Whole genome shotgun phylogenomics resolves the pattern and timing of swallowtail butterfly evolution

Rémi Allio, Celine Scornavacca, Benoit Nabholz, Anne-Laure Clamens, Felix A. H. Sperling, Fabien F. Condamine

Running head

Shotgun phylogenomics and molecular dating

Title proposal

Whole genome shotgun phylogenomics resolves the pattern and timing of swallowtail butterfly evolution

Authors

Rémi Allio[1]*, Céline Scornavacca[1,2], Benoit Nabholz[1], Anne-Laure Clamens[3,4], Felix A.H. Sperling[4], and Fabien L. Condamine[1,4]*

Affiliations

[1]*Institut des Sciences de l'Evolution de Montpellier (Université de Montpellier | CNRS | IRD | EPHE), Place Eugène Bataillon, 34095 Montpellier, France;*

[2]*Institut de Biologie Computationnelle (IBC), Montpellier, France;*

[3]*INRA, UMR 1062 Centre de Biologie pour la Gestion des Populations (INRA, IRD, CIRAD, Montpellier SupAgro), 755 avenue du Campus Agropolis, 34988 Montferrier-sur-Lez, France;*

[4]*University of Alberta, Department of Biological Sciences, Edmonton T6G 2E9, AB, Canada.*

Corresponding authors (*): *rem.allio@yahoo.fr*; *fabien.condamine@gmail.com*

Abstract

Evolutionary relationships have remained unresolved in many well-studied groups, even though advances in next-generation sequencing and analysis, using approaches such as transcriptomics, anchored hybrid enrichment, or ultraconserved elements, have brought systematics to the brink of whole genome phylogenomics. Recently, it has become possible to sequence the entire genomes of numerous non-biological models in parallel at reasonable cost, particularly with shotgun sequencing. Here we identify orthologous coding sequences from whole-genome shotgun sequences, which we then use to investigate the relevance and power of phylogenomic relationship inference and time-calibrated tree estimation. We study an iconic group of butterflies - swallowtails of the family Papilionidae - that has remained phylogenetically unresolved, with continued debate about the timing of their diversification. Low-coverage whole genomes were obtained using Illumina shotgun sequencing for all genera. Genome assembly coupled to BLAST-based orthology searches allowed extraction of 6,621 orthologous protein-coding genes for 45 Papilionidae species and 16 outgroup species (with 32% missing data after cleaning phases). Supermatrix phylogenomic analyses were performed with both maximum-likelihood (IQ-TREE) and Bayesian mixture models (PhyloBayes) for amino acid sequences, which produced a fully resolved phylogeny providing new insights into controversial relationships. Species tree reconstruction from gene trees was performed with ASTRAL and SuperTriplets and recovered the same phylogeny. We estimated gene site concordant factors to complement traditional node-support measures, which strengthens the robustness of inferred phylogenies. Bayesian estimates of divergence times based on a reduced dataset (760 orthologs and 12% missing data) indicate a mid-Cretaceous origin of Papilionoidea around 99.2 million years ago (Ma) (95% credibility interval: 68.6-142.7 Ma) and Papilionidae around 71.4 Ma (49.8-103.6 Ma), with subsequent diversification of modern lineages well after the Cretaceous-Paleogene event. These results

show that shotgun sequencing of whole genomes, even when highly fragmented, represents a powerful approach to phylogenomics and molecular dating in a group that has previously been refractory to resolution. [Computational limitations; cross contamination; divergence times; exon capture; fragmented genomes; low-coverage whole genomes; orthology; Papilionidae; shotgun sequencing; supermatrix; supertree.]

Introduction

Next-generation sequencing (NGS) provides vast amounts of data, and effective extraction of its phylogenetic signal has become a key challenge in systematics (Metzker 2010; McCormack et al. 2013). Methods that sequence hundreds or thousands of loci are now cost-efficient and have proven useful for constructing robust phylogenies (Metzker 2010; McCormack et al. 2013). Consequently, phylogenomics has fundamentally changed how we address questions in evolutionary biology, even as NGS methods continue to develop.

Two sequencing methods have risen to the forefront of phylogenomics: transcriptomics (Oakley et al. 2012; Misof et al. 2014; Garrison et al. 2016) and hybrid enrichment (Faircloth et al. 2012; Lemmon et al. 2012; Lemmon and Lemmon 2013), and a third, shotgun sequencing, has recently become attractive (Allen et al. 2017). Transcriptomics relies on sequencing of expressed RNAs, and no knowledge of targeted gene regions is required. However, the availability of fresh or properly stored tissues limits the number of taxa included in such phylogenetic studies (Lemmon and Lemmon 2013; McCormack et al. 2013). In contrast, hybrid enrichment uses DNA probes to hybridize and selectively capture targets from a genome, which requires prior knowledge of the desired targets (Lemmon and Lemmon 2013; McCormack et al. 2013). An advantage of hybrid enrichment techniques is the ease of using ethanol-preserved tissues, old DNA extractions, and in some cases, old museum specimens (e.g. Guschanski et al. 2013; Blaimer et al. 2016). This can greatly increase the number of taxa in a phylogenomic study. However, later studies mining the original data are limited to the conserved regions of the hybrid enrichment. The third sequencing method - shotgun sequencing - can readily provide similar amounts of genomic data as the two other methods (Staden 1979; Anderson 1981; Gardner et al. 1981; Fuentes-Pardo and Ruzzante 2017). This method breaks up template DNA sequences across the genome into many small fragments before sequencing them, which has been used for both

high-level and low-divergence phylogenomic analyses (Harkins et al. 2016; Allen et al. 2017; Pouchon et al. 2018; Zhang et al. 2019). Three main approaches for reconstructing phylogenetic relationships from whole genome shotgun sequencing have recently been developed (Allen et al. 2015; Schwartz et al. 2015; Hughes and Teeling 2018; Pouchon et al. 2018; Zhang et al. 2019). The first involves a search for shared conserved sequences in different species without focus on coding sequences (Schwartz et al. 2015; Pouchon et al. 2018). Both Schwartz et al. (2015) and Pouchon et al. (2018) rely on selecting reads with high similarity with respect to reference contigs to create a *de novo* sequence (i.e. mapping methods). This method is more suitable for low divergence datasets, since mapping to more divergent datasets can result in difficulties when identifying homologous data (Schwartz et al 2015). The second approach is to extract sequences from *de novo* assemblies via a set of predefined orthologous gene clusters (Hughes and Teeling 2018; Zhang et al. 2019). This approach allows focusing on genes of interest while avoiding difficulties in orthology detection, but its use is confined to groups with suitable genomic resources that provide an adequate initial set of orthologous genes. However, orthologous datasets are not available for some groups. Therefore, to make better use of less suitable genomic resources, a third approach was developed by Allen et al. (2015). The advantage of this approach lies in the assembly of predefined targeted genes by selecting reads with an optimized BLAST search step (a standard all-to-all BLAST search would have been impractical due to the number of reads in shotgun sequencing). Extending the rationale of Allen et al. (2015), we used a custom-designed BLAST method to directly annotate *de novo* assemblies of highly fragmented genomes instead of selecting reads. Additionally, rather than using predefined orthologous genes to annotate *de novo* genomes (Allen et al. 2017), we used all genes available from the reference genome. Orthology detection was then performed specifically on our dataset, which is likely to generate more specific data (and potentially a larger amount of

data) than from a restricted focus on a predefined list of genes. This approach allows annotation of divergent and highly fragmented genomes, with the potential to resolve complex phylogenomic relationships and contribute to analyses like molecular dating.

With 18,000+ described species (van Nieukerken et al. 2011), butterflies (Papilionoidea) represent an evolutionarily successful lineage of phytophagous insects in terms of species richness, morphological diversity and ecological habits. Butterflies include numerous biological models and represent some of the most popular invertebrates, demonstrating that lepidopteran phylogeny and evolution are of both scientific and public interest. Attempts to resolve the higher-level phylogeny of butterflies have been based on varied taxonomic sampling and molecular datasets ranging from multi-gene Sanger data (Regier et al. 2009; Mutanen et al. 2010; Heikkilä et al. 2012) to genomic data (Kawahara and Breinholt 2014; Breinholt et al. 2018; Espeland et al. 2018), providing considerable resolution of the higher phylogeny of butterflies.

Swallowtail butterflies (Papilionidae) represent a charismatic and well-known family of butterflies, with colorful wing patterns and extensive morphological diversity - such as wingspans ranging from 2-3 cm (the tiny dragontail butterflies, *Lamproptera*) to 20 cm (the world's largest butterflies, *Ornithoptera*). Their global distribution currently includes 32 genera comprising at least 550 described species (Collins and Morris 1985; Tyler et al. 1994; Scriber et al. 1995). Most species are found in tropical regions, where they reach their greatest species richness within the true swallowtails (*Papilio*, Wallace 1865; Condamine et al. 2012), while mountain-adapted apollo butterflies occur on temperate and cold climates (*Parnassius*, Condamine et al. 2018a). Papilionidae include model organisms that have contributed to fundamental studies in biogeography (Wallace 1865; Condamine et al. 2013), insect-plant interactions (Ehrlich and Raven 1964; Berenbaum and Feeny 2008), speciation (Dupuis and Sperling 2015, 2016), and other areas of evolution and ecology (Scriber et al. 1995; Kunte

2009; Condamine et al. 2012; Kunte et al. 2014). Although numerous studies have investigated the phylogeny of this group (Munroe 1961; Hancock 1983; Igarashi 1984; Miller 1987; Tyler et al. 1994; Caterino et al. 2001; Zakharov et al. 2004; Nazari et al. 2007; Simonsen et al. 2011; Condamine et al. 2012, 2018b), the phylogenetic backbone of Papilionidae has not been resolved, potentially constraining our understanding of global biogeographic processes like those affecting the divergence of key clades of swallowtail butterflies in the Southern Hemisphere (Condamine et al. 2013).

Although phylogenomic studies have examined relationships among lineages of Lepidoptera (Breinholt and Kawahara 2013; Bazinet et al. 2017; Breinholt et al. 2018) and butterflies (Kawahara and Breinholt 2014; Espeland et al. 2018), few have employed comprehensive taxon sampling for swallowtail butterflies. The latest phylogenomic study of butterflies included 14 swallowtail butterflies in 12 genera and 352 loci obtained with anchored hybrid enrichment (Espeland et al. 2018). Most of their inferred relationships were congruent with previous studies, including Baroniinae as sister to the remainder of the family. However, Papilioninae was found to be a strongly supported polyphyletic group, which has never been proposed before (Munroe 1961; Hancock 1983; Miller 1987; Simonsen et al. 2011; Condamine et al. 2012, 2018b). All possible relationships between the four tribes of Papilioninae have been supported by previous studies, although Leptocircini is most often found (albeit not always highly supported) as the sister group to the remainder of the Papilioninae. Non-monophyly of Papilioninae has important implications for our understanding of their evolutionary history. For instance, study of the latitudinal diversity gradient revealed significant differences in diversification rates between tropical and temperate clades and these insights relied on Parnassiinae and Papilioninae being monophyletic sister groups (Condamine et al. 2012). As for other groups, the lack of resolution of phylogenetic relationships within the swallowtail butterflies with molecular and

morphological data can be attributed to (i) evolutionary processes like ancient and rapid diversification of lineages (e.g. birds: Jarvis et al. 2014; Prum et al. 2015; Suh 2016) or ancient hybridization (e.g. living cats: Li et al. 2016), and/or (ii) methodological and sampling artifacts such as missing data, low taxon sampling, or long branch attraction (Nabhan and Sarkar 2012; Roure et al. 2013). Phylogenetic patterns that are not due to artifacts can be important signatures of patterns of diversification, revealing links to events that were responsible for the current diversity of butterflies.

In recent dating studies, butterflies have been found to originate in the mid-Cretaceous, ca. 100-110 million years ago (Ma; Heikkilä et al. 2012; Wahlberg et al. 2013; Espeland et al. 2018). Lineages leading to extant families had all diverged rapidly from each other by 90 Ma, with Papilionidae being the first to diverge from the common ancestor of all butterflies, Nymphalidae diverging from Lycaenidae and Riodinidae about 102 Ma, Hedylidae diverging from Hesperiidae about 99 Ma, and finally Riodinidae diverging from Lycaenidae about 88 Ma. Interestingly, the most recent common ancestor of each butterfly family originated in the Late Cretaceous (70 to 90 Ma), but extant lineages began diversifying only after the K-Pg event at 66 Ma. Estimating a dated phylogenetic hypothesis for more than 18,000 species of butterflies is currently impractical. Just as for vertebrates dated trees that include large clades (Jetz et al. 2012), one solution for dealing with large datasets is to infer a higher-level phylogenomic tree for the main butterfly lineages as a backbone, then perform separate analyses that include all sampled species for each main lineage, and finally to link each clade into the backbone tree.

Our study presents a procedure for inferring fully resolved, strongly supported and complete genus-level phylogenies from low-coverage genome data, here applied to swallowtail butterflies. We perform Illumina shotgun sequencing of whole genomes using both newly-collected and museum specimens that represent all swallowtail butterfly genera.

This analytical pipeline builds on existing methods to (i) generate 41 *de novo* low-coverage whole genomes using shotgun techniques, (ii) build a genome dataset by including other swallowtail (4 in total) and outgroup (16 in total) genomes, (iii) check for cross-contamination, (iv) retrieve orthologous (protein-coding) genes based on a single reference genome, and (v) reconstruct a robust time-calibrated phylogenomic tree. Without needing to restrict our analysis to preselected genes, this thorough pipeline has the potential to extract thousands of orthologous genes (6,621 in our case) from fragmented genomes. Using maximum likelihood, Bayesian phylogenetic analyses and supertree analyses, we evaluate the utility of low-coverage whole genomes for phylogenomics at two systematic levels: across the entire superfamily Papilionoidea and within the family Papilionidae (the main focus of this study). We then test the effect of different protein models of evolution, partitioning strategies, missing data, and measures of node support on the inference of phylogenetic relationships. Finally, we infer the origin of butterflies by estimating divergence times using a relaxed molecular clock calibrated with fossils. This study provides a phylogenomic foundation for evaluating hypotheses on higher-level relationships within Papilionidae and assesses the enigmatic and long-debated status of some genera and tribes. It also gives a timescale for investigating hypotheses on the early evolutionary history of this group, and will ultimately allow better assessments of trait evolution.

Materials and Methods

Taxon Sampling

In order to be phylogenetically informative about the most ancient relationships, our taxon sampling incorporates all described genera in the family Papilionidae (32 genera *sensu* Scriber et al. 1995; Simonsen et al. 2011; Condamine et al. 2012, 2018b). We sampled 41 species representing all subfamilies and all genera of Papilionidae (**Table 1**). We also

included four genomes in the analyses that were already available for swallowtail butterflies (*Papilio glaucus*, Cong et al. 2015a; *P. machaon*, Li et al. 2015; *P. polytes*, Nishikawa et al. 2015; *P. xuthus*, Li et al. 2015). In our taxon sampling, we also included *Papilio joanae* (from the USA), a species of the *machaon* group (Dupuis and Sperling 2015), which we compare to the available *P. machaon* (from China, Li et al. 2015) as a control for our approach. Based on the latest phylogenies of Papilionoidea (Heikkilä et al. 2012; Kawahara and Breinholt 2014; Breinholt et al. 2018), we selected 16 outgroups, of which 14 are families closely related to Papilionidae including: one Hesperiidae (*Lerema accius*, Cong et al. 2015b), one Pieridae (*Phoebis sennae*, Cong et al. 2016a), one Lycaenidae (*Calycopis cecrops*, Cong et al. 2016b), and 11 Nymphalidae (*Heliconius melpomene*, Davey et al. 2016; *Laparus doris*; *Eueides tales*; *Agraulis vanillae*; *Dryas iulia*; *Junonia coenia*; *Melitaea cinxia*, Ahola et al. 2014; *Polygonia c-album*, de la Paz Celorio-Mancera et al. 2013; *Bicyclus anynana*, Nowell et al. 2017; *Pararge aegeria*, Carter et al. 2013*; Danaus plexippus*, Zhan et al. 2011); in addition, two moth species in the families Bombycidae (*Bombyx mori*, Mita et al. 2004), and Tortricidae (*Choristoneura fumiferana*, *de-novo* sequencing) were used to root the phylogeny as these families are distant outgroups of the Papilionoidea (Wahlberg et al. 2013). The lepidopteran data was recovered from Lepbase (http://lepbase.org/). In total, the taxon sampling represents 61 taxa (45 ingroup and 16 outgroup species).

DNA Extractions, Library Preparation and Shotgun Sequencing

For butterfly samples, DNA extractions were obtained using legs or the thorax. Total genomic DNA extraction was performed with DNeasy Blood and Tissue Kits (Qiagen®), digested overnight with proteinase K following manufacturer recommendations, and eluted with AE buffer to either 50 or 100 µl; this method recovered DNA with a concentration of 3-50 ng/µl.

We used the Illumina® Nextera DNA Sample Preparation Kit to provide a fast and easy library preparation workflow delivering whole-genome sequencing libraries. The approach relies on an engineered transposome to simultaneously fragment and tag ("tagment") the input DNA, adding unique adapter sequences in the process. The Nextera library preparation kit is well suited for insect DNA extractions as it only requires 50 ng of DNA as input. A limited-cycle PCR reaction uses these adapter sequences to amplify the insert DNA. The PCR reaction also adds index sequences on both ends of the DNA, thus enabling dual-indexed sequencing of pooled libraries on any Illumina Sequencing System. Based on results of preliminary tests, we optimized the tagmentation and PCR clean-up steps by increasing DNA input from the recommended 50 to 70 ng, and transposome volume from 3.5 to 5 μL. We also modified clean-up of the tagmented DNA by using 35 μL of AMPure® magnetic beads instead of the Zymo® kit as recommended by Illumina. A second clean-up was performed with 30 μL of AMPure beads at the end of library preparations prior to sequencing (sizing of fragments to the desired 400-500 bp size for NextSeq).

For library sequencing, we relied on the NextSeq® series of sequencing systems, which are fast, flexible, high-throughput desktop sequencers. They support a broad range of sequencing applications, with fast turnaround time and moderate output compared to the MiSeq and HiSeq platforms (generating up to 800 million reads pair-ended, 100-120 Gb of data in less than 30 hours). Since prior work showed genome size of swallowtail butterflies to be about 300 Mb (Cong et al. 2015), we multiplexed between 11 and 15 butterfly samples per NextSeq run to give about 10 Gb DNA sequence per sample and obtain low-coverage whole genomes at a sequence depth of about 30x. We used the NextSeq 500/550 High Output v2 kit (300 cycles, 2 x 150 bp) for a total of four NextSeq sequencing runs. We also added several negative controls for each sequencing run, including sham DNA extractions and library preparations, to allow potential removal of reads belonging to laboratory contaminants from

analyses and facilitate assemblies of genomes. The choice of Nextera and NextSeq technology is based on the need to generate numerous mid-size DNA fragments at an affordable cost (compared to HiSeq).

Assembly of Low-Coverage Whole Genomes

The full analytical pipeline is illustrated in **Fig. 1**, and the scripts necessary to reproduce the study are available in the Supplementary Material that accompanies this article, as well as on Dryad (at http://dx.doi.org/10.5061/[NNNN], Appendix S1).

From reads to coding DNA sequences. Using NGS technology (Illumina© NextSeq, paired-end reads with an averaged insert size of 500 bp), we sequenced and assembled 41 new low-coverage whole genomes of Papilionidae (added to four genomes on GenBank). In addition, we sequenced and assembled a new low-coverage whole genome for *Choristoneura fumiferana*, and assembled five outgroup genomes from raw reads available on the Lepbase database (added to ten genomes on GenBank). For these 47 genomes, raw reads were cleaned using Trimmomatic 0.33 (Bolger et al. 2014) by removing low quality bases from their beginning (LEADING:3) and the end (TRAILING:3), by removing reads below 50 bp (MINLEN:50), and by evaluating read quality with a sliding window approach (SLIDINGWINDOW:4:15). Quality was measured for sliding windows of 4 base pairs and had to be greater than 15 on average. A plethora of methods now exists for *de novo* genome assembly (e.g. ALLPATHS-LG, Gnerre et al. 2011; SOAPdenovo, Luo et al. 2012; MaSuRCA, Zimin et al. 2013; Platanus, Kajitani et al. 2014). Here we assembled the genomes using SOAPdenovo-63mer 2.04 (Luo et al. 2012). Several kmer size values (between 27 and 39) were tested for ten genome assemblies, which lead to no substantial difference for the N50 of our assemblies (median of 96 bp of difference between the lowest

and highest N50). Kmer size of 31 was selected for further analysis. Then, we closed gaps emerging during the scaffolding process with SOAPdenovo, using the abundant pair relationships of short reads with GapCloser 1.12 (Bolger et al. 2014) (**Fig. 1**). *Papilio* genomes have recently been successfully assembled using Platanus (Cong et al. 2015), a tool designed to handle highly heterozygous genomes. In fact, when heterozygosity is too elevated, some assemblers split homologous haplotypes into different contigs. We quantified the impact of heterozygosity on our assemblies with a BLAST (Basic Local Alignment Search Tool) search of our contigs against themselves (96% similarity or higher). We found that duplicated portions of the genomes (found in two or more contigs) amount to only about 1% of the genome on average (including repeated elements); this indicates that the level of heterozygosity did not cause abundant artifactual contig duplications in our assemblies. Nonetheless, to deal with potential alleles still present in separate contigs in our assembly (due to heterozygosity, for example), our annotation approach makes a consensus sequence for ambiguous sites (see below and consensus step in **Fig. 1**). Duplicated contigs could also be the result of recent real duplications but we opted for a more conservative approach since our focus is on the deeper phylogeny of the family.

To annotate the sequences of all genomes, we performed a BLAST search using all available proteins for *Papilio xuthus* (**Fig. 1**). We used the tblastn function to annotate nucleotide sequences with reference protein sequences of *Papilio xuthus* (Altschul et al. 2010). Only scaffolds with 60% or more similarity with the reference protein were selected. Several thresholds were tested for our dataset, and we retained 60% because this threshold provided the best trade-off between missing too many nucleotides versus including spurious nucleotides in the sequences. For example, for a threshold of 80% only highly-conserved regions (with less phylogenetic signal) were generally kept, while for a threshold of 40%, a larger proportion of presumably non-orthologous nucleotides were included. For each species,

all scaffolds selected for a single coding DNA sequence (CDS) were aligned with *Papilio xuthus* with TranslatorX (Abascal et al. 2010) to generate a consensus (**Fig. 1**). This approach relies on amino acid translations to generate multiple alignments of nucleotides. All sites showing intraspecific variation were set to N, to conservatively avoid false informative sites. For example, recently duplicated genes could match (BLAST step) the same reference protein-coding gene. In this case, all divergent sites between the two copies of genes are replaced by N in the consensus, which avoids creating false informative sites due to a recent duplication event.

Check for cross-contaminations. Cross-contamination is a known but largely neglected issue (Ballenghien et al. 2017). Using shotgun sequencing, we were particularly exposed to the risk of cross-contamination since we multiplexed between 11 and 15 butterfly samples per sequencing run. Before creating the datasets (**Fig. 1**), we checked the cross-contamination level in our different sequencing runs using CroCo 0.1 (Simion et al. 2018), which was developed for identifying and removing cross contaminants from assembled transcriptomes. For any given focal species, CroCo identifies CDS that have significantly higher coverage (number of reads mapped to the CDS) in another species than the focal one, with each species of the dataset successively considered as focal. To measure relative coverage between two species, CroCo implements a metric, called Fragments per Kilobase Million (FPKM; Mortazavi et al. 2008), that is used to estimate relative coverage for each gene and is directly comparable between genes because the value is normalized by sequencing depth and size of each gene. Originally developed for transcriptomic data, this method can also be applied to CDS annotated in whole genome sequences. CroCo is thereby used to estimate relative coverage for each CDS of each species and to identify CDS that are suspiciously similar among species. CroCo was set to default parameters, i.e. the option -R to use the tool RapMap

for mapping (Srivastava et al. 2016), with values between 0.2 and 300 for minimum and maximum coverage. Any contigs suspected of being contaminated were then discarded in subsequent analyses.

To test the effect of not controlling for cross contamination in orthology assignment and phylogenomic reconstructions, the analyses were performed on both the contaminated and the non-contaminated datasets.

Orthology assignment and phylogenomic datasets. Orthologous proteins were identified with OrthoFinder 2.2.0 (Emms and Kelly 2015). The method produces orthogroups, which are sequence clusters containing genes that descended via speciation from a single gene in the last common ancestor of the species whose genes are being analysed, although some paralogs may be included (mostly in-paralogs). Orthogroups are suitable for phylogenomic datasets, and we selected only orthogroups with one gene per species, to limit gene duplication problems (**Fig. 1**).

We used HMMCleaner 1.8 (Di Franco et al. 2019) to clean CDS alignments from misaligned sequences (gene by gene). This method cleans an alignment by first building a Hidden Markov Model profile of the alignment, and then measuring the score of the different sequence regions along this profile. After that, the sites present in at least two thirds of the sampled species were selected for the phylogenomic dataset. Finally, we performed a last cleaning step using trimAl 1.2rev59 (Capella-Gutiérrez et al. 2009), which is designed to trim alignments for large-scale phylogenomic analyses. We adopted a stringent approach by selecting all CDS for each species that have at least 30% of sites overlapping with 75% of the rest of the sequences (-seqoverlap 30 and -resoverlap 0.75 options).

After these steps, we built two amino-acid phylogenomic datasets to test the impact of missing data (Roure et al. 2013). In *Dataset 1*, we kept all genes present in at least 95% of

species. For the *Dataset 2*, we selected all genes present in at least four species. The two amino acid matrices concatenated hundreds (*Dataset 1*) or thousands (*Dataset 2*) of selected orthologous genes. In addition, since phylogenomic incongruences between amino-acid and nucleotide datasets have been observed (e.g. in spider flies, Gillung et al. 2018), we also created two nucleotide-based versions of *Dataset 1* and *2* (*Datasets 3* and *4*, respectively). Final alignments are available on Dryad (at http://dx.doi.org/10.5061/[NNNN], Appendices S2, S3, S4, and S5).

Phylogenomic Analyses with a Supermatrix Approach

Phylogenomic analyses were performed using both maximum likelihood (ML) and Bayesian Inference (BI) methods on concatenated amino-acid datasets of selected orthologous proteins. ML and Bayesian analyses were implemented with IQ-TREE 1.6.6 (Nguyen et al. 2015) and PhyloBayes MPI 1.8 (Lartillot et al. 2013), respectively.

For *Dataset 1*, a ML analysis with IQ-TREE was first performed using a single LG model for amino acids (Le and Gascuel 2008) including four matrices, each corresponding to one discrete gamma rate category ($+\Gamma_4$ option; Le et al. 2012), and empirical amino acid frequencies estimated from the data (+F option). Node supports were calculated with 100 non-parametric bootstrap (BS) replicates. To compare node supports, a second ML analysis with IQ-TREE was carried out under the same conditions but with 1,000 ultrafast bootstrap (UFBS) replicates (Minh et al. 2013; Hoang et al. 2018). BS values and UFBS values were considered strong when higher than 70% and 95%, respectively. These ML analyses assumed a single rate matrix for the whole dataset; however, rate heterogeneity is widespread in phylogenomic datasets (Yang 1996; Jia et al. 2014) and must be taken into account. IQ-TREE provides a number of site specific frequency models such as the posterior mean site frequency (PMSF) model as a rapid approximation to the time- and memory-consuming profile mixture

models C10 to C60 (Le et al. 2008; a variant of the CAT model in PhyloBayes, Lartillot and Philippe 2004). PMSF is the amino-acid profile for each alignment site computed from an input mixture model and a guide tree, and the PMSF model is much faster and requires much less memory than C10 to C60 models (Wang et al. 2018), regardless of the number of mixture classes. Moreover, simulations and empirical phylogenomic data analyses have shown that PMSF models can be effective against long branch attraction artefacts (Wang et al. 2018). We performed IQ-TREE analyses with the C50 model as well as the PMSF model. The C50 analysis required 466 Gb of memory and more than five days to infer the ML tree, so we did not perform bootstrap analysis. However, we ran 1,000 UFBS replicates for the PMSF analysis. For all IQ-TREE analyses, we estimated the most likely tree with 100 separate ML searches, as well as 100 searches using the –t RANDOM option, which after initial model optimization on a parsimony tree uses 100 random tree topologies as starting trees for each search.

Bayesian phylogenetic reconstruction was conducted using PhyloBayes MPI (Lartillot et al. 2013) under the CAT+F81+Γ_4 mixture model (Lartillot and Philippe 2004). The CAT model allowed us to take into account the across-site heterogeneities in the amino-acid replacement process (Lartillot and Philippe 2004), and has proven to perform well on large molecular datasets (e.g. Chiari et al. 2012). PhyloBayes MPI has been run as follows: two independent Markov chains Monte Carlo (MCMC) analyses starting from a random tree were run until we generated at least 5,000 cycles after convergence (maximum allowed 10,000 cycles), with trees and associated model parameters sampled every cycle. After checking for convergence in both likelihood and model parameters (*tracecomp* subprogram), the trees sampled in each MCMC run before reaching convergence were discarded as burn-in. The 50% majority-rule Bayesian consensus tree and associated posterior probabilities (PP) were

then computed from the remaining trees (*bpcomp* subprogram). We consider node support with PP ≥ 0.95 to be robust.

The size of *Dataset 2* precluded Bayesian analyses. Instead we performed two ML analyses with IQ-TREE and 1,000 UFBS replicates, one using the protein LG+Γ_4+F model for the whole matrix (Le and Gascuel 2008), and one using the mixture PMSF model (Wang et al. 2018).

For both *Datasets 3* and *4*, ML analyses were performed with IQ-TREE with the same settings as above, except that one partition per gene was specified and a best-fitting substitution model for each partition was identified using ModelFinder implemented in IQ-TREE (option MFP, Kalyaanamoorthy et al. 2017). Node supports were evaluated with 1,000 UFBS replicates.

Phylogenomic Analyses with a Supertree Approach

Several studies (e.g. Jeffroy et al. 2006; Kumar et al 2012) have pointed out that high support values can hide statistically significant incongruences at the gene level, with concatenation analyses returning fully-resolved and well-supported trees even when the level of gene incongruence is high. Also, concatenation can be statistically inconsistent with respect to incomplete lineage sorting (ILS, Roch and Steel 2015). We thus decided to perform a supertree analysis on *Dataset 2*. Supertree analyses can be more robust to ILS and better show conflicts among genes and involve two steps: first, partially overlapping, source phylogenetic trees are inferred from primary data, then they are assembled into a larger, more comprehensive tree, called the *supertree*. Thus, we started our analysis by performing phylogenetic inference with IQ-TREE using the LG+Γ_4+F model for protein sequences for each gene in *Dataset 2*. Node supports were calculated with 100 non-parametric BS replicates.

We first used ASTRAL-III 5.6.3 (Mirarab et al. 2014; Zhang et al. 2018), a state-of-the art supertree method for unrooted gene trees that is robust to ILS, on the collection of all unrooted gene trees, having previously collapsed branches with a BS value lower than 70. We estimated quartet support per each internal branch of the ASTRAL supertree (t -1 option). Second, we used SuperTriplets 1.1 (Ranwez et al. 2010), an extremely fast and accurate supertree method based on a triplet-based representation of rooted input trees that is robust to ILS (Warnow 2017). We selected trees containing either *Choristoneura fumiferana* or *Bombyx mori* and rooted them with bppReRoot, which is provided within the BppSuite (https://github.com/BioPP/bppsuite) implemented in Bio++ (Guéguen et al. 2013). Branches with a BS value lower than 70 were collapsed. The resulting rooted trees were given as input to SuperTriplets, which permits a rooted supertree to be built and, alternatively, a given tree to be scored. This package was used to reconstruct a supertree and score the consensus tree previously inferred with IQ-TREE and PhyloBayes. The advantage of SuperTriplets, compared to ASTRAL, is that it permits information from gene tree rooting to be used; more than 80% of gene trees in our dataset contained one of the outgroup species.

Estimation of Gene and Site Concordance Factors

As noted in the previous section, concatenation analyses can return fully-resolved and well-supported trees even when the level of gene incongruence is high (e.g., Jeffroy et al. 2006; Kumar et al. 2012). As recommended in Minh et al. (2018), we measured gene concordant (gCF) and site concordant (sCF) factors to complement traditional bootstrap node-support measures for *Datasets 1* and *3* (760 loci). First, using the concatenation of all 760 loci, a reference tree was inferred with IQ-TREE with a search for substitution partition for each locus via ModelFinder (Kalyaanamoorthy et al. 2017). Second, we inferred a gene tree for

each locus alignment using IQ-TREE with a model selection. Finally, gCF and sCF were calculated using the specific option -scf and -gcf in IQ-TREE (Minh et al. 2018).

Estimation of Divergence Times

The genomic datasets generated in this study, although large and informative, can represent computational encumbrances that render phylogenomic dating intractable over reasonable timeframes (dos Reis et al. 2016; Collins and Hrbek 2018; Smith et al. 2018). Molecular dating analyses were thus performed with *Dataset 1* (amino acids) under a Bayesian relaxed molecular framework using PhyloBayes 4.1c (Lartillot et al. 2009). We enforced the tree topology as the consensus tree previously inferred with IQ-TREE and PhyloBayes. Dating analyses were conducted by partitioning the dataset using the site heterogeneous CAT+GTR+Γ_4 mixture model, as recommended by Lartillot et al. (2009), with a birth–death prior on divergence times (Gernhard 2008), and a relaxed clock model that was set to an uncorrelated lognormal model (Drummond et al. 2006). Fossil calibrations were assigned to a uniform prior distribution with soft bounds (Yang and Rannala 2006).

Constraints on swallowtail clade ages were enforced by fossil calibrations with systematic position assessed using phylogenetic analyses (Condamine et al. 2018a). Four unambiguous and informative fossils belong to Papilionidae, two of which are Parnassiinae (Nazari et al. 2007). The first is †*Thaites ruminiana* (Scudder 1875), a compression fossil from limestone in the Niveau du gypse d'Aix Formation of France (Bouches-du-Rhône, Aix-en-Provence) within the Chattian (23.03–28.1 Ma) of the late Oligocene (Sohn et al. 2012). †*Thaites* was often recovered as sister to Parnassiini, and occasionally as sister to Luehdorfiini + Zerynthiini. Thus, we constrained the crown age of Parnassiinae with a uniform distribution bounded by a minimum age of 23.03 Ma. The second is †*Doritites bosniaskii* (Rebel 1898), an exoskeleton and compression fossil from Italy (Tuscany) from the

Messinian (5.33–7.25 Ma, late Miocene; Sohn et al. 2012). †*Doritites* was reconstructed as sister to *Archon* (Luehdorfiini), in agreement with Carpenter (1992). The crown of Luehdorfiini was thus constrained for divergence time estimation using a uniform distribution bounded with 5.33 Ma. Third is the genus †*Praepapilio,* with two fossil species †*P. colorado* and †*P. gracilis* (Durden and Rose 1978) from the early Lutetian (Eocene) of the Green River Formation (Colorado, U.S.A.). This fossil was used to constrain the crown age of Papilionidae with a uniform distribution bounded by a minimum age of 47.8 Ma (Smith et al. 2003; de Jong 2007).

For the rest of butterflies, we used the recently described fossil of Hesperiidae, †*Protocoeliades kristenseni* (de Jong 2016, 2017) from the Island of Fur, northwest Jutland, Denmark. It is the oldest butterfly fossil, and is related to the subfamily Coeliadinae, which is the first clade to branch off within Hesperiidae (Warren et al. 2009). Since the taxon sampling included one genome of Hesperiidae (*Lerema accius*), we calibrated the stem of Hesperiidae with a minimum age of 55 Ma. Finally, we relied on the oldest non-ambiguous fossil of Nymphalidae to constrain the crown of the family. The taxon †*Prolibythea vagabonda* from the Florissant formation in Colorado (late Eocene: Priabonian 33.9–38.0 Ma), found to be sister to extant *Libytheana* in a phylogenetic analysis (Kawahara 2009), was used to calibrate the crown age of Nymphalidae with a minimum age of 33.9 Ma.

We were unable to use other fossil calibrations, although suitable butterfly fossils exist for other families (e.g. Wahlberg et al. 2009; Sohn et al. 2012), because the corresponding nodes to which the fossil calibrations could be assigned were not present in our phylogeny. In particular, the families Lycaenidae and Riodinidae have few representatives. Moreover, four fossils have been used to date the phylogeny of Pieridae (Braby et al. 2006) but their identification and phylogenetic assignment is doubtful (de Jong 2007, 2016).

PhyloBayes requires a calibration for the root. Since no fossils are available for the root of Papilionoidea, we did not set an *a priori* minimum age for the root of butterflies but we set the maximum age of the root with a uniform prior bounded by the inferred age of angiosperms. Because most butterflies, and the potential closest relatives, all feed on angiosperms, it is unlikely that they originated earlier than their main host plants. Alternative age estimates have been inferred for angiosperms (e.g. 189 Ma, Bell et al. 2010; 140 Ma, Magallón et al. 2015; 221 Ma, Foster et al. 2017) but these ages are close to the estimated age of Lepidoptera (e.g. Wahlberg et al. 2013; Rainford et al. 2014), and are therefore not appropriate for the root of the butterflies. A survey of nine recent dating analyses that estimated 95% credibility intervals (CI) of the crown age of butterflies yielded a mean maximum age of 128.5 Ma, based on the nine following ages: 129.5 Ma (Chazot et al. 2019), 143 Ma (Espeland et al. 2018), 116 Ma (Wahlberg et al. 2009), 128 Ma (Heikkilä et al. 2012), 114 Ma (Wahlberg et al. 2013), 126 Ma (Rainford et al. 2014), 110 Ma (Tong et al. 2015), 162 Ma (Cong et al. 2017), and 128 Ma (Talla et al. 2017). Thus we set a conservative maximum age of 150 Ma for the Papilionoidea. Uniform distributions of internal fossil calibrations were also maximally bounded at 150 Ma. The bound of the uniform distribution is soft and does not prohibit the inferred age to be older than the set maximum if suggested by the data (Yang and Rannala 2006).

All PhyloBayes calculations were conducted by running three independent MCMC until we generated at least 5,000 cycles after convergence (maximum allowed 10,000 cycles), with sampling posterior rates and dates collected every cycle. After checking for convergence in both likelihood and model parameters (*tracecomp*), posterior estimates of divergence times were then retrieved from the sampled trees of each chain after the burn-in period to compute the Bayesian time-calibrated tree and associated 95% CI (*readdiv* subprogram). As

recommended by Brown and Smith (2018), we compared prior and posterior distributions to determine whether signal is coming from the data or the prior.

Results

Low-Coverage Whole Genomes and Phylogenomic Datasets

Illumina sequencing returned a median of 67.6 million quality-filtered reads per species (60.3 million reads after cleaning). **Table 1** presents statistics for all genomes generated and used for this study. The cost per genome in 2015 was USD 458.6 (404.3€) on average including library preparation, NextSeq sequencing, and all laboratory consumables. Our 41 *de-novo* genomes of Papilionidae (plus *Choristoneura fumiferana*) are highly fragmented, as indicated by their low N50 values (median of 526) and high number of scaffolds (median of 1,372,876). On average, 78,468 scaffolds per species were assigned by BLAST using *Papilio xuthus* as the reference for protein-coding genes. Of these, 35,090 scaffolds with at least 60% similarity with the reference protein were selected. On average, three scaffolds were assigned to each protein, and the different scaffolds were aligned to the reference protein to make a consensus. An average of 10,071 proteins of the 15,131 known proteins in *Papilio xuthus* were recovered per genome.

The cross-contamination check using CroCo recovered a low level of cross contamination with a median of 26 out of 10,000 (0.26%) contigs contaminated by species (**Table 1**). Despite a very low level of species cross-contamination on average, we found that this level was significantly higher for *Parnassius imperator* (26.71% of the contigs). All contaminations were removed for downstream analyses.

OrthoFinder was used to find 30,043 orthogroups, where an orthogroup is a set of genes originating by speciation of a gene present in the last common ancestor. Among these orthogroups, we selected those having only one copy per species. The selected groups were

then filtered again, with the genes present in at least 95% of the species comprising *Dataset 1*, while the orthologous genes present in at least four species formed *Dataset 2*. These sets of genes were used to create both nucleotide-based and amino-acid-based matrices. In the smallest matrix, we obtained 760 genes, which represent 288,446 amino acids, 162,859 variable sites (56.5%), and 100,994 parsimony-informative sites (35%). We found an average of 96% of genes per species and a median of 12% missing data per species (gaps and undetermined sites in the supermatrix). In the largest matrix, we obtained 6,621 genes, which represent 1,656,028 amino acids, 1,020,365 variable sites (61.6%), and 608,399 parsimony-informative sites (36.7%). Here we found an average of 65% of genes per species and a median of 31.6% missing data per species.

All orthologous genes identified with OrthoFinder and selected to create *Datasets 1* and *2* were also used to create nucleotide matrices (*Datasets 3* and *4*, respectively). Nucleotide matrices were cleaned independently leading to the fact that the nucleotide and the amino acids dataset are largely, but not completely, overlapping. In *Dataset 3*, we obtained 889,191 nucleotides for 760 genes with a median of 971 bp (average 1171 bp) per gene altogether containing 651,305 variable sites (73.2%) and 449,010 parsimony-informative sites (50.5%), including a median of 11.6% missing data. For *Dataset 4*, we obtained 5,267,461 nucleotides for 6,407 genes with a median of 594 bp (average 822 bp) per gene altogether containing 3,372,338 variable sites (64%) and 2,581,850 parsimony-informative sites (49%), including a median of 32.2% missing data. Due to the redundancy of the genetic code, similarity between species is higher in amino acids sequences than in nucleotide sequences. This had a direct impact in the cleaning step and accounts for the difference in the number of genes in *Dataset 4* compared to *Dataset 2*.

Supermatrix Phylogenomics

We evaluated the robustness of phylogenomic relationships obtained from *Datasets 1* and *2* by testing the impact of the number of genes (760 vs 6,621 CDS), percentage of missing data in the supermatrix (12% vs 32%), effect of the protein model used for the analysis (LG vs PMSF vs CAT), and analytical framework (ML in IQ-TREE vs BI in PhyloBayes).

For *Dataset 1* (760 CDS, 288,446 amino acids, 61 species), the first two analyses (BS and UFBS) used the LG+Γ_4+F model (total CPU time = 13284h:32m/208h:6m, and memory = 5/10 Gb for BS and UFBS analyses, respectively). The inferred topology recovered *Baronia brevicornis* (Baroniinae) as the sister species to all Papilionidae, followed by a clade comprising Papilioninae and Parnassiinae, both of which were monophyletic (**Fig. 2**). When multiple species were sequenced for a genus (*Graphium*, *Ornithoptera*, *Papilio*, *Parnassius*), they were also monophyletic in the analyses (**Fig. 2**, Appendices S6 and S7 available on Dryad). Taking into account site heterogeneity in the supermatrix, the third analysis with the PMSF model (total CPU time = 229h:38m, and memory = 11 Gb) and the fourth analysis with the CAT+F81+Γ_4 model with PhyloBayes reached convergence after 1,500 cycles (total cycles = 6,510 cycles, total CPU time = 161,280h) and provided identical topologies, differing only slightly in branch length estimates (**Fig. 2**, Appendices S6 and S7).

For *Dataset 2* (6,621 CDS, 1,656,028 amino acids, 61 species), we performed only ML reconstructions with IQ-TREE and tested the effect of the protein model (LG+Γ_4+F vs PMSF). The ML analyses with the LG+Γ_4+F model (total CPU time = 1066h:37m, and memory = 57 Gb) yielded the same topology as obtained with the analyses of *Dataset 1*. ML analyses with the PMSF model (total CPU time = 4016h:00m, and memory = 70 Gb) provided a very similar topology, except for the branching of *Parnassius imperator*, which was retrieved as sister to *P. orleans* (Appendices S6 and S7).

For *Dataset 3* (760 CDS, 889,191 nucleotides, 61 species), and *Dataset 4* (6,407 CDS, 5,267,461 nucleotides, 61 species), the best substitution model for each gene was selected

with ModelFinder followed by ML analyses (total CPU time = 127h:26m, and memory = 15 Gb for *Dataset 3*, and total CPU time = 884h:33m, and memory = 76 Gb for *Dataset 4*). The ML analyses provided the same topology as the one obtained with *Datasets 1* and *2*, except for the relationships of *Iphiclides* and *Lamproptera* and the relationships of *Papilio antimachus* and *Papilio polytes*, which were not recovered as sister taxa in the nucleotide-based analyses (Appendices S6 and S7).

Node support was either evaluated with non-parametric bootstrap (BS), ultrafast bootstrap (UFBS) or posterior probabilities (PP, CAT model). The results show maximal support for an average of 96.7% of nodes in Papilionidae for all phylogenomic analyses (**Fig. 2**). All backbone nodes were always supported with maximal values. Both species of the *machaon* group (*P. machaon* from GenBank and our *de-novo* genome of *P. joanae*) were always found as sister groups with small branch lengths. Only two nodes did not have maximal nodal support and were located within *Papilio* (the sister relationships between *P. antimachus* and *P. polytes*: BS = 98, UFBS = 99, PP = 1) and within *Parnassius* (the placement of *P. imperator*: BS = 37, UFBS = 55, PP = 0.78). The inferred phylogeny is thus statistically robust.

Supertree Phylogenomics

The phylogenetic trees obtained by ASTRAL and SuperTriplets had the same topology and pattern of quartet and triplet supports for the nodes (**Fig. 3**), demonstrating the robustness of the supertree analysis. Indeed, the topology was invariant to the method chosen to reconstruct the supertree and whether rooted or unrooted information was used. The SuperTriplets analysis took 13s on a 3,2 GHz Intel Core i3 with 8 Gb RAM in a single thread, while the ASTRAL analysis took 55m on the same computer. Moreover, the supertree topology only

differs from the concatenation tree in the placement of *Parnassius imperator,* showing the robustness to gene-level scrutiny of the phylogenetic analyses performed in this paper.

Gene and Site Concordance Factors

IQ-TREE with ModelFinder returned the same topology as the one obtained in previous analyses (**Fig. 2**)(total CPU time = 965h:11m/441h:04m/5809h:27m for *Datasets 1* and *3* [760 genes] and *2* [6,621 genes] respectively). Concordance factors for each locus were compared with discordance factors, which relate to the proportion of genes (gDF) or sites (sDF) that support a different resolution of the node (Appendix S8). For each node, the most common resolution inferred in the gene trees is the one we obtained with supermatrix and supertree inferences. In fact, gCF is always higher than gDF1 and gDF2. Concerning the sCF and sDF, all but six nodes were supported by more sites than the other configurations (sCF > sDF but slightly, Appendix S8). Interestingly, for three out of the six nodes with a sDF higher than the sCF, UFBS values were not maximal (67, 97 and 97). For the three other nodes, the results highlight interesting nodes of the phylogeny (red squares in **Fig. 3**).

Molecular Dating

Bayesian analyses of divergence times performed with the CAT-GTR model in PhyloBayes reached convergence between 1,500-2,000 cycles (total CPU time [1 thread per chain; 3 chains] = between 6 and 8 months). For a conservative estimate of posterior node ages, 1,500 cycles were discarded as burn-in (Appendices S9 and S10 available on Dryad). Dating analysis results for swallowtails and outgroups are shown in **Fig. 4**. The crown group of butterflies (Papilionoidea) began diversifying in the Late Cretaceous at 99.2 Ma (95% CI: 68.6-142.7 Ma), and swallowtails (Papilionidae) originated in the end of the Late Cretaceous at 71.4 Ma (95% CI: 49.8-103.6 Ma). Subfamilies Papilioninae and Parnassiinae began to

diversify at 52.9 Ma (95% CI: 36.7-77.4 Ma) and at 53.6 Ma (95% CI: 36.9-79.2 Ma), respectively. We recovered early Oligocene to mid-Miocene origins for the species-rich genera: *Papilio* at 22.8 Ma (95% CI: 14.9-34.6 Ma), *Graphium* at 17.5 Ma (95% CI: 9.9-28.7 Ma), and *Parnassius* at 21.2 Ma (95% CI: 12.4-35.2 Ma). Comparison of the prior (uniform) distributions and the posterior (normal) distributions of node ages indicates that the priors did not influence the posteriors (Appendix S11).

Sensitivity analyses performed with and without outgroups yielded very similar median estimates of divergence times, with maximum age differences of two million years (**Table 2**, Appendices S9 and S10). However, we found that including the outgroups reduced the 95% CI by an average of about 40%. Finally, including or excluding *Parnassius imperator* did not affect the median age estimates for the swallowtail groups except for the crown age of *Parnassius*, which had a difference of 6.5 million years (**Table 2**, Appendices S9 and S10).

Cross-Contamination Issues

When cross-contamination checks (with CroCo) were not applied, we retrieved 29,792 orthogroups with OrthoFinder (Emms and Kelly 2015), and *Datasets 1* and *2* contained 959 and 2,993 genes, respectively. Phylogenomic reconstructions provided the same topology as the one obtained after the cross-contamination process, except for Bayesian inference on *Dataset 2* where *Parnassius* was not monophyletic (Appendix S12). We also found that cross contamination impacted phylogenomic inferences by overestimating branch length for several taxa (Appendix S13).

Discussion

Using Shotgun Sequencing for Phylogenomics

Shotgun sequencing is one of the simplest and most affordable of sequencing approaches, requiring minimum sample preparation before sequencing and yielding data that is evenly spread across the genome (Staden 1979; Anderson 1981; Gardner et al. 1981). With current NGS tools (Metzker 2010), this sequencing approach represents an opportunity to rapidly increase phylogenomic sampling. However, one limitation is that shotgun sequencing may require high sequencing effort to obtain useable read coverage, as well as more intensive bioinformatics analyses to find loci of interest compared to other sequencing approaches like capture methods (for which fewer reads are required to obtain sufficient loci coverage due to more specific reads).

Although the use of low-coverage whole-genome data often results in fragmented genomes, it has become a fast-moving field, as shown by the recent development of several pipelines to handle this kind of data. Pipelines like aTRAM (Allen et al. 2015, 2017) and AGILE (Hughes and Teeling 2018) aim to mine and annotate coding sequences from a fragmented target genome that uses a set of predefined orthologous reference genes from a closely related taxon. Other recently-described approaches based on shotgun sequencing (Schwartz et al. 2015; Pouchon et al. 2018) extract nuclear regions shared between species of interest. For example, Pouchon et al. (2018) extracted 1,877 metacontings shared by at least one outgroup and three other taxa, highlighting the usefulness of this approach for phylogenomic reconstruction and subsequent applications.

Here, we meet the challenge of phylogenomic reconstruction by orthologous CDS identification from contigs obtained with whole-genome shotgun sequencing. The method is designed for highly fragmented and low-coverage genomes and requires the availability of a single (related) reference genome. Despite the low-coverage nature of our data, we were able to cost-effectively identify more than 10,000 CDS for 41 newly sequenced species (plus *Choristoneura fumiferana*). Applying a rigorous cleaning procedure, we extracted 6,621

orthologous genes and assembled four genomic datasets including 100,994 (35%) and 608,399 (36.7%) informative amino-acid sites for *Datasets 1* and *2*, respectively; and 449,010 (50.5%) and 2,581,850 (49%) informative nucleotide sites for *Datasets 3* and *4*, respectively. This amount of informative data for phylogenomic analyses is comparable to sequence-capture datasets like UCEs (854 UCE loci for stinging wasps including 143,608 [70.7%] informative nucleotide sites, Branstetter et al. 2017), which now constitute the most widely used approach in phylogenomics (McCormack et al. 2013). Our BLAST-based annotation and orthologous detection was validated because two closely-related species of the *machaon* group were consistently found as sister lineages and had short branch lengths. In addition, both ML and Bayesian phylogenies agreed with several established studies (Simonsen et al. 2011; Condamine et al. 2012) and uncover new relationships (see below). Remarkably, even with poor-quality libraries (*Allancastria cerisyi, Hypermnestra helios* and *Parnassius imperator*), our approach correctly places these species in the same position as in a fully sampled tree of Parnassiinae (Condamine et al. 2018a), although with low support for *Parnassius imperator*.

Our approach could be enhanced by the use of multiple reference genomes, preferably distributed across the phylogeny (i.e. one per tribe), for the BLAST-based annotation step. Using several related species for annotation should increase the number of annotated genes for all species, and thus increase the number of orthologous CDS in the final dataset. Note that our BLAST-based annotation permits the use of divergent genomes as references. However, the use of highly divergent genomes can result in a loss of information due to non-identification of genes that are too divergent.

The Importance of Controlling for Cross Contamination

An increasing number of publications have warned about the effect of cross contamination on phylogenomic inferences (Ballenghien et al. 2017; Philippe et al. 2017; Simion et al. 2018). As previously shown in plants (Laurin-Lemay et al. 2012), we found that cross contamination not only impacts phylogenomic inference with artefactual relationships (Appendix S12) and over-estimated branch lengths (Appendix S13), but it also has an impact on orthology detection (Table 1). Indeed, by using CroCo (Simion et al. 2018) for cross-contamination cleaning, we were able to obtain substantially more 1:1 orthologous genes: 6,621 instead of 2,993 for *Dataset 2*. This may be explained by spurious sequences leading OrthoFinder to incorrectly infer clusters of orthogroups in the similarity graph, reducing the number of 1:1 orthologous groups. We consequently recommend that phylogenomic studies using shotgun sequencing (with multiplexing steps) should carefully check for cross contamination to obtain as many good-quality genes as possible in the final dataset.

Using Shotgun Sequencing for Dating

The explosion in genomic sequences brings new challenges for inferring divergence times (Jarvis et al. 2014; Misof et al. 2014; Tong et al. 2015; dos Reis et al. 2016). Phylogenomic datasets raise two distinct problems: (i) the volume of data makes inference of the entire dataset increasingly more challenging, and (ii) the extent of underlying topological and rate heterogeneity across genes makes model mis-specification a serious concern (Smith et al. 2018). Dating of phylogenomic trees can be performed with methods that rely on a molecular matrix (e.g. BEAST, MCMCTree, PhyloBayes) or on branch lengths of previously inferred gene trees (e.g. PATHd8, r8s, treePL). This choice strongly impacts the computational time to infer a dated tree: branch-length-based methods usually run in minutes while the former take weeks to months. Even though the size of *Dataset 1* was substantial, we were able to use a

molecular-matrix-based method (PhyloBayes), which took at least six months on a computer cluster.

Molecular dating in phylogenomic studies is generally performed with BEAST and MCMCTree (e.g. dos Reis et al. 2012; Misof et al. 2014; dos Reis et al. 2015; Prum et al. 2015; Branstetter et al. 2017; Espeland et al. 2018). Only a few studies have used PhyloBayes to estimate divergence times with genomic data (but see Chiari et al. 2012). We hope that our study will encourage other researchers to also use PhyloBayes for molecular dating analyses. Our study demonstrates that PhyloBayes can scale up to genomic data while appropriately accounting for the site specific heterogeneities of genomic datasets via the CAT model (Lartillot and Philippe 2004). Indeed, the CAT model has been shown to better take into account the heterogeneity in the data than traditional partitioning approaches (sometimes for a limited number of genes) or no partitioning at all, when dating with BEAST and MCMCTree (dos Reis et al. 2012; Misof et al. 2014; dos Reis et al. 2015; Prum et al. 2015; Branstetter et al. 2017; Espeland et al. 2018). Yet, partitioning of the molecular dataset may improve divergence time estimates (shown with simulations and real data in Angelis et al. 2018), which has been demonstrated in a dating analysis using mitogenomes of butterflies (Condamine et al. 2018b).

The main limitation we encountered with PhyloBayes, as it is currently implemented, is that it runs on a single MCMC (although independent MCMC can be launched and mixed); a limitation that also pertains to MCMCTree. It would be useful to have a multi-core version of these programs with Metropolis coupled MCMC. This would increase the number of MCMC to simultaneously explore the landscape of models and parameters and jump to another landscape area to avoid a chain becoming marooned in a local optimum (Altekar et al. 2004).

Computational Limitations for Phylogenomics

The genomic datasets generated in this study and others (e.g. Jarvis et al. 2014; Misof et al. 2014; Branstetter et al. 2017; Breinholt et al. 2018) are so large that some analyses become intractable over time frames that are realistic. We compared the computational time of ML (IQ-TREE) and Bayesian (PhyloBayes) inferences, and found a significant difference between ML analyses running for less than two weeks on 18 threads and Bayesian analyses running for more than three months on 64 threads. Both ML and Bayesian inferences gave identical topologies and similar branch lengths (Appendices S6 and S7). Although Bayesian inference is generally recognized as the gold-standard of phylogenetic analyses, our study shows that ML analyses, as implemented in IQ-TREE, performed just as well for the focal group as Bayesian analyses. In addition, the speed of IQ-TREE allows us to test and compare a vast range of datasets and associated settings in a matter of weeks. With genomic datasets becoming increasingly large (e.g. Jarvis et al. 2014; Misof et al. 2014; Branstetter et al. 2017), methods that intersect with Bayesian inferences, such as by including more sophisticated models like the ML approximation of the Bayesian mixture model (CAT for Bayesian inferences, Lartillot and Philippe 2004; PMSF for ML inferences, Wang et al. 2018), represent an interesting avenue to explore.

Confirming and Uncovering Phylogenomic Relationships within Papilionidae

Using shotgun sequencing of whole genomes, we have provided genomic data for all genera of Papilionidae, a dataset that is potentially useful for more diverse evolutionary questions than those normally encompassed by a family tree. Despite the fragmented nature of the genomes, we obtained a resolved and strongly supported phylogeny displaying the relationships of all extant swallowtail genera. The tree is noteworthy for its node support, with only one node not supported, and that being partly due to the poor quality library of the

species *Parnassius imperator*. Supertree methods (SuperTriplets and ASTRAL) gave the same topology as supermatrix methods, indicating that this topology is robust (**Fig. 3**), which is also confirmed by gene concordance factors (Appendix S8). All phylogenomic analyses showed that *Baronia* is sister to all remaining Papilionidae with maximal node support in Bayesian and ML analyses (**Fig. 2**, Appendix S6). In previous studies *Baronia* has not always been recovered as sister to other Papilionidae, but our result benefits from the largest molecular dataset ever assembled for swallowtail genera and also agrees with the latest Sanger-based phylogenies (Simonsen et al. 2011; Condamine et al. 2012) and a mitogenomic study (Condamine et al. 2018b).

Parnassiinae have previously been found to be paraphyletic using both morphological and molecular data (e.g. Ford 1944; Yagi et al. 1999; Caterino et al. 2001; Michel et al. 2008). Here, Bayesian and ML analyses recovered Parnassiinae, as well as the three included tribes, as monophyletic with maximal support (**Fig. 2**). We further found that Parnassiini is sister to Zerynthiini and Luehdorfiini. These results confirm recent densely-sampled Sanger-based phylogenies (Condamine et al. 2012, 2018a, 2018b) on which biogeographic and diversification analyses have been performed.

Interestingly, our topology conflicts with a recent phylogenomic study of butterflies based on 352 loci, which recovered Papilioninae as non-monophyletic due to the strongly supported inclusion of Parnassiinae between Leptocircini and the rest of Papilioninae (Espeland et al. 2018). Non-monophyly of Papilioninae has never been proposed before, and has important ramifications for the understanding of swallowtail evolutionary history (e.g. evolution of host-plant association, latitudinal diversity gradient). However, regardless of the dataset, our phylogenomic analyses recovered Papilioninae as monophyletic and this result is consistent across the concatenated, quartet-based and triplet-based methods with maximal nodal support, but also with gene and site concordance factors (**Figs. 2 and 3**, Appendices S6

and S13), in agreement with previous studies (e.g. Simonsen et al. 2011; Condamine et al. 2012, 2018b). It is possible that the non-monophyly of Papilioninae in Espeland et al. (2018) arose from their limited taxon sampling in Papilionidae. Indeed, Leptocircini contain 140 species and seven genera, and Parnassiinae comprise 85 species and eight genera. We sampled all genera while Espeland et al. (2018) sampled only two genera for Leptocircini and three genera for Parnassiinae. The lack of key genera that diverged early in Leptocircini (*Iphiclides* and *Lamproptera*) or Parnassiinae (*Hypermnestra* and *Sericinus*) may have led to the apparent non-monophyly of Papilioninae based on exon-capture data. Alternatively, it is possible that our analyses recovered the monophyly of Papilioninae because our datasets rely on two- (*Dataset 1*) and eighteen-fold (*Dataset 2*) more genes than Espeland et al. (2018). Also, previous studies relying on few genes always recovered Papilioninae as monophyletic (e.g. Simonsen et al. 2011; Condamine et al. 2012), and the same is true for studies with morphological characters (e.g. Munroe 1961; Hancock 1983; Miller 1987; Parsons 1996). This suggests that dense taxon sampling is essential to phylogenomic tree reconstruction, since insufficient sampling may lead to highly supported clade relationships that are wrong.

Systematic debates have surrounded the phylogenetic positions of enigmatic genera like *Meandrusa* and *Teinopalpus*, *Cressida* and *Euryades*, or *Iphiclides* and *Lamproptera* (Ford 1944; Hancock 1983; Miller 1987; Tyler et al. 1994; Parsons 1996; Simonsen et al. 2011; Condamine et al. 2012). In the first case, we found strong support for *Teinopalpus* as the sister group of Troidini and Papilionini, with *Meandrusa* as the sister group of *Papilio* and both together forming the tribe Papilionini. This result was suggested by a mitogenome study (Condamine et al. 2018b), but not recovered with Sanger-based phylogenies (Simonsen et al. 2011; Condamine et al. 2012). In the second case, *Cressida* was recovered as sister to *Parides* and *Euryades* with supermatrix and supertree analyses but not with a mitogenomic study, which showed *Cressida* as sister to *Euryades* (Condamine et al. 2018b). This latter result

seems unlikely given that *Cressida* is an Australasian genus, while *Euryades* and *Parides* are both Neotropical, and the divergence of these three lineages dates back to the early-middle Miocene (**Fig. 4**). This combined with the fact that both low node supports are obtained with supertree approaches and site concordance factor is lower than site discordance factors may indicate effects of ILS or hybridization in these parts of the tree, or the effect of model misspecification when reconstructing gene trees or even hidden paralogy. For the third case, *Iphiclides* is found as sister to *Lamproptera* with amino-acid-based phylogenomic analyses, with both being sister to all Leptocircini (**Fig. 2**), but we found *Lamproptera* as sister to all Leptocircini in nucleotide-based phylogenomic analyses (Appendix S6). Gene-tree analyses provide insights into this supermatrix-driven discrepancy with supertree methods showing low node support for the sister relationship (**Fig. 3**), and site concordance factors are lower than site discordance factors despite gene concordance factors being higher than gene discordance factors (although all factors for these branches have low values, Appendix S8), which means that this relationship remains unclear even using the information provided by this large genomic dataset. Our study demonstrates the need for more specific studies to clarify the phylogeny of Leptocircini, which represents a phylogenetic impediment within Papilionidae. Interestingly, two similar topological issues within the genus *Papilio* are revealed for the placement of *P. alexanor*, and for the relationship between *P. antimachus* and *P. polytes,* but may be an artefact of low taxon sampling (10 out of 200 species are sampled in *Papilio*; Nabhan and Sarkar 2012). Further work with more comprehensive taxon sampling is needed to identify the causes of these low supports and is beyond the scope of this study. Such a comprehensive topology will have important evolutionary implications in terms of trait evolution like host-plant associations or historical biogeography.

Cretaceous Origin of Papilionoidea and Paleogene Diversification of Papilionidae

It has been notoriously difficult to date the origin and diversification events of butterflies, due to the scarcity of their fossil record (Sohn et al. 2012, 2015; de Jong 2017) as well as limited taxon and/or molecular sampling. However, a consensus is emerging from recent analyses relying on comprehensive taxon sampling (Chazot et al. 2019) or large genomic sampling (Espeland et al. 2018). Genome-based estimates of divergence times reveal that butterflies (Papilionoidea) originated around 99.2 Ma in the Late Cretaceous (**Fig. 4, Table 2**, Appendix S9). This result largely agrees with the mean age of 106.6 Ma (end of Early Cretaceous) calculated from a survey of ten recent dating analyses estimating the crown age of butterflies (Wahlberg et al. 2009; Heikkilä et al. 2012; Wahlberg et al. 2013; Rainford et al. 2014; Tong et al. 2015; Cong et al. 2017; Talla et al. 2017; Condamine et al. 2018b; Espeland et al. 2018; Chazot et al. 2019). These studies, combined with our genome-based estimates, propose that butterflies appeared in the mid-Cretaceous (ca. 100 Ma), which is biologically plausible given their association with angiosperm host-plants (Ehrlich and Raven 1964). Angiosperms diversified rapidly and rose to ecological dominance in the Cretaceous between 125 and 80 Ma (a.k.a. the Cretaceous rise of angiosperms, Bell et al. 2010; Magallón et al. 2015; Foster et al. 2017). Our dating analyses suggest an origin of butterflies that is concurrent with the global radiation of angiosperms, and subsequent diversification in the extant butterfly families in the Late Cretaceous when angiosperms dominated ecosystems. Angiosperms thus likely acted as a mid-Cretaceous resource-driven enhancer of insect-plant associational diversity that created new opportunities for insect herbivores and pollinators (Labandeira and Currano 2013). Still, these time-calibrated trees indicate a 45-million-year gap (ghost lineage) between the oldest butterfly fossil (a 55-million-year-old hesperiid, de Jong 2016) and the estimated origin of butterflies based on molecular data.

Within butterflies, most extant lineages diverged after the K-Pg boundary (**Fig. 4**, Appendix S9), suggesting that this event had a major impact on the evolutionary history of

butterflies, with lineages possibly going extinct (Wahlberg et al. 2009; Heikkilä et al. 2012). We infer that the most recent common ancestor of the Papilionidae lived in the Late Cretaceous ca. 71.4 Ma, but the divergence of ancestors of all other extant lineages lagged 10 million years behind the end-Cretaceous catastrophe (**Fig. 4**), and likely survived in Northern Hemisphere regions (Condamine et al. 2012, 2013). Such a pattern of diversification suggests clade extinctions at the K-Pg boundary and subsequent diversification of extant clades in the Cenozoic (52.9 Ma for Papilioninae and 53.6 Ma for Parnassiinae, **Fig. 4, Table 2**). Subsequent diversification within the two subfamilies occurred in the Eocene, with almost all lineages leading to currently recognized tribes originating in the early Oligocene at 33.5 Ma on average (ranging from 37.5 Ma for Papilionini to 22.4 Ma for Luehdorfiini) and most genera diverging from sister genera in the Miocene (**Fig. 4, Table 2**). This diversification pattern is similar to that shown in Nymphalidae (Wahlberg et al. 2009), Riodinidae (Espeland et al. 2015) and Hesperiidae (Sahoo et al. 2017), suggesting that common drivers or causes have shaped butterfly diversification dynamics through time.

Conclusion

The utility of whole genomes for building and dating phylogenies has never been more auspicious than today. The successful development of powerful analytical tools, in conjunction with the rapid and massive increase in the availability of genomic data (Fuentes-Pardo and Ruzzante 2017), allows us to resolve and understand evolutionary histories that are more and more complex. We still face important limitations in data accessibility (too few genomes are available) and methodological shortcomings (orthology assessment, running time). However, our approach (and analytical pipeline) has empowered the use of low-coverage and highly fragmented whole genomes, providing productive perspectives for future investigations of other model groups. Applied to an insect radiation, we were able to produce

a much-needed stable backbone for a revised classification of swallowtail butterflies through a fully resolved phylogenomic framework unveiling novel relationships and confirming previous hypotheses. The resulting time-calibrated tree also permits a much better understanding of the major events of Papilionidae diversification for interpreting future comparative studies ranging from ecology to genome evolution.

Acknowledgements

We thank two anonymous referees and associate editor Matthew Hahn who provided excellent and constructive comments that greatly improved the paper. We are grateful to Sophie Dang, Troy Locke and Corey Davis at the Molecular Biology Service Unit of the University of Alberta for their help, assistance and advice on next-generation sequencing. We also thank Frédéric Delsuc for his helpful comments on early talks and drafts. The analyses benefited from the Montpellier Bioinformatics Biodiversity (MBB) platform services. This is contribution ISEM 2019-079 of the Institut des Sciences de l'Evolution de Montpellier.

Supplementary Material

Data available from the Dryad Digital Repository: http://dx.doi.org/10.5061/[NNNN].

Funding

This work was supported by the Marie Curie Action (EU 7th Framework Programme) BIOMME project, IOF-627684 to F.L.C. (jointly supervised by F.A.H.S. and Isabel Sanmartín), by a PICS grant of the CNRS (PASTA project) to F.L.C., by a Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery Grant to F.A.H.S., and by the French ANR (BirdIslandGenomic project, ANR-14-CE02-0002) to B.N.

References

Abascal F., Zardoya R., Telford M.J. 2010. TranslatorX: multiple alignment of nucleotide sequences guided by amino acid translations. Nucleic Acids Res. 38:W7-13.

Ahola V., Lehtonen R., Somervuo P., Salmela L., Koskinen P., Rastas P., Välimäki N., Paulin L., Kvist J., Wahlberg N., Tanskanen J., Hornett E.A., Ferguson L.C., Luo S., Cao Z., de Jong M.A., Duplouy A., Smolander O.P., Vogel H., McCoy R.C., Qian K., Chong W.S., Zhang Q., Ahmad F., Haukka J.K., Joshi A., Salojärvi J., Wheat C.W., Grosse-Wilde E., Hughes D., Katainen R., Pitkänen E., Ylinen J., Waterhouse R.M., Turunen M., Vähärautio A., Ojanen S.P., Schulman A.H., Taipale M., Lawson D., Ukkonen E., Mäkinen V., Goldsmith M.R., Holm L., Auvinen P., Frilander M.J., Hanski I. 2014. The Glanville fritillary genome retains an ancient karyotype and reveals selective chromosomal fusions in Lepidoptera. Nat. Commun. 5:4737.

Allen J.M., Huang D.I., Cronk Q.C., Johnson K.P. 2015. aTRAM - automated target restricted assembly method: a fast method for assembling loci across divergent taxa from next-generation sequencing data. BMC Bioinformatics 16:98.

Allen J.M., Boyd B., Nguyen N.P., Vachaspati P., Warnow T., Huang D.I., Grady P.G.S., Bell K.C., Cronk Q.C.B., Mugisha L., Pittendrigh B.R., Leonardi M.S., Reed D.L., Johnson K.P. 2017. Phylogenomics from whole genome sequences using aTRAM. Syst. Biol. 66:786-798.

Altekar G., Dwarkadas S., Huelsenbeck J.P., Ronquist F. 2004. Parallel Metropolis coupled Markov chain Monte Carlo for Bayesian phylogenetic inference. Bioinformatics 20:407-415.

Altschul S.F., Wootton J.C., Zaslavsky E., Yu YK. 2010. The construction and use of log-odds substitution scores for multiple sequence alignment. PLoS Comput. Biol. 6:e1000852.

Anderson S. 1981. Shotgun DNA sequencing using cloned DNase I-generated fragments. Nucleic Acids Res. 9:3015–3027.

Angelis K., Álvarez-Carretero S., dos Reis M., Yang Z. 2017. An evaluation of different partitioning strategies for Bayesian estimation of species divergence times. Syst. Biol. 67:61-77.

Ballenghien M., Faivre N., Galtier N. 2017. Patterns of cross-contamination in a multispecies population genomic project: detection, quantification, impact, and solutions. BMC Biol. 15:25.

Bazinet A.L., Mitter K.T., Davis D.R., Van Nieukerken E.J., Cummings M.P., Mitter C. 2017. Phylotranscriptomics resolves ancient divergences in the Lepidoptera. Syst. Entomol. 42:82–93.

Bell C.D., Soltis D.E., Soltis P.S. 2010. The age and diversification of the angiosperms re-revisited. Am. J. Bot. 97:1296-1303.

Berenbaum M.R., Feeny P.P. 2008. Chemical mediation of host-plant specialization: the Papilionid paradigm. In: Specialization, Speciation, and Radiation: The Evolutionary Biology of Herbivorous Insects (ed. Tilmon K.J.). California University Press, Berkeley, pp. 3–19.

Blaimer B.B., Lloyd M.W., Guillory W.X., Brady S.G. 2016. Sequence capture and phylogenetic utility of genomic ultraconserved elements obtained from pinned insect specimens. PLoS One 11:e0161531.

Bolger A.M., Lohse M., Usadel B. 2014. Trimmomatic: a exible trimmer for Illumina sequence data. Bioinformatics 30:2114–2120.

Branstetter M.G., Danforth B.N., Pitts J.P., Faircloth B.C., Ward P.S., Buffington M.L., Gates M.W., Kula R.R., Brady S.G. 2017. Phylogenomic insights into the evolution of stinging wasps and the origins of ants and bees. Curr. Biol. 27:1019-1025.

Branstetter M.G., Longino J.T., Ward P.S., Faircloth B.C. 2017. Enriching the ant tree of life: enhanced UCE bait set for genome-scale phylogenetics of ants and other Hymenoptera. Meth. Ecol. Evol. 8:768-776.

Breinholt J.W., Kawahara, A.Y. 2013. Phylotranscriptomics: saturated third codon positions radically in uence the estimation of trees based on next-gen data. Genome Biol. Evol. 5:2082–2092.

Breinholt J.W., Earl C., Lemmon A.R., Lemmon E.M., Xiao L., Kawahara A.Y. 2018. Resolving relationships among the megadiverse butterflies and moths with a novel pipeline for anchored phylogenomics. Syst. Biol. 67:78–93.

Brown J.W., Smith S.A. 2018. The past sure is tense: on interpreting phylogenetic divergence time estimates. Syst. Biol. 67:340-353.

Capella-Gutiérrez S., Silla-Martínez J.M., Gabaldón T. 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics 25:1972-1973.

Carpenter F.M. 1992. Treatise on invertebrate paleontology, Part R, Arthropoda 3–4. Boulder (CO): Geological Society of America.

Caterino M.S., Reed R.D., Kuo M.M., Sperling F.A.H. 2001. A partitioned likelihood analysis of swallowtail butterfly phylogeny (Lepidoptera: Papilionidae). Syst. Biol. 50:106–127.

Carter J.M., Baker S.C., Pink R., Carter D.R., Collins A., Tomlin J., Gibbs M., Breuker C.J. 2013. Unscrambling butterfly oogenesis. BMC Genom. 14:283.

Chazot N., Wahlberg N., Freitas A.V.L., Mitter C., Labandeira C.C., Sohn J.-C., Sahoo R.K., Seraphim N., de Jong R., Heikkila M. 2019. Priors and posteriors in Bayesian timing of divergence analyses: The age of butterflies revisited. Syst. Biol. doi.org/10.1093/sysbio/syz002

Chiari Y., Cahais V., Galtier N., Delsuc F. 2012. Phylogenomic analyses support the position of turtles as the sister group of birds and crocodiles (Archosauria). BMC Biol. 10:65.

Collins N.M., Morris M.G. 1985. Threatened swallowtail butterflies of the world. The Cambridge: IUCN Red Data Book.

Collins R.A., Hrbek, T. 2018. An *in silico* comparison of protocols for dated phylogenomics. Syst. Biol. 67:633-650.

Condamine F.L., Sperling F.A.H., Wahlberg N., Rasplus J.-Y., Kergoat G.J. 2012. What caused the latitudinal gradient of species diversity in swallowtail butterflies? Ecol. Lett. 15:267–277.

Condamine F.L., Sperling F.A.H., Kergoat G.J. 2013. Global biogeographical pattern of swallowtail diversification demonstrates alternative colonization routes in the Northern and Southern hemispheres. J. Biogeogr. 40:9-23.

Condamine F.L., Rolland J., Höhna S., Sperling F.A.H., Sanmartín I. 2018a. Testing the role of the Red Queen and Court Jester as drivers of the macroevolution of Apollo butterflies. Syst. Biol. 67:940-964.

Condamine F.L., Nabholz B., Clamens A.-L., Dupuis J.R., Sperling F.A.H. 2018b. Mitochondrial phylogenomics, the origin of swallowtail butterflies, and the impact of the number of clocks in Bayesian molecular dating. Syst. Entomol. 43:460-480.

Cong Q., Borek D., Otwinowski Z., Grishin N.V. 2015a. Tiger swallowtail genome reveals mechanisms for speciation and caterpillar chemical defense. Cell Rep. 10:910-919.

Cong Q., Borek D., Otwinowski Z., Grishin N.V. 2015b. Skipper genome sheds light on unique phenotypic traits and phylogeny. BMC Genomics 16:639.

Cong Q., Shen J., Warren A.D., Borek D., Otwinowski Z., Grishin N.V. 2016a. Speciation in cloudless sulphurs gleaned from complete genomes. Genome Biol. Evol. 8:915-931.

Cong Q., Shen J., Borek D., Robbins R.K., Otwinowski Z., Grishin N.V. 2016b. Complete genomes of Hairstreak butterflies, their speciation, and nucleo-mitochondrial incongruence. Sci. Rep. 6:24863.

Cong Q., Shen J., Li W., Borek D., Otwinowski Z., Grishin N.V. 2017. The first complete genomes of metalmarks and the classification of butterfly families. Genomics 109:485-493.

Davey J.W., Chouteau M., Barker S.L., Maroja L., Baxter S.W., Simpson F., Merrill R.M., Joron M., Mallet J., Dasmahapatra K.K., Jiggins C.D. 2016. Major improvements to the *Heliconius melpomene* genome assembly used to confirm 10 chromosome fusion events in 6 million years of butterfly evolution. Genes Genomes Genet. 6:695-708.

Di Franco A., Poujol R., Baurain D., Philippe H. 2019. Evaluating the usefulness of alignment filtering methods to reduce the impact of errors on evolutionary inferences. BMC Evol. Biol. 19:21.

dos Reis M., Inoue J., Hasegawa M., Asher R.J., Donoghue P.C., Yang Z. 2012. Phylogenomic datasets provide both precision and accuracy in estimating the timescale of placental mammal phylogeny. Proc. R. Soc. B 279:3491-3500.

dos Reis M., Thawornwattana Y., Angelis K., Telford M.J., Donoghue P.C., Yang Z. 2015. Uncertainty in the timing of origin of animals and the limits of precision in molecular timescales. Curr. Biol. 25:2939-2950.

dos Reis M., Donoghue P.C.J., Yang Z. 2016. Bayesian molecular clock dating of species divergences in the genomics era. Nat. Rev. Genet. 17:71–80.

Drummond A.J., Ho S.Y.W., Phillips M.J., Rambaut A. 2006. Relaxed phylogenetics and dating with confidence. PLoS Biol. 4:e88.

Dupuis J.R., Sperling F.A.H. 2015. Repeated reticulate evolution in North American *Papilio machaon* group swallowtail butterflies. PLoS One 10:e0141882.

Dupuis J.R., Sperling F.A.H. 2016. Hybrid dynamics in a species group of swallowtail butterflies. J. Evol. Biol. 29:1932-1951.

Durden C.J., Rose H. 1978. Butterflies from the middle Eocene: the earliest occurrence of fossil Papilionidae. Prarce-Sellards Ser. Tax. Mem. Mus. 29:1–25.

Ehrlich P.R., Raven P.H. 1964. Butterflies and plants: a study in coevolution. Evolution 18:586-608.

Emms D.M., Kelly S. 2015. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. Genome Biol. 16:157.

Espeland M., Hall J.P., DeVries P.J., Lees D.C., Cornwall M., Hsu Y.F., Wu L.W., Campbell D.L., Talavera G., Vila R., Salzman S., Ruehr S., Lohman D.J., Pierce N.E. 2015. Ancient Neotropical origin and recent recolonisation: Phylogeny, biogeography and diversification of the Riodinidae (Lepidoptera: Papilionoidea). Mol. Phylogenet. Evol. 93:296-306.

Espeland M., Breinholt J., Willmott K.R., Warren A.D., Vila R., Toussaint E.F.A., Maunsell S.C., Aduse-Poku K., Talavera G., Eastwood R., Jarzyna M.A., Guralnick R., Lohman D.J., Pierce N.E., Kawahara A.Y. 2018. A comprehensive and dated phylogenomic analysis of butterflies. Curr. Biol. 28:770-778.

Faircloth B.C., McCormack J.E., Crawford N.G., Harvey M.G., Brumfield R.T., Glenn T.C. 2012. Ultraconserved elements anchor thousands of genetic markers spanning multiple evolutionary timescales. Syst. Biol. 61:717-726.

Faircloth B.C., Branstetter M.G., White N.D., Brady S.G. 2015. Target enrichment of ultraconserved elements from arthropods provides a genomic perspective on relationships among Hymenoptera. Mol. Ecol. Res. 15:489-501.

Faircloth B.C. 2017. Identifying conserved genomic elements and designing universal bait sets to enrich them. Meth. Ecol. Evol. 8:1103-1112.

Ford E.B. 1944. Studies on the chemistry of pigments in the Lepidoptera, with reference to their bearing on systematics. 4. The classification of the Papilionidae. Trans. R. Entomol. Soc. L. 94:201-223.

Foster C.S., Sauquet H., Van der Merwe M., McPherson H., Rossetto M., Ho S.Y. 2017. Evaluating the impact of genomic data and priors on Bayesian estimates of the angiosperm evolutionary timescale. Syst. Biol. 66:338–351.

Fuentes-Pardo A.P., Ruzzante D.E. 2017. Whole-genome sequencing approaches for conservation biology: Advantages, limitations and practical recommendations. Mol. Ecol. 26:5369-5406.

Gardner R.C., Howarth A.J., Hahn P., Brown-Luedi M., Shepherd R.J., Messing J. 1981. The complete nucleotide sequence of an infectious clone of cauliflower mosaic virus by M13mp7 shotgun sequencing. Nucleic Acids Res. 9:2871–2888.

Garrison N.L., Rodriguez J., Agnarsson I., Coddington J.A., Griswold C.E., Hamilton C.A., Hedin M., Kocot K.M., Ledford J.M., Bond J.E. 2016. Spider phylogenomics: untangling the Spider Tree of Life. PeerJ 4:e1719.

Gernhard T. 2008. The conditioned reconstructed process. J. Theoret. Biol. 253:769–778.

Gillung J.P., Winterton S.L., Bayless K.M., Khouri Z., Borowiec M.L., Yeates D., Kimsey L.S., Misof B., Shin S., Zhou X., Mayer C., Petersen M., Wiegmann B.M. 2018. Anchored phylogenomics unravels the evolution of spider flies (Diptera, Acroceridae) and reveals discordance between nucleotides and amino acids. Mol. Phylogenet. Evol. 128:233-245.

Gnerre S., Maccallum I., Przybylski D., Ribeiro F.J., Burton J.N., Walker B.J., Sharpe T., Hall G., Shea TP., Sykes S., Berlin A.M., Aird D., Costello M., Daza R., Williams L.,

Nicol R., Gnirke A., Nusbaum C., Lander E.S., Jaffe D.B. 2011. High-quality draft assemblies of mammalian genomes from massively parallel sequence data. Proc. Natl Acad. Sci. USA 108:1513-1518.

Guéguen L., Gaillard S., Boussau B., Gouy M., Groussin M., Rochette N.C., Bigot T., Fournier D., Pouyet F., Cahais V., Bernard A., Scornavacca C., Nabholz B., Haudry A., Dachary L., Galtier N., Belkir K., Dutheil J.Y. 2013. Bio++: efficient extensible libraries and tools for computational molecular evolution. Mol. Biol. Evol. 30:1745-1750.

Guschanski K., Krause J., Sawyer S., Valente L.M., Bailey S., Finstermeier K., Sabin R., Gilissen E., Sonet G., Nagy Z.T., Lenglet G., Mayer F., Savolainen V. 2013. Next-generation museomics disentangles one of the largest primate radiations. Syst. Biol. 62: 539–554.

Hancock D.L. 1983. Classification of the Papilionidae (Lepidoptera): a phylogenetic approach. Smithersia 2:1–48.

Harkins K.M., Schwartz R.S., Cartwright R.A., Stone A.C. 2016. Phylogenomic reconstruction supports supercontinent origins for *Leishmania*. Infect. Genet. Evol. 38:101-109.

Heikkilä M., Kaila L., Mutanen M., Peña C., Wahlberg N. 2012. Cretaceous origin and repeated tertiary diversification of the redefined butterflies. Proc. R. Soc. B 279:1093–1099.

Hoang D.T., Chernomor O., von Haeseler A., Minh B.Q., Vinh L.S. 2018. UFBoot2: improving the ultrafast bootstrap approximation. Mol. Biol. Evol. 35:518-522.

Hughes G.M., Teeling E.C. 2018. AGILE: an assembled genome mining pipeline. Bioinformatics 35:1252-1254.

Igarashi S. 1984. The classification of the Papilionidae mainly based on the morphology of their immature stages. Trans. Lepido. Soc. Japan 34:41–96.

Jarvis E.D., Mirarab S., Aberer A.J., Li B., Houde P., Li C., Ho S.Y., Faircloth B.C., Nabholz B., Howard J.T., Suh A., Weber C.C., da Fonseca R.R., Li J., Zhang F., Li H., Zhou L., Narula N., Liu L., Ganapathy G., Boussau B., Bayzid M.S., Zavidovych V., Subramanian S., Gabaldón T., Capella-Gutiérrez S., Huerta-Cepas J., Rekepalli B., Munch K., Schierup M., Lindow B., Warren W.C., Ray D., Green R.E., Bruford M.W., Zhan X., Dixon A., Li S., Li N., Huang Y., Derryberry E.P., Bertelsen M.F., Sheldon F.H., Brumfield R.T., Mello C.V., Lovell P.V., Wirthlin M., Schneider M.P., Prosdocimi F., Samaniego J.A., Vargas Velazquez A.M., Alfaro-Núñez A., Campos P.F., Petersen B., Sicheritz-Ponten T., Pas A., Bailey T., Scofield P., Bunce M., Lambert D.M., Zhou Q., Perelman P., Driskell A.C., Shapiro B., Xiong Z., Zeng Y., Liu S., Li Z., Liu B., Wu K., Xiao J., Yinqi X., Zheng Q., Zhang Y., Yang H., Wang J., Smeds L., Rheindt F.E., Braun M., Fjeldsa J., Orlando L., Barker F.K., Jønsson K.A., Johnson W., Koepfli K.P., O'Brien S., Haussler D., Ryder O.A., Rahbek C., Willerslev E., Graves G.R., Glenn T.C., McCormack J., Burt D., Ellegren H., Alström P., Edwards S.V., Stamatakis A., Mindell D.P., Cracraft J., Braun E.L., Warnow T., Jun W., Gilbert M.T., Zhang G. 2014. Whole-genome analyses resolve early branches in the tree of life of modern birds. Science 346:1320–1331.

Jeffroy O., Brinkmann H., Delsuc F., Philippe H. 2006. Phylogenomics: the beginning of incongruence? Trends Genet. 22:225-231.

Jetz W., Thomas G.H., Joy J.B., Hartmann K., Mooers A.O. 2012. The global diversity of birds in space and time. Nature 491:444-448.

Jia F., Lo N., Ho S.Y. 2014. The impact of modelling rate heterogeneity among sites on phylogenetic estimates of intraspecific evolutionary rates and timescales. PLoS One 9:e95722.

de Jong R. 2003. Are there butterflies with Gondwanan ancestry in the Australian region? Invert. Syst. 17:143–156.

de Jong R. 2007. Estimating time and space in the evolution of the Lepidoptera. Tijdschrift voor Entomologie, 150:319–346.

de Jong R. 2016. Reconstructing a 55-million-year-old butterfly (Lepidoptera: Hesperiidae). European J. Entomol. 113:423-428.

de Jong R. 2017. Fossil butterflies, calibration points and the molecular clock (Lepidoptera: Papilionoidea). Zootaxa 4270:1–63.

Kajitani R., Toshimoto K., Noguchi H., Toyoda A., Ogura Y., Okuno M., Yabana M., Harada M., Nagayasu E., Maruyama H., Kohara Y., Fujiyama A., Hayashi T., Itoh T. 2014. Efficient de novo assembly of highly heterozygous genomes from whole-genome shotgun short reads. Genome Res. 24:1384-1395.

Kalyaanamoorthy S., Minh B.Q., Wong T.K., von Haeseler A., Jermiin L.S. 2017. ModelFinder: fast model selection for accurate phylogenetic estimates. Nat. Meth. 14:587-589.

Kawahara A.Y. 2009. Phylogeny of snout butter ies (Lepidoptera: Nymphalidae: Libytheinae): combining evidence from morphology of extant, fossil, and recently extinct taxa. Cladistics 25:263–278.

Kawahara A.Y., Breinholt J.W. 2014. Phylogenomics provides strong evidence for relationships of butterflies and moths. Proc. R. Soc. B 281:20140970.

Kunte K. 2009. The diversity and evolution of Batesian mimicry in *Papilio* swallowtails butterflies. Evolution 63:2707–2716.

Kunte K., Zhang W., Tenger-Trolander A., Palmer D.H., Martin A., Reed R.D., Mullen S.P., Kronforst M.R. 2014. Doublesex is a mimicry supergene. Nature 507:229–232.

Lartillot N., Philippe H. 2004. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. Mol. Biol. Evol. 21:1095–1109.

Lartillot N., Lepage T., Blanquart S. 2009. PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating. Bioinformatics 25:2286-2288.

Lartillot N., Rodrigue N., Stubbs D., Richer J. 2013. PhyloBayes MPI: phylogenetic reconstruction with infinite mixtures of profiles in a parallel environment. Syst. Biol. 62:611–615.

Laurin-Lemay S., Brinkmann H., Philippe H. 2012. Origin of land plants revisited in the light of sequence contamination and missing data. Curr. Biol. 22:593-594.

Le S.Q., Gascuel O. 2008. An improved general amino acid replacement matrix. Mol. Biol. Evol. 25:1307-1320.

Le S.Q., Gascuel O., Lartillot N. 2008. Empirical profile mixture models for phylogenetic reconstruction. Bioinformatics 24:2317-2323.

Le S.Q., Dang C.C., Gascuel O. 2012. Modeling protein evolution with several amino acid replacement matrices depending on site rates. Mol. Biol. Evol. 29:2921-2936.

Lemmon E.M., Lemmon A.R. 2013. High-throughput genomic data in systematics and phylogenetics. Annu. Rev. Ecol. Evol. Syst. 44:99-121.

Lemmon A.R., Emme S.A., Lemmon E.M. 2012. Anchored hybrid enrichment for massively high-throughput phylogenomics. Syst. Biol. 61:727-744.

Li X., Fan D., Zhang W., Liu G., Zhang L., Zhao L., Fang X., Chen L., Dong Y., Chen Y., Ding Y., Zhao R., Feng M., Zhu Y., Feng Y., Jiang X., Zhu D., Xiang H., Feng X., Li S., Wang J., Zhang G., Kronforst M.R., Wang W. 2015. Outbred genome sequencing and CRISPR/Cas9 gene editing in butterflies. Nat. Commun. 6:8212.

Li G., Davis B.W., Eizirik E., Murphy W.J. 2016. Phylogenomic evidence for ancient hybridization in the genomes of living cats (Felidae). Genome Res. 26:1-11.

Luo R., Liu B., Xie Y., Li Z., Huang W., Yuan J., He G., Chen Y., Pan Q., Liu Y., Tang J., Wu G., Zhang H., Shi Y., Liu Y., Yu C., Wang B., Lu Y., Han C., Cheung D.W., Yiu S.M., Peng S., Xiaoqian Z., Liu G., Liao X., Li Y., Yang H., Wang J., Lam T.W., Wang J. 2012. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. GigaScience 1:18.

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

McCormack J.E., Hird S.M., Zellmer A.J., Carstens B.C., Brumfield R.T. 2013. Applications of next-generation sequencing to phylogeography and phylogenetics. Mol. Phylogenet. Evol. 66:526-538.

Metzker M.L. 2010. Sequencing technologies—the next generation. Nat. Rev. Genet. 11:31-46.

Michel F., Rebourg C., Cosson E., Descimon H. 2008. Molecular phylogeny of Parnassiinae butterflies (Lepidoptera: Papilionidae) based on the sequences of four mitochondrial DNA segments. Ann. Soc. Entomol. Fr. 44:1-36.

Miller J.S. 1987. Phylogenetic studies in the Papilioninae (Lepidoptera: Papilionidae). Bull. Am. Mus. Nat. Hist. 186:365–512.

Minh B.Q., Hahn M., Lanfear R. 2018. New methods to calculate concordance factors for phylogenomic datasets. bioRxiv, 487801.

Minh B.Q., Nguyen M.A.T., von Haeseler A. 2013. Ultrafast approximation for phylogenetic bootstrap. Mol. Biol. Evol. 30:1188-1195.

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

Misof B., Liu S., Meusemann K., Peters R.S., Donath A., Mayer C., Frandsen P.B., Ware J., Flouri T., Beutel R.G., Niehuis O., Petersen M., Izquierdo-Carrasco F., Wappler T., Rust J., Aberer A.J., Aspöck U., Aspöck H., Bartel D., Blanke A., Berger S., Böhm A., Buckley T.R., Calcott B., Chen J., Friedrich F., Fukui M., Fujita M., Greve C., Grobe P., Gu S., Huang Y., Jermiin L.S., Kawahara A.Y., Krogmann L., Kubiak M., Lanfear R., Letsch H., Li Y., Li Z., Li J., Lu H., Machida R., Mashimo Y., Kapli P., McKenna D.D., Meng G., Nakagaki Y., Navarrete-Heredia J.L., Ott M., Ou Y., Pass G., Podsiadlowski L., Pohl H., von Reumont B.M., Schütte K., Sekiya K., Shimizu S., Slipinski A., Stamatakis A., Song W., Su X., Szucsich N.U., Tan M., Tan X., Tang M., Tang J., Timelthaler G., Tomizuka S., Trautwein M., Tong X., Uchifune T., Walzl MG., Wiegmann B.M., Wilbrandt J., Wipfler B., Wong T.K., Wu Q., Wu G., Xie Y., Yang S., Yang Q., Yeates D.K., Yoshizawa K., Zhang Q., Zhang R., Zhang W., Zhang Y., Zhao J., Zhou C., Zhou L., Ziesmann T., Zou S., Li Y., Xu X., Zhang Y., Yang H., Wang J., Wang J., Kjer K.M., Zhou X. 2014. Phylogenomics resolves the timing and pattern of insect evolution. Science 346:763-767.

Mita K., Kasahara M., Sasaki S., Nagayasu Y., Yamada T., Kanamori H., Namiki N., Kitagawa M., Yamashita H., Yasukochi Y., Kadono-Okuda K., Yamamoto K., Ajimura M., Ravikumar G., Shimomura M., Nagamura Y., Shin-I T., Abe H., Shimada T., Morishita S., Sasaki T. 2004. The genome sequence of silkworm, *Bombyx mori*. DNA Res. 11:27-35.

Mortazavi A., Williams B.A., McCue K., Schaeffer L., Wold B. 2008. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat. Methods 5:621-628.

Munroe E. 1961. The classification of the Papilionidae (Lepidoptera). Canad. Entomologist: Suppl. 17:1–51.

Mutanen M., Wahlberg N., Kaila L. 2010. Comprehensive gene and taxon coverage elucidates radiation patterns in moths and butterflies. Proc. R. Soc. B 277:2839–2848.

Nabhan A.R., Sarkar I.N. 2012. The impact of taxon sampling on phylogenetic inference: a review of two decades of controversy. Brief. Bioinfo. 13:122-134.

Nazari V., Zakharov E.V., 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.

Nguyen L.T., Schmidt H.A., von Haeseler A., Minh B.Q. 2015. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol. Biol. Evol. 32:268-274.

van Nieukerken E.J., Kaila L., Kitching I.J. et al. 2011. Order Lepidoptera Linnaeus 1758. Animal biodiversity: An outline of higher-level classification and survey of taxonomic richness (ed. by Z.Q. Zhang). Zootaxa 3148:212–221.

Nishikawa H., Iijima T., Kajitani R., Yamaguchi J., Ando T., Suzuki Y., Sugano S., Fujiyama A., Kosugi S., Hirakawa H., Tabata S., Ozaki K., Morimoto H., Ihara K., Obara M., Hori H., Itoh T., Fujiwara H. 2015. A genetic mechanism for female-limited Batesian mimicry in *Papilio* butterfly. Nature Genet. 47:405-409.

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., Aslam M.L., Kumar S., Challis R.J., Monteiro A., Brakefield P.M., Blaxter M. 2017. A high-coverage draft genome of the mycalesine butterfly *Bicyclus anynana*. GigaScience 6:1-7.

Oakley T.H., Wolfe J.M., Lindgren A.R., Zaharoff A.K. 2012. Phylotranscriptomics to bring the understudied into the fold: Monophyletic Ostracoda, fossil placement and Pancrustacean phylogeny. Mol. Biol. Evol. 30:215–233.

Parsons M.J. 1996. Gondwanan evolution of the troidine swallowtails (Lepidoptera: Papilionidae): cladistic reappraisals using mainly immature stage characters, with focus on the birdwings *Ornithoptera* Boisduval. Bull. Kitakyushu Mus. Nat. Hist. 15:43–118.

de la Paz Celorio-Mancera M., Wheat C. W., Vogel H., Söderlind L., Janz N., Nylin S. 2013. Mechanisms of macroevolution: polyphagous plasticity in butterfly larvae revealed by RNA-Seq. Mol. Ecol. 22:4884-4895.

Philippe H., Vienne D.M.D., Ranwez V., Roure B., Baurain D., Delsuc F. 2017. Pitfalls in supermatrix phylogenomics. European J. Taxon. 283:1-25.

Pouchon C., Fernández A., Nassar J.M., Boyer F., Aubert S., Lavergne S., Mavárez J. 2018. Phylogenomic analysis of the explosive adaptive radiation of the *Espeletia* complex (Asteraceae) in the tropical Andes. Syst. Biol. 67:1041-1060.

Prum R.O., Berv J.S., Dornburg A., Field D.J., Townsend J.P., Lemmon E.M., Lemmon A.R. 2015. A comprehensive phylogeny of birds (Aves) using targeted next-generation DNA sequencing. Nature 526:569–573.

Rainford J.L., Hofreiter M., Nicholson D.B., Mayhew P.J. 2014. Phylogenetic distribution of extant richness suggests metamorphosis is a key innovation driving diversification in insects. PLoS One 9:e109085.

Ranwez V., Criscuolo A. Douzery E.J.P. 2010. SuperTriplets: a triplet-based supertree approach to phylogenomics. Bioinformatics 26:i115-i123.

Rebel H. 1898. Fossile Lepidopteren aus der Miocänformation von Gabbro. Sitzungsberichte der Kaiserlichen Akademie der Wissenschaften. Mathematisch-Naturwissenschaftliche Classe 107:731–745.

Regier J.C., Zwick A., Cummings M.P., Kawahara A.Y., Cho S., Weller S., Roe A., Baixeras J., Brown J.W., Parr C., Davis D.R., Epstein M., Hallwachs W., Hausmann A., Janzen D.H., Kitching I.J., Solis M.A., Yen S.H., Bazinet A.L., Mitter C. 2009. Toward reconstructing the evolution of advanced moths and butterflies (Lepidoptera: Ditrysia): an initial molecular study. BMC Evol. Biol. 9:280.

Roch S., Steel M. 2015. Likelihood-based tree reconstruction on a concatenation of aligned sequence data sets can be statistically inconsistent. Theoret. Pop. Biol. 100:56-62.

Roure B., Baurain D., Philippe H. 2013. Impact of missing data on phylogenies inferred from empirical phylogenomic data sets. Mol. Biol. Evol. 30:197-214.

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

Schwartz R.S., Harkins K.M., Stone A.C., Cartwright R.A. 2015. A composite genome approach to identify phylogenetically informative data from next-generation sequencing. BMC Bioinformatics 16:193.

Scriber J.M., Tsubaki Y., Lederhouse R.C. 1995. Swallowtail butterflies: their ecology and evolutionary biology. Gainesville (FL): Scientific Publishers.

Scudder S.H. 1875. Fossil butterflies. Mem. Am. Assoc. Advanc. Sci. 1:1–99.

Simion P., Belkhir K., François C., Veyssier J., Rink J.C., Manuel M., Philippe H., Telford M.J. 2018. A software tool 'CroCo' detects pervasive cross-species contamination in next generation sequencing data. BMC Biol. 16:28.

Simonsen T.J., Zakharov E.V., Djernaes M., Cotton A.M., Vane-Wright R.I., 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.

Smith M.E., Singer B., Carroll A. 2003. $^{40}Ar/^{39}Ar$ geochronology of the Eocene Green River Formation, Wyoming. Geol. Soc. Am. Bull. 115:549–565.

Smith S.A., Brown J.W., Walker J.F. 2018. So many genes, so little time: A practical approach to divergence-time estimation in the genomic era. PLoS One 13:e0197433.

Sohn J.-C., Labandeira C.C., Davis D., Mitter C. 2012. An annotated catalog of fossil and subfossil Lepidoptera (Insecta: Holometabola) of the world. Zootaxa 3286:1–132.

Srivastava A., Sarkar H., Gupta N., Patro R. 2016. RapMap: a rapid, sensitive and accurate tool for mapping RNA-seq reads to transcriptomes. Bioinformatics 32:i192-i200.

Staden R. 1979. A strategy of DNA sequencing employing computer programs. Nucleic Acids Res. 6:2601–2610.

Suh A. 2016. The phylogenomic forest of bird trees contains a hard polytomy at the root of Neoaves. Zool. Scripta 45:50–62.

Talla V., Suh A., Kalsoom F., Dincă V., Vila R., Friberg M., Wiklund C., 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.

Tong K.J., Duchêne S., Ho S.Y., Lo N. 2015. Comment on "Phylogenomics resolves the timing and pattern of insect evolution". Science 349:487.

Tyler H.A., Brown K.S., Wilson K. 1994. Swallowtail butterflies of the Americas: a study in biological dynamics, ecological diversity, biosystematics and conservation. Gainesville (FL): Scientific Publishers.

Wahlberg N., Leneveu J., Kodandaramaiah U., Peña C., Nylin S., Freitas A.V., Brower, A.V. 2009. Nymphalid butterflies diversify following near demise at the Cretaceous/Tertiary boundary. Proc. R. Soc. B 276:4295-4302.

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

Wallace A.R. 1865. On the phenomena of variation and geographical distribution as illustrated by the Papilionidae of the Malayan region. Trans. Linn. Soc. London 25:1–71.

Wang H.C., Minh B.Q., Susko E., Roger A.J. 2018. Modeling site heterogeneity with posterior mean site frequency profiles accelerates accurate phylogenomic estimation. Syst. Biol. 67:216-235.

Warnow T. 2017. Computational phylogenetics: An introduction to designing methods for phylogeny estimation. Cambridge University Press.

Warren A.D., Ogawa J.R., Brower A.V. 2009. Revised classification of the family Hesperiidae (Lepidoptera: Hesperioidea) based on combined molecular and morphological data. Syst. Entomol. 34:467–523.

Yagi T., Sasaki G., Takebe H. 1999. Phylogeny of Japanese papilionid butterflies inferred from nucleotide sequences of the mitochondrial ND5 gene. J. Mol. Evol. 48:42-48.

Yang Z. 1996. Among-site rate variation and its impact on phylogenetic analyses. Trends Ecol. Evol. 11:367-372.

Yang Z., Rannala B. 2006. Bayesian estimation of species divergence times under a molecular clock using multiple fossil calibrations with soft bounds. Mol. Biol. Evol. 23:212-226.

Zakharov E.V., Caterino M.S., Sperling F.A.H. 2004. Molecular phylogeny, historical biogeography, and divergence time estimates for swallowtail butterflies of the genus *Papilio* (Lepidoptera: Papilionidae). Syst. Biol. 53:193–215.

Zhan S., Merlin C., Boore J.L., Reppert S.M. 2011. The monarch butterfly genome yields insights into long-distance migration. Cell 147:1171-1185.

Zhang C., Rabiee M., Sayyari E., Mirarab S. 2018. ASTRAL-III: polynomial time species tree reconstruction from partially resolved gene trees. BMC Bioinfo. 19:153.

Zhang F., Ding Y., Zhu C., Zhou X., Orr M.C., Scheu S., Luan Y.X. 2019. Phylogenomics from low-coverage whole-genome sequencing. Meth. Ecol. Evol. 10 :507-517.

Zimin A.V., Marçais G., Puiu D., Roberts M., Salzberg S.L., Yorke J.A. 2013. The MaSuRCA genome assembler. Bioinformatics 29:2669-2677.

Table legends

Table 1. Taxon sampling and genomic results of swallowtail butterfly specimens subjected to shotgun sequencing. Butterfly and moth outgroups are included, along with a new low-coverage genome for *Choristoneura fumiferana.* All voucher specimens are deposited at the University of Montpellier in the Institut des Sciences de l'Evolution de Montpellier or at the Sperling lab of the University of Alberta.

Table 2. Results of Bayesian dating of main nodes in butterflies. Using 760-gene data**,** four Bayesian analyses were conducted to test the impact of outgroups (59/58 spp vs 45/44 spp) or the exclusion of *Parnassius imperator* (59/45 versus 58/44 species) on node age estimates. Large 95% credibility intervals (CI) were obtained for analyses without outgroups compared to analyses with outgroups, and a large difference was found in the crown age of *Parnassius* when *Parnassius imperator* was excluded from the analysis.

Figure legends

Figure 1. Conceptualization of the shotgun sequencing pipeline used to construct and analyze the *Dataset 1* (760 genes in amino acids), the *Dataset 2* (6,621 genes in amino acids), the *Dataset 3* (760 genes in nucleotides) and the *Dataset 4* (6,407 genes in nucleotides).

Figure 2. Phylogenomic relationships of Papilionidae based on supermatrix analyses. All nodes have maximal BS, UFBS and PP support, except for two nodes with circles and support values in colored boxes, explained in the lower left corner legend. The topology reflects the results of all phylogenetic analyses, except the IQ-TREE analysis based on 6,621-gene data and a PMSF model that differs in placing *Parnassius imperator* as sister to *Parnassius orleans* (Appendix S6). Colors highlight tribes of Papilionidae.

Figure 3. Phylogenomic relationships of Papilionidae based on a) supertree analyses and b) gene and site concordance of supermatrix analyses. The supertree topology is inferred by ASTRAL and SuperTriplets with 6,621 genes and 5,367 rooted gene trees, respectively. For those analyses, nodes from source trees with bootstrap support lower than 70 were collapsed (quarter/triplet support is reported for each node). The supermatrix topology is inferred with IQ-TREE (see Fig. 2) while estimating gene and site concordance factors (reported for each node). Red squares highlight nodes with sCF lower than sDF. Colors highlight tribes of Papilionidae, with grey for other butterfly families. Images of extant butterfly species (indicated with asterisks by their taxon names) are interspersed in the tree to serve as illustrative markers for major lineages.

Figure 4. Bayesian time-calibrated phylogeny of butterflies. The dated tree was obtained with PhyloBayes analyses of *Dataset 1* (excluding *Bombyx mori* and *Choristoneura fumiferana*)

using the CAT-GTR model, a birth-death model, and an uncorrelated clock model constrained with five fossil calibrations (three Papilionidae and two within outgroups). The tree shows median ages obtained from the posterior distribution of Bayesian analyses (95% credibility intervals are reported in Table 2). Sensitivity analyses are presented in Appendix S9. Colored taxon names highlight tribes of Papilionidae and butterfly outgroups. Images of extant butterfly species (indicated with asterisks by their taxon names) are interspersed in the tree to serve as illustrative markers for major lineages.

Supplementary Information

Appendix S1. Scripts used to perform the analyses presented in this study.

Appendix S2. Phylogenomic dataset of Papilionoidea including 760 orthologous genes in amino acid format (*Dataset 1*).

Appendix S3. Phylogenomic dataset of Papilionoidea including 6,621 orthologous genes in amino acid format (*Dataset 2*).

Appendix S4. Phylogenomic dataset of Papilionoidea including 760 orthologous genes in nucleotide format (*Dataset 3*).

Appendix S5. Phylogenomic dataset of Papilionoidea including 6,621 orthologous genes in nucleotide format (*Dataset 4*).

Appendix S6. Phylogenomic trees of Papilionoidea inferred with both maximum-likelihood and Bayesian inference using 760 orthologous genes (*Dataset 1*) and 6,621 orthologous genes (*Dataset 2*).

Appendix S7. Tree files of the molecular phylogenomic analyses of Papilionoidea as inferred with IQ-TREE and PhyloBayes.

Appendix S8. Gene and site concordance and discordance factors estimated with the *Datasets 1* and *3* (760 genes) and *Dataset 2* (6,621 genes).

Appendix S9. Bayesian dated trees of Papilionoidea inferred with the 760-gene dataset and the mixture model CAT-GTR (PhyloBayes).

Appendix S10. Tree files of molecular divergence-time estimation of Papilionoidea as inferred with the PhyloBayes CAT-GTR model following four different analyses of the 760-gene dataset.

Appendix S11. Comparison of prior and posterior distributions for nodes with set fossil calibrations. Bayesian posterior distributions are not driven by the uniform prior distributions used to calibrate the five nodes with fossil calibrations.

Appendix S12. Phylogenomic tree of Papilionoidea inferred with the Bayesian mixture model using an amino-acid dataset comprised of 2,993 orthologous genes selected without the cross-contamination check.

Appendix S13. Correlations of branch lengths as inferred with the 2,993-gene (without CroCo) versus the 6,621-gene (with CroCo) datasets (a), and as inferred with the 760-gene versus the 6,621-gene datasets (both with CroCo) (b). Units are the number of substitutions per site per branch. Note the higher correlation (R^2) obtained when comparing branch lengths between the 760-gene and 6,621-gene datasets with cross-contamination excluded.

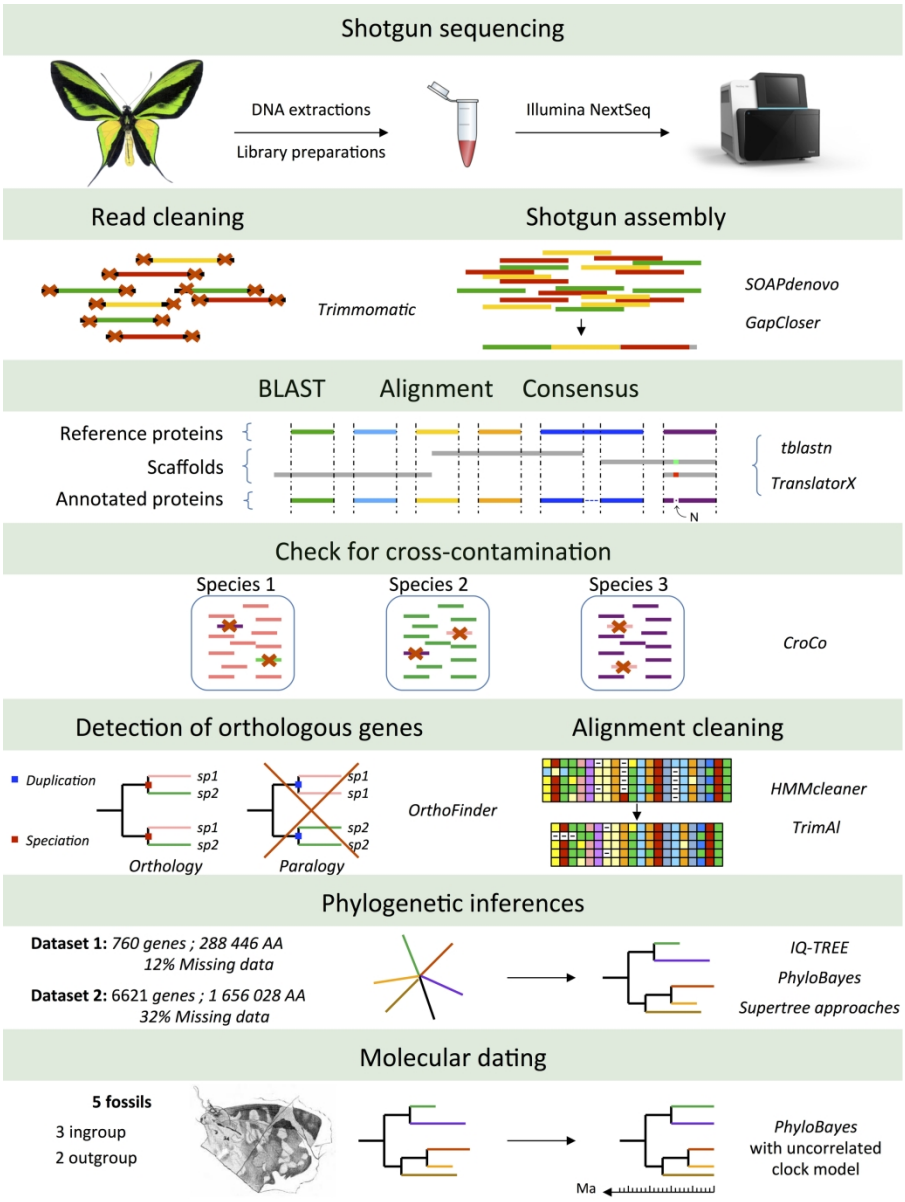

Figure 1. Conceptualization of the shotgun sequencing pipeline used to construct and analyze the Dataset 1 (760 genes in amino acids), the Dataset 2 (6,621 genes in amino acids), the Dataset 3 (760 genes in nucleotides) and the Dataset 4 (6,407 genes in nucleotides).

209x279mm (300 x 300 DPI)

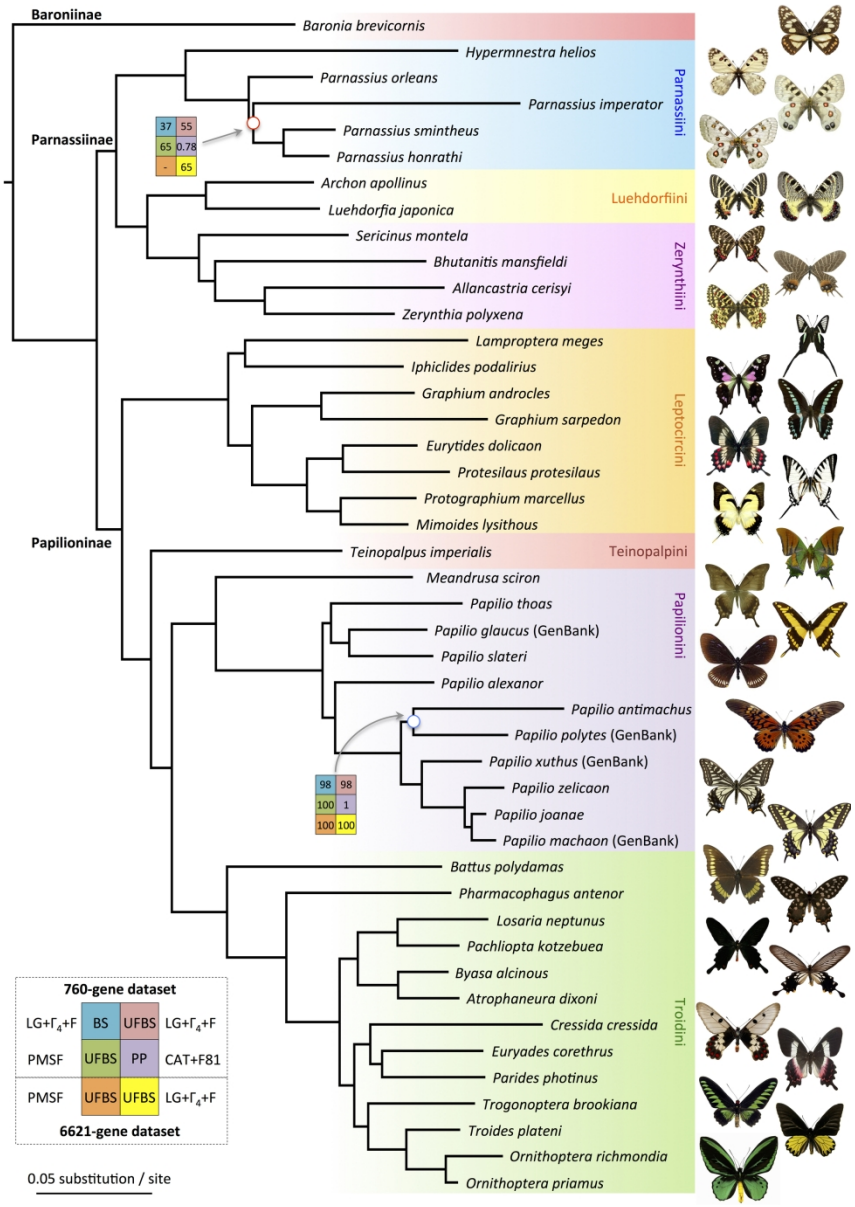

Figure 2. Phylogenomic relationships of Papilionidae based on supermatrix analyses. All nodes have maximal BS, UFBS and PP support, except for two nodes with circles and support values in colored boxes, explained in the lower left corner legend. The topology reflects the results of all phylogenetic analyses, except the IQ-TREE analysis based on 6,621-gene data and a PMSF model that differs in placing Parnassius imperator as sister to Parnassius orleans (Appendix S6). Colors highlight tribes of Papilionidae.

209x296mm (300 x 300 DPI)

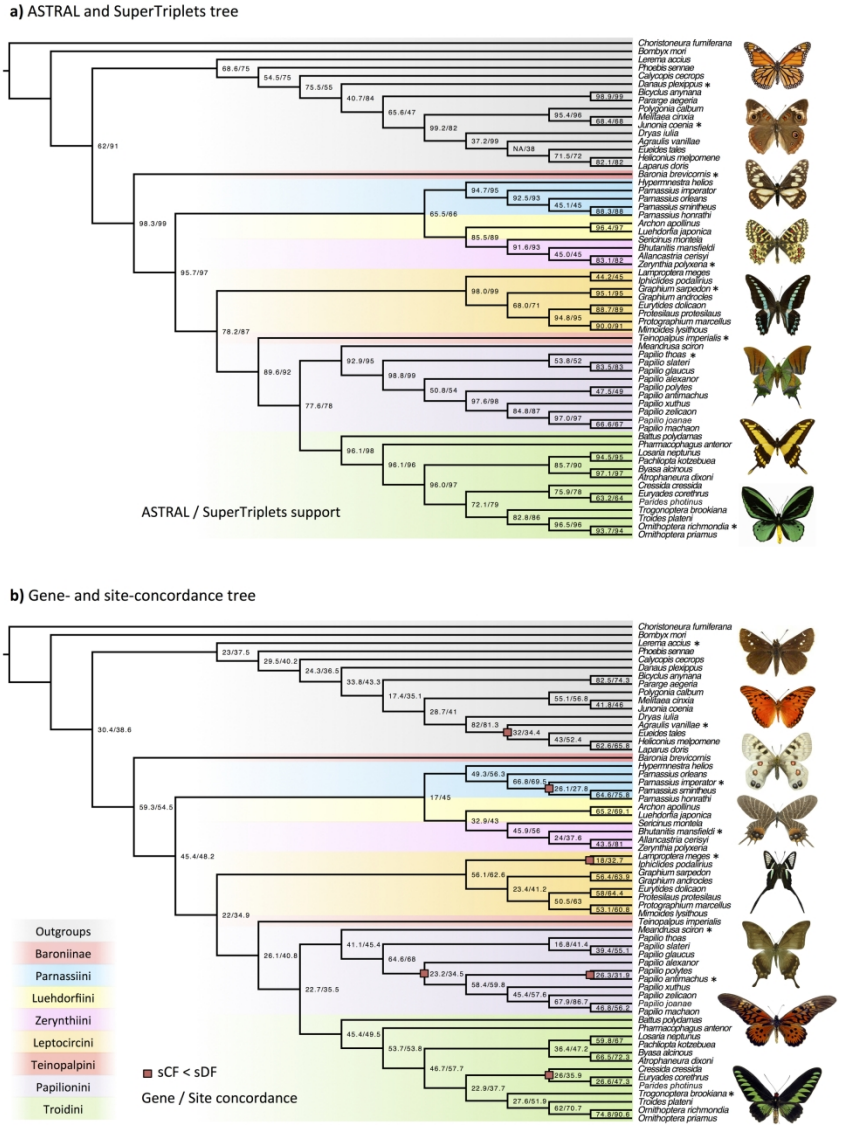

Figure 3. Phylogenomic relationships of Papilionidae based on a) supertree analyses and b) gene and site concordance of supermatrix analyses. The supertree topology is inferred by ASTRAL and SuperTriplets with 6,621 genes and 5,367 rooted gene trees, respectively. For those analyses, nodes from source trees with bootstrap support lower than 70 were collapsed (quarter/triplet support is reported for each node). The supermatrix topology is inferred with IQ-TREE (see Fig. 2) while estimating gene and site concordance factors (reported for each node). Red squares highlight nodes with sCF lower than sDF. Colors highlight tribes of Papilionidae, with grey for other butterfly families. Images of extant butterfly species (indicated with asterisks by their taxon names) are interspersed in the tree to serve as illustrative markers for major lineages.

209x296mm (300 x 300 DPI)

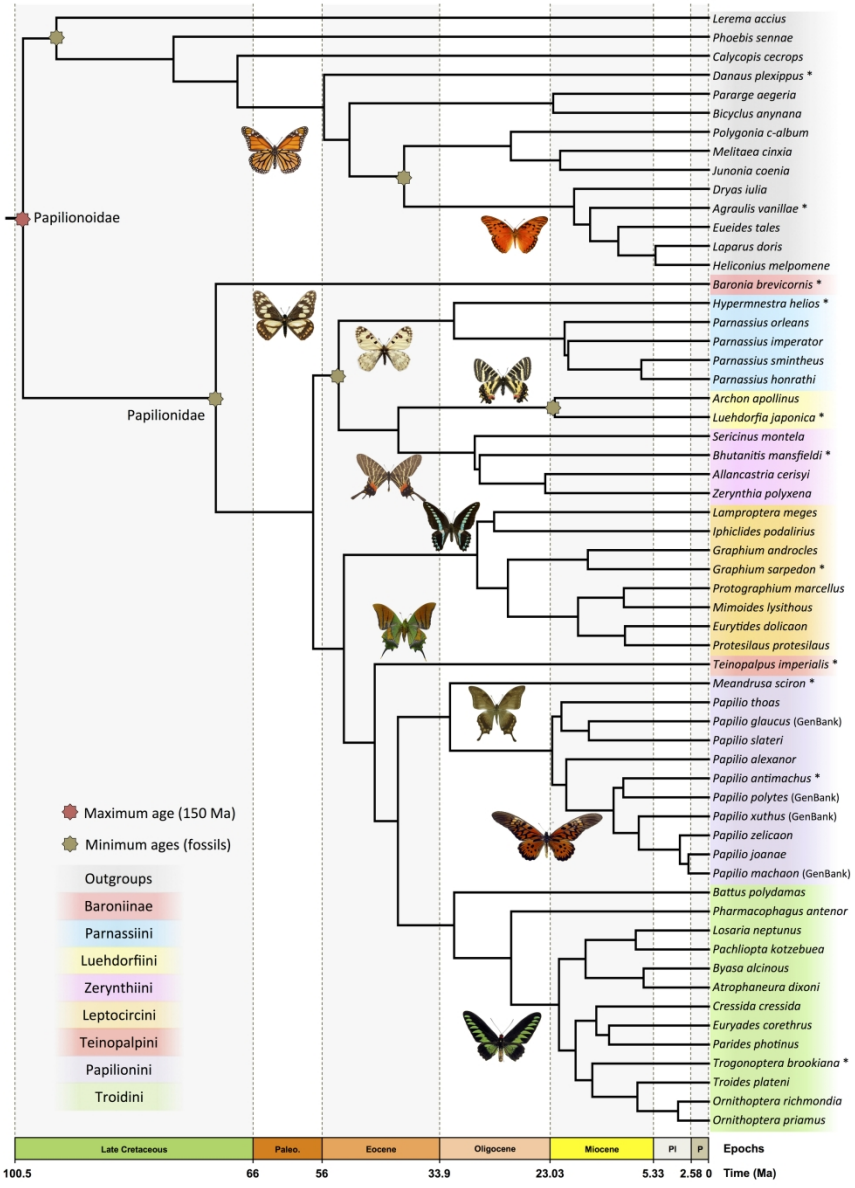

Figure 4. Bayesian time-calibrated phylogeny of butterflies. The dated tree was obtained with PhyloBayes analyses of Dataset 1 (excluding Bombyx mori and Choristoneura fumiferana) using the CAT-GTR model, a birth-death model, and an uncorrelated clock model constrained with five fossil calibrations (three Papilionidae and two within outgroups). The tree shows median ages obtained from the posterior distribution of Bayesian analyses (95% credibility intervals are reported in Table 2). Sensitivity analyses are presented in Appendix S9. Colored taxon names highlight tribes of Papilionidae and butterfly outgroups. Images of extant butterfly species (indicated with asterisks by their taxon names) are interspersed in the tree to serve as illustrative markers for major lineages.

209x296mm (300 x 300 DPI)

Table 1. Taxon sampling and genomic results of swallowtail butterfly specimens subjected to shotgun sequencing. Butterfly and moth outgroups are included. along with a new low-coverage genome for *Choristoneura fumiferana*. All voucher specimens are deposited at the University of Montpellier in the Institut des Sciences de l'Evolution de Montpellier or at the Sperling lab of the University of Alberta.

ID		Data		Genome assembly					Cross-contamination analyses			Final *Dataset 1*		Final *Dataset 2*		SRA number
Species name	Voucher	Number of raw reads (M)	Reads after cleaning (M)	Assembly size (M)	N50	Number of scaffolds	Average length	Mean %GC	Number of genes before CroCo Dataset 1	Number of genes before CroCo Dataset 2	% contamination	Number of genes	% missing data	Number of genes	% missing data	
Allancastria cerisyi	2081	73.342	60.072	146.973	150	905391	162.3	36.30	328	485	0.169	104	93.84	543	95.52	SRR8954514
Archon apollinus	2124	67.361	62.930	787.111	398	2752266	286	35.69	954	1840	0.242	748	18.29	4189	40.17	SRR8954513
Atrophaneura dixoni	2008	59.715	55.536	369.571	1616	785140	470.7	39.77	957	1913	0.327	758	5.22	4708	25.3	SRR8954516
Baronia brevicornis	167	77.182	72.038	488.028	1169	973148	501.5	35.21	959	1757	0.472	760	7.72	4344	33.39	SRR8954515
Battus polydamas	308	76.666	69.456	453.655	427	1499801	302.5	33.23	957	1902	0.010	758	12.21	4506	33.2	SRR8954510
Bhutanitis mansfieldi	2097	44.295	33.719	317.424	200	1637993	193.8	36.28	890	1526	0.011	675	44.12	2746	68.7	SRR8954509
Byasa alcinous	93	94.005	86.018	319.164	2081	443395	719.8	33.82	959	1942	0.094	760	4.72	4760	24.3	SRR8954512
Cressida cressida	1827	66.593	59.979	459.866	558	1414336	325.1	32.48	954	1876	0.231	754	10.18	4515	31.6	SRR8954511
Euryades corethrus	3716	51.927	45.178	343.287	301	1372876	250	33.15	957	1912	0.010	758	9.66	4595	30.42	SRR8954508
Eurytides dolicaon	11002	76.265	68.143	589.135	354	2185856	269.5	34.77	957	1873	9.303	720	18.22	3885	44.7	SRR8954507
Graphium androcles	1904	62.412	56.862	933.096	495	3089220	302	33.14	957	1833	2.464	751	13.4	4262	37.67	SRR8954549
Graphium sarpedon	2021	36.692	34.228	590.915	221	2775869	212.9	33.04	938	1718	0.122	748	27.6	3719	52.04	SRR8954548
Hypermnestra helios	Hyp.01-01a	17.743	10.851	164.657	150	1040790	158.2	36.27	887	1540	0.000	645	53.52	2500	75.51	SRR8954546
Hypermnestra helios	2069	33.169	29.511	292.277	161	1741883	167.8	37.74			4.278					SRR8954547
Iphiclides podalirius	6	38.605	34.373	390.909	599	1041389	375.4	35.92	959	1876	0.105	759	14.08	4378	36.37	SRR8954545
Lamproptera meges	1955	59.278	53.733	512.606	275	2153366	238	33.64	957	1824	0.650	755	16.52	4232	40.68	SRR8954544
Losaria neptunus	3706	31.248	27.061	248.117	449	785959	315.7	36.02	939	1794	0.054	751	17.01	4164	41.5	SRR8954543
Luehdorfia japonica	356	82.816	74.878	543.466	360	1930389	281.5	37.07	957	1901	0.182	757	13.31	4455	33.91	SRR8954542
Meandrusa sciron	1896	42.681	39.647	509.369	479	1699899	299.6	38.02	959	1996	0.009	759	10.48	4740	28.87	SRR8954541
Mimoides lysithous	3789	92.328	83.586	541.013	730	1377484	392.8	35.65	958	1889	0.057	759	7.13	4595	28.92	SRR8954540
Ornithoptera priamus	1919	62.251	57.565	419.015	1305	762773	549.3	36.78	958	1931	0.009	760	4.66	4770	23.97	SRR8954527
Ornithoptera richmondia	1938	70.462	60.269	480.793	315	1827473	263.1	39.20	957	1891	1.979	749	15.41	4326	38.12	SRR8954528
Pachliopta kotzebuea	1920	94.744	87.123	304.109	5273	161975	1877.5	35.38	958	1920	8.394	747	6.3	4372	30.55	SRR8954525
Papilio alexanor	Alx.01-01a	161.124	128.084	409.954	3716	565709	724.7	36.11	959	2238	0.034	759	4.77	5252	18.32	SRR8954526
Papilio antimachus	Dru.01-01a	195.146	144.268	388.176	2842	646388	600.5	40.50	959	2270	1.083	750	9.18	5146	23.63	SRR8954523
Papilio glaucus (GenBank)	3	-	-	374.816	230841	60470	6198.4	35.47	959	2224	-	760	4.6	5206	19.08	-
Papilio joanae	295	61.752	57.302	408.782	643	1010689	404.5	32.72	958	2382	0.123	760	6.43	5465	19.01	SRR8954524
Papilio machaon (GenBank)	1	-	-	278.436	1174287	63187	4406.5	33.77	959	2413	-	759	3.2	5376	14.87	-
Papilio polytes (GenBank)	2	-	-	227.021	3672263	3874	58601.2	33.77	946	2233	-	744	8.67	5098	22.16	-
Papilio slateri	1796	60.549	54.703	356.588	855	811786	439.3	36.67	959	2213	0.377	758	7.81	5094	23.68	SRR8954521
Papilio thoas	Her.01-01a	153.287	122.570	339.705	1002	748457	453.9	33.34	959	2192	2.229	744	11.24	4991	27.04	SRR8954522
Papilio xuthus (GenBank)	-	-	-	-	-	-	-	-	959	2482	-	760	1.16	5730	9.76	-
Papilio zelicaon	Pap.01-01a	199.531	164.682	399.147	764	947128	421.4	33.57	959	2351	1.982	755	13.48	5151	26.89	SRR8954529
Parides photinus	149	67.923	63.876	380.543	835	871292	436.8	33.64	959	1916	1.010	757	5.61	4684	26.28	SRR8954530
Parnassius honrathi	Parn.01-01a	178.536	150.484	994.537	955	2130202	466.9	36.56	959	1872	0.106	758	6.43	4578	28.51	SRR8954539
Parnassius imperator	Kai.01-01a	35.528	21.955	483.372	171	2712024	178.2	37.86	841	1307	26.714	394	72.94	1368	86.98	SRR8954538
Parnassius orleans	Drio.01-01a	156.174	120.544	814.564	485	2954078	275.7	35.38	955	1908	8.955	708	14.37	4145	38.93	SRR8954532
Parnassius smintheus	1706	103.125	95.149	846.707	379	3198623	264.7	35.92	957	1866	0.260	759	11.45	4417	33.94	SRR8954531
Pharmacophagus antenor	1646	120.153	107.409	387.987	4329	399998	970	40.99	958	1874	1.034	760	5.32	4612	26.79	SRR8954535
Protesilaus protesilaus	3711	79.486	72.676	901.946	234	4111085	219.4	33.97	954	1774	10.172	706	27.2	3434	55.15	SRR8954534
Protographium marcellus	7	64.339	58.553	636.959	320	2472899	257.6	35.25	955	1872	0.560	757	14.06	4408	35.48	SRR8954537
Sericinus montela	Ser.01-01a	137.724	109.471	526.009	1843	750961	700.4	35.74	959	1887	2.622	750	7.82	4535	29.83	SRR8954536
Teinopalpus imperialis	1566	64.583	58.453	530.299	848	1161826	456.4	35.12	958	1956	0.158	759	6.92	4737	26.84	SRR8954519
Trogonoptera brookiana	2023	35.422	30.401	282.281	423	897158	314.6	34.22	956	1888	10.976	699	25.74	3865	47.93	SRR8954533
Troides plateni	Tr.20-01h	168.313	134.462	449.338	2905	693452	648	42.36	959	1904	0.892	758	5.63	4664	25.82	SRR8954518
Zerynthia polyxena	2122	32.378	29.101	461.127	255	1969767	234.1	34.39	939	1751	0.091	738	30.92	3638	54.68	SRR8954517
Agraulis vanillae	NA	-	-	390.758	21413	45022	8679.3	31.80	946	1547	-	754	12.05	3702	44.52	-
Bicyclus anynana	NA	-	-	-	-	-	-	-	941	2273	-	754	9.87	5157	27.72	-
Bombyx mori	NA	-	-	-	-	-	-	-	938	2191	-	739	16.39	5044	31.44	-
Calycopis cecrops	NA	-	-	-	-	-	-	-	875	1826	-	714	17.17	4365	38.84	-
Choristoneura fumiferana	399	155.399	139.886	570.774	606	1574167	362.6	36.59	946	1548	0.000	752	15.63	3652	47.38	SRR8954520
Danaus plexippus	NA	-	-	-	-	-	-	-	952	2480	-	757	10.52	5564	22.97	-
Dryas iulia	NA	-	-	566.959	21916	53067	10683.8	34.64	950	1555	-	757	12.06	3725	44.17	-
Eueides tales	NA	-	-	539.115	32552	49376	10918.6	33.40	951	1541	-	751	13.03	3669	45.16	-
Heliconius melpomene	NA	-	-	-	-	-	-	-	957	2507	-	756	6.45	5703	18.64	-
Junonia coenia	NA	-	-	-	-	-	-	-	883	2135	-	739	12.78	5012	31.28	-
Laparus doris	NA	-	-	404.741	46843	31556	12826.1	35.41	948	1533	-	758	11.95	3690	45.02	-
Lerema accius	NA	-	-	-	-	-	-	-	905	2239	-	732	13.3	5066	30.09	-
Melitaea cinxia	NA	-	-	-	-	-	-	-	881	2169	-	712	21.81	4660	41.37	-
Pararge aegeria	NA	-	-	406.226	5576	118444	3429.7	36.24	943	1483	-	754	14.46	3449	49.35	-
Phoebis sennae	NA	-	-	-	-	-	-	-	946	2241	-	754	8.55	5124	25.99	-
Polygonia calbum	NA	-	-	345.882	2829	177472	1948.9	32.85	940	1512	-	741	19.71	3442	50.67	-
Median		67.642	60.269	409.954	643	1010689	294	35.43	957	1891	0.260	754	12.05	4515	31.6	-

Table 2. Results of Bayesian dating of main nodes in butterflies. Using 760-gene data**,** four Bayesian analyses were conducted to test the impact of outgroups (59/58 spp vs 45/44 spp) or the exclusion of *Parnassius imperator* (59/45 versus 58/44 species) on node age estimates. Large 95% credibility intervals (CI) were obtained for analyses without outgroups compared to analyses with outgroups, and a large difference was found in the crown age of *Parnassius* when *Parnassius imperator* was excluded from the analysis.

Sampling	Papilionoidea		Papilionidae		Parnassiinae		Parnassiini		*Parnassius*		Luehdorfiini		Zerynthiini	
	Median age	**95% CI**	**Median age**	**95% CI**	**Median age**	**95% CI**	**Median age**	**95% CI**	**Median age**	**95% CI**	**Median age**	**95% CI**	**Median age**	**95% CI**
59 species	99.2	68.6-142.7	71.4	49.8-103.6	53.6	36.9-79.2	36.9	22.8-57.9	21.2	12.4-35.2	22.4	9.9-39.9	34.0	21.3-52.1
58 species	100.4	70.6-142.5	72.0	50.4-103.5	53.7	37.0-77.2	35.3	21.1-55.0	14.5	8.0-25.4	22.5	11.0-37.4	34.6	21.7-52.9
45 species	NA	NA	71.9	43.7-139.8	58.1	31.7-115.6	40.2	19.9-83.4	23.3	11.2-49.6	26.0	10.6-56.3	36.8	18.3-74.8
44 species	NA	NA	71.7	43.9-139.9	56.5	31.2-111.1	37.0	17.8-76.8	15.8	6.9-34.7	24.6	10.8-52.8	36.1	18.4-73.1

Sampling	Papilioninae		Leptocircini		*Graphium*		Teinopalpini (stem)		Papilionini		*Papilio*		Troidini	
	Median age	**95% CI**	**Median age**	**95% CI**	**Median age**	**95% CI**	**Median age**	**95% CI**	**Median age**	**95% CI**	**Median age**	**95% CI**	**Median age**	**95% CI**
59 species	52.9	36.7-77.4	33.4	21.1-50.9	17.5	9.9-28.7	48.4	33.6-71.3	37.5	25.3-56.3	22.8	14.9-34.6	37.0	25.2-55.1
58 species	52.6	36.7-75.3	33.2	21.7-49.6	17.5	10.2-27.9	48.3	33.5-69.5	37.4	25.3-54.8	22.4	14.7-33.4	36.8	25.3-53.4
45 species	57.8	31.7-114.9	36.7	18.6-75.1	19.1	8.6-40.4	53.0	29.0-105.6	40.1	21.8-82.2	24.9	13.1-50.7	40.6	21.8-81.4
44 species	55.6	31.0-109.6	35.4	18.7-72.0	18.4	8.6-38.4	51.0	28.4-100.9	39.5	21.6-78.4	23.8	12.6-47.9	39.0	21.2-77.8