host repertoires in two different ways, with (1) a *traditional representation* that only considers ancestral pairs of plant-butterfly interactions that exceed a specified probability threshold; and (2) a new *probabilistic representation* that makes fuller use of the posterior distribution of ancestral states. Reconstructing ancestral networks in these ways, we show how host shifts, host range expansions, and recolonizations of ancestral hosts have shaped the Pieridae-angiosperm network.

Methods

Pierid Butterflies and Angiosperm Hosts

We reconstructed historical interactions between Pieridae butterflies and their host plants using a Bayesian phylogenetic approach (Braga et al. 2020). Interaction data between butterfly genera and plant families were gathered from the literature (see Supplementary Information). We used previously published time-calibrated phylogenies for 66 Pieridae genera (Edger et al. 2015, Fig. S1) and angiosperm families (Edger et al. 2015; Magallón et al. 2015). We pruned the host angiosperm phylogeny, keeping all 33 angiosperm families that are known to be hosts of pierid butterflies, then collapsing increasingly ancestral nodes until only 50 terminal branches were left. This increased the chance that all ancestral angiosperm lineages that might have been used as hosts in the past were included in the analysis, while keeping the analysis computationally tractable.

Model of Host Repertoire Evolution

We modeled host repertoire evolution across Pieridae as a continuous-time Markov chain (CTMC) that describes gain and loss of individual hosts. In the model, the host repertoire of a given parasite is represented as a binary vector of length 50, where each element within the vector describes the interaction between the parasite and a given host plant family. Hosts (i.e. vector elements) can assume one of two states: 0 (non-host) or 1 (host). We assumed that each parasite must have at least one host at any given time. Thus, the state space (i.e. the number of state combinations that a host repertoire can assume) for this model includes $2^{50} - 1 \approx 1.13 \times 10^{15}$ unique repertoires. We used a Bayesian data augmentation approach (Robinson et al. 2003; Landis et al. 2013; Quintero and Landis 2019; Braga et al. 2020) to sample evolutionary character histories under this large state space. We did not consider uncertainty in the host or parasite phylogenies to facilitate the inference of model parameters under our data augmentation method. Note that the original model described in Braga et al. (2020) included three states (non-host, potential host and actual host), but because our data set does not report information on potential hosts, model performance was poor under the 3-state model.

In a 2-state model, two types of events can change the host repertoire: host gain ($0 \rightarrow 1$) occurs with the rate λ_{01}, and host loss ($1 \rightarrow 0$) occurs with the rate λ_{10}. These rates allow us to compute the probability of any given coevolutionary history based on the instantaneous-rate matrix that defines the CTMC. This matrix is constructed such that only one host in the repertoire is allowed to change in state at a time. Relative gain and loss rates are constrained between 0 and 1, which are multiplied by global rate scaling parameter, μ, to produce absolute rates of gain and loss.

Our model allowed for phylogenetic relatedness among hosts to influence how easily a butterfly might expand its host repertoire to include a new host species. Specifically, host gain rates were further multiplied by a phylogenetic-distance rate modifier, which is defined as $e^{-\beta d_{ij}}$, where d_{ij} measures the relative phylogenetic distance between the currently parasitized host i and the newly gained host j and β rescales the magnitude of d_{ij} (see Braga et al. (2020) for details). That is, if $\beta > 0$, parasites prefer to colonize new hosts that are phylogenetically similar to currently parasitized hosts. If $\beta = 0$, the gain rates are not affected by the host tree. Following Braga et al. (2020), we measured phylogenetic distance between host lineages in two different ways: (1) using what we call the *anagenetic tree*, where distances reflect time-calibrated divergence times among hosts, and (2) using a modified *cladogenetic* tree, where all host branch lengths were set to 1, approximating phylogenetic distances that are proportional to the number of older (i.e. family-level) cladogenetic events that separate two taxa.

Summarizing ecological interactions through time

Ancestral interactions were estimated by regularly sampling histories of host repertoire evolution during the Bayesian Markov chain Monte Carlo (MCMC) analysis (described below), meaning interaction histories were sampled alongside the joint posterior distribution of model parameters. We first summarized the sampled histories using a traditional representation of ancestral states (e.g. Nylin et al. 2014). To do so, we calculated marginal posterior probabilities for interactions between each host plant and each internal node in the Pieridae phylogeny, based on the frequency with which state 1 was sampled during MCMC for the given host at the given internal node. Interactions with marginal posterior probability of > 0.9 were treated as 'true' occurrences, with all other interactions being treated as 'false'. This traditional approach has three important limitations: (1) it only considers states at internal nodes, ignoring what happens along the branches of the butterfly tree; (2) by focusing on the highest-probability butterfly-plant interactions, it filters out ancestral interactions with middling probabilities; and (3) it is blind to how joint sets of interactions might have evolved together, as it is based on marginal probabilities of pairwise host-parasite interactions. We discuss each of these three items in detail below and explore new ways to summarize host repertoire evolution.

Viewing ecological histories as networks.— To resolve the first limitation, we reconstructed the host repertoires of all extant butterfly lineages at eight time slices, from 80 Ma to 10 Ma. Thus, instead of reconstructing the host repertoire of internal nodes in the butterfly tree, we reconstructed ancestral Pieridae-host plant networks at different ages throughout the diversification of Pieridae. This way we capture more information about the system at specific time slices and, most importantly, we can quantify changes in network structure over time, as contrasting hypotheses of eco-evolutionary dynamics are expected to generate similarly contrasting structures in ecological networks (Braga et al. 2018).

Summarizing posterior distributions of networks with point estimates.— In order to investigate how much information is lost when we only consider the highest-posterior interactions (limitation 2), we compared three kinds of summary networks for each time slice: one binary (presence/absence) and two weighted (quantitative) networks. In the binary networks, only interactions with at least 0.9 marginal posterior probability were considered to be present, while all other interactions were considered absent. In the weighted networks, plant-butterfly interactions were assigned weights equal to their posterior probabilities, but interactions with probabilities under a threshold were assigned the weight of 0 (absent). The two weighted networks differed in this threshold: one excluded only interactions with very low probability (< 0.1), while the other excluded all interactions with probability < 0.5.

To characterize the structure of extant and ancestral (inferred) networks, we used two standard metrics: modularity and nestedness. Modularity measures the degree to which the network is divided in sets of nodes with high internal connectivity and low external connectivity (Olesen et al. 2007), which, in our case, identify plants and butterflies that interact more with each other than with other taxa in the network.

Nestedness measures the degree to which the partners of poorly connected nodes form a subset of partners of highly connected nodes (Bascompte et al. 2003). To measure modularity, we used the Beckett (2016) algorithm, which works for both binary and weighted networks, as implemented in the function *computeModules* from the package *bipartite* (Dormann et al. 2008) in R version 3.6.2 (R Core Team 2019). This algorithm assigns plants and butterflies to modules and computes the modularity index, Q. To measure nestedness, we computed the nestedness metric based on overlap and decreasing fill, NODF (Almeida-Neto et al. 2008; Almeida-Neto and Ulrich 2011), as implemented for binary and weighted networks in the function *networklevel* also in the R package *bipartite*. To test when Q and NODF scores were significant, we computed standardized Z-scores that can be compared across networks of different sizes and complexities using the R package *vegan* (Oksanen et al. 2019) (details in Supplement).

We emphasize that our method does not estimate the first ages of origin for modularity or nestedness, but rather it estimates the first ages for which these network features can be detected. The difficulty of detecting topological features increases with geological time, in part because phylogenetic reconstructions become less certain as time increases, but also because time-calibrated phylogenies of extant organisms are represented by fewer lineages as time rewinds. For these reasons, our statistical power to infer the age of origin for the oldest ecological interactions is limited. When interpreting our results, we focus on the ages that we first detect modularity and nestedness among surviving lineages, where first-detection times are assumed to follow origination times for these network features.

Finally, we compared these estimates to the posterior distribution of Z-scores and statistical significance by calculating Q and NODF for 100 samples from the MCMC and 100 null networks for each sample. This comparison was done to test if the three summary networks accurately represent the posterior distribution of ancestral networks in terms of modularity and nestedness.

Posterior support for ecological modules.— Defining eco-evolutionary groupings as modules allows us to visualize when those modules first appeared and how they changed over time. But in contrast to indices that are calculated for the entire network, the information about module configuration is not easily summarized into a posterior distribution. To circumvent this problem, we used one of the summary networks (probability threshold = 0.5) to characterize the modules across time, and then validate these modules with the posterior probability that two nodes belong to the same module (see below). This weighted network includes many more interactions than the binary network, while preventing very improbable interactions from implying spurious modules.

After identifying the modules for the summary network at each age, we assigned fixed identities to modules based on the host plant(s) with most interactions within the module. We then validated the modules in the eight summary networks (one for each time slice) using 100 networks sampled during MCMC, i.e. snapshots of character histories sampled during MCMC. We first decomposed each network of ancestral interactions sampled during MCMC into modules, and then calculated the frequency with which each pair of nodes in the summary network (butterflies and plants) were assigned to the same module across samples; that is, the posterior probability that two nodes belong to the same module.

Bayesian inference method

Bayesian MCMC was used to estimate the joint posterior distribution of the parameters in the model of host repertoire evolution described above. All analyses were performed in RevBayes (Höhna et al. 2016) using the inference strategy described in Braga et al. (2020). We ran four independent MCMC analyses (two with the anagenetic distance and two with the cladogenetic distance between hosts), each set to run for 2×10^5 cycles, sampling parameters and node histories every 50 cycles, and discarding the first 2×10^4 as burnin. Prior distributions were $\mu \sim Exponential(10)$, $\beta \sim Exponential(1)$, and $\lambda \sim Dirichlet(1, 1)$, where elements of λ follow the marginal distribution,

$\lambda_{i,j} \sim \text{Beta}(1,1)$. To verify that MCMC analyses converged to the same posterior distribution, we applied the Gelman diagnostic (Gelman and Rubin 1992) as implemented in the R package *coda* (Plummer et al. 2006). Results from a single MCMC analysis are presented.

To test whether the phylogenetic relatedness between hosts had an important effect on the host gain rate, we computed the Bayes factor using the Savage-Dickey ratio (Verdinelli and Wasserman 1995; Suchard et al. 2001; Marin and Robert 2010), defined as the ratio between the prior and posterior probability that $\beta = 0$. We then followed the guidelines of Jeffreys (1961) to interpret the resulting Bayes factor, as also done in Braga et al. (2020).

Code availability

Our RevBayes and R scripts are available at `https://github.com/maribraga/pieridae_hostrep`. Our R scripts additionally depend on a suite of generalized R tools we designed for analyzing ancestral ecological network structures `https://github.com/maribraga/evolnets`.

Results

Posterior estimates of Pieridae host repertoire evolution were partially sensitive to whether we measured distances between host lineages in units of geological time or in units of major cladogenetic events (Fig. 1). When measuring anagenetic distances between host lineages, posterior mean (95% highest posterior density; HPD95) estimates were: global rate scaling factor for host repertoire evolution $\mu = 0.02$ (0.015 - 0.026), phylogenetic-distance power $\beta = 2.1$ (0.017 - 3.82), relative host gain rate $\lambda_{01} = 0.035$ (0.022 - 0.047), and relative host loss rate $\lambda_{10} = 0.965$ (0.95 - 0.98). Mean estimates were similar when distances between hosts were measured in units of cladogenetic events: $\mu = 0.019$ (0.014 - 0.024), $\beta = 1.48$ (1.02 - 1.97), $\lambda_{01} = 0.027$ (0.017 - 0.036), and $\lambda_{10} = 0.97$ (0.96 - 0.98). An important difference between the two inferences is that the HPD95 for β under cladogenetic distance excludes $\beta = 0$, whereas β estimated under anagenetic distance assigns a non-zero probability (≈ 0.1) to $\beta = 0$. The decisive support for $\beta > 0$ when using cladogenetic distance led us to focus primarily on this reconstruction throughout the main text (Fig. S2 for results with anagenetic distance). Because rate parameters can be difficult to interpret, we also calculated the average number of proposed events across MCMC samples, which was 148, being 75 host gains and 73 host losses throughout the diversification of Pieridae.

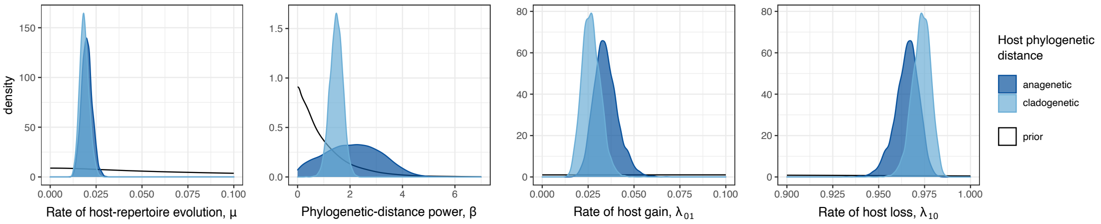

Figure 1: Estimated marginal posterior densities for parameters in the host-repertoire evolution model using two different representations of the phylogenetic distance between host-plant families: anagenetic (time) or cladogenetic (number of branches).

With the traditional approach for ancestral state reconstruction, that is, focusing on the highest-probability hosts at internal nodes of the butterfly tree, we can describe the general patterns of evolution of interactions between Pieridae butterflies and their host plants (Fig. 2). We can confidently say that: (1) the most recent common ancestor (MRCA) of all Pieridae butterflies used a Fabaceae host, (2) all ancestral Coliadinae and Dismorphiinae used Fabaceae, (3) the MRCA of, and early Pierinae (Pierina + Aporina + Anthocharidini + Teracolini) used a Capparaceae host, (4) Brassicaceae and Loranthaceae were used by one Anthocharidini clade each, (5) early Aporina used both Loranthaceae and Santalaceae, and (6) the MRCA of, and early Pierina used three host families: Capparaceae, Brassicaceae and Tropaeolaceae.

While the traditional ancestral state reconstruction described above tells us relevant and important pieces of the history of interaction between pierid butterflies and

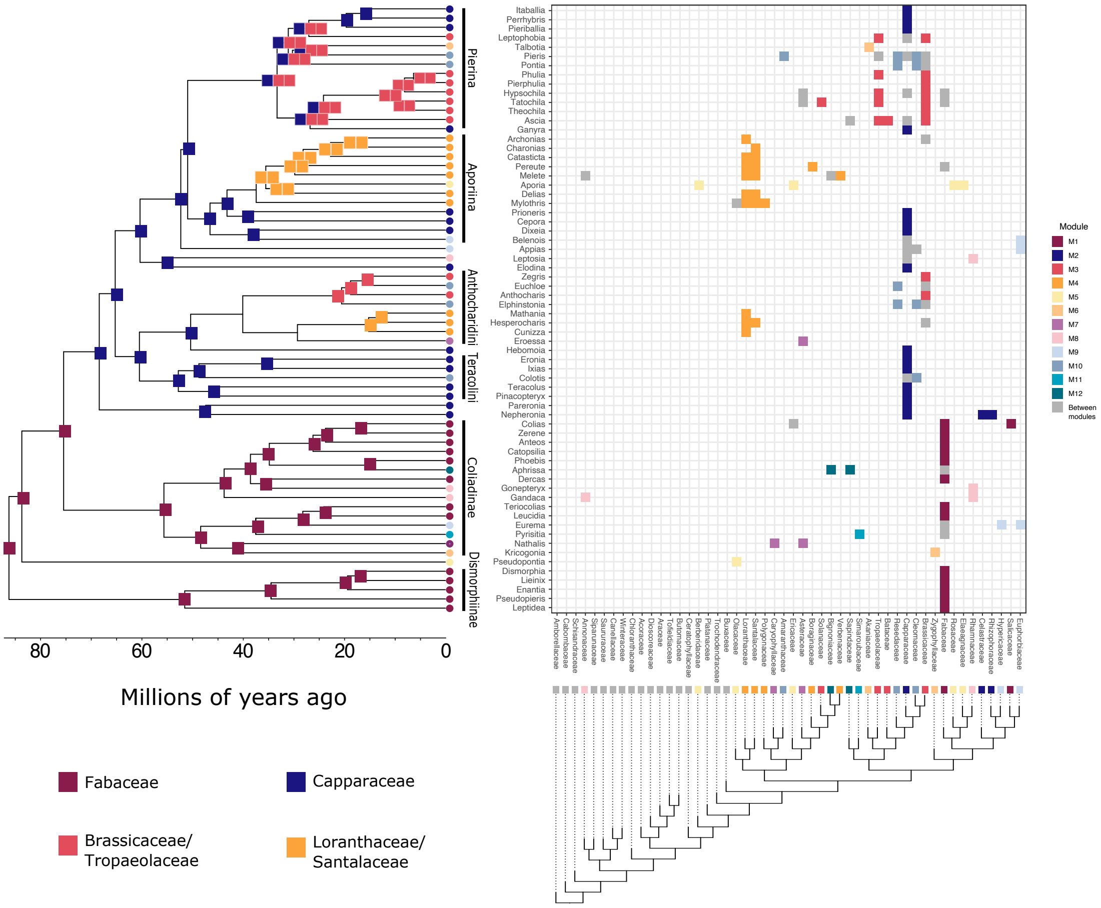

Figure 2: Ancestral state reconstruction showing interactions with marginal posterior probability ≥ 0.9. The model reconstructs how host repertoire evolved along the Pieridae phylogeny (left), based on the observed butterfly-plant interactions (top-right), and the cladogenetic distance between hosts (measured as the number of branches separating the hosts; bottom-right). The color of the symbols at the tips of both trees shows to which module the butterfly genus or plant family belongs (modules from the present-day network). Each square at internal nodes of the butterfly tree represents one plant family and is colored by the module to which the plant belongs. The matrix in the top-right shows the observed interactions between butterflies (rows) and plant families (columns). Rows and columns are ordered to match the phylogenetic trees. Interactions between butterflies and plants within modules are colored by module, whereas interactions between modules are in grey.

their host plants, it represents only a part of the posterior distribution of ancestral interactions. The remaining analyses provide more detailed information on the inferred host repertoire evolution. Instead of reconstructing ancestral host repertoires at internal nodes of the butterfly tree, we looked at eight time slices along the diversification of Pieridae: every 10 Myr, from 80 Ma to 10 Ma.

According to the posterior distribution of Q and NODF based on networks sampled from the MCMC, modularity and nestedness were first detectable 30 Ma (Fig. 3; for raw Q and NODF values see Fig. S3). But while the support for modularity has not changed much in the last 30 Myr, support for nestedness has increased linearly in the past 50 Myr. Overall, the summary networks have overestimated the presence of modularity, and only the weighted summary network with the 0.1-threshold correctly estimated that significant modularity appeared 30 Ma (Fig. 3 upper panel). On the other hand, the summary networks underestimated the existence of nestedness in ancestral networks (Fig. 3 lower panel), with several networks being significantly less nested than expected by chance, especially with the binary networks.

The present-day Pieridae-angiosperm network is characterized by both higher modularity (M = 0.64, p $\leq$ 0.001, Z-score = 3.62) and nestedness (NODF = 14.8, p $\leq$ 0.001, Z-score = 11.21) than expected by chance. Most of the butterfly lineages within Dismorphiinae and Coliadinae are associated with Fabaceae hosts (module M1), while Pierinae butterflies use many other host families (Fig. 2), the most common being Capparaceae (module M2), Brassicaceae + Tropaeolaceae (M3) and Loranthaceae + Santalaceae (M4). Interestingly, some Pierinae butterflies recolonized Fabaceae and others colonized new hosts while keeping the old host in their repertoire, resulting in among-module interactions that connected the whole network and produced signal for nestedness. By exploring the posterior distribution of ancestral interactions, we were able to characterize how this network was assembled throughout the diversification of Pieridae butterflies, as described below.

At 80 Ma, M1 and M2 are already recognized as separate modules based on

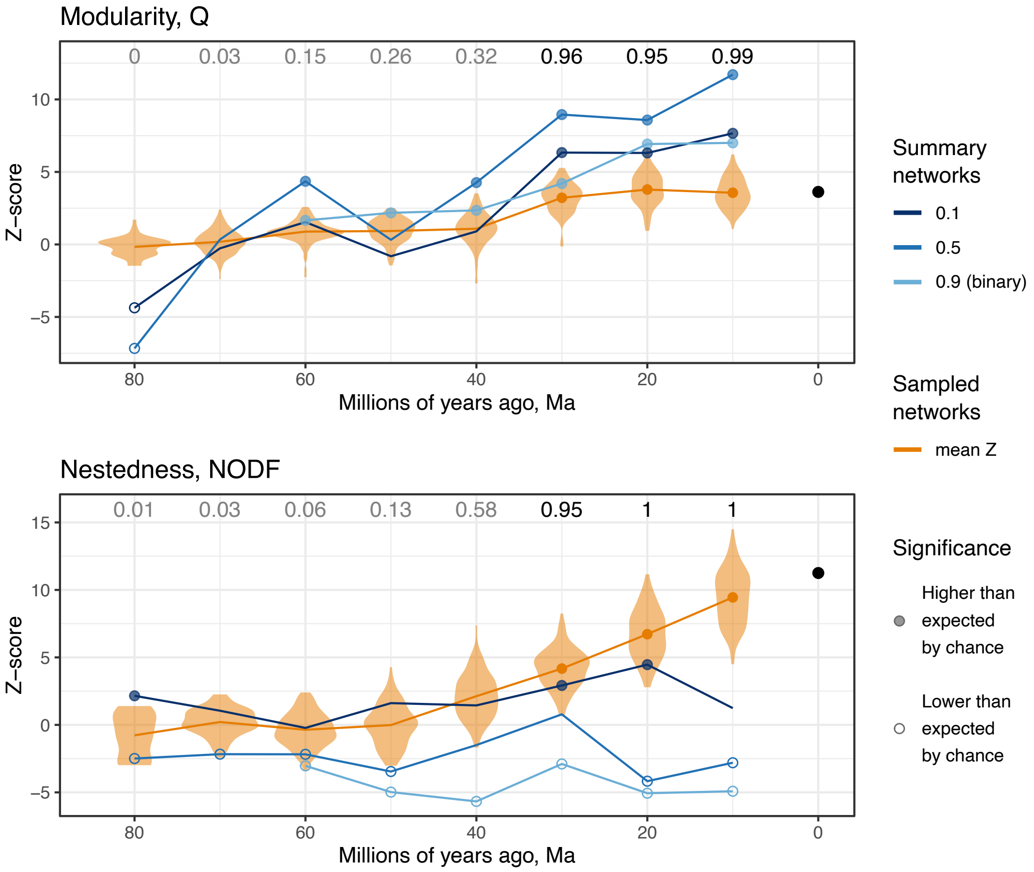

Figure 3: Structure of the Pieridae-angiosperm network over time. Z-scores for (a) modularity and (b) nestedness for summary (blues) and sampled networks (orange) from 80 Ma to 10 Ma, and for the observed present-day network (black circles). Each orange violin represents the distribution of Z-scores for sampled networks at each time slice and the orange line shows the mean Z-score. Indices (Q or NODF) higher than expected under the null model are shown with a filled circle, while indices lower than expected are shown with an empty circle. Numbers at the top of each graph show the proportion of sampled networks that were significantly modular or nested. In all cases, the significance level $\alpha = 0.05$.

marginal posterior probabilities of interactions (weighted summary network with probability threshold of 0.5, Fig. 4a). However, these modules were not validated by joint probabilities of two nodes being assigned to the same module across MCMC samples. Nodes that were assigned to different modules in the summary network were placed in the same module in many MCMC samples (grey cells in Fig. 5a). For example, Fabaceae and Capparaceae were assigned to the same module in 75 of the 100 MCMC samples, suggesting that at 80 Ma there was only one module, including both Fabaceae and Capparaceae. Then, between the Late Cretaceous (represented by 70 Ma) and the Middle Eocene (represented by 50 Ma), Pieridae formed two distinct sets of ecological relationships with their angiosperm host plants: one set of pierid lineages feeding primarily on Fabaceae (M1), and a second set that first diversified between 70 and 60 Ma feeding primarily on Capparaceae (M2; Fig. 4b–d). During that time, as more butterfly lineages accumulated within the Fabaceae and Capparaceae modules, the only plant lineages in the two modules were Fabaceae and Capparaceae themselves. Besides the two main modules, a small module was formed around 50 Ma including the ancestor of *Pseudopontia* and Olacaceae.

Between 40 and 30 million years ago, coinciding with the onset of the Oligocene, two new modules emerged, one composed of butterflies that shared interactions with Brassicaceae and/or Tropaeolaceae (M3), and another of lineages that interacted with Loranthaceae and/or Santalaceae (M4; Figs. 4e–f and 5e–f). At the end of this period, M1 had expanded due to butterfly diversification and colonization of new host plants; M2 and M3 expanded and became more connected, as the first Pierina diversified while using both the ancestral host Capparaceae and the more recent host Brassicaceae. Entering into the Miocene at 20 Ma and 10 Ma, as the sizes of modules grew, so did the number of interactions between modules. Modules M6, M7 and M8 appeared for the first time, and the remaining modules, M7–M12, appeared between 10 Ma and the present.

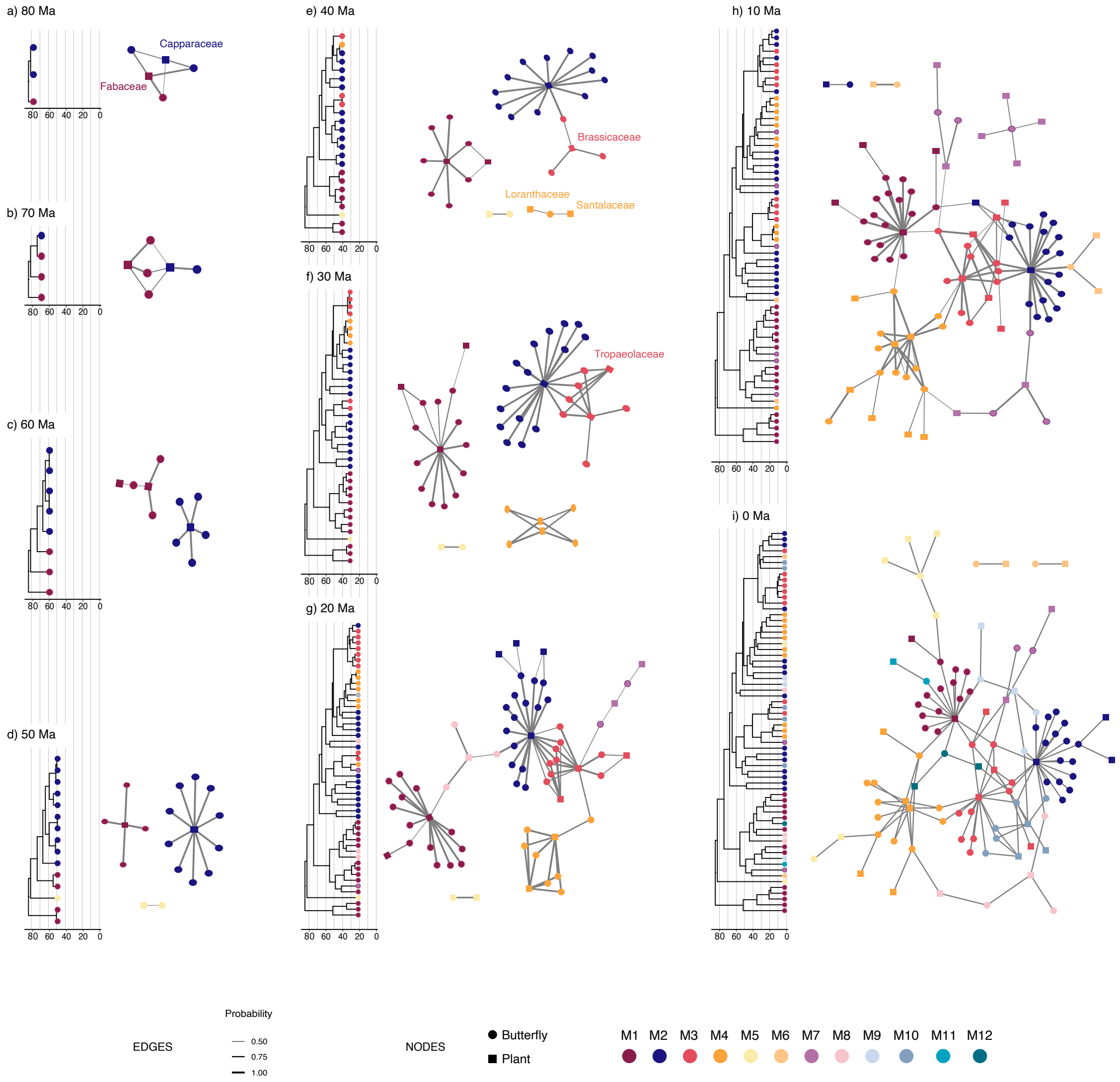

Figure 4: Evolution of the Pieridae-angiosperm network across nine time slices from 80 Ma to the present. Each panel (a-i) shows the butterfly lineages extant at a time slice (left) and the estimated network (right) of interactions with at least 0.5 posterior probability. Edge width is proportional to interaction probability. Nodes of the network and tips of the trees are colored by module, which were identified for each network separately and then matched across networks using the main host plant as reference. Names of the six main host-plant families are shown at the time when they where first colonized by Pieridae.

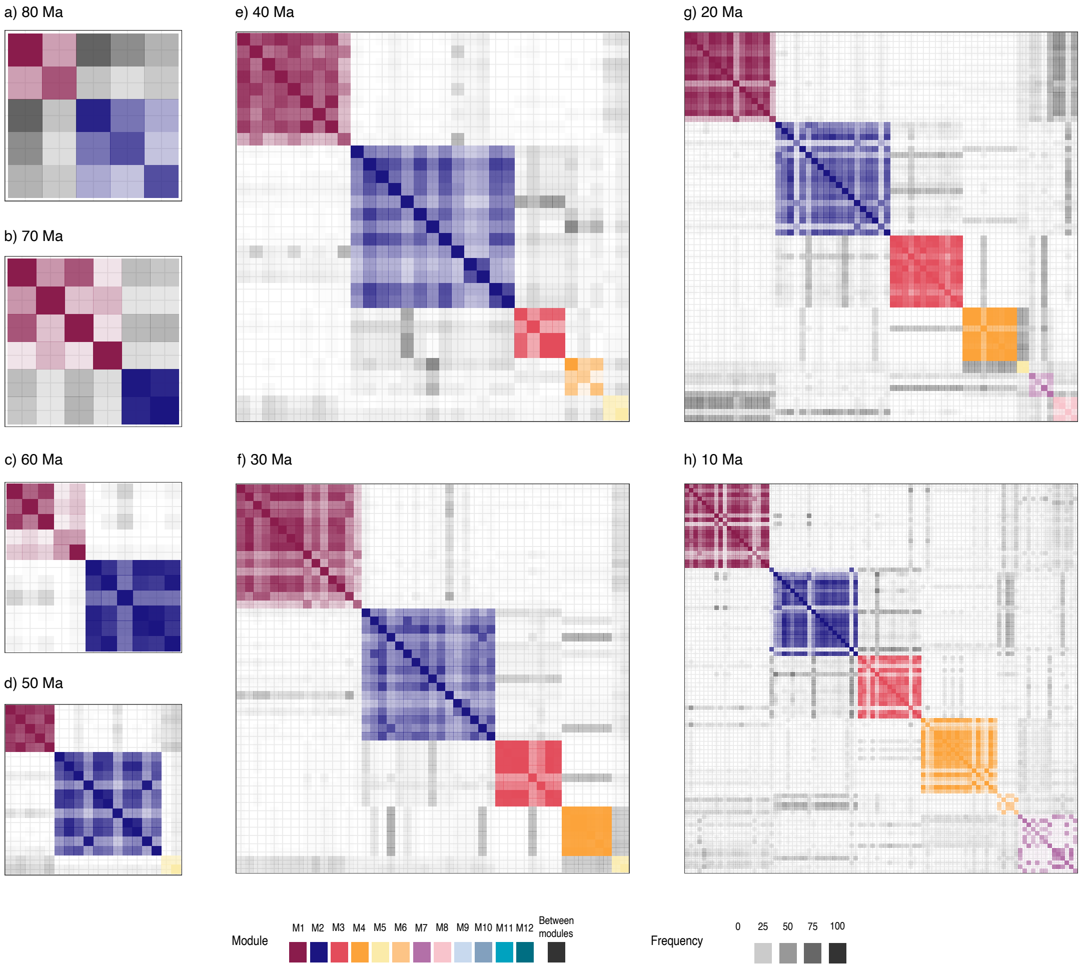

Figure 5: Heatmap of frequency with which each pair of nodes (butterflies and host plants) was assigned to the same module across 100 networks sampled throughout MCMC. In each panel, rows and columns contain all nodes included in the weighted summary network with probability threshold of 0.5 at the given time slice (depicted in Fig. 4). Rows and columns are ordered by module. When the nodes in the row and in the column are in the same module in the weighted summary network (Fig. 4), the cell takes the color of the module; otherwise, the cell is grey. The opacity of the cell is proportional to the posterior probability that the two nodes (row and column) belong to the same module.

Discussion

Given the recent developments in model-based statistical inference of historical ecological interactions, it is now possible to explicitly test complex mechanistic models of evolution of host-parasite interactions. Previously, these phenomena could only be addressed indirectly, for instance, through network analysis of extant interactions and phylogenetic comparative methods. In this paper, we use these novel methods to reconstruct the evolutionary history of the association between Pieridae butterflies and their host plants over time, with two goals in mind. First, we contribute to these new methods by developing new ways to explore the posterior distribution produced by Bayesian analysis of an explicit mechanistic model of host repertoire evolution. Second, we provide a powerful test of the ideas proposed in Braga et al. (2018) regarding the evolution of networks of host-parasite interactions. Our findings support the conclusions of the original study, while providing detailed insights into the underlying evolutionary processes.

One of the main ideas the new methods allowed us to test was that the evolution of butterfly-plant networks is driven by their repeated probing of new hosts combined with phylogenetic conservatism in host-use abilities. We estimated the rate of repertoire evolution in Pieridae to be near 6 host-use events for every 100 million years of butterfly evolution (per lineage). For comparison, the evolution of host repertoire in Nymphalini butterflies was estimated to be 20 times faster in the only previous analysis using this methodological framework (Braga et al. 2020). Genus-level rates for Pieridae are difficult to compare to species-level rates for Nymphalini, still, it is likely that pierids have been considerably more conservative in their host repertoires than the Nymphalini. Of all the estimated events, about half were host gains and half host losses (75 gains and 73 losses). Of these, a small subset of seven gains of five plant families had the strongest effect on the structure of the Pieridae-angiosperm network, creating and connecting the main modules: Capparaceae (gained once), Loranthaceae (twice), Santalaceae (once), Brassicaceae (twice), and Tropaeolaceae (once).

Based on extant interactions and phylogenetic information, Braga et al. (2018) suggested that the evolution of butterfly-plant interactions is shaped by a combination of processes consistent with the escape-and-radiate hypothesis (Ehrlich and Raven 1964) and processes consistent with the oscillation hypothesis (Janz and Nylin 2008). More specifically, they suggested that three types of host gains leave unique signatures in the network structure. First, a complete host shift (i.e. gain of new host followed by loss of ancestral host) produces a new module isolated from the rest of the network. Second, host-range expansion (i.e. colonization of new host without loss of ancestral host) increases the size of the module and creates nestedness within the module. And third, recolonizations (i.e. gain of a host that has been used in the past) connect different modules, increasing nestedness in the whole network. Besides these three types of host gain, host loss can also change the structure of the network. Host specialization (or host range contraction, i.e. loss of part of the host repertoire) can create new modules by breaking up the original module. We discuss the role of each one of these processes in the evolution of the Pieridae-angiosperm network below.

In agreement with previous studies, our analysis provided strong support for interactions between the first butterflies in the Pierinae subfamily and Brassicales hosts. The diversification of Pierinae was first explained as a radiation following the colonization of the chemically well-defended Brassicales plants by Ehrlich and Raven (1964). More recent studies identified the origins of defense and counter-defense mechanisms, which support the idea of an arms-race during Pierinae-Brassicales coevolution (Wheat et al. 2007; Edger et al. 2015). Both our reconstructions (Figs. 2 and 4) support the hypothesis that the colonization of Capparaceae (Brassicales) and subsequent loss of Fabaceae (Fabales) – the ancestral host – by early Pierinae butterflies created a new module in the network (M2 in Figs. 2, 4 and 5). All evidence from the present and the previous studies mentioned above suggest that the host shift from Fabaceae to Capparaceae was completed between 70 Ma and 60 Ma, which overlaps with the Cretaceous-Paleogene (K-Pg) extinction event. This period also coincides with an estimated increase in Brassicales diversification rate (Edger et al. 2015). Even though we cannot draw any conclusions about the relative roles of the K-Pg extinction event and of the coevolutionary arms-race on the shift in host use by Pieridae, all these factors were likely involved in the origin of the Pierinae-Brassicales association.

While during the first half of Pieridae diversification the Pieridae-angiosperm network was structured in two modules – M1 (basal pierids using Fabaceae) and M2 (Pierinae using Capparaceae) – during the second half many other plant families were added to the host repertoires of pierids. In the Late Eocene, there was a second significant change in the structure of the pierid-angiosperm network. We reconstructed the origin of modules M3 and M4 at 40 Ma, as a consequence of two host shifts and one host-range expansion. During the early diversification of Aporiina butterflies, one lineage started using the closely related Loranthaceae and Santalaceae, creating module M4, and seem to have completely lost Capparaceae from their host repertoire, given that we have no record of extant descendants feeding on Capparaceae. Around the same time, early Anthocharidini (the sister clade to *Hebomoia*) shifted from Capparaceae to the early Brassicaceae, creating part of module M3. The other part of M3 was composed of the emerging Pierina. One feature of the model of host-repertoire evolution used here is that it permits ancestral butterflies to have fed on any combination of ancestral plant hosts. This is evident in the reconstructed host repertoires of early Pierina, which include three plant families: Capparaceae and – the two newly acquired – Brassicaceae and Tropaeolaceae (Fig. 2). This host-range expansion coincides with the origin of the Core Brassicaceae and increases in diversification rates in both Pierina and Brassicaceae (Edger et al. 2015), thus having a major effect on the network structure.

Besides the detection of two large modules, between 40 Ma and 30 Ma is also when the network became both modular and nested (Fig. 3). Modularity likely increased because of the two new modules in the network (M3 and M4), while nestedness likely emerged because of the retention of Capparaceae in the repertoire of early Pierina, which connected modules M2 and M3. Even though the network increased considerably in the last 30 Myr, the general structure remained the same: most interactions are within the four largest modules (M1–4) and are organized in a modular and nested structure. However, while the level of modularity stayed almost constant, nestedness increased linearly over time (Fig. 3). This happened because most of the seven modules that were first detected in the past 30 Myr are connected to at least one, but often two, of the large modules. In other words, as butterflies gained new hosts and formed new modules, a subset of these butterflies retained or recolonized the ancestral host (Fabaceae, Capparaceae, Brassicaceae, Tropaeolaceae, Loranthaceae or Santalaceae, depending on the butterfly clade), preserving connectivity to the original modules. Thus, host-range expansions and recolonizations promoted a phase transition in the basic structure of the network, which went from a disconnected network composed of small, isolated modules, to a connected network with a giant component that connects most species through direct or indirect pathways (Guimares Jr. 2020). This is an important example of a mechanism for the emergence of a giant component in ecological networks, whose main consequence is the propagation of eco-evolutionary feedbacks across multiple species in the system.

In summary, the diversification and evolution of host repertoire of Pieridae butterflies can indeed be explained by a combination of the escape-and-radiate (Ehrlich and Raven 1964) and the oscillation hypothesis (Janz and Nylin 2008). Even though the Pierinae-Brassicales association has been a model system for research on the genetics of one-to-one coevolution, by allowing more complex coevolutionary histories, more of the dynamics can be explained. Here, we provide evidence for the mechanistic basis of host-repertoire evolution that underlie the patterns revealed by phylogenetic network analysis of butterfly-host plant interactions. Our results demonstrate the power of combining network analysis with Bayesian inference of host repertoire evolution in understanding how complex species interactions change over time. Future avenues of research should explore the extent to which host shifts, host range expansions, and host recolonizations characterize the evolution of other host-parasite systems.

Acknowledgements.— We thank Paulo R. Guimarães Jr. and Christopher W. Wheat for discussing and commenting on earlier versions of this paper. The Swedish Research Council supported [2015-04218 to S.N.] and [2014-05901 to F.R.].

*

Almeida-Neto, M., P. Guimarães, P. R. Guimarães, R. D. Loyola, and W. Ulrich. 2008. A consistent metric for nestedness analysis in ecological systems: reconciling concept and measurement. Oikos 117:1227–1239.

Almeida-Neto, M. and W. Ulrich. 2011. A straightforward computational approach for measuring nestedness using quantitative matrices. Environmental Modelling & Software 26:173 – 178.

Bascompte, J., P. Jordano, C. J. Melián, and J. M. Olesen. 2003. The nested assembly of plant-animal mutualistic networks. Proceedings of the National Academy of Sciences 100:9383–9387.

Beckett, S. J. 2016. Improved community detection in weighted bipartite networks. Royal Society Open Science 3:140536.

Braga, M. P., P. R. Guimarães Jr, C. W. Wheat, S. Nylin, and N. Janz. 2018. Unifying host-associated diversification processes using butterfly-plant networks. Nature communications 9.

Braga, M. P., M. J. Landis, S. Nylin, N. Janz, and F. Ronquist. 2020. Bayesian Inference of Ancestral Host-parasite Interactions under a Phylogenetic Model of Host Repertoire Evolution. Systematic biology .

Darwin, C. R. 1877. On the various contrivances by which British and foreign orchids are fertilised by insects. John Murray.

Dormann, C. F., B. Gruber, and J. Fruend. 2008. Introducing the bipartite package: Analysing ecological networks. R News 8:8–11.

Edger, P. P., H. M. Heidel-Fischer, M. Bekaert, J. Rota, G. Gloeckner, A. E. Platts, D. G. Heckel, J. P. Der, E. K. Wafula, M. Tang, J. A. Hofberger, A. Smithson, J. C.

Hall, M. Blanchette, T. E. Bureau, S. I. Wright, C. W. dePamphilis, M. E. Schranz, M. S. Barker, G. C. Conant, N. Wahlberg, H. Vogel, J. C. Pires, and C. W. Wheat. 2015. The butterfly plant arms-race escalated by gene and genome duplications. Proceedings of the National Academy of Sciences 112:8362–8366.

Ehrlich, P. R. and P. H. Raven. 1964. Butterflies and plants: a study in coevolution. Evolution 18:586.

Ferrer-Paris, J. R., A. Snchez-Mercado, . L. Viloria, and J. Donaldson. 2013. Congruence and Diversity of Butterfly-Host Plant Associations at Higher Taxonomic Levels. PLoS ONE 8:e63570.

Fordyce, J. A. 2010. Host shifts and evolutionary radiations of butterflies. Proceedings of the Royal Society B: Biological Sciences 277:3735–3743.

Forister, M. L., L. A. Dyer, M. S. Singer, J. O. I. Stireman, and J. T. Lill. 2012. Revisiting the evolution of ecological specialization, with emphasis on insect-plant interactions. Ecology 93:981–991.

Forister, M. L., V. Novotny, A. K. Panorska, L. Baje, Y. Basset, P. T. Butterill, L. Cizek, P. D. Coley, F. Dem, I. R. Diniz, P. Drozd, M. Fox, A. E. Glassmire, R. Hazen, J. Hrcek, J. P. Jahner, O. Kaman, T. J. Kozubowski, T. A. Kursar, O. T. Lewis, J. Lill, R. J. Marquis, S. E. Miller, H. C. Morais, M. Murakami, H. Nickel, N. A. Pardikes, R. E. Ricklefs, M. S. Singer, A. M. Smilanich, J. O. Stireman, S. Villamarn-Cortez, S. Vodka, M. Volf, D. L. Wagner, T. Walla, G. D. Weiblen, and L. A. Dyer. 2015. The global distribution of diet breadth in insect herbivores. Proceedings of the National Academy of Sciences 112:442 – 447.

Gelman, A. and D. B. Rubin. 1992. Inference from Iterative Simulation Using Multiple Sequences. Statistical Science 7:457–472.

Guimarães, P. R., P. Jordano, and J. N. Thompson. 2011. Evolution and coevolution in mutualistic networks. Ecology Letters 14:877–885.

Guimares Jr., P. R. 2020. The Structure of Ecological Networks Across Levels of Organization. Annual Review of Ecology, Evolution, and Systematics 51:1–28.

Höhna, S., M. J. Landis, T. A. Heath, B. Boussau, N. Lartillot, B. R. Moore, J. P. Huelsenbeck, and F. Ronquist. 2016. RevBayes: Bayesian Phylogenetic Inference Using Graphical Models and an Interactive Model-Specification Language. Systematic Biology 65:726–736.

Janz, N. 2011. Ehrlich and Raven Revisited: Mechanisms Underlying Codiversification of Plants and Enemies. Annual Review of Ecology, Evolution, and Systematics 42:71–89.

Janz, N. and S. Nylin. 2008. The oscillation hypothesis of host-plant range and speciation. Pages 203–215 *in* Specialization, speciation, and radiation: the evolutionary biology of herbivorous insects (K. J. Tilmon, ed.). . . . , California.

Jeffreys, H. 1961. The Theory of Probability. OUP Oxford.

Jurado-Rivera, J. and E. Petitpierre. 2015. New contributions to the molecular systematics and the evolution of host-plant associations in the genus chrysolina (coleoptera, chrysomelidae, chrysomelinae). ZooKeys 547:165–192.

Landis, M. J., N. J. Matzke, B. R. Moore, and J. P. Huelsenbeck. 2013. Bayesian analysis of biogeography when the number of areas is large. Systematic Biology 62:789–804.

Magallón, S., S. Gómez-Acevedo, L. L. Sánchez-Reyes, and T. Hernández-Hernández. 2015. A metacalibrated time-tree documents the early rise of flowering plant phylogenetic diversity. New Phytologist 207:437–453.

Marin, J.-M. and C. P. Robert. 2010. On resolving the Savage–Dickey paradox. Electronic Journal of Statistics 4:643–654.

Navaud, O., A. Barbacci, A. Taylor, J. P. Clarkson, and S. Raffaele. 2018. Shifts in diversification rates and host jump frequencies shaped the diversity of host range among sclerotiniaceae fungal plant pathogens. Molecular Ecology 27:1309–1323.

Nylin, S., S. Agosta, S. Bensch, W. A. Boeger, M. P. Braga, D. R. Brooks, M. L. Forister, P. A. Hambäck, E. P. Hoberg, T. Nyman, A. Schäpers, A. L. Stigall, C. W. Wheat, M. Österling, and N. Janz. 2018. Embracing Colonizations: A New Paradigm for Species Association Dynamics. Trends in Ecology & Evolution 33:4–14.

Nylin, S., J. Slove, and N. Janz. 2014. Host plant utilization, host range oscillations and diversification in nymphalid butterflies: a phylogenetic investigation. Evolution 68:105–124.

Oksanen, J., F. G. Blanchet, M. Friendly, R. Kindt, P. Legendre, D. McGlinn, P. R. Minchin, R. B. O'Hara, G. L. Simpson, P. Solymos, M. H. H. Stevens, E. Szoecs, and H. Wagner. 2019. vegan: Community Ecology Package. R package version 2.5-6.

Olesen, J. M., J. Bascompte, Y. L. Dupont, and P. Jordano. 2007. The modularity of pollination networks. Proceedings of the National Academy of Sciences of the United States of America 104:19891 – 19896.

Plummer, M., N. Best, K. Cowles, K. Vines, and 2006. 2006. CODA: convergence diagnosis and output analysis for MCMC. R News 6:7–11.

Quintero, I. and M. J. Landis. 2019. Interdependent Phenotypic and Biogeographic Evolution Driven by Biotic Interactions. Systematic Biology Syz082.

R Core Team. 2019. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing Vienna, Austria.

Robinson, D. M., D. T. Jones, H. Kishino, N. Goldman, and J. L. Thorne. 2003. Protein Evolution with Dependence Among Codons Due to Tertiary Structure. Molecular Biology and Evolution 20:1692–1704.

Suchard, M. A., R. E. Weiss, and J. S. Sinsheimer. 2001. Bayesian selection of continuous-time Markov chain evolutionary models. Molecular Biology and Evolution 18:1001–1013.

Tsang, L. M., K. H. Chu, Y. Nozawa, and C. Benny. 2014. Morphological and host specificity evolution in coral symbiont barnacles (balanomorpha: Pyrgomatidae) inferred from a multi-locus phylogeny. Molecular Phylogenetics and Evolution 77.

Verdinelli, I. and L. Wasserman. 1995. Computing Bayes Factors Using a Generalization of the Savage-Dickey Density Ratio. Journal of the American Statistical Association 90:614–618.

Vienne, D. M. d., G. Refregier, M. Lopez-Villavicencio, A. Tellier, M. E. Hood, and T. Giraud. 2013. Cospeciation vs host-shift speciation: methods for testing, evidence from natural associations and relation to coevolution. New Phytologist 198:347 – 385.

Wheat, C. W., H. Vogel, U. Wittstock, M. F. Braby, D. Underwood, and T. Mitchell-Olds. 2007. The genetic basis of a plant-insect coevolutionary key innovation. Proceedings of the National Academy of Sciences of the United States of America 104:20427–20431.