Report

Current Biology

A Comprehensive and Dated Phylogenomic Analysis of Butterflies

Highlights

- Phylogenomic data provide a novel view of broad butterfly evolutionary relationships
- Most current diversity originated after the K-Pg mass extinction
- Many accepted higher taxa are para- or polyphyletic
- Ant association originated three times independently in blues and metalmarks

Authors

Marianne Espeland, Jesse Breinholt, Keith R. Willmott, ..., David J. Lohman, Naomi E. Pierce, Akito Y. Kawahara

Correspondence

m.espeland@leibniz-zfmk.de (M.E.), kawahara@flmnh.ufl.edu (A.Y.K.)

In Brief

Espeland et al. inferred a dated, molecular phylogeny of butterflies based on the largest and most taxonomically extensive phylogenomic dataset ever assembled for the group: 352 loci from 207 species representing all butterfly families and subfamilies and 98% of the tribes. This tree will provide a robust framework for future comparative analyses.

Espeland et al., 2018, Current Biology *28*, 770–778
March 5, 2018
https://doi.org/10.1016/j.cub.2018.01.061

CellPress

CellPress

Current Biology
Report

A Comprehensive and Dated Phylogenomic Analysis of Butterflies

Marianne Espeland,[1,2,3,11,*] Jesse Breinholt,[2,4] Keith R. Willmott,[2] Andrew D. Warren,[2] Roger Vila,[5] Emmanuel F.A. Toussaint,[2] Sarah C. Maunsell,[3] Kwaku Aduse-Poku,[6] Gerard Talavera,[3,5] Rod Eastwood,[7] Marta A. Jarzyna,[8] Robert Guralnick,[2] David J. Lohman,[6,9,10] Naomi E. Pierce,[3] and Akito Y. Kawahara[2,*]

[1]Arthropoda Department, Zoological Research Museum Alexander Koenig, Adenauer Allee 160, 53113 Bonn, Germany
[2]Florida Museum of Natural History, University of Florida, Gainesville, FL 32611, USA
[3]Museum of Comparative Zoology, Harvard University, Cambridge, MA 02138, USA
[4]RAPiD Genomics 747 SW 2nd Avenue IMB#14, Gainesville, FL 32601, USA
[5]Institut de Biologia Evolutiva (CSIC- Universitat Pompeu Fabra), Passeig Marítim de la Barceloneta 37, ESP-08003 Barcelona, Spain
[6]Biology Department, City College of New York, New York, NY 10031, USA
[7]Institute for Agricultural Sciences, Biocommunication & Entomology, Eidgenössische Technische Hochschule Zürich (ETHZ), WEV E26.1, Weinbergstrasse 56-58, 8092 Zürich, Switzerland
[8]Department of Ecology and Evolutionary Biology, Yale University, New Haven, CT 06511-8934, USA
[9]PhD Program in Biology, Graduate Center, City University of New York, New York, NY 10016, USA
[10]Entomology Section, National Museum of the Philippines, Manila 1000, Philippines
[11]Lead Contact
*Correspondence: m.espeland@leibniz-zfmk.de (M.E.), kawahara@flmnh.ufl.edu (A.Y.K.)
https://doi.org/10.1016/j.cub.2018.01.061

SUMMARY

Butterflies (Papilionoidea), with over 18,000 described species [1], have captivated naturalists and scientists for centuries. They play a central role in the study of speciation, community ecology, biogeography, climate change, and plant-insect interactions and include many model organisms and pest species [2, 3]. However, a robust higher-level phylogenetic framework is lacking. To fill this gap, we inferred a dated phylogeny by analyzing the first phylogenomic dataset, including 352 loci (> 150,000 bp) from 207 species representing 98% of tribes, a 35-fold increase in gene sampling and 3-fold increase in taxon sampling over previous studies [4]. Most data were generated with a new anchored hybrid enrichment (AHE) [5] gene kit (BUTTERFLY1.0) that includes both new and frequently used (e.g., [6]) informative loci, enabling direct comparison and future dataset merging with previous studies. Butterflies originated around 119 million years ago (mya) in the late Cretaceous, but most extant lineages diverged after the Cretaceous-Paleogene (K-Pg) mass-extinction 65 mya. Our analyses support swallowtails (Papilionidae) as sister to all other butterflies, followed by skippers (Hesperiidae) + the nocturnal butterflies (Hedylidae) as sister to the remainder, indicating a secondary reversal from diurnality to nocturnality. The whites (Pieridae) were strongly supported as sister to brush-footed butterflies (Nymphalidae) and blues + metalmarks (Lycaenidae and Riodinidae). Ant association independently evolved once in Lycaenidae and twice in Riodinidae. This study overturns prior notions of the taxon's evolutionary history, as many long-recognized subfamilies and tribes are para- or polyphyletic. It also provides a much-needed backbone for a revised classification of butterflies and for future comparative studies including genome evolution and ecology.

RESULTS AND DISCUSSION

Probe Set and Additional Data

We sequenced 425 loci from two outgroups and 173 species of butterflies representing all subfamilies and approximately 98% of all tribes (missing Hypotheclini, Heliophorini, Sertaniini and Zetherini) (Data S1A) using our anchored hybrid enrichment (AHE) butterfly kit (BUTTERFLY1.0), of which 352 captured more than 75% of the taxa. These 352 loci were included in downstream phylogenetic analyses (151, 377 bp). On average, 86% of taxa were captured per locus. Locus characteristics can be found in Data S1B. Even though our kit was designed specifically for butterflies, most loci also captured well for outgroups (*Stegasta bosqueella*, Gelechiidae = 79%; *Striglina strigifera*, Thyrididae = 81%), indicating broader applicability. We supplemented our dataset with additional taxa based on 12 available genomes, 15 transcriptomes, and AHE data for three taxa (Data S1C), as well as newly generated low-coverage genome data for *Jalmenus evagoras* and *Lepidochrysops patricia* (Lycaenidae) and *Parnassius apollo* (Papilionidae). Sequence reads for genetic data generated here as well as alignments used for probe design can be found on Dryad https://doi.org/10.5061/dryad.v1h63 and NCBI: PRJNA432311).

Evolution of the Butterflies

Phylogenies were inferred using maximum likelihood ([ML]; IQ-TREE v.1.4.2 [7]) and coalescent summary methods (ASTRAL-II [8], ASTRID [9]) based on 13 different datasets to

Table 1. Dataset Statistics

Dataset abbreviation	NQS Astral	NQS Astrid	NQS ML	#loci	# sites (bp)	Informative sites (%)	Missing data (%)	# taxa	Average locus length (bp)	Average pairwise identity (%)	Average GC content (%)
ALL*	0.766	0.766	0.765	352	151377	48.3	10.3	207	430	79.86	45.18
GC1	0.82	0.819	0.818	81	41298	49.2	11.2	207	510	79.74	37.4
GC2	0.759	0.757	0.757	254	105072	47.8	9.83	207	414	79.72	46.34
GC3	0.619	0.592	0.611	17	5002	52.5	12.82	207	294	82.64	64.84
LE1	0.696	0.695	0.694	164	37988	47.9	10.78	207	231	80.28	46.88
LE2	0.791	0.791	0.789	88	34262	48.4	9.5	207	389	79.93	44.64
LE3*	0.857	0.856	0.855	100	78888	57.4	10.41	207	791	79.12	42.87
PI1	0.793	0.792	0.792	181	85148	52.2	12.24	207	470	77.52	44.34
PI2	0.742	0.741	0.74	171	66224	43.4	7.8	207	387	82.34	46.06
GCLE1*	0.831	0.832	0.83	182	110762	48.4	10.08	207	609	79.39	42.95
AA	0.657	0.656	0.656	214	92445 (30815 AA)	38.3	12.3	207	146 (AA)	89.66	N/A
LE4M*	0.865	0.864	0.863	100	78888	56.1	6.32	188	791	79.81	42.43
13L	0.848	0.841	0.834	13	10893	48.3	13.29	207	838	80.17	44.58

Normalized quartet score (NQS) calculated in ASTRAL-II for trees inferred using ASTRAL, Astrid, and IQ-TREE (ML), for different combinations of loci, and various dataset statistics. The four datasets mainly discussed in the text are denoted by an asterisk. Dataset abbreviations: ALL, nucleotide alignments of all loci; GC1, nucleotide alignments of loci with average GC content lower than 40%; GC2, nucleotide alignments of loci with average GC content between 40% and 60%; GC3, nucleotide alignments of loci with average GC content higher than 60%; LE1, loci shorter than 300 bp long; LE2, loci between 300 and 499 bp long; LE3, loci longer than 499 bp; PI1, nucleotide alignments of loci with average pairwise identity between 70% and 80%; PI2, nucleotide alignments of loci with average pairwise identity of 80% and higher; GCLE1, nucleotide alignments of loci longer than 200 bp and with average GC content lower than 60; AA, amino acid alignments of loci with average pairwise identity below 95%; LE4M, nucleotide alignments of loci longer than 499 bp with taxa with more than 25% missing data removed (Data S1D); 13L, 13 commonly used loci in butterfly phylogenetics (locus 1–13 in Data S1B); bp, base pair; GC, guanine-cytosine. See also Figure S1.

assess the effects of nucleotide bias, evolutionary rate, phylogenetic signal, missing data, and incomplete lineage sorting (ILS) on tree topology (Table 1). Results from analyses of the various datasets (see Quantification and Statistical Analysis; Figure S1) suggest that ILS is very limited, and we therefore mainly discuss the ML results, unless otherwise noted. Support was calculated using nonparametric bootstrap (bs) [10] and quartet support (QS) [11]. Given the combined taxonomic and molecular sampling, this is the most comprehensive phylogenetic dataset for butterflies thus far. All datasets and trees can be found on Dryad https://doi.org/10.5061/dryad.v1h63.

The phylogenetic relationships of major butterfly lineages have been controversial for decades [4, 12] (and references therein). The most recent family-level phylogenies [13–17] are all based on Sanger sequencing and a limited number of genes (3–10), often suffer from low support, and have short branches and conflicting relationships at higher levels. Specifically, we test previous hypotheses about relationships among the main lineages of butterflies, such as whether the "nocturnal butterflies" (Hedylidae), belong within or outside of the butterflies and whether ant association is ancestral [18, 19] or has evolved multiple times [20] in Lycaenidae+Riodinidae. Many previously recognized taxa are paraphyletic or polyphyletic in our tree including Papilioninae (Papilionidae); Polyommatinae, Theclinae (both Lycaenidae) and many tribes therein; and Coeini, Elymniini, and Limenitidini (Nymphalidae), suggesting the need for higher-level revision.

The ML tree from the full dataset (Table 1) was dated using nine calibration points based on fossils and a secondary calibration of the root based on the age of the angiosperms (see Method Details). Most previous studies included few taxa, few genes, and/or based their calibrations on very few fossils that in several cases were misplaced in the phylogeny (Table 2; [21]). The most comprehensive higher-level butterfly study so far [4], for example, used two fossils for calibration in addition to several secondary calibrations. One of these fossil species, *Oligodonta florissantensis*, which was used as a calibration point within Pieridae (also in other studies, see Table 2) is now considered to be a nymphalid [22]. With the most extensive use of butterfly fossils to date (Table 2), we found that the butterflies originated around 119 mya (92–142 mya) in the lower Cretaceous (Figures 1, tree with credibility intervals on Dryad https://doi.org/10.5061/dryad.v1h63, and S2, undated ML tree with outgroups). This result broadly agrees with previous studies, although the estimated age has varied from 41 [23], a clear underestimate given it is 14 million years (Ma) younger than the oldest butterfly fossil [24], to over 200 Ma (Table 2). Most of the current butterfly diversity originated after the K-Pg boundary, suggesting that this event had a major impact on the evolutionary history of butterflies [4, 16], with many lineages possibly going extinct.

Papilionoidea was strongly supported as monophyletic, with the Papilionidae as the sister group to the remaining butterflies (bs = 100, QS = 0.88–1 for the two datasets with lowest gene tree-species tree discordance, LE3 and LE4M, Table 1), congruent with prior studies [4, 25–28]. The Papilionidae traditionally has been divided into three extant subfamilies: Baroniinae, Parnassiinae, and Papilioninae [29], with the monotypic Baroniinae placed as sister group to the two other subfamilies (e.g., [4, 17, 30]) or Papilioninae as sister to Baroniinae and Parnassiinae [31]. We find that Papilionidae started diversifying around 84 mya (63–109) in the upper Cretaceous, similar to [4],

Table 2. Ages of Butterflies and Butterfly Families Inferred Here and in Previous Studies

Studies	Crown Ages of Butterfly Clades								# Genes	# Butterfly taxa	# Butterfly Fossils
	Papilionoidea	Papilionidae	Hesperiidae	Hedylidae	Pieridae	Nymphalidae	Lycaenidae	Riodinidae			
This study	119 (91–143)	84 (63–109)	79 (60–99)	29 (18–42)	87 (67–108)	91 (71–112)	78 (60–96)	73 (56–92)	352	195	9
Wahlberg et al. (2009) [16][1]	104 (93–116)	63 (52–76)	N/A	N/A	73 (57–86)	94 (84–104)	75 (63–86)	65 (56–76)	10	429	7
Heikkilä et al. (2012) [4][2]	110 (92–128)	75 (62–88)	65 (54–79)	N/A	80 (67–97)	87 (74–101)	73 (58–84)	72 (57–83)	8	71	2
Pohl et al. (2009) [54][3]	113–219	N/A	N/A	N/A	N/A	86–158	N/A	N/A	5	27	0
Wahlberg et al. (2013) [55][4]	104 (95–114)	N/A	N/A	N/A	N/A	N/A	N/A	N/A	8	7	0
Misof et al. (2014) [23][5]	41 (36–50)	N/A	N/A	N/A	N/A	N/A	N/A	N/A	1478	2	1
Rainford et al. (2014) [56][6]	105 (66–126)	N/A	N/A	N/A	N/A	N/A	N/A	N/A	8	7	1
Tong et al. (2015) [57][7]	70 (55–110)	N/A	N/A	N/A	N/A	N/A	N/A	N/A	1478	2	1
Wu et al. (2015) [32][8]	69 (61–78)	57 (52–65)	N/A	N/A	N/A	N/A	N/A	N/A	4	87	3
Cong et al. (2017) [39][9]	133 (102–162)	N/A	90 (33–113)	N/A	N/A	87 (69–104)	N/A	N/A	1624	13	1
Talla et al. (2017) [58][10]	115 (95–128)	N/A	N/A	N/A	N/A	N/A	N/A	N/A	243	16	0
Nazari et al. (2007) [31][11]	103	96 (94–99)	N/A	N/A	N/A	N/A	N/A	N/A	7	43	3
Simonsen et al. (2010) [30]	N/A	68 (53–87)	N/A	N/A	N/A	N/A	N/A	N/A	7	35	2
Condamine et al. (2012) [17][12]	N/A	52 (46–63)	N/A	N/A	N/A	N/A	N/A	N/A	3	221	5
Sahoo et al. (2017) [44][13]	N/A	N/A	94	N/A	N/A	N/A	N/A	N/A	10	341	1
Wahlberg & Wheat (2013) [59][14]	N/A	N/A	N/A	N/A	84 (60–109)	77 (50–104)	N/A	N/A	42	8	0
Braby et al. (2006) [45][15]	N/A	N/A	N/A	N/A	95 (83–112)	N/A	N/A	N/A	4	90	4
Espeland et al. (2015) [15][16]	N/A	N/A	N/A	N/A	N/A	N/A	77 (65–89)	81 (72–94)	5	201	2
Seraphim et al. (2018) [52][17]	N/A	N/A	N/A	N/A	N/A	N/A	N/A	56 (52–61)	9	304	3

Credibility or confidence intervals, depending on dating method used, are shown in parentheses where available. See footnotes for additional details.

[1]Several of the fossils used likely placed too low in the tree, and too old minimum age used for fossils in dominican amber [21]. Rooted with a hesperiid, which is now known to be incorrect.

[2]Fossil used to calibrate within Pierinae is now considered to be a nymphalid [22]. Ages shown here are approximate, based on the figure in the paper, since no exact values were given.

[3]Ranges of dates based on multiple datasets. Used various combination of single and multi-copy genes. Ages based on molecular clock rates.

[4]No butterfly fossils used, only a secondary calibration. Additionally, six other Lepidoptera fossils were used.

[5]Phylogeny of the insects. Only two butterfly taxa included. Calibrated with 37 insect fossils. Inferred age of butterflies is younger than the oldest butterfly fossil [24].

[6]Phylogeny of hexapods. Calibrated with 85 fossils.

[7]Reanalysis of data from [23] with additional non-Lepidopteran fossils. Ages shown here are approximate, based on the figure in the paper, since no exact values were given.

[8]Fossil used for calibration within Pieridae is no longer considered to be a pierid. One fossil likely placed too low in the tree [21].

[9]Ignores Hedylidae. Credibility interval for Papilionoidea is approximate, based on the figure in the paper, since no exact values are given.

[10]Based only on secondary calibrations. Ages shown are approximate, based on the figure in the paper, since no exact values were given.

[11]Youngest of multiple age estimates. *Praepapilio* fossil placed too low in phylogeny [21].

[12]Fossil used to calibrate crown Pieridae no longer considered to be a pierid [22].

[13]Unclear what fossil was used since age and placement of calibration (25 Ma, root of Hesperiinae) does not match the fossil described in the provided reference (stem of Coeliadinae, 55 Ma) [24].

[14]No papilionids, hesperiids, and hedylids included. Phylogeny of Arthropods.

[15]Three of the fossils used for calibration within Pieridae are no longer considered to be pierids ([21, 22]).

[16]Fossils likely placed too low in the tree [21].

[17]*Riodinella* fossil used for calibration of Riodininae cannot confidently be placed in Riodinidae according to [21] (and references therein), but see [52].

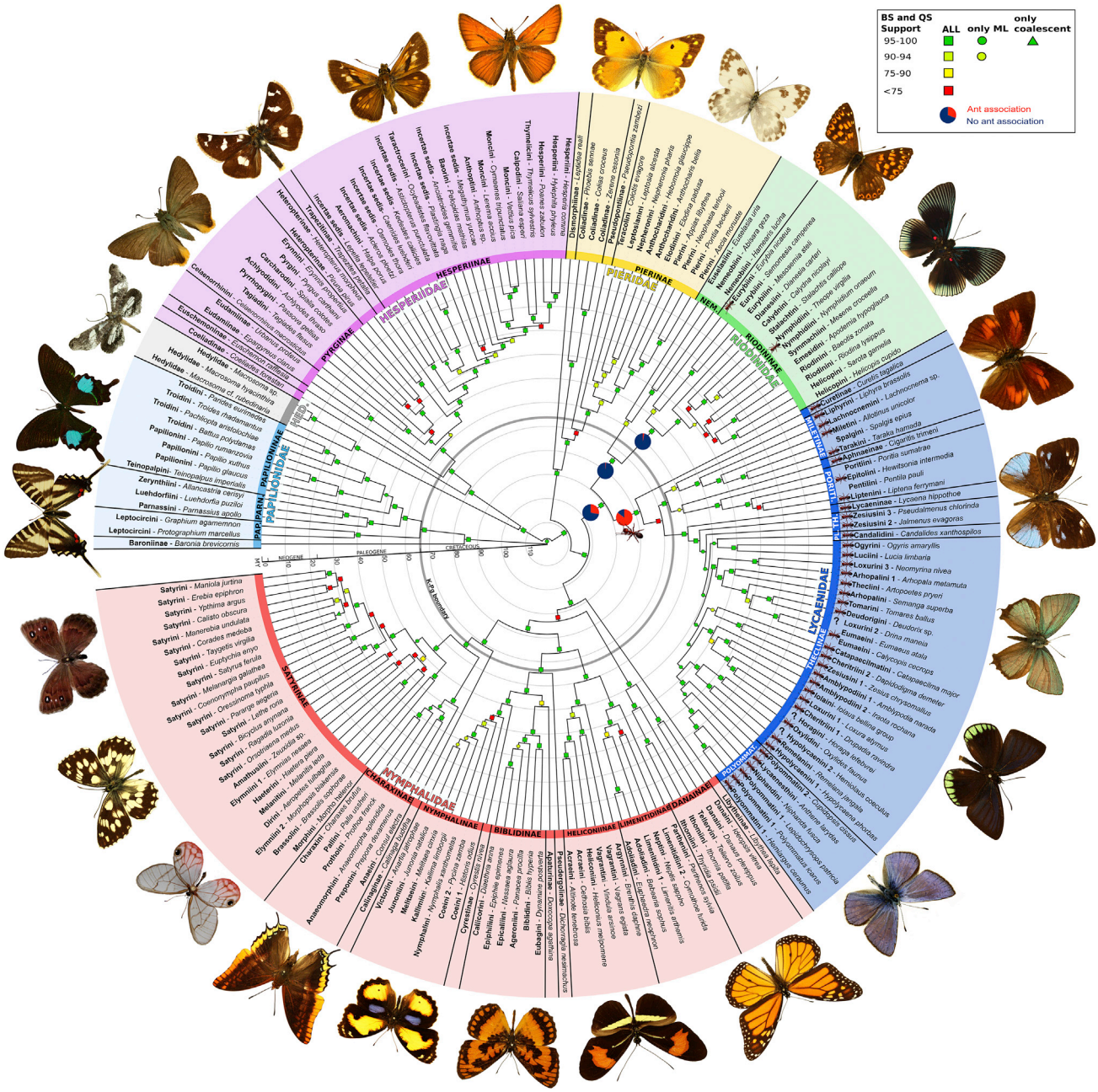

Figure 1. Dated Phylogeny Inferred Using Maximum Likelihood for the Full Dataset (ALL, 207 taxa, 352 loci, >151,000 bps), Related to Table 1

Ages for certain nodes can be found in Table 2; tree with credibility intervals for all nodes on Dryad https://doi.org/10.5061/dryad.v1h63; undated ML tree with outgroups in Figure S2. Branch support is color coded according to the support thresholds indicated on the legend, and the shape indicates whether the branch was supported using one or more methods (see legend). Pie charts on certain nodes in Lycaenidae and Riodinidae indicate the probability of ant association (red) and no ant association (blue) based on ancestral state reconstruction in BayesTraits. Ants before tribe/subfamily names indicate at least one ant-associated species in the group. Question marks denote taxa for which no data on ant association could be found. POLYOMMAT, Polyommatinae; PL, Polyommatinae; TH, Theclinae; PORITI, Poritiinae; NEM, Nemeobiinae; HED, Hedylidae; PARN, Parnassiinae; PAP, Papilionidae.

but more than 25 Ma older than inferred by [17, 32], based on more extensive papilionid taxon sampling, but only three and four genes, respectively, and both with placement of the same fossil species (*Stolopsyche libytheoides* = *Oligodonta florissantensis* [22]) in the wrong family, as in [4] (Table 2). Baroniinae is recovered as sister to the remainder of the family with strong support (bs = 100, QS = 0.82–1), followed by Leptocircini (Papilioninae), Parnassiinae, and the remaining Papilioninae, rendering the latter polyphyletic. Other studies largely place Leptocircini as the sister group to the remainder of the Papilioninae but not always with high support [17, 30, 31]. A non-monophlyetic Papilioninae is here well supported (bs = 100, QS = 0.61–1) but has,

to our knowledge, never previously been proposed. Leptocircini feed solely on Annonaceae, not utilized by any other members of Papilioninae, which mainly feed on Aristolochiaceae, Lauraceae, Rutaceae, and Apiaceae [17]. All possible relationships between the three remaining tribes of Papilioninae have been supported by previous studies. We found that Teinopalpini was recovered as sister to Papilionini and Troidini (bs = 100, QS = 0.62–1), agreeing with [30] but contradicting [17], who found Papilionini and Teinopalpini to be sister groups, and [32], who found Teinopalpini and Troidini to be most closely related. Within Parnassiinae, Parnassiini was sister to Zerynthiini and Luehdorfiini as in previous studies [17, 30, 32].

The diverse skippers (Hesperiidae) with more than 4,000 described species [1] have previously been treated as a superfamily and long been considered to be the sister group to all butterflies (e.g., [29, 33–35]). In contrast, more recent studies supported the "nocturnal butterflies" (Hedylidae) as sister group to the butterflies [36–38]. Our study strongly supports the inclusion of both Hesperiidae and Hedylidae within Papilionoidea, corroborating the most recent molecular studies [4, 28], although inclusion of Hedylidae among the butterflies is still ignored in some studies (e.g., [39]). Alternative topologies with Hedylidae as sister group to the remaining butterflies followed by Hesperiidae (AU test, deltaL = 221.0, p = 0.00) or with Hesperiidae and Hedylidae as sister group to the rest (AU test, deltaL = 525.6, p = 0.03) could be rejected. Hesperiidae and Hedylidae diverged from each other around 106 mya (81–129 mya) in the mid-Cretaceous (bs = 100, QS = 0.97–1). Hedylids reversed back to nocturnality in the upper Cretaceous or the Paleogene and evolved several unique characteristics among butterflies related to nocturnal life such as ultrasonic hearing organs on their wings to combat bat attack [40] and eyes adapted to a nocturnal lifestyle similar to moths [41]. Additionally, they do not have "clubbed" antennae, a feature found in all other butterflies. Rather, they have filiform or pectinate antennae adapted to detect pheromones, otherwise associated with moths. This is likely a consequence of mate attraction via pheromones rather than by sight as in other butterflies [42, 43]. Hedylidae nested within the butterflies contrasts the popular belief that all butterflies are diurnal but is also supported by morphology of all life stages [43]. The exclusively neotropical Hedylidae (c. 40 described species [42]) started diversifying around 29 mya (18–42 mya) in the mid-Oligocene. This is the first crown age estimate of the family, although only based on three taxa (see Table 2). Hedylidae currently consists of the single genus *Macrosoma*, and our inferred age indicates that this is a very old genus or that it should be split up upon closer inspection. Hesperiidae started to diversify significantly earlier than its sister group, around 79 mya (60–99 mya). This is about 14 Ma older than found by [4] but 15 Ma younger than inferred by [44] using one fossil and one secondary calibration, although it is unclear what fossil they actually used (see Table 2). Based on eight loci, [13] found two equally supported topologies for relationships among Hesperiidae, one of which, namely a monophyletic Pyrginae with Coeliadinae as sister group to the remainder of the family followed by Euschemoninae, was supported by our study (Figure 1). Congruent with [13], we found Heteropterinae as sister to Trapezitinae and Hesperiinae.

Pieridae is a medium-sized butterfly family with >1,100 species [1] that started diversifying around 87 mya (67–108 mya) in the upper Cretaceous. Pierinae and all other subfamilies were monophyletic, and as in [14], the Dismorphiinae were well supported as the sister group to the remainder of the family. Other studies [4, 45], on the other hand, found Dismorphiinae as sister group to Pseudopontiinae, a relationship never supported in any of our analyses. According to an AU test, the placement of Pseudopontiinae as sister to the remaining Pieridae excluding Dismorphiinae (deltaL = 26.5, p = 0.185) was not significantly different from Pseudopontiinae being sister group to Pierinae (logL –8927996.6) as found in the ML tree based on ALL). Trees with Pseudopontiinae as the sister group to the remaining pierids (deltaL = 483.5, p = 0.001) or with Pseudopontiinae+Dismorphiinae followed by the remainder of the family (deltaL = 371.5, p = 0.001) could, however, both be rejected. The largest subfamily, Pierinae, has evolved a glucosinolate detoxification mechanism thought to be a coevolutionary key innovation, enabling them to feed on cabbage and relatives (Brassicales), which are poisonous to most other insects [46], and consequently must have evolved after the origin of glucosinolate-containing Brassicales 80–92 mya [47, 48]. We found that Pierinae started diversifying around 60 mya in the early Paleocene, which indicates a longer gap between the origin of the Brassicales and the origin of the detoxification mechanism than found previously [46, 47]. Within Pierinae the tribes form two clades, one consisting of Teracolini as sister to Leptosianini followed by Nepheroniini and Anthocharidini, which is moderately to well supported in all datasets and methods, and the other consisting of Elodini and Pierini, which is well supported. This relationship has remained unresolved in previous molecular analyses [14, 45].

Approximately 75% of the 5,200 lycaenid species [1] and 20% of the 1,500 riodinids [1] are involved in mainly mutualistic associations with ants. The caterpillars possess specialized glands on their cuticle, which provide the ants with a food source rich in sugars and amino acids, and the ants in return protect the caterpillars against parasites and predators [49]. Lycaenids and riodinids are sister groups, but whether ant association is ancestral [18, 19] or evolved independently in these two families [20, 50] remains unknown, since broader studies on butterfly-ant associations have been hindered by the lack of a higher-level phylogeny of lycaenids. By including all subfamilies and nearly all tribes, we here provide the first lycaenid backbone tree. Riodinidae and Lycaenidae split around 90 mya (70–110 mya), and diversification of the extant lineages started around 74 and 78 mya, respectively. We found that ant association must have arisen three times independently, once in the ancestor of Lycaenidae around 78 mya (probability of ant association = 0.831) and twice more recently in the Riodinidae (Figure 1), since the probabilities of ant association in the ancestor of Riodinidae and in the most recent ancestor of the two riodinid ant-associated clades (subtribe Eurybiina, Eurybiini, and Nymphidiini) both were very low (0.007 and 0.02, respectively). This would indicate not only non-homology of ant organs in Lycaenidae and Riodinidae as suggested by [50] but additionally non-homology of the organs within Riodinidae. Our results will also help revise lycaenid tribal classification, which still largely follows the arrangement of Eliot [51]; we found that many of his tribes were polyphyletic and that the polyphyletic blues (Polyommatinae) were nested within the hairstreaks (Theclinae). Our lycaenid backbone showed some resemblance to that inferred by [4], based on much lower taxon sampling, but there Lycaeninae was moving

around, and the backbone in general had little support. Subfamily relationships within Riodinidae agreed with those of [15, 52], although in the latter, Euselasiinae is treated as a tribe within Nemeobiinae. Relationships within the large neotropical radiation containing Helicopini, Nymphidiini, Stalachtini, Symmachiini, Dianesiini, Calydnini, and Emesidini were still largely unsupported in all studies, and branch lengths were extremely short, indicating an explosive radiation, although seemingly not associated with a shift in diversification rate [15]. Stalachtini was the sister group to Nymphidiini, which disagrees with [15, 52] where Stalachtini was found within Nymphidiini, but does agree with morphology since Stalachtini lack all the synapomorphies related to ant association found in Nymphidiini, including the vibratory papillae on the prothorax of the caterpillar, which allow caterpillars to communicate with ants [53].

The most species-rich butterfly family, the Nymphalidae, with more than 6,100 described species [1], originated around 90 mya. This age has remained stable, with most studies inferring ages between 87 and 94 Ma (Table 2). The most comprehensive nymphalid phylogeny to date was published by Wahlberg et al. [16] using 10 genes and 250 morphological characters and including ~75% of the genera. Even with this impressive taxon and gene sampling, the support for the backbone remained low, and relationships among many subfamilies, such as the placement of Libytheinae (snouts) and Danainae (Monarchs and relatives) could not be resolved. We found two equally likely topologies, but neither of them was well supported. Libytheinae was either the sister group to the remainder of Nymphalidae followed by Danainae for trees inferred from the four main datasets (Table 1) (bs = 63–80, QS = 0.73–0.89) or the sister group to Danainae (ML trees for datasets AA, PI1, and LE2), with bootstrap of 77, 100, and 86, respectively. The two alternative topologies were not significantly different (AU test, p = 0.309, deltaL = 5.062). The remainder of the nymphalid backbone was well supported (Figure 1) and largely concordant with the relationships that were indicated but not well supported in [16]. The main exception was the placement of Morphini and Brassolini, which in that study were nested within Satyrinae and here were found as sister group to the remainder of the subfamily.

Conclusion

Our study confirms the power of phylogenomic approaches to resolve challenging arthropod phylogenetic relationships [28]. Adding more than 340 genes to the 10 used previously [6, 16] and tripling the number of taxa included in previous studies confirmed some formerly poorly supported nodes and indicated many novel relationships. A well-supported phylogeny with broad coverage across tribes enables tests of existing hypotheses about higher-level relationships and identification of areas needing further study. Critically, it also serves as a needed scaffold for testing entirely new questions about the tempo and mode of butterfly evolution, such as associations between butterfly and plant clades and the impact of the K-Pg mass-extinction event. Moreover, the phylogeny provides the needed framework for broad comparative studies of the origins of key innovations, such as caterpillar-ant symbioses and other hypothesized drivers of lineage diversification, that have shaped the evolution of this highly studied insect group.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- CONTACT FOR REAGENT AND RESOURCE SHARING
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
  - Taxon sampling
- METHOD DETAILS
  - Probe design
  - Molecular methods
  - Data assembly and cleanup
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Phylogenetic analyses
  - Molecular dating
  - Ant association
- DATA AND SOFTWARE AVAILABILITY

SUPPLEMENTAL INFORMATION

Supplemental Information includes two figures, one table, and one data file and can be found with this article online at https://doi.org/10.1016/j.cub.2018.01.061.

ACKNOWLEDGMENTS

We thank all relevant authorities and countries for permits and G. Adelson, Y. Ang, J. Armstrong, J. Barber, M. Cornwall, V. Dinca, O. Espinoza, S. Fisher, G. Gallice, A.J. Gardiner, J.P.W. Hall, D.J. Hallbritter, J. Hayden, A. Heath, K. Hemeon, M.K. Hojo, N. Homziak, W. Hürter, J. Imamura, Y.-J. Jang, A. Keener, T. Kell, P. Kongnoo, B. Leavel, V. Lukhtanov, D.J. Martins, R. Marx, L. Matthews, C. McBride, M. Miko, C.J. Muller, T. Nagano, S. Nakahara, C. Nano, J. Nielsen, L. Ramírez, L. Reeves, S. Sáfián, J.G. Sanders, G. Sihezar, M.W. Tan, C. Tello, J. Tennent, W. Tiren, J.D. Turner, J. van Itterbeeck, G. Watson, J.C.R. Willmott, J.I.R. Willmott, D.L. Yong, and local collectors for collecting samples. We acknowledge the staff at the Butterfly Rainforest at the Florida Museum of Natural History for providing additional samples. Kelly Dexter, Samantha Epstein, Vincent Ficarrotta, Ryan St Laurent, Lei Xiao, and Rachel Hawkins assisted with lab work and preparation of specimen vouchers. Jaqueline Miller helped advise the first author during her tenure at University of Florida. A.Y.K., N.E.P., J.B., M.E., M.A.J., K.R.W., R.G., and D.J.L. acknowledge the US National Science Foundation (NSF) for financially supporting parts of this study (DEB-1541500, 1541557, 1541560, 1256742). Specimen collection by D.J.L. was funded by NSF DEB-1120380 and Grant 9285-13 from the Committee for Exploration and Research of the National Geographic Society. Specimen collection by M.E. was supported by a Putnam Expedition grant (Museum of Comparative Zoology), the Research Council of Norway (#204308), and the Hintelmann Scientific Award for Zoological Systematics. Specimen collection by R.V. was supported by two Putnam Expedition grants (Museum of Comparative Zoology) and the Spanish Ministry grant CGL2007-60516/BOS. We acknowledge Research Computing at the University of Florida and at the Faculty of Arts and Sciences, Harvard University for providing computational resources and support that contributed to results reported in this study (https://www.rc.ufl.edu, https://www.rc.fas.harvard.edu). We thank Alan Lemmon and Emily M. Lemmon for organizing the sequencing of *Jalmenus evagoras* and *Lepidochrysops patricia*, Ulrike Kleikamp and France Gimnich at ZFMK photographed butterflies for Figure 1. Stefan Schröder kindly identified the *Arhopala* specimen. Finally, we thank three anonymous reviewers for helpful comments and suggestions.

AUTHOR CONTRIBUTIONS

M.E. and A.Y.K. conceived the study. M.E. performed and supervised the molecular work. M.E. and J.B. did the bioinformatic work required for probe

design and data cleanup. M.E. did the phylogenetic analyses. A.Y.K., J.B., R.G., and M.E. wrote the paper with M.E. taking the lead. M.E., J.B., K.R.W., A.D.W., R.V., R.E., D.J.L., N.E.P., and A.Y.K. provided data. All authors were involved in discussions and contributed to the writing of the final manuscript.

DECLARATION OF INTERESTS

The authors declare no conflict of interest.

Received: September 17, 2017
Revised: November 21, 2017
Accepted: January 19, 2018
Published: February 15, 2018

REFERENCES

1. van Nieukerken, E., Kaila, L., Kitching, I., Kristensen, N.P., Lees, D., Minet, J., Mitter, J., Mutanen, M., Regier, J., Simonsen, T., et al. (2011). Order Lepidoptera Linnaeus, 1758. Zootaxa *3148*, 212–221.
2. Boggs, C.L., Watt, W.B., and Ehrlich, P.R. (2003). Butterflies: Ecology and Evolution Taking Flight (University of Chicago Press).
3. Roe, A., Weller, S., Baixeras, J., Brown, J., Cummings, M., Davis, D., Kawahara, A., Parr, C., Regier, J., Rubinoff, D., et al. (2009). Evolutionary framework for Lepidoptera model systems. In Molecular Biology and Genetics of the Lepidoptera, M. Goldsmith, ed. (CRC Press).
4. Heikkilä, M., Kaila, L., Mutanen, M., Peña, C., and Wahlberg, N. (2012). Cretaceous origin and repeated tertiary diversification of the redefined butterflies. Proc Biol Sci. *279*, 1093–1099.
5. Lemmon, A.R., Emme, S.A., and Lemmon, E.M. (2012). Anchored hybrid enrichment for massively high-throughput phylogenomics. Syst Biol. *61*, 727–744.
6. Wahlberg, N., and Wheat, C.W. (2008). Genomic outposts serve the phylogenomic pioneers: designing novel nuclear markers for genomic DNA extractions of lepidoptera. Syst Biol. *57*, 231–242.
7. Nguyen, L.-T., Schmidt, H.A., von Haeseler, A., and Minh, B.Q. (2015). IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol Biol Evol. *32*, 268–274.
8. Mirarab, S., and Warnow, T. (2015). ASTRAL-II: coalescent-based species tree estimation with many hundreds of taxa and thousands of genes. Bioinformatics *31*, i44–i52.
9. Vachaspati, P., and Warnow, T. (2015). ASTRID: Accurate Species TRees from Internode Distances. BMC Genomics *16* (*Suppl 10*), S3.
10. Felsenstein, J. (1985). Confidence limits on phylogenies: An approach using the bootstrap. Evolution *39*, 783–791.
11. Sayyari, E., and Mirarab, S. (2016). Fast coalescent-based computation of local branch support from quartet frequencies. Mol Biol Evol. *33*, 1654–1668.
12. Mitter, C., Davis, D.R., and Cummings, M.P. (2017). Phylogeny and evolution of Lepidoptera. Annu Rev Entomol. *62*, 265–283.
13. Sahoo, R.K., Warren, A.D., Wahlberg, N., Brower, A.V.Z., Lukhtanov, V.A., and Kodandaramaiah, U. (2016). Ten genes and two topologies: an exploration of higher relationships in skipper butterflies (Hesperiidae). PeerJ *4*, e2653.
14. Wahlberg, N., Rota, J., Braby, M.F., Pierce, N.E., and Wheat, C.W. (2014). Revised systematics and higher classification of pierid butterflies (Lepidoptera: Pieridae) based on molecular data. Zool Scr. *43*, 641–650.
15. Espeland, M., Hall, J.P.W., DeVries, P.J., Lees, D.C., Cornwall, M., Hsu, Y.-F., Wu, L.-W., Campbell, D.L., Talavera, G., Vila, R., et al. (2015). Ancient Neotropical origin and recent recolonisation: Phylogeny, biogeography and diversification of the Riodinidae (Lepidoptera: Papilionoidea). Mol Phylogenet Evol. *93*, 296–306.
16. Wahlberg, N., Leneveu, J., Kodandaramaiah, U., Peña, C., Nylin, S., Freitas, A.V.L., and Brower, A.V.Z. (2009). Nymphalid butterflies diversify following near demise at the Cretaceous/Tertiary boundary. Proc Biol Sci. *276*, 4295–4302.
17. Condamine, F.L., Sperling, F.A.H., Wahlberg, N., Rasplus, J.Y., and Kergoat, G.J. (2012). What causes latitudinal gradients in species diversity? Evolutionary processes and ecological constraints on swallowtail biodiversity. Ecol Lett. *15*, 267–277.
18. Scott, J.A., and Wright, D.M. (1990). Butterfly phylogeny and fossils. In Butterflies of Europe, O. Kudrna, ed. (Wiesbaden: Aula-Verlag), pp. 152–208.
19. Hinton, H.E. (1951). Myrmecophilous Lycaenidae and other Lepidoptera - a summary. Proc Lond Entomol Nat Hist Soc., 1949–50, 111–175.
20. Fiedler, K. (1991). Systematic, evolutionary, and ecological implications of myrmecophily within the Lycaenidae (Insecta: Lepidoptera: Papilionoidea) (Bonn: Zoologisches Forschungsinstitut und Museum Alexander Koenig).
21. Jong, R. (2017). Fossil butterflies, calibration points and the molecular clock (Lepidoptera: Papilionoidea). Zootaxa *4270*, 1–63.
22. Kawahara, A.Y. (2013). Systematic revision and review of the extant and fossil snout butterflies (Lepidoptera: Nymphalidae: Libytheinae). Zootaxa *3631*, 1–74.
23. Misof, B., Liu, S., Meusemann, K., Peters, R.S., Donath, A., Mayer, C., Frandsen, P.B., Ware, J., Flouri, T., Beutel, R.G., et al. (2014). Phylogenomics resolves the timing and pattern of insect evolution. Science *346*, 763–767.
24. De Jong, R. (2016). Reconstructing a 55-million-year-old butterfly (Lepidoptera: Hesperiidae). Eur J Entomol. *113*, 423–428.
25. Mutanen, M., Wahlberg, N., and Kaila, L. (2010). Comprehensive gene and taxon coverage elucidates radiation patterns in moths and butterflies. Proc Biol Sci. *277*, 2839–2848.
26. Regier, J.C., Mitter, C., Zwick, A., Bazinet, A.L., Cummings, M.P., Kawahara, A.Y., Sohn, J.-C., Zwickl, D.J., Cho, S., Davis, D.R., et al. (2013). A large-scale, higher-level, molecular phylogenetic study of the insect order Lepidoptera (moths and butterflies). PLoS ONE *8*, e58568.
27. Kawahara, A.Y., and Breinholt, J.W. (2014). Phylogenomics provides strong evidence for relationships of butterflies and moths. Proc Biol Sci. *281*, 20140970.
28. Breinholt, J.W., Earl, C., Lemmon, A.R., Lemmon, E.M., Xiao, L., and Kawahara, A.Y. (2018). Resolving relationships among the megadiverse butterflies and moths with a novel pipeline for Anchored Phylogenomics. Syst Biol. *67*, 78–93.
29. Ehrlich, P.R. (1958). The comparative morphology, phylogeny and higher classification of the butterflies (Lepidoptera: Papilionoidea). Univ Kans Sci Bull. *39*, 305–370.
30. Simonsen, T.J., Zakharov, E.V., Djernaes, M., Cotton, A.M., Vane-Wright, R.i., and Sperling, F.A.H. (2011). Phylogenetics and divergence times of Papilioninae (Lepidoptera) with special reference to the enigmatic genera *Teinopalpus* and *Meandrusa*. Cladistics *27*, 113–137.
31. Nazari, V., Zakharov, E.V., and Sperling, F.A.H. (2007). Phylogeny, historical biogeography, and taxonomic ranking of Parnassiinae (Lepidoptera, Papilionidae) based on morphology and seven genes. Mol Phylogenet Evol. *42*, 131–156.
32. Wu, L.-W., Yen, S.-H., Lees, D.C., Lu, C.-C., Yang, P.-S., and Hsu, Y.-F. (2015). Phylogeny and Historical Biogeography of Asian *Pterourus* Butterflies (Lepidoptera: Papilionidae): A Case of Intercontinental Dispersal from North America to East Asia. PLoS ONE *10*, e0140933.
33. Kristensen, N.P. (1976). Remarks on the family-level phylogeny of butterflies (Insecta, Lepidoptera, Rhopalocera). J Zoological Syst Evol Res. *14*, 25–33.
34. Scott, J.A. (1984). The phylogeny of butterflies (Papilionoidea and Hesperioidea). J Res Lepid. *23*, 241–281.
35. De Jong, R., Vane-Wright, R.I., and Ackery, P.R. (1996). The higher classification of butterflies (Lepidoptera): problems and prospects. Insect Syst Evol. *27*, 65–101.
36. Minet, J. (1991). Tentative reconstruction of the ditrysian phylogeny (Lepidoptera: Glossata). Insect Syst. Evol. *22*, 69–95.

37. Weller, S.J., and Pashley, D.P. (1995). In search of butterfly origins. Mol Phylogenet Evol. *4*, 235–246.

38. Wahlberg, N., Braby, M.F., Brower, A.V., de Jong, R., Lee, M.-M., Nylin, S., Pierce, N.E., Sperling, F.A., Vila, R., Warren, A.D., and Zakharov, E. (2005). Synergistic effects of combining morphological and molecular data in resolving the phylogeny of butterflies and skippers. Proc Biol Sci. *272*, 1577–1586.

39. Cong, Q., Shen, J., Li, W., Borek, D., Otwinowski, Z., and Grishin, N.V. (2017). The first complete genomes of Metalmarks and the classification of butterfly families. Genomics *109*, 485–493.

40. Yack, J.E., and Fullard, J.H. (2000). Ultrasonic hearing in nocturnal butterflies. Nature *403*, 265–266.

41. Yack, J.E., Johnson, S.E., Brown, S.G., and Warrant, E.J. (2007). The eyes of *Macrosoma* sp. (Lepidoptera: Hedyloidea): a nocturnal butterfly with superposition optics. Arthropod Struct Dev. *36*, 11–22.

42. Scoble, M.J. (1986). The structure and affinities of the Hedyloidea: a new concept of the butterflies. Bull Br Mus Nat Hist Entomol. *53*, 251–286.

43. Scoble, M.J., and Aiello, A. (1990). Moth-like butterflies (Hedylidae: Lepidoptera): a summary, with comments on the egg. J Nat Hist. *24*, 159–164.

44. Sahoo, R.K., Warren, A.D., Collins, S.C., and Kodandaramaiah, U. (2017). Hostplant change and paleoclimatic events explain diversification shifts in skipper butterflies (Family: Hesperiidae). BMC Evol Biol. *17*, 174.

45. Braby, M.F., Vila, R., and Pierce, N.E. (2006). Molecular phylogeny and systematics of the Pieridae (Lepidoptera: Papilionoidea): higher classification and biogeography. Zool J Linn Soc. *147*, 239–275.

46. Wheat, C.W., Vogel, H., Wittstock, U., Braby, M.F., Underwood, D., and Mitchell-Olds, T. (2007). The genetic basis of a plant-insect coevolutionary key innovation. Proc Natl Acad Sci USA *104*, 20427–20431.

47. Edger, P.P., Heidel-Fischer, H.M., Bekaert, M., Rota, J., Glöckner, G., Platts, A.E., Heckel, D.G., Der, J.P., Wafula, E.K., Tang, M., et al. (2015). The butterfly plant arms-race escalated by gene and genome duplications. Proc Natl Acad Sci USA *112*, 8362–8366.

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

49. Pierce, N.E., Braby, M.F., Heath, A., Lohman, D.J., Mathew, J., Rand, D.B., and Travassos, M.A. (2002). The ecology and evolution of ant association in the Lycaenidae (Lepidoptera). Annu Rev Entomol. *47*, 733–771.

50. DeVries, P.J. (1991). Ecological and evolutionary patterns in myrmecophilous riodinid butterflies. In Ant-Plant Interactions, C. Huxley, and D.F. Cutler, eds. (Oxford University Press), pp. 143–156.

51. Eliot, J.N. (1974). The higher classification of the Lycaenidae (Lepidoptera) : a tentative arrangement. Bull Br Mus Nat Hist Entomol. *28*, 371–505.

52. Seraphim, N., Kaminski, L.A., Devries, P.J., Penz, C., Callaghan, C., Wahlberg, N., Silva-Brandão, K.L., and Freitas, A.V.I. (2018). Molecular phylogeny and higher systematics of the metalmark butterflies (Lepidoptera: Riodinidae). Syst Entomol. https://doi.org/10.1111/syen.12282.

53. Devries, P.J. (1990). Enhancement of symbioses between butterfly caterpillars and ants by vibrational communication. Science *248*, 1104–1106.

54. Pohl, N., Sison-Mangus, M.P., Yee, E.N., Liswi, S.W., and Briscoe, A.D. (2009). Impact of duplicate gene copies on phylogenetic analysis and divergence time estimates in butterflies. BMC Evol Biol. *9*, 99.

55. Wahlberg, N., Wheat, C.W., and Peña, C. (2013). Timing and patterns in the taxonomic diversification of Lepidoptera (butterflies and moths). PLoS ONE *8*, e80875.

56. Rainford, J.L., Hofreiter, M., Nicholson, D.B., and Mayhew, P.J. (2014). Phylogenetic distribution of extant richness suggests metamorphosis is a key innovation driving diversification in insects. PLoS ONE *9*, e109085.

57. Tong, K.J., Duchêne, S., Ho, S.Y.W., and Lo, N. (2015). INSECT PHYLOGENOMICS. Comment on "Phylogenomics resolves the timing and pattern of insect evolution''. Science *349*, 487.

58. Talla, V., Suh, A., Kalsoom, F., Dinca, V., Vila, R., Friberg, M., Wiklund, C., and Backström, N. (2017). Rapid Increase in Genome Size as a consequence of transposable element hyperactivity in wood-white (*Leptidea*) butterflies. Genome Biol Evol. *9*, 2491–2505.

59. Wheat, C.W., and Wahlberg, N. (2013). Phylogenomic insights into the cambrian explosion, the colonization of land and the evolution of flight in arthropoda. Syst. Biol. *62*, 93–109.

60. Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., and Madden, T.L. (2009). BLAST+: architecture and applications. BMC Bioinformatics *10*, 421.

61. Katoh, K., and Standley, D.M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol. *30*, 772–780.

62. Edgar, R.C. (2010). Search and clustering orders of magnitude faster than BLAST. Bioinformatics *26*, 2460–2461.

63. Krueger, F. (2015). Trim Galore. A wrapper tool around Cutadapt and FastQC to consistently apply quality and adapter trimming to FastQ files, www.bioinformatics.babraham.ac.uk/projects/trim_galore/.

64. Kück, P., and Longo, G.C. (2014). FASconCAT-G: extensive functions for multiple sequence alignment preparations concerning phylogenetic studies. Front Zool. *11*, 81.

65. Kück, P., Meusemann, K., Dambach, J., Thormann, B., von Reumont, B.M., Wägele, J.W., and Misof, B. (2010). Parametric and non-parametric masking of randomness in sequence alignments can be improved and leads to better resolved trees. Front Zool. *7*, 10.

66. Misof, B., and Misof, K. (2009). A Monte Carlo approach successfully identifies randomness in multiple sequence alignments: a more objective means of data exclusion. Syst Biol. *58*, 21–34.

67. Lanfear, R., Frandsen, P.B., Wright, A.M., Senfeld, T., and Calcott, B. (2017). PartitionFinder 2: New methods for selecting partitioned models of evolution for molecular and morphological phylogenetic analyses. Mol Biol Evol. *34*, 772–773.

68. Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics *30*, 1312–1313.

69. Sukumaran, J., and Holder, M.T. (2015). SumTrees: Phylogenetic tree summarization, https://github.com/jeetsukumaran/DendroPy.

70. Sukumaran, J., and Holder, M.T. (2010). DendroPy: a Python library for phylogenetic computing. Bioinformatics *26*, 1569–1571.

71. Mirarab, S., Reaz, R., Bayzid, M.S., Zimmermann, T., Swenson, M.S., and Warnow, T. (2014). ASTRAL: genome-scale coalescent-based species tree estimation. Bioinformatics *30*, i541–i548.

72. Jombart, T., Kendall, M., Almagro-Garcia, J., and Colijn, C. (2017). treespace: Statistical exploration of landscapes of phylogenetic trees. Mol Ecol Resour. *17*, 1385–1392.

73. Yang, Z. (2007). PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol. *24*, 1586–1591.

74. Rambaut, A., Suchard, M.A., and Drummond, A.J. (2014). Tracer 1.6, http://tree.bio.ed.ac.uk/software/tracer/.

75. Zhan, S., Merlin, C., Boore, J.L., and Reppert, S.M. (2011). The monarch butterfly genome yields insights into long-distance migration. Cell *147*, 1171–1185.

76. Kersey, P.J., Allen, J.E., Armean, I., Boddu, S., Bolt, B.J., Carvalho-Silva, D., Christensen, M., Davis, P., Falin, L.J., Grabmueller, C., et al. (2016). Ensembl Genomes 2016: more genomes, more complexity. Nucleic Acids Res. *44* (D1), D574–D580.

77. Knowles, L.L., and Kubatko, L.S. (2011). Estimating species trees: Practical and theoretical aspects (Hoboken, NJ: John Wiley and Sons).

78. Camargo, A., Avila, L.J., Morando, M., and Sites, J.W., Jr. (2012). Accuracy and precision of species trees: effects of locus, individual, and base pair sampling on inference of species trees in lizards of the

Liolaemus darwinii group (Squamata, Liolaemidae). Syst Biol. *61*, 272–288.

79. Nishikawa, H., Iijima, T., Kajitani, R., Yamaguchi, J., Ando, T., Suzuki, Y., Sugano, S., Fujiyama, A., Kosugi, S., Hirakawa, H., et al. (2015). A genetic mechanism for female-limited Batesian mimicry in *Papilio* butterfly. Nat Genet. *47*, 405–409.
80. Heliconius Genome Consortium (2012). Butterfly genome reveals promiscuous exchange of mimicry adaptations among species. Nature *487*, 94–98.
81. Ahola, V., Lehtonen, R., Somervuo, P., Salmela, L., Koskinen, P., Rastas, P., Välimäki, N., Paulin, L., Kvist, J., Wahlberg, N., et al. (2014). The Glanville fritillary genome retains an ancient karyotype and reveals selective chromosomal fusions in Lepidoptera. Nat Commun. *5*, 4737.
82. Nowell, R.W., Elsworth, B., Oostra, V., Zwaan, B.J., Wheat, C.W., Saastamoinen, M., Saccheri, I.J., Van't Hof, A.E., Wasik, B.R., Connahs, H., et al. (2017). A high-coverage draft genome of the mycalesine butterfly *Bicyclus anynana*. Gigascience *6*, 1–7.
83. Carter, J.-M., Baker, S.C., Pink, R., Carter, D.R., Collins, A., Tomlin, J., Gibbs, M., and Breuker, C.J. (2013). Unscrambling butterfly oogenesis. BMC Genomics *14*, 283.
84. O'Neil, S.T., Dzurisin, J.D.K., Williams, C.M., Lobo, N.F., Higgins, J.K., Deines, J.M., Carmichael, R.D., Zeng, E., Tan, J.C., Wu, G.C., et al. (2014). Gene expression in closely related species mirrors local adaptation: consequences for responses to a warming world. Mol Ecol. *23*, 2686–2698.
85. Kozak, K.M., Wahlberg, N., Neild, A.F.E., Dasmahapatra, K.K., Mallet, J., and Jiggins, C.D. (2015). Multilocus species trees show the recent adaptive radiation of the mimetic *heliconius* butterflies. Syst Biol. *64*, 505–524.
86. Cong, Q., Shen, J., Borek, D., Robbins, R.K., Otwinowski, Z., and Grishin, N.V. (2016). Complete genomes of Hairstreak butterflies, their speciation, and nucleo-mitochondrial incongruence. Sci Rep. *6*, 24863.
87. International Silkworm Genome Consortium (2008). The genome of a lepidopteran model insect, the silkworm *Bombyx mori*. Insect Biochem Mol Biol. *38*, 1036–1045.
88. Kriventseva, E.V., Tegenfeldt, F., Petty, T.J., Waterhouse, R.M., Simão, F.A., Pozdnyakov, I.A., Ioannidis, P., and Zdobnov, E.M. (2015). OrthoDB v8: update of the hierarchical catalog of orthologs and the underlying free software. Nucleic Acids Res. *43*, D250–D256.
89. Ratnasingham, S., and Hebert, P.D.N. (2007). bold: The Barcode of Life Data System (http://www.barcodinglife.org). Mol Ecol Notes 7, 355–364.
90. Cho, S., Epstein, S.W., Mitter, K., Hamilton, C.A., Plotkin, D., Mitter, C., and Kawahara, A.Y. (2016). Preserving and vouchering butterflies and moths for large-scale museum-based molecular research. PeerJ *4*, e2160.
91. Lanfear, R., Calcott, B., Kainer, D., Mayer, C., and Stamatakis, A. (2014). Selecting optimal partitioning schemes for phylogenomic datasets. BMC Evol Biol. *14*, 82.
92. Chernomor, O., von Haeseler, A., and Minh, B.Q. (2016). Terrace aware data structure for phylogenomic inference from supermatrices. Syst. Biol. *65*, 997–1008.
93. Liu, L., and Yu, L. (2011). Estimating species trees from unrooted gene trees. Syst Biol. *60*, 661–667.
94. Robinson, D.F., and Foulds, L.R. (1981). Comparison of phylogenetic trees. Math Biosci. *53*, 131–147.
95. Ward, J.H., Jr. (1963). Hierarchical grouping to optimize an objective function. J Am Stat Assoc. *58*, 236–244.
96. Hillis, D.M., Heath, T.A., and St John, K. (2005). Analysis and visualization of tree space. Syst Biol. *54*, 471–482.
97. Shimodaira, H. (2002). An approximately unbiased test of phylogenetic tree selection. Syst Biol. *51*, 492–508.
98. Spearman, C. (1904). The proof and measurement of association between two things. Am J Psychol. *15*, 72–101.
99. Šidák, Z. (1967). Rectangular confidence regions for the means of multivariate normal distributions. J Am Stat Assoc. *62*, 626–633.
100. Romiguier, J., Ranwez, V., Delsuc, F., Galtier, N., and Douzery, E.J.P. (2013). Less is more in mammalian phylogenomics: AT-rich genes minimize tree conflicts and unravel the root of placental mammals. Mol Biol Evol. *30*, 2134–2144.
101. Durden, C.J., and Rose, H. (1978). Butterflies from the middle Eocene: The earliest occurrence of fossil Papilionoidea (Lepidoptera). The Pearce-Sellards Series. Texas Memorial Museum *29*, 1–25.
102. Nel, A., and Nel, J. (1986). A propos de deux Lépidoptères fossiles du Stampien d'Aix- en-Provence (Lepidoptera, Hesperiidae et Nymphalidae Satyrinae). Bull Mus Natl Hist Nat. *8*, 343–350.
103. Warren, A.D., Ogawa, J.R., and Brower, A.V.Z. (2009). Revised classification of the family Hesperiidae (Lepidoptera: Hesperioidea) based on combined molecular and morphological data. Syst Entomol. *34*, 467–523.
104. Scudder, S.H. (1875). Fossil butterflies. Mem Am Assoc Adv Sci. *1*, 1–99.
105. DeVries, P.J., and Poinar, G.O. (1997). Ancient butterfly–ant symbiosis: direct evidence from Dominican amber. Proc Biol Sci. *264*, 1137–1140.
106. Peñalver, E., and Grimaldi, D.A. (2006). New data on Miocene butterflies in Dominican amber (Lepidoptera: Riodinidae and Nymphalidae) with the description of a new Nymphalid. Am Mus Novit. *3519*, 1–17.
107. Nel, A., Nel, J., and Balme, C. (1993). Un nouveau Lépidoptère Satyrinae fossile de l'Oligocène du Sud-Est de la France (Insecta, Lepidoptera, Nymphalidae). Linn Belg. *14*, 20–36.
108. Miller, J.Y., Miller, L.D., and Ivie, M.A. (2012). A new lepidopteran fossil from the Canyon Ferry Reservoir deposit in southwestern Montana (Lepidoptera: Nymphalidae: Satyrinae). Bull Allyn Mus. *167*, 1–12.
109. Peña, C., and Wahlberg, N. (2008). Prehistorical climate change increased diversification of a group of butterflies. Biol Lett. *4*, 274–278.
110. Bell, C.D., Soltis, D.E., and Soltis, P.S. (2010). The age and diversification of the angiosperms re-revisited. Am J Bot. *97*, 1296–1303.
111. Foster, C.S.P., Sauquet, H., van der Merwe, M., McPherson, H., Rossetto, M., and Ho, S.Y.W. (2017). Evaluating the impact of genomic data and priors on bayesian estimates of the Angiosperm evolutionary timescale. Syst Biol. *66*, 338–351.
112. Eastwood, R., and Fraser, A. (1999). Associations between lycaenid butterflies and ants in Australia. Aust J Ecol. *24*, 503–537.
113. Heath, A., and Claassens, A.J.M. (2003). Ant association among Southern African Lycaenidae. J Lepid Soc. *57*, 1–16.
114. Clark, G.C., and Dickson, C.G.C. (1971). Life Histories of the South African Lycaenid Butterflies (Cape Town: Purnell).
115. Meade, A., and Pagel, M. (2017). BayesTraits, http://www.evolution.rdg.ac.uk/BayesTraitsV3/BayesTraitsV3.html.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Biological Samples		
Lepidoptera tissue samples	This paper	See Data S1A
Chemicals, Peptides, and Recombinant Proteins		
GlycoBlue Coprecipitant	ThermoFisher Scientific	cat # AM9516
Chloroform, molecular grade > 99.5%	Sigma-Aldrich	cat # C2432-500ML
Isopropanol, molecular grade > 99%	Sigma-Aldrich	cat # I9516-500ML
Ethanol 99.5%	VWR	cat # 89125-186
Critical Commercial Assays		
Omniprep kit	G-Biosciences	cat # 786-136
QIAGEN DNEASY Blood & Tissue Kit	QIAGEN	cat # 69506
Qubit dsDNA HS Assay Kit	ThermoFisher Scientific	cat # Q32854
Qubit dsDNA BR Assay Kit	ThermoFisher Scientific	cat # Q32853
Repli-g Mini Kit	QIAGEN	cat # 150025
Deposited Data		
WGS data	This paper	NCBI Bioproject: PRJNA432311
Alignments used for probe design	This paper	https://doi.org/10.5061/dryad.v1h63
Alignments used for phylogenetic analyses	This paper	https://doi.org/10.5061/dryad.v1h63
Trees produced from all datasets	This paper	https://doi.org/10.5061/dryad.v1h63
Published sequence data	Multiple	See Table S1C
Software and Algorithms		
genome_getprobe_BLAST.py	This study	https://doi.org/10.5061/dryad.v1h63
IBA pipeline	[28]	http://datadryad.org/resource/doi:10.5061/dryad.rf7g5
BLASTN v. 2.2.31	[60]	https://blast.ncbi.nlm.nih.gov/Blast.cgi
MAFFT v.7.0.1	[61]	https://mafft.cbrc.jp/alignment/software/
USEARCH v.7.0	[62]	https://www.drive5.com/usearch/
Trim Galore! v.0.4.0	[63]	https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/
FASconCAT-G	[64]	https://www.zfmk.de/en/research/research-centres-and-groups/fasconcat-g
Geneious v. 9	Biomatters	https://www.geneious.com
Aliscore v. 2.2	[65, 66]	https://www.zfmk.de/en/research/research-centres-and-groups/aliscore
PartitionFinder v. 2.0.0-pre11	[67]	http://www.robertlanfear.com/partitionfinder/
RAxML v. 8.2.8	[68]	https://sco.h-its.org/exelixis/web/software/raxml/index.html
IQ-TREE v. 1.4.2	[7]	http://www.iqtree.org
SumTrees module, DendroPy Python library	[69, 70]	https://github.com/jeetsukumaran/DendroPy
ASTRAL v. 4.10.8	[8, 71]	https://github.com/smirarab/ASTRAL
Astrid v. 1.4	[9]	https://github.com/pranjalv123/ASTRID
R package treespace v. 1.10.17	[72]	https://cran.r-project.org/web/packages/treespace/index.html
MCMCtree v. 4.9	[73]	http://abacus.gene.ucl.ac.uk/software/paml.html#download
Tracer v. 1.6	[74]	http://tree.bio.ed.ac.uk/software/tracer/

CONTACT FOR REAGENT AND RESOURCE SHARING

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Marianne Espeland (m.espeland@leibniz-zfmk.de).

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Taxon sampling

Ingroup taxa were sampled to include all butterfly families, subfamilies and as many tribes as possible. In several cases more than one member was included for a tribe, especially when previous studies indicated that tribes might not be monophyletic (e.g [16]) or for lycaenids where no molecular phylogeny is available and tribes are still based on the tentative morphological assessment of [51]. Outgroups were mainly selected based on data available from other studies, but two additional species from two families (Gelechiidae and Thyrididae) previously shown to possibly be relatively close to the butterflies (e.g., [28]) were sequenced here (see below). Voucher information can be found in Data S1A.

METHOD DETAILS

Probe design

Loci included in the BUTTERFLY1.0 kit were 336 of the 855 selected from Lepidoptera kit 'LEP1' by [28], plus 89 new loci that were not in this kit. Loci with little or no variation (average pairwise identity > 95%) within the probe region for butterflies were removed, as was any locus not found in all three of the butterflies (*Neophasia, Troides, Euptychia)* included in [28]. Loci for which the nucleotide sequences of butterflies mapped back to multiple locations in the reference sequence (*Danaus plexippus* [75]), with an e-value of 1e-5 or lower, were removed. The probe regions were extended to include larger parts of the exons until a splicing site or the exon boundary was reached for at least one taxon. Exon boundaries were based on those found for *Bombyx mori*, *D. plexippus, Heliconius melpomene* and *Melitaea cinxia* in the Ensembl Metazoa database [76]. Locus length as a proxy of phylogenetic signal can influence phylogeny (e.g., [77, 78]), and we therefore chose to create probe regions that were as long as possible. A few of the loci in LEP1 partially overlapped after extension, and were thus combined. Sequence data from taxa with available genomes [75, 79–82] and transcriptomes [23, 27, 83, 84] as well as Sanger sequencing data from GenBank were added for each locus from [28], while all other Lepidoptera were removed, to make butterfly specific alignments. To pull loci from additional genomes and transcriptomes (Data S1C) we used a custom python script (*genome_getprobe_BLAST.py,* available on Dryad https://doi.org/10.5061/dryad.v1h63) to identify possible blast hits to target loci and then validated orthology of the sequences from these hits following the pipeline of [28] Additionally, we included sequences from a low coverage genome (*Jalmenus evagoras*, Lycaenidae) by assembling the target loci from the raw reads using IBA.py [28], and validated orthology of assembled sequences following the pipeline by [28]. We chose to include 89 additional loci in the butterfly kit to make it more widely applicable (Data S1B). Nuclear loci frequently included in butterfly phylogenetic studies [6, 26, 85], were included to enable direct comparison with previous studies and allow future studies to merge datasets. We also included seven of the proposed nuclear barcoding genes from [86] to make them easily available for phylogenomic studies, as well as several loci with potential biological implications. Finally, the 658 bp barcoding region of cytochrome c oxidase subunit I (COI) was included in the kit for its phylogenetic utility, as a means of easily identifying spurious samples, merging existing datasets for future projects, and assessing possible contamination. We downloaded the corresponding sequences of these 89 additional loci for the three butterflies (*D. plexippus, H. melpomene, M. cinxia)* and the silk moth *B. mori* [87] present on OrthoDB v. 8 [88] to assess orthology and copy number. Sequences of other butterfly species (see alignments for each locus on Dryad https://doi.org/10.5061/dryad.v1h63) from GenBank were then mapped to the *D. plexippus* reference genome using BLASTN of the BLAST+ suite v. 2.2.31 [60], only including hits with an e value lower than 1e-5. Sequences for each locus were aligned using MAFFT v.7.0.1 [61] and used as basis for probe design. Probes of 120 bp were tiled across sequences of all reference taxa for each probe region with 2x tiling. Probes that were $\geq$ 95% identical or with $\geq$ 75bp overlap were collapsed to centroids using the *-cluster_fast* algorithm in USEARCH v. 7.0 [62] to reduce the number of redundant probes. To avoid over-capture of the mitochondrial COI gene compared to the less abundant nuclear genes we only used two probes placed 144 bp apart in regions that appeared to have the least amount of variation across an alignment of publicly available COI barcode data from BOLD [89]. The COI alignments from the BOLD data were trimmed to the two desired probe regions and then collapsed at 90% similarity using USEARCH and the resulting centroids were then used for COI probes. Since flanking regions are also captured the entire barcoding region should be captured using this approach. A total of 56,470 probes were included in the Agilent SureSelect Target Enrichment kit. The 425 alignments used for probe generation can be accessed on Dryad (https://doi.org/10.5061/dryad.v1h63). See Data S1B for locus information.

Molecular methods

DNA was extracted from butterfly tissues preserved either in 80%–100% ethanol and stored at –80°C, stored dry in envelopes in silica gel, or pinned. Voucher information can be found in Data S1A. The amount of tissue ranged from a single leg to approximately 10 mg thoracic tissue, depending on availability. DNA was extracted using the Omniprep DNA extraction kit (G-Biosciences) with several modifications for low quality tissue: Up to 10mg tissue was added to 500μg lysis buffer, ground using a pestle. Thereafter 7μl proteinase K was added and the samples were incubated at 56°C over night. Samples were allowed to cool to room temperature before adding 200μl chloroform, mixing, and centrifuging for 11 minutes at 14,000 g. The upper phase was carefully removed and added to a clean Eppendorf tube before adding 50μl DNA stripping solution, and incubating the samples at 56°C for 8 minutes. Then 100μl precipitation solution was added to the sample followed by centrifugation for 5.5 minutes at 14,000 g. The supernatant was transferred to a new tube and 500μl ice cold (–20°C) isopropanol and 1μl GlycoBlue (ThermoFisher Scientific) were added before incubating the sample for 30 minutes on ice to precipitate the DNA, and centrifuging for 5.5 minutes at 14,000 g. After discarding the

isopropanol, 700μl ice cold 70% ethanol was added to wash the DNA followed by 1.5 minute centrifugation at 14000 g. The ethanol was discarded and the DNA pellet was left to dry for 7-10 minutes at room temperature. Finally, 50-100 μl TE buffer was added to elute the DNA, with the addition of 0.5μl RNase A to remove traces of RNA. DNA was quantified using a Qubit dsDNA HS or BR Assay kit on a Qubit 2.0 fluorometer (ThermoFisher Scientific), and 5 μl of selected samples were run on a 1% agarose gel to assess fragmentation. Samples with DNA concentration lower than 8 ng/μl were whole-genome-amplified using the Repli-g Mini Kit (QIAGEN) (half reactions) and quantified again. Specimen vouchers used are preserved in the Florida Museum of Natural History, Gainesville FL, following published protocols [90], and at the Museum of Comparative Zoology, Cambridge MA.

DNA extracts were submitted to RAPiD Genomics (Gainesville, FL, USA) where libraries were constructed by random mechanical shearing of DNA to an average size of 300 bp followed by an end-repair reaction, ligation of an adenine residue to the 3' end of the blunt-end fragments to allow the ligation of barcoded adapters and PCR-amplification of the library. After library construction custom SureSelect probes (Agilent Technologies) were used for solution-based target enrichment of a pool containing 16 libraries following the SureSelect^{xt} Target Enrichment System for Illumina Paired-End Multiplexed Sequencing Library protocol. Following enrichment pooled libraries were sequenced using Illumina HiSeq 3000 to generate paired-end 100-bp reads.

Specimens for whole genome sequencing (WGS, *Parnassius apollo, Jalmenus evagoras, Lepidochrysops patricia*) were extracted using the QIAGEN DNEASY Blood & Tissue kit (QIAGEN), and concentration was assessed and libraries prepared as above.

Parnassius apollo was sequenced at Starseq (Mainz, Germany), and *Jalmenus evagoras* and *Lepidochrysops patricia* at HudsonAlpha Institute for Biotechnology (Huntsville, Alabama). All other samples were sequenced at RAPiD Genomics (Gainesville, FL).

Data assembly and cleanup

Data assembly and clean up followed the pipeline by [28] for anchored phylogenomics. Paired-end Illumina data were cleaned with Trim Galore! v. 0.4.0 [63] allowing a minimum read size of 30 bp and trimming to remove bases with a Phred score < 20. For each taxon, all loci were assembled using iterative baited assembly (IBA.py, [28]) using only reads with a forward and reverse read that passed filtering. Assembled reads from the probe region were searched against to the *D. plexippus* genome and BLAST results were used for single hit and orthology filtering using the python script s_hit_checker.py with a bit score threshold of 0.90 following [28]. Orthologs were assessed and screened for contamination following Breinholt et al. [28] to identify and remove sequences that are nearly identical at different taxonomic levels (family and genus). Lastly, loci were aligned with MAFFT v. 7 [61] and a strict consensus was made of any assembled isoform (different copies of DNA in heterozygous individuals) using FASconCAT-G [64].

QUANTIFICATION AND STATISTICAL ANALYSIS

Phylogenetic analyses

In the phylogenetic analyses we only included loci that had been captured for at least 75% of the included taxa, so the full matrix consisted of 207 taxa, 352 loci and 151,372bp, of which 10.3% were missing data. Eleven taxa had more than 40% missing data. Average pairwise identity of loci varied from 71 to 91%, and average % GC from 29.2 to 71.2 (Table 1, Data S1B). The shortest locus contained 150bp and the longest 1857bp. All loci are coding, and the reading frame was identified using Geneious 9 (Biomatters) before realigning by translation also in Geneious, and trimming to frame prior to further analyses. All alignments are available on Dryad https://doi.org/10.5061/dryad.v1h63.

To assess eventual randomly evolving sites caused by saturation or ambiguous alignment we ran Aliscore v. 2.0 [65, 66], to check for randomly evolving sites, and none were found.

Loci were concatenated using FASconCAT-G [64], and the best partitioning scheme (partitioned by locus and codon position) was found using PartitionFinder v. 2.0.0-pre11 [67] with the GTR+G model, RAxML [68] and the rcluster algorithm [91], with the rcluster-percent set to 20, under the AICc criterion. The best partitioning scheme was then used as input for model selection in IQ-TREE 1.4.2 using the parameter TESTNEWONLY, [7, 92]. All partitions shared the same set of branch lengths, but could have their own substitution model and rate of evolution (-spp option). This was followed by 250 independent likelihood searches from a random starting tree, each with a single standard nonparametric bootstrap replicate, also run in IQ-TREE v. 1.4.2. Bootstrap trees were summarized on the tree with the highest likelihood (maximum likelihood, ML) using the SumTrees module [69] of the DendroPy Python library [70]. Additionally, we inferred a tree based on amino acids for all loci with pairwise identify lower than 95% (dataset AA). Partitioning scheme was inferred as above, and model selection was performed in IQ-TREE while restricting the possible models to DAYHOFF, WAG, LG, JTT and JTTDCMut, with and without rate heterogeneity, to decrease computational time. This was followed by the same likelihood search and bootstrap scheme as above. *Plutella xylostella* (Plutellidae), which in previous studies (e.g., [28]) has been identified as the most distant relative to the butterflies included in our study was used as an outgroup to root the trees.

To account for possible incomplete lineage sorting species trees were also inferred using the coalescent summary methods ASTRAL-II v. 4.10.8 [8, 71] and Astrid v. 1.4. [9], since these are among the few methods that can handle a large amount of data and also missing taxa in the gene trees. We applied both methods since they are based on fundamentally different principles. Whereas Astrid is a faster implementation of Njst [93] (species relationships are based on internode distances), ASTRAL finds the species tree with the maximum quartet support. Loci were partitioned to codon positions, and partitioning schemes and substitution models for each individual locus were estimated in IQ-TREE followed by 50 likelihood tree searches. The resulting ML gene trees were used as input for coalescent summary methods. Branch support for all species trees was calculated using the quartet score [11], which measures the proportion of quartets in the scored tree found in all the gene trees, (QS) in ASTRAL.

In addition to inferring trees from the total 352 loci dataset we partitioned the data in various ways to study the effects of nucleotide bias (GC content), phylogenetic signal (locus length), evolutionary rate (pairwise identity) and missing data on topology (Table 1). The normalized quartet score (NQS) of all trees inferred using IQ-TREE, ASTRAL-II and ASTRID on all data partitions was calculated in ASTRAL-II (Table 1) to assess topological discordance. To assess effects of missing data we analyzed one dataset where we removed all taxa with more than 25% missing data (19 taxa, Data S1D) from dataset LE3 (LE4M), which had the highest NQS score (Table 1), and thus potential effects could be investigated without the issues leading to discordance found in other datasets. To compare tree topologies generated by the different datasets and methods, the unweighted Robinson-Foulds (RF) distance [94] was calculated for all the combinations of methods and datasets outlined in Table 1 where all taxa were present (two principal components retained), using the R package treespace v. 1.10.17 [72]. This method maps differences between phylogenetic trees into a small number of dimensions for easy identification of clusters. Then we used hierarchical clustering with Ward's method [95] on the principal components to identify groups of similar trees (threshold set to 75). The result was visualized using multidimensional scaling [96]. The approximate unbiased (AU) test [97] in IQ-TREE with 10000 bootstrap replicates AU test was used to test previous hypotheses of butterfly relationships, by comparing tree topologies congruent with the previous hypotheses with the tree with the highest overall likelihood (ML tree, ALL).

Discordance between gene trees and species trees was quantified by calculating unweighted RF-distances between each gene tree and the species trees (from IQ-TREE, Astrid and ASTRAL-II) inferred from the full dataset in RAxML v. 8.2.8 while taking into account that gene trees might only include a subset of the taxa in the species tree (Data S1B). Correlation between the RF distances and locus length, GC content, pairwise identity and missing data was calculated using Spearman's rank correlation [98] taking multiple testing into account using Sidak's adjustment [99] (Table S1). Trees inferred using the different datasets and methods show some topological differences (Figure S1), and the trees cluster into seven groups, with most maximum likelihood trees found in one cluster and the summary coalescent trees displaying more varied topologies. Loci with high nucleotide bias (high guanine-cytosine (GC) content, dataset GC3), especially, produced tree topologies widely different from those inferred by other datasets and showed high gene tree – species tree discordance as measured by the NQS (Table 1). Limited phylogenetic performance of GC-rich genes has also been shown for mammals and was attributed to variation in recombination rates in the genome [100]. Gene tree-species tree discordance was highly correlated with nucleotide bias, evolutionary rate and phylogenetic signal, but not with the amount of missing data (Table S1). While the dataset with high GC content (GC3) had the lowest NQS score, followed by short loci (LE1), removing missing data (LE4M) only negligibly improved the NQS ($\leq$0.01), also indicating that missing data has little impact on discordance in our data. Gene tree discordance as measured by NQS was considerably lower for datasets without potential problematic loci than for the full dataset (Table 1), indicating that these loci are suboptimal for phylogenetic inference using summary coalescent methods, and potentially should be removed prior to analysis. Short loci are also responsible for incongruence in our data. The Papilionoidea is, for example, always, monophyletic using ML (with the exception of dataset GC3), while the summary coalescent methods only support a monophyletic Papilionoidea for datasets LE3 and LE4M where short loci have been removed (see trees on Dryad: https://doi.org/10.5061/dryad.v1h63). These were also the two datasets with the overall lowest gene tree discordance (highest NQS, Table 1). Removing most of the problematic loci still leaves some discordance (highest NQS = 0.865; 1.0 would mean that all quartets in the species tree were found in the gene trees), but the NQS of the ML tree, which does not account for ILS and coalescent trees which do account for ILS, is almost identical, indicating that ILS might not be the main cause of the remaining discordance. Most topological discordance can in this study be explained by factors other than ILS, illustrating the importance of assessing factors related to discordance before applying summary coalescent methods.

Molecular dating

The ML tree based on the full dataset (ALL) without outgroups was used as input tree for dating analyses in MCMCtree v. 4.9 [73] using likelihood approximation with calculation of the gradient and Hessian matrix of the branch lengths to speed up computation. Due to the size of the dataset it was left unpartitioned and analyzed under the F84 substitution model with gamma with five rate parameters. The clock was set to independent, the gamma prior for the transition/transversion ratio (kappa) was set to $\alpha = 6$, $\beta = 2$, the gamma prior for variable rates among sites (alpha) to $\alpha = 1$, $\beta = 1$, the Dirichlet-gamma prior for the mean substitution rate to $\alpha = 2$, $\beta = 20$, and the Dirichlet-gamma prior for the rate drift parameter (sigma^2) to $\alpha = 1$, $\beta = 10$. Node age priors for nodes without calibrations were set to uniform (1 1 0.1). The tree prior was set to birth-death and the rate prior to lognormal. To check that the priors were sensible we first ran an analysis without computing the likelihood by setting usedata = 0.

For time calibration of the tree we used nine fossil calibration points discussed in the recent publication by [21]. Fossil calibrations were set as minimum ages, with soft bounds, using a truncated Cauchy distribution, with an offset (p) of 0.1, a scale parameter (c) of 1, and a left tail probability of 0.025 (default in MCMCtree).

Praepapilio colorado (Papilionidae) from from the Green River formation in Colorado, USA has been dated to the early Lutetian in the middle Eocene [101]. It has been used for calibration in three previous studies of the Papilionidae [17, 30, 31]. As suggested by [17, 21] we place this fossil at the root of Papilionidae to calibrate the crown papilionids to a minimum age of 48 Ma.

Pamphilites abdita (Hesperiidae) from Bouches-du-Rhône, Aix-de-Provence, France, has been dated to the Chattinian-Aquitanian, Late Oligocene-early Miocene. Wing venation and markings places it within the subfamily Hesperiinae [102] and following [21] we use this fossil to calibrate the Hesperiinae to a minimum age of 23 Ma. Based on the age and placement of the calibration point given in the paper, [44] apparently also used this fossil as a minimum age of Hesperiinae to date their hesperid phylogeny, but they mentioned that a recently described fossil was used and cited [24], which describes a much older fossil that should be placed elsewhere (see below), so what fossil they used remains somewhat unclear.

Protocoeliades kristenseni (Hesperiidae) is from the late Paleocene-early Eocene of Denmark (Jutland, Island of Fur, 55 Ma) [24]. Wing venation places it within the Hesperiidae, but not within any particular group [103], and following [21] we place it at the stem of Coeliadinae to calibrate the Hesperiidae to a minimum of 55 Ma.

Vanessa pluto (Pieridae) from the Burdigalian, early Miocene of Radoboj, Croatia was considered by [104] to be a pierid based on wing venation that is similar to what is found in many genera in the subfamilies Pierinae and Coliadinae. Again, following [21] we place this fossil at the root of Pierinae+Coliadinae with a minimum age of 16 Ma.

Theope sp. (Riodinidae) is a caterpillar from Dominican amber dated to the Burdigalian, Early Miocene [105]. This fossil was also used by [15] as one of several calibration points to date the Riodinidae. We place this at the root of the tribe Nymphidiini with a minimum age of 15 Ma.

Dynamine alexa (Nymphalidae) is another fossil from Dominican amber with the same age as above. It was placed in the extant genus *Dynamine* based on wing pattern, but it is also similar to other genera in Biblidinae [106]. This fossil was used to calibrate the Biblidinae by [16], but with a too old date (20 Ma) according to [21], and we set the root of Biblidinae to a minimum age of 15 Ma as suggested by [21].

Jupiteria charon (Nymphalidae) from the Florissant formation in Colorado, dated to late Priabonian, in the late Eocene can be placed in the Nymphalinae based on characters of the wing [21]. We therefore placed it at the base of Nymphalinae with a minimum age of 34 Ma.

Lethe montana and *Lethe ? corbieri* (Nymphalidae) from the Canyon Ferry reservoir in Montana, USA and Céreste, France, respectively, were both dated to the Rupelian in the early Oligocene and placed in or close to the extant genus *Lethe* (Satyrinae) [107, 108]. [16, 109] used *L. corbieri* as a calibration point for *Lethe* +(*Satyrodes, Enodia*) and *Lethe+Neope*, respectively. [21] mentioned that both of these fossils are difficult to place and should only be used as a minimum calibration for the root of the Satyrinae. We therefore applied these fossils to calibrate the root of the Satyrinae to a minimum age of 28 Ma.

Chlorippe wilmattae (Nymphalidae) derives from the Florissant formation in Colorado. Based on the open forewing cell it can be placed in a group consisting of Biblidinae, Nymphalinae, Apaturinae and Cyrestinae [16], and following [21] we place it at the stem of this group with a minimum age of 34 Ma.

[21] discuss the use of several fossils in addition to those mentioned above. We chose not include these remaining fossils since they were either of uncertain affinity, too young to be informative, or redundant, i.e., older fossils exists for the same node.

MCMCtree requires a calibration for the root. Since no fossils are available for the root of Papilionoidea we set the maximum age of the root (uniform prior) to the median age of the Angiosperms (139.4 Ma) inferred by [48]. This bound is soft and does not prohibit the inferred age to be higher than the set maximum if suggested by the data. Since most butterflies, and the potential closest relatives, all feed on Angiosperms it is unlikely that the originated earlier than their main host plants. Initial tests using an alternative age of the Angiosperms (189 Ma [110]) showed that this calibration had little effect on the overall ages of the tree, with the root age only being around 2 Ma older than when using the younger age. Even older ages estimated for the Angiosperms (e.g., 221 Ma [111]), were not used since these are similar to the age estimated for Lepidoptera (e.g., [55]), and therefore deemed too old to be appropriate for the root of the butterflies.

The first 4000 generations of the MCMC chains were discarded as burnin and then trees were sampled every 400 generations until a posterior distribution of 10,000 was reached (4,004,000 generations). This was repeated three times and convergence was assessed in Tracer v. 1.6 [74]. Trees from all runs were combined and summarized as a majority rule consensus tree again using MCMCtree.

Ant association

Ant association for each tribe or subfamily was scored as a binary character (presence/absence) based on the literature [20, 49, 112–114]. The ML tree from the full dataset was trimmed down to only contain Lycaenidae and Riodinidae, and maximum one member of each tribe. Any tribe/subfamily where at least one myrmecophilous species is known was coded as myrmecophilous (marked with an ant in Figure 1). This somewhat inflates the amount of ant association in Lycaenidae since tribes or subfamilies with a large fraction of not ant associated species (e.g., Lycaeninae, Candalidini, Eumaeini) are coded as associated, but suits our purpose since our null hypothesis is that ant association is ancestral [19] and if we, with data biased toward ant association, see the opposite we have even stronger reasons for rejecting the null hypothesis.

The analyses were run using the Multistate model and Markov chain Monte Carlo (MCMC) in BayesTraits v. 3.0 [115]. All priors were set using a hyper-prior approach by seeding the mean of the exponential prior from a uniform distribution with interval 0-10. Three runs were run for 10 million generations for each node of interest (See Figure 1) to assure convergence, with the first 1 million removed as burn-in and sampling every 1,000 generations. Probabilities of ant association were then summarized as the mean of the remaining generations.

All analyses were run on the Hipergator2 cluster at the University of Florida and the Odyssey cluster at Harvard University.

DATA AND SOFTWARE AVAILABILITY

Sequences reads, the locus set used for probe design, alignments for all loci and datasets as well as all trees can be found on Dryad, https://doi.org/10.5061/dryad.v1h63. Accession number for the WGS data reported in this study is NCBI Bioproject PRJNA432311.