

A global phylogeny of butterflies reveals their evolutionary history, ancestral hosts and biogeographic origins

Received: 8 May 2022

Accepted: 16 March 2023

Published online: 15 May 2023

 Check for updates

A list of authors and their affiliations appears at the end of the paper

Butterflies are a diverse and charismatic insect group that are thought to have evolved with plants and dispersed throughout the world in response to key geological events. However, these hypotheses have not been extensively tested because a comprehensive phylogenetic framework and datasets for butterfly larval hosts and global distributions are lacking. We sequenced 391 genes from nearly 2,300 butterfly species, sampled from 90 countries and 28 specimen collections, to reconstruct a new phylogenomic tree of butterflies representing 92% of all genera. Our phylogeny has strong support for nearly all nodes and demonstrates that at least 36 butterfly tribes require reclassification. Divergence time analyses imply an origin ~100 million years ago for butterflies and indicate that all but one family were present before the K/Pg extinction event. We aggregated larval host datasets and global distribution records and found that butterflies are likely to have first fed on Fabaceae and originated in what is now the Americas. Soon after the Cretaceous Thermal Maximum, butterflies crossed Beringia and diversified in the Palaeotropics. Our results also reveal that most butterfly species are specialists that feed on only one larval host plant family. However, generalist butterflies that consume two or more plant families usually feed on closely related plants.

Butterflies have long captivated naturalists, scientists and the public, and they have played a central part in studies of speciation, community ecology, plant–insect interactions, mimicry, genetics and conservation. Despite being the most intensely studied insect group, the evolutionary history and drivers of butterfly diversification remain poorly understood^{1,2}. Butterflies are thought to have diversified in relation to multiple abiotic and biotic factors, including adaptations to novel climates and species interactions, with caterpillar–host interactions and geographic history playing a major role³. However, these hypotheses have not been tested because a robust phylogenetic framework at the taxonomic scale that would be needed to examine their evolution has not been available. Furthermore, host plant and distribution data have largely been scattered across literature, museum collections, and

local databases, limiting our ability to conduct broad, comparative macroevolutionary studies.

We sequenced 391 genes from nearly 2,300 butterfly species to reconstruct a new phylogenomic tree of butterflies representing 92% of all genera (Fig. 1 and Supplementary Fig. 1), assembled a comprehensive host association dataset and aggregated global distribution records. Using our tree, we inferred the evolutionary timing, patterns of host use, and biogeographic history of butterflies. We addressed three long-standing questions related to butterfly evolution: (1) did butterflies originate in the northern (Laurasia) or southern (Gondwana) hemisphere⁴; (2) what plants did the ancestor of butterflies feed on⁵; and (3) are host repertoires (that is, diets) of butterfly species and clades constrained by host phylogeny^{6,7}?

✉ e-mail: kawahara@flmnh.ufl.edu; npierce@oeb.harvard.edu; dlohman@ccny.cuny.edu

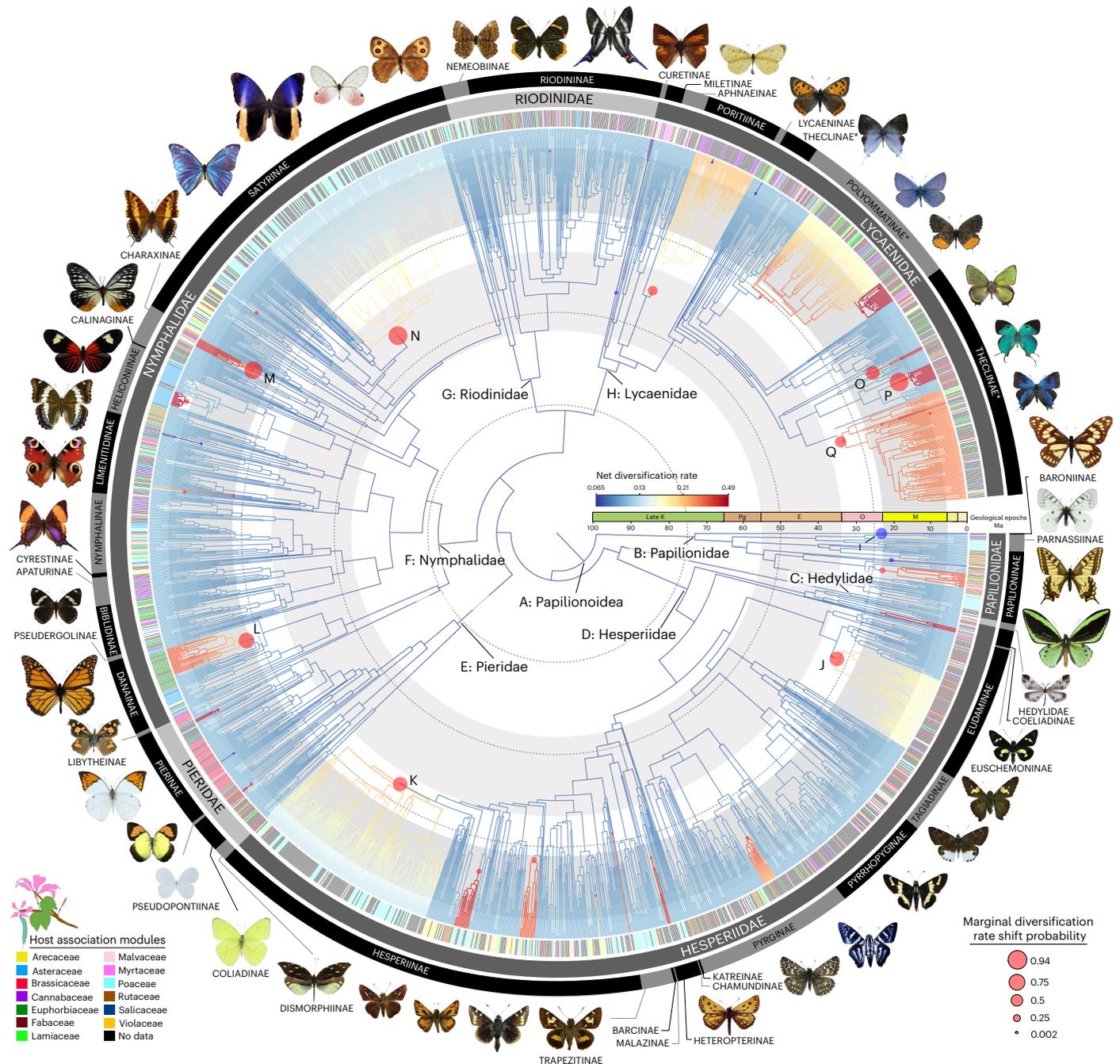


Fig. 1 | Evolutionary relationships and diversification patterns of butterflies. Time-calibrated tree of 2,244 butterfly species based on 391 loci and 150 amino acid partitions. Branches show distinct changes in diversification (circles) as estimated by clade-specific models. Letters at nodes refer to clades with significant rate shifts (see section 6 of Supplementary Results). Coloured lines

in the outer ring beside tips indicate association with one of the 13 host modules (see section 17 of Extended Online Methods). Black lines in the host association ring indicate species without data, and asterisks denote non-monophyletic subfamilies. Supplementary Fig. 1 shows this tree with visible species names and ages for all nodes.

Results and discussion

To elucidate patterns of global butterfly diversification in space and time, we used targeted exon capture⁸ to assemble a dataset of 391 gene regions (161,166 nucleotides and 53,722 amino acids) from 2,244 butterfly species (Supplementary Table 1). The majority (1,914 specimens) of butterflies sampled were newly sequenced for this study, representing all families, subfamilies and tribes, and 92% of recognized genera, from 90 countries. These were obtained from 28 specimen collections across the world (see section 2 of the Extended Online Methods). Phylogenomic trees were inferred with nucleotides

or translated amino acids with nine different subsets and partitioning schemes. Our trees were highly congruent, with strong support for the monophyly of all families and nearly all subfamilies with branch support metrics (SH-aLRT, ultrafast bootstrap) and multispecies coalescent species tree analyses (Supplementary Table 2). We also conducted four-cluster likelihood mapping to identify potentially conflicting signals in our datasets (Supplementary Table 3). Our results strongly support the need for revision of the classification of at least 36 butterfly tribes (27% of total) as currently circumscribed (Supplementary Table 2).

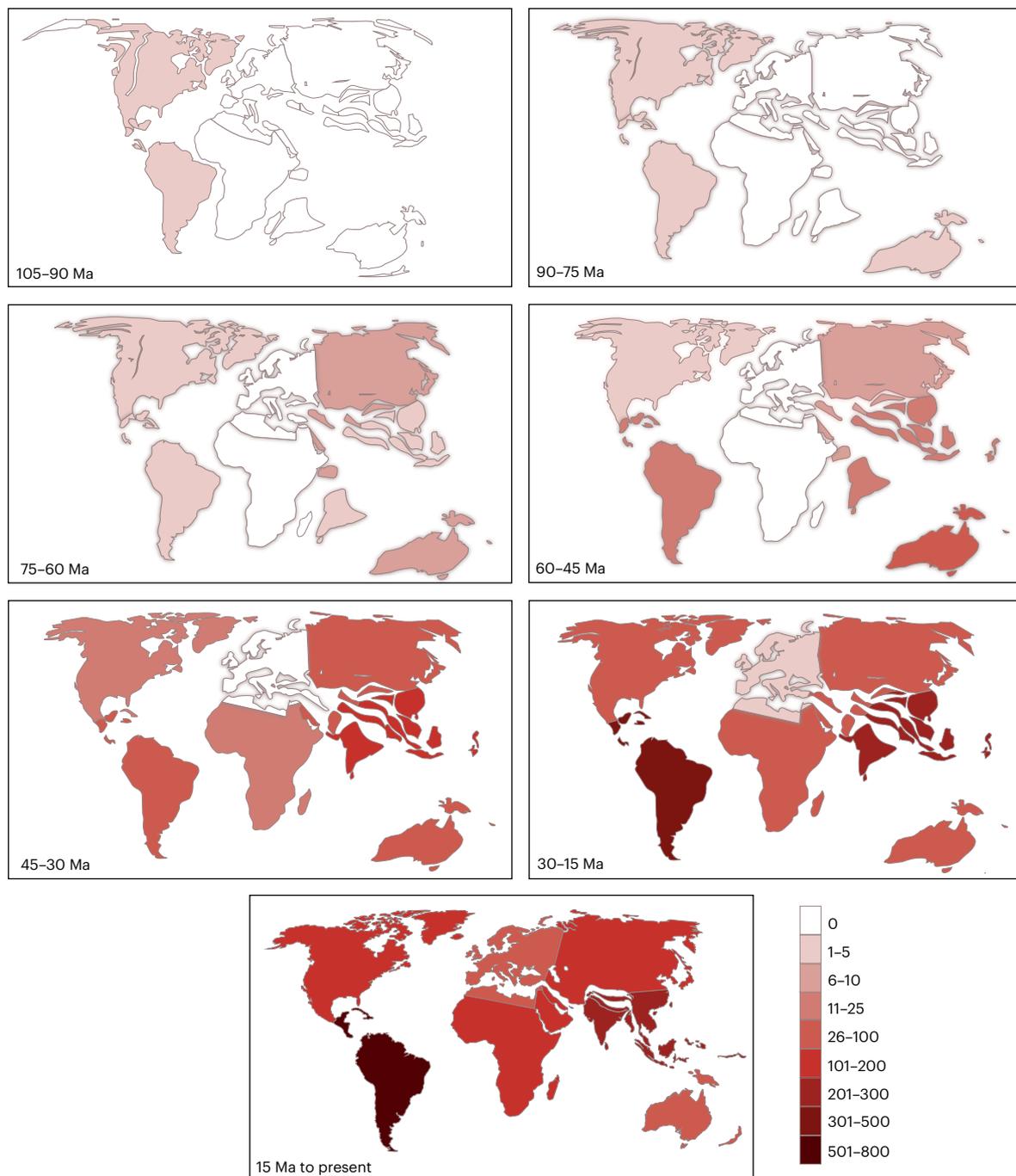


Fig. 2 | Distribution of butterflies over time. Bioregion shading indicates the number of butterfly lineages that were associated with that bioregion during that time period, as determined by BioGeoBEARS ancestral state reconstruction. Each map corresponds to a 15-Ma interval of butterfly evolution. Results are based on data from this study.

We conducted 24 dating analyses using different fossil and secondary calibration schemes along with sensitivity analyses to assess the impact of analytical and sampling bias. Across analyses, our results revealed largely congruent timing of butterfly divergence events (Supplementary Table 4). Butterflies originated from nocturnal, herbivorous moth ancestors around 101.4 million years ago (Ma) (102.5–100.0 Ma), providing evidence for a mid-Cretaceous origin of butterflies^{2,9}.

To determine the geographic origin of butterflies, we used our dated tree (Fig. 1) to conduct a global biogeographic analysis with 15,764 newly aggregated country-level distribution records (Supplementary

Table 5). Modelling with three different area categorizations, models of range evolution and parameters (adjacency matrices, time slices, etc.) consistently recovered butterflies as originating in the Americas, in what is present-day western North America or Central America (Fig. 2 and Supplementary Tables 6 and 7). All extant butterfly families excluding the Neotropical Hedyliidae diversified -10–30 Ma after the Cretaceous Thermal Maximum, -90 Ma, when the global climate cooled by nearly 5 °C (ref. 10) (Figs. 1 and 2). During the Cretaceous, butterflies dispersed out of the Neotropics at a much higher rate than that of any other dispersal route (Supplementary Fig. 2). As new butterfly lineages became established in other bioregions, interbioregion

Our study provides a robust baseline for future studies of this model insect lineage. The consistency of results obtained using different approaches for each of our analyses suggests that our conclusions are robust. Our data support the hypothesis that butterflies originated in the Americas in the late Cretaceous, 100 million years after the origin of angiosperms, and that they first fed on legumes. Butterflies dispersed from the Americas to the Eastern Palearctic across Beringia ~75 Ma before diversifying in the Palaeotropics. Although our analyses point to a Nearctic origin, evidence for a North American versus a Central American origin is not strong, and we therefore tentatively conclude that a Laurasian origin is likely. Larval host plants played an important part in the evolution of butterflies, and some groups became host specific whereas others retained a wide host breadth. The molecular, host plant and geographic data provided here serve as a baseline for future comparative analyses of butterflies.

Methods

Taxon sampling and sequence acquisition

A total of 2,248 butterfly specimens representing 2,244 species in 1,644 genera were included for the molecular component of this study, along with ten outgroups from other lepidopteran superfamilies (Supplementary Table 1). The ingroup included genera from all families, subfamilies and tribes of butterflies according to the current classification. We aimed to include at least one species from every valid genus and sequenced the type species of each genus whenever possible. We obtained 92% of all described valid butterfly genera when the initial dataset was assembled (July 2019).

We obtained marker loci for phylogenetic analysis by (1) anchored hybrid enrichment exon capture of DNA extracts and subsequent Illumina sequencing³⁰ or (2) bioinformatically removing these sequences from published genomes and transcriptomes. We used the BUTTERFLY1.0 probe set⁸ and selected a 391-locus subset that was captured reliably in at least 60% of samples. We chose this approach because it has been proven to resolve relationships of many different butterfly groups^{31–34}. The BUTTERFLY1.0 probe set includes 13 genes (12 nuclear genes and the COI mitochondrial gene) that have been widely used in butterfly phylogenetics^{9,35}, also termed ‘legacy genes’³⁶, and additional protein-coding genes that may be used to address broad questions pertaining to butterfly biology, such as vision, host use and olfaction⁸.

Specimens were collected in 90 countries over a 70-year period by over 300 people and deposited in one of the 28 specimen collections from which we obtained tissue samples (Supplementary Table 1). We successfully captured and sequenced DNA from decades-old museum specimens³⁷, which enabled us to include taxa that are rare or live in areas where collecting fresh material is difficult. The oldest sample was a pinned specimen collected on 22 April 1946: *Dira clytus* (Nymphalidae) (LEP79391). Images of 460 representative voucher specimens are shown in Supplementary Data 2, and specimen repositories are listed in Supplementary Table 1. All voucher specimens, at minimum, had their wings and genitalia retained for identification and future research.

We obtained sequence data from 343 published genomes and transcriptomes. Ten of these were outgroups representing nine moth families that are closely related to butterflies according to published studies on lepidopteran phylogeny^{9,38–42}.

We extracted DNA from 1,915 specimens that were (1) stored in ethanol and frozen; (2) dried and stored in glassine envelopes under ambient conditions (papered); or (3) dried, spread and pinned in a museum collection. Locus assembly and sequence clean-up followed the pipeline of Breinholt et al.⁴². Published sequences comprised (1) genome assemblies, (2) genomic reads, and (3) paired or (4) single-end transcriptomes. Three sequence datasets were created for this study: a nucleotide dataset with all codon positions (nt123); a nucleotide dataset that excludes all synonymous changes (degen), created using the Perl script Degen1 v.1.4 (refs. 43,44); and an amino acid (aa) dataset translated from the nt123 dataset (Supplementary Data 3).

Phylogenetic analysis and dating

Maximum likelihood (ML) tree inference was conducted on all three datasets (nt123, degen and aa) in IQ-TREE 2.0 (ref. 45); parameter settings for each analysis can be found in Supplementary Table 14. Branch support was calculated with 1,000 ultrafast bootstrap replicates (UFBS; ‘-B 1000’ command)^{46,47} and Shimodaira–Hasegawa approximate likelihood ratio tests (SH-aLRT; ‘-alrt 1000’ command)⁴⁸. Quartet sampling was performed on the degen359 and aa154 trees with the highest likelihood score. Four-cluster likelihood mapping analyses⁴⁹ were performed on the degen and aa datasets to assess the placement of particular butterfly clades that have been the subject of previous phylogenetic studies. We applied this approach in addition to standard branch support metrics, because the latter can be subject to inflated estimates⁴⁹.

We obtained divergence time estimates using a penalized-likelihood based approach implemented in treePL⁵⁰. We implemented three different methods for calibrating trees and assessed similarities among results. Method 1 involved dating with secondary calibrations only. We used the 95% credibility intervals of Lepidoptera ages from Fig. S12 of Kawahara et al.³⁸ to assign minimum and maximum ages to 27 ingroup and six outgroup nodes in our tree. Method 2 involved dating with fossils and one secondary root calibration. In this approach, we followed the guidelines of Parham et al.⁵¹ by calibrating nodes with 11 butterfly fossils that could be assigned to the geological age of a butterfly lineage with confidence as verified by de Jong⁵². None of the outgroup nodes could be calibrated because reliable fossils associated with our non-butterfly Lepidoptera were too young to influence deeper node ages representing multisuperfamily clades. Consequently, preliminary treePL analyses yielded highly dubious age estimates for deep nodes on the tree, hundreds of millions of years older than expected based on the literature. We therefore added a single secondary calibration to the root of the tree. Although combining secondary and fossil calibrations in a single analysis can create redundancy that negatively affects the resulting age estimates⁵³, the limited fossil record of Lepidoptera made it a necessity to obtain comparable results derived primarily from fossils. We ran two versions of this method, each with a different root calibration. Method 2A used a maximum-age estimate of 139.4 Ma, based on the angiosperm age estimate of Smith and Brown¹⁷. Method 2B used a more conservative maximum-age estimate of 251 Ma, based on the older end of the credibility interval for the age of angiosperms in Foster et al.⁵⁴. Both calibrations were used under the assumption that butterflies diverged from their moth ancestors after their most frequently used host plants, angiosperms, were already present^{55,56}. Method 3 involved secondary calibrations and six fossils. In this approach, we combined the 33 secondary calibrations from Method 1 with six fossil calibrations, including some of the fossils used in Method 2. Fossils previously used to calibrate trees of Kawahara et al.³⁸ were excluded from this analysis to avoid circularity and redundancy with secondary calibrations. Whenever possible, redundant fossil calibrations from Method 2 were replaced with calibrations from unrelated fossils that could be associated with a different node in the same clade.

Diversification rate analyses

We performed a Bayesian analysis of macroevolutionary mixtures using the program BAMM v.1.10.4 (ref. 57) to detect shifts in diversification rates between clades. Reversible-jump Markov chain Monte Carlo was run for 50 million generations and sampled every 50,000 generations. Priors were estimated with the R package BAMMtools v.2.1.6 (ref. 58) using the command ‘setBAMMpriors’. The tree was trimmed in Mesquite v.3.6 (ref. 59) to remove all outgroups. Six analyses were performed using different priors for expected numbers of shifts (5, 10, 20, 30, 40 and 50 shifts).

We conducted a series of analyses in HiSSE (Hidden State Speciation and Extinction) and a BiSSE-like (Binary State Speciation and Extinction) implementation of HiSSE⁶⁰ in the R package hisse⁶¹ to

evaluate whether there is a correlation between butterfly and plant diversification. We pruned outgroups from the aa154 dated tree (Strategy A) and compared 20 HiSSE models and BiSSE-like implementations of HiSSE. The BiSSE equivalent of HiSSE tests whether there are different diversification rates associated with the two host plant use states. Other models were built in the HiSSE framework to test alternative combinations of the presence or absence of hidden state and host plant use associations while also considering different transition rate matrices, net turnover rates, τ_i (speciation plus extinction: $\lambda_i + \mu_i$) and extinction fractions, e_i (extinction divided by speciation: μ_i/λ_i) (Supplementary Table 15). We tested whether diversification rates were linked to feeding (A) as a larval specialist or generalist (Supplementary Table 16); (B) on Poales (Supplementary Table 17) in Papilionoidea, Hesperioidea and Nymphalidae; (C) on Fabales (Supplementary Table 18) in Papilionoidea and Nymphalidae; (D) on Brassicales (Supplementary Table 19) in Papilionoidea and Pieridae; (E) on Fagales (Supplementary Table 20); (F) on the Poaceae module (Supplementary Table 21); (G) on the Fabaceae module (Supplementary Table 22); and (H) on Fabaceae in Eudaminae (Supplementary Tables 22 and 23). We compared these different models of HiSSE and BiSSE-like implementations to account for hidden states to alleviate concerns that SSE models can lead to a high incidence of false positive results⁶².

Fraction files of clade-based taxonomic diversity estimates were created for all HiSSE runs to account for taxonomic sampling bias (Supplementary Table 24). We set the total number of extant butterfly species as 19,500, which is an ~8% increase compared with the butterfly species richness estimate of van Nieukerken et al.⁶³. We added this diversity correction based on many recent new butterfly species descriptions (for example, by Cong et al.⁶⁴) and morphospecies that we are aware of that have not yet been formally described. We estimated the total number of generalist and specialist species by calculating the percentage of generalists and specialists in our dataset at the family level. We standardized the proportion of species richness in that family compared to all butterflies, based on diversity estimates of van Nieukerken et al.⁶³. For example, 78.61% of all sampled Hesperioidea that had host data were specialists, and Hesperioidea comprise 21.91% of all butterfly species richness; thus, we estimated Hesperioidea specialists as $19,500 \times 0.2191 \times 0.7861 = 3,359$ species. Applying these calculations for all families yielded totals of 12,969 specialist species and 6,531 generalist species (Supplementary Table 25); these numbers were used to estimate fractions of generalists and specialists in our dataset.

Calculating the fraction of species sampled within each host plant module proved more challenging. To estimate the true butterfly species richness for each module, we used unpublished estimates of species richness for all butterfly genera by G.L. and assumed that if a species was known to belong to a module, so would some of its congeners. These calculations were revised because some genera had large host ranges with species assigned to multiple modules. For example, the three species of *Vanessa* with host records in our dataset were assigned to three different modules. As there is an estimated total of 24 *Vanessa* species, we calculated that approximately $24/3 = 8$ *Vanessa* species belonged in each of those modules. Calculations for all genera in all modules, and the resulting estimates of module totals and fractions sampled, are provided in Supplementary Table 26.

Biogeography

To reconstruct the biogeographic history of butterflies, we aggregated global distribution data from multiple sources to create a butterfly checklist for each country. Data sources included: (1) the Lepidoptera and Other Life Forms Database (http://ftp.funet.fi/index/Tree_of_life/insecta/lepidoptera); (2) WikiSpecies (<https://species.wikimedia.org>); and (3) the type locality of each species or subspecies in our list of valid butterfly names, which was obtained from 1, above. This initial global checklist was vetted using published country checklists and the ButterflyNet Trait Database⁶⁵. Trait data from ca. 100

comprehensive and country-specific field guides have been entered into this database, allowing us to generate species lists to cross-validate checklists assembled⁶⁶.

We designated 14 biogeographic regions across the globe (Supplementary Fig. 11 and Supplementary Table 27), determined which of these regions were occupied by each species in our tree and developed a 14-state character matrix. Six countries (Canada, China, Indonesia, Mexico, Russia, US) spanned two or three bioregions, which required manual evaluation of whether species in these countries were found in one or more of the adjoining bioregions. US and Canadian species were assigned to East and/or West Nearctic bioregions based on the palaeogeographic history of North America (that is, whether the species were east or west of the continental divide) with reference to locality records from Butterflies and Moths of North America (<https://www.butterfliesandmoths.org>). Russian species were assigned to Eastern and/or Western Palaearctic bioregions based on locality records assembled by the Lepidoptera and Other Life Forms Database⁶⁷. Some countries did not have complete distribution lists and were thus evaluated manually by coauthors. Chinese species were assigned to Eastern Palaearctic and Oriental bioregions by H.W. Indonesian species were assigned to Oriental, Wallacean and Australian bioregions by D.J.L. and D.P. Mexican species were assigned to East Nearctic, West Nearctic and Central American bioregions by J.I.M.

The majority of butterfly species are distributed in fewer than five bioregions. Some species are more widespread, but we found that this was often due to recent anthropogenic introductions. Consequently, a final round of data cleaning was performed in which records of species found in at least five bioregions were manually verified and edited to accurately reflect true native species' ranges. Cleaned bioregion and tropicality data were converted to character matrices to be used for subsequent distribution analyses (Supplementary Tables 28 and 29).

We estimated the ancestral area of origin and geographic range evolution for butterflies using two approaches: the ML approach of the DECX model⁶⁸ as implemented in the C++ version^{69,70} (<https://github.com/champost/DECX>); and the program BioGeoBEARS v.1.1.2 (ref. 71). DECX uses a time-calibrated tree, the modern 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 evolutionary history of clade of interest⁷².

We also ran BioGeoBEARS with seven and eight areas to estimate immigration and emigration rates (Supplementary Figs. 12 and 13 and Supplementary Table 27). BioGeoBEARS could not be run with 14 states owing to the complexity of our dataset (2,248 tree tips). The seven and eight bioregions largely corresponded to the biogeographic realms defined by Udvardy⁷³. We implemented both the Dispersal Extinction Cladogenesis (DEC)^{68,74} and the Likelihood equivalent of the Dispersal-Vicariance approach (DIVALIKE)⁷⁵ models and different adjacency matrices (Supplementary Data 4). Both approaches gave largely consistent results, regardless of the model and parameters used (Supplementary Tables 6 and 30).

We performed biogeographic stochastic mapping to examine in situ speciation, immigration and emigration between the seven bioregions in BioGeoBEARS. We followed Li et al.⁷⁶ and ran 1,000 simulations with the DEC model, and calculated relative mean dispersal rates between all permutations of bioregions (Fig. 3 and Supplementary Data 5). These mean dispersal rates represent dispersal of butterfly lineages throughout the entire evolutionary history of Papilionoidea and thus cannot reveal changes in rates over time. To look at historical biogeography of butterflies during different epochs, rates along all possible interbioregion colonization rates were calculated at specific time intervals of 5 million years (Supplementary Table 31). These relative rates were averaged to represent relevant geological time periods (Supplementary Figs. 2–4).

Larval host plant analyses

Larval host records were compiled from nine sources: (1) the Database of the World's Lepidopteran Hostplants (HOSTS)²¹, which summarizes data from ~270 other sources; (2) the Lepidoptera and Other Life Forms Database (http://ftp.funet.fi/index/Tree_of_life/insecta/lepidoptera/); (3) 40 years of food plant rearing records from Costa Rica by D.H.J., W.H., and colleagues (<http://janzen.sas.upenn.edu/>); (4) the ButterflyNet Trait Database⁶⁵, which includes host plant records from 109 butterfly field guides and other resources; (5) a comprehensive database for host records for all butterflies in Japan⁷⁷; (6) a set of papers documenting the hosts of butterflies in India^{78–84}; (7) a database of hosts and ant symbionts of larval Lycaenidae and Riodinidae compiled from 85 literature sources by N.E.P. and members of her laboratory; (8) a database of butterfly host records from Ecuador based on field observations and literature records compiled by K.R.W.; and (9) 88 papers from the primary literature or relevant websites (Supplementary Table 9 and Supplementary Data 6). Whenever possible, we retained the following information for each host record, if available: (1) the taxon and taxonomic authority of butterfly to the lowest available taxonomic level (family, subfamily, tribe, genus, species or subspecies); (2) the taxon and taxonomic authority of host to the lowest available taxonomic level (family, genus, species, subspecies or variety); (3) plant part eaten; (4) record certainty (novel plant accepted in captivity, oviposition record with no observation of herbivory, etc.); (5) geographic location of observation; and (6) relevant information on all non-plant hosts. The extensive data recorded in the host (food plant) database of D.H.J., W.H., and colleagues were simplified to retain the fields of butterfly genus and specific epithet, as well as plant family, genus and specific epithet, together with an indication of whether the plant was introduced to Costa Rica. This database contains many records of informal, non-ICZN-compliant names of butterfly cryptic species. Rather than discarding the large number of records that would not be compatible with any other data source, we regarded these as the nominal species (for example, *Battus polydamas* instead of *Battus polydamas*DHJ01). The number of records for each butterfly species × plant species interaction was recorded.

We examined relationships between individual butterfly species and host families that are consumed by their larvae. For these analyses, we chose the rank of plant family because it has been adopted as the standard taxonomic rank for examining host use evolution^{6,85}. For each plant-feeding butterfly species in our tree, we quantified host plant richness and phylogenetic distance using six different metrics implemented in the R package *picante* v.1.8.2 (ref. 86). To calculate these metrics, we used the calibrated tree of seed plants from Smith and Brown¹⁷.

As the number of host groups in our dataset was too large for an ancestral state reconstruction (approximately 200 of the 300 known host plant families²¹ plus host insects), we first reduced the number of host groups by using a network analysis. The Beckett algorithm⁸⁷, as implemented in the function 'computeModules' from the package *bipartite*⁸⁸ in R v.3.6.2 (ref. 89), assigns plants and butterflies to modules and computes the modularity index, Q. By maximizing Q, the algorithm finds groups of butterflies and hosts that interact more with each other than with other taxa in the network. Thus, hosts that are assigned to the same module tend to be used by the same butterflies. We found 13 modules for butterfly host associations in our module analysis (Supplementary Tables 32 and 33). We then conducted three larval host ancestral state reconstruction analyses using stochastic character mapping with SIMMAP in phytools v.0.7-70 (refs. 90,91) using the 'make.simmap' command. We reconstructed the ancestral state of (A) generalist versus specialist feeding (two states, Supplementary Data 7); (B) plant, lichen, Hemiptera or Hymenoptera as a food source (four states, Supplementary Data 8); and (C) plant module (13 states, Supplementary Data 9).

Reporting summary

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

Data availability

All supplementary data archives are available on Figshare (<https://doi.org/10.6084/m9.figshare.21774899>). Genomic data for all newly sequenced specimens in this study have been uploaded to GenBank as part of BioProject PRJNA714105. Individual BioSample accession numbers for each specimen are provided in Supplementary Table 1.

Code availability

All new code developed to aid with the analyses in this study has been uploaded to GitHub and made publicly available. GitHub URLs for specific scripts are provided in the Methods and Extended Online Methods sections.

References

- Chazot, N. et al. Priors and posteriors in Bayesian timing of divergence analyses: the age of butterflies revisited. *Syst. Biol.* **68**, 797–813 (2019).
- Allio, R. et al. Whole genome shotgun phylogenomics resolves the pattern and timing of swallowtail butterfly evolution. *Syst. Biol.* **69**, 38–60 (2020).
- Boggs, C. L., Watt, W. B. & Ehrlich, P. R. *Butterflies: Ecology and Evolution Taking Flight* (University of Chicago Press, 2003).
- Braby, M. F., Trueman, J. W. H. & 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).
- Janz, N. & Nylin, S. Butterflies and plants: a phylogenetic study. *Evolution* **52**, 486–502 (1998).
- Braga, M. P., Landis, M. J., Nylin, S., Janz, N. & Ronquist, F. Bayesian inference of ancestral host–parasite interactions under a phylogenetic model of host repertoire evolution. *Syst. Biol.* **69**, 1149–1162 (2020).
- Braga, M. P., Janz, N., Nylin, S., Ronquist, F. & Landis, M. J. Phylogenetic reconstruction of ancestral ecological networks through time for pierid butterflies and their host plants. *Ecol. Lett.* **24**, 2134–2145 (2020).
- Espeland, M. et al. A comprehensive and dated phylogenomic analysis of butterflies. *Curr. Biol.* **28**, 770–778.e5 (2018).
- Wahlberg, N., Wheat, C. W. & Peña, C. Timing and patterns in the taxonomic diversification of Lepidoptera (butterflies and moths). *PLoS ONE* **8**, e80875 (2013).
- Linnert, C. et al. Evidence for global cooling in the Late Cretaceous. *Nat. Commun.* **5**, 4194 (2014).
- Domingo, L., Tomassini, R. L., Montalvo, C. I., Sanz-Pérez, D. & Alberdi, M. T. The Great American Biotic Interchange revisited: a new perspective from the stable isotope record of Argentine Pampas fossil mammals. *Sci. Rep.* **10**, 1608 (2020).
- Carrillo, J. D. et al. Disproportionate extinction of South American mammals drove the asymmetry of the Great American Biotic Interchange. *Proc. Natl Acad. Sci. USA* **117**, 26281–26287 (2020).
- Rolland, J., Condamine, F. L., Beeravolu, C. R., Jiguet, F. & Morlon, H. Dispersal is a major driver of the latitudinal diversity gradient of Carnivora. *Glob. Ecol. Biogeogr.* **24**, 1059–1071 (2015).
- Condamine, F. L., Silva-Brandão, K. L., Kergoat, G. J. & Sperling, F. A. H. Biogeographic and diversification patterns of neotropical Troidini butterflies (Papilionidae) support a museum model of diversity dynamics for Amazonia. *BMC Evol. Biol.* **12**, 82 (2012).
- Chazot, N. et al. Conserved ancestral tropical niche but different continental histories explain the latitudinal diversity gradient in brush-footed butterflies. *Nat. Commun.* **12**, 5717 (2021).
- 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).

17. Smith, S. A. & Brown, J. W. Constructing a broadly inclusive seed plant phylogeny. *Am. J. Bot.* **105**, 302–314 (2018).
18. Li, H.-T. et al. Origin of angiosperms and the puzzle of the Jurassic gap. *Nat. Plants* **5**, 461–470 (2019).
19. Heikkilä, M., Kaila, L., Mutanen, M., Peña, C. & Wahlberg, N. Cretaceous origin and repeated tertiary diversification of the redefined butterflies. *Proc. R. Soc. B Biol. Sci.* **279**, 1093–1099 (2011).
20. Braga, M. P., Guimarães, P. R., Wheat, C. W., Nylin, S. & Janz, N. Unifying host-associated diversification processes using butterfly–plant networks. *Nat. Commun.* **9**, 5155 (2018).
21. Robinson, G. S., Ackery, P. R., Kitching, I. J., Beccaloni, G. W. & Hernández, L. M. *HOSTS – A Database of the World's Lepidopteran Hostplants* (accessed 1 August 2020); <http://www.nhm.ac.uk/our-science/data/hostplants/>
22. Kaliszewska, Z. A. et al. When caterpillars attack: biogeography and life history evolution of the Miletinae (Lepidoptera: Lycaenidae). *Evolution* **69**, 571–588 (2015).
23. Pierce, N. E. et al. The ecology and evolution of ant association in the Lycaenidae (Lepidoptera). *Annu. Rev. Entomol.* **47**, 733–771 (2002).
24. Moreau, C. S. & Bell, C. D. Testing the museum versus cradle tropical biological diversity hypothesis: phylogeny, diversification, and ancestral biogeographical range evolution of the ants. *Evolution* **67**, 2240–2257 (2013).
25. Forister, M. L. et al. The global distribution of diet breadth in insect herbivores. *Proc. Natl Acad. Sci. USA* **112**, 442 (2015).
26. Kelly, C. K. & Southwood, T. R. E. Species richness and resource availability: a phylogenetic analysis of insects associated with trees. *Proc. Natl Acad. Sci. USA* **96**, 8013–8016 (1999).
27. Kennedy, C. E. J. & Southwood, T. R. E. The number of species of insects associated with British trees: a re-analysis. *J. Anim. Ecol.* **53**, 455–478 (1984).
28. Rutz, A. et al. The LOTUS initiative for open knowledge management in natural products research. *eLife* **11**, e70780 (2022).
29. Ehrlich, P. R. & Raven, P. H. Butterflies and plants: a study of coevolution. *Evolution* **18**, 586–608 (1965).
30. Lemmon, A. R., Emme, S. A. & Lemmon, E. M. Anchored hybrid enrichment for massively high-throughput phylogenomics. *Syst. Biol.* **61**, 727–744 (2012).
31. Valencia-Montoya, W. A. et al. Evolutionary trade-offs between male secondary sexual traits revealed by a phylogeny of the hyperdiverse tribe Eumaeini (Lepidoptera: Lycaenidae). *Proc. R. Soc. B Biol. Sci.* **288**, 1–10 (2021).
32. Toussaint, E. F. A. et al. Anchored phylogenomics illuminates the skipper butterfly tree of life. *BMC Evol. Biol.* **18**, 101 (2018).
33. Espeland, M. et al. Four hundred shades of brown: higher level phylogeny of the problematic Euptychiina (Lepidoptera, Nymphalidae, Satyrinae) based on hybrid enrichment data. *Mol. Phylogenet. Evol.* **131**, 116–124 (2019).
34. Carvalho, A. P. S. et al. Diversification is correlated with temperature in white and sulfur butterflies. Preprint at *bioRxiv* <https://doi.org/10.1101/2022.09.22.509088> (2022).
35. Kawahara, A. Y. et al. Phylogenetics of moth-like butterflies (Papilionoidea: Hedyliidae) based on a new 13-locus target capture probe set. *Mol. Phylogenet. Evol.* **127**, 600–605 (2018).
36. Toussaint, E. F. A. et al. Afrotropics on the wing: phylogenomics and historical biogeography of awl and policeman skippers. *Syst. Entomol.* **46**, 172–185 (2021).
37. Nunes, R. et al. Predictors of sequence capture in a large-scale anchored phylogenomics project. *Front. Ecol. Evol.* **10**, 943361 (2022).
38. Kawahara, A. Y. et al. Phylogenomics reveals the evolutionary timing and pattern of butterflies and moths. *Proc. Natl Acad. Sci. USA* **116**, 22657–22663 (2019).
39. Regier, J. C. et al. A large-scale, higher-level, molecular phylogenetic study of the insect order Lepidoptera (moths and butterflies). *PLoS ONE* **8**, 1–23 (2013).
40. Mayer, C. et al. Adding leaves to the Lepidoptera tree: capturing hundreds of nuclear genes from old museum specimens. *Syst. Entomol.* **46**, 649–671 (2021).
41. Rota, J. et al. The unresolved phylogenomic tree of butterflies and moths (Lepidoptera): assessing the potential causes and consequences. *Syst. Entomol.* **47**, 531–550 (2022).
42. Breinholt, J. W. et al. Resolving relationships among the megadiverse butterflies and moths with a novel pipeline for anchored phylogenomics. *Syst. Biol.* **67**, 78–93 (2018).
43. Regier, J. C. et al. Arthropod relationships revealed by phylogenomic analysis of nuclear protein-coding sequences. *Nature* **463**, 1079–1083 (2010).
44. Zwick, A. Degeneracy coding web service. *GitHub* <https://github.com/carlosp420/degenerate-dna> (2010).
45. Minh, B. Q. et al. IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. *Mol. Biol. Evol.* **37**, 1530–1534 (2020).
46. Minh, B. Q., Nguyen, M. A. T. & von Haeseler, A. Ultrafast approximation for phylogenetic bootstrap. *Mol. Biol. Evol.* **30**, 1188–1195 (2013).
47. Hoang, D. T., Chernomor, O., von Haeseler, A., Minh, B. Q. & Vinh, L. S. UFBoot2: improving the ultrafast bootstrap approximation. *Mol. Biol. Evol.* **35**, 518–522 (2018).
48. Guindon, S. et al. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst. Biol.* **59**, 307–321 (2010).
49. Strimmer, K. & von Haeseler, A. Likelihood-mapping: a simple method to visualize phylogenetic content of a sequence alignment. *Proc. Natl Acad. Sci. USA* **94**, 6815 (1997).
50. Smith, S. A. & O'Meara, B. C. TreePL: divergence time estimation using penalized likelihood for large phylogenies. *Bioinformatics* **28**, 2689–2690 (2012).
51. Parham, J. F. et al. Best practices for justifying fossil calibrations. *Syst. Biol.* **61**, 346–359 (2012).
52. de Jong, R. Fossil butterflies, calibration points and the molecular clock (Lepidoptera: Papilionoidea). *Zootaxa* **4270**, 1–63 (2017).
53. Graur, D. & Martin, W. Reading the entrails of chickens: molecular timescales of evolution and the illusion of precision. *Trends Genet.* **20**, 80–86 (2004).
54. Foster, C. S. et al. Evaluating the impact of genomic data and priors on Bayesian estimates of the angiosperm evolutionary timescale. *Syst. Biol.* **66**, 338–351 (2017).
55. Shields, O. Fossil butterflies and the evolution of Lepidoptera. *J. Res. Lepid.* **15**, 132–143 (1976).
56. Wahlberg, N. That awkward age for butterflies: insights from the age of the butterfly subfamily Nymphalinae (Lepidoptera: Nymphalidae). *Syst. Biol.* **55**, 703–714 (2006).
57. Rabosky, D. L. Automatic detection of key innovations, rate shifts, and diversity-dependence on phylogenetic trees. *PLoS ONE* **9**, e89543–e89543 (2014).
58. 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).
59. Maddison, W. P. & Maddison, D. R. Mesquite: a modular system for evolutionary analysis. *Mesquite* <https://www.mesquiteproject.org> (2018).
60. Beaulieu, J. M. & O'Meara, B. C. Detecting hidden diversification shifts in models of trait-dependent speciation and extinction. *Syst. Biol.* **65**, 583–601 (2016).
61. Beaulieu, J., O'Meara, B., Caetano, D., Boyko, J. & Vasconcelos, T. Package 'hisse'. *CRAN* <https://CRAN.R-project.org/package=hisse> (2021).

62. Rabosky, D. L. & Goldberg, E. E. Model inadequacy and mistaken inferences of trait-dependent speciation. *Syst. Biol.* **64**, 340–355 (2015).
63. van Nieuwerkerken, E. J. et al. Order Lepidoptera Linnaeus, 1758. In: Zhang, Z.-Q. (Ed.) *Animal biodiversity: an outline of higher-level classification and survey of taxonomic richness*. *Zootaxa* **3148**, 212 (2011).
64. Cong, Q., Zhang, J., Shen, J. & Grishin, N. V. Fifty new genera of HesperIIDae (Lepidoptera). *Insecta Mundi* **2019**, 0731 (2019).
65. Shirey, V. et al. LepTraits 1.0 A globally comprehensive dataset of butterfly traits. *Sci. Data.* **9**, 382 (2022).
66. Pinkert, S., Barve, V., Guralnick, R. P. & Jetz, W. Global geographical and latitudinal variation in butterfly species richness captured through a comprehensive country-level occurrence database. *Glob. Ecol. Biogeogr.* **31**, 830–839 (2022).
67. Savela, M. Lepidoptera and some other life forms. *FUNET* <https://www.funet.fi/pub/sci/bio/life/intro.html> (2021).
68. 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).
69. 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).
70. Beeravolu, C. R. & Condamine, F. L. An extended maximum likelihood inference of geographic range evolution by dispersal, local extinction and cladogenesis. Preprint at *bioRxiv* <https://doi.org/10.1101/038695> (2016).
71. Matzke, N. J. BioGeoBEARS: biogeography with Bayesian (and likelihood) evolutionary analysis in R scripts (GitHub, 2018).
72. Allio, R. et al. Genome-wide macroevolutionary signatures of key innovations in butterflies colonizing new host plants. *Nat. Commun.* **12**, 354 (2021).
73. Udvardy, M. D. F. A classification of the biogeographical provinces of the world. Morges (Switzerland): International Union of Conservation of Nature and Natural Resources. IUCN Occasional Paper no. 18 (IUCN, 1975).
74. Ree, R. H. Detecting the historical signature of key innovations using stochastic models of character evolution and cladogenesis. *Evolution* **59**, 257–265 (2005).
75. Ronquist, F. Dispersal-vicariance analysis: a new approach to the quantification of historical biogeography. *Syst. Biol.* **46**, 195–203 (1997).
76. Li, X. et al. A diversification relay race from Caribbean-Mesoamerica to the Andes: historical biogeography of *Xylophanes* hawkmoths. *Proc. R. Soc. B Biol. Sci.* **289**, 20212435 (2022).
77. Saito, M. U., Jinbo, U., Yago, M., Kurashima, O. & Ito, M. Larval host records of butterflies in Japan. *Ecol. Res.* **31**, 491–491 (2016).
78. Veenakumari, K., Mohanraj, P. & Sreekumar, P. Host plant utilization by butterfly larvae in the Andaman and Nicobar Islands (Indian Ocean). *J. Insect Conserv.* **1**, 235–246 (1997).
79. Kunte, K. Additions to known larval host plants of Indian butterflies. *J. Bombay Nat. Hist. Soc.* **103**, 119–122 (2006).
80. Kalesh, S. & Prakash, S. K. Additions to larval host plants of butterflies of the Western Ghats, Kerala, Southern India (Rhopalocera, Lepidoptera). Part 1. *J. Bombay Nat. Hist. Soc.* **104**, 235–238 (2007).
81. Kalesh, S. & Prakash, S. K. Additions to larval host plants of butterflies of the Western Ghats, Kerala, Southern India (Rhopalocera, Lepidoptera). Part 2. *J. Bombay Nat. Hist. Soc.* **112**, 111–113 (2015).
82. Naik, D. & Mustak, M. S. Additions to larval host plants of Indian butterflies (Lepidoptera). *J. Bombay Nat. Hist. Soc.* **112**, 181–183 (2015).
83. Karmakar, T. et al. Early stages and larval host plants of some northeastern Indian butterflies. *J. Threat. Taxa* **10**, 11780–11799 (2018).
84. Nitin, R. et al. Larval host plants of the butterflies of the Western Ghats, India. *J. Threat. Taxa* **10**, 11495–11550 (2018).
85. Edger, P. P. et al. The butterfly plant arms-race escalated by gene and genome duplications. *Proc. Natl Acad. Sci. USA* **112**, 8362 (2015).
86. Kembel, S. W. et al. Picante: R tools for integrating phylogenies and ecology. *Bioinformatics* **26**, 1463–1464 (2010).
87. Beckett, S. J. Improved community detection in weighted bipartite networks. *R. Soc. Open Sci.* **3**, 140536 (2016).
88. Dormann, C. F., Gruber, B. & Fründ, J. Introducing the bipartite package: analysing ecological networks. *R News* **8**, 8–11 (2008).
89. R Core Team. R: a language and environment for statistical computing (R Foundation for Statistical Computing, 2009).
90. Revell, L. J. phytools: an R package for phylogenetic comparative biology (and other things). *Methods Ecol. Evol.* **3**, 217–223 (2012).
91. Wahlberg, N. et al. Synergistic effects of combining morphological and molecular data in resolving the phylogeny of butterflies and skippers. *Proc. R. Soc. B Biol. Sci.* **272**, 1577–1586 (2005).

Acknowledgements

We thank M. Kuhn and E. Mavrodiev for assembling trait data. S. Epstein, T. Girard-Ang, P. Pezzi and L. Xiao assisted with laboratory work. J. Barber, M. Brownlee, N. Chazot, S. Cinel, J. Daniels, R. Godfrey, A. Gomez, H. Gough, C. Hamilton, G. Hill, P. Houlihan, C. Huang, J. Miller, K. Miner, C. Mitter, K. Mitter, A. Renevier-Faure, J. Rubin, M. Scallion, R. Singh, Y. Sondhi and L. Wu provided specimens or helped to improve the manuscript. C. Wheat and N. Wahlberg kindly provided input when the project began. RAPiD Genomics (Gainesville, FL, USA) conducted sequencing. K. Meusemann (1KITE) provided four-cluster likelihood mapping scripts. T. Barve, K. Casarella, A. Clark, C. Couch, H. Dansby, R. Merritt, L. Nguyen and X. Zheng helped create figures. Maps in Fig. 3 and in all supplementary figures were created with mapchart.net; these maps are licensed under a Creative Commons Attribution-ShareAlike 4.0 International License (<https://creativecommons.org/licenses/by-sa/4.0/>). High-performance clusters at Brigham Young University, Harvard University, Smithsonian Institution, University of Florida and Zoological Research Museum Alexander Koenig provided computational support. Funding came from the US National Science Foundation (NSF) GoLife ‘ButterflyNet’ collaborative grant (DEB-1541500, 1541557, 1541560) to A.Y.K., R.P.G., D.J.L. and N.E.P. Specimen collection and preservation was funded by NSF DBI-1349345, 1601369, DEB-1557007 and IOS-1920895 (A.Y.K.), NSF DEB-1120380 (D.J.L.), grants 9285-13 and WW-227R-17 from the National Geographic Society (D.J.L.), NSF DBI-1256742 (A.Y.K. and K.R.W.), NSF DEB-0639861 (K.R.W.) and NSF SES-0750480, DEB-0447244 and DEB-9615760 (N.E.P.). M.E. was supported by the Research Council of Norway (no. 204308) and the Hintelmann Scientific Award for Zoological Systematics. F.L.C. was supported by the European Research Council under the European Union’s Horizon 2020 research and innovation programme (project GAIA, no. 851188). M.P.B. was supported by the Swedish Research Council (IPG no. 2020-06422). R.V. was supported by the Spanish Ministry of Science and Innovation grant PID2019-107078GB-I00/AEI/10.13039/501100011033. G.T. was supported by the Spanish Ministry of Science and Innovation (grants PID2020-117739GA-I00/AEI/10.13039/501100011033 and RYC2018-025335-I). V.L. was supported by the Russian Science Foundation (grant 19-14-00202) and by the Ministry of Science and Higher Education of the Russian Federation (grant 075-15-2021-1069). M.Y. was supported by MEXT KAKENHI no. 19916010 and JSPS KAKENHI grants 13010131, 23570111, 26440207, 17K07528 and 21H02215. A.B.B.M., H.P.R. and N.O.M. were supported by CNPQ grants proc 563332/2010-7 and 304273/2014-7. We are thankful for the continuous support from the Museum of

Comparative Zoology for fieldwork (Putnam Expedition Grants), and providing funds to support Open Access (Wetmore-Colles Fund).

Author contributions

A.P.S.C., A.Y.K., C.E., C.S., D.M.P., E.A.E., E.F.A.T., F.L.C., H.L.O., J.W.B., M.E., M.P.B., P.B.F., R.A.S. and X.L. contributed analysis. A.Y.K., D.J.L., L.R., J.W.B., M.E., N.E.P. and R.P.G. conceptualized the study. A.Y.K., D.J.L., M.A.J., N.E.P., R.P.G. and W.J. acquired funding. A.Y.K., C.S., D.J.L., D.M.P., E.A.E., E.A.L., F.L.C., H.W., J.P.W.H., J.W.B., L.R., M.A.K.N., M.E., M.G.A.I., N.E.P., R.P.G., V.B., V.S. and W.A.V.-M. performed data assembly. A.Y.K., D.J.L. and N.E.P. were responsible for project administration. A.B.B.M., A.P.S.C., A.Y.K., D.J.L., E.F.A.T., G.T., H.P.R., J.I.M., J.W.B., M.E., M.Y., N.E.P., R.A.S., R.V., P.V., Y.-F.H. and Z.F.F. performed sampling. C.S., J.W.B. and M.E. were responsible for sequence workflow. A.D.W., A.Markee, A.P.S.C., A.Y.K., D.J.L., D.P., H.P.R., J.I.M., J.V.L., K.M.D., M.E., M.Y., N.O.M., R.A.S., R.M., S.C.M., T.D., Y.-L.N. and Z.F.F. identified and prepared specimens. A.Y.K., D.J.L., J.W.B., M.E., N.E.P. and R.P.G. provided supervision. D.J.L., G.L., R.P.G. and V.B. were responsible for taxonomy and curation. A.Markee, A.P.S.C., A.Y.K., D.J.L., F.L.C. and X.L. performed visualizations. A.P.S.C., A.Y.K., D.J.L. and D.M.P. wrote the original draft of the article. A.B.B.M., A.D.W., A.H., A.Markee, A.Monastyrskii, A.P.S.C., A.S., A.Y.K., C.A.M., C.E., C.J.M., C.S., D.H.J., D.J.L., D.J.M., D.M.P., D.P., D.S., E.A.E., E.A.L., E.D., E.F.A.T., F.L.C., G.L., G.N., G.T., H.A.B., H.L.O., H.P.R., H.W., I.J.K., J.A.T.B., J.D.H., J.I.M., J.P.W.H., J.V.L., J.W.B., K.A.-P., K.K., K.M., K.M.D., K.R.W., L.C., L.R., M.A.J., M.A.K.N., M.E., M.F.B., M.G.A.I., M.J.L., M.P.B., M.S., M.Y., N.B.H., N.E.P., N.O.M., N.T.H., P.A.C.B., P.B.F., P.S.P., P.V., R.A.R.C., R.A.S., R.E., R.J.G., R.L.H.S., R.M., R.P.G., R.V., S.C.C., S.C.M., S.Sáfián, S.Schröder, T.B.L., T.D., V.B., V.L., V.S., W.-P.C., W.A.V.-M., W.H., W.J., X.L., Y.-F.H., Y.-L.N., Y.I. and Z.F.F. reviewed and edited the article.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41559-023-02041-9>.

Correspondence and requests for materials should be addressed to Akito Y. Kawahara, Naomi E. Pierce or David J. Lohman.

Peer review information *Nature Ecology & Evolution* thanks Nicolas Chazot and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at 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) 2023

Akito Y. Kawahara^{1,2,3}✉, Caroline Storer¹, Ana Paula S. Carvalho¹, David M. Plotkin^{1,2}, Fabien L. Condamine⁴, Mariana P. Braga^{5,6}, Emily A. Ellis¹, Ryan A. St Laurent^{1,7}, Xuankun Li^{1,8}, Vijay Barve¹, Liming Cai^{9,10}, Chandra Earl^{1,11}, Paul B. Frandsen¹², Hannah L. Owens^{11,13,14}, Wendy A. Valencia-Montoya⁹, Kwaku Aduse-Poku^{15,16}, Emmanuel F. A. Toussaint^{1,17}, Kelly M. Dexter¹, Tenzing Doleck^{15,18}, Amanda Markee¹, Rebecca Messcher¹, Y-Lan Nguyen¹⁵, Jade Aster T. Badon¹⁹, Hugo A. Benítez^{20,21}, Michael F. Braby^{22,23}, Perry A. C. Buenavente^{10,24}, Wei-Ping Chan⁹, Steve C. Collins²⁵, Richard A. Rabideau Childers⁹, Even Dankowicz⁹, Rod Eastwood⁹, Zdenek F. Fric²⁶, Riley J. Gott^{1,2}, Jason P. W. Hall⁷, Winnie Hallwachs²⁷, Nate B. Hardy²⁸, Rachel L. Hawkins Sipe⁹, Alan Heath^{9,29}, Jomar D. Hinolan³⁰, Nicholas T. Homziak^{1,2}, Yu-Feng Hsu³¹, Yutaka Inayoshi³², Micael G. A. Itliong¹⁵, Daniel H. Janzen²⁷, Ian J. Kitching³³, Krushnamegh Kunte³⁴, Gerardo Lamas³⁵, Michael J. Landis⁶, Elise A. Larsen³⁶, Torben B. Larsen⁶⁶, Jing V. Leong^{15,26,37}, Vladimir Lukhtanov³⁸, Crystal A. Maier⁹, Jose I. Martinez^{1,2}, Dino J. Martins³⁹, Kiyoshi Maruyama⁴⁰, Sarah C. Maunsell⁹, Nicolás Oliveira Mega⁴¹, Alexander Monastyrskii^{42,43}, Ana B. B. Morais⁴⁴, Chris J. Müller⁴⁵, Mark Arcebal K. Naive^{46,47,48}, Gregory Nielsen⁴⁹, Pablo Sebastián Padrón^{1,50}, Djunijanti Peggie⁵¹, Helena Piccoli Romanowski⁴¹, Szabolcs Sáfián⁵², Motoki Saito⁵³, Stefan Schröder⁵⁴, Vaughn Shirey³⁶, Doug Soltis¹¹, Pamela Soltis¹¹, Andrei Sourakov¹, Gerard Talavera^{9,55}, Roger Vila⁵⁶, Petr Vlasanek⁵⁷, Houshuai Wang⁵⁸, Andrew D. Warren¹, Keith R. Willmott¹, Masaya Yago⁵⁹, Walter Jetz^{60,61}, Marta A. Jarzyna^{60,62,63}, Jesse W. Breinholt^{11,64}, Marianne Espeland⁶⁵, Leslie Ries³⁶, Robert P. Guralnick¹¹, Naomi E. Pierce⁹✉ & David J. Lohman^{15,18,24}✉

¹McGuire Center for Lepidoptera and Biodiversity, Florida Museum of Natural History, University of Florida, Gainesville, FL, USA. ²Entomology and Nematology Department, University of Florida, Gainesville, FL, USA. ³Department of Biology, University of Florida, Gainesville, FL, USA. ⁴CNRS, Institut des Sciences de l'Evolution de Montpellier (Université de Montpellier), Montpellier, France. ⁵Department of Ecology, Swedish University of Agricultural Sciences, Uppsala, Sweden. ⁶Department of Biology, Washington University in St. Louis, St. Louis, MO, USA. ⁷Department of Entomology, National Museum of Natural History, Smithsonian Institution, Washington, DC, USA. ⁸Center for Biodiversity Research, Department of Biological Sciences, University of Memphis, Memphis, TN, USA. ⁹Department of Organismic and Evolutionary Biology and Museum of Comparative Zoology, Harvard University, Cambridge, MA, USA. ¹⁰Department of Botany and Plant Sciences, University of California, Riverside, Riverside, CA, USA. ¹¹Florida Museum

of Natural History, University of Florida, Gainesville, FL, USA. ¹²Department of Plant and Wildlife Sciences, Brigham Young University, Provo, UT, USA. ¹³Center for Global Mountain Biodiversity, Globe Institute, University of Copenhagen, Copenhagen, Denmark. ¹⁴Center for Macroecology, Evolution, and Climate, GLOBE Institute, University of Copenhagen, Copenhagen, Denmark. ¹⁵Biology Department, City College of New York, City University of New York, New York, NY, USA. ¹⁶Department of Life and Earth Sciences, Perimeter College, Georgia State University, Decatur, GA, USA. ¹⁷Department of Entomology, Natural History Museum of Geneva, Geneva, Switzerland. ¹⁸PhD Program in Biology, Graduate Center, City University of New York, New York, NY, USA. ¹⁹Animal Biology Division, Institute of Biological Sciences, University of the Philippines Los Baños, Laguna, Philippines. ²⁰Laboratorio de Ecología y Morfometría Evolutiva, Centro de Investigación de Estudios Avanzados del Maule, Universidad Católica del Maule, Talca, Chile. ²¹Millennium Institute Biodiversity of Antarctic and Subantarctic Ecosystems (BASE), Santiago, Chile. ²²Division of Ecology and Evolution, Research School of Biology, The Australian National University, Acton, Canberra, Australian Capital Territory, Australia. ²³Australian National Insect Collection, Canberra, Australian Capital Territory, Australia. ²⁴Entomology Section, National Museum of Natural History, Manila, Philippines. ²⁵African Butterfly Research Institute, Nairobi, Kenya. ²⁶Biology Centre CAS, České Budějovice, Czech Republic. ²⁷Department of Biology, University of Pennsylvania, Philadelphia, PA, USA. ²⁸Department of Entomology and Plant Pathology, Auburn University, Auburn, AL, USA. ²⁹Iziko South African Museum, Cape Town, South Africa. ³⁰Botany and National Herbarium Division, National Museum of the Philippines, Manila, Philippines. ³¹College of Life Science, National Taiwan Normal University, Taipei, Taiwan. ³²Chiang Mai, Thailand. ³³Natural History Museum, London, UK. ³⁴National Centre for Biological Sciences, Tata Institute of Fundamental Research, Bengaluru, India. ³⁵Museo de Historia Natural, Universidad Nacional Mayor de San Marcos, Lima, Peru. ³⁶Department of Biology, Georgetown University, Washington, DC, USA. ³⁷Faculty of Science, Department of Zoology, University of South Bohemia, České Budějovice, Czech Republic. ³⁸Department of Karyosystematics, Zoological Institute of Russian Academy of Sciences, St. Petersburg, Russia. ³⁹Turkana Basin Institute, Stony Brook University, Stony Brook, NY, USA. ⁴⁰Hachiōji, Japan. ⁴¹Departamento de Zoologia, Universidade Federal do Rio Grande do Sul, Porto Alegre, Brazil. ⁴²Vietnam Programme, Fauna & Flora International, Hanoi, Vietnam. ⁴³Vietnam National Museum of Nature, Vietnam Academy of Science and Technology, Hanoi, Vietnam. ⁴⁴Centro de Ciências Naturais e Exatas, Pós-Graduação em Biodiversidade Animal, Universidade Federal de Santa Maria, Santa Maria, Brazil. ⁴⁵Australian Museum, Sydney, New South Wales, Australia. ⁴⁶Center for Integrative Conservation, Xishuangbanna Tropical Botanical Garden, Chinese Academy of Sciences, Mengla, China. ⁴⁷University of Chinese Academy of Sciences, Beijing, China. ⁴⁸College of Arts and Sciences, Jose Rizal Memorial State University, Tampilisan, Philippines. ⁴⁹Aquapro, Villavicencio, Colombia. ⁵⁰Entomology Laboratory, Museo de Zoología, Universidad del Azuay, Cuenca, Ecuador. ⁵¹Research Center for Biosystematics and Evolution, National Research and Innovation Agency (BRIN), Cibinong-Bogor, Indonesia. ⁵²Institute of Silviculture and Forest Protection, University of West Hungary, Sopron, Hungary. ⁵³The Research Institute of Evolutionary Biology (Insect Study Division), Setagaya, Japan. ⁵⁴Köln, Germany. ⁵⁵Institut Botànic de Barcelona (IBB, CSIC-Ajuntament de Barcelona), Barcelona, Spain. ⁵⁶Institut de Biologia Evolutiva (CSIC-Univ. Pompeu Fabra), Barcelona, Spain. ⁵⁷T.G. Masaryk Water Research Institute, Prague, Czech Republic. ⁵⁸Department of Entomology, College of Plant Protection, South China Agricultural University, Guangzhou, China. ⁵⁹The University Museum, The University of Tokyo, Tokyo, Japan. ⁶⁰Department of Ecology & Evolutionary Biology, Yale University, New Haven, CT, USA. ⁶¹Center for Biodiversity and Global Change, Yale University, New Haven, CT, USA. ⁶²Translational Data Analytics Institute, The Ohio State University, Columbus, OH, USA. ⁶³Department of Evolution, Ecology and Organismal Biology, The Ohio State University, Columbus, OH, USA. ⁶⁴RAPID Genomics, Gainesville, FL, USA. ⁶⁵Leibniz Institute for the Analysis of Biodiversity Change, Zoological Research Museum Alexander Koenig, Bonn, Germany. ⁶⁶Deceased: Torben B. Larsen. ✉ e-mail: kawahara@flmnh.ufl.edu; npierce@oeb.harvard.edu; dlohman@ccny.cuny.edu

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our [Editorial Policies](#) and the [Editorial Policy Checklist](#).

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

- | n/a | Confirmed |
|-------------------------------------|---|
| <input checked="" type="checkbox"/> | <input type="checkbox"/> The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> The statistical test(s) used AND whether they are one- or two-sided
<i>Only common tests should be described solely by name; describe more complex techniques in the Methods section.</i> |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> A description of all covariates tested |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals) |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> For null hypothesis testing, the test statistic (e.g. F , t , r) with confidence intervals, effect sizes, degrees of freedom and P value noted
<i>Give P values as exact values whenever suitable.</i> |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Estimates of effect sizes (e.g. Cohen's d , Pearson's r), indicating how they were calculated |

Our web collection on [statistics for biologists](#) contains articles on many of the points above.

Software and code

Policy information about [availability of computer code](#)

Data collection	No software was used
Data analysis	<p>The following programs and code were used for data analysis: R, TrimGalore!, MAFFT, FASconCAT, Iq-TREE, published scripts from Breinholt et al. (2018), SPAdes, BLAST+, FastTree, PhyloTreePruner, Trimal, MACSE, PartitionFinder2, FastTree, catfasta2phymI, MonoPhy, Degen, ASTRAL, TreePL, BMM, BMMtools, RevBayes, picante, bipartite, Phytools, hisse, BioGeoBEARS, DECC.</p> <p>A new R script for running analyses in picante was created by co-authors of this manuscript and is publicly available on GitHub (https://github.com/lmcai/BNet-PD-analysis)</p> <p>A new R script for partially automating TreePL analyses was created by co-authors of this manuscript and is publicly available on GitHub (https://github.com/sunray1/treepl)</p> <p>References for all existing programs, and github links for all novel code, have been provided in the main text and/or the Extended Online Methods.</p>

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio [guidelines for submitting code & software](#) for further information.

Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our [policy](#)

All supplementary figures, tables, and data archives are available on Figshare (<https://doi.org/10.6084/m9.figshare.21774899>). Genomic data for all newly sequenced specimens in this study have been uploaded to GenBank as part of BioProject PRJNA714105. Individual BioSample accession numbers for each specimen are provided in Table S1.

Human research participants

Policy information about [studies involving human research participants and Sex and Gender in Research](#).

Reporting on sex and gender

this information has not been collected

Population characteristics

see above

Recruitment

see above

Ethics oversight

see above

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

Life sciences Behavioural & social sciences Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see [nature.com/documents/nr-reporting-summary-flat.pdf](https://www.nature.com/documents/nr-reporting-summary-flat.pdf)

Ecological, evolutionary & environmental sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description

We sequenced 391 genes from nearly 2,300 butterfly species to reconstruct a new, phylogenomic tree of butterflies

Research sample

A total of 2,248 butterfly specimens representing 2,244 species in 1,644 genera was included for the molecular component of this study, along with 10 outgroups from other lepidopteran superfamilies. We aimed to include at least one species from every valid genus and sequenced the type species of each genus whenever possible.

Sampling strategy

Specimens were collected in 90 countries over a 70-year period by over 300 people, preserved and deposited in one of the 28 specimen collections from which we obtained new tissue samples.
We also obtained sequence data from 343 published genomes and transcriptomes.

Data collection

We extracted DNA from 1,915 specimens that were 1) stored in ethanol and frozen; 2) dried and stored in glassine envelopes under ambient conditions (papered); or 3) dried, spread and pinned in a museum collection. For ethanol-preserved and enveloped specimens, we prioritized tissue from the thorax, but if thoracic tissue was unavailable, we used legs or tissue from the proximal region of the abdomen. We removed abdomens and/or up to three legs from museum specimens. In all cases, when multiple specimens of the same species were available, the most-recently collected specimen was used for DNA extraction.
All tissues were extracted with the OmniPrep DNA Extraction Kit for High Quality Genomic DNA Extraction
All DNA extractions were stored in 1.5 mL microcentrifuge tubes at -20° C prior to sequencing. Most extractions were performed at the McGuire Center for Lepidoptera and Biodiversity, Florida Museum of Natural History, University of Florida; the Museum of Comparative Zoology, Harvard University; or the Biology Department, City College of New York, City University of New York. Chinese samples that could not be exported to the US were sequenced at Annoroad Gene Technology Co., Ltd, Beijing, China.

Timing and spatial scale

DNA extractions took place between 2015-2019

Data exclusions

Multiple tests were conducted to determine whether sequences were contaminated. All samples that were included in the final dataset were checked first by BLAST, comparing its COI locus via the command-line blastn in Genbank. A minimum e-value of 0.0001 and maximum of five target sequences were specified to filter the search. The length of the sequence matching the query, its bit score, and taxonomy of resulting matched sequences were used to determine contamination. Samples with a low bit score were

flagged, and if the top three hits for that sequence was a species in a different butterfly family, the voucher specimen was examined for confirmation. Identification was confirmed by sharing specimen images with taxonomists, and in some cases, genitalia were dissected for verification.

We took three additional steps to identify contaminated or mislabeled sequences. (A) We generated gene trees in FastTree v2.1.744 for every locus. Along with visual inspection, branch tip lengths were compared in all gene trees using LongBranchFinder.pl50. If a particular gene tree branch was found to be significantly long (i.e., ≥ 8 SD from the mean branch length) the corresponding sequence for that taxon was removed from that locus (Table S31). (B) The R package MonoPhy v1.351 was used to perform a thorough search for any non-monophyletic groups present in an initial unpartitioned ML tree generated in IQ-TREE52, using our subfamily, tribe, and genus-level butterfly taxonomy. Questionable placements of taxa in the tree were verified by expert taxonomists and removed if there was a significant concern for contamination. (C) We calculated Quartet Fidelity (QF) scores⁵³ and specimens that were deemed outliers were removed.

Reproducibility

Input files used in our analyses (as described in sections 9-20 of Extended Online Methods), have been provided in supplementary Data Archives and uploaded to Figshare, for the purposes of reproducibility.

Randomization

Random allocation of samples into groups is not necessary prior to a phylogenetic analysis. A random starting tree is automatically generated as part of the tree-searching algorithm in the program used for our primary phylogenetic analysis, IQ-tree.

Blinding

Blinding is not relevant to our study. The program used for our primary phylogenetic analysis, IQ-tree, is not provided any information on previous phylogenetic hypotheses, it interprets the dataset strictly based on the DNA sequences and partition information in the input files. No prior taxonomic knowledge was used when performing this analysis.

Did the study involve field work? Yes No

Field work, collection and transport

Field conditions

The majority of specimens that were used for the genetic portion of this study were obtained from legs and abdomens of museum specimens. These were collected before enforcement of the Nagoya Protocol on Access to Genetic Resources and the Fair and Equitable Sharing of Benefits Arising from their Utilization to the Convention on Biological Diversity (<https://www.cbd.int/abs/>). However, some specimens were obtained more recently through collecting efforts by our team and collaborators. Some countries, such as the USA, Chile, and many European countries, do not require permits for collecting non-protected species outside of protected areas, and some of our specimens were obtained in this manner. Other specimens were obtained in areas where permits were required. We list permits and their relevant countries and regions in Table S29; a pdf of permits can be found in Data S10.

Location

Locations where each specimen in our study was collected are provided in Table S1. As our dataset was comprised of over 2,000 specimens, there is not enough space to exhaustively list all locations here.

Access & import/export

We list permits and their relevant countries and regions in Table S29; a pdf of permits can be found in Data S10.

Disturbance

Butterfly specimens were collected individually with nets or by hand; this minimizes disturbance and prevents the death of other surrounding organisms that may otherwise happen with mass insect trapping techniques

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

- | n/a | Involved in the study |
|-------------------------------------|---|
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Antibodies |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Eukaryotic cell lines |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Palaeontology and archaeology |
| <input type="checkbox"/> | <input checked="" type="checkbox"/> Animals and other organisms |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Clinical data |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Dual use research of concern |

Methods

- | n/a | Involved in the study |
|-------------------------------------|---|
| <input checked="" type="checkbox"/> | <input type="checkbox"/> ChIP-seq |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Flow cytometry |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> MRI-based neuroimaging |

Animals and other research organisms

Policy information about [studies involving animals](#); [ARRIVE guidelines](#) recommended for reporting animal research, and [Sex and Gender in Research](#)

Laboratory animals	the study did not involve laboratory animals
Wild animals	All wild insects collected in the field were killed via a painless chemical killing agent (e.g., potassium cyanide) or via flash-freezing if such equipment was available at the field site. All insects were deceased prior to shipping.
Reporting on sex	this information has can be inferred via examination of the voucher specimens, but is not relevant to the study and not reported here. Sex is not relevant to a genus-level phylogenetic analysis.
Field-collected samples	all field-collected samples were deceased prior to being brought into a laboratory setting. No special parameters were necessary, beyond the basic sterilization and hygiene protocols used to minimize cross-contamination and ensure safety of people working in the lab.
Ethics oversight	no ethical approval was required for working with non-endangered invertebrate specimens.

Note that full information on the approval of the study protocol must also be provided in the manuscript.