






# Genome-wide macroevolutionary signatures of key innovations in butterflies colonizing new host plants

Rémi Allio <sup>1</sup>✉, Benoit Nabholz<sup>1</sup>, Stefan Wanke <sup>2</sup>, Guillaume Chomicki<sup>3</sup>, Oscar A. Pérez-Escobar<sup>4</sup>, Adam M. Cotton <sup>5</sup>, Anne-Laure Clamens<sup>6</sup>, Gaël J. Kergoat <sup>6</sup>, Felix A. H. Sperling<sup>7</sup> & Fabien L. Condamine <sup>1,7</sup>✉

The mega-diversity of herbivorous insects is attributed to their co-evolutionary associations with plants. Despite abundant studies on insect-plant interactions, we do not know whether host-plant shifts have impacted both genomic adaptation and species diversification over geological times. We show that the antagonistic insect-plant interaction between swallowtail butterflies and the highly toxic birthworts began 55 million years ago in Beringia, followed by several major ancient host-plant shifts. This evolutionary framework provides a valuable opportunity for repeated tests of genomic signatures of macroevolutionary changes and estimation of diversification rates across their phylogeny. We find that host-plant shifts in butterflies are associated with both genome-wide adaptive molecular evolution (more genes under positive selection) and repeated bursts of speciation rates, contributing to an increase in global diversification through time. Our study links ecological changes, genome-wide adaptations and macroevolutionary consequences, lending support to the importance of ecological interactions as evolutionary drivers over long time periods.

<sup>1</sup> CNRS, IRD, EPHE, Institut des Sciences de l'Evolution de Montpellier, Université de Montpellier, Place Eugène Bataillon, 34095 Montpellier, France. <sup>2</sup> Institut für Botanik, Technische Universität Dresden, Zellescher Weg 20b, 01062 Dresden, Germany. <sup>3</sup> Department of Bioscience, Durham University, Stockton Road, Durham DH1 3LE, UK. <sup>4</sup> Royal Botanic Gardens, Kew TW9 3AB, UK. <sup>5</sup> 86/2 Moo 5, Tambon Nong Kwai, Hang Dong Chiang Mai, Thailand. <sup>6</sup> CBGP, INRAE, CIRAD, IRD, Montpellier SupAgro, Univ. Montpellier, Montpellier, France. <sup>7</sup> Department of Biological Sciences, University of Alberta, Edmonton T6G 2E9 AB, Canada. ✉email: [rem.allio@yahoo.fr](mailto:rem.allio@yahoo.fr); [fabien.condamine@gmail.com](mailto:fabien.condamine@gmail.com)

Plants and phytophagous insects account for the majority of the documented species of terrestrial organisms<sup>1,2</sup>. To explain the high diversity of insects, a long held hypothesis states that their diversification is directly related to that of plants<sup>3,4</sup>. More than half a century ago, Ehrlich and Raven<sup>5</sup> proposed a model in which a continual arms race of attacks by herbivorous insects and new defences by their host plants is linked to species diversification via the creation of new adaptive zones, later termed the ‘escape-and-radiate’ model<sup>6</sup>. According to Ehrlich and Raven<sup>5</sup>, these developments mainly correspond to toxic secondary compounds in plants, and the associated detoxification mechanisms in insects. This model would apply to all plants and plant-eating insects and could explain why these groups represent an important part of global biodiversity<sup>7,8</sup>.

Study of insect–plant interactions has progressed tremendously since then through a focus on host chemistry<sup>9</sup>, phylogenetics<sup>10,11</sup> and genomics<sup>12–15</sup>. Divergence of key gene families<sup>13–16</sup> and high speciation rates<sup>17–19</sup> have been identified after host–plant shifts, with one example linking duplication of key genes to the ability to feed on new plants and increase diversification<sup>13</sup>. The emerging consensus from most phylogenetic studies indicates (1) strong phylogenetic conservatism of host–plant associations (related insect species tend to feed on plants that are also related), suggesting ancient and specialized biotic interactions<sup>20</sup>, and (2) enhanced diversification rates for clades shifting to new host–plant groups compared to those remaining on ancestral plants. Despite high levels of conservatism and specialization, bursts of insect diversification appear to mainly be a consequence of host shifts<sup>21</sup>, and this somewhat paradoxical conclusion can be understood by considering ecological as well as genetic mechanisms behind host shifts<sup>12,15</sup>. There are several ways—both direct and indirect—that interactions can influence speciation<sup>22</sup>, with or without host–plant-based divergent selection on reproductive barriers. One current debate is on the relative importance of radiations following shifts to new adaptive zones and elevated rates of speciation in groups with plastic and diverse host use<sup>23–25</sup>. Increasingly sophisticated use of time-calibrated phylogenies is being made to investigate the actual timing and rate of diversification and to link such events more conclusively to other factors that may have been important, whether biotic or abiotic<sup>18,19</sup>.

Genomic aspects of adaptation by herbivorous insects to their host plants have received significant attention<sup>26</sup>, but few studies have put their genomic data into phylogenetic perspectives. A seminal study by Edger et al.<sup>13</sup> on the evolutionary arms race between Pierinae butterflies and their Brassicales host plants showed that shifts in diversification within the plants and their butterflies are associated with gradual changes in plant chemical defences and insect molecular counter adaptations. They identified the genomic mechanisms (gene and genome duplications) explaining the evolution of biosynthetic pathways associated with this arms race. More clues for host-encoded digestive and detoxification mechanisms come from a cross-taxonomic comparison of the gut microbiome of caterpillars with other insects and vertebrates<sup>27</sup>. The microbes in caterpillar guts are unusually at low densities, and reflect the abundance and composition of leaf-associated microbes in the caterpillar faeces, with high pH, simple gut structure, and fast transit times potentially preventing microbial colonization.

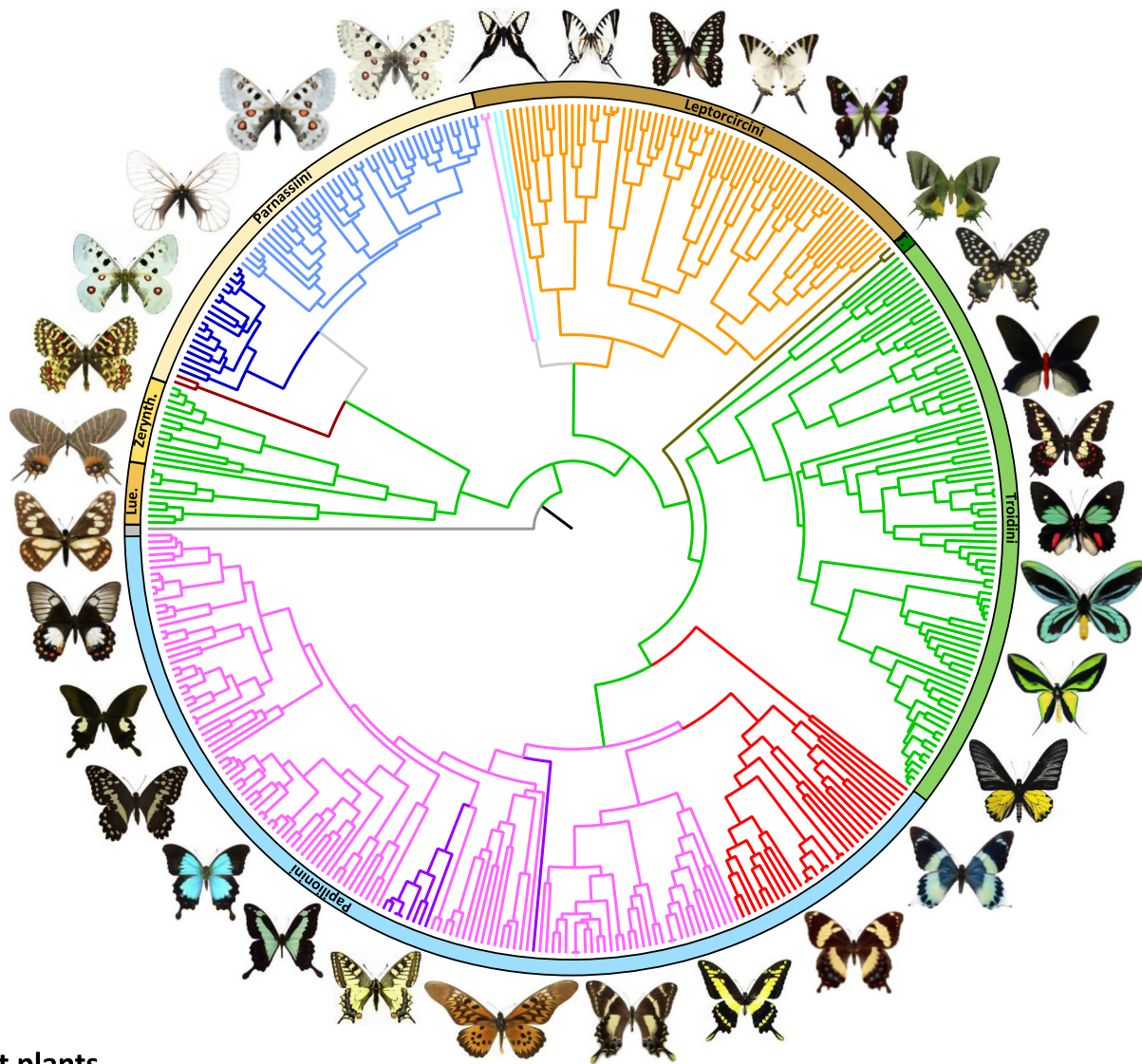
These recent results have illustrated the need for a multi-disciplinary approach to studying the evolution of insect–plant interactions within a macroevolutionary and genomic framework. However, a major knowledge gap lies in our understanding of the evolutionary links and drivers of host–plant shifts, genome-wide signatures of adaptations and processes of species diversification<sup>28</sup>. As noted by Hembry and Weber<sup>29</sup>, this implies that the questions of if, when and how coevolution has an impact on

macroevolutionary dynamics remain open challenges. Here we address this gap with an emblematic group that was instrumental in Ehrlich and Raven’s model—the swallowtail butterflies (Lepidoptera: Papilionidae). Swallowtail caterpillars feed on a range of different flowering families<sup>30</sup>, but a third of all species, including the tribes Zerynthiini (Parnassiinae), Luehdorfiini (Parnassiinae) and Troidini (Papilioninae), feeds exclusively on the birthwort family (Aristolochiaceae), which is one of the most toxic plant groups<sup>31</sup>. The Aristolochiaceae notoriously contain toxic aristolochic acids, which are known to be carcinogenic to many organisms, and Papilionidae are among the few that can feed on these plants<sup>32,33</sup>. By eating these toxic plants, the caterpillars sequester aristolochic acids that render both the caterpillars and the adults unpalatable for predators<sup>31</sup>. Interestingly, previous phylogenetic estimations of ancestral states indicated either that Aristolochiaceae was the ancestral host plant of Papilionidae<sup>34</sup> or that Aristolochiaceae was colonized twice<sup>35</sup>, suggesting that the host–plant shifts have ancient origins and seem to be highly constrained as shown by the high level of host conservatism. Moreover, the arms race between Papilionidae and their host plants has been demonstrated at the molecular level with the evolution of a cytochrome P450 gene that plays a role in the detoxification of secondary plant compounds<sup>36</sup>. Some mutations can bypass the toxic defences of certain plants, providing survival and diversification on certain plants (and not others). Further studies have shown how changes in the use of host plants are associated with changes in the sequence, structure and function of P450. Results provide evidence that new P450 copies can appear for herbivores that colonize new hosts, supporting the hypothesis that interaction between herbivores and their host plants contributed to P450–gene diversification<sup>37</sup>. These studies provide convincing examples of host–plant shifts that may result in increased net diversification rate<sup>18,34</sup> and specific changes in key genes that confer new abilities to feed on toxic plants<sup>36–38</sup>.

Here, we study the insect–plant interactions at macroevolutionary scale using genomic and diversification approaches within a phylogenetic context. Given the complexity of shifting to a new host plant, we can expect more widespread effects across the entire genome<sup>15,39,40</sup>, but this has remained difficult to demonstrate. Indeed, both comprehensive species-level phylogeny and genomic data are necessary to disentangle the origin of the arms race and to understand the underlying mechanisms of insect–plant interaction as a major driver of diversification. The swallowtail model offers a relevant opportunity to better understand the role played by ecological interactions over the long timescales shaping the astonishing diversity of herbivores<sup>41</sup>.

## Results and discussion

**Co-phylogenetic history of an insect–plant antagonistic interaction.** First, we created an extensive phylogenetic dataset including seven genetic markers for 71% of swallowtail species diversity (408 of ~570 described species, see ‘Methods’). This dataset leads to the assembly of the most complete and well-resolved dated phylogeny of swallowtail butterflies (79% of nodes with strong bootstrap support defined as  $\geq 95\%$ ; Supplementary Figs. 1–3). Both tribe- and genus-level relationships are mostly consistent with previous results using multilocus datasets<sup>18,34,35,42–45</sup>. However, our species tree benefits from a phylogenomic backbone that we recently inferred at the genus level for the Papilionidae using genome-scale data<sup>46</sup>. Second, we compiled host–plant preferences for each swallowtail species in the dataset, and we performed ancestral-state estimations (see ‘Methods’). Phylogenetic estimates of ancestral host–plant preferences indicate that Aristolochiaceae were either the food plant of ancestral Papilionidae<sup>34</sup> or were colonized twice<sup>35</sup>, suggesting



**Host plants**

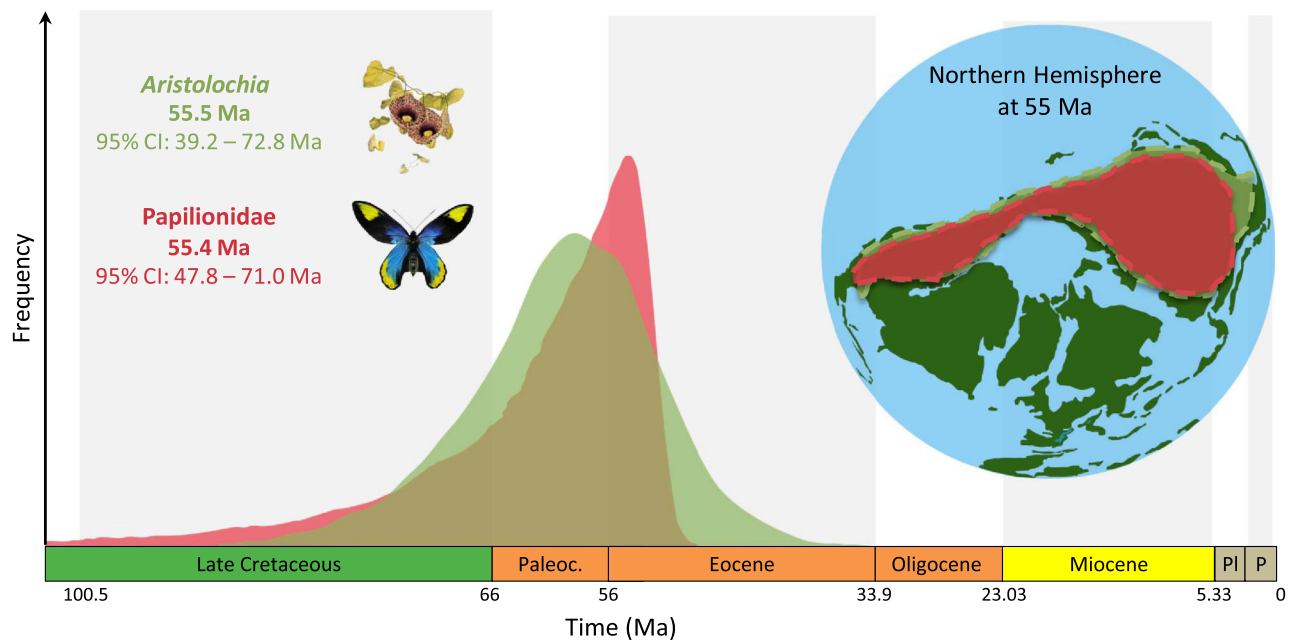
- Aristolochiaceae
- Rutaceae
- Annonaceae
- Crassulaceae + Saxifragaceae
- Papaveraceae
- Lauraceae
- Apiaceae
- Rosaceae
- Hernandiaceae
- Magnoliaceae
- Zygophyllaceae
- Fabaceae (*Acacia*)

**Fig. 1 Evolution of host-plant association through time shows strong host-plant conservatism across swallowtail butterflies.** Phylogenetic relationships of swallowtail butterflies, with coloured branches mapping the evolution of host-plant association, as inferred by a maximum-likelihood model (Supplementary Figs. 4 and 6). Additional analyses with two other maximum-likelihood and Bayesian models inferred the same host-plant associations across the phylogeny (Supplementary Fig. 5). Lue. Luehedorfiini, Zerynth. Zerynthiini, T. Teinopalpini. Pictures of butterflies made by Fabien Condamine.

an ancient and highly conserved association with Aristolochiaceae throughout swallowtail butterflies evolution. Using this robust time-calibrated phylogeny (Supplementary Figs. 1–3), we have traced the evolutionary history of food-plant use and infer that the family Aristolochiaceae was the ancestral host for Papilionidae (Fig. 1; relative probabilities = 0.915, 0.789 and 0.787 with three models; Supplementary Figs. 4 and 5). We further show that the genus *Aristolochia* was the ancestral host plant, as almost all Aristolochiaceae-associated swallowtails feed on *Aristolochia* (Supplementary Fig. 6). Across the swallowtail phylogeny, we recover only 14 host-plant shifts at the plant family level (14 nodes out of 407; Supplementary Figs. 4 and 5), suggesting strong evolutionary host-plant conservatism.

With the ancestor of swallowtails feeding on birthworts, evidence for synchronous temporal and geographical origins

further links the genus *Aristolochia* and the family Papilionidae and supports the escape-and-radiate model. Reconstructions of co-phylogenetic history for other insect–plant antagonistic interactions have shown either synchronous diversification<sup>11</sup> or herbivore diversification lagging behind that of their host plants<sup>10,47</sup>. We assembled a molecular dataset for ~49% of the species diversity of Aristolochiaceae (247 of ~502 described species; see ‘Methods’) and reconstructed their phylogeny (Supplementary Fig. 7), which is in agreement with previous works<sup>48–52</sup>. Divergence time estimates strongly suggest synchronous radiations of Papilionidae (55.4 million years ago [Ma], 95% credibility intervals (CIs): 47.8–71.0 Ma) and *Aristolochia* (55.5 Ma, 95% CIs: 39.2–72.8 Ma) since the early Eocene (Fig. 2 and Supplementary Figs. 3, 8 and 9). This result is robust to known biases in inferring divergence times, with slightly older ages

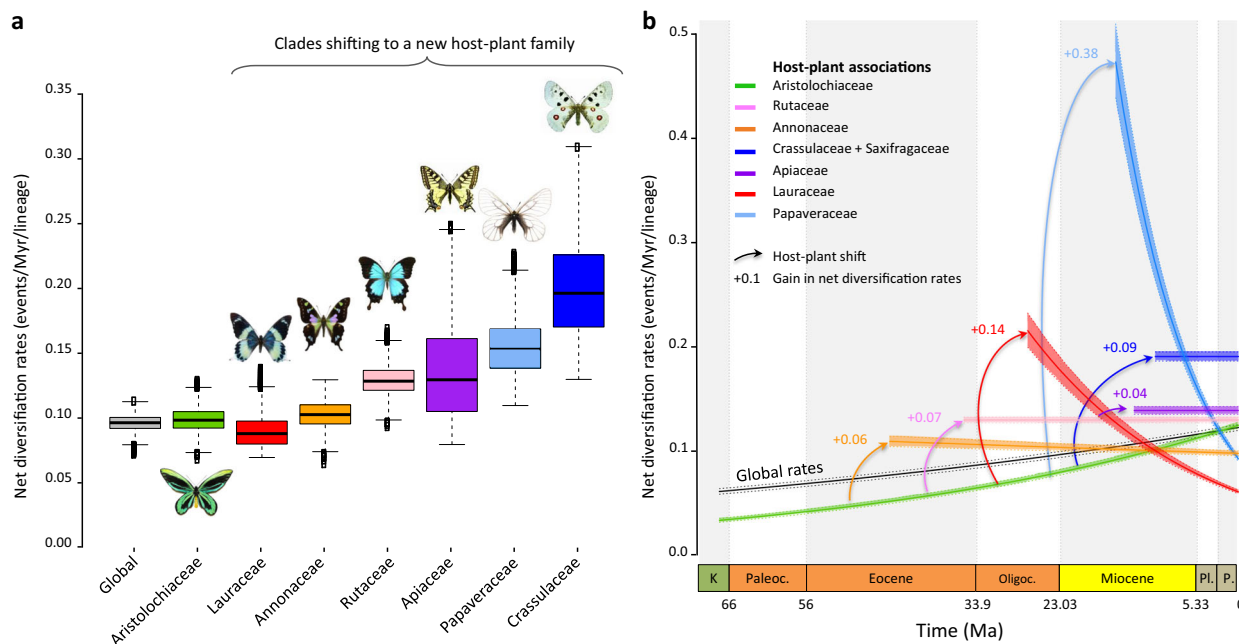


**Fig. 2 Synchronous temporal and geographic origin for swallowtails and birthworts.** Bayesian molecular divergence times with exponential priors estimate an early Eocene origin (~55 Ma) for both swallowtails and *Aristolochia* (alternatively, analyses with a uniform prior estimated an origin around 67 Ma for swallowtails and 64 Ma for *Aristolochia*; Supplementary Figs. 3, 8 and 9). Biogeographical maximum-likelihood models infer an ancestral area of origin comprising West Nearctic, East Palearctic and Central America for both swallowtails and birthworts (Supplementary Figs. 10 and 11). Paleoc Paleocene, PI Pliocene, P Pleistocene, Ma million years ago. Pictures of the plant and butterfly made by Fabien Condamine, and the world map made by Rémi Allio.

inferred for both groups when using more conservative priors on clade ages (Supplementary Fig. 9). Such temporal congruence between *Aristolochia* and Papilionidae raises the question of whether both clades had similar geographical origins and dispersal routes. To characterize the macroevolutionary patterns of the *Aristolochia*/Papilionidae arms race in space, we assembled two datasets of current geographic distributions for all species included in the phylogenies of both Aristolochiaceae and Papilionidae. We reconstructed the historical biogeography of both groups, taking into account palaeogeographical events throughout the Cenozoic (see ‘Methods’). Along with the known fossil record of both groups<sup>53–57</sup>, these results suggest that both Papilionidae and *Aristolochia* were ancestrally co-distributed throughout a region, including West Nearctic (WN), East Palearctic (EP), and Central America (CA) in the early Eocene, when Asia and North America were connected by the Bering land bridge (Fig. 2 and Supplementary Figs. 10 and 11). This combination of close temporal and spatial congruence provides strong evidence that Papilionidae and *Aristolochia* diversified concurrently through time and space until several swallowtail lineages shifted to the new host-plant families in the middle Eocene.

*Host-plant shifts confer higher rates of diversification.* Our ancestral-state estimates and biogeographic analyses are consistent with a sustained arms race between *Aristolochia* and Papilionidae in the past 55 million years. According to the escape-and-radiate model, a host-plant shift should confer higher rates of species diversification for herbivores through the acquisition of novel resources to radiate into<sup>5,6</sup> and/or the lack of competitors (Aristolochiaceae-feeder swallowtails have almost no competitors<sup>31</sup>). We tested the hypothesis that increases in diversification rates occurred in swallowtail lineages that shifted to new host plants. Given the uncertainty surrounding the inferences of macroevolutionary rates from phylogenies of extant species, we

applied a suite of birth–death models to cross-validate the estimated rates of diversification LASER (Likelihood Analysis of Speciation and Extinction Rates), MuSSE (Multiple State Speciation Extinction), RPANDA (R: Phylogenetic ANALyses of Diversification), BAMB (Bayesian analysis of macroevolutionary mixtures), CoMET (CPP on Mass-Extinction Times) and RevBayes; see ‘Methods’). We find evidence for (1) increases of diversification at host–plant shifts with trait-dependent birth–death models (as inferred with: MuSSE, Fig. 3a and Supplementary Fig. 12; RPANDA, Supplementary Fig. 13; and LASER, Supplementary Table 1) and (2) host–plant shifts contributing to a global increase through time with clade- and time-dependent birth–death models (as inferred with: RPANDA, Fig. 3b and Supplementary Fig. 13; BAMB, Supplementary Fig. 14; RevBayes, Supplementary Fig. 15; and CoMET, Supplementary Fig. 16). Although we should be cautious about the estimations of macroevolutionary rates<sup>58–63</sup>, all models concur that diversification rates increase through time either globally or due to recurrent host–plant shifts. Interestingly, these results contrast with the slowdown of diversification that is classically recovered in most phylogenies, often attributed to ecological limits and niche filling processes<sup>63</sup>. This sustained and increasing diversification during the Cenozoic may be explained by ecological opportunities not decreasing, due to a steady increase in host breadth for Papilionidae with new host–plant families colonized through time (Supplementary Fig. 17). Opening up new niches, which can also expand due to diversification increases of the host–plant families through time<sup>64–66</sup>, would allow a continuous increase in diversification rates through time in a dynamic biotic environment, lending support to the primary role of ecological interactions in clade diversification over long timescales—a long-contentious issue<sup>29</sup>. Nonetheless, when taking into account the possibility that rates may have been heterogeneous across the phylogeny, we find that the diversification of three lineages (those feeding on Annonaceae, Lauraceae and Papaveraceae) had early



**Fig. 3** Host-plant shifts lead to repeated bursts in diversification rates and a sustained overall increase in diversification through time. **a** Diversification tends to be higher for clades shifting to new host plants, as estimated by trait-dependent diversification models. Boxplots represent Bayesian estimates of net diversification rates for clades feeding on particular host plants (see also Supplementary Fig. 12). **b** A global increase in diversification is recovered with birth-death models estimating time-dependent diversification (see also Supplementary Figs. 14 and 15). Taking into account rate heterogeneity by estimating host-plant and clade-specific diversification indicates positive gains of net diversification after shifting to new host plants (see also Supplementary Fig. 13). K Cretaceous, Paleoc. Paleocene, Oligoc. Oligocene, Pl. Pliocene, P Pleistocene, Ma million years ago. Pictures of butterflies made by Fabien Condamine.

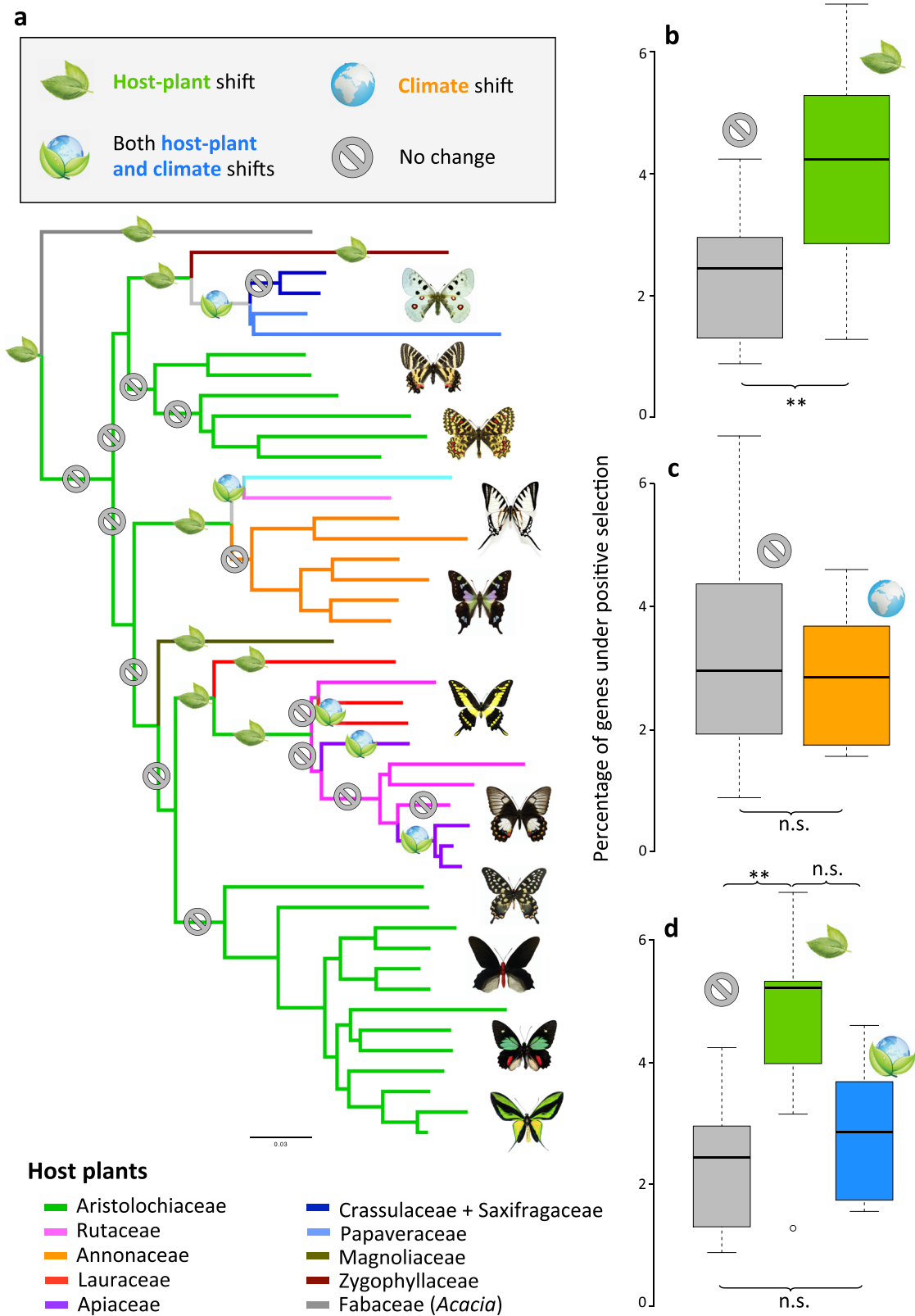
rates of speciation that are higher than the ancestral rates, but slowed down through time.

Interestingly, not all host-plant shifts led to evolutionary success in terms of extant species diversity. Given our rate estimations, we found significantly lower diversification rates than the rates on the ancestral host-plant Aristolochiaceae for three host-plant shifts (to Fabaceae, Magnoliaceae and to Zygophyllaceae; Supplementary Fig. 12). Altogether, these three host switches correspond to a very low proportion (~1%) of the total swallowtail diversity today. Indeed, a single species (*Baronia brevicornis*) feeds on the Fabaceae, the genus *Hypermnestra* (two species) feeds on Zygophyllaceae and the genus *Teinopalpus* (two species) feeds on Magnoliaceae. Hence, these are unsuccessful host-plant shifts from an evolutionary perspective (i.e. evolutionary dead-ends).

**Genome-wide adaptations to host-plant shifts.** Key innovations are often considered to underlie ecological opportunities and/or evolutionary success<sup>67</sup>, particularly in the case of chemically mediated interactions between butterflies and their host plants<sup>13</sup>. Studies on Papilionidae have provided strong examples of specific changes in key genes that confer new abilities to feed on toxic plants and allow host-plant shifts<sup>36,37</sup>. Adaptations of swallowtails to their hosts have particularly been assessed through the study of cytochrome P450 monooxygenases (P450s), which have a major role in detoxifying secondary plant compounds. New P450s appear to arise in swallowtails that colonize new hosts to bypass toxic defences, providing survival and diversification on some but not all plants<sup>15,36,37</sup>. This supports the hypothesis that insect-plant interactions contributed to P450-gene family diversification, with P450s being key innovations that explain the evolutionary and ecological success of phytophagous insects<sup>14,15,36,38,68,69</sup>. However, host-plant shifts not only alter

single genes but may also influence unlinked genes<sup>40</sup>. Moreover, host-plant shifts can accompany changes of the abiotic environment, which may, in turn, require further biotic adaptation (new predators and/or competitors). But the macroevolutionary and genomic consequences of the evolutionary dynamics of host-plant shifts have not yet been demonstrated.

Relying on a genomic dataset comprising 45 genomes covering all swallowtail genera<sup>46,70-72</sup>, we constructed two specific datasets (Dataset 1: 520 genes and Dataset 2: 1533 genes; mean gene coverage = 26.7x; see ‘Methods’ and Supplementary Data 1). To test whether there are any genomic signatures of positive selection caused by host-plant shifts within swallowtails, we performed a comparative genomic survey of molecular adaptation between swallowtail lineages that shifted to new host plants compared to non-shifting lineages (see ‘Methods’). We selected 14 phylogenetic branches representing a host-plant shift and 14 phylogenetic branches with no change as negative controls<sup>73,74</sup> (Fig. 4a). For a fair molecular comparison, each branch selected as a negative control was chosen to be as close as possible to a test branch representing a host-plant shift (i.e. sister groups; Supplementary Fig. 18). Among branches with host-plant shifts, five branches also had a shift in climate preference (represented by distributional changes from tropical to temperate conditions). Using a maximum-likelihood (ML) method, we estimated the ratio of non-synonymous substitutions (dN) over synonymous substitutions (dS) in all branches where a host-plant shift was identified relative to branches with no host-plant shift<sup>75,76</sup> (see ‘Methods’). The dN/dS analyses on branches with host-plant shifts (combined or not with environmental shifts) showed more genes with a subset of codons evolving under positive selection (dN/dS > 1) in lineages shifting to a new plant family, although the difference was marginally non-significant for the smallest dataset and highly significant for the second dataset containing more genes (Fig. 4b, Supplementary Fig. S19 and Supplementary



**Fig. 4 Host-plant shifts promote higher molecular adaptations.** **a** Genus-level phylogenomic tree displaying branches with and without host-plant shifts, on which genome-wide analyses of molecular evolution are performed. **b** Number of genes under positive selection ( $dN/dS > 1$ ) for swallowtail lineages shifting to new host-plant families ( $n = 14$ , green) or not ( $n = 14$ , grey). **c** Number of genes under positive selection for swallowtail lineages undergoing climate shifts ( $n = 5$ , orange) or not ( $n = 23$ , grey). **d** Number of genes under positive selection for swallowtail lineages shifting to new host plants ( $n = 9$ , green), shifting both host-plant and climate ( $n = 5$ , blue) or not ( $n = 14$ , grey). The proportion of genes was estimated with Dataset 2 (1533 genes, see Supplementary Fig. 19 for the results with Dataset 1 and 520 genes). This demonstrates genome-wide signatures of adaptations in swallowtail lineages shifting to new host-plant families. Genes under positive selection did not contain over- or under-represented functional GO categories (Supplementary Data 2). Wilcoxon rank-sum test: n.s. = not significant ( $P > 0.05$ ),  $*P \leq 0.05$ ,  $**P \leq 0.01$ . Pictures and icons made by Fabien Condamine.

Table 2,  $P = 0.0501/0.0079$  for the two datasets, respectively, Wilcoxon rank-sum test, see ‘Methods’ for the definition of the datasets). However, dN/dS analyses on branches with environmental shifts indicated a balanced number of genes under positive selection (Fig. 4c, Supplementary Fig. S19 and Supplementary Table 2,  $P = 0.336/0.8162$  for the two datasets, respectively, Wilcoxon rank-sum test), suggesting a lower impact of environmental shifts than host–plant shifts. We then performed dN/dS analyses for branches with host–plant shifts only (not followed by environmental shifts) and found that swallowtail lineages shifting to a new host–plant family had significantly more genes under positive selection (4.41%/3.98% of genes under positive selection for the two datasets, respectively; Supplementary Table 2) than non-shifting lineages (3.02%/2.43% of genes under positive selection for the two datasets, respectively, Fig. 4d, Supplementary Fig. S19 and Supplementary Table 2,  $P = 0.0071/0.00156$  for the two datasets, respectively, Wilcoxon rank-sum test). Surprisingly, the dual changes in environment and host–plant preferences did not spur molecular adaptation across swallowtail lineages compared to control branches ( $P = 1/0.4439$  for the two datasets, respectively, Wilcoxon rank-sum test; Fig. 4d, Supplementary Fig. S19 and Supplementary Table 2). Comparing the proportion of genes under positive selection between the branches with dual changes and branches with host–plant shifts only shows a marginally significant difference with Dataset 1 and no difference with Dataset 2 ( $P = 0.0327/0.1471$  for the two datasets, respectively, Wilcoxon rank-sum test; Fig. 4d and Supplementary Fig. S19). However, this result might be an artefact due to the use of a few branches to perform the statistical comparison. Although we did not control for the effect of multi-nucleotide mutations<sup>77</sup>, which should affect dN/dS analyses equally for control and host–plant shift branches, we checked individually the gene alignments and performed sensitivity analyses, which showed that our results are not driven either by an excess of misaligned regions or missing data and GC-content variations among species (see ‘Methods’ and Supplementary Figs. 20–26). Finally, given that fixing the topology for CodeML (see ‘Methods’) can spuriously inflate substitution rates on some branches<sup>78</sup>, we computed the proportion of genes under positive selection by selecting the gene trees from the largest dataset (Dataset 2) for which the focal branches were recovered (in agreement with the species tree). These analyses confirmed the previous results suggesting more genes under positive selection during host–plant shifts ( $P = 0.0444$ , Wilcoxon rank test; Supplementary Table 2).

We further studied the functional categories of positively selected genes by using gene ontology (GO) analyses (PANTHER and EggNOG; see ‘Methods’). Applied to the high-quality genomes of *Papilio xuthus*<sup>71</sup> and *Heliconius melpomene*<sup>79</sup>, we found that ~70% of the genes could be associated with a gene function and ~30% lacked annotation, which suggests a gap of knowledge in the current insect database of gene function. Among the annotated genes, we found that genes under positive selection along branches with host shifts did not contain over- or under-represented functional GO categories: 252 out of 1213 GO categories represented by genes under positive selection ( $P > 0.05$ , Fisher’s exact test after false discovery rate correction; Supplementary Data 2). These results support the hypothesis that genome-wide signatures of adaptations are associated with host–plant shifts, and encourage enlarging the hypothesis that changes in only one or a few candidate gene families could be enough to act as key innovations for adaptation to new resources<sup>13,17</sup>. Despite a weak signal, it is striking that host–plant shifts left stronger genome-wide signatures than were associated with changing climate preferences. This result further suggests that the success of phytophagous insects involved widespread adaptations to biotic interactions than for shifts in the abiotic environment.

To conclude, establishing evolutionary links between ecological adaptations, genomic changes and species diversification over geological timescales remains a tremendous challenge<sup>28,80,81</sup> with, for instance, important limitations due to the lack of knowledge in functional gene annotations in insects. However, the successful development of powerful analytical tools in conjunction with the increasing availability of insect genomes and improvements in genomic analyses<sup>82</sup> have allowed the detection of more genes than those already known to be involved in detoxification pathways playing a role in long-term relationships between plants and insects. Our genome-wide analyses have also generated a list of candidate genes potentially involved in plant–insect interactions. This opens new research avenues for finding the functionality of genes potentially linked with the adaptation and diversification of phytophagous insects. We hope that our study will help move in that direction, and that it will provide perspectives for future investigations of other model groups.

Over a half-century ago, Ehrlich and Raven<sup>5</sup> proposed that insect–plant interactions driven by diffuse coevolution over long evolutionary periods can be a major source of terrestrial biodiversity. Applied to a widely appreciated case in the insect–plant interactions theory, our study has been able to investigate genome-wide adaptive processes and corresponding macroevolutionary consequences in a comprehensive framework, suggesting that more genes could be involved in host–plant shifts than previously studied in the diversification of herbivorous insects. This result confirms the general belief in the insect–plant community that host–plant shifts are complex and would thus require a number of adaptations, which likely affect various genes beyond those directly linked to detoxification of the plant compounds<sup>36,39,40</sup>. By expanding the possible genes and gene families and identifying more adaptations than those gene families in detoxification pathways that were detected through antagonist interactions<sup>39</sup>, we show genomically wide-ranging co-evolutionary consequences<sup>40,83</sup> for close relationships between insects and their larval host plants. Hence, genome-wide macroevolutionary consequences of key adaptations in new insect–plant interactions may be a general feature of the co-evolutionary interactions that have generated Earth’s diversity.

## Methods

**Time-calibrated phylogeny of Papilionidae.** We assembled a supermatrix dataset with available data extracted from GenBank as of May 2017 (most of which has been generated by our research group), using five mitochondrial genes (*COI*, *COII*, *NDI*, *ND5* and *rRNA 16S*) and two nuclear markers (*EF-1a* and *Wg*) for 408 Papilionidae species (~71% of the total species diversity) and 20 outgroup species. We aligned the DNA sequences for each gene using MAFFT 7.110<sup>84</sup> with default settings (E-INS-i algorithm), and the alignments were checked for codon stops and eventually refined by eye with Mesquite 3.1 (available at: [www.mesquiteproject.org](http://www.mesquiteproject.org)). The best-fit partitioning schemes and substitution models for phylogenetic analyses were determined with PartitionFinder 2.1.1<sup>85</sup> using the greedy search algorithm and the Bayesian Information Criterion. All gene alignments were concatenated in a supermatrix, which is available in Figshare (see Data availability).

Phylogenetic relationships were estimated with both ML and Bayesian inference. ML analyses were carried out with IQ-TREE 1.6.8<sup>86</sup>. We set the best-fit partitioning scheme (-ssp option) and used ModelFinder to determine the best-fit substitution model for each partition<sup>87</sup> and then estimated model parameters separately for every partition<sup>88</sup> such that all partitions shared the same set of branch lengths, but we allowed each partition to have its own evolution rate (-m TESTNEW option). For tree search parameters, we relied on a more thorough and slower nearest-neighbour interchange search to consider all possible nearest-neighbour interchanges instead of only those in the vicinity previously applied (-allnni option). Following the recommendation of IQ-TREE developers, we also set smaller perturbation strength (-pers 0.2) and a larger number of stop iterations (-nstop 500) to avoid local optima. We performed 2000 ultrafast bootstrap replicates to investigate nodal support across the topology, considering values  $\geq 95$  as strongly supported nodes<sup>89</sup>.

Estimating phylogenetic relationships for such a dataset is computationally intensive with Bayesian inference. The ML tree inferred with IQ-TREE was used as a starting tree for Bayesian inference as implemented in MrBayes 3.2.6<sup>90</sup>. Rather than using a single substitution model per molecular partition, we sampled across

the entire substitution-model space<sup>91</sup> using reversible-jump Markov Chain Monte Carlo (rj-MCMC). Two independent analyses with one cold chain and seven heated chains, each run for 50 million generations, sampled every 5000 generations. Convergence and performance of Bayesian runs were evaluated using Tracer 1.7.1<sup>92</sup>, the average deviation of split frequencies (ADSFs) between runs, the effective sample size (ESS) and the potential scale reduction factor (PSRF) values for each parameter. The runs had to have values of ADSF approaching zero, PSRF close to 1.0 and ESS >200 to be considered convergent. A 50% majority-rule consensus tree was built after conservatively discarding 25% of sampled trees as burn-in. Node support was evaluated with posterior probability considering values  $\geq 0.95$  as strong support<sup>93</sup>. All analyses were performed on the CIPRES Science Gateway computer cluster<sup>94</sup>, using BEAGLE<sup>95</sup>.

Dating inferences were performed using Bayesian relaxed-clock methods accounting for rate variation across lineages<sup>96</sup>. MCMC analyses implemented in BEAST 1.8.4<sup>97</sup> were employed to approximate the posterior distribution of rates and divergence times and infer their CIs. Estimation of divergence times relied on constraining clade ages through fossil calibrations. Swallowtail fossils are scarce, but five can unambiguously be attributed to the family. The oldest fossil occurrences of Papilionidae are the fossils †*Praepapilio colorado* and †*Praepapilio gracilis*<sup>53</sup>, both from the Green River Formation (Colorado, USA). The Green River Formation encompasses a 5 million years period between ~48.5 and 53.5 Ma, which falls within the Ypresian (47.8–56 Ma) in the early Eocene<sup>98</sup>. These fossils can be phylogenetically placed at the crown of the family as they share synapomorphies with all extant subfamilies<sup>55,99</sup>, and have proven to be reliable calibration points for the crown group<sup>18,34,46</sup>. Two other fossils belong to Parnassiinae, whose systematic position was assessed using phylogenetic analyses based on both morphological and molecular data in a total-evidence approach<sup>18</sup>. The first is †*Thaites ruminiana*<sup>100</sup>, a compression fossil from limestone in the Niveau du gypse d'Aix Formation of France (Bouches-du-Rhône, Aix-en-Provence, France) within the Chattian (23.03–28.1 Ma) of the late Oligocene<sup>54,101</sup>. †*Thaites* is sister to Parnassiini, and occasionally sister to Luehdorfiini + Zerynthiini<sup>18</sup>. Thus, we constrained the crown age of Parnassiinae with a uniform distribution bounded by a minimum age of 23.03 Ma. The second is †*Dorittes bosniakii*<sup>102</sup>, an exoskeleton and compression fossil from Italy (Tuscany) from the Messinian (5.33–7.25 Ma, late Miocene)<sup>54</sup>. †*Dorittes* is sister to *Archon* (Luehdorfiini<sup>18</sup>), in agreement with Carpenter<sup>103</sup>. The crown of Luehdorfiini was thus constrained for divergence time estimation using a uniform distribution bounded with 5.33 Ma. Absolute ages of geological formations were taken from the latest update of the geological time scale.

We used a conservative approach to apply calibration priors with the selected fossil constraints by setting uniform priors bounded with a minimum age equal to the youngest age of the geological formation where each fossil was found. All uniform calibration priors were set with an upper bound equal to the estimated age of angiosperms (150 Ma<sup>104</sup>), which is more than three times older than the oldest Papilionidae fossil. This upper age is intentionally set as ancient to allow exploration of potentially old ages for the clade. Since the fossil record of butterflies is incomplete and biased<sup>105</sup>, caution is needed in using these fossil calibrations (effect shown in burying beetles<sup>106</sup>).

After enforcing the fossil calibrations, we set the following settings and priors: a partitioned dataset (after the best-fitting PartitionFinder scheme) was analysed using the uncorrelated log-normal distribution clock model, with the mean set to a uniform prior between 0 and 1, and an exponential prior ( $\lambda = 0.333$ ) for the standard deviation. The branching process prior was set to a birth–death<sup>107</sup> process, using the following uniform priors: the birth–death mean growth rate ranged between 0 and 10 with a starting value at 0.1, and the birth–death relative death rate ranged between 0 and 1 (starting value = 0.5). We performed four independent BEAST analyses for 100 million generations, sampled every 10,000th, resulting in 10,000 samples in the posterior distribution, of which the first 2500 samples were discarded as burn-in. All analyses were performed on the CIPRES Science Gateway computer cluster<sup>94</sup>, using BEAGLE<sup>95</sup>. Convergence and performance of each MCMC run were evaluated using Tracer 1.7.1<sup>92</sup> and the ESS for each parameter (ESS > 200). We combined the four runs using LogCombiner 1.8.4<sup>97</sup>. A maximum-clade credibility (MCC) tree was reconstructed, with median ages and 95% CIs. The BEAST files generated for this study are available in Figshare (see Data availability).

**Estimating ancestral host–plant association.** We inferred the temporal evolution of host–plant association up to the ancestral host plant(s) at the root of Papilionidae using three approaches: the ML implementation of the Markov k-state (Mk) model<sup>108</sup>, the ML Dispersal-Extinction-Cladogenesis (DEC) model<sup>109</sup>, and the Bayesian approach in BayesTraits<sup>110</sup>. These approaches require a time-calibrated tree and a matrix of character states (current host–plant preference) for each species in the tree. An extensive bibliographic survey was conducted to obtain primary larval host plants at the family level<sup>5,30,111–113</sup>. The host associations of species were categorized using the following 12 character states: (1) Annonaceae, (2) Apiaceae, (3) Aristolochiaceae, (4) Crassulaceae or Saxifragaceae (core Saxifragales), (5) Fabaceae, (6) Hernandiaceae, (7) Lauraceae, (8) Magnoliaceae, (9) Papaveraceae, (10) Rosaceae, (11) Rutaceae, and (12) Zygophyllaceae. The host–plant matrix of Papilionidae is available in Figshare (see Data availability).

Ancestral states for host–plant association were first reconstructed using the Mk model (one rate for all transitions between states) allowing any host shift to be equally probable. The Mk model does not allow multiple states for a species. The few species that use multiple host families were thus scored with the most frequent host association. The Mk model was performed with Mesquite 3.1 (available at: [www.mesquiteproject.org](http://www.mesquiteproject.org)). To estimate the support of any one character state over another, the most likely state was selected according to a decision threshold, such that if the log likelihoods between two states differ by two log-likelihood units, the one with lower likelihood is rejected<sup>108</sup>.

The DEC model was also used to reconstruct ancestral host–plant states<sup>109,114</sup>. As with the Mk model, we assumed that host–plant shifts occurred at equivalent probabilities between plant families and through time, which may not be true given that the host–plant families of Papilionidae did not originate at the same time (e.g. Aristolochiaceae originated ~108.07 Ma [95% CIs: 81.01–132.66 Ma]<sup>115</sup>, and Annonaceae originated ~98.94 Ma [95% CIs: 84.78–113.70 Ma]<sup>115</sup>). We used the estimated molecular ages of the different host–plant groups to constrain our inferences of ancestral host plants a posteriori. We preferred such an approach compared to a more constrained one in which the DEC model is informed with a matrix of host–plant appearances based on their estimated ages by implementing matrices of presence/absence of the character states through time (equivalent to the time-stratified palaeogeographic model, see below for inference of biogeographical history).

Finally, the Bayesian approach implemented in BayesTraits 3.0.1<sup>110</sup> was performed to provide a cross-validation of ML analyses. This approach automatically detects shifts in rates of evolution for multistate data using rj-MCMC. The number of parameters and priors was set by default. We ran the rj-MCMC for ten million generations and sampled states and parameters every 1000 generations (burn-in of 10,000 generations). We specifically estimated ancestral states at 21 nodes as well as at the root of Papilionidae. For this analysis, we used a set of 100 trees randomly taken from the dating analysis to probe the robustness of our ancestral-state estimation across topological uncertainty.

The results of these inferences determined the host–plant family(ies) that was (were) the most likely ancestral host(s) at the origin of Papilionidae, indicating (1) which plant phylogeny to reconstruct for studying the macroevolution of the arms race, and (2) the evolution of ancestral host–plant association along the phylogeny to identify the tree branches where shifts occurred and test for genome-wide changes.

The Mk and BayesTraits models always inferred with high support (relative probability = 0.915 and 0.789, respectively) that Aristolochiaceae is the ancestral host plant at the crown of Papilionidae. With the unconstrained DEC model, we found that the ancestral host–plant preference for Papilionidae was always composed of Aristolochiaceae, but also included another family (either Fabaceae, Hernandiaceae or Zygophyllaceae, which are only fed upon by *Baronia*, *Lamproptera* and *Hypermnestra*, respectively). As the sister lineage to all other Papilionidae, *Baronia* is the only species that feeds on Fabaceae. More precisely, only one species of Fabaceae is consumed: *Vachellia cochliacantha* (formerly *Acacia cochliacantha*; recent changes in *Acacia* taxonomy<sup>116</sup>). However, *Vachellia* diverged from its sister clade in the Eocene, ~50 Ma, and diversified in the Miocene between 13 and 17 Ma<sup>117</sup>, which substantially postdate the origin of Papilionidae. Therefore, this result suggests that the family Aristolochiaceae represents the most likely candidate as the ancestral host plant of Papilionidae. Hernandiaceae are consumed by *Lamproptera* (occasionally by *Papilio homerus*, *Graphium codrus*, *G. doson* and *G. empedovana*<sup>113</sup>). More precisely, the host plants of *Lamproptera* belong to the genus *Illigera*. This plant genus diverged from its sister genus 48 Ma<sup>115</sup> and started diversifying 27 Ma<sup>118</sup>. The derived phylogenetic position of *Lamproptera* and the age of its use as a host plant make it very unlikely that Hernandiaceae could constitute the ancestral host plant for Papilionidae. Similarly, the family Zygophyllaceae is consumed by *Hypermnestra*, most specifically it feeds on the genus *Zygophyllum* in Central Asia. The genus *Zygophyllum* is not monophyletic, but Asian *Zygophyllum* appeared 19.6 Ma<sup>119</sup>. Applying the same rationale, we are able to discard Zygophyllaceae as a candidate ancestral host plant for Papilionidae. To further refine our ancestral host–plant estimates, we built a presence–absence matrix of plant families based on clade origins estimated in molecular dating studies. Thereby, the age of the different plants can be used to constrain the inference of ancestral host plants. Under such a constrained model, Aristolochiaceae is always recovered as the most likely ancestral host plant for Papilionidae. It is also interesting that almost all Aristolochiaceae feeders have *Aristolochia* as host plants, and tests to determine which genus of Aristolochiaceae was originally consumed by Papilionidae showed that it was *Aristolochia*.

**Time-calibrated phylogeny of the ancestral host: the Aristolochiaceae.** Estimation of ancestral host–plant relationships indicated that the family Aristolochiaceae was the ancestral host for Papilionidae. We refer to Aristolochiaceae in its traditional circumscription including the genera *Asarum*, *Saruma*, *Thottea* and *Aristolochia*. The Angiosperm Phylogeny Group<sup>120</sup> proposes that Aristolochiaceae also includes the holoparasitic genera *Hydnora* and *Prospanche* (Hydnoraceae), as well as the monotypic family Lactoridaceae from the Juan Fernandez Islands of Chile (*Lactoris fernandeziana*). The conclusion of Angiosperm Phylogeny Group (APG)<sup>120</sup> is based on an online survey<sup>121</sup> rather than on primary data and this is why we disagree with their argumentation as well as the resulting conclusion of APG given available



resilient primary molecular phylogenomic data. However, arguments based on morphology and anatomy<sup>122–125</sup>, genetics<sup>49,50,126–129</sup>, molecular divergence time<sup>115,129</sup>, and conservation considerations (Tod Stuessy, personal communication with S.W., July 2019) favour splitting them into four families: Aristolochiaceae (*Aristolochia* and *Thottea*), Asaraceae (*Asarum* and *Saruma*), Hydnoraceae (*Hydnora* and *Prosopanche*), and Lactoridaceae (*Lactoris*), collectively called the perianth-bearing Piperales. Therefore, we extracted and assembled a supermatrix dataset with available data from GenBank for the perianth-bearing Piperales and its sister lineage, the perianth-less Piperales including Saururaceae and Piperaceae (as of May 2017, most of which has been generated by our research group). We obtained four chloroplast genes (*matK*, *rbcL*, *trnL* and *trnL-trnF*) and one nuclear marker (*ITS*) for 247 species of perianth-bearing Piperales (~49% of the total species diversity<sup>130</sup>) and six outgroups from perianth-less Piperales. We could not include the two genera *Hydnora* and *Prosopanche* (Hydnoraceae) because available genetic data do not overlap those of perianth-bearing Piperales<sup>126,128,131,132</sup>. We applied the same analytical procedure that we did for Papilionidae. DNA sequences for each gene were aligned using MAFFT 7.110<sup>84</sup> with default settings (E-INS-i algorithm and Q-INS-I to take into account secondary structure). Resulting alignments were checked for codon stops and eventually refined by eye with Mesquite 3.1 (available at: [www.mesquiteproject.org](http://www.mesquiteproject.org)). The best-fit partitioning schemes and substitution models for phylogenetic analyses were determined with PartitionFinder 2.1.1<sup>85</sup>. All gene alignments were concatenated into a supermatrix; the final dataset is available in Figshare (see Data availability).

Phylogenetic relationships were estimated with Bayesian inference as implemented in MrBayes 3.2.6<sup>90</sup>. Rather than using a single substitution model per molecular partition, we sampled across the entire substitution-model space<sup>91</sup> using rj-MCMC. Two independent analyses with one cold chain and seven heated chains, each was run for 50 million generations, sampled every 5000 generations. Convergence and performance of Bayesian runs were evaluated using Tracer 1.7.1<sup>92</sup> and the ESS, ADSF and PSRF criteria. Once convergence was achieved, a 50% majority-rule consensus tree was built after discarding 25% of the sampled trees as burn-in.

Bayesian relaxed-clock methods were used that accounted for rate variation across lineages<sup>96</sup>. MCMC analyses implemented in BEAST 1.8.4<sup>97</sup> were employed to approximate the posterior distribution of rates and divergences times and infer their CIs. Estimation of divergence times relied on constraining clade ages through fossil calibrations. Three unambiguous fossils from perianth-bearing Piperales (*Aristolochiaceae sensu lato*), and one corresponding to the family Saururaceae were used. First, we relied on the fossil record of the monotypic family Lactoridaceae (*L. fernandeziana*)<sup>126,129</sup>, a shrub endemic to the cloud forest of the Juan Fernández Islands archipelago of Chile. The oldest pollen fossil for the group is †*Lactoripollenites africanus*<sup>133,134</sup> from the Turonian/Campanian (72.1–89.8 Ma) of the Orange Basin in South Africa. This fossil confers a minimum age of 72.1 Ma for the stem node of *L. fernandeziana*. Second, the oldest and only pollen record of the Aristolochiaceae was recently described from Late Cretaceous sediments of Siberia: †*Aristolochiacidites viluensis*<sup>56</sup> from the Timerdyakh Formation of the latest Campanian to earliest Maastrichtian (66–72.1 Ma) in the Vilui Basin (Russia). Because inaperturate pollen grains in combination with this unique exine configuration and fitting size can be observed in extant members of Aristolochiaceae, this fossil provides a minimum age of 66 Ma for the family. The third fossil belongs to the genus *Aristolochia* and described as †*Aristolochia austriaca*<sup>57</sup> from the Pannonian (late Miocene) in the Hollabrunn–Mistelbach Formation (Austria). Based on a thorough morphological leaf comparison, this fossil is assigned to a species group including *Aristolochia baetica* and *Aristolochia rotunda*, which then confers a minimum age of 7.25 Ma for the clade. Finally, we used the fossil †*Saururus tuckeri*<sup>135</sup> from the Princeton Chert of Princeton in British Columbia (Canada), which is part of the Princeton Group, Allenby Formation dated with stable isotopes to the middle Eocene<sup>136</sup>. This fossil has been phylogenetically placed as sister to extant *Saururus* species<sup>136</sup>, hence providing a minimum age of 44.3 Ma for the stem node of *Saururus*. Absolute ages of geological formations were taken from the latest update of the geological time scale.

We set the following settings and priors: a partitioned dataset (after the best-fitting PartitionFinder scheme) was analysed using the uncorrelated log-normal distribution clock model, with the mean set to a uniform prior between 0 and 1, and an exponential prior ( $\lambda = 0.333$ ) for the standard deviation. The branching process prior was set to a birth–death<sup>107</sup> process, using the following uniform priors: the birth–death mean growth rate ranged between 0 and 10 with a starting value at 0.1, and the birth–death relative death rate ranged between 0 and 1 (starting value = 0.5). We performed four independent BEAST analyses for 100 million generations, sampled every 10,000th, resulting in 10,000 samples in the posterior distribution of which the first 2500 samples were discarded as burn-in. All analyses were performed on the CIPRES Science Gateway computer cluster<sup>94</sup>, using BEAGLE<sup>95</sup>. Convergence and performance of each MCMC run were evaluated using Tracer 1.7.1<sup>92</sup> and the ESS for each parameter. We combined the four runs using LogCombiner 1.8.4<sup>97</sup>. The MCC tree was reconstructed with median age and 95% CI. The BEAST files generated for this study are available in Figshare (see Data availability).

**Dual biogeographic history of Papilionidae and Aristolochiaceae.** We estimated the ancestral area of origin and geographic range evolution for both clades using

the ML approach of DEC model<sup>109</sup> as implemented in the C++ version<sup>137,138</sup> that is available at: <https://github.com/champost/DECX>. To infer the biogeographic history of a clade, DEC requires a time-calibrated tree, the current distribution of each species for a set of geographic areas, and a time-stratified geographic model that is represented by connectivity matrices for specified time intervals spanning the entire evolutionary history of the group.

The geographic distribution for each species in Papilionidae<sup>30,112,113</sup> and Aristolochiaceae was categorized as present or absent in each of the following areas: (1) WN, (2) East Nearctic, (3) CA, (4) South America, (5) West Palaearctic, (6) EP, (7) Madagascar, (8) Indonesia and Wallacea, (9) India, (10) Africa and (11) Australasia. The resulting matrices of species distribution for the two groups are available in Figshare (see Data availability).

A time-stratified geographic model was built using connectivity matrices that take into account palaeogeographic changes through time, with time slices indicating the possibility or not for a species to access a new area<sup>138</sup>. Based on palaeogeographical reconstructions<sup>139–141</sup>, we created a connectivity matrix for each geological epoch that represented a period bounded by major changes in tectonic and climatic conditions thought to have affected the distribution of organisms. The following geological epochs were selected: (1) 0–5.33 Ma (Pliocene to present), (2) 5.33–23.03 Ma (Miocene), (3) 23.03–33.9 Ma (Oligocene), (4) 33.9–56 Ma (Eocene) and (5) 56 Ma to the origin of the clade (Palaeocene to Late Cretaceous). For each of these five time intervals, we specified constraints on area connectivity by coding 0 if any two areas are not connected or 1 if they are connected in a given time interval. We assumed a conservative dispersal matrix with equal dispersal rates between areas through time<sup>142</sup>.

**Impact of host–plant shifts on swallowtail diversification.** We tested the effect of host–plant association on diversification by estimating speciation and extinction rates with five methods to cross-test hypotheses and corroborate results. Analyses were performed on 100 dated trees randomly sampled from the Bayesian dating analyses to take into account the uncertainty in age estimates. We used the following approaches: (1) ML-based trait-dependent diversification<sup>143,144</sup>; (2) ML-based time-dependent diversification<sup>145</sup>; (3) Bayesian analysis of macroevolutionary mixture<sup>146</sup>; (4) Bayesian branch-specific diversification rates<sup>147</sup>; and (5) Bayesian episodic birth–death model<sup>148</sup>. It is worth mentioning that each method differs at several points in their estimation of speciation and extinction rates. For instance, trait-dependent birth–death models estimate constant speciation and extinction rates<sup>144</sup>, whereas time-dependent birth–death models estimate clade-specific speciation and extinction rates and their variation through time<sup>145,147</sup>. Therefore, we expect some differences in the values of estimated diversification rates that are inherent to each approach. Our diversification analyses should be seen as complementary to the inferred diversification trend rather than corroborating the values and magnitude of speciation and extinction rates.

First, we computed the probability of obtaining a clade as large as size  $n$ , given the crown age of origin, the overall net diversification rate of the family, and an extinction rate as a fraction of speciation rate following the approach in Condamine et al.<sup>34</sup> relying on the method of moments<sup>149</sup>. We used the R-package LASER 2.3<sup>150</sup> to estimate the net diversification rates of Papilionidae and six clades shifting to new host plants with the *bd.ms* function (providing crown age and total species diversity). Then, we used the *crown.limits* function to estimate the mean expected clade size for each clade shifting to new host plants given clades' crown age and overall net diversification rates, and we finally computed the probability to observe such clade size using the *crown.p* function. All rate estimates were calculated with three  $\epsilon$  values ( $\epsilon = 0/0.5/0.9$ ), knowing that the extinction rate in swallowtails is usually low<sup>34</sup> (supported by the results of this study).

Second, we relied on the state-dependent speciation and extinction (SSE) model, in which speciation and extinction rates are associated with phenotypic evolution of a trait along a phylogeny<sup>143</sup>. In particular, we used the MuSSE<sup>144</sup> implemented in the R-package *diversitree* 0.9–10<sup>151</sup>, which allows multiple character states to be studied. Larval host–plant data were taken from previous works<sup>5,18,30,34,112,113,152</sup>. The following ten host–plant character states and corresponding ratios of sampled species in the tree of all known species for each character (sampling fractions) were used: 1 = Aristolochiaceae (110/152), 2 = Annonaceae (69/138), 3 = Lauraceae (33/39), 4 = Apiaceae (9/10), 5 = Rutaceae (119/163), 6 = Crassulaceae (19/19), 7 = Papaveraceae (44/44), 8 = Fabaceae (1/1), 9 = Zygophyllaceae (2/2), and 10 = Magnoliaceae (2/2). Data at a lower taxonomic level than plant family were not used because of the large number of multiple associations exhibited by genera that could alter the phylogenetic signal. We assigned a single state to each species by selecting the food plant with the maximum number of collections for each species. We did not employ multiple states per species, which represents a lesser problem because (1) few swallowtail species feed on multiple plant families, (2) current shared-state models can only model two states, and (3) the addition of multi-plant states to the MuSSE analysis would have greatly increased the number of parameters. We performed both ML and Bayesian MCMC analyses (10,000 steps) performed using an exponential ( $1/(2 \times \text{net diversification rate})$ ) prior with starting parameter values obtained from the best-fitting ML model and resulting speciation, extinction and transition rates. After a burn-in of 500 steps, we estimated posterior density distribution for speciation, extinction and transition rates. There have been concerns about the power of SSE models to infer diversification dynamics from a distribution of

species traits<sup>153–155</sup>, hence other birth–death models were used to corroborate the results obtained with SSE models.

Third, to provide an independent assessment of the relationship between diversification rates and host specificity, we used the ML approach of Morlon et al.<sup>145</sup> implemented in the R-package *RPANDA* 1.3<sup>156</sup>. This is a birth–death method in which speciation and/or extinction rates may change continuously through time. This method has the advantage of not assuming a constant extinction rate over time (unlike BAMM<sup>146</sup>), and allows clades to have declining diversity since extinction can exceed speciation, meaning that diversification rates can be negative<sup>145</sup>. For each clade that shifted to a new host family, we designed and fitted six diversification models: (1) a Yule model, where speciation is constant and extinction is null; (2) a constant birth–death model, where speciation and extinction rates are constant; (3) a variable speciation rate model without extinction; (4) a variable speciation rate model with constant extinction; (5) a rate-constant speciation and variable extinction rate model; and (6) a model in which both speciation and extinction rates vary. Models were compared by computing the ML estimate of each model and the resulting Akaike information criterion corrected by sample size. We then plotted rates through time with the best-fit model for each clade, and the rates for the family as a whole for comparison purpose.

Fourth, we performed models that allow diversification rates to vary among clades across the whole phylogeny. BAMM 2.5<sup>146,157</sup> was used to explore for differential diversification dynamic regimes among clades differing in their host–plant feeding. BAMM can automatically detect rate shifts and sample distinct evolutionary dynamics that explain the diversification dynamics of a clade without a priori hypotheses on how many and where these shifts might occur. Evolutionary dynamics can involve time-variable diversification rates; in BAMM, speciation is allowed to vary exponentially through time while extinction is maintained constant: subclades in a tree may diversify faster (or slower) than others. This Bayesian approach can be useful in detecting shifts of diversification potentially associated with key innovations<sup>157</sup>. BAMM analyses were run with four MCMC for 20 million generations, sampling every 20,000th and with three different values (1, 5 and 10; Supplementary Table 3) of the compound Poisson prior (CPP) to ensure the posterior is independent of the prior<sup>158</sup>. We accounted for non-random incomplete taxon sampling using the implemented analytical correction; we set a sampling fraction per genus based on the known species diversity of each genus. Mixing and convergence among runs (ESS > 200 after 15% burn-in) were assessed with the R-package *BAMMtools* 2.1<sup>159</sup> to estimate (1) the mean global rates of diversification through time, (2) the estimated number of rate shifts evaluating alternative diversification models comparing priors and posterior probabilities and (3) the clade-specific rates through time when a distinct macroevolutionary regime is identified.

Fifth, BAMM has been criticized for incorrectly modelling rate shifts on extinct lineages, that is, unobserved (extinct or non-sampled) lineages inherit the ancestral diversification process and cannot experience subsequent diversification-rate shifts<sup>158,160</sup>. To solve this, we used a Bayesian approach implemented in RevBayes 1.0.10<sup>161</sup> that models rate shifts consistently on extinct lineages by using the SSE framework<sup>147,158</sup>. Although there is no information of rate shifts for unobserved/extinct lineages in a phylogeny including extant species only, these types of events must be accounted for in computing the likelihood. The number of rate categories is fixed in the analysis but RevBayes allows any number to be specified, thus allowing direct comparison of different macroevolutionary regimes.

Finally, we evaluated the impact of abrupt changes in diversification using the Bayesian episodic birth–death model of CoMET<sup>148</sup> implemented in the R-package *TESS* 2.1<sup>162</sup>. These models allow detection of discrete changes in speciation and extinction rates concurrently affecting all lineages in a tree, and estimate changes in diversification rates at discrete points in time, but can also infer mass extinction events (sampling events in which the extant diversity is reduced by a fraction<sup>163</sup>). Speciation and extinction rates can change at those points, but remain constant within time intervals. In addition, *TESS* uses independent CPPs to simultaneously detect mass extinction events and discrete changes in speciation and extinction rates, while *TreePar* estimates the magnitude and timing of speciation and extinction changes independently to the occurrence of mass extinctions (i.e. the three parameters cannot be estimated simultaneously due to parameter identifiability issues<sup>163</sup>). We performed two independent analyses allowing and disallowing mass extinction events. Bayes factor comparisons were used to assess the model fit between models with varying number and time of changes in speciation/extinction rates and mass extinctions.

**Detecting genome-wide adaptations during host–plant shifts.** We analysed genomic sequence data in swallowtail butterflies that have independently shifted to new ecological (biological) traits. Similar approaches have been conducted on mammals<sup>164,165</sup> and birds<sup>166</sup>, but have been rarely implemented on arthropod groups over such a long geological time scale. Here, we estimated swallowtail molecular evolution with whole-genome data and compared selection regimes on protein-coding genes along independent branches with or without host–plant shift and/or environmental shift.

For these analyses, we studied 45 whole genomes<sup>46</sup> covering all 32 genera of the family Papilionidae: 41 of which were previously generated by our research group added to four genomes already available<sup>70–72</sup>. In summary, raw reads (Sequence Read Archive: SRR8954507–SRR8954549) were cleaned using Trimmomatic 0.33<sup>167</sup>, and assembled into contigs and scaffolds with SOAPdenovo-63mer 2.04<sup>168</sup>

to obtain whole-genome assemblies (30× average read depth<sup>46</sup>). All coding DNA sequences (CDS) were retrieved from the high-quality annotated genome of *P. xuthus*<sup>71</sup>. To annotate the sequences of all our genomes, a BLAST search using all available CDS of *P. xuthus* was performed at the amino-acid level (using *tblastn*). For each species, the recovered genes were aligned one by one with *P. xuthus* using TranslatorX<sup>169</sup>. This method performs alignment at the amino-acid level and preserves the open reading frame. All sites showing intraspecific variation were set to N, to conservatively avoid false informative sites. Any contamination was removed using CroCo 0.1<sup>170</sup> and orthologous proteins were identified with OrthoFinder 2.2.0<sup>171</sup>. Finally, CDS alignments were strongly cleaned from misaligned sequences (gene by gene) using HMMCleaner 1.8<sup>172</sup>. A last cleaning step was performed using trimAl 1.2rev59<sup>173</sup>, which is designed to trim alignments for large-scale phylogenomic analyses. The resulting dataset comprised 6621 genes in at least four sampled species (median of 32% of missing data), which was used to reconstruct a robust phylogenomic tree of Papilionidae<sup>46</sup> (Supplementary Fig. 18).

We used this genomic dataset of 45 species representing all genera in which the resulting genus-level swallowtail phylogenomic tree<sup>46</sup> accurately represents the evolutionary associations with host plants as estimated using the ancestral-state analyses applied to the species-level phylogeny<sup>34</sup> (Fig. 1 and Supplementary Figs. 4 and 5). We thus transferred the inference of ancestral host–plant shifts on the phylogenomic tree and selected the branches representing a host–plant shift and/or a shift of climate preference (in general from tropical to temperate conditions; Supplementary Fig. 10). We also selected branches with no change as negative controls<sup>73</sup>. As a result, 14 branches are selected to measure the impact of a host–plant shift and 14 branches are selected as controls (Supplementary Fig. 18). Within these 14 branches with an ecological change, nine branches represent host–plant shifts only, and five branches correspond to shifts in both host plant and environment (from tropical to temperate conditions). To test the impact of these different changes on the genomes, two datasets were created, *Dataset 1* and *2*. Given the low quality of the genomes of *Allancastris cerisyi* and *Parnassius imperator*, these two genomes were discarded for the downstream analyses. We first selected the genes from the 6621-gene dataset for each focal branch using three criteria: (1) the dataset is composed only of orthologous protein-coding genes (OrthoFinder 2.2<sup>171</sup>), (2) the species needed to accurately define the branch were available (i.e. crown node of the clade) and (3) for each branch, one species per tribe was available, and therefore include a different number of genes per branch. Thus, for *Dataset 1*, only the genes containing sequences for the species needed to generate all focal branches were selected. This stringent selection leads to *Dataset 1*, comprising only 520 genes but the same genes for all branches (no missing genes). For *Dataset 2*, the genes were selected for each branch independently (i.e. for a given branch, a gene was selected if the sequence needed to generate that branch was present). This second selection leads to 1439 genes per branch on average among a total of 1533 genes, which were selected at least once for one branch. The genomic dataset is available in Figshare (see Data availability).

We studied the ratio ( $\omega$ ) of dN/dS to find genes under positive selection<sup>76,174</sup>. The dN/dS ratio is traditionally used to estimate selective pressure from protein-coding sequences. If host–plant shifts have no effect on the selection of a given gene, we expect a dN/dS = 1 and the selective regime is considered neutral. However, if host–plant shifts result in positive selection on coding genes, the ratio increases such that dN/dS > 1. Finally, it is possible that host–plant shifts lead to purifying selection, thus reducing the number of dN and resulting in dN/dS < 1. Here we focused on the adaptation of Papilionidae to host–plant shifts, that is, outgroups are not studied. We tested if branches representing inferred host–plant shifts along the phylogeny of swallowtails have more genes with dN/dS > 1 than lineages that did not have an inferred shift. The *branch-site* models allow  $\omega$  to vary both among sites in the protein and across branches on the tree and aim to detect positive selection affecting a few sites along particular lineages. The approach described by Zhang et al.<sup>175</sup> was chosen to determine genome-wide selection regimes as performed with two ML models: (1) a null model assuming two site classes, one with dN/dS < 1 and one with dN/dS = 1 (model = 2, NSsites = 2, fix\_omega = 1, omega = 1) and (2) an alternative model adding a third site class with dN/dS > 1 (model = 2, NSsites = 2, fix\_omega = 0, omega = 1.5). The fit for including positive selection is tested using a likelihood ratio test comparing the null model with the alternative model with one degree of freedom<sup>76,176</sup>. If the alternative model is better suited to host-shift branches, it is more likely the gene was under positive selection during the host–plant shifts. For each gene and for each branch, both the null and alternative models using CodeML were implemented in PAML 4<sup>177</sup> with a fixed topology (as inferred with the phylogenomic dataset<sup>46</sup>) and the nucleotide alignment of each gene. To test the robustness of the estimations, we used a false discovery rate test to control false positives<sup>178</sup>. Finally, for each branch, we reported the number of genes under positive selection (i.e. for which the alternative model including the site class with dN/dS > 1 have a better likelihood) on the total gene number. The proportion of genes under positive selection was compared with associated control branches for branches representing host–plant shifts, environmental shifts or both plant and environmental shifts using the non-parametric, Wilcoxon rank-sum test<sup>179</sup>.

**Sensitivity analyses.** We performed several control analyses to ensure that the signal of more genes under positive selection in host–plant shifts branches is not artefactual.

First, it has been shown that the choice of the tree is an important factor for the branch-site analysis of positive selection<sup>180</sup>. Indeed, constraining the topology for a given gene may lead to overestimating the number of substitution events for the constrained branches<sup>78</sup> and so could lead to overestimating the dN/dS ratio. Estimating dN/dS over thousands of gene trees would make the branch comparison not equal between control and test branches. Indeed, in a given gene tree it is likely and expected that the species topology is not always recovered, which results in a different number of branches compared to the species tree. For instance, the host-plant shift to Annonaceae might disappear in certain proportions of genes. We thus decided to estimate dN/dS on a fixed species tree topology for all genes to be sure to be able to measure this ratio for each gene that must be present in the topology for the focal branches. However, given that this issue can lead to a bias in our analysis, we decided to compute the number of gene trees that did not recover the branches of interest. We then checked whether the branches leading to a host-plant shift were more often unrecovered than the control branches without shift. Overall the control branches were less often recovered than host-plant shift branches ( $P = 0.030$ , Wilcoxon rank-sum test; data presented in Supplementary Table 4), which suggests that if gene tree/species tree discordance leads to an overestimation of positive selection, then this overestimation is higher for control branches than for host-plant shift branches. Finally, we filtered out the gene trees for which the focal branches were recovered in agreement with the species tree and used these genes to re-estimate the proportion of genes under positive selection among this new set of genes. We found that the  $P$  value remains significant ( $P = 0.0444$ , Wilcoxon rank-sum test; more genes during host-plant shifts than along control branches, Supplementary Table 2). Then, we specifically focused on missing data and GC-content variation among genes known to bias dN/dS estimations. Missing data are prone to introducing misaligned regions that could create false positives in branch-site likelihood method for detecting positive selection<sup>181–183</sup>. Variations in GC content are known to impact the estimation of dN/dS mainly through the process of GC-biased gene conversion (gBGC<sup>184–186</sup>).

The number of missing data ('N' and '-') sites and GC content at the third codon position (GC3) were computed using a home-made C++ program created with BIO++ library<sup>187</sup>. Mean GC content and missing data were calculated per gene and for each branch. For a given branch, mean GC3 and missing data were computed for the species of a clade for which the branch is the root. All statistics and graphical representations were performed using the R-packages *tidyverse*<sup>188</sup> and *cowplot*<sup>189</sup>. We found that genes under positive selection ( $PS_{\text{genes}} n_{\text{Dataset 1}} = 142$ ,  $n_{\text{Dataset 2}} = 407$ ) have significantly more missing data and GC3 than genes not under positive selection ( $NS_{\text{genes}} n_{\text{Dataset 1}} = 378$ ,  $n_{\text{Dataset 2}} = 1126$ ;  $P = 0.001/0.02$  for the two datasets, respectively, Mann-Whitney test; Supplementary Fig. 20). This result confirms that branch-site likelihood methods for detecting positive selection are sensitive to missing data, probably because of misaligned sites<sup>181,182</sup>, and that GC content that may be influenced by gBGC<sup>184,185</sup>.

Missing data were, however, heterogeneously distributed among species, ranging from <1% in *P. xuthus* to 45% in *Hypermnestra helios* (Supplementary Fig. 21). The difference in missing data between branches with ( $n = 14$ , mean missing<sub>Dataset 1</sub> = 13.4%, mean missing<sub>Dataset 2</sub> = 14.1%) or without host-plant shifts ( $n = 14$ , mean missing<sub>Dataset 1</sub> = 12.8%, mean missing<sub>Dataset 2</sub> = 12.7%) is not significant ( $P = 0.83/1.00$  for the two datasets, respectively, Mann-Whitney test; Supplementary Fig. 22). In addition, there is no correlation between the number of genes under positive selection and the amount of missing data ( $P = 0.33/0.20$  for the two datasets, respectively, Spearman's correlation test; Supplementary Fig. 23). For GC3, we also found variation between species ranging from 37% in *Parnassius smintheus* to 44% in *Papilio antimachus* (Supplementary Fig. 24). Similarly to missing data, we found no significant difference between plant-shift and no plant-shift branches ( $P = 0.63/0.63$  for the two datasets, Mann-Whitney test; Supplementary Fig. 25) and there is no correlation between the number of genes under positive selection and GC3 ( $P = 0.20/0.1362$  for the two datasets, respectively, Spearman's correlation test; Supplementary Fig. 26).

Despite the known fact that false positives can increase with the amount of missing data, our control analyses indicate that variations in missing data and GC content do not drive the signal that more genes are under positive selection in branches that have undergone a host-plant shift. Additionally to these controls, we checked by eyes all the gene alignments at the amino-acid level for genes under positive selection in branches with and without host-plant shifts using SeaView 4<sup>190</sup>. Misaligned regions, which could lead to biased dN/dS ratios<sup>191</sup>, were not significantly more detected for genes under positive selection in branches with host-plant shifts. In some cases, we found ourselves in complicated situations to discriminate between false- and true-positive selected genes.

Overall, given our alignment checks and sensitivity analyses, we do not see any reason for biased dN/dS ratios in genes along branches with or without host-plant shifts. False-positive and false-negative genes can be present in the two categories of branches, but, in any cases, the general pattern observed is likely to remain conserved.

**Gene ontology.** To annotate proteins of our alignment, we used the two different approaches implemented in PANTHER 14<sup>192</sup> (available at: <http://pantherdb.org/>) and EggNOG 5.0<sup>193,194</sup> (available at: <http://eggnog5.embl.de/#/app/home>). We used the HMM Scoring tool to assign PANTHER family (library version 14.1<sup>192</sup>) to the protein of *P. xuthus* (assembly Pxut\_1.0); similar results were obtained using

another high-quality annotated genome (from *H. melpomene*) as reference (assembly ASM31383v2). We performed the statistical overrepresentation test implemented on the PANTHER online website, relying on the GO categories in the PANTHER GO-Slim annotation dataset including Molecular function, Biological process and Cellular component. First, we tested if positively selected genes have over- or under-represented functional GO categories as compared to the whole set of genes (option "PANTHER Generic Mapping"). Second, we tested if positively selected genes involving a host-plant shift along the 14 branches have over- or under-represented functional categories. These statistical comparisons were performed with the Fisher's exact test using the false discovery rate correction to control for false positives. Independently, we used the eggNOG-mapper v2<sup>193</sup> (<https://github.com/eggnogdb/eggnog-mapper>) and the associated Lepidoptera database (LepNOG, including the genomes of *Bombyx mori*, *Danaus plexippus* and *Heliconius melpomene*<sup>194</sup>) to annotate the proteins of our dataset. EggNOG uses precomputed orthologous groups and phylogenies from the database to transfer functional information from fine-grained orthologs only. We used the diamond method as recommended<sup>193</sup>. Finally, we reported the GO families inferred for the proteins of the Dataset 2.

**Reporting summary.** Further information on research design is available in the Nature Research Reporting Summary linked to this article.

## Data availability

Supermatrix datasets (for phylogenetic analyses), phylogenetic trees, host-plant preferences, species geographic distributions, and gene alignments (for dN/dS analyses) that are necessary for repeating the analyses described here have been made available through the Figshare digital data repository (<https://doi.org/10.6084/m9.figshare.12278402>). Source data are provided with this paper.

## Code availability

Bioinformatic scripts used to perform the analyses described here are available through the Figshare digital data repository (<https://doi.org/10.6084/m9.figshare.12278402>).

Received: 23 June 2020; Accepted: 3 December 2020;

Published online: 13 January 2021

## References

- Becerra, J. X. On the factors that promote the diversity of herbivorous insects and plants in tropical forests. *Proc. Natl Acad. Sci. USA* **112**, 6098–6103 (2015).
- Stork, N. E. How many species of insects and other terrestrial arthropods are there on earth? *Annu. Rev. Entomol.* **63**, 31–45 (2018).
- Grimaldi, D. A. & Engel, M. S. *Evolution of the Insects* (Cambridge University Press, 2005).
- Strong, D. R., Lawton, J. H. & Southwood, R. *Insects on Plants: Community Patterns and Mechanisms* (Harvard University Press, 1984).
- Ehrlich, P. R. & Raven, P. H. Butterflies and plants: a study in coevolution. *Evolution* **18**, 586–608 (1964).
- Thompson, J. N. Concepts of coevolution. *Trends Ecol. Evol.* **4**, 179–183 (1989).
- Mitter, C., Farrell, B. & Wiegmann, B. The phylogenetic study of adaptive zones: has phytophagy promoted insect diversification? *Am. Nat.* **132**, 107–128 (1988).
- Farrell, B. D. 'Inordinate fondness' explained: why are there so many beetles? *Science* **281**, 555–559 (1998).
- Berenbaum, M. & Specialization, P. F. *Chemical Mediation of Host-plant Specialization: The Papilionid Paradigm. Specialization, Speciation, and Radiation: The Evolutionary Biology of Herbivorous Insects* (University of California Press, 2008).
- Winter, S., Friedman, A. L. L., Astrin, J. J., Gottsberger, B. & Letsch, H. Timing and host plant associations in the evolution of the weevil tribe Apionini (Apioninae, Brentidae, Curculionioidea, Coleoptera) indicate an ancient co-diversification pattern of beetles and flowering plants. *Mol. Phylogenet. Evol.* **107**, 179–190 (2017).
- Kergoat, G. J. et al. Opposite macroevolutionary responses to environmental changes in grasses and insects during the Neogene grassland expansion. *Nat. Commun.* **9**, 5089 (2018).
- Wheat, C. W. et al. The genetic basis of a plant-insect coevolutionary key innovation. *Proc. Natl Acad. Sci. USA* **104**, 20427–20431 (2007).
- Edger, P. P. et al. The butterfly plant arms-race escalated by gene and genome duplications. *Proc. Natl Acad. Sci. USA* **112**, 8362–8366 (2015).
- Calla, B. et al. Cytochrome P450 diversification and hostplant utilization patterns in specialist and generalist moths: Birth, death and adaptation. *Mol. Ecol.* **26**, 6021–6035 (2017).

15. Nallu, S. et al. The molecular genetic basis of herbivory between butterflies and their host plants. *Nat. Ecol. Evol.* **2**, 1418–1427 (2018).
16. Karageorgi, M. et al. Genome editing retraces the evolution of toxin resistance in the monarch butterfly. *Nature* **574**, 409–412 (2019).
17. Sahoo, R. K., Warren, A. D., Collins, S. C. & Kodandaramaiah, U. Hostplant change and paleoclimatic events explain diversification shifts in skipper butterflies (Family: Hesperidae). *BMC Evol. Biol.* **17**, 174 (2017).
18. Condamine, F. L., Rolland, J., Höhna, S., Sperling, F. A. H. & Sanmartín, I. Testing the role of the red queen and court jester as drivers of the macroevolution of apollo butterflies. *Syst. Biol.* **67**, 940–964 (2018).
19. Letsch, H. et al. Climate and host-plant associations shaped the evolution of ceutorhynch weevils throughout the Cenozoic. *Evolution* **72**, 1815–1828 (2018).
20. Forister, M. L. et al. The global distribution of diet breadth in insect herbivores. *Proc. Natl Acad. Sci. USA* **112**, 442–447 (2015).
21. Winkler, I. S., Mitter, C. & Scheffer, S. J. Repeated climate-linked host shifts have promoted diversification in a temperate clade of leaf-mining flies. *Proc. Natl Acad. Sci. USA* **106**, 18103–18108 (2009).
22. Chomicki, G., Weber, M., Antonelli, A., Bascompte, J. & Kiers, E. T. The impact of mutualisms on species richness. *Trends Ecol. Evol.* **34**, 698–711 (2019).
23. Janz, N. Ehrlich and Raven revisited: mechanisms underlying codiversification of plants and enemies. *Annu. Rev. Ecol. Syst.* **42**, 71–89 (2011).
24. Suchan, T. & Alvarez, N. Fifty years after Ehrlich and Raven, is there support for plant–insect coevolution as a major driver of species diversification? *Entomol. Exp. Appl.* **157**, 98–112 (2015).
25. Endara, M.-J. et al. Coevolutionary arms race versus host defense chase in a tropical herbivore–plant system. *Proc. Natl Acad. Sci. USA* **114**, E7499–E7505 (2017).
26. Simon, J.-C. et al. Genomics of adaptation to host-plants in herbivorous insects. *Brief. Funct. Genomics* **14**, 413–423 (2015).
27. Hammer, T. J., Janzen, D. H., Hallwachs, W., Jaffe, S. P. & Fierer, N. Caterpillars lack a resident gut microbiome. *Proc. Natl Acad. Sci. USA* **114**, 9641–9646 (2017).
28. Hua, X. & Bromham, L. Darwinism for the genomic age: connecting mutation to diversification. *Front. Genet.* **8**, 12 (2017).
29. Hembry, D. H. & Weber, M. G. Ecological interactions and macroevolution: a new field with old roots. *Annu. Rev. Ecol. Syst.* **51**, (2020).
30. Scriber, J. M., Tsubaki, Y. & Lederhouse, R. C. *Swallowtail Butterflies: Their Ecology and Evolutionary Biology* (Scientific Publishers, 1995).
31. Nishida, R. Sequestration of defensive substances from plants by Lepidoptera. *Annu. Rev. Entomol.* **47**, 57–92 (2002).
32. Schmeiser, H. H., Stiborová, M. & Arlt, V. M. Chemical and molecular basis of the carcinogenicity of *Aristolochia* plants. *Curr. Opin. Drug Discov. Dev.* **12**, 141–148 (2009).
33. Poon, S. L. et al. Genome-wide mutational signatures of aristolochic acid and its application as a screening tool. *Sci. Transl. Med.* **5**, 197ra101 (2013).
34. Condamine, F. L., Sperling, F. A. H., Wahlberg, N., Rasplus, J.-Y. & Kergoat, G. J. What causes latitudinal gradients in species diversity? Evolutionary processes and ecological constraints on swallowtail biodiversity. *Ecol. Lett.* **15**, 267–277 (2012).
35. Simonsen, T. J. et al. Phylogenetics and divergence times of Papilioninae (Lepidoptera) with special reference to the enigmatic genera *Teinopalpus* and *Meandrusa*. *Cladistics* **27**, 113–137 (2011).
36. Berenbaum, M. R., Favret, C. & Schuler, M. A. On defining ‘Key Innovations’ in an adaptive radiation: cytochrome P450s and Papilionidae. *Am. Nat.* **148**, S139–S155 (1996).
37. Cohen, M. B., Schuler, M. A. & Berenbaum, M. R. A host-inducible cytochrome P-450 from a host-specific caterpillar: molecular cloning and evolution. *Proc. Natl Acad. Sci. USA* **89**, 10920–10924 (1992).
38. Li, W., Schuler, M. A. & Berenbaum, M. R. Diversification of furanocoumarin-metabolizing cytochrome P450 monooxygenases in two papilionids: specificity and substrate encounter rate. *Proc. Natl Acad. Sci. USA* **100**(Suppl.), 14593–14598 (2003).
39. Thompson, J. N. Variation in preference and specificity in monophagous and oligophagous swallowtail butterflies. *Evolution* **42**, 118–128 (1988).
40. Thompson, J. N., Wehling, W. & Podolsky, R. Evolutionary genetics of host use in swallowtail butterflies. *Nature* **344**, 148–150 (1990).
41. Berenbaum, M. R. & Feeny, P. P. in *Specialization, Speciation, and Radiation: The Evolutionary Biology of Herbivorous Insects* (ed. Tilmon, K.) 2–19 (University of California Press, 2008).
42. Zakharov, E. V., Caterino, M. S. & Sperling, F. A. H. Molecular phylogeny, historical biogeography, and divergence time estimates for swallowtail butterflies of the genus *Papilio* (Lepidoptera: Papilionidae). *Syst. Biol.* **53**, 193–215 (2004).
43. Braby, M., Trueman, J. & Eastwood, R. When and where did troidine butterflies (Lepidoptera: Papilionidae) evolve? Phylogenetic and biogeographic evidence suggests an origin in remnant Gondwana in the Late Cretaceous. *Invertebr. Syst.* **19**, 113–143 (2005).
44. Condamine, F. L., Silva-Brandão, K. L., Kergoat, G. J. & Sperling, F. A. Biogeographic and diversification patterns of Neotropical Troidini butterflies (Papilionidae) support a museum model of diversity dynamics for Amazonia. *BMC Evol. Biol.* **12**, 82 (2012).
45. Condamine, F. L. et al. Deciphering the evolution of birdwing butterflies 150 years after Alfred Russel Wallace. *Sci. Rep.* **5**, 11860 (2015).
46. Allio, R. et al. Whole genome shotgun phylogenomics resolves the pattern and timing of swallowtail butterfly evolution. *Syst. Biol.* **69**, 38–60 (2020).
47. McKenna, D. D., Sequeira, A. S., Marvaldi, A. E. & Farrell, B. D. Temporal lags and overlap in the diversification of weevils and flowering plants. *Proc. Natl Acad. Sci. USA* **106**, 7083–7088 (2009).
48. Takahashi, D. & Setoguchi, H. Molecular phylogeny and taxonomic implications of *Asarum* (Aristolochiaceae) based on ITS and *matK* sequences. *Plant Species Biol.* **33**, 28–41 (2018).
49. Wanke, S. et al. Evolution of Piperales—*matK* gene and *trnK* intron sequence data reveal lineage specific resolution contrast. *Mol. Phylogenet. Evol.* **42**, 477–497 (2007).
50. Neinhuis, C., Wanke, S., Hilu, K. W., Müller, K. & Borsch, T. Phylogeny of Aristolochiaceae based on parsimony, likelihood, and Bayesian analyses of *trnL-trnF* sequences. *Plant Syst. Evol.* **250**, 7–26 (2005).
51. Wanke, S., González, F. & Neinhuis, C. Systematics of pipevines: combining morphological and fast-evolving molecular characters to investigate the relationships within subfamily Aristolochioideae. *Int. J. Plant Sci.* **167**, 1215–1227 (2006).
52. González, F. et al. Present trans-Pacific disjunct distribution of *Aristolochia* subgenus *Isotrema* (Aristolochiaceae) was shaped by dispersal, vicariance and extinction. *J. Biogeogr.* **41**, 380–391 (2014).
53. Durden, C. J. & Rose, H. *Butterflies from the Middle Eocene: The Earliest Occurrence of Fossil Papilionoidea (Lepidoptera)* (Prace-Sellards Ser. Tax. Mem. Mus., 1978).
54. Sohn, J., Labandeira, C., Davis, D. & Mitter, C. An annotated catalog of fossil and subfossil Lepidoptera (Insecta: Holometabola) of the world. *Zootaxa* **3286**, 1–132 (2012).
55. de Jong, R. Estimating time and space in the evolution of the Lepidoptera. *Tijdschr. voor Entomol.* **150**, 319–346 (2007).
56. Hofmann, C.-C. & Zetter, R. Upper Cretaceous sulcate pollen from the Timerdyakh formation, Vilui Basin (Siberia). *Grana* **49**, 170–193 (2010).
57. Meller, B. The first fossil *Aristolochia* (Aristolochiaceae, Piperales) leaves from Austria. *Palaeontol. Electron* **17**, 1–17 (2014).
58. Nee, S., May, R. M. & Harvey, P. H. The reconstructed evolutionary process. *Philos. Trans. R. Soc. Lond. Ser. B* **344**, 305–311 (1994).
59. Nee, S. Birth-death models in macroevolution. *Annu. Rev. Ecol. Syst.* **37**, 1–17 (2006).
60. Rabosky, D. L. & Lovette, I. J. Explosive evolutionary radiations: Decreasing speciation or increasing extinction through time? *Evolution* **62**, 1866–1875 (2008).
61. Crisp, M. D. & Cook, L. G. Explosive radiation or cryptic mass extinction? Interpreting signatures in molecular phylogenies. *Evolution* **63**, 2257–2265 (2009).
62. Quantal, T. B. & Marshall, C. R. Diversity dynamics: molecular phylogenies need the fossil record. *Trends Ecol. Evol.* **25**, 434–441 (2010).
63. Morlon, H. Phylogenetic approaches for studying diversification. *Ecol. Lett.* **17**, 508–525 (2014).
64. Xue, B. et al. Accelerated diversification correlated with functional traits shapes extant diversity of the early divergent angiosperm family Annonaceae. *Mol. Phylogenet. Evol.* **142**, 106659 (2020).
65. Folk, R. A. et al. Rates of niche and phenotype evolution lag behind diversification in a temperate radiation. *Proc. Natl Acad. Sci. USA* **116**, 10874–10882 (2019).
66. Sun, M. et al. Recent accelerated diversification in rosids occurred outside the tropics. *Nat. Commun.* **11**, 3333 (2020).
67. Losos, J. B. Adaptive radiation, ecological opportunity, and evolutionary determinism. *Am. Nat.* **175**, 623–639 (2010).
68. Cheng, T. et al. Genomic adaptation to polyphagy and insecticides in a major East Asian noctuid pest. *Nat. Ecol. Evol.* **1**, 1747–1756 (2017).
69. Rane, R. V. et al. Detoxifying enzyme complements and host use phenotypes in 160 insect species. *Curr. Opin. Insect Sci.* **31**, 131–138 (2019).
70. Cong, Q., Borek, D., Otwinowski, Z. & Grishin, N. V. Tiger swallowtail genome reveals mechanisms for speciation and caterpillar chemical defense. *Cell Rep.* **10**, 910–919 (2015).
71. Li, X. et al. Outbred genome sequencing and CRISPR/Cas9 gene editing in butterflies. *Nat. Commun.* **6**, 8212 (2015).
72. Nishikawa, H. et al. A genetic mechanism for female-limited Batesian mimicry in *Papilio* butterfly. *Nat. Genet.* **47**, 405–409 (2015).

73. Thomas, G. W. C. & Hahn, M. W. Determining the null model for detecting adaptive convergence from genomic data: a case study using echolocating mammals. *Mol. Biol. Evol.* **32**, 1232–1236 (2015).
74. Zou, Z. & Zhang, J. No genome-wide protein sequence convergence for echolocation. *Mol. Biol. Evol.* **32**, 1237–1241 (2015).
75. Kimura, M. *The Neutral Theory of Molecular Evolution* (Cambridge University Press, 1983).
76. Yang, Z. *Computational Molecular Evolution* (Oxford University Press, 2006).
77. Venkat, A., Hahn, M. W. & Thornton, J. W. Multinucleotide mutations cause false inferences of lineage-specific positive selection. *Nat. Ecol. Evol.* **2**, 1280–1288 (2018).
78. Mendes, F. K. & Hahn, M. W. Gene tree discordance causes apparent substitution rate variation. *Syst. Biol.* **65**, 711–721 (2016).
79. Dasmahapatra, K. K. et al. Butterfly genome reveals promiscuous exchange of mimicry adaptations among species. *Nature* **487**, 94–98 (2012).
80. Walden, N. et al. Nested whole-genome duplications coincide with diversification and high morphological disparity in Brassicaceae. *Nat. Commun.* **11**, 3795 (2020).
81. McGee, M. D. et al. The ecological and genomic basis of explosive adaptive radiation. *Nature* **586**, 75–79 (2020).
82. Thomas, G. W. C. et al. Gene content evolution in the arthropods. *Genome Biol.* **21**, 15 (2020).
83. de Medeiros, B. A. S. & Farrell, B. D. Evaluating species interactions as a driver of phytophagous insect divergence. *bioRxiv* <https://doi.org/10.1101/842153> (2019).
84. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).
85. Lanfear, R., Frandsen, P. B., Wright, A. M., Senfeld, T. & Calcott, B. PartitionFinder 2: new methods for selecting partitioned models of evolution for molecular and morphological phylogenetic analyses. *Mol. Biol. Evol.* **34**, 772–773 (2016).
86. Nguyen, L.-T., Schmidt, H. A., von Haeseler, A. & Minh, B. Q. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* **32**, 268–274 (2015).
87. Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., von Haeseler, A. & Jermini, L. S. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat. Methods* **14**, 587–589 (2017).
88. Chernomor, O., von Haeseler, A. & Minh, B. Q. Terrace aware data structure for phylogenomic inference from supermatrices. *Syst. Biol.* **65**, 997–1008 (2016).
89. Minh, B. Q., Nguyen, M. A. T. & von Haeseler, A. Ultrafast approximation for phylogenetic bootstrap. *Mol. Biol. Evol.* **30**, 1188–1195 (2013).
90. Ronquist, F. et al. MrBayes 3.2: efficient bayesian phylogenetic inference and model choice across a large model space. *Syst. Biol.* **61**, 539–542 (2012).
91. Huelsenbeck, J. P., Larget, B. & Alfaro, M. E. Bayesian phylogenetic model selection using reversible jump Markov Chain Monte Carlo. *Mol. Biol. Evol.* **21**, 1123–1133 (2004).
92. Rambaut, A., Drummond, A. J., Xie, D., Baele, G. & Suchard, M. A. Posterior summarization in bayesian phylogenetics using Tracer 1.7. *Syst. Biol.* **67**, 901–904 (2018).
93. Douady, C. J., Delsuc, F., Boucher, Y., Doolittle, W. F. & Douzery, E. J. P. Comparison of bayesian and maximum likelihood bootstrap measures of phylogenetic reliability. *Mol. Biol. Evol.* **20**, 248–254 (2003).
94. Miller, M. A. et al. A RESTful API for access to phylogenetic tools via the CIPRES Science Gateway. *Evol. Bioinforma.* **11**, EBO.S21501 (2015).
95. Ayres, D. L. et al. BEAGLE: an application programming interface and high-performance computing library for statistical phylogenetics. *Syst. Biol.* **61**, 170–173 (2012).
96. Drummond, A. J., Ho, S. Y. W., Phillips, M. J. & Rambaut, A. Relaxed phylogenetics and dating with confidence. *PLoS Biol.* **4**, e88 (2006).
97. Drummond, A. J., Suchard, M. A., Xie, D. & Rambaut, A. Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Mol. Biol. Evol.* **29**, 1969–1973 (2012).
98. Smith, M. E., Singer, B. & Carroll, A. <sup>40</sup>Ar/<sup>39</sup>Ar geochronology of the Eocene Green River Formation, Wyoming. *Geol. Soc. Am. Bull.* **115**, 549–565 (2003).
99. de Jong, R. Fossil butterflies, calibration points and the molecular clock (Lepidoptera: Papilionoidea). *Zootaxa* **4270**, 1–63 (2017).
100. Scudder, S. H. Fossil butterflies. *Mem. Am. Assoc. Adv. Sci.* **1**, 1–99 (1875).
101. Rasnitsyn, A. P. & Zherikhin, V. V. in *History of Insects* 437–446 (Kluwer Academic Publishers, 2002).
102. Rebel, H. *Doritites bosniaskii*. Sitzungsberichte der akademie der wissenschaften. Mathematischen-Naturwissenschaftliche classe. *Abt. 1 Mineral. Biol. Erdkd.* **1**, 734–741 (1898).
103. Carpenter, F. *Treatise on Invertebrate Paleontology: Arthropoda 4. Superclass Hexapoda* (Geological Society of America, 1992).
104. Magallón, S., Gómez-Acevedo, S., Sánchez-Reyes, L. L. & Hernández-Hernández, T. A metacalibrated time-tree documents the early rise of flowering plant phylogenetic diversity. *N. Phytol.* **207**, 437–453 (2015).
105. Sohn, J.-C., Labandeira, C. C. & Davis, D. R. The fossil record and taphonomy of butterflies and moths (Insecta, Lepidoptera): implications for evolutionary diversity and divergence-time estimates. *BMC Evol. Biol.* **15**, 12 (2015).
106. Toussaint, E. F. A. & Condamine, F. L. To what extent do new fossil discoveries change our understanding of clade evolution? A cautionary tale from burying beetles (Coleoptera: *Nicrophorus*). *Biol. J. Linn. Soc.* **117**, 686–704 (2016).
107. Gernhard, T. The conditioned reconstructed process. *J. Theor. Biol.* **253**, 769–778 (2008).
108. Lewis, P. O. A likelihood approach to estimating phylogeny from discrete morphological character data. *Syst. Biol.* **50**, 913–925 (2001).
109. Ree, R. H. & Smith, S. A. Maximum likelihood inference of geographic range evolution by dispersal, local extinction, and cladogenesis. *Syst. Biol.* **57**, 4–14 (2008).
110. Pagel, M. & Meade, A. Bayesian analysis of correlated evolution of discrete characters by reversible-jump Markov chain Monte Carlo. *Am. Nat.* **167**, 808–825 (2006).
111. Igarashi, S. The classification of the Papilionidae mainly based on the morphology of their immature stages. *Lepid. Sci.* **34**, 41–96 (1984).
112. Collins, N. M. & Morris, M. *Threatened Swallowtail Butterflies of the World: the IUCN Red Data Book* (IUCN, 1985).
113. Tyler, H. A., Brown, K. S. & Wilson, K. H. *Swallowtail Butterflies of the Americas: A Study in Biological Dynamics, Ecological Diversity, Biosystematics, and Conservation* (Scientific Publishers, 1994).
114. Ree, R. H., Moore, B. R., Webb, C. O. & Donoghue, M. J. A likelihood framework for inferring the evolution of geographic range on phylogenetic trees. *Evolution* **59**, 2299–2311 (2005).
115. Massoni, J., Couvreur, T. L. & Sauquet, H. Five major shifts of diversification through the long evolutionary history of Magnoliidae (Angiosperms). *BMC Evol. Biol.* **15**, 49 (2015).
116. Kyalangalilwa, B., Boatwright, J. S., Daru, B. H., Maurin, O. & van der Bank, M. Phylogenetic position and revised classification of *Acacia* s.l. (Fabaceae: Mimosoideae) in Africa, including new combinations in *Vachellia* and *Senegalia*. *Bot. J. Linn. Soc.* **172**, 500–523 (2013).
117. Miller, J. T., Murphy, D. J., Ho, S. Y. W., Cantrill, D. J. & Seigler, D. Comparative dating of *Acacia*: combining fossils and multiple phylogenies to infer ages of clades with poor fossil records. *Aust. J. Bot.* **61**, 436–445 (2013).
118. Michalak, I., Zhang, L.-B. & Renner, S. S. Trans-Atlantic, trans-Pacific and trans-Indian Ocean dispersal in the small Gondwanan Laurales family Hernandiaceae. *J. Biogeogr.* **37**, 1214–1226 (2010).
119. Wu, S.-D. et al. Evolution of asian interior arid-zone biota: Evidence from the diversification of asian *Zygophyllum* (Zygophyllaceae). *PLoS ONE* **10**, e0138697 (2015).
120. Chase, M. W. et al. An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG IV. *Bot. J. Linn. Soc.* **181**, 1–20 (2016).
121. Christenhusz, M. J. M., Vorontsova, M. S., Fay, M. F. & Chase, M. W. Results from an online survey of family delimitation in angiosperms and ferns: recommendations to the Angiosperm Phylogeny Group for thorny problems in plant classification. *Bot. J. Linn. Soc.* **178**, 501–528 (2015).
122. Gonzáles, F., Rudall, P. J. & Furness, C. A. Microsporogenesis and systematics of Aristolochiaceae. *Bot. J. Linn. Soc.* **137**, 221–242 (2001).
123. González, F. & Rudall, P. The questionable affinities of *Lactoris*: evidence from branching pattern, inflorescence morphology, and stipule development. *Am. J. Bot.* **88**, 2143–2150 (2001).
124. Isnard, S. et al. Growth form evolution in Piperales and its relevance for understanding angiosperm diversification: An integrative approach combining plant architecture, anatomy, and biomechanics. *Int. J. Plant Sci.* **173**, 610–639 (2012).
125. Wagner, S. T. et al. Major trends in stem anatomy and growth forms in the perianth-bearing Piperales, with special focus on *Aristolochia*. *Ann. Bot.* **113**, 1139–1154 (2014).
126. Nickrent, D. L. et al. Molecular data place Hydnoraceae with Aristolochiaceae. *Am. J. Bot.* **89**, 1809–1817 (2002).
127. Kelly, L. M. & González, F. Phylogenetic relationships in Aristolochiaceae. *Syst. Bot.* **28**, 236–249 (2003).
128. Naumann, J. et al. Single-copy nuclear genes place haustorial Hydnoraceae within piperales and reveal a cretaceous origin of multiple parasitic angiosperm lineages. *PLoS ONE* **8**, e79204 (2013).
129. Salomo, K. et al. The emergence of earliest angiosperms may be earlier than fossil evidence indicates. *Syst. Bot.* **42**, 607–619 (2017).
130. Christenhusz, M. J. M. & Byng, J. W. The number of known plants species in the world and its annual increase. *Phytotaxa* **261**, 201–217 (2016).

131. Naumann, J. et al. Detecting and characterizing the highly divergent plastid genome of the nonphotosynthetic parasitic plant *Hydnora visseri* (Hydnoraceae). *Genome Biol. Evol.* **8**, 345–363 (2016).
132. Jost, M., Naumann, J., Rocamundi, N., Cocucci, A. A. & Wanke, S. The first plastid genome of the Holoparasitic genus *Prosopanche* (Hydnoraceae). *Plants* **9**, 306 (2020).
133. Zavada, M. S. & Benson, J. M. First fossil evidence for the primitive angiosperm family Lactoridaceae. *Am. J. Bot.* **74**, 1590–1594 (1987).
134. Gamarro, J. C. & Barreda, V. New fossil record of Lactoridaceae in southern South America: a palaeobiogeographical approach. *Bot. J. Linn. Soc.* **158**, 41–50 (2008).
135. Smith, S. Y. & Stockey, R. A. Establishing a fossil record for the perianthless Piperales: *Saururus tuckerae* sp. nov. (Saururaceae) from the Middle Eocene Princeton Chert. *Am. J. Bot.* **94**, 1642–1657 (2007).
136. Massoni, J., Doyle, J. & Sauquet, H. Fossil calibration of Magnoliidae, an ancient lineage of angiosperms. *Palaeontol. Electron.* **18**, 1–25 (2015).
137. Smith, S. A. Taking into account phylogenetic and divergence-time uncertainty in a parametric biogeographical analysis of the Northern Hemisphere plant clade Caprifoliaceae. *J. Biogeogr.* **36**, 2324–2337 (2009).
138. Beeravolu, C. R. & Condamine, F. L. An extended maximum likelihood inference of geographic range evolution by dispersal, local extinction and cladogenesis. *bioRxiv* <https://doi.org/10.1101/038695> (2016).
139. Scotese, C. R. A continental drift flipbook. *J. Geol.* **112**, 729–741 (2004).
140. Blakey, R. C. Gondwana paleogeography from assembly to breakup—a 500 m. y. odyssey. *Geol. Soc. Am. Spec. Pap.* **441**, 1–28 (2008).
141. Seton, M. et al. Global continental and ocean basin reconstructions since 200 Ma. *Earth Sci. Rev.* **113**, 212–270 (2012).
142. Chacón, J. & Renner, S. S. Assessing model sensitivity in ancestral area reconstruction using Lagrange: a case study using the Colchicaceae family. *J. Biogeogr.* **41**, 1414–1427 (2014).
143. Maddison, W. P., Midford, P. E. & Otto, S. P. Estimating a binary character's effect on speciation and extinction. *Syst. Biol.* **56**, 701–710 (2007).
144. FitzJohn, R. G., Maddison, W. P. & Otto, S. P. Estimating trait-dependent speciation and extinction rates from incompletely resolved phylogenies. *Syst. Biol.* **58**, 595–611 (2009).
145. Morlon, H., Parsons, T. L. & Plotkin, J. B. Reconciling molecular phylogenies with the fossil record. *Proc. Natl Acad. Sci. USA* **108**, 16327–16332 (2011).
146. Rabosky, D. L. et al. Rates of speciation and morphological evolution are correlated across the largest vertebrate radiation. *Nat. Commun.* **4**, 1958 (2013).
147. Höhna, S. et al. A Bayesian approach for estimating branch-specific speciation and extinction rates. *bioRxiv* <https://doi.org/10.1101/555805> (2019).
148. May, M. R., Höhna, S. & Moore, B. R. A Bayesian approach for detecting the impact of mass-extinction events on molecular phylogenies when rates of lineage diversification may vary. *Methods Ecol. Evol.* **7**, 947–959 (2016).
149. Magallon, S. & Sanderson, M. J. Absolute diversification rates in angiosperm clades. *Evolution* **55**, 1762–1780 (2001).
150. Rabosky, D. L. Likelihood methods for detecting temporal shifts in diversification rates. *Evolution* **60**, 1152–1164 (2006).
151. FitzJohn, R. G. Diversitree: comparative phylogenetic analyses of diversification in R. *Methods Ecol. Evol.* **3**, 1084–1092 (2012).
152. Scriber, J. M. in *Chemical Ecology of Insects* (eds Bell, W. J. & Cardé, R. T.) 159–202 (Springer US, 1984).
153. Davis, M. P., Midford, P. E. & Maddison, W. Exploring power and parameter estimation of the BiSSE method for analyzing species diversification. *BMC Evol. Biol.* **13**, 38 (2013).
154. Maddison, W. P. & FitzJohn, R. G. The unsolved challenge to phylogenetic correlation tests for categorical characters. *Syst. Biol.* **64**, 127–136 (2015).
155. Rabosky, D. L. & Goldberg, E. E. Model inadequacy and mistaken inferences of trait-dependent speciation. *Syst. Biol.* **64**, 340–355 (2015).
156. Morlon, H. et al. RPANDA: an R package for macroevolutionary analyses on phylogenetic trees. *Methods Ecol. Evol.* **7**, 589–597 (2016).
157. Rabosky, D. L. Automatic detection of key innovations, rate shifts, and diversity-dependence on phylogenetic trees. *PLoS ONE* **9**, e89543 (2014).
158. Moore, B. R., Höhna, S., May, M. R., Rannala, B. & Huelsenbeck, J. P. Critically evaluating the theory and performance of Bayesian analysis of macroevolutionary mixtures. *Proc. Natl Acad. Sci. USA* **113**, 9569–9574 (2016).
159. Rabosky, D. L. et al. BAMMtools: an R package for the analysis of evolutionary dynamics on phylogenetic trees. *Methods Ecol. Evol.* **5**, 701–707 (2014).
160. Rabosky, D. L., Mitchell, J. S. & Chang, J. Is BAMM flawed? Theoretical and practical concerns in the analysis of multi-rate diversification models. *Syst. Biol.* **66**, 477–498 (2017).
161. Höhna, S. et al. RevBayes: Bayesian phylogenetic inference using graphical models and an interactive model-specification language. *Syst. Biol.* **65**, 726–736 (2016).
162. Höhna, S., May, M. R. & Moore, B. R. TESS: an R package for efficiently simulating phylogenetic trees and performing Bayesian inference of lineage diversification rates. *Bioinformatics* **32**, 789–791 (2016).
163. Stadler, T. Mammalian phylogeny reveals recent diversification rate shifts. *Proc. Natl Acad. Sci. USA* **108**, 6187–6192 (2011).
164. Partha, R. et al. Subterranean mammals show convergent regression in ocular genes and enhancers, along with adaptation to tunneling. *eLife* **6**, e25884 (2017).
165. Wu, J., Yonezawa, T. & Kishino, H. Rates of molecular evolution suggest natural history of life history traits and a Post-K-Pg nocturnal bottleneck of placentals. *Curr. Biol.* **27**, 3025–3033 (2017).
166. Zhang, G. et al. Comparative genomics reveals insights into avian genome evolution and adaptation. *Science* **346**, 1311–1320 (2014).
167. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
168. Luo, R. et al. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *Gigascience* **1**, 18 (2012).
169. Abascal, F., Zardoya, R. & Telford, M. J. TranslatorX: multiple alignment of nucleotide sequences guided by amino acid translations. *Nucleic Acids Res.* **38**, W7–W13 (2010).
170. Simion, P. et al. A software tool 'CroCo' detects pervasive cross-species contamination in next generation sequencing data. *BMC Biol.* **16**, 28 (2018).
171. Emms, D. M. & Kelly, S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol.* **20**, 238 (2019).
172. Di Franco, A., Poujol, R., Baurain, D. & Philippe, H. Evaluating the usefulness of alignment filtering methods to reduce the impact of errors on evolutionary inferences. *BMC Evol. Biol.* **19**, 21 (2019).
173. Capella-Gutierrez, S., Silla-Martinez, J. M. & Gabaldon, T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972–1973 (2009).
174. Yang, Z. & Nielsen, R. Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. *Mol. Biol. Evol.* **17**, 32–43 (2000).
175. Zhang, J., Nielsen, R. & Yang, Z. Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. *Mol. Biol. Evol.* **22**, 2472–2479 (2005).
176. Yang, Z. Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. *Mol. Biol. Evol.* **15**, 568–573 (1998).
177. Yang, Z. PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24**, 1586–1591 (2007).
178. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc.* **57**, 289–300 (1995).
179. Bauer, D. F. Constructing confidence sets using rank statistics. *J. Am. Stat. Assoc.* **67**, 687–690 (1972).
180. Diekmann, Y. & Pereira-Leal, J. B. Gene tree affects inference of sites under selection by the branch-site test of positive selection. *Evol. Bioinforma.* **11**, 11–17 (2015).
181. Mallick, S., Gnerre, S., Muller, P. & Reich, D. The difficulty of avoiding false positives in genome scans for natural selection. *Genome Res.* **19**, 922–933 (2009).
182. Fletcher, W. & Yang, Z. The effect of insertions, deletions, and alignment errors on the branch-site test of positive selection. *Mol. Biol. Evol.* **27**, 2257–2267 (2010).
183. Jordan, G. & Goldman, N. The effects of alignment error and alignment filtering on the sitewise detection of positive selection. *Mol. Biol. Evol.* **29**, 1125–1139 (2012).
184. Duret, L. & Galtier, N. Biased gene conversion and the evolution of mammalian genomic landscapes. *Annu. Rev. Genomics Hum. Genet.* **10**, 285–311 (2009).
185. Galtier, N. & Duret, L. Adaptation or biased gene conversion? Extending the null hypothesis of molecular evolution. *Trends Genet.* **23**, 273–277 (2007).
186. Ratnakumar, A. et al. Detecting positive selection within genomes: the problem of biased gene conversion. *Philos. Trans. R. Soc. Ser. B* **365**, 2571–2580 (2010).
187. Guéguen, L. et al. Bio++: efficient extensible libraries and tools for computational molecular evolution. *Mol. Biol. Evol.* **30**, 1745–1750 (2013).
188. Wickham, H. & Grolemund, G. *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data* (O'Reilly Media, Inc., Canada, 2016).
189. Wilke, C. O. cowplot: streamlined plot theme and plot annotations for 'ggplot2.' *CRAN Repos.* **2**, R2 (2016).
190. Gouy, M., Guindon, S. & Gascuel, O. SeaView version 4: a multiplatform graphical user interface for sequence alignment and phylogenetic tree building. *Mol. Biol. Evol.* **27**, 221–224 (2010).
191. Redelings, B. Erasing errors due to alignment ambiguity when estimating positive selection. *Mol. Biol. Evol.* **31**, 1979–1993 (2014).
192. Mi, H., Muruganujan, A., Ebert, D., Huang, X. & Thomas, P. D. PANTHER version 14: More genomes, a new PANTHER GO-slim and improvements in enrichment analysis tools. *Nucleic Acids Res.* **47**, D419–D426 (2019).
193. Huerta-Cepas, J. et al. Fast genome-wide functional annotation through orthology assignment by eggNOG-mapper. *Mol. Biol. Evol.* **34**, 2115–2122 (2017).

194. Huerta-Cepas, J. et al. eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res.* **47**, D309–D314 (2019).

### Acknowledgements

This project has received funding from the Marie Curie International Outgoing Fellow under the European Union's Seventh Framework Programme (project BIOMME, agreement no. 627684), a PICS grant from the CNRS (project PASTA), an 'Investissement d'Avenir' grant from the Agence Nationale de la Recherche (project CASMA, CEBA, ref. ANR-10-LABX-25-01) and the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (project GAIA, agreement no. 851188) to F.L.C.; a Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery Grant (RGPIN-2018-04920) to F.A.H.S.; and a German Research Foundation grant (WA 2461/9-1) to S.W. We are grateful to Sophie Dang, Troy Locke and Corey Davis at the Molecular Biology Service Unit of the University of Alberta for their help, assistance, and advice on next-generation sequencing. The analyses benefited from the Montpellier Bioinformatics Biodiversity (MBB) platform services. Finally, we are grateful to Seth Bybee, Frédéric Delsuc, Claude dePamphilis, Krushna-megh Kunte, Conrad Labandeira, Harald Letsch, Sören Nylin, Timothy O'Hara, Susanne Renner and Chris Wheat for helpful comments and discussions on earlier drafts of the study.

### Author contributions

F.L.C. and F.A.H.S. designed and conceived the research. R.A. and F.L.C. assembled the phylogenetic data for swallowtail butterflies. S.W., O.A.P.-E., G.C., F.L.C. and R.A. assembled the phylogenetic data for birthworts. R.A. and F.L.C. analysed the phylogenetic data. R.A. and F.L.C. performed the estimations of the ancestral states. F.L.C. performed the diversification analyses. A.-L.C. and F.L.C. generated the genomic data. R.A. and B.N. assembled and analysed the genomic data. All authors contributed to the interpretation and discussion of results. R.A. and F.L.C. drafted the paper with substantial input from all authors.

### Competing interests

The authors declare no competing interests.

### Additional information

**Supplementary information** is available for this paper at <https://doi.org/10.1038/s41467-020-20507-3>.

**Correspondence** and requests for materials should be addressed to R.A. or F.L.C.

**Peer review information** *Nature Communications* thanks Niklas Wahlberg, and the other, anonymous, reviewers for their contribution to the peer review of this work. Peer reviewer reports are available.

**Reprints and permission information** is available at <http://www.nature.com/reprints>

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2021