

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 JANZ¹, SÖREN NYLIN¹, FREDRIK RONQUIST³, AND
MICHAEL J LANDIS²

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.

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

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

36 In recent years, there has been a clear trend from a somewhat simplified
37 escape-and-radiate hypothesis to more complex models of coevolution, shifting from
38 one-to-one associations to diffuse coevolution, from tight to more loosely connected
39 evolutionary trajectories, and from interacting species-pairs to networks of interacting
40 species (Guimarães et al. 2011). In line with this trend, Braga et al. (2018) recently
41 suggested that coevolving host-parasite associations in general may be characterized by

42 processes fitting *both* the escape-and-radiate *and* the oscillation hypotheses. This was
43 based on network and phylogenetic analyses of two butterfly families, Nymphalidae and
44 Pieridae. Specifically, it was shown that these alternative diversification scenarios
45 generate different structural patterns in the networks that characterize extant
46 interactions between butterfly families and their host plants. The escape-and-radiate
47 scenario generates network modularity (Olesen et al. 2007; Braga et al. 2018), where
48 each module is composed of a given host taxon and closely-related butterflies that
49 diversified after shifting to the given host. Conversely, oscillations in host range
50 produces a specialist-generalist gradient in both trophic levels, where specialized
51 butterflies use a subset of the host plants used by closely-related generalists. Thus, the
52 oscillation hypothesis generates network nestedness (Bascompte et al. 2003; Braga et al.
53 2018).

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

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

82 METHODS

83 *Pierid Butterflies and Angiosperm Hosts*

84 We reconstructed historical interactions between Pieridae butterflies and their host
85 plants using a Bayesian phylogenetic approach (Braga et al. 2020). Interaction data
86 between butterfly genera and plant families were gathered from the literature (see
87 Supplementary Information). We used previously published time-calibrated phylogenies
88 for 66 Pieridae genera (Edger et al. 2015, Fig. S1) and angiosperm families (Edger et al.
89 2015; Magallón et al. 2015). We pruned the host angiosperm phylogeny, keeping all 33
90 angiosperm families that are known to be hosts of pierid butterflies, then collapsing
91 increasingly ancestral nodes until only 50 terminal branches were left. This increased
92 the chance that all ancestral angiosperm lineages that might have been used as hosts in
93 the past were included in the analysis, while keeping the analysis computationally
94 tractable.

Model of Host Repertoire Evolution

95

96 We modeled host repertoire evolution across Pieridae as a continuous-time
97 Markov chain (CTMC) that describes gain and loss of individual hosts. In the model,
98 the host repertoire of a given parasite is represented as a binary vector of length 50,
99 where each element within the vector describes the interaction between the parasite and
100 a given host plant family. Hosts (i.e. vector elements) can assume one of two states: 0
101 (non-host) or 1 (host). We assumed that each parasite must have at least one host at
102 any given time. Thus, the state space (i.e. the number of state combinations that a
103 host repertoire can assume) for this model includes $2^{50} - 1 \approx 1.13 \times 10^{15}$ unique
104 repertoires. We used a Bayesian data augmentation approach (Robinson et al. 2003;
105 Landis et al. 2013; Quintero and Landis 2019; Braga et al. 2020) to sample evolutionary
106 character histories under this large state space. We did not consider uncertainty in the
107 host or parasite phylogenies to facilitate the inference of model parameters under our
108 data augmentation method. Note that the original model described in Braga et al.
109 (2020) included three states (non-host, potential host and actual host), but because our
110 data set does not report information on potential hosts, model performance was poor
111 under the 3-state model.

112 In a 2-state model, two types of events can change the host repertoire: host gain
113 ($0 \rightarrow 1$) occurs with the rate λ_{01} , and host loss ($1 \rightarrow 0$) occurs with the rate λ_{10} . These
114 rates allow us to compute the probability of any given coevolutionary history based on
115 the instantaneous-rate matrix that defines the CTMC. This matrix is constructed such
116 that only one host in the repertoire is allowed to change in state at a time. Relative
117 gain and loss rates are constrained between 0 and 1, which are multiplied by global rate
118 scaling parameter, μ , to produce absolute rates of gain and loss.

119 Our model allowed for phylogenetic relatedness among hosts to influence how
120 easily a butterfly might expand its host repertoire to include a new host species.
121 Specifically, host gain rates were further multiplied by a phylogenetic-distance rate
122 modifier, which is defined as $e^{-\beta d_{ij}}$, where d_{ij} measures the relative phylogenetic

123 distance between the currently parasitized host i and the newly gained host j and β
124 rescales the magnitude of d_{ij} (see Braga et al. (2020) for details). That is, if $\beta > 0$,
125 parasites prefer to colonize new hosts that are phylogenetically similar to currently
126 parasitized hosts. If $\beta = 0$, the gain rates are not affected by the host tree. Following
127 Braga et al. (2020), we measured phylogenetic distance between host lineages in two
128 different ways: (1) using what we call the *anagenetic tree*, where distances reflect
129 time-calibrated divergence times among hosts, and (2) using a modified *cladogenetic*
130 tree, where all host branch lengths were set to 1, approximating phylogenetic distances
131 that are proportional to the number of older (i.e. family-level) cladogenetic events that
132 separate two taxa.

133 *Summarizing ecological interactions through time*

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

149 interactions. We discuss each of these three items in detail below and explore new ways
150 to summarize host repertoire evolution.

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

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

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

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

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

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

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

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

221 *Bayesian inference method*

222 Bayesian MCMC was used to estimate the joint posterior distribution of the parameters
223 in the model of host repertoire evolution described above. All analyses were performed
224 in RevBayes (Höhna et al. 2016) using the inference strategy described in Braga et al.
225 (2020). We ran four independent MCMC analyses (two with the anagenetic distance
226 and two with the cladogenetic distance between hosts), each set to run for 2×10^5
227 cycles, sampling parameters and node histories every 50 cycles, and discarding the first
228 2×10^4 as burnin. Prior distributions were $\mu \sim Exponential(10)$, $\beta \sim Exponential(1)$,
229 and $\lambda \sim Dirichlet(1, 1)$, where elements of λ follow the marginal distribution,

230 $\lambda_{i,j} \sim \text{Beta}(1,1)$. To verify that MCMC analyses converged to the same posterior
231 distribution, we applied the Gelman diagnostic (Gelman and Rubin 1992) as
232 implemented in the R package *coda* (Plummer et al. 2006). Results from a single
233 MCMC analysis are presented.

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

240 *Code availability*

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

245 RESULTS

246 Posterior estimates of Pieridae host repertoire evolution were partially sensitive
247 to whether we measured distances between host lineages in units of geological time or in
248 units of major cladogenetic events (Fig. 1). When measuring anagenetic distances
249 between host lineages, posterior mean (95% highest posterior density; HPD95)
250 estimates were: global rate scaling factor for host repertoire evolution $\mu = 0.02$ (0.015 -
251 0.026), phylogenetic-distance power $\beta = 2.1$ (0.017 - 3.82), relative host gain rate
252 $\lambda_{01} = 0.035$ (0.022 - 0.047), and relative host loss rate $\lambda_{10} = 0.965$ (0.95 - 0.98). Mean
253 estimates were similar when distances between hosts were measured in units of

254 cladogenetic events: $\mu = 0.019$ (0.014 - 0.024), $\beta = 1.48$ (1.02 - 1.97), $\lambda_{01} = 0.027$ (0.017
255 - 0.036), and $\lambda_{10} = 0.97$ (0.96 - 0.98). An important difference between the two
256 inferences is that the HPD95 for β under cladogenetic distance excludes $\beta = 0$, whereas
257 β estimated under anagenetic distance assigns a non-zero probability (≈ 0.1) to $\beta = 0$.
258 The decisive support for $\beta > 0$ when using cladogenetic distance led us to focus
259 primarily on this reconstruction throughout the main text (Fig. S2 for results with
260 anagenetic distance). Because rate parameters can be difficult to interpret, we also
261 calculated the average number of proposed events across MCMC samples, which was
262 148, being 75 host gains and 73 host losses throughout the diversification of Pieridae.

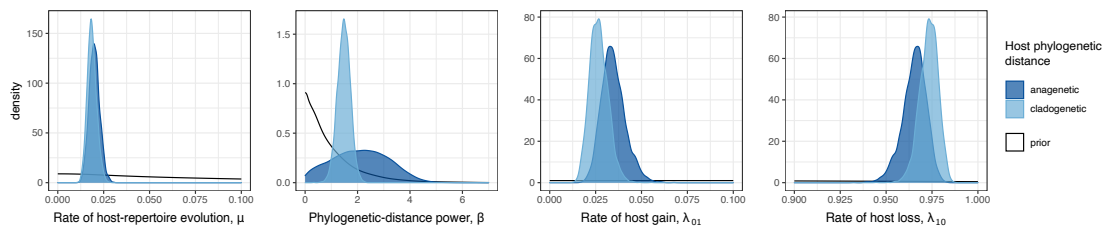


Figure 1: Estimated marginal posterior densities for parameters in the host-repertoire evolution model using two different representations of the phylogenetic distance between host-plant families: anagenetic (time) or cladogenetic (number of branches).

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

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

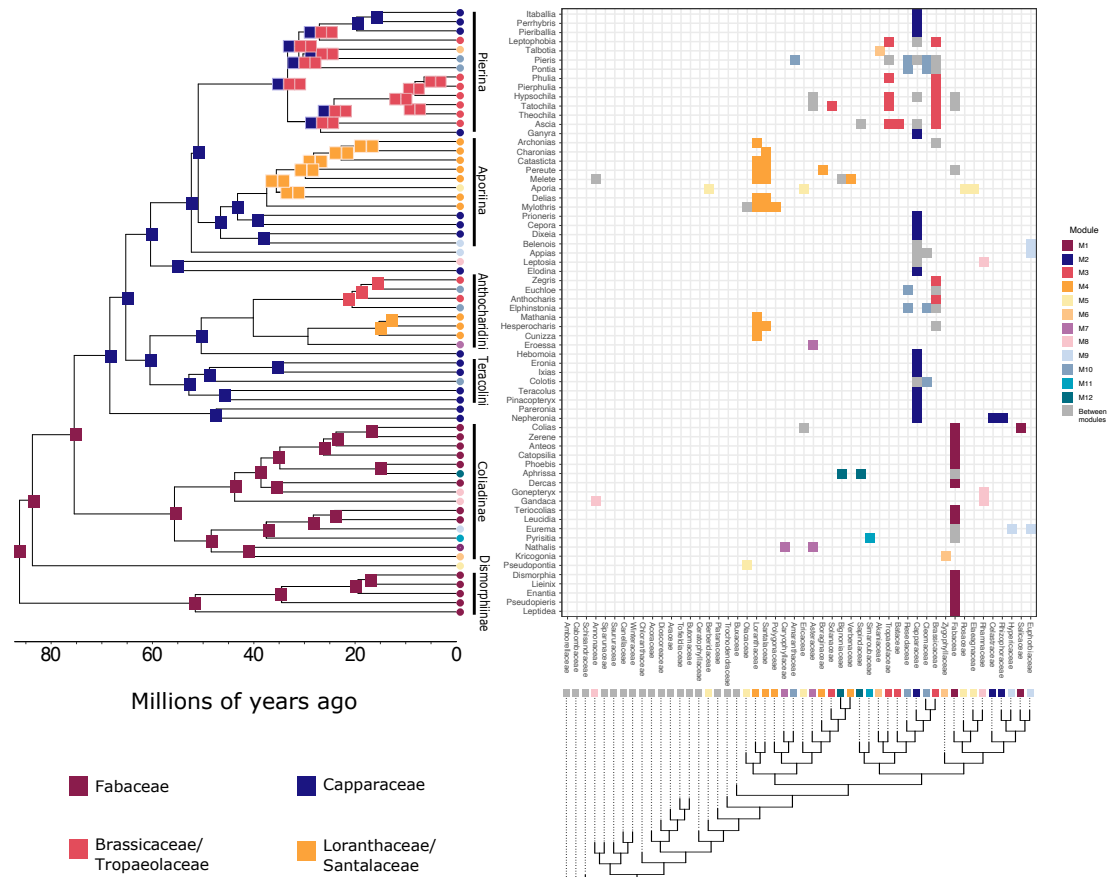


Figure 2: Ancestral state reconstruction showing interactions with marginal posterior probability ≥ 0.9 . The model reconstructs how host repertoire evolved along the Pieridae phylogeny (left), based on the observed butterfly-plant interactions (top-right), and the cladogenetic distance between hosts (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.

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

280 According to the posterior distribution of Q and NODF based on networks
281 sampled from the MCMC, modularity and nestedness were first detectable 30 Ma (Fig.
282 3; for raw Q and NODF values see Fig. S3). But while the support for modularity has
283 not changed much in the last 30 Myr, support for nestedness has increased linearly in
284 the past 50 Myr. Overall, the summary networks have overestimated the presence of
285 modularity, and only the weighted summary network with the 0.1-threshold correctly
286 estimated that significant modularity appeared 30 Ma (Fig. 3 upper panel). On the
287 other hand, the summary networks underestimated the existence of nestedness in
288 ancestral networks (Fig. 3 lower panel), with several networks being significantly less
289 nested than expected by chance, especially with the binary networks.

290 The present-day Pieridae-angiosperm network is characterized by both higher
291 modularity ($M = 0.64$, $p \leq 0.001$, Z -score = 3.62) and nestedness (NODF = 14.8, p
292 ≤ 0.001 , Z -score = 11.21) than expected by chance. Most of the butterfly lineages
293 within Dismorphiinae and Coliadinae are associated with Fabaceae hosts (module M1),
294 while Pierinae butterflies use many other host families (Fig. 2), the most common being
295 Capparaceae (module M2), Brassicaceae + Tropaeolaceae (M3) and Loranthaceae +
296 Santalaceae (M4). Interestingly, some Pierinae butterflies recolonized Fabaceae and
297 others colonized new hosts while keeping the old host in their repertoire, resulting in
298 among-module interactions that connected the whole network and produced signal for
299 nestedness. By exploring the posterior distribution of ancestral interactions, we were
300 able to characterize how this network was assembled throughout the diversification of
301 Pieridae butterflies, as described below.

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

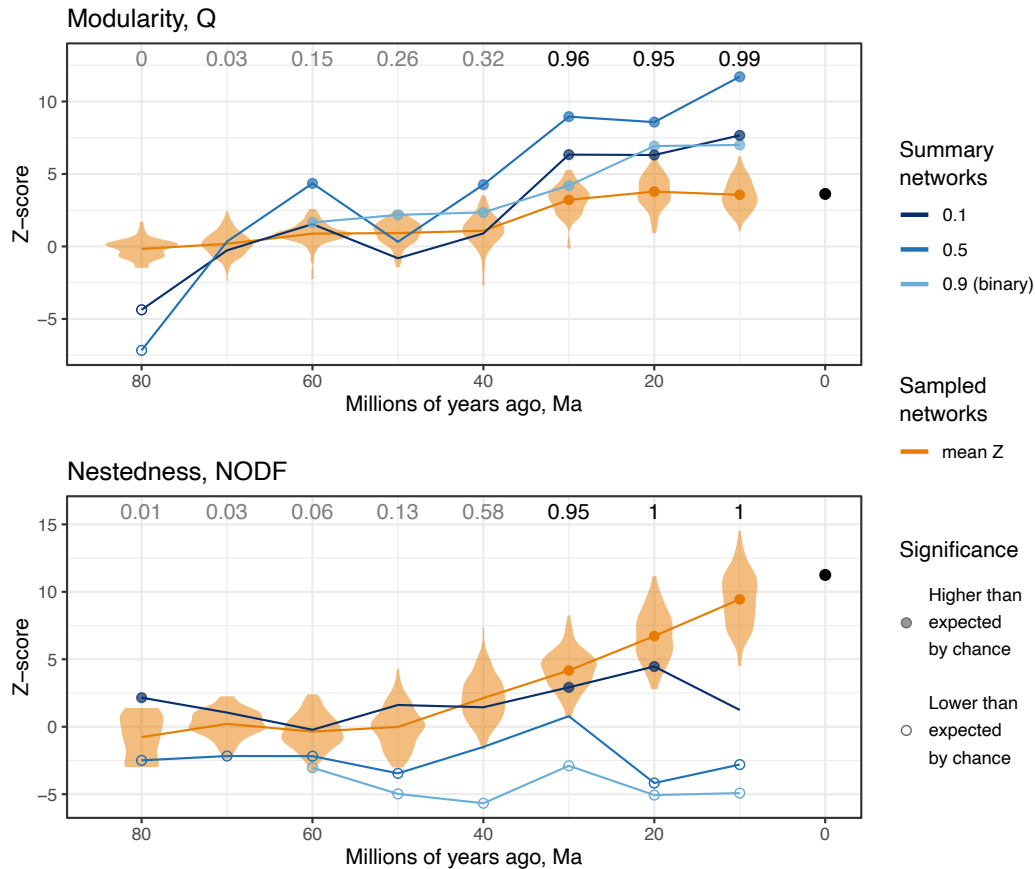


Figure 3: Structure of the Pieridae-angiosperm network over time. Z -scores for (a) modularity and (b) nestedness for summary (blues) and sampled networks (orange) from 80 Ma to 10 Ma, and for the observed present-day network (black circles). Each orange violin represents the distribution of Z -scores 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 $\alpha = 0.05$.

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

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

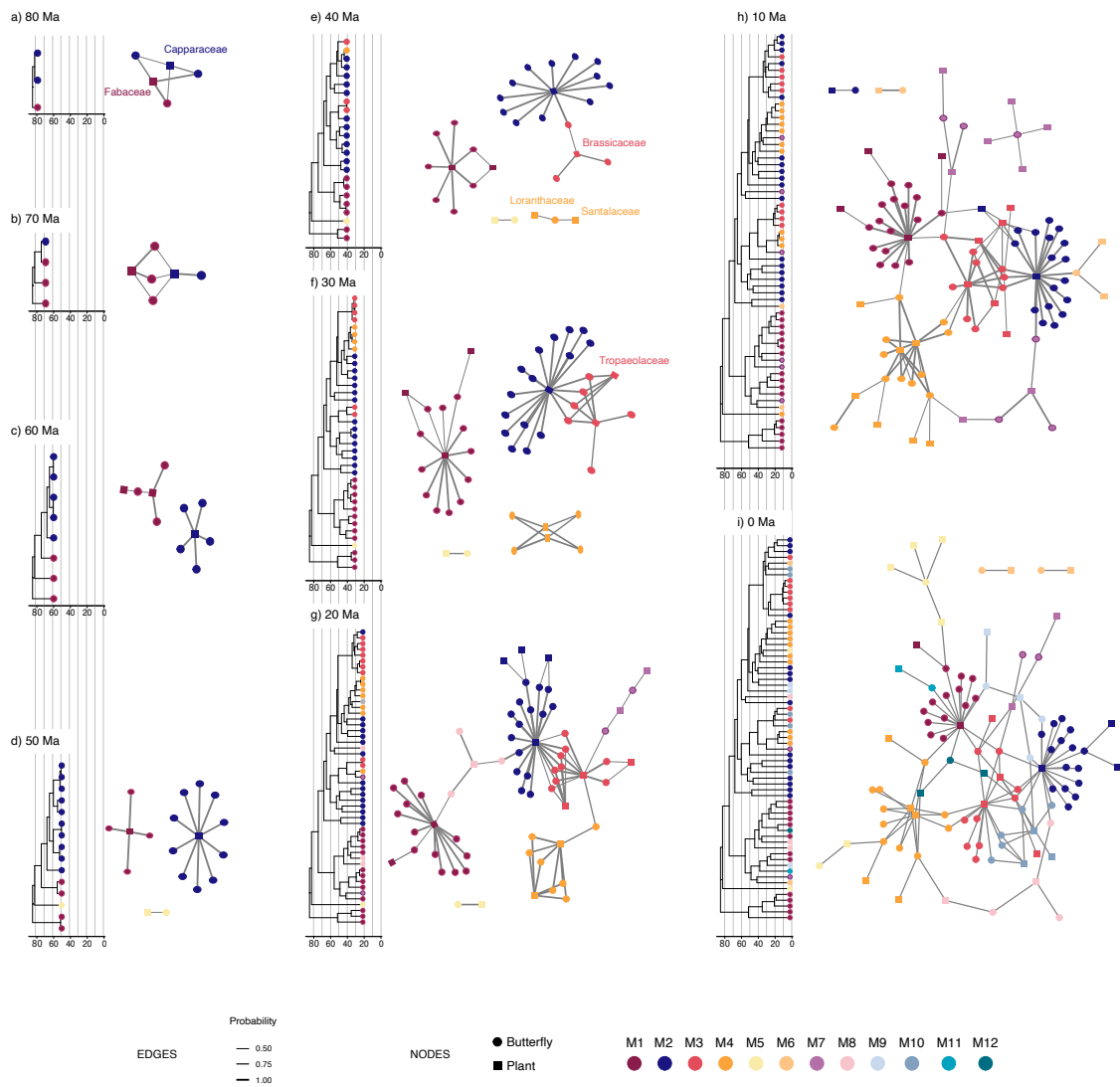


Figure 4: Evolution of the Pieridae-angiosperm network across nine time slices from 80 Ma to the present. Each panel (a-i) shows the butterfly lineages extant at a time slice (left) and the estimated network (right) of interactions with at least 0.5 posterior probability. Edge width is proportional to 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 were first colonized by Pieridae.

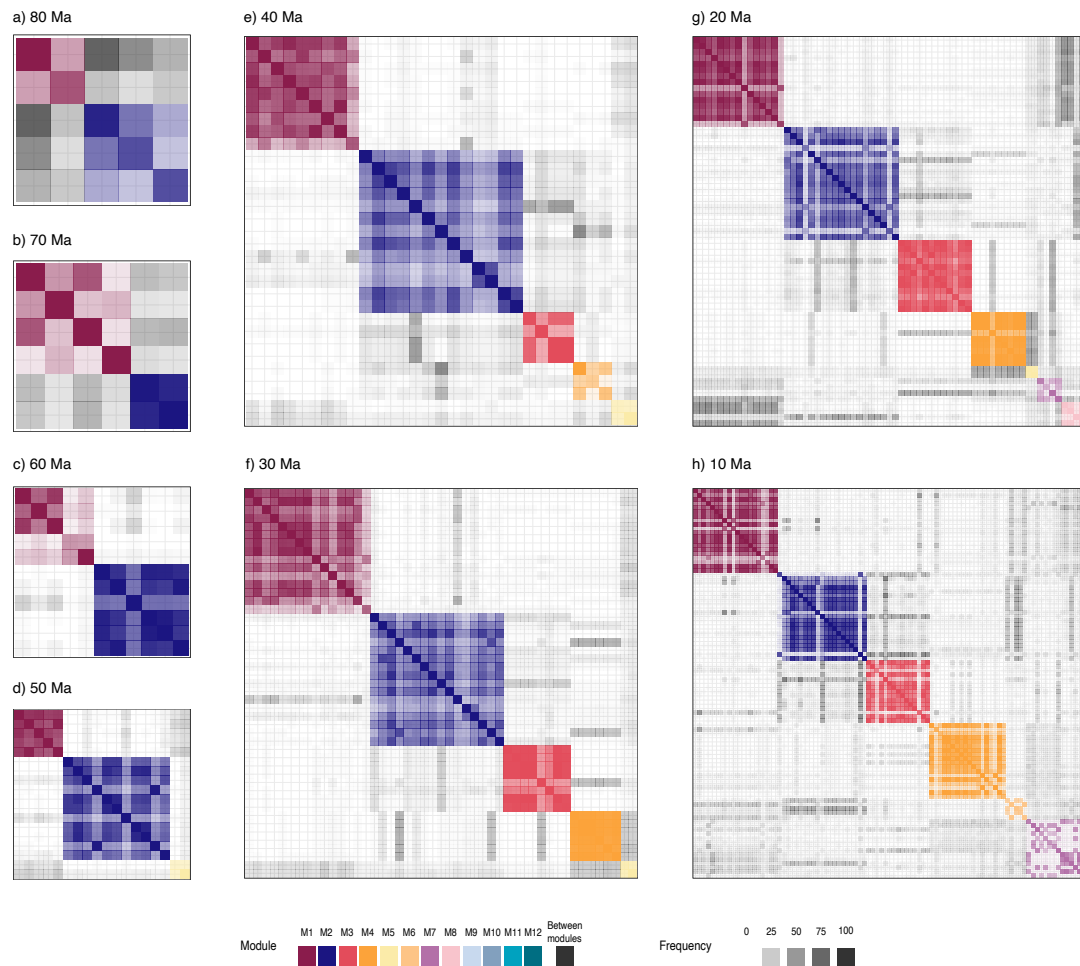


Figure 5: Heatmap of frequency with which each pair of nodes (butterflies and host plants) was assigned to the same module across 100 networks sampled throughout MCMC. In each panel, rows and columns contain all nodes included in the weighted summary network with probability threshold of 0.5 at the given 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

330

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

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

357 Santalaceae (once), Brassicaceae (twice), and Tropaeolaceae (once).

358 Based on extant interactions and phylogenetic information, Braga et al. (2018)
359 suggested that the evolution of butterfly-plant interactions is shaped by a combination
360 of processes consistent with the escape-and-radiate hypothesis (Ehrlich and Raven
361 1964) and processes consistent with the oscillation hypothesis (Janz and Nylin 2008).
362 More specifically, they suggested that three types of host gains leave unique signatures
363 in the network structure. First, a complete host shift (i.e. gain of new host followed by
364 loss of ancestral host) produces a new module isolated from the rest of the network.
365 Second, host-range expansion (i.e. colonization of new host without loss of ancestral
366 host) increases the size of the module and creates nestedness within the module. And
367 third, recolonizations (i.e. gain of a host that has been used in the past) connect
368 different modules, increasing nestedness in the whole network. Besides these three types
369 of host gain, host loss can also change the structure of the network. Host specialization
370 (or host range contraction, i.e. loss of part of the host repertoire) can create new
371 modules by breaking up the original module. We discuss the role of each one of these
372 processes in the evolution of the Pieridae-angiosperm network below.

373 In agreement with previous studies, our analysis provided strong support for
374 interactions between the first butterflies in the Pierinae subfamily and Brassicales hosts.
375 The diversification of Pierinae was first explained as a radiation following the
376 colonization of the chemically well-defended Brassicales plants by Ehrlich and Raven
377 (1964). More recent studies identified the origins of defense and counter-defense
378 mechanisms, which support the idea of an arms-race during Pierinae-Brassicales
379 coevolution (Wheat et al. 2007; Edger et al. 2015). Both our reconstructions (Figs. 2
380 and 4) support the hypothesis that the colonization of Capparaceae (Brassicales) and
381 subsequent loss of Fabaceae (Fabales) – the ancestral host – by early Pierinae butterflies
382 created a new module in the network (M2 in Figs. 2, 4 and 5). All evidence from the
383 present and the previous studies mentioned above suggest that the host shift from
384 Fabaceae to Capparaceae was completed between 70 Ma and 60 Ma, which overlaps

385 with the Cretaceous-Paleogene (K-Pg) extinction event. This period also coincides with
386 an estimated increase in Brassicales diversification rate (Edger et al. 2015). Even
387 though we cannot draw any conclusions about the relative roles of the K-Pg extinction
388 event and of the coevolutionary arms-race on the shift in host use by Pieridae, all these
389 factors were likely involved in the origin of the Pierinae-Brassicales association.

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

409 Besides the detection of two large modules, between 40 Ma and 30 Ma is also
410 when the network became both modular and nested (Fig. 3). Modularity likely
411 increased because of the two new modules in the network (M3 and M4), while
412 nestedness likely emerged because of the retention of Capparaceae in the repertoire of

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

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

441 recolonizations characterize the evolution of other host-parasite systems.

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

*

445

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

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

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

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

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

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

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

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

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

- 469 Hall, M. Blanchette, T. E. Bureau, S. I. Wright, C. W. dePamphilis, M. E. Schranz,
470 M. S. Barker, G. C. Conant, N. Wahlberg, H. Vogel, J. C. Pires, and C. W. Wheat.
471 2015. The butterfly plant arms-race escalated by gene and genome duplications.
472 *Proceedings of the National Academy of Sciences* 112:8362–8366.
- 473 Ehrlich, P. R. and P. H. Raven. 1964. Butterflies and plants: a study in coevolution.
474 *Evolution* 18:586.
- 475 Ferrer-Paris, J. R., A. Snchez-Mercado, . L. Vilorio, and J. Donaldson. 2013.
476 Congruence and Diversity of Butterfly-Host Plant Associations at Higher Taxonomic
477 Levels. *PLoS ONE* 8:e63570.
- 478 Fordyce, J. A. 2010. Host shifts and evolutionary radiations of butterflies. *Proceedings*
479 *of the Royal Society B: Biological Sciences* 277:3735–3743.
- 480 Forister, M. L., L. A. Dyer, M. S. Singer, J. O. I. Stireman, and J. T. Lill. 2012.
481 Revisiting the evolution of ecological specialization, with emphasis on insect-plant
482 interactions. *Ecology* 93:981–991.
- 483 Forister, M. L., V. Novotny, A. K. Panorska, L. Baje, Y. Basset, P. T. Butterill,
484 L. Cizek, P. D. Coley, F. Dem, I. R. Diniz, P. Drozd, M. Fox, A. E. Glassmire,
485 R. Hazen, J. Hrcek, J. P. Jahner, O. Kaman, T. J. Kozubowski, T. A. Kursar, O. T.
486 Lewis, J. Lill, R. J. Marquis, S. E. Miller, H. C. Morais, M. Murakami, H. Nickel,
487 N. A. Pardikes, R. E. Ricklefs, M. S. Singer, A. M. Smilanich, J. O. Stireman,
488 S. Villamarn-Cortez, S. Vodka, M. Volf, D. L. Wagner, T. Walla, G. D. Weiblen, and
489 L. A. Dyer. 2015. The global distribution of diet breadth in insect herbivores.
490 *Proceedings of the National Academy of Sciences* 112:442 – 447.
- 491 Gelman, A. and D. B. Rubin. 1992. Inference from Iterative Simulation Using Multiple
492 Sequences. *Statistical Science* 7:457–472.
- 493 Guimarães, P. R., P. Jordano, and J. N. Thompson. 2011. Evolution and coevolution in
494 mutualistic networks. *Ecology Letters* 14:877–885.

- 495 Guimares Jr., P. R. 2020. The Structure of Ecological Networks Across Levels of
496 Organization. *Annual Review of Ecology, Evolution, and Systematics* 51:1–28.
- 497 Höhna, S., M. J. Landis, T. A. Heath, B. Boussau, N. Lartillot, B. R. Moore, J. P.
498 Huelsenbeck, and F. Ronquist. 2016. RevBayes: Bayesian Phylogenetic Inference
499 Using Graphical Models and an Interactive Model-Specification Language. *Systematic
500 Biology* 65:726–736.
- 501 Janz, N. 2011. Ehrlich and Raven Revisited: Mechanisms Underlying Codiversification
502 of Plants and Enemies. *Annual Review of Ecology, Evolution, and Systematics*
503 42:71–89.
- 504 Janz, N. and S. Nylin. 2008. The oscillation hypothesis of host-plant range and
505 speciation. Pages 203–215 *in* Specialization, speciation, and radiation: the
506 evolutionary biology of herbivorous insects (K. J. Tilmon, ed.). . . . , California.
- 507 Jeffreys, H. 1961. *The Theory of Probability*. OUP Oxford.
- 508 Jurado-Rivera, J. and E. Petitpierre. 2015. New contributions to the molecular
509 systematics and the evolution of host-plant associations in the genus *chrysolina*
510 (*coleoptera*, *chrysomelidae*, *chrysomelinae*). *ZooKeys* 547:165–192.
- 511 Landis, M. J., N. J. Matzke, B. R. Moore, and J. P. Huelsenbeck. 2013. Bayesian
512 analysis of biogeography when the number of areas is large. *Systematic Biology*
513 62:789–804.
- 514 Magallón, S., S. Gómez-Acevedo, L. L. Sánchez-Reyes, and T. Hernández-Hernández.
515 2015. A metacalibrated time-tree documents the early rise of flowering plant
516 phylogenetic diversity. *New Phytologist* 207:437–453.
- 517 Marin, J.-M. and C. P. Robert. 2010. On resolving the Savage–Dickey paradox.
518 *Electronic Journal of Statistics* 4:643–654.

- 519 Navaud, O., A. Barbacci, A. Taylor, J. P. Clarkson, and S. Raffaele. 2018. Shifts in
520 diversification rates and host jump frequencies shaped the diversity of host range
521 among sclerotiniaceae fungal plant pathogens. *Molecular Ecology* 27:1309–1323.
- 522 Nylin, S., S. Agosta, S. Bensch, W. A. Boeger, M. P. Braga, D. R. Brooks, M. L.
523 Forister, P. A. Hambäck, E. P. Hoberg, T. Nyman, A. Schäpers, A. L. Stigall, C. W.
524 Wheat, M. Österling, and N. Janz. 2018. Embracing Colonizations: A New Paradigm
525 for Species Association Dynamics. *Trends in Ecology & Evolution* 33:4–14.
- 526 Nylin, S., J. Slove, and N. Janz. 2014. Host plant utilization, host range oscillations and
527 diversification in nymphalid butterflies: a phylogenetic investigation. *Evolution*
528 68:105–124.
- 529 Oksanen, J., F. G. Blanchet, M. Friendly, R. Kindt, P. Legendre, D. McGlenn, P. R.
530 Minchin, R. B. O’Hara, G. L. Simpson, P. Solymos, M. H. H. Stevens, E. Szoecs, and
531 H. Wagner. 2019. *vegan: Community Ecology Package*. R package version 2.5-6.
- 532 Olesen, J. M., J. Bascompte, Y. L. Dupont, and P. Jordano. 2007. The modularity of
533 pollination networks. *Proceedings of the National Academy of Sciences of the United*
534 *States of America* 104:19891 – 19896.
- 535 Plummer, M., N. Best, K. Cowles, K. Vines, and 2006. 2006. CODA: convergence
536 diagnosis and output analysis for MCMC. *R News* 6:7–11.
- 537 Quintero, I. and M. J. Landis. 2019. Interdependent Phenotypic and Biogeographic
538 Evolution Driven by Biotic Interactions. *Systematic Biology* Syz082.
- 539 R Core Team. 2019. *R: A Language and Environment for Statistical Computing*. R
540 Foundation for Statistical Computing Vienna, Austria.
- 541 Robinson, D. M., D. T. Jones, H. Kishino, N. Goldman, and J. L. Thorne. 2003. Protein
542 Evolution with Dependence Among Codons Due to Tertiary Structure. *Molecular*
543 *Biology and Evolution* 20:1692–1704.

- 544 Suchard, M. A., R. E. Weiss, and J. S. Sinsheimer. 2001. Bayesian selection of
545 continuous-time Markov chain evolutionary models. *Molecular Biology and Evolution*
546 18:1001–1013.
- 547 Tsang, L. M., K. H. Chu, Y. Nozawa, and C. Benny. 2014. Morphological and host
548 specificity evolution in coral symbiont barnacles (balanomorpha: Pyrgomatidae)
549 inferred from a multi-locus phylogeny. *Molecular Phylogenetics and Evolution* 77.
- 550 Verdinelli, I. and L. Wasserman. 1995. Computing Bayes Factors Using a Generalization
551 of the Savage-Dickey Density Ratio. *Journal of the American Statistical Association*
552 90:614–618.
- 553 Vienne, D. M. d., G. Refregier, M. Lopez-Villavicencio, A. Tellier, M. E. Hood, and
554 T. Giraud. 2013. Cospeciation vs host-shift speciation: methods for testing, evidence
555 from natural associations and relation to coevolution. *New Phytologist* 198:347 – 385.
- 556 Wheat, C. W., H. Vogel, U. Wittstock, M. F. Braby, D. Underwood, and
557 T. Mitchell-Olds. 2007. The genetic basis of a plant-insect coevolutionary key
558 innovation. *Proceedings of the National Academy of Sciences of the United States of*
559 *America* 104:20427–20431.