THÈSE POUR OBTENIR LE GRADE DE DOCTEUR DE L'UNIVERSITÉ DE MONTPELLIER

En Génétique et Génomique

École doctorale GAIA – Biodiversité, Agriculture, Alimentation, Environnement, Terre, Eau

Unité de recherche – Institut des Sciences de l'Evolution de Montpellier (ISEM) – UMR 5554

A comprehensive phylogenomic approach to study diversification, host-plant shifts and biogeography in two tribes of swallowtail butterflies (Lepidoptera: Papilionidae)

Présentée par Eliette L. REBOUD

Le 7 Décembre 2023

Sous la direction de Fabien L. Condamine et Benoit Nabholz

Devant le jury composé de

Violaine LLAURENS, Directrice de Recherche, CNRS, Paris, France	Rapportrice
Niklas WAHLBERG, Professeur, Université de Lund, Lund, Suède	Rapporteur
Thomas COUVREUR, Directeur de Recherche, IRD, Montpellier, France	Examinateur
Rodolphe ROUGERIE, Maître de Conférences, MNHN, Paris, France	Examinateur
Emmanuelle JOUSSELIN, Directrice de Recherche, INRAE, Montferrier-sur-Lez, France	Présidente du jury
Benoit NABHOLZ, Professeur, Université de Montpellier, Montpellier, France	Co-directeur de thèse
Fabien CONDAMINE, Chargé de Recherche, CNRS, Montpellier, France	Co-directeur de thèse

UNIVERSITÉ DE MONTPELLIER

À Louis Reboud, pour le début, le milieu, et la fin.

TABLE OF CONTENTS

ACKNOWLEDGMENTS – REMERCIEMENTS 8

GENERAL INTRODUCTION 12

0. CONTEXT OF PHYLOGENETIC ANALYSES - GOALS AND CHALLENGES 13
1- THE FIRST STEP: DEFINING SPECIES 17
2. RECONSTRUCTING ROBUST PHYLOGENIES: THE SECOND STEP 20
3. DATING PHYLOGENIES TO STUDY MACROEVOLUTIONARY PROCESSES 23
4. STUDY DIVERSIFICATION HYPOTHESES WITH AN ENTIRE INSECT CLADE 28
5. PRESENTATION OF GOALS AND STRUCTURE OF THIS THESIS 34
REFERENCES 38

CHAPTER 1. GENOMICS, POPULATION DIVERGENCE AND HISTORICAL DEMOGRAPHY OF THE WORLD'S LARGEST AND ENDANGERED BUTTERFLY, THE QUEEN ALEXANDRA'S BIRDWING 55

INTRODUCTION 57
RESULTS AND DISCUSSION 60
MATERIALS AND METHODS 72
ACKNOWLEDGMENTS 80
AUTHOR CONTRIBUTIONS 80
DATA AVAILABILITY 80
REFERENCES 81
SUPPLEMENTARY INFORMATION 89

CHAPTER 2.1. DECIPHERING THE SYSTEMATICS AND EVOLUTIONARY HISTORY OF THE RECALCITRANT TRIBE LEPTOCIRCINI (LEPIDOPTERA: PAPILIONIDAE) WITH WHOLE-GENOME DATA 100

INTRODUCTION 103
MATERIALS AND METHODS 109
RESULTS 117
DISCUSSION 124
ACKNOWLEDGMENTS 133
REFERENCES 133
SUPPLEMENTARY INFORMATION 140

CHAPTER 2.2. SPECIES-LEVEL GENOME-WIDE PHYLOGENY OF THE GIANT SWALLOWTAIL BUTTERFLIES OF THE TRIBE TROIDINI (LEPIDOPTERA: PAPILIONIDAE) 162

INTRODUCTION 165
MATERIALS AND METHODS 168
RESULTS AND DISCUSSION 172
ACKNOWLEDGMENTS 183
REFERENCES 184
SUPPLEMENTARY INFORMATION 189

CHAPTER 3. COMPARATIVE MACROEVOLUTIONARY HISTORY OF TWO TRIBES OF SWALLOWTAIL BUTTERFLIES: DID HOST-PLANT SHIFTS PLAY A DOMINANT ROLE IN DIVERSIFICATION? 206

ABSTRACT 208
INTRODUCTION 209
MATERIALS AND METHODS 214
RESULTS AND DISCUSSION 220
ACKNOWLEDGMENTS 236
REFERENCES 237
SUPPLEMENTARY INFORMATION 247

GENERAL DISCUSSION 260

0. FOREWORD 261
1. STUDYING WHOLE GENOME DATA 261
2. INTERPRETING AND USING PHYLOGENIES DERIVED FROM A MULTITUDE OF MARKERS 268
3. DIVERSIFICATION OF PHYTOPHAGOUS INSECTS 270
REFERENCES 278

FRENCH SUMMARY 284

CONTEXTE 285
CHAPITRE 1 – ETUDE SUR LA GÉNOMIQUE, LA DÉMOGRAPHIE HISTORIQUE ET LA DIVERGENCE DES POPULATIONS DU PLUS GRAND PAPILLON DU MONDE, L'ORNITHOPTÈRE DE LA REINE ALEXANDRA. 291
CHAPITRE 2.1 - DÉCRYPTAGE DE LA SYSTÉMATIQUE ET DE L'HISTOIRE ÉVOLUTIVE DE LA TRIBU LEPTOCIRCINI (LEPIDOPTERA : PAPILIONIDAE) À L'AIDE DE DONNÉES GÉNOMIQUES COMPLÈTES. 294
CHAPITRE 2.2 - PHYLOGÉNIE À L'ÉCHELLE DE L'ESPÈCE DES PAPILLONS GÉANTS DE LA TRIBU TROIDINI (LEPIDOPTERA : PAPILIONIDAE) 296
CHAPITRE 3 – ETUDE MACROÉVOLUTIVE COMPARÉE DE DEUX TRIBUS DE PAPILIONIDAE : LES CHANGEMENTS DE PLANTES HÔTES ONT-ILS JOUÉ UN RÔLE DOMINANT DANS LEUR DIVERSIFICATION ? 296
PERSPECTIVES 297

APPENDICES **298**

TABLE OF CONTENTS **299**

Abstract

Understanding why biodiversity is not evenly distributed on Earth across space, time, and lineages is a key question in macroevolutionary studies. In particular, the great diversity of insects raises the question of the role of biotic factors and key innovations, such as the ability to circumvent plant defenses, in the diversification of organisms. Although this is one of the main theories for the diversification of phytophagous insects, numerous climatic, morphological or contingent factors could have also played an important role in their evolutionary history. To address these questions, it is necessary to build a robust and well-sampled phylogenetic framework at the species level, which can be challenging in the case of non-model organisms. In this thesis, I address this issue by reconstructing nearly complete, robust, and dated phylogenies of two tribes of swallowtail butterflies, a charismatic family of Lepidoptera with a worldwide distribution. I revise the taxonomy of the two groups using genomic approaches for species delimitation at the population level, production of reference genomes for several key genera and over 280 whole-genomes, and in-depth phylogenetic studies. Based on this work, I infer the evolutionary history of the two tribes in a biotic (host plant) and biogeographic framework. This comparative study represents a major advance in our understanding of the evolutionary history of Papilionidae, and emphasizes the importance of well-sampled studies in accurately unraveling the factors behind their diversification.

Key words: Biogeography, Diversification, Host-plant shifts, Leptocircini, Molecular dating, Phylogenomics, Reference genomes, Troidini

Résumé

La diversité des espèces vivantes sur Terre n'est pas répartie équitablement dans l'espace, le temps, et entre lignées. L'origine de cette disparité est un questionnement phare des études macroévolutives. La grande diversité des insectes, en particulier, pose la question du rôle des facteurs biotiques et des innovations clefs tel que le contournement de défense des plantes dans la diversification des organismes. Bien que cela constitue l'une des théories principales de la diversification des insectes phytophages, de nombreux facteurs climatiques, morphologiques, ou contingents pourraient également avoir été des acteurs majeurs dans leur histoire évolutive. Pour explorer ces questions, il est nécessaire de se placer dans un cadre phylogénétique robuste et bien échantillonné, ce qui peut représenter un défi important pour des organismes non-modèles. Dans cette thèse, je traite cette question en reconstruisant les phylogénies quasi exhaustives, robustes et datées de deux tribus de la famille des Papilionidae, des papillons charismatiques à la répartition mondiale. A l'aide d'approches génomiques de délimitation d'espèces à l'échelle populationnelle, de productions de génomes complets et de génomes de référence, ainsi que des études phylogénomiques poussées, je révise la taxonomie des deux groupes, et infère leur histoire évolutive d'un point de vue biotique (plante hôte) et biogéographique. Cette étude comparative représente une avancée majeure pour la compréhension de l'histoire évolutive de la famille des Papilionidae, et souligne l'importance d'un échantillonnage exhaustif pour appréhender précisément les facteurs de leur diversification.

Mots-clefs : Biogéographie, Datation moléculaire, Diversification, Génomes de référence, Interactions plantes-insectes, Leptocircini, Phylogénomiques, Troidini

Acknowledgments – Remerciements

The fundamentals

First, I would like to thank all the members of my jury for accepting to read this thesis with such enthusiasm. I hope you really enjoy it and I look forward to discussing it with you.

I would also like to thank the European Research Council (ERC), through the GAIA project of which Fabien is the holder, for funding this thesis and for the pleasant working environment, both in terms of funding and equipment, that this has created for me.

Cette thèse n'aurait donc pas pu avoir lieu sans mon directeur de thèse Fabien Condamine que je remercie en premier lieu pour la confiance qu'il m'a accordée en me faisant signer ce fameux papier d'après-nuit-blanche de plein Covid. Ça valait le coup. Merci pour ton enthousiasme scientifique débordant, pour ta connaissance infinie de la biblio (et désolée de te l'avoir fait réviser pendant ces longues heures de mise en forme). Merci pour ta curiosité et ta générosité, aussi bien scientifique qu'humaine. Ton soutien dans ce dernier mois d'écriture a été inestimable. J'ai hâte de continuer !

Merci à Benoit Nabholz, qui a accepté de co-diriger cette thèse. Benoit, tu es une source d'inspiration scientifique à bien des égards, tant par l'apparente santé que tu sembles entretenir avec ton travail, que par l'immensité de tes connaissances naturalistes et scientifiques et le plaisir que tu y prends. Merci pour tous les échanges que nous avons eus, pour ton entrain indéfectible, pour la simplicité de ces interactions et leur positivité à toute épreuve.

Vous avez tous les deux fait de ma direction de thèse un environnement stimulant, bienveillant et scientifiquement riche, merci.

L'écosystème isémien

De manière générale, l'ISEM, et en particulier le couloir du rez-de-chaussée, a été pour moi un bon endroit pour passer ma thèse, et j'en mesure ma chance. Les diverses discussions avec des membres de l'équipe phylevolmol (Fred, Manu, Sébastien, Nicolas, Anna-Sophie, Jonathan, pour n'en citer que quelques-uns), mais également avec l'équipe paléonto (Laurent, PierrO) ont élargie ma vision et fait prendre du recul quand il y en avait besoin. Je remercie par cette même occasion les membres de mon comité de thèse pour les échanges que l'on a pu avoir et la stimulation scientifique que cela a généré à chaque fois.

Merci aux membres de l'ISEM qui soutiennent tous ces travaux de recherche, notamment le service informatique, et les membres du labo au bâtiment 21, pour leur soutien administratif et leur aide à organiser des missions. Merci Florence, en particulier, pour ton efficacité.

Merci à tous les non-permanents de l'ISEM pour l'ambiance, le soutien, et le plaisir qu'ils génèrent à venir au labo chaque jour. Les week-ends et les évènements de l'ISEM sont peuplés d'un nombre toujours aussi effarant de doctorants isémiens, et je ne les remercierais donc pas tous ici, mais je tiens à saluer particulièrement ici Alba, Alexis, Alex, Arthur,

Bérénice, Elodie, Gwen, Jean-Loup, Laura, Léa, Louise, Lucas, Manue, Marie, Marie, Noémie et Sophie.

Sophie, tu mérites une double mention ici pour avoir été une guide administrative exemplaire au cours de cette écriture et rendu de thèse ! Tes conseils ont toujours été lumineux dans cette sombre forêt administrative.

Merci à Marie-ka, Manue et Bérénice pour la production de la matière première de cette thèse (après les papillons eux-mêmes). Mon travail n'aurait pas pu exister sans vos mains magiques, vos intuitions et connaissances incroyables des drôles d'humeur de l'ADN.

Merci à la Team Pap (qui ne cesse de s'agrandir !) pour tous ces moments passés, pour toutes ces missions réalisées et pour toutes celles à venir. Manue, saches que je ne me lasserai jamais des vidéos et enregistrements audio de nos missions.

Qu'aurais-je fait sans vous ?

Merci à Jean-Yves Rasplus pour cet accueil chez lui et au CBGP, il y a maintenant de nombreuses années, qui a sans nul doute continué à alimenter mon émerveillement déjà grand pour la systématique des hyménoptères et autres insectes, pour l'arrière-pays montpellierain, la phylogénie et les aventures de terrain.

Je tiens à remercier Pierre Arnal pour les bases bioinformatiques qu'il m'a transmises de manière si efficace au début de mon stage de Master 2. Certains de mes scripts, ainsi que la capacité (rare je n'en doute pas) à utiliser mcmctree sans avoir (trop) envie de casser son clavier, proviennent de ses précieux conseils. Sur cette même note, je tiens à également remercier Méril Reboud, Nathan Mazet, Benoit Nabholz, Rémi Allio et Mathilde Barthe pour leur aide dans les divers langages informatiques que j'ai pu utiliser au cours de cette thèse.

Merci à mes co-bureaux Nathan et Mathilde pour tous ces moments vécus, qu'ils aient été de faits de joie, de fébrilité ou même simplement de routine. Merci pour votre humour et tous les quiproquos qui s'en sont ensuivis, et merci pour ces discussions quasiment aussi riches que nos gâteaux (c'est dire !).

Je souhaite remercier tous les amis qui m'ont soutenu depuis tant d'année et qui ont subi d'une manière ou d'une autre mes emplois du temps trop chargés. Merci Alaïs, d'être là, encore, et de l'avoir toujours été. Merci aux amis de la 51, le rendez-vous est pris pour les 10 ans à Marseillan. Merci à la Greu' et autres charmant.e.s coupain.e.s, j'ai hâte de redormir dans une yourte avec vous. Merci aux Dada pour ces années passées (si vite, d'ailleurs), merci pour l'entraide, les partages de graphiques très moches et autres memes scientifiques (ou pas), merci pour les vacances, les jeux de société et autres soirées ludiques… Venez, on continue ?

La famille, les familles, ma famille

Merci à la coloc (*sensu lato*). J'ai été heureuse de vivre entourée de cerveaux tout aussi étrangement branchés que le mien. Merci Mathilde et Erwan pour cette imagination naturelle dans les expressions qui n'existent pas, et merci pour votre capacité à filer des métaphores

scientifiques pour parler de choses tout à fait courantes de la vie. Merci Ju et Foué pour votre joie communicative (et dessins de qualité) lors de nos soirées de jeu.

Merci à la famille Brée pour cet accueil chaleureux auprès de vous, et merci particulièrement à votre compréhension et vos encouragements pour cet été d'écriture. La sérénité apportée par ces pauses café n'avait d'égal que la douceur du climat normand.

Enfin, merci ma famille, cette famille de savants. J'estime qu'une grande partie des valeurs scientifiques et humaines qui m'ont été transmises provient de celles de mes grands-parents, chacun à leur manière, et je les remercie pour cela. Pa pour l'humilité et la rigueur du raisonnement, Ma pour la ténacité et la force, Mano pour l'écriture et l'importance de la communication, et Palou pour la curiosité et la passion. Il ne parait alors pas si étonnant que mes oncles, cousin.e.s, parents et frères suivent à leur manière aussi ces mêmes chemins. Papa et Maman, merci pour tout. Mais évidemment ici, en particulier, merci pour votre soutien continuel et encouragements dans la voie des « petites bêtes » que j'ai choisie. Méril et Nathan, merci, vous me manquez.

J'ai hâte que les nouvelles petites dames de la famille découvrent à quel point les insectes, c'est stylé.

Mathilde, j'ai décidé que tu serais remerciée partout, et en même temps, quand on travaille dans le même bureau, et qu'on vie dans la même maison, forcément… Merci infiniment à toi pour ton soutien inébranlable, ton abnégation (oui pardon c'est un terme pompeux mais il marche vraiment bien ici), et ton émerveillement spontané aux choses de la vie (comme devant chacune des mésanges qui est passée devant notre bureau par exemple), qui ont fait de cette thèse une chouette aventure, et de cette fin de thèse, une chose…surmontable.

Bab, il serait difficile d'écrire un merci assez représentatif de tout ce que tu es pour moi, mais j'espère que tu en mesures l'ampleur. « Merci d'être » me semble en faire un assez bon résumé, cela dit. Un élément quantifiable tout de même : merci pour cette playlist de 4220 titres, 273h et 54 minutes (last accessed 09/10/2023), elle est (à quelques exceptions techno-boum près) tout à fait parfaite pour mettre en forme l'intégralité d'une thèse et reprendre chaque virgule de biblio.

Les précieux anonymes

I would like to conclude by thanking all the people, the taxonomists, fieldworkers, amateurs, all the local and indigenous communities... who continue to publish information about species and their environment, and who share their knowledge in one way or another through websites and free resources, books, articles, or by guiding us through the Amazon rainforest. It's such precise, unique, rich, and indispensable field knowledge that I'm admiring and grateful for.

General Introduction



Le vol du *Byasa* - Eliette Reboud

0. Context of phylogenetic analyses - goals and challenges

Phylogenetic trees, also known as phylogenies, are graphical representations of the evolutionary history and relationships between or within groups of organisms (Kjer et al. 2016). They have been employed to illustrate and understand the immense diversity of living beings for over two centuries now (Aubier 1801, Lamarck 1815, Eichwald 1829 following Pallas' idea 1766; **Figure 1A-B-C**). As taxonomy (science of naming living things) was a response to the necessity of identifying and categorizing the diversity in which humans are immersed, and thus attempting to gain a better comprehension of their environment (Raven et al. 1971, Manktelow 2010), it is likely that a phylogenetic tree constituted an intuitive graphical tool for arranging these different taxonomic levels at a wider scale and acquiring a comprehensive picture of the "tree of life" (but see Ragan 2009 and Figure 1D-E for other alternative representations, such as scales, networks, circles, stars, polyhedron...). Yet, the concept of genealogical relationships between the different parts of the graphical representation was previously limited to a vaguely defined notion of "affinity" (Ragan 2009). Only with the advent of Darwin and Wallace's theory of evolution, and the work of Haeckel (Levit and Hossfeld 2019), did phylogenetic representation acquire their genealogical evolutionary dimension, that we now instinctively interpret (**Figure 1F-G**; but note the genealogical dimension of Buffon network of dogs' breeds **Figure 1D**).

Phylogenies thus became models to understand the evolution of this diversity across groups of organisms, across lineages, and even across time. The methodological and conceptual framework subsequently developed by Hennig (1950, 1966, introducing concepts such as "cladistics, synapomorphy, outgroups…") have made phylogenies an indispensable tool for addressing a broad range of biological questions, and they are employed in a variety of fields to study topics as diverse as epidemic dynamics (Wang et al. 2015, Worobey et al. 2020; **Figure 2A**), evolutionary pathways in the development of diseases such as cancers (Abu-Asab et al. 2008; **Figure 2B**), composition and resilience of ecological communities and ecosystems (Tucker et al. 2017; **Figure 2C**), relationships among or within species groups (from population to phylum level, e.g. Hug et al. 2016; **Figure 2D**), evolution of genes and gene families (Thornton and Rob DeSalle 2000, e.g. Hansen et al. 2021; **Figure 2E**), impact of geological event and paleocrises (e.g. Meredith et al. 2011, Paterson et al. 2019; **Figure 2F**).

The nature and scope of the scientific question are important, as they determine the nature of the building blocks used to construct phylogeny. The study of the evolution of gene families (evolution of gene function), of lineages of organisms or populations (SARS-Cov

strains, human cohorts, animal populations), or of entire groups of organisms (entire genera, families, orders, phyla... i.e. the reconstruction of what is known as the “tree of life”) is based on the same methodological framework, but with very different elementary bricks. Studying entire groups of organisms by reconstructing "species-level phylogenies" (each tip of the tree is a species) is a valuable tool for studying the global patterns that may have shaped current species diversity (Barraclough and Vogler 2000, Morlon 2014). However, the myriad of steps that separate the collection and sequencing of individuals to the study and use of a robust and dated phylogeny is a very long journey (this thesis being probably one of the witnesses). As it seems important to present the objectives of each of these steps, as well as the hypotheses and challenges that they represent, I will proceed in the first section of this introduction to present each of them in three main parts: 1) defining species, 2) reconstructing phylogenies using different markers, and 3) estimating divergence times using these phylogenies in the context of evolutionary questions. In a second section, I will focus on 4) how these tools can help to understand co-evolution and the evolution of phytophagous insects and 5) how this question could be addressed by studying two tribes of Papilionidae.

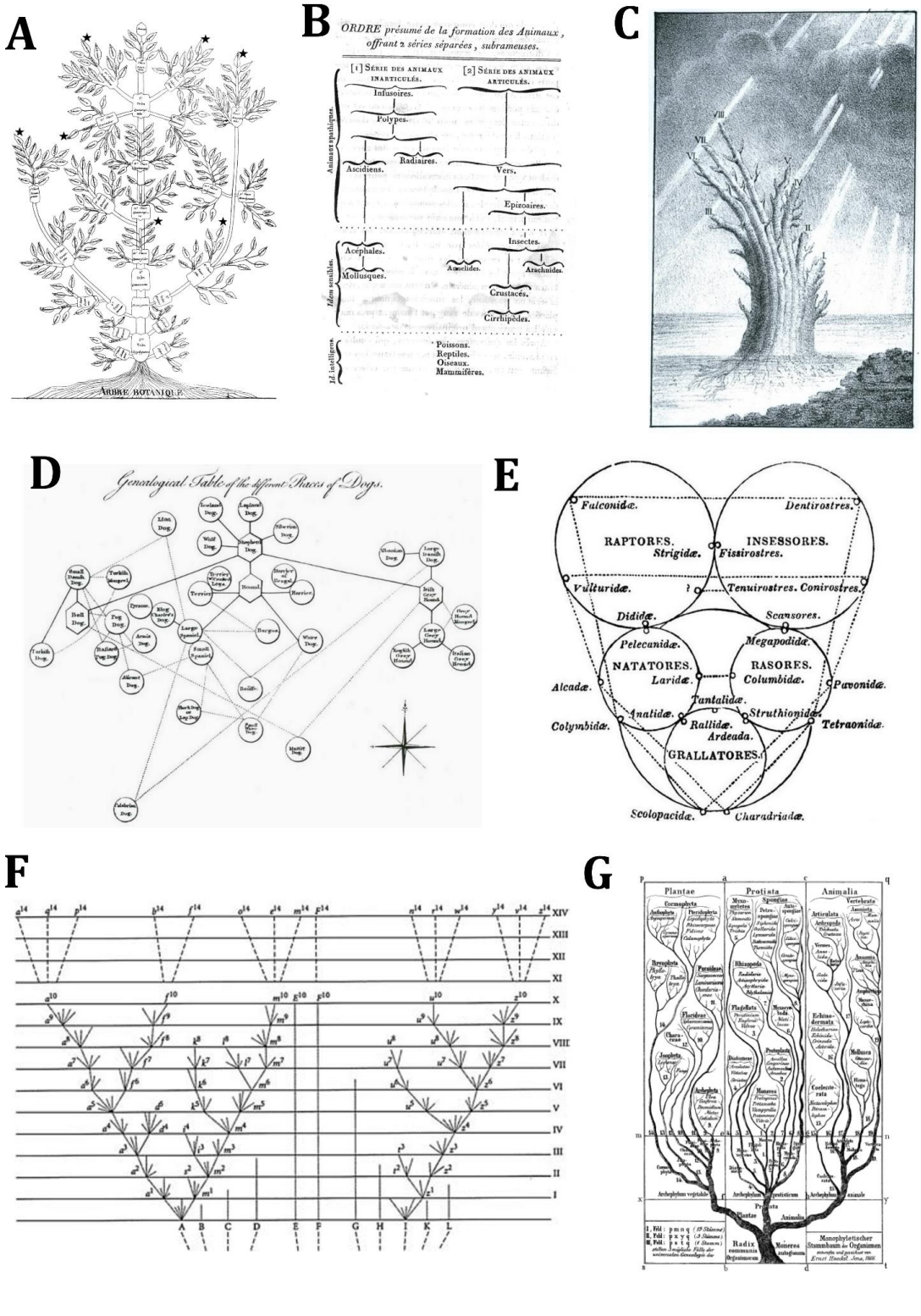


Figure 1. Phylogenetic representations from the 19th century. A. Botanical tree of life by Aubier (1801), B. Tree depicting two branching series of animal origins from Lamarck (1815), C. Tree of animal life, from Eichwald (1829), D. Network of dogs breeds from Buffon (1753), E. Quinarian system of birds from Swainson (1837), F. Great Tree of Life from Darwin (1859) and G. Genealogical tree depicting three kingdoms of life from Haeckel (1866).

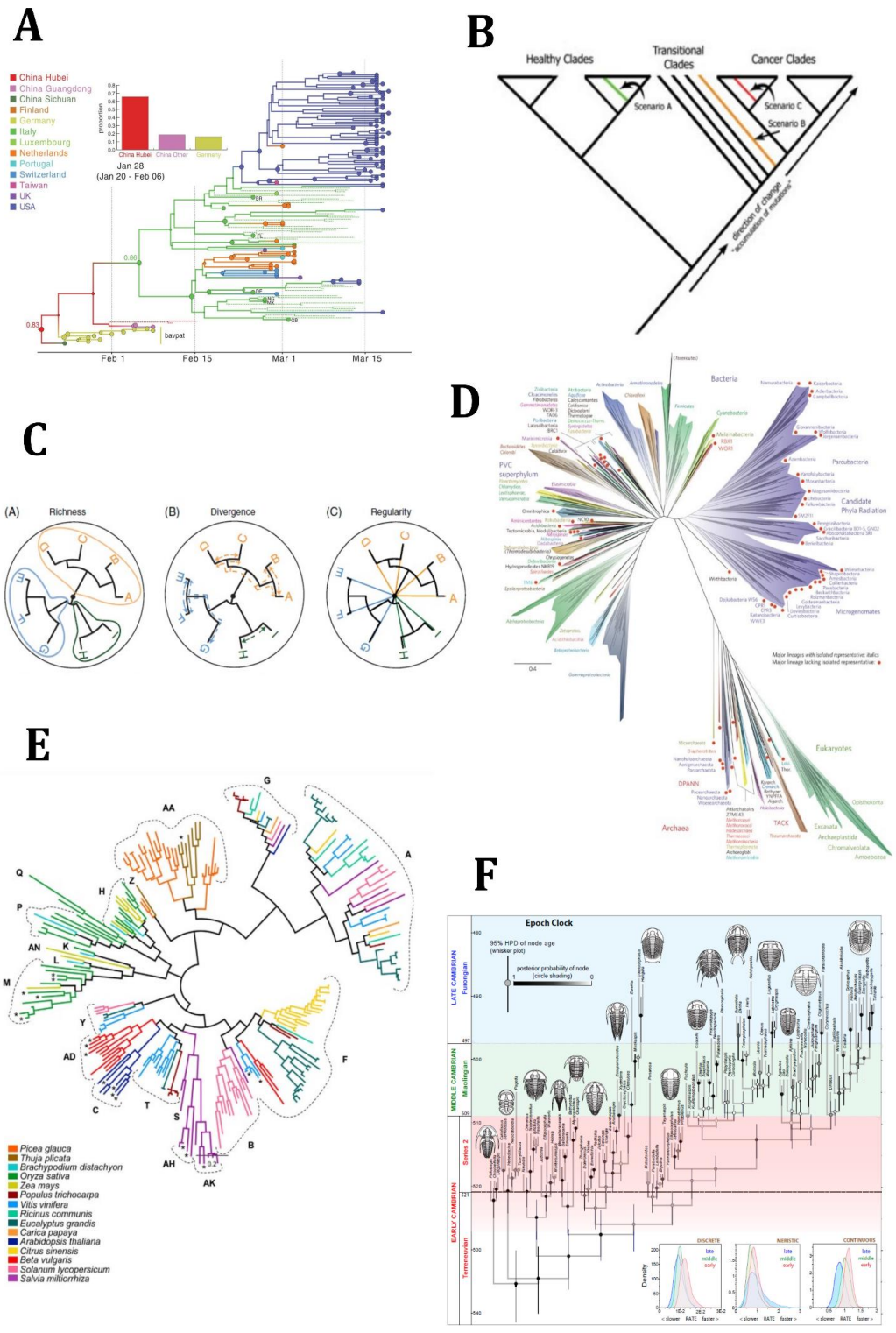


Figure 2. Some representations of modern phylogenies. A. Maximum Clade Credibility tree of SARS-CoV-2 entry into Europe (Worobey et al. 2020), B. Schematic topology of a typical proteomic cladogram of a cancer analysis (Abu-Asab et al. 2008), C. Conceptual diagrams illustrating three dimensions of phylogenetic information for Ecology: richness, divergence, and regularity (Tucker et al. 2017), D. Tree of life, including bacterial and archaeal and Eukaryotic groups (Hug et al. 2016, E. Tree representing the dynamics of the CYP76 family (Hansen et al. 2021), F. Dated time tree of Cambrian trilobites (Paterson et al. 2019).

1- The first step: defining species

Studying entire groups of organisms to understand global patterns of diversity (i.e. macroevolutionary questions) involves reconstructing species-level phylogenies. However, defining what are species, in particular differentiating them clearly from subspecies or populations of the same species, represents a thorny issue (sometimes referred to as the "species problem") and constitutes in itself an entire field of biology (Hey 2001, de Queiroz 2007, Stankowski and Ravinet 2021a). This is yet all the more essential as species are considered by many scientists and research fields as one of the fundamental units of biology, even granting them an importance comparable to that of genes, cells and organisms, some of the fundamental units at lower levels of biological organization (e.g. Mayr 1982, de Queiroz 2005, 2007). Yet biologists still do not agree on a single, exact concept of species, and some of the different variations of the species concept are even being at least partially incompatible (Mayden 1997, de Queiroz 1998a, Harrison 1998). Recently, Stankowski and Ravinet (2021a) have reviewed that dozens of species concepts are currently recognized and frequently used, and that two randomly chosen biologists will most likely disagree on the nature of species. Visions of the species concept generally vary around three main levels of concepts: 1) the biological species concept, where species are recognized on the basis of reproductive isolation, 2) the evolutionary species concept, where species are cohesive evolutionary units (genetically and phenotypically), and 3) the phylogenetic species concept, where species are historically related groups (e.g. smallest monophyletic group, diagnosability) (Sangster 2013, Stankowski and Ravinet 2021a). In fact, a specific concept is all the more likely to be adopted and shared by researchers who have similar objectives, similar training, and a studied type of organism in common (Stankowski and Ravinet 2021a). Moreover, the evolutionary scale of research is one of the main factors of disagreement, and researchers in microevolution, who study population-level changes over short time scales, will thus be much more likely to use a biological species concept than researchers in macroevolution, who study changes over long periods and in large taxonomic groups, and whose comparative macroevolutionary and phylogenetic analyses generally consider speciation at this scale to be a discrete event, and estimate its rate of occurrence over time, for example (Stankowski and Ravinet 2021b). De Queiroz (2007) has pointed out that many of these concepts and their associated definitions are incompatible in that they can lead to different conclusions regarding species delineation and numbers. Thus, in his view, the problem of the species concept - i.e. current disagreements over the theoretical

concept of species - is closely linked to the question of species delimitation - i.e. how to determine species limits and numbers based on empirical data (de Queiroz 2007).

Species are difficult to define due to the continuous and still partially unpredictable process of speciation on a microevolutionary scale (Stankowski and Ravinet 2021b). This "speciation continuum" is sometimes simply defined as a continuum of reproductive isolation (Stankowski and Ravinet 2021b, **Figure 3**), i.e. reproductive isolation is established by the reduction in potential gene flow between populations caused by isolating barriers. Speciation can then be considered to correspond to the transition from a state of free gene exchange within populations to the interruption of these gene flows (i.e. reproductive isolation; Galtier 2019). This transition takes place through the accumulation of mutations at certain points in the genome (also known as barrier loci), which lead to reproductive incompatibility (no reproduction or reduction of hybrid fitness) and thus reduce gene flow (Barton and Bengtsson 1986, Rieseberg et al. 1999, Lowry et al. 2008)

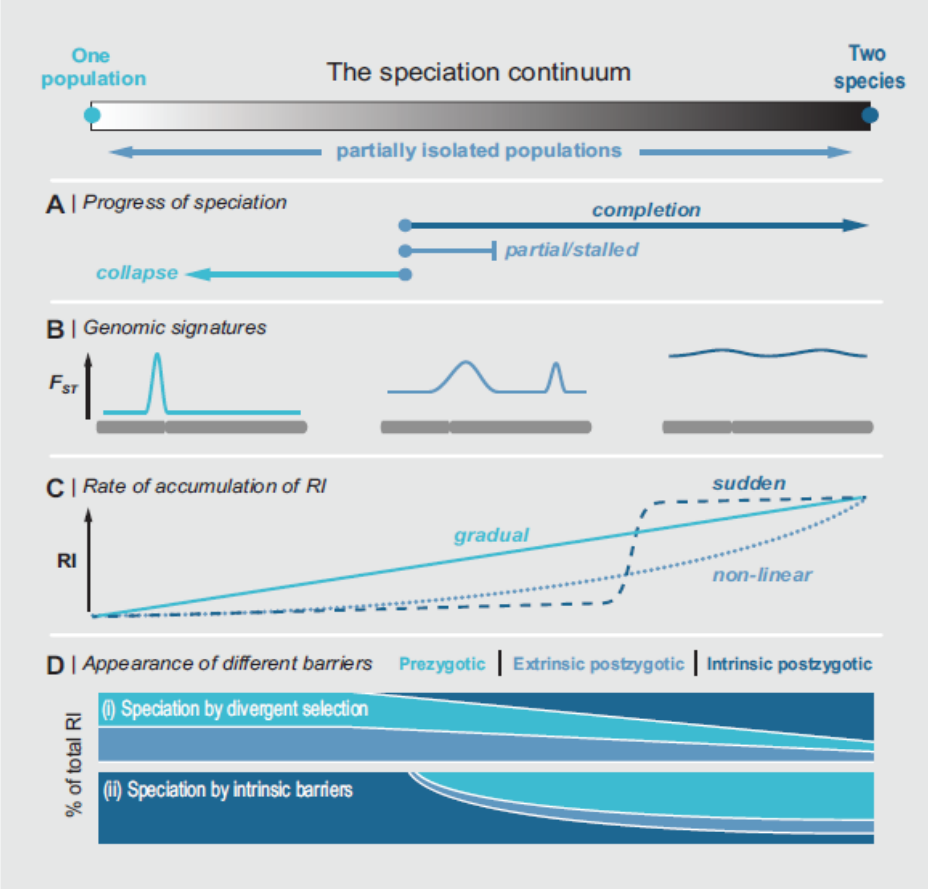


Figure 3. Graphical depiction of the speciation continuum and some problems that it has helped address, from Stankowski and Ravinet (2021b).

This means that species delimitation at our human scale requires positioning on the speciation continuum, and this position is likely to depend on many factors such as the objective and focus of the study, the organisms studied and their intrinsic characteristics (mode of reproduction, mode of displacement...) and the history and overall consistency of the taxonomy of the studied group. For example, studies of speciation with gene flow have largely focused on populations with low levels of reproductive isolation, mainly driven by divergent selection on adaptive traits (e.g. Turner et al. 2005, Marques et al. 2016). In contrast, most classical work on the genetic basis of hybrid sterility and unviability has focused on the end of the speciation process (e.g. Presgraves 2002, Phadnis et al. 2015). In addition, other aspects not directly biological are sometimes involved in taxonomic decision making, since species are commonly considered as units for conservation and policy makers (Delić et al. 2017, Coates et al. 2018), one of the well-known examples of the influence of conservation on species description being the Galapagos tortoises Loire and Galtier 2021).

In summary, the objectives of biologists are extremely varied, and the diversity of concepts can be seen as a solution to needs that vary according to the diversity of approaches. The real "problem" arises when the concept used is not clearly explained, thus preventing future authors, scientists, and others from adapting when interpreting results or making decisions (Stankowski and Ravinet 2021a). There might be nothing wrong with disagreeing with the species delimitation decision, as long as the different decisions made are clearly stated and allow following works to adapt (Sangster 2014, Stankowski and Ravinet 2021a). Given the complexity and continuity of the species concept, species description or revision needs to be done with care and a lot of evidence (shared with the community) so that taxonomy doesn't become an unsteady ground on shaky elementary bricks (Christenhusz 2020, Padial and De la Rive 2021). For this thesis, I tried to follow two principles in an attempt to limit the difficulties inherent in the concept of species: 1) integrate as much accessible data as possible to define species (this is the concept of "integrative taxonomy"; Padial et al. 2010, Schlick-Steiner et al. 2010, Yeates et al. 2011, Sangster 2014), and 2) I tried to justify each of my choices as much as possible and make my data as accessible as possible to others so that future studies can easily revise my taxonomic choices if necessary.

2. Reconstructing robust phylogenies: the second step

Concepts and bases

Cladistics seeks to answer the following question: *given any group of more than three taxa, which taxa are most closely related to other taxa?* (Funk 1995). The possibility of answering this question is based on the notion of heritability of characters (Mendel 1866), on the homology of the characters considered within a group of organisms (Jardine 1969, Sneath, 1969, De Pinna 1991), but also on the variability of these characters within the group studied. If the selected characters are informative (sufficient variability) and inherited (not the result of convergence), it is possible to determine relatedness between taxa by identifying the sharing of one or more unique derived characters that other taxa outside the group do not possess, known as apomorphy (Funk 1995).

The theory and concepts behind modern phylogenies (such as monophyly and synapomorphy) were first formalized by Hennigs (1966, 1969) on the basis of phenotypic characters (morphological characters), and this field is still under development today (Scornavacca et al. 2020). Another major turning point has been the advent of DNA sequencing and genomic markers (Metzker 2010). Whereas the coding of morphological traits is difficult and time-consuming, and could be the subject of convergence, DNA emerged as a more objective and numerous sources of information for phylogenetic reconstructions (but see Parker et al. 2013, Thomas and Hahn 2015, Zou and Zhang 2015 for example of a debate on molecular convergence).

In a nutshell, the history of cladistics and phylogeny has seen three main periods: morphological phylogenetics, when morphology was the only source of data (roughly before 1990); the Sanger sequencing period, when a few genes dominated most studies (roughly between 1990 and 2005 *sensu* Kjer et al. 2016); and the current state of knowledge, with datasets so large that traditional methods of analysis are no longer feasible (Kjer et al. 2016). Numerous references describe this fantastic journey in detail (e.g. Delsuc et al. 2005, Kjer et al. 2016) and would be far more comprehensive and documented than I can do here. Nevertheless, what is important to understand is that thanks to the widespread use of Sanger sequencing, the number of genetic markers and the number of samples in phylogenetic inference have increased significantly, improving our understanding of the evolutionary history of many groups on the tree of life (Hinchliff et al. 2015).

However, phylogenetic studies based on a handful of molecular markers can still result in poorly supported phylogenetic relationships, particularly for certain groups that have

diverged for a long time or for groups showing high diversity or rapid radiation (Kingston et al. 2009, Nyakatura and Bininda-Emonds 2012, Alda et al. 2021). Yet recent technological developments ("next-generation sequencing" and "third-generation sequencing" technologies) have enabled the access to large quantities of genome-wide data, and it has become possible to sequence the entire genome of a growing number of organisms at ever lower cost (Bonetta 2010, Goodwin et al. 2016, Mardis 2017). The field of phylogenetic reconstruction has entered the genomic era, posing at the same time new challenges (Delsuc et al. 2005, Scornavacca et al. 2020).

The use of genomic data - compromise between fantastic power and computational shortage

It has now long been shown that phylogenetic studies based on a handful of molecular markers could result in poorly supported phylogenetic relationships, and that use of genomic data have been shown to alleviate some of these issues by expanding the number of characters that can be used in phylogenetic analysis from a few thousand to tens of thousands (Delsuc et al. 2005, Philippe and Blanchette 2007), and help to produce better resolved phylogenetic trees and especially robust backbones. Yet, this new armament comes with new challenges (Eytan et al. 2015, Alda et al. 2021). First, analyzing such large amounts of sequencing data is difficult from the computational, statistical and modeling points of view (Philippe et al. 2019, Simion et al. 2020, Allio 2021). On another note, the complexity of data sets grows with size, the potential for poor model fit to bias inferences also grows. Increasing data set size may reduce stochastic error, but it can also exacerbate systematic error and lead to high confidence in erroneous phylogenies (Delsuc et al. 2005, Kumar et al. 2012, Richards et al. 2018, Alda et al. 2021).

Classically, the reliability of a phylogeny is assessed by the robustness of its nodes, which is usually estimated by the statistical representation of a given node in the various samples or resamples of an analysis. These node support measures, such as the well-known bootstrap (Felsenstein 1985, Minh et al. 2013) and Bayesian posterior probabilities (Ronquist and Huelsenbeck 2003), are indeed crucial for assessing the robustness of inferences made from phylogenetic analyses. However, while these measures provide useful information about the statistical support of a given node, none capture the heterogeneity or topological variation present in the underlying data (Kumar et al. 2012, Minh et al. 2020). It has been shown that traditional statistical measures of nodal support are prone to saturation in phylogenomic data sets (Minh et al. 2020, **Figure 4**) and thus, different genomic datasets can support conflicting phylogenetic hypotheses with very high statistical support (Rokas et al. 2003, Galtier 2004,

Bleidorn et al. 2009, Minh et al. 2020), sometimes implying very different scenarios for the evolution of important traits (e.g. the metazoan nervous system: Dunn et al. 2008, Hejnol et al. 2009, Moroz et al. 2014, Ryan 2014, Borowiec et al. 2015, Simion et al. 2017). However, the relative roles of biological processes and systematic errors in causing this conflict still require much theoretical and methodological work to be distinguished (Richards et al. 2018).

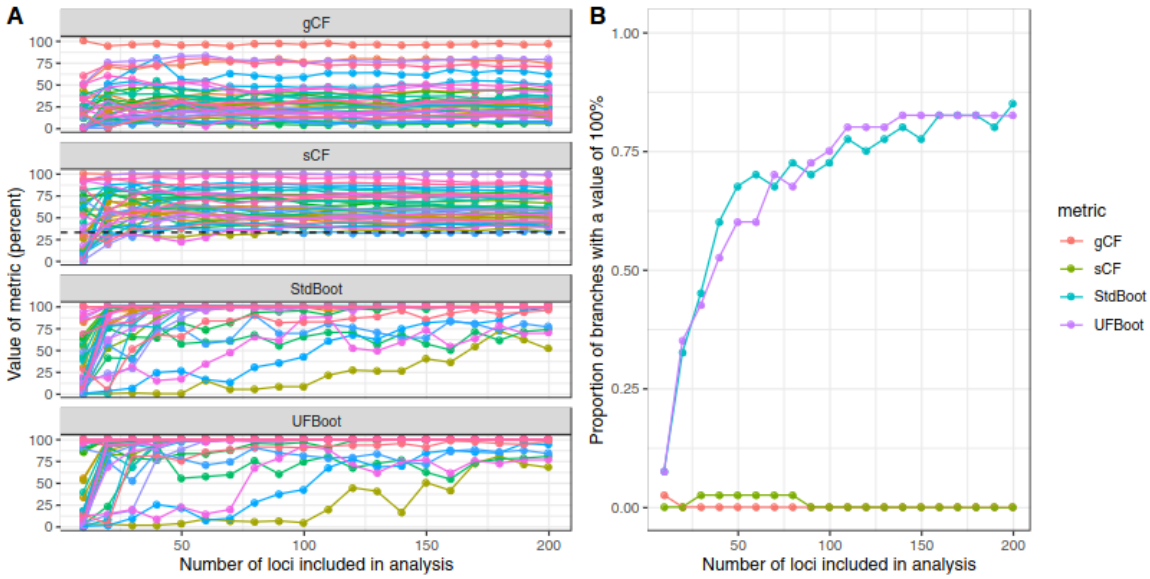


Figure 4. Illustration from Minh et al. 2020, showing that concordance factors remain relatively stable as loci are added to an analysis, whereas bootstrap values continue to increase toward 100%. A. Results of reanalysis with ten new loci added to the analysis, with each colored line representing a branch in the tree. B. For each of the four metrics considered, the number of loci included in the analysis affects the proportion of the branches that have a value of 100%.

On the one hand, numerous errors or too strong approximations could be made along the process of reconstructing a phylogeny with genomic data. For example, the increasing accessibility and automation of sequencing numerous whole series of species brings the risks of cross-contamination (Simion et al. 2020), or the large size of the studied groups, and sometimes the great phylogenetic distance that separates them, makes genes homology prediction all the more difficult (Thornton and Rob DeSalle 2000, Fernández et al. 2020, Simion et al. 2020). Thus, hidden paralogy, alignment problems and substitution model mis-specification are, among others, well-documented sources of discrepancy between gene trees and species trees (Scornavacca et al. 2020, Galtier 2023).

In addition, biological phenomena such as genetic introgression or incomplete lineage sorting could also lead to complex gene evolution and incongruent gene trees (Funk and

Omland 2003, Jennings and Edwards 2005, Rheindt et al. 2009, Mallet et al. 2016, Galtier 2023). Although incongruences in gene trees were initially associated with weak phylogenetic information for each locus, many recent examples confirm the biological causes of this topological variation in recently diverged species (e.g. Hobolth et al. 2011, Fontaine et al. 2015, Pease et al. 2016, Meyer et al. 2017, Edelman et al. 2019, Rogers et al. 2019, Meleshko et al. 2021, Rivas-González et al. 2023). Specific methodological approaches that take into account the distinct evolutionary history of each marker have been developed to address this problem (e.g. Rannala and Yang 2003, Wen and Nakhleh 2018, Flouri et al. 2020, Rannala et al. 2020, Minh et al. 2020, Galtier 2023). Instead of using concatenation approaches that muddle the heterogeneity signal underlying large datasets, the idea is to acknowledge that different genes of an organism may have different evolutionary histories and to use this information with tree-reconciliation (also called "coalescent") approaches to accurately account for the independent evolutionary history of each marker and ultimately make sense of these phylogenetic networks (Rannala and Yang 2003, Rannala et al. 2020).

Overall, the progressive integration of solutions discovered to address the many challenges posed by phylogenomic datasets continues to improve phylogenomic inference today. It has led to promising techniques and perspectives that use both robust phylogenetic inferences or more complex phylogenomic networks, as well as the enormous amount of information stored in genomes to better understand the evolution of entire clades. For example, genomic analyses of adaptive selection signatures, combined with robust phylogenetic context, have recently made it possible to associate evolutionary events with molecular signatures, representing a promising strategy for linking molecular evolution to life historical traits (Partha et al. 2017, Allio et al. 2021). Our capacity to generate highly accurate and robust phylogenetic trees is therefore crucial to understand the role of different natural processes in shaping biodiversity.

3. Dating phylogenies to study macroevolutionary processes

Dating phylogenies is a crucial but challenging step…

A phylogenetic tree represents the evolutionary relationships and genetic distances between taxa. Although it is already an enormous amount of information by itself, it does not directly provide information about absolute evolutionary time, referred to as time of divergence between the taxa. Yet, some questions need to be translated in absolute (geological) time to be

explored. This is for example the case of historical biogeography (comparing the evolutionary history of a group with global processes of the Earth), inference of diversification rates (estimating speciation and extinction rates across time), trait reconstructions and comparative methods (inferring ancestral states or testing the correlation between traits), or co-diversification studies (testing whether two clades of interacting organisms, such as host and parasites or plants and pollinators, have diversified synchronously).

The concept of "molecular dating" was initially proposed and developed in the 1960s by a series of articles (Zuckerkandl and Pauling 1962, 1965, Doolittle and Blombäck 1964, Margoliash and Smith 1965). They developed the idea that the time of divergence between two species could be measured by the number of differences between two molecular sequences, thus relying on the hypothesis of a "molecular clock", making the assumption that the rate of molecular evolution remains constant over time (Sauquet 2013), and that the fossil record, for example, could be used to calibrate this molecular clock. While this may seem simple and self-evident at first glance (assuming that the more genetically distinct the taxa, the older their common ancestor), the idea was quickly described as too simple a solution for the "too complex and variable" process that is molecular evolution (Mayr 1965; in Zuckerkandl and Pauling 1965, Ho 2020). Indeed, it has been shown that the rates of molecular evolution are neither constant over time, nor between genes, nor between taxa. As a result, ‘relaxed’ molecular clocks were developed to account for this rate heterogeneity (Sanderson et al. 2004, Drummond et al. 2006, Rutschmann 2006). Today, molecular dating has become a commonplace tool for many evolutionary biologists ranging from ecologists to epidemiologists (Sauquet 2013).

Technically, phylogenomic tree dating can be achieved using methods that rely on a molecular matrix (e.g. BEAST, MCMCTree and PhyloBayes) or on branch lengths of previously inferred gene trees (e.g. PATHd8, r8s and treePL) (Allio et al. 2020). In both cases, it is usually necessary to calibrate the model using external information, and calibration has been shown to have a critical impact on estimated divergence times (Ho et al. 2008, Goodall-Copestake et al. 2009, Inoue et al. 2010, Magallón et al. 2013, Meredith et al. 2011, Ruane et al. 2011, Sauquet et al. 2012, Sauquet 2013). A well-furnished fossil record is generally considered the most reliable source of calibration, even if it presents additional difficulties, such as managing phylogenetic uncertainties and accurate fossil positioning (Ho and Phillips 2009, Sauquet 2013). However, many groups of organisms (such as bats: Brown et al. 2019; birds: Brocklehurst et al. 2012; Lepidoptera: Sohn et al. 2015) do not have a good enough fossil record to date trees and therefore require other sources of calibration. Dates obtained in a previous molecular dating analysis (known as secondary calibration) are commonly used as a

last-resort calibration when no other reliable sources are available (Sauquet 2013, Schenk 2016), although they are considered to increase error and uncertainty in estimated ages (Graur and Martin 2004). Consequently, particular attention must be paid to the quality of the original analysis and the use of multiple secondary calibration sources (Sauquet 2013).

From a practical point of view, a few genes and taxa might be sufficient for molecular dating, but, as with phylogenetic reconstruction, significant improvements in precision and accuracy (robustness) could be achieved by using as many genes and taxa as possible (Sauquet 2013, Bromham et al. 2018). Yet the genomic datasets generated today represent computational bottlenecks that make phylogenomic dating unfeasible in reasonable timeframes (dos Reis et al. 2016, Collins and Hrbek 2018, Smith et al. 2018, Allio et al. 2020). Bayesian methods, which are today considered the most reliable (such as BEAST or PhyloBayes), are still constrained by the number of taxa or the amount of data to work within a reasonable time (Smith et al. 2018). The development of computationally efficient methods seems to be the next crucial step in this field to keep pace with the continuing development of phylogenomics (To et al. 2016).

… but also an indispensable tool to study drivers of diversification

However, once you have good reasons to believe that a phylogeny is robust and well dated, it becomes a powerful tool to investigate the relation between apparition of lineages (speciation) and disappearance of species (extinction) through time, in relation with intrinsic and extrinsic characters of the studied group of organisms. By studying tempo of diversification of the living things, macroevolution tries to understand three main things: *what are the drivers that make the number of species varying over time* (e.g. mass extinction crises), *across space* (diversity hotspot) *and across lineages* (e.g. adaptive radiation, mammals vs. plants diversity)*?*

For centuries now, the question of temporal heterogeneity has been addressed with paleontological data, notably for famous groups such as the emblematic dinosaurs (Brusatte et al. 2015), or the extinction of megafauna (O'Keefe et al. 2023), and the diversification of microfauna (Benson et al. 2021). Such research into the distant past is obviously facilitated for groups with extensive fossil data, a direct witness of the past and an important indicator of richness over time. However, beyond the dependence on direct evidence of the past (with fossils in the first instance, although we will see later that phylogenies themselves possess a certain signature of these past extinctions), the question of spatial heterogeneity turned out to be more directly accessible and a subject of scientific curiosity for many clades, since it is directly visible in the number and repartition of species today (Rosenzweig, 1995; Gaston,

2000). When Myers et al. (2000) and Mittermeier et al. (2011) presented the different global biodiversity hotspots, the question of spatial heterogeneity seems to be directly perceptible and underlying: *Why is the distribution so uneven, and why does diversity and endemism vary so much from one region to another?*

Interestingly, distribution data of species associated with relationships and divergence times in a phylogeny allow to infer the biogeographical history of living lineages (Ree 2005, Ree and Smith 2008, Matzke 2014, Beeravolu and Condamine 2016). Historical biogeography inferences aim at understanding the origin of species or clades and tries to identify the factors that have shaped the current distribution of biodiversity. On a global scale for example, the so-called "latitudinal gradient of biodiversity", i.e. the increase in the number of species from the poles to the equator, has been an important source of macroevolutionary research and debates (Willig et al. 2003, Hillebrand 2004, Cardillo et al. 2005, Barthlott et al. 2007, Condamine et al. 2012, Mannion et al. 2014, Rabosky et al. 2018, Meseguer and Condamine 2020). At finer scales, numerous studies over the last decades have attempted to understand the difference in specific richness within certain regions, and to determine the patterns and evolution over time. For example, the Neotropics are one of the most important and emblematic regions including seven hotspots and one of the three wilderness areas, the Amazon Forest, which has been the subject of much research (Antonelli et al. 2009, Lagomarsino et al. 2016, Pérez-Escobar et al. 2017, Antonelli et al. 2018, Chazot et al. 2019). Indeed, some Neotropical groups have occupied scientists because they currently show a pattern of species richness centered in Amazonia, with relatively few species outside, while others show a quite different pattern, centered in the Andes and extending across Central America, but grouping few species in the lowland rainforest of Amazonia (Gentry 1982, Antonelli and Sanmartin 2011, Meseguer et al. 2022). Some authors have hypothesized that these distribution patterns of species richness could be explained by geological events that took place in this region. Thanks to geological and paleontological resources, such as the reconstruction of the paleogeographic sequence of events (Boschman 2021) or paleoclimatic conditions such as temperature and oxygen content (Hoorn and Wesselingh 2011), scientists have shown that the Andean orogeny, in particular, profoundly affected regional climate and rainfall cycles, as well as the landscape evolution and connectivity between regions over the last 60 million years. Thus, the dynamics caused by the uplift of the Andes was found to have had a strong effect on diversification of clades as diverse as birds (e.g. McGuire et al. 2014), frogs and lizards (Boschman and Condamine 2022), butterflies (Elias et al. 2009, Chazot et al. 2019) and angiosperms (e.g. Hughes and Eastwood 2006, Antonelli et al. 2009, Madriñán et al. 2013, Lagomarsino et al. 2016).

Similarly, in a completely different area, Wallacea, the South-East Asian tropical zone that includes all the islands between Sundaland, the Philippines, northern Australia and New Guinea, has been the subject of numerous studies that attempt to establish relationships between changes in sea level and island connectivity to understand the distribution of species, their more or less recent isolation, their high diversity or rate of endemism, etc… (de Bruyn et al. 2014, Condamine et al. 2015, Esquerré et al. 2020, Letsch et al. 2020).

Thus, environmental changes and geological events, seen as extrinsic factors, are important candidates for explaining changes in the rate of diversification (Condamine et al. 2019a). However, this hypothesis, known as the "Court Jester", is not the only one that can be considered. The "Red Queen" hypothesis, which states that biotic factors affect diversification (such as the role of biological key innovation, ecological limits or adaptive radiation), is also prominent in the debate (Benton 2009) and has been demonstrated in a large range of clades (e.g. McPeek 2008, Philimore and Price 2008). Yet, the two types of factors undoubtedly have an effect on diversification and the final support and conclusions may depend on what is actually being tested (e.g. Silvestro et al. 2015, Liow et al. 2015, Condamine et al. 2019b, 2020, Guo et al. 2023). Increasingly, the integration of these two types of factors (Condamine et al. 2019a, Condamine et al. 2021), and even the existence of a synergy between the two (Lagomarsino et al. 2016, Condamine et al. 2019b), is now being proposed. The issue is therefore more complex than it seems, and the simultaneous interplay of these factors makes correlation analyses between the two all the more difficult.

While historical biogeography attempts to retrace the evolutionary history of clades from a spatial point of view, this is not the only information that can be gleaned from a phylogeny. Indeed, in the course of a clade's evolutionary history, lineages appear, and others disappear. The diversification rate is the balance between this tempo of appearance and disappearance, i.e. the speciation and extinction rates. The estimation of these rates usually relies on so-called "birth-death" models applied to phylogenies. Different models have been developed to infer these diversification rates in relation with various factors, such as time (Morlon et al. 2011), environmental variables (Condamine et al. 2013, 2019a), clade specific traits (Beaulieu and O'Meara 2016), or to model diversification along branches (Rabosky et al. 2013, Rabosky 2014, Höhna et al. 2016, Maliet et al. 2019), among clades (Mazet et al. 2023), or as tree-wide sudden changes in diversification rates (Stadler 2011, May et al. 2016). By comparing models, we can then estimate which modalities of parameter variation are more likely to explain branching processes in phylogenies. The idea is to choose the model that best explains the current diversity of a group and the type and timing of branching observed in the

phylogenetic tree, and thus reconstruct a hypothetical scenario for the diversification of a clade. Since these models were first published and used, a number of studies have suggested that such deductions should be made with great caution (Kubo and Iwasa 1995, Crisp and Cook 2009, Antonelli and Sanmartin 2011, Louca and Pennell 2020), by highlighting the difficulty of reliably estimating model parameters of the models. However, Helmstetter et al. (2022 = **Appendix I**) suggested that the issue of non-identifiability may only apply to clade-independent and time-dependent models for which exact values of speciation and extinction rates cannot be estimated, but that this may not apply to clade- or trait-specific models, concluding that we should continue to explore the use of these models, while being aware of their limitations (Helmstetter et al. 2022). In addition, certain conditions, such as an assumption-based framework that focuses more on the relative fit of different scenarios than on exact parameter values, or such as a good fossil record, low extinction rates or exhaustive specific sampling, could reduce the risk of inferring erroneous diversification rates (Morlon et al. 2022, Helmstetter et al. 2022).

As the most diverse eukaryotic group on Earth (Grimaldi and Engel 2005, Stork 2018), insects are a particularly interesting group to study macroevolutionary questions. In insects, the fossil record suggests a low probability of extinction (Labandeira and Sepkoski 1993, Condamine et al. 2016). On the other hand, gathering dense specific sampling for a highly diverse insect group can prove more demanding. Whereas insects are currently the eukaryotic group with the most described species (Stork 2018), they are comparatively absent from many macroevolutionary and biogeographic studies (Mittelbach et al. 2007, Antonelli et al. 2018, Condamine et al. 2019a, Meseguer et al. 2022; but see Hoorn et al. 2010). Still few molecular phylogenetic studies include the vast majority of species to infer robust or extensive insect evolutionary history and diversification (but see Condamine et al. 2018, Kergoat et al. 2018, Chazot et al. 2019).

4. Study diversification hypotheses with an entire insect clade

Insects are an ancient phylum (Rainford et al. 2014), and numerous lineages may have undergone major paleoclimatic changes that have undoubtedly affected their diversification and spatial distribution (Labandeira and Sepkoski 1993, Jouault et al. 2022). In that way, several insect groups have been the subject of biogeographic studies to study the diversification of insects as a function of extrinsic factors such as paleogeographic or paleoclimatic events

(e.g. Coleoptera: Condamine et al. 2013, Toussaint et al. 2016, 2017; Lepidoptera: Condamine et al. 2012, Chazot et al. 2019, Rougerie et al. 2022; bees: Almeida et al. 2019, 2023, Bossert et al. 2022). Nevertheless, as insects are an extremely rich and diverse group, other studies have sought to understand whether key innovations (which are intrinsic biotic factors) during the evolution of this group may also have played a major role in their diversification (Mayhew 2007, Nicholson et al. 2014, Blaimer et al. 2023). These key innovations are novel phenotypic traits leading to the subsequent radiation and success of a taxonomic group, by increasing specialization, allowing the access to new niches, or increasing fitness compared to other taxonomic groups (Heard and Hauser 1995). In insects, it may have been morphological traits such as the acquisition of flight (de Queiroz 1998b, Nicholson et al. 2014), physiological traits, such as metamorphosis (Nicholson et al. 2014, Rainford et al. 2014) or behavioral traits such as the evolution of sociality (Legendre and Condamine 2018, Chouvenc et al. 2021), or parasitoidism (Blaimer et al. 2023).

As a matter of fact, as a large proportion of insect species feed on plants (Jaenike, 1990), some authors have wondered if the apparition(s) of phytophagy in this group could have been one of these key innovations (Ehrlich and Raven 1964, Mitter et al. 1988, Farrell 1998, Rainford et al. 2015). Indeed, Mitter et al. (1988) showed that phytophagous clades are almost systematically more diverse than their non-phytophagous sister groups. Studying the extremely diverse plant-feeding beetles (Chrysomeloidea and Curculionoidea), Farrell (1998) hypothesized that the diversification of the flowering plants (angiosperms) that rapidly expanded in taxonomic and morphological diversity in the Cretaceous (Oberprieler et al. 2007, Shin et al. 2018, Ramírez-Barahona et al. 2020, Benton et al. 2022) was the principal factor accounting for the species richness of these two groups today. Since then, this hypothesis received various support, with different insect groups studied with phylogenetic inferences at various taxonomic scales (e.g. Peccoud et al. 2009, Nosil et al. 2012, Cardinal and Danforth 2013, Arhens et al. 2014, Letsch et al. 2018, Shin et al. 2018, Allio et al. 2021). Yet, in most groups, it remains difficult to clearly demonstrate the coevolution and the diversification triggered by the novel trait (de Queiroz 2002, Nicholson et al. 2014, Rabosky 2017, Jousselin and Elias 2019). Only a few groups like the well-known fig wasps display such an extreme pattern of diversity dependence and intertwined life traits with plants that it constitutes in itself a proof of the interlinked evolution of the two protagonists (Cruaud et al. 2012, **Figure 5**).

Sixty years ago, Ehrlich and Raven (1964) proposed a model, later called the "escape-and-radiate" model, in which an arms race between attacks by herbivorous insects and new defenses of their host plants induces species diversification through the creation of new

adaptive zones. To develop this theory of coevolution, the authors took as their object of study the butterflies (order Lepidoptera, superfamily Papilionoideae), which is a large group of more than 19,000 species described to date that originated around 100 million years ago in the Cretaceous (Kawahara et al. 2019, 2023). Since Ehrlich and Raven (1964), numerous studies have focused on different aspects of plant-insect interactions for various butterfly clades (e.g. Wheat et al. 2007, Pieridae: Edger et al. 2015; Hesperiidae: Sahoo et al. 2017; Nymphalidae: Janz et al. 2001, 2006, Nylin and Wahlberg 2008, Pena and Espeland 2015, Hamm and Fordyce 2016), some taking the example of the swallowtail butterflies (family Papilionidae), which belongs to the Papilionoideae, comprising over 600 species (Nakae 2021, Condamine et al. 2012, 2023) and appeared around 67 million years ago (Allio et al. 2021). These butterflies are distributed throughout the world, although most of their diversity is found in tropical regions, mainly in Southeast Asia, South America and Africa. The high diversity and specificity of host plants for Papilionidae caterpillars, and the wealth of knowledge gained about them in recent decades (e.g. Collins and Morris 1985, Tyler et al. 1994, Scriber et al. 1995, Condamine et al. 2012, Allio et al. 2021, Nakae 2021), make this family an ideal group to study plant-insect interactions in a macroevolutionary framework.

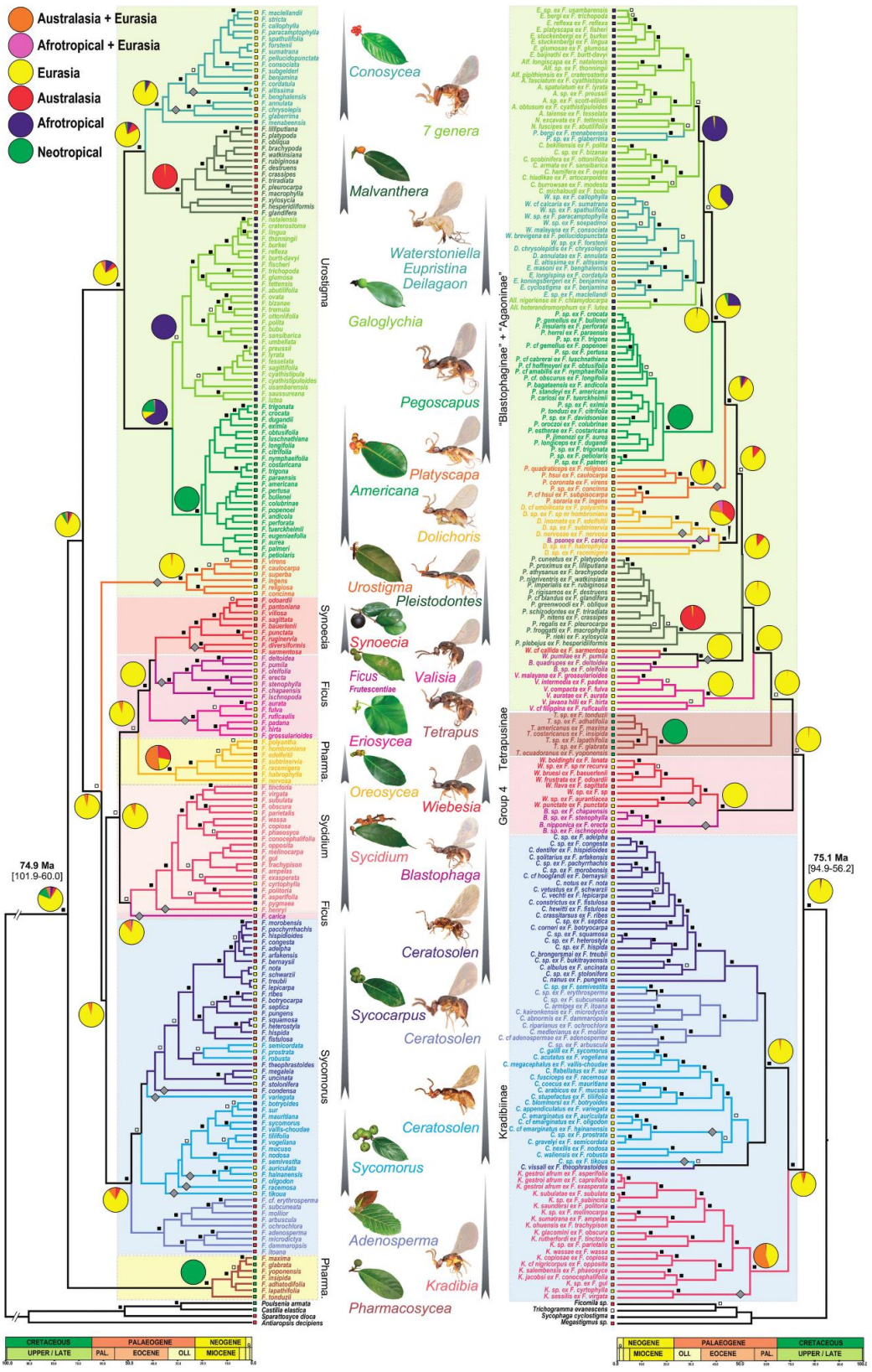


Figure 5. Phylogenetic chronograms of the evolutionary history of figs and fig wasps, from Cruaud et al. (2012). Groups of figs and their associated genera of pollinators are represented using the same color. Pie charts at main nodes show the likelihood of different geographic areas of origin.

The coevolution of phytophagous insects and their specific host plants is an antagonistic relationship, often resulting in physicochemical defenses on the part of the plant and circumvention of these defenses by the insects (Simon et al. 2015, Jousselin and Elias 2019). In the study by Ehrlich and Raven (1964), the authors noted several cases of phytophagous insects for which the larvae are unable to eat some plants, not because of the chemical compounds but because of the mechanical characteristics of the plant (tough or spiky leaves). Other studies have observed adaptive strategies in response to these barriers, ranging from oviposition strategies and plant consumption (Fordyce 2003, Pinto et al. 2009, Dimarco and Fordyce 2013), to avoid chemical defenses (toxin sequestration with benefits, Sime et al. 2000) and genetic adaptations (such as the enzymatic activity of Glucosinolate sulfatase, Heidel-Fischer et al. 2019; or the P450 gene complex in *Papilio*, Cohen et al. 1992, Berenbaum et al. 1996, Li et al. 2003). So, in addition to the mechanical defenses and behavioral or physiognomic (external physical characteristics) responses of phytophagous insects, most of the literature focuses on the chemical defenses of plants and the physiological and genetic modifications induced in phytophagous insects by the presence of these generally toxic chemical elements (e.g. in Pieridae: Steiner et al. 2019, Okamura et al. 2019, Griese et al. 2021). A well-known and well-developed example is the case of furocoumarins, contained in Rutaceae and Apiaceae, which are known to be highly photosensitizers, mutagenic and to inactivate the cytochrome P450 (detoxifier) of many organisms (Diawara et al. 1993, Del Río et al. 2014). *Papilio*, whose larvae feed on these two plant families, has been shown to possess modified P450 genes that is conserved within the genus (Cohen et al. 1992, Berenbaum et al. 1996, Li et al. 2003) and enable the detoxification of furocoumarins (Hung et al. 1995, Li et al. 2001). We thus have some examples of key gene evolution associated with key innovations in terms of plant-insect interactions, and some underlying mechanisms highlighted.

More recently, Allio et al. (2021) have demonstrated that a much larger genome-wide signature appears to be involved in plant-insect coevolutionary innovations with higher diversification rates after host-plant shifts associated with more genome-wide adaptive molecular evolution (i.e. more genes under positive selection). In this study, Allio et al. (2021) also tested whether the concomitance of host-plant changes with climate-preference changes could also have participated in increasing diversification rates (Becerra and Venable 1999). Surprisingly, when comparing the ratio of non-synonymous mutations to synonymous mutations (dN/dS) along specific branches that had undergone host plant and/or environmental changes in Papilionidae, they did not detect any particular effects due to environmental changes (Allio et al. 2021, **Figure 6**). However, the number of whole genomes in this study was

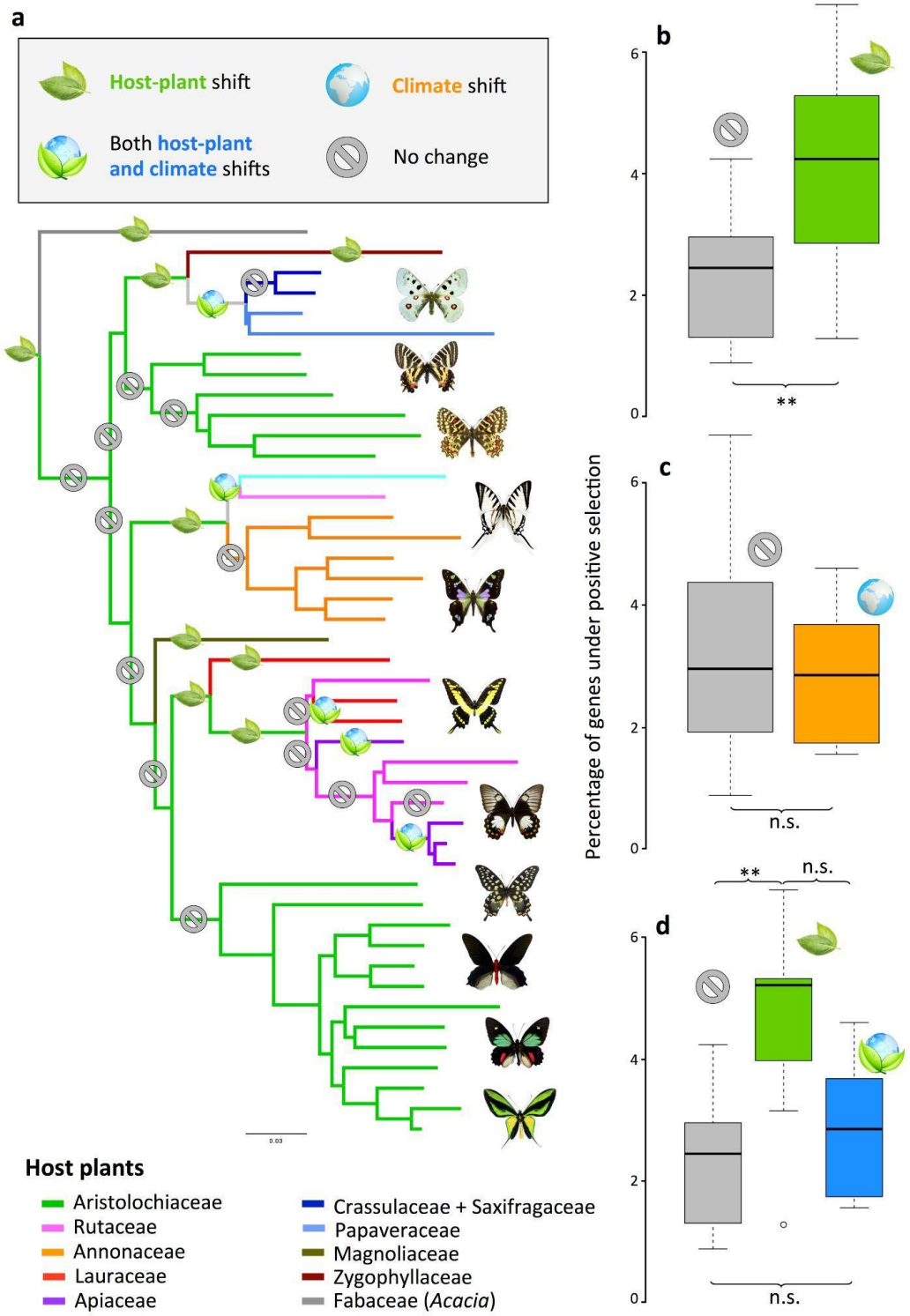


Figure 6. Phylogenomic study of host–plant shifts in swallowtail butterflies (Papilionidae) from Allio et al. (2021). a. Genus-level phylogenomic tree displaying branches with and without host–plant shifts. b. Number of genes under positive selection for lineages shifting to new host–plant families (n = 14) or not (n = 14). c. Number of genes under positive selection for lineages undergoing climate shifts (n =5) or not (n = 23). d. Number of genes under positive selection for lineages shifting to new host plants (n = 9), shifting both host–plant and climate (n = 5) or not (n = 14).

restricted to the genus level (except for genus *Papilio* where ten lineages were sampled to represent the main host-plant shifts), resulting in few branches tested (*n*=14) across the whole swallowtail tree of life. Similar analyses but with a higher statistical power could alter this conclusion. On the other hand, the classification of environmental changes was limited to "temperate" vs. "tropical", while the possible effects of other changes (biomes, altitude, region, morphological change...) had not been studied.

5. Presentation of goals and structure of this thesis

To explore the divers diversification hypotheses, a solid phylogenomic framework is needed (Jousselin and Elias 2019). Although Allio et al. (2021) produced one of the most complete and robust phylogeny of Papilionidae, some gray zones remain unresolved. In particular, the tribe Leptocircini, accounting for a quarter of all Papilionidae diversity, was less sampled than the rest of the tree and resulted in a poorly resolved clade. This is a limitation in the case of testing plant association hypothesis as several host plant shifts happened in this tribe. My thesis is part ERC GAIA, which aims at testing the hypothesis that the arms race between plants and insects may have favored their diversification. My PhD thesis constituted the first steps in exploring this question, by producing both new whole-genome data and a robust and almost exhaustive phylogeny of two tribes of similar age of emergence and richness within the Papilionidae: Leptocircini and Troidini. Indeed, whereas Leptocircini have undergone several changes of host plant during their evolutionary history (**Figure 7**), Troidini feed exclusively on Aristolochiaceae (**Figure 7**), a plant considered to be the ancestral host plant within the Papilionidae (Condamine et al. 2012, Allio et al. 2021). In order to achieve this goal, this thesis have involved various aspects such as field data acquisition, long- and short-read sequencing, reference genome assemblies, population-scale genomic and analyses and historical demography, systematics, phylogenomic analyses and macroevolutionary analyses (biogeography and diversification).

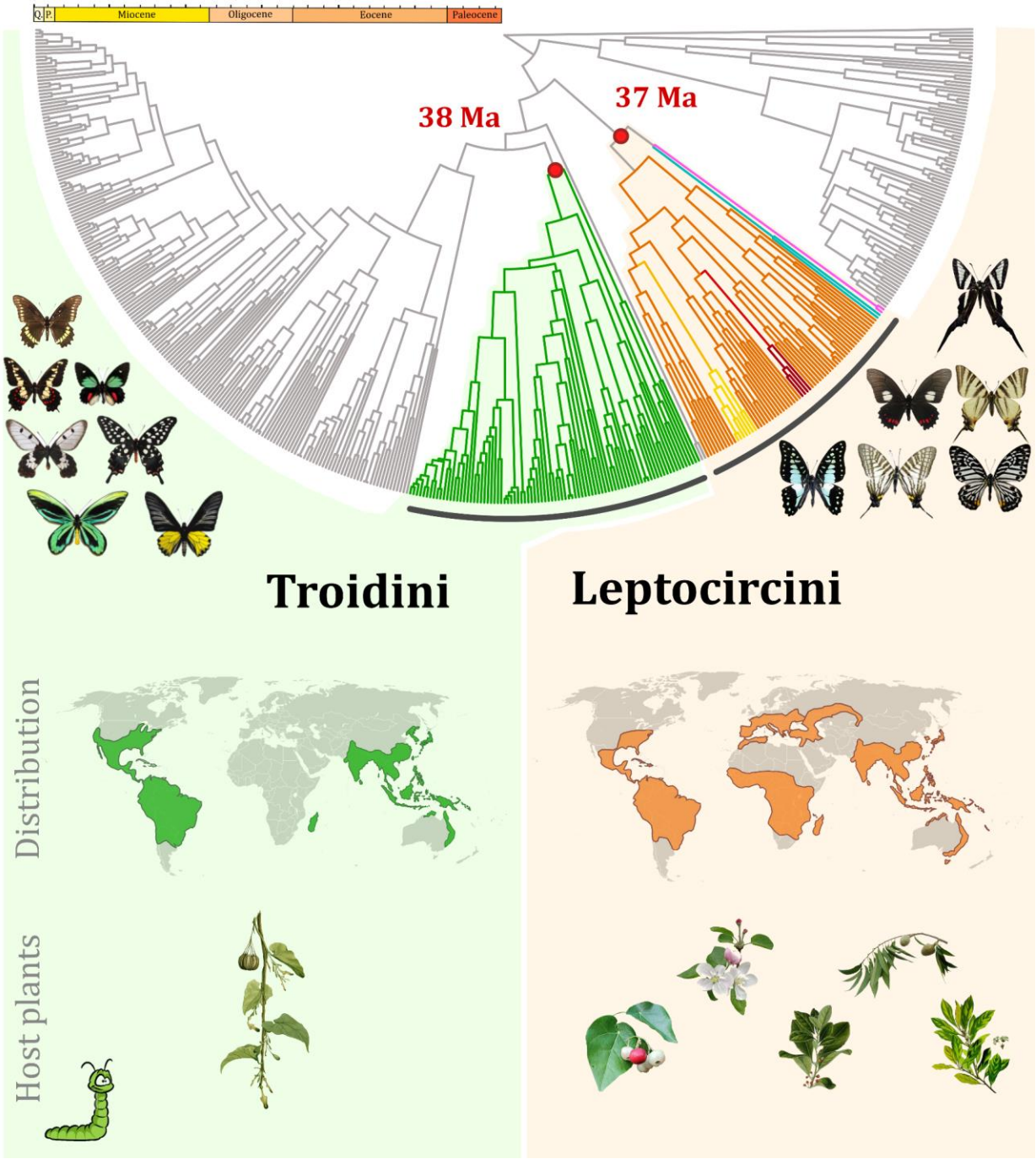


Figure 7. Comparison of the Leptocircini and Troidini. The phylogeny represented is from Allio et al. (2021), with the crown age estimate of both tribes. Tribe Troidini is colored green and tribe Leptocircini is colored orange. The distribution of the Troidini and Leptocircini are shown, as well as the main host plants. For Leptocircini, the host plant families represented are Annonaceae, Rosaceae, Hernandiaceae, Lauraceae and Aquifoliaceae. Photo *Graphium doson* (Alex Dumchus); *Lamproptera paracurius* (Adam M. Cotton); *Battus polydamas* (Cosmin O. Manci); *Parides childrenae* (Andrey Zheludev, CC BY-NC 2.5); *Cressida cressida* (CC BY-NC 4.0, JKDw); *Euryades duponchelli, Pharmacophagus antenor, Trogonoptera brookiana, Troides aeacus, Ornithoptera alexandrae* (Fabien L. Condamine); *Atrophaneura sycorax, Pachliopta pandiyana, Byasa alcinous* (Eliette L. Reboud.)

Hence, the structure of my thesis is articulated around the various steps and challenges of macroevolution, as described throughout this introduction. In **Chapter 1**, I performed **genomic approaches at the population scale to determine the taxonomic status** of two geographically isolated populations of the troidine species *Ornithoptera alexandrae*, which bear slight differences in phenotype. This species, one of the world's largest butterflies, is microendemic to Papua New Guinea and is classified as endangered by the IUCN, despite long-standing conservation efforts. By sequencing and assembling *de novo* reference genomes, including two closely related species, I investigated the population structure of this species and its implications for taxonomy. Beside, in addition to its application in phylogenomic research, reliable genomic information also represents a powerful tool for examining past and contemporary population structure and demography (Garner et al. 2016, Morin et al. 2021, Formenti et al. 2022). This resource is invaluable for species that are naturally rare and challenging to study (e.g. Westbury et al. 2018, Van der Valk et al. 2020, Morin et al. 2021). Therefore, this chapter also presented an opportunity to examine the historical demography of this species and assess the condition of their genomes and populations, providing new data for the policy makers to adapt their conservation strategy of this species. This study implied methodological consideration on the use of historical population demography by polymorphism-based methods on low-polymorphic invertebrates.

In the same vein as this population-based study, I collaborated with Corné van der Linden (PhD student) and Dr. Sabrina Simon (at Wageningen University and Research, The Netherlands), on the species delimitation of *Troides* species (the sister genus of *Ornithoptera*) leveraging the wealth of whole-genome data for the four species of the *amphrysus* group and museomic data at the subspecies and population levels (**Appendix III**).

In **Chapter 2,** I conducted **the species-level phylogenomic analyses of the tribes Leptocircini and Troidini**. By developing a bioinformatic pipeline with whole-genome data, I carried out two nearly complete species-level phylogenies (>93% of their species richness). Based on reference genomes for each genus, and shotgun whole genomes for each species, I extracted thousands of orthologous genes to perform phylogenomic analyses using supermatrix and tree-reconciliation approaches. The newly inferred species-level phylogenies allowed revisiting the systematics of the two groups. In that respect, the case of Leptocircini is particularly developed given its historical systematic recalcitrance within Papilionidae (**Chapter 2.1**). Accordingly, I highlight various aspects of a phylogenomic study on relatively difficult or little-known groups including suspicions of taxonomic problems or taxonomic inflation, the lack of available data in the literature, production of new reference genomes. The

genomic dataset associated with the species-level phylogenies also allowed testing for possible introgression and incomplete lineage sorting in certain parts of the phylogeny where taxonomic problems were suspected. In **Chapter 2.2**, I repeat this protocol for the second focal lineage of my PhD thesis: the tribe Troidini. In addition to providing a novel nearly complete species-level phylogeny for this iconic butterfly group, I take advantage of the study on Leptocircini to discuss differences in each phylogenetic stage for a closely-related group.

Benefiting from the data and analyses made in Chapter 2, I carried out **a comparative macroevolutionary study with the two swallowtail tribes** in **Chapter 3**. As explained above, molecular dating represents a necessary step to conduct such a study. The two species-level phylogenies are dated using a Bayesian relaxed-clock inference calibrated with secondary calibrations. Because of computational limits and methodological challenges, the whole genomic dataset could not be used to perform the molecular dating. As a consequence, I proposed an approach consisting of a random selection of 50 orthologous genes per species and replicated 10 times to better account for rate heterogeneity and take advantage of the genomic data. With the dated trees, I then aim at studying the evolutionary history of each group and comparing their biogeographical history and diversification dynamics through time. The main hypothesis tested is the effect of host plant shifts on group diversification. Other alternative hypotheses for diversification will also be examined or discussed, such as the effect of mimicry (e.g. *Paranticopsis*, *Mimoides*) or the effect of environment or geography.

References

Abu-Asab M, Chaouchi M, Amri H. 2008. Evolutionary medicine: a meaningful connection between omics, disease, and treatment. *PROTEOMICS–Clinical Applications,* **2**:122-134.

Ahrens D, Schwarzer J, Vogler AP. 2014. The evolution of scarab beetles tracks the sequential rise of angiosperms and mammals. *Proc. R. Soc. B: Biol. Sci.* **281**:20141470.

Alda F, Ludt WB, Elías DJ, McMahan CD, Chakrabarty P. 2021. Comparing ultraconserved elements and exons for phylogenomic analyses of Middle American cichlids: when data agree to disagree. *Genome Biol. Evol.* **13**:evab161.

Allio R, Scornavacca C, Nabholz B, Clamens AL, Sperling FA, Condamine FL. 2020. Whole genome shotgun phylogenomics resolves the pattern and timing of swallowtail butterfly evolution. *Syst. Biol.* **69**:38-60.

Allio R. (2021). Phylogénomique et génomique comparative chez les mammifères myrmécophages. Doctoral dissertation, Montpellier.

Allio R, Nabholz B, Wanke S, Chomicki G, Pérez-Escobar OA, Cotton AM, Clamens AL, Kergoat GJ, Sperling FAH, Condamine FL. 2021. Genome-wide macroevolutionary signatures of key innovations in butterflies colonizing new host plants. *Nat. Commun.* **12**:354.

Almeida EA, Bossert S, Danforth BN, Porto DS, Freitas FV, Davis CC, *et al.* 2023. The evolutionary history of bees in time and space. *Curr. Biol.* **33**:3409-3422.

Almeida EA, Packer L, Melo GA, Danforth BN, Cardinal SC, Quinteiro FB, Pie MR. 2019. The diversification of neopasiphaeine bees during the Cenozoic (Hymenoptera: Colletidae). *Zoologica Scripta* **48**:226-242.

Antonelli A, Sanmartín I. 2011. Why are there so many plant species in the Neotropics? *Taxon* **60**:403-414.

Antonelli A, Nylander JAA, Persson C, Sanmartín I. 2009. Tracing the impact of the Andean uplift on Neotropical plant evolution. *Proc. Natl Acad. Sci. USA* **106**:9749–9754.

Antonelli A, Zizka A, Carvalho FA, Scharn R, Bacon CD, Silvestro D, Condamine FL. 2018. Amazonia is the primary source of Neotropical biodiversity. *Proc. Natl Acad. Sci. USA* **115**:6034–6039.

Augier A. 1801. Essai d'une nouvelle Classification des Végétaux conforme à l'Ordre que la Nature paraît avoir suivi dans le Règne Végétal: d'où Résulte une Méthode qui conduit à la Connaissance des Plantes et de leur Rapports naturels Lyon: Bruyset Ainé.

Barraclough TG, Vogler AP. 2000. Detecting the geographical pattern of speciation from species-level phylogenies. *Am. Nat.* **155**:419-434.

Barthlott W, Hostert A, Kier G, Kueper W, Kreft H, Mutke J, Daud Rafiqpoor M, Sommer JH. 2007. Geographic patterns of vascular plant diversity at continental to global scales (Geographische Muster der Gefäßpflanzenvielfalt im kontinentalen und globalen Maßstab). *Erdkunde* 305-315.

Barton N, Bengtsson BO. 1986. The barrier to genetic exchange between hybridising populations. *Heredity* **57**:357-376

Beaulieu JM, O'Meara BC. 2016. Detecting hidden diversification shifts in models of trait-dependent speciation and extinction. *Syst. Biol.* **65**:583-601.

Becerra JX, Venable DL. 1999. Macroevolution of insect–plant associations: the relevance of host biogeography to host affiliation. *Proc. Natl Acad. Sci. USA* **96**:12626-12631.

Beeravolu CR, Condamine FL. 2016) An extended maximum likelihood inference of geographic range evolution by dispersal, local extinction and cladogenesis. *bioRxiv* https://doi.org/10.1101/038695

Benson RB, Butler R, Close RA, Saupe E, Rabosky DL. 2021. Biodiversity across space and time in the fossil record. *Curr. Biol.* **31**:R1225-R1236.

Benton MJ, Wilf P, Sauquet H. 2022. The Angiosperm Terrestrial Revolution and the origins of modern biodiversity. *New Phytol.* **233**:2017-2035.

Benton MJ. 2009. The Red Queen and the Court Jester: species diversity and the role of biotic and abiotic factors through time. *Science* **323**:728–732.

Berenbaum MR, Favret C, Schuler MA. 1996. On defining "key innovations" in an adaptive radiation: Cytochrome P450s and papilionidae. *Am. Nat.* **148**:S139-S155.

Blaimer BB, Santos BF, Cruaud A, Gates MW, Kula RR, Mikó I, Rasplus JY, Smith DR, Talamas EJ, Brady SG, *et al.* 2023. Key innovations and the diversification of Hymenoptera. *Nat. Commun.* **14**:1212.

Bleidorn C, Podsiadlowski L, Zhong M, Eeckhaut I, Hartmann S, Halanych KM, Tiedemann R. 2009. On the phylogenetic position of Myzostomida: can 77 genes get it wrong? *BMC Evol. Biol.* **9**:150.

Bonetta L. 2010. Whole-genome sequencing breaks the cost barrier. *Cell* **141**:917-919.

Borowiec ML, Lee EK, Chiu JC, Plachetzki DC. 2015. Extracting phylogenetic signal and accounting for bias in whole-genome data sets supports the Ctenophora as sister to remaining Metazoa. *BMC Genomics* **16**:987.

Boschman LM. 2021. Andean mountain building since the Late Cretaceous: A paleoelevation reconstruction. *Earth-Sci. Rev.* **220**:103640.

Boschman LM, Condamine FL. 2022. Mountain radiations are not only rapid and recent: Ancient diversification of South American frog and lizard families related to Paleogene Andean orogeny and Cenozoic climate variations. *Global Planet. Change* **208**:103704.

Bossert S, Wood TJ, Patiny S, Michez D, Almeida EA, Minckley RL, Packer L, Neff JL, Copeland RS, Straka J, *et al.* 2022. Phylogeny, biogeography and diversification of the mining bee family Andrenidae. *Syst. Entomol.* **47**:283-302.

Brocklehurst N, Upchurch P, Mannion PD, O'Connor J. 2012. The completeness of the fossil record of Mesozoic birds: implications for early avian evolution. *PLoS One* **7**:e39056.

Bromham L, Duchêne S, Hua X, Ritchie AM, Duchêne DA, Ho SYW. 2018. Bayesian molecular dating: opening up the black box. *Biol. Rev.* **93**:1165-1191.

Brown EE, Cashmore DD, Simmons NB, Butler RJ. 2019. Quantifying the completeness of the bat fossil record. *Palaeontology* **62**:757-776.

Cardillo M, Orme CDL, Owens IP. 2005. Testing for latitudinal bias in diversification rates: an example using New World birds. *Ecology* **86**:2278-2287.

Cardinal S, Danforth BN 2013. Bees diversified in the age of eudicots. *Proc. R. Soc. B: Biol. Sci.* **280**:20122686.

Chazot N, Willmott KR, Lamas G, Freitas AVL, Piron-Prunier F, Arias CF, Mallet J, De-Silva DL, Elias M. 2019. Renewed diversification following Miocene landscape turnover in a Neotropical butterfly radiation. *Global Ecol. Biogeogr.* **28**:1118-1132.

Chouvenc T, Šobotník J, Engel MS, Bourguignon T. 2021. Termite evolution: mutualistic associations, key innovations, and the rise of Termitidae. *Cell. Mol. Life Sci.* **78**:2749-2769.

Christenhusz MJ. 2020. On species concepts, phylogenetics and the science of natural history—three current issues facing taxonomy. *Megataxa* **1**:67-72.

Coates DJ, Byrne M, Moritz C. 2018. Genetic diversity and conservation units: dealing with the species-population continuum in the age of genomics. *Frontiers Ecol. Evol.* **6**:165.

Cohen MB, Schuler MA, Berenbaum MR. 1992. A host-inducible cytochrome P-450 from a host-specific caterpillar: molecular cloning and evolution. *Proc. Natl Acad. Sci. USA* **89**:10920-10924.

Collins NM, Morris MG. 1985. *Threatened Swallowtail Butterflies of the World*. The IUCN Red Data Book, Cambridge.

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

Condamine FL, Allio R, Reboud EL, Dupuis JR, Toussaint EFA, Mazet N, *et al.* 2023. A comprehensive phylogeny and revised taxonomy illuminate the origin and diversification of the global radiation of *Papilio* (Lepidoptera: Papilionidae). *Mol. Phylogenet. Evol.* **183**:107758.

Condamine FL, Guinot G, Benton MJ, Currie PJ. 2021. Dinosaur biodiversity declined well before the asteroid impact, influenced by ecological and environmental pressures. *Nat. Commun.* **12**:3833.

Condamine FL, Silvestro D, Koppelhus EB, Antonelli A. 2020. The rise of angiosperms pushed conifers to decline during global cooling. *Proc. Natl Acad. Sci. USA* **117**:28867-28875.

Condamine FL, Rolland J, Morlon H. 2019a. Assessing the causes of diversification slowdowns: temperature-dependent and diversity-dependent models receive equivalent support. *Ecol. Lett.* **22**:1900-1912.

Condamine FL, Romieu J, Guinot G. 2019b. Climate cooling and clade competition likely drove the decline of lamniform sharks. *Proc. Natl Acad. Sci. USA* **116**:20584-20590.

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

Condamine FL, Clapham ME, Kergoat GJ. 2016. Global patterns of insect diversification: towards a reconciliation of fossil and molecular evidence? *Sci. Rep.* **6**:19208.

Condamine FL, Toussaint EFA, Clamens AL, Genson G, Sperling FAH, Kergoat GJ. 2015. Deciphering the evolution of birdwing butterflies 150 years after Alfred Russel Wallace. *Sci. Rep.* **5**:11860.

Condamine FL, Soldati L, Clamens AL, Rasplus JY, Kergoat GJ. 2013. Diversification patterns and processes of wingless endemic insects in the Mediterranean Basin: historical biogeography of the genus *Blaps* (Coleoptera: Tenebrionidae). *J. Biogeogr.* **40**:1899-1913.

Condamine FL, Sperling FAH, Wahlberg N, Rasplus JY, Kergoat GJ. 2012. What caused the latitudinal gradient of species diversity in swallowtail butterflies? *Ecol. Lett.* **15**:267–277.

Crisp MD, Cook LG. 2009. Explosive radiation or cryptic mass extinction? Interpreting signatures in molecular phylogenies. *Evolution* **63**:2257-2265.

Cruaud A, Rønsted N, Chantarasuwan B, Chou LS, Clement WL, Couloux A, *et al.* 2012. An extreme case of plant–insect codiversification: figs and fig-pollinating wasps. *Syst. Biol.* **61**:1029-1047.

de Bruyn M, Stelbrink B, Morley RJ, Hall R, Carvalho GR, Cannon CH, van den Bergh G, Meijaard E, Metcalfe I, Boitani L, *et al.* 2014. Borneo and Indochina are major evolutionary hotspots for Southeast Asian biodiversity. *Syst. Biol.* **63**:879-901.

De Pinna MC. 1991. Concepts and tests of homology in the cladistic paradigm. *Cladistics* **7**:367-394.

De Queiroz K. 1998a. The general lineage concept of species, species criteria, and the process of speciation. In: *Endless Forms: Species and Speciation*. Oxford University Press, USA, pp. 57-75.

De Queiroz A. 1998b. Interpreting sister-group tests of key innovation hypotheses. *Syst. Biol.* **47**:710-718.

De Queiroz A. 2002. Contingent predictability in evolution: key traits and diversification. *Syst. Biol.* **51**:917-929.

De Queiroz K. 2005. Ernst Mayr and the modern concept of species. *Proc. Natl Acad. Sci. USA* **102**:6600-6607.

De Queiroz K. 2007. Species concepts and species delimitation. *Syst. Biol.* **56**:879-886.

Del Río JA, Díaz L, García-Bernal D, Blanquer M, Ortuno A, Correal E, Moraleda JM. 2014. Furanocoumarins: Biomolecules of therapeutic interest. *Studies Natl Prod. Chem.* **43**:145-195.

Delić T, Trontelj P, Rendoš M, Fišer C. 2017. The importance of naming cryptic species and the conservation of endemic subterranean amphipods. *Sci. Rep.* **7**:3391.

Delsuc F, Brinkmann H, Philippe H. 2005. Phylogenomics and the reconstruction of the tree of life. *Nat. Rev. Genet.* **6**:361–375.

Diawara MM, Trumble JT, White KK, Carson WG, Martinez LA. 1993. Toxicity of linear furanocoumarins to *Spodoptera exigua*: evidence for antagonistic interactions. *J. Chem. Ecol.* **19**:2473-2484.

Dimarco RD, Fordyce JA. 2013. Larger clutches of chemically defended butterflies reduce egg mortality: evidence from *Battus philenor*. *Ecological Entomol.* **38**:535-538.

Doolittle RF, Blombäck B. 1964. Amino-acid sequence investigations of fibrinopeptides from various mammals: evolutionary implications. *Nature* **202**:147-152.

Dos Reis M, Donoghue PC, Yang Z. 2016. Bayesian molecular clock dating of species divergences in the genomics era. *Nat. Rev. Genet.* **17**:71-80.

Drummond AJ, Ho SYW, Phillips MJ, Rambaut A. 2006. Relaxed phylogenetics and dating with confidence. *PLoS Biol.* **4**:e88.

Dunn CW, Hejnol A, Matus DQ, Pang K, Browne WE, Smith SA, Seaver E, Rouse GW, Obst M, Edgecombe GD, *et al.* 2008. Broad phylogenomic sampling improves resolution of the animal tree of life. *Nature* 452:745–749.

Edelman NB, Frandsen PB, Miyagi M, Clavijo B, Davey J, Dikow RB, García-Accinelli G, Van Belleghem SM, Patterson N, Neafsey DE, *et al.* 2019. Genomic architecture and introgression shape a butterfly radiation. *Science* **366**:594-599.

Edger PP, Heidel-Fischer HM, Bekaert M, Rota J, Glöckner G, Platts AE, *et al.* 2015. The butterfly plant arms-race escalated by gene and genome duplications. *Proc. Natl Acad. Sci. USA* **112**:8362-8366.

Ehrlich PR, Raven PH. 1964. Butterflies and plants: a study in coevolution. *Evolution* **18**:586-608.

Eichwald E. 1829. Zoologia specialis quam expositis animalibus tum vivis, tum fossilibus potissimum Rossiae in universum, et Poloniae in species, in usum lectionum publicarum in Universitate Caesarea Vilnensi habendarum. Parsprior. Propaedeuticam zoologiae atque specialem Heterozoorum expositionem continens Vilnae: Josephus Zawadzki.

Elias M, Joron M, Willmott K, Silva-Brandão KL, Kaiser V, Arias CF, Gomez Piñerez LM, Uribe S, Brower AVZ, Freiras AVL, *et al.* 2009. Out of the Andes: patterns of diversification in clearwing butterflies. *Mol. Ecol.* **18**:1716-1729.

Esquerré D, Donnellan S, Brennan IG, Lemmon AR, Moriarty Lemmon E, Zaher H, Grazziotin FG, Keogh JS. 2020. Phylogenomics, biogeography, and morphometrics reveal rapid phenotypic evolution in pythons after crossing Wallace's line. *Syst. Biol.* **69**:1039-1051.

Eytan RI, Evans BR, Dornburg A, Lemmon AR, Lemmon EM, Wainwright PC, Near TJ. 2015. Are 100 enough? Inferring acanthomorph teleost phylogeny using Anchored Hybrid Enrichment. *BMC Evol. Biol.* **15**:1-20.

Farrell BD. 1998. "Inordinate Fondness" explained: why are there So many beetles? *Science* **281**:555-559.

Felsenstein J. 1985. Confidence limits on phylogenies: an approach using the bootstrap. *Evolution* **39**:783-791.

Fernández R, Gabaldon T, Dessimoz C. 2020. Orthology: definitions, prediction, and impact on species phylogeny inference. In: *Phylogenetics in the Genomic Era* (Eds. Scornavacca C, Delsuc D, Galtier N). Chapter 2–4. No commercial publisher. Authors open access book.

Flouri T, Jiao X, Rannala B, Yang Z. 2020. A Bayesian implementation of the multispecies coalescent model with introgression for phylogenomic analysis. *Mol. Biol. Evol.* **37**:1211-1223

Fontaine MC, Pease JB, Steele A, Waterhouse RM, Neafsey DE, Sharakhov IV, Jiang X, Hall AB, Catteruccia F, Kakani E, Mitchell SN, Wu YC, Smith HA, Love RR, Lawniczak MK, Slotman MA, Emrich SJ, Hahn MW, Besansky NJ. 2015. Extensive introgression in a malaria vector species complex revealed by phylogenomics. *Science* **347**:1258524.

Fordyce JA. 2003. Aggregative feeding of pipevine swallowtail larvae enhances hostplant suitability. *Oecologia* **135**:250-257.

Formenti G, Theissinger K, Fernandes C, Bista I, Bombarely A, Bleidorn C, Ciofi C, Crottini A, Godoy JA, Höglund J, *et al.* 2022. The era of reference genomes in conservation genomics. *Trends Ecol. Evol.* **37**:197–202.

Funk DJ, Omland KE. 2003. Species-level paraphyly and polyphyly: frequency, causes, and consequences, with insights from animal mitochondrial DNA. *Annu. Rev. Ecol. Evol. Syst.* **34**:397-423.

Funk VA. 1995. Cladistic methods. In: *Hawaiian Biogeography: Evolution on a Hot Spot Archipelago*. (eds. Funk VA, Wagner WL). Washington, Smithsonian Institution Press, pp. 30-38.

Galtier N. 2004. Sampling properties of the bootstrap support in molecular phylogeny: influence of nonindependence among sites. *Syst. Biol.* **53**:38-46.

Galtier N. 2019. Delineating species in the speciation continuum: A proposal. *Evol. Appli.* **12**:657-663.

Galtier N. 2023. Phylogenetic conflicts: distinguishing gene flow from incomplete lineage sorting. *bioRxiv*, https://doi.org/10.1101/2023.07.06.547897

Garner BA, Hand BK, Amish SJ, Bernatchez L, Foster JT, Miller KM, Morin PA, Narum SR, O'Brien SJ, Roffler G, et al. 2016. Genomics in conservation: case studies and bridging the gap between data and application. *Trends Ecol Evol.* **31**:81–83.

Gaston KJ. 2000. Global patterns in biodiversity. *Nature* **405**:220–227.

Gentry AH. 1982. *Patterns of Neotropical Plant Species Diversity*. In: *Evolutionary Biology* (ed. by Hecht MK, Wallace B, Prance GT), Springer US, Boston, MA, pp. 1–84.

Goodall-Copestake WP, Harris DJ, Hollingsworth PM. 2009. The origin of a mega-diverse genus: dating *Begonia* (Begoniaceae) using alternative datasets, calibrations and relaxed clock methods. *Bot. J. Linn. Soc.* **159**:363–380.

Goodwin S, McPherson JD, McCombie WR. 2016. Coming of age: ten years of next-generation sequencing technologies. *Nat. Rev. Genet.* **17**:333-351.

Graur D, Martin W. 2004. Reading the entrails of chickens: molecular timescales of evolution and the illusion of precision. *Trends Genet.* **20**:80-86.

Griese E, Caarls L, Bassetti N, Mohammadin S, Verbaarschot P, Bukovinszkine'Kiss G, *et al.* 2021. Insect egg-killing: a new front on the evolutionary arms-race between brassicaceous plants and pierid butterflies. *New Phytol.* **230**:341-353.

Grimaldi D, Engel MS. 2005. *Evolution of the Insects*. Cambridge University Press.

Guo Z, Flannery-Sutherland JT, Benton MJ, Chen ZQ. 2023. Bayesian analyses indicate bivalves did not drive the downfall of brachiopods following the Permian-Triassic mass extinction. *Nat. Commun.* **14**:5566.

Hamm CA, Fordyce JA. 2016. Greater host breadth still not associated with increased diversification rate in the Nymphalidae-A response to Janz et al. *Evolution* **70**:1156-1160.

Hansen CC, Nelson DR, Møller BL, Werck-Reichhart D. 2021. Plant cytochrome P450 plasticity and evolution. *Mol. Plant* **14**:1244-1265.

Heidel-Fischer HM, Kirsch R, Reichelt M, Ahn SJ, Wielsch N, Baxter SW, *et al.* 2019. An insect counteradaptation against host plant defenses evolved through concerted neofunctionalization. *Mol. Biol. Evol.* **36**:930-941.

Hejnol A, Obst M, Stamatakis A, Ott M, Rouse GW, Edgecombe GD, Martinez P, Baguñà J, Bailly X, Jondelius U., *et al.* 2009. Assessing the root of bilaterian animals with scalable phylogenomic methods. *Proc. R. Soc. B: Biol. Sci.* 276:4261-4270.

Helmstetter AJ, Glemin S, Käfer J, Zenil-Ferguson R, Sauquet H, de Boer H, Dagallier LPMJ, Mazet N, Reboud EL, Couvreur TLP, *et al.* 2022. Pulled diversification rates, lineages-through-time plots, and modern macroevolutionary modeling. *Syst. Biol.* **71**:758-773.

Hennig W. 1950. *Grundzüge einer Theorie der Phylogenetischen Systematik*. Berlin: Deutscher Zentralverlag.

Hennig W. 1966. *Phylogenetic Systematics*. Urbana, Illinois, USA: University of Illinois Press.

Hennig W. 1969. Die *Stammesgeschichte der Insekten*. Frankfurt am Main: Waldemar Kramer.

Hey J. 2001. The mind of the species problem. *Trends Ecol. Evol.* **16**:326-329.

Hillebrand H. 2004. On the generality of the latitudinal diversity gradient. *Am. Nat.* **163**:192-211.

Hinchliff CE, Smith SA, Allman JF, Burleigh JG, Chaudhary R, Coghill LM, *et al.* 2015. Synthesis of phylogeny and taxonomy into a comprehensive tree of life. *Proc. Natl Acad. Sci. USA* **112**:12764-12769.

Ho SYW. 2020. *The Molecular Evolutionary Clock*. Cham: Springer.

Ho SYW, Phillips MJ. 2009. Accounting for calibration uncertainty in phylogenetic estimation of evolutionary divergence times. *Syst. Biol.* **58**:367–380.

Ho SYW, Saarma U, Barnett R, Haile J, Shapiro B. 2008. The effect of inappropriate calibration: three case studies in molecular ecology. *PLoS One* **3**:e1615.

Hobolth A, Dutheil JY, Hawks J, Schierup MH, Mailund T. 2011. Incomplete lineage sorting patterns among human, chimpanzee, and orangutan suggest recent orangutan speciation and widespread selection. *Genome Res.* **21**: 349-356.

Höhna S, Landis MJ, Heath TA, Boussau B, Lartillot N, Moore BR, Huelsenbeck JP, Ronquist F. 2016. RevBayes: Bayesian phylogenetic inference using graphical models and an interactive model-specification language. *Syst. Biol.* **65**:726–736.

Hoorn C, Wesselingh FP, ter Steege H, Bermudez MA, Mora A, Sevink J, *et al.* (2010) Amazonia through time: Andean uplift, climate change, landscape evolution, and biodiversity. *Science* **330**:927–931.

Hoorn C, Wesselingh F. (Eds.). 2011. *Amazonia: Landscape and Species Evolution: A Look Into the Past*. John Wiley and Sons.

Howard DJ, Berlocher SH. (Eds.). 1998. *Endless Forms: Species and Speciation*. Oxford University Press, USA.

Hug LA, Baker BJ, Anantharaman K, Brown CT, Probst AJ, Castelle CJ, Butterfield CN, Hernsdorf AW, Amano Y, Ise K, *et al.* 2016. A new view of the tree of life. *Nat. Microbiol.* **1**:16048.

Hughes C, Eastwood R. 2006. Island radiation on a continental scale: exceptional rates of plant diversification after uplift of the Andes. *Proc. Natl Acad. Sci. USA* **103**:10334-10339.

Hung CF, Prapaipong H, Berenbaum MR, Schuler MA. 1995. Differential induction of cytochrome P450 transcripts in *Papilio polyxenes* by linear and angular furanocoumarins. *Insect Biochem. Mol. Biol.* **25**:89-99.

Inoue J, Donoghue PCJ, Yang Z. 2010. The impact of the representation of fossil calibrations on Bayesian estimation of species divergence times. *Syst. Biol.* **59**:74–89.

Jaenike J. 1990. Host specialization in phytophagous insects. *Annu. Rev. Ecol. Syst.* **21**:243-273.

Janz N, Nyblom K, Nylin S. 2001. Evolutionary dynamics of host-plant specialization: A case study of the tribe Nymphalini. *Evolution* **55**:783-796.

Janz N, Nylin S, Wahlberg N. 2006. Diversity begets diversity: host expansions and the diversification of plant-feeding insects. *BMC Evol. Biol*. **6**:4.

Jardine N. 1969. The observational and theoretical components of homology: a study on the morphology of the dermal skull-roofs of rhidpidistian fishes. *Biol. J. Linn. Soc.* **1**:327–361.

Jennings WB, Edwards SV. 2005. Speciational history of Australian grass finches (*Poephila*) inferred from thirty gene trees. *Evolution* **59**:2033-2047.

Jouault C, Nel A, Perrichot V, Legendre F, Condamine FL. 2022. Multiple drivers and lineage-specific insect extinctions during the Permo–Triassic. *Nat. Commun.* **13**:7512.

Jousselin E, Elias M. 2019. Testing host-plant driven speciation in phytophagous insects: a phylogenetic perspective. *PCi Evol. Biol.* https://hal.science/hal-02400722.

Kawahara AY, Plotkin D, Espeland M, Meusemann K, Toussaint EFA, Donath A, *et al.* 2019. Phylogenomics reveals the evolutionary timing and pattern of butterflies and moths. *Proc. Natl Acad. Sci. USA* **116**:22657-22663.

Kawahara AY, Storer C, Carvalho APS, Plotkin DM, Condamine FL, Braga MP, *et al.* 2023. A global phylogeny of butterflies reveals their evolutionary history, ancestral hosts and biogeographic origins. *Nat. Ecol. Evol.* **7**:903-913.

Kergoat GJ, Condamine FL, Toussaint EF, Capdevielle-Dulac C, Clamens AL, Barbut J, Goldstein PZ, Le Ru B. 2018. Opposite macroevolutionary responses to environmental changes in grasses and insects during the Neogene grassland expansion. *Nat. Commun.* **9**:5089.

Kingston SE, Adams LD, Rosel PE. 2009. Testing mitochondrial sequences and anonymous nuclear markers for phylogeny reconstruction in a rapidly radiating group: molecular systematics of the Delphininae (Cetacea: Odontoceti: Delphinidae). *BMC Evol. Biol.* **9**:1-19.

Kjer KM, Simon C, Yavorskaya M, Beutel RG. 2016.. Progress, pitfalls and parallel universes: a history of insect phylogenetics. *J. R. Soc. Interface* **13**:20160363.

Kubo T, Iwasa Y. 1995. Inferring the rates of branching and extinction from molecular phylogenies. *Evolution.* **49**:694-704.

Kumar S, Filipski AJ, Battistuzzi FU, Kosakovsky Pond SL, Tamura K. 2012. Statistics and truth in phylogenomics. *Mol. Biol. Evol.* **29**:457-472.

Labandeira CC, Sepkoski Jr, JJ. 1993. Insect diversity in the fossil record. *Science* **261**:310-315.

Lagomarsino LP, Condamine FL, Antonelli A, Mulch A, Davis CC. 2016 The abiotic and biotic drivers of rapid diversification in Andean bellflowers (Campanulaceae). *New Phytol.* **210**:1430–1442.

Lamarck J-B. 1815. Histoire Naturelle des Animaux sans Vertèbres, présentant les Caractères Généraux et Particuliers de ces Animaux, leur Distribution, leurs Classes, leurs Familles, leurs Genres, et la Citation des principales Espèces qui s'y Rapportent; Précédée d'une Introduction offrant la Détermination des Caractères essentiels de l'Animal, sa Distinction du Végétal et des autres Corps naturels, enfin, l'Exposition des Principes Fondamentaux de la Zoologie. Tome Premier Paris: Verdière

Legendre F, Condamine FL. 2018. When Darwin's special difficulty promotes diversification in insects. *Syst. Biol.* **67**:873-887.

Letsch H, Balke M, Toussaint EFA, Narakusumo RP, Fiedler K, Riedel A. 2020. Transgressing Wallace's Line brings hyperdiverse weevils down to earth. *Ecography* **43**:1329-1340.

Letsch H, Gottsberger B, Metzl C, Astrin J, Friedman AL, McKenna DD, Fiedler K. 2018. Climate and host-plant associations shaped the evolution of ceutorhynch weevils throughout the Cenozoic. *Evolution* **72**:1815-1828.

Levit GS, Hossfeld U. 2019. Ernst Haeckel in the history of biology. *Curr. Biol.* **29**:R1276-R1284.

Li W, Berenbaum MR, Schuler MA. 2001. Molecular analysis of multiple CYP6B genes from polyphagous *Papilio* species. *Insect Biochem. Mol. Biol.* **31**:999-1011.

Li W, Schuler MA, Berenbaum MR. 2003. Diversification of furanocoumarin-metabolizing cytochrome P450 monooxygenases in two papilionids: specificity and substrate encounter rate. *Proc. Natl Acad. Sci. USA* **100**:14593-14598.

Liow LH, Reitan T, Harnik PG. 2015. Ecological interactions on macroevolutionary time scales: clams and brachiopods are more than ships that pass in the night. *Ecol. Lett.* **18**:1030-1039.

Loire E, Galtier N. 2021. Lacking conservation genomics in the giant Galápagos tortoise. *Peer Community J.* **1**:e18

Louca S, Pennell MW. 2020. Extant timetrees are consistent with a myriad of diversification histories. *Nature* **580**:502–505.

Lowry DB, Modliszewski JL, Wright KM, Wu CA, Willis JH. 2008. The strength and genetic basis of reproductive isolating barriers in flowering plants. *Philos. Trans. R. Soc. B: Biol. Sci.* **363**:3009-3021.

Madriñán S, Cortés AJ, Richardson JE. 2013. Páramo is the world's fastest evolving and coolest biodiversity hotspot. *Frontiers Genet.* **4**:192.

Magallón S, Hilu KW, Quandt D. 2013. Land plant evolutionary timeline: gene effects are secondary to fossil constraints in relaxed clock estimation of age and substitution rates. *Am. J. Bot.* **100**:556–573.

Maliet O, Hartig F, Morlon H. 2019. Amodel with many small shifts for estimating species-specific diversification rates. *Nat. Ecol. Evol.* **3**:1086–1092.

Mallet J, Besansky N, Hahn MW. 2016. How reticulated are species? *BioEssays* **38**:140–149.

Manktelow M. 2010. History of taxonomy. Lecture from Dept. of Syst. Biol., Uppsala University, 29.

Mannion PD, Upchurch P, Benson RB, Goswami A. 2014. The latitudinal biodiversity gradient through deep time. *Trends Ecol. Evol.* **29**:42-50.

Mardis ER. 2017. DNA sequencing technologies: 2006–2016. *Nat. Protocols* **12**:213-218.

Margoliash E, Smith EL. 1965. Structural and functional aspects of cytochrome c in relation to evolution. In Evolving genes and proteins (pp. 221-242). Academic Press.

Matzke NJ. 2014. Model selection in historical biogeography reveals that founder-event speciation is a crucial process in island clades. *Syst. Biol.*, **63**:951-970.

May MR, Höhna S, Moore BR. 2016. A Bayesian approach for detecting the impact of mass-extinction events on molecular phylogenies when rates of lineage diversification may vary. *Meth. Ecol. Evol.* **7**:947-959.

Mayden RL. 1997. A hierarchy of species concepts: the denouement in the saga of the species problem.

Mayhew PJ. 2007. Why are there so many insect species? Perspectives from fossils and phylogenies. *Biol. Rev.* **82**:425-454.

Mayr E. 1982. The growth of biological thought: Diversity, evolution, and inheritance. Harvard University Press.

Mazet N, Fabre P-H, Morlon H, Condamine FL. 2023. Estimating cladespecific diversification rates and palaeodiversity dynamics from reconstructed phylogenies. *Meth. Ecol. Evol.* doi.org/10.1101/2022.05.10.490920.

McGuire JA, Witt CC, Remsen JV, Corl A, Rabosky DL, Altshuler DL, Dudley R. 2014. Molecular phylogenetics and the diversification of hummingbirds. *Curr. Biol.* **24**:910-916.

McPeek MA. 2008. The ecological dynamics of clade diversification and community assembly. *Am. Nat*. **172**:E270-E284.

Meleshko O, Martin MD, Korneliussen TS, Schröck C, Lamkowski P, Schmutz J, Healey A, Piatkowski BT, Shaw AJ, Weston DJ, *et al.* 2021. Extensive genome-wide phylogenetic discordance is due to incomplete lineage sorting and not ongoing introgression in a rapidly radiated bryophyte genus. *Mol. Biol. Evol.* **38**:2750-2766.

Mendel G. 1866. Versuche uber pflanzen-hybriden, Verhandlungen des naturforschenden Vereins in Brunn fur 4 .

Meredith RW, Janečka JE, Gatesy J, Ryder OA, Fisher CA, Teeling EC, Goodbla A, Eizirik E, Simão TLL, Stadler T, Rabosky DL, *et al.* 2011. Impacts of the Cretaceous terrestrial revolution and KPg extinction on mammal diversification. *Science* **334**:521–524.

Meseguer AS, Michel A, Fabre PH, Escobar OAP, Chomicki G, Riina R, Antonelli A, Antoine PO, Delsuc F, Condamine FL. 2022. Diversification dynamics in the Neotropics through time, clades, and biogeographic regions. *eLife* **11**:e74503.

Meseguer AS, Condamine FL. 2020. Ancient tropical extinctions at high latitudes contributed to the latitudinal diversity gradient. *Evolution* **74**:1966-1987.

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

Meyer BS, Matschiner M, Salzburger W. 2017. Disentangling incomplete lineage sorting and introgression to refine species-tree estimates for lake tanganyika cichlid fishes. *Syst. Biol.* **66**:531-550.

Minh BQ, Hahn MW, Lanfear R. 2020. New methods to calculate concordance factors for phylogenomic datasets. *Mol. Biol. Evol.* **37**:2727-2733.

Minh BQ, Nguyen MAT, Von Haeseler A. 2013. Ultrafast approximation for phylogenetic bootstrap. *Mol. Biol. Evol.* **30**:1188-1195.

Mittelbach GG, Schemske DW, Cornell HV, Allen AP, Brown JM, Bush MB, Harrison SP, Hurlbert AH, Knowlton N, Lessios HA, *et al.* 2007. Evolution and the latitudinal diversity gradient: speciation, extinction and biogeography. *Ecol. Lett.* **10**:315-331.

Mitter C, Farrell B, Wiegmann B. 1988. The phylogenetic study of adaptive zones: has phytophagy promoted insect diversification? *Am. Nat.* **132**:107-128.

Mittermeier RA, Turner WR, Larsen FW, Brooks TM, Gascon C. 2011 *Global Biodiversity Conservation: The Critical Role of Hotspots*. *Biodiversity Hotspots* (ed. by Zachos FE, Habel JC), pp. 3–22. Springer Berlin Heidelberg, Berlin, Heidelberg.

Morin PA, Archer FI, Avila CD, Balacco JR, Bukhman YV, Chow W, Fedrigo O, Formenti G, Fronczek JA, Fungtammasan A, *et al.* 2021. Reference genome and demographic history of the most endangered marine mammal, the vaquita. *Mol. Ecol. Res.* **21**:1008–1020.

Morlon H, Robin S, Hartig F. 2022. Studying speciation and extinction dynamics from phylogenies: addressing identifiability issues. *Trends Ecol. Evol.* **37**:497-506.

Morlon H, Parsons TL, Plotkin JB. 2011. Reconciling molecular phylogenies with the fossil record. *Proc. Natl. Acad. Sci. USA* **108**:16327–16332.

Morlon H. 2014. Phylogenetic approaches for studying diversification. *Ecol. Lett.* **17**:508-525.

Moroz L. L., Kocot K. M., Citarella M. R., Dosung S., Norekian T. P., Povolotskaya I. S., Grigorenko A. P., Dailey C., Berezikov E., Buckley K. M., *et al.* 2014. The ctenophore genome and the evolutionary origins of neural systems. *Nature* **510**:109-114.

Myers N, Mittermeier RA, Mittermeier CG, da Fonseca GAB, Kent J. 2000 Biodiversity hotspots for conservation priorities. *Nature* **403**:853–858.

Nakae M. 2021. Papilionidae of the World. Tokyo: Roppon-Ashi Entomological Books.

Nicholson DB, Ross AJ, Mayhew PJ. 2014. Fossil evidence for key innovations in the evolution of insect diversity. *Proc. R. Soc. B: Biol. Sci.* **281**:20141823.

Nixon KC, Carpenter JM. 2012. On homology. *Cladistics* **28**:160-169.

Nosil P. 2012. Ecological speciation. Oxford: Oxford University Press.

Nyakatura K, Bininda-Emonds OR. 2012. Updating the evolutionary history of Carnivora (Mammalia): a new species-level supertree complete with divergence time estimates. *BMC Biol.* **10**:1-31.

Nylin S, Wahlberg N. 2008. Does plasticity drive speciation? Host-plant shifts and diversification in nymphaline butterflies (Lepidoptera: Nymphalidae) during the tertiary. *Biol. J. Linn. Soc.* **94**:115-130.

O'Keefe FR, Dunn RE, Weitzel EM, Waters MR, Martinez LN, Binder WJ, Cohen JE, Meachen JA, Desantis LRG, Kirby ME, *et al.* 2023. Pre–Younger Dryas megafaunal extirpation at Rancho La Brea linked to fire-driven state shift. *Science* **381**:eabo3594.

Oberprieler R, Marvaldi A, Anderson RS. 2007. Weevils, weevils, weevils everywhere. *Zootaxa* **1668**:491-520.

Okamura YU, Sato AI, Tsuzuki N, Sawada Y, Hirai MY, Heidel-Fischer H, *et al.* 2019. Differential regulation of host plant adaptive genes in Pieris butterflies exposed to a range of glucosinolate profiles in their host plants. *Sci. Rep.* **9**:7256.

Padial JM, De la Riva I. 2021. A paradigm shift in our view of species drives current trends in biological classification. *Biol. Rev.* **96**:731-751.

Padial JM, Miralles A, De la Riva I, Vences M 2010. The integrative future of taxonomy. *Front. Zool.* **7**:1-14.

Pallas PS. 1766. Elenchus zoophytorum sistens generum adumbrationes generaliores et specierum cognitarum succinctas descriptiones cum selectis auctorum synonymis Hagae-Comitum [The Hague]: Petrum van Cleef.

Parker J, Tsagkogeorga G, Cotton JA, Liu Y, Provero P, Stupka E, Rossiter SJ. 2013. Genome-wide signatures of convergent evolution in echolocating mammals. *Nature* **502**:228–231.

Partha R, Chauhan BK, Ferreira Z, Robinson JD, Lathrop K, Nischal KK, *et al.* 2017. Subterranean mammals show convergent regression in ocular genes and enhancers, along with adaptation to tunneling. *eLife* **6**:e25884.

Paterson JR, Edgecombe GD, Lee MS. 2019. Trilobite evolutionary rates constrain the duration of the Cambrian explosion. *Proc. Natl Acad. Sci. USA* **116**:4394-4399.

Pease JB, Haak DC, Hahn MW, Moyle LC. 2016. Phylogenomics reveals three sources of adaptive variation during a rapid radiation. *PLoS Biol.* **14**:e1002379

Peccoud J, Ollivier A, Plantegenest M, Simon JC. 2009. A continuum of genetic divergence from sympatric host races to species in the pea aphid complex. *Proc. Natl Acad. Sci. USA* **106**:7495-7500.

Pena C, Espeland M. 2015. Diversity dynamics in Nymphalidae butterflies: effect of phylogenetic uncertainty on diversification rate shift estimates. *PLoS One* **10**:e0120928.

Pérez-Escobar OA, Chomicki G, Condamine FL, Karremans AP, Bogarín D, Matzke NJ, Silvestro D, Antonelli, A. 2017. Recent origin and rapid speciation of Neotropical orchids in the world's richest plant biodiversity hotspot. *New Phytol.* **215**:891-905.

Phillimore AB, Price TD. 2008. Density-dependent cladogenesis in birds. *PLoS Biol.* **6**:e71.

Philippe H, Blanchette M. 2007. Overview of the first phylogenomics conference. *BMC Evol. Biol.* **7**:S1.

Philippe H, Poustka AJ., Chiodin M, Hoff KJ, Dessimoz C, Tomiczek B, Schiffer PH, Müller S, Domman D, Horn M, Kuhl H, *et al.* 2019. Mitigating anticipated effects of systematic errors supports sister-group relationship between Xenacoelomorpha and Ambulacraria. *Curr. Biol.* **29**:1818-1826.

Pinto CF, Troncoso AJ, Urzua A, Niemeyer HM. 2009. Aristolochic acids affect the feeding behaviour and development of *Battus polydamas archidamas* larvae (Lepidoptera: Papilionidae: Troidini). *Eur. J. Entomol.* **106**:357-361.

Rabosky D.L. 2014. Automatic detection of key innovations, rate shifts, and diversity-dependence on phylogenetic trees. *PLoS One* **9**:e89543.

Rabosky DL, Santini F, Eastman JM, Smith SA, Sidlauskas B, Chang J, Alfaro ME. 2013. Rates of speciation and morphological evolution are correlated across the largest vertebrate radiation. *Nat. Commun.* **4**:1958

Rabosky DL. 2017. Phylogenetic tests for evolutionary innovation: the problematic link between key innovations and exceptional diversification. *Philos. Trans. R. Soc. B: Biol. Sci.* **372**:20160417.

Rabosky DL, Chang J, Title PO, Cowman PF, Sallan L, Friedman M, *et al.* (2018). An inverse latitudinal gradient in speciation rate for marine fishes. *Nature* **559**:392-395.

Ragan MA 2009. Trees and networks before and after Darwin. *Biol. Direct* **4**:1-38.

Rainford JL, Mayhew PJ. 2015. Diet evolution and clade richness in Hexapoda: a phylogenetic study of higher taxa. *Am. Nat.* **186**:777-791.

Rainford JL, Hofreiter M, Nicholson DB, Mayhew PJ. 2014. Phylogenetic distribution of extant richness suggests metamorphosis is a key innovation driving diversification in insects. *PLoS One* **9**:e109085.

Ramirez-Barahona S, Sauquet H, Magallon S. 2020. The delayed and geographically heterogeneous diversification of flowering plant families. *Nat. Ecol. Evol.* **4**:1232-1238.

Rannala B, Edwards S, Leaché A. 2020. The multi-species coalescent model and species tree inference. In: *Phylogenetics in the Genomic Era* (Eds. Scornavacca C, Delsuc D, Galtier N). Chapter 3–3. No commercial publisher. Authors open access book.

Rannala B, Yang Z. 2003. Bayes estimation of species divergence times and ancestral population sizes using DNA sequences from multiple loci. *Genetics* **164**:1645-1656.

Raven PH, Berlin B, Breedlove DE. 1971. The origins of taxonomy: A review of its historical development shows why taxonomy is unable to do what we expect of it. *Science* **174**:1210-1213.

Ree RH, Moore BR, Webb CO, Donoghue MJ. 2005. A likelihood framework for inferring the evolution of geographic range on phylogenetic trees. *Evolution* **59**:2299-2311.

Ree RH, Smith SA. 2008. Maximum likelihood inference of geographic range evolution by dispersal, local extinction, and cladogenesis. *Syst. Biol.* **57**:4–14.

Rheindt FE, Christidis L, Norman JA. 2009. Genetic introgression, incomplete lineage sorting and faulty taxonomy create multiple cases of polyphyly in a montane clade of tyrant-flycatchers (Elaenia, Tyrannidae). *Zoologica Scripta* **38**:143-153.

Rieseberg LH, Whitton J, Gardner K. 1999. Hybrid zones and the genetic architecture of a barrier to gene flow between two sunflower species. *Genetics* **152**:713-727.

Rivas-González I, Rousselle M, Li F, Zhou L, Dutheil JY, Munch K, Shao Y, Wu D, Schierup MH, Zhang G. 2023. Pervasive incomplete lineage sorting illuminates speciation and selection in primates. *Science* **380**:eabn4409.

Rogers J, Raveendran M, Harris RA, Mailund T, Leppälä K, Athanasiadis G, Schierup MH, Cheng J, Munch K, Walker JA, *et al.* 2019. The comparative genomics and complex population history of *Papio* baboons. *Sci. Adv.* **5**:eaau6947.

Rokas A, Williams BL, King N, Carroll SB. 2003. Genome-scale approaches to resolving incongruence in molecular phylogenies. *Nature* **425**:798-804.

Ronquist F, Huelsenbeck JP. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* **19**:1572-1574.

Rosenzweig M.L. 1995. *Species Diversity in Space and Time.* Cambridge University Press.

Rougerie R, Cruaud A, Arnal P, Ballesteros-Mejia L, Condamine FL, Decaëns T, Elias M, Grey D, Hebert PDN, Kitching IJ, *et al.* 2022. Phylogenomics illuminates the evolutionary history of wild silkmoths in space and time (Lepidoptera: Saturniidae). *bioRxiv* https://doi.org/10.1101/2022.03.29.486224.

Ruane S, Pyron RA, Burbrink FT. 2011. Phylogenetic relationships of the Cretaceous frog *Beelzebufo* from Madagascar and the placement of fossil constraints based on temporal and phylogenetic evidence. *J. Evol. Biol.* **24**:274–285.

Rutschmann F. 2006. Molecular dating of phylogenetic trees: a brief review of current methods that estimate divergence times. *Divers. Distribut.* **12**:35–48.

Ryan JF. 2014. Did the ctenophore nervous system evolve independently? *Zoology* **117**:225-226.

Sahoo RK, Warren AD, Collins SC, Kodandaramaiah U. 2017. Hostplant change and paleoclimatic events explain diversification shifts in skipper butterflies (Family: Hesperiidae). *BMC Evol. Biol.* **17**:1-9.

Sanderson MJ, Thorne JL, Wikstrom N, Bremer K, 2004. Molecular evidence on plant divergence times. *Am. J. Bot.* **91**:1656–1665.

Sangster G. 2014. The application of species criteria in avian taxonomy and its implications for the debate over species concepts. *Biol. Rev.* **89**:199-214.

Sauquet H. 2013. A practical guide to molecular dating. *Comptes Rendus Palevol.* **12**:355-367.

Sauquet H, Ho SYW, Gandolfo MA, Jordan GJ, Wilf P, Cantrill DJ, Bayly MJ, Bromham L, Brown GK, Carpenter RJ, *et al.* 2012. Testing the impact of calibration on molecular

divergence times using a fossil-rich group: the case of *Nothofagus* (Fagales). *Syst. Biol.* **61**:289–313.

Schenk JJ. 2016. Consequences of secondary calibrations on divergence time estimates. *PLoS One* **11**:e0148228.

Schlick-Steiner BC, Steiner FM, Seifert B, Stauffer C, Christian E, Crozier RH. 2010. Integrative taxonomy: a multisource approach to exploring biodiversity. *Annu. Rev. Entomol.* **55**:421-438.

Scornavacca C, Delsuc F, Galtier N. 2020. *Phylogenetics in the Genomic Era.* No commercial publisher. Authors open access book. 568 pp.

Scriber JM, Tsubaki Y, Lederhouse RC. 1995. *Swallowtail Butterflies: Their Ecology and Evolutionary Biology*. Scientific Publishers, Gainesville, Florida.

Shin S, Clarke DJ, Lemmon AR, Moriarty Lemmon E, Aitken AL, Haddad S, *et al.* 2018. Phylogenomic data yield new and robust insights into the phylogeny and evolution of weevils. *Mol. Biol. Evol.* **35**:823-836.

Silvestro D, Antonelli A, Salamin N, Quental TB. 2015. The role of clade competition in the diversification of North American canids. *Proc. Natl Acad. Sci. USA* **112**:8684-8689.

Sime KR, Feeny PP, Haribal MM. 2000. Sequestration of aristolochic acids by the pipevine swallowtail, *Battus philenor* (L.): evidence and ecological implications. *Chemoecology* **10**:169-178.

Simion P, Delsuc F, Philippe H. 2020. To what extent current limits of phylogenomics can be overcome? In: *Phylogenetics in the Genomic Era* (Eds. Scornavacca C, Delsuc D, Galtier N). Chapter 2–1. No commercial publisher. Authors open access book.

Simion P, Philippe H, Baurain D, Jager M, Richter DJ, Di Franco A, Roure B, Satoh N, Queinnec E, Ereskovsky A, *et al.* 2017. A large and consistent phylogenomic dataset supports sponges as the sister group to all other animals. *Curr. Biol.* **27**:958-967.

Simon JC, d'Alencon E, Guy E, Jacquin-Joly E, Jaquiery J, Nouhaud P, *et al.* 2015. Genomics of adaptation to host-plants in herbivorous insects. *Briefings Functional Genom.* **14**:413-423.

Smith SA, Brown JW, Walker JF. 2018. So many genes, so little time: a practical approach to divergence-time estimation in the genomic era. *PLoS One* **13**:e0197433.

Sneath PHA. 1969. Recent trends in numerical taxonomy. *Taxon* **18**:14-20.

Sohn JC, Labandeira CC, Davis DR. 2015. The fossil record and taphonomy of butterflies and moths (Insecta, Lepidoptera): implications for evolutionary diversity and divergence-time estimates. *BMC Evol. Biol.* **15**:1-15.

Stadler T. 2011. Mammalian phylogeny reveals recent diversification rate shifts. *Proc. Natl Acad. Sci. USA*, **108**:6187-6192.

Stankowski S, Ravinet M. 2021a. Quantifying the use of species concepts. *Curr. Biol.* **31**:R428-R429.

Stankowski S, Ravinet M. 2021b. Defining the speciation continuum. *Evolution* **75**:1256-1273.

Steiner AM, Busching C, Vogel H, Wittstock U. 2018. Molecular identification and characterization of rhodaneses from the insect herbivore *Pieris rapae*. *Sci. Rep.* **8**:10819.

Heard SB, Hauser DL. 1995. Key evolutionary innovations and their ecological mechanisms. *Hist. Biol.* **10**:151-173.

Stork NE. 2018. How many species of insects and other terrestrial arthropods are there on Earth? *Annu. Rev. Entomol.* **63**:31-45.

Thomas GWC, Hahn MW. 2015. Determining the null model for detecting adaptive convergence from genomic data: A case study using echolocating mammals. *Mol. Biol. Evol.* **32**:1232–1236.

Thornton JW, DeSalle R. 2000. Gene family evolution and homology: genomics meets phylogenetics. *Annu. Rev. Genom. Human Genet.* **1**:41-73.

To TH, Jung M, Lycett S, Gascuel O. 2016. Fast dating using least-squares criteria and algorithms. *Syst. Biol.* **65**:82-97.

Toussaint EFA, Beutel RG, Morinière J, Jia F, Xu S, Michat MC, *et al.* 2016. Molecular phylogeny of the highly disjunct cliff water beetles from South Africa and China (Coleoptera: Aspidytidae). *Zool. J. Linn. Soc.* **176**:537-546.

Toussaint EFA, Seidel M, Arriaga-Varela E, Hájek J, Kral D, Sekerka L, *et al.* 2017. The peril of dating beetles. *Syst. Entomol.* **42**:1-10.

Tucker CM, Cadotte MW, Carvalho SB, Davies TJ, Ferrier S, Fritz SA., Grenyer R, Helmus MR, Jin LS, Mooers AO, *et al.* 2017. A guide to phylogenetic metrics for conservation, community ecology and macroecology. *Biol. Rev.* **92**:698-715.

Tyler HA, Brown KS, Wilson K. 1994. *Swallowtail Butterflies of the Americas: A Study in Biological Dynamics, Ecological Diversity, Biosystematics, and Conservation*. Scientific Publishers, Gainesville, Florida.

Van Der Valk T, Gonda CM, Silegowa H, Almanza S, Sifuentes-Romero I, Hart TB, Hart JA, Detwiler KM, Guschanski K. 2020. The genome of the endangered dryas monkey provides new insights into the evolutionary history of the vervets. *Mol. Biol. Evol.* **37**:183–194.

Wang X, Jordan IK, Mayer LW. 2015. A phylogenetic perspective on molecular epidemiology. *Mol. Medical Microbiol.* 517-536.

Wen D, Nakhleh L. 2018. Coestimating reticulate phylogenies and gene trees from multilocus sequence data. *Syst Biol.* **67**:439-457.

Westbury MV, Hartmann S, Barlow A, Wiesel I, Leo V, Welch R, Parker DM, Sicks F, Ludwig A, Dalén L, *et al.* 2018. Extended and continuous decline in effective population size results in low genomic diversity in the world's rarest hyena species, the brown hyena. *Mol. Biol. Evol.* **35**:1225–1237.

Wheat, CW, Vogel H, Wittstock U, Braby MF, Underwood D, Mitchell-Olds T. 2007. The genetic basis of a plant-insect coevolutionary key innovation. *Proc. Natl Acad. Sci. USA* **104**:20427-20431.

Willig MR, Kaufman DM, Stevens RD. 2003. Latitudinal gradients of biodiversity: pattern, process, scale, and synthesis. *Annu. Rev. Ecol. Evol. Syst.* **34**: 273-309.

Worobey M, Pekar J, Larsen BB, Nelson MI, Hill V, Joy JB, Rambaut A, Suchard MA, Wertheim JO, Lemey P. 2020. The emergence of SARS-CoV-2 in Europe and North America. *Science* **370**:564-570.

Yeates DK, Seago A, Nelson L, Cameron SL, Joseph LEO, Trueman JW. 2011. Integrative taxonomy, or iterative taxonomy? *Syst. Entomol.* **36**:209-217.

Zou Z, Zhang J. 2015. No genome-wide protein sequence convergence for echolocation. *Mol. Biol. Evol.* **32**:1237–1241.

Zuckerkandl E, Pauling L. 1965. Evolutionary divergence and convergence in proteins. In: *Evolving genes and proteins* (Eds. Bryson V, Vogel HJ). Academic Press, New York, pp. 97-166.

Zuckerkandl E, Pauling L. 1962. Molecular disease, evolution, and genetic heterogeneity. In: *Horizons in Biochemistry* (Eds. Kasha M, Pullman B). Academic Press, New York, pp. 189–225.

CHAPTER 1

Foreword

The following chapter deals with the species delimitation of *Ornithoptera alexandrae*, a Papua New Guinea birdwing butterfly of the tribe Troidini, for which two allopatric populations of uncertain status were known. As outlined in the **General Introduction**, this work was part of one of the first steps in phylogenomics: defining valid species. However, as this butterfly is listed as Critically Endangered by the IUCN and breeding programs are currently underway, this was also an important conservation matter. It is from this perspective that the article is presented.



Ornithoptera alexandrae mâle - Eliette Reboud

This article is published in *Genome Biology and Evolution* (Accepted: 24 February 2023) - https://doi.org/10.1093/gbe/evad040

Genomics, population divergence and historical demography of the world's largest and endangered butterfly, the Queen Alexandra's birdwing

Eliette L. Reboud[1]*, Benoit Nabholz[1,2], Emmanuelle Chevalier[1], Marie-ka Tilak[1], Darren Bito[3] & Fabien L. Condamine[1]*

[1] *Institut des Sciences de l'Evolution de Montpellier (Université Montpellier | CNRS | IRD | EPHE), Place Eugène Bataillon, 34095 Montpellier, France*
[2] *Institut Universitaire de France (IUF), Paris, France*
[3] *Pacific Adventist University, Private Mail Bag, BOROKO 111, National Capital District, Papua New Guinea*

Corresponding authors (*)

Eliette L. Reboud: eliette.reboud@gmail.com
Fabien L. Condamine: fabien.condamine@gmail.com

Running head

Genomics of the world's largest butterfly

Abstract

The world's largest butterfly is the microendemic Papua New Guinean *Ornithoptera alexandrae*. Despite years of conservation efforts to protect its habitat and breed this up-to-28-cm butterfly, this species still figures as endangered in the IUCN Red List and is only known from two allopatric populations occupying a total of only ~140 km². Here we aim at assembling reference genomes for this species to investigate its genomic diversity, historical demography and determining whether the population is structured, which could provide guidance for conservation programs attempting to (inter)breed the two populations. Using a combination of long and short DNA reads and RNA sequencing, we assembled six reference genomes of the tribe Troidini, with four annotated genomes of *O. alexandrae* and two genomes of related species *O. priamus* and *Troides oblongomaculatus*. We estimated the genomic diversity of the three species, and we proposed scenarios for the historical population demography using two polymorphism-based methods taking into account the characteristics of low-polymorphic invertebrates. Indeed, chromosome-scale assemblies reveal very low levels of nuclear heterozygosity across Troidini, which appears to be exceptionally low for *O. alexandrae* (lower than 0.01%). Demographic analyses demonstrate low and steadily declining *Ne* throughout *O. alexandrae* history, with a divergence into two distinct populations about 10,000 years ago. These results suggest that *O. alexandrae* distribution has been microendemic for a long time. It should also make local conservation programs aware of the genomic divergence of the two populations, which should not be ignored if any attempt is made to cross the two populations.

Key words: Conservation genomics, heterozygosity, low genetic diversity, *Ornithoptera alexandrae*, reference genome.

Significance statement

Despite its charisma, little is known about the demographic trends and taxonomic status of the two populations from the giant endangered birdwing butterfly *Ornithoptera alexandrae*. By sampling and sequencing individuals of this species and two closely related species, we study whether and how the population is structured, and we investigate the genomic diversity of the species and the "health" of their genomes and populations (e.g., demographic trend, evidence of inbreeding). Overall, the very low genomic diversity and steadily declining trend inferred by this study suggests that efforts need to be reinforced to conserve this amazing Papua New Guinean insect.

Introduction

When in January 1906, Alfred S. Meek saw an enormous butterfly flying high above him in the canopy of this forest of the Northern Province of Papua New Guinea, two days walk from the coast, he took his rifle and shot down the beast. This is what one can read in the letter he sent to his correspondent Karl Jordan at the Natural History Museum of Tring (England) (letter n°155 of Meek's communications, Meek 1906; Ackery 1997; Tennent 2021). Meek let his funder, Lord Walter Rothschild describe in 1907 for the first time *Ornithoptera alexandrae* (Papilionidae: Troidini), known as the Queen Alexandra's birdwing butterfly, based on this female whose wing still bears the stigma of this extraordinary hunt. This "trophy" and the even larger congeners that followed have become representatives of the world's largest known butterfly species to date and contribute to the continuing amazement of scientists at the incredible diversity, size, and beauty of Papua New Guinea's insects (Parsons 1992; Mitchell et al. 2016). Indeed, many naturalists have been studying *O. alexandrae*, culminating with the comprehensive review on this butterfly by Mitchell et al. (2016) that serves as a basis for this work.

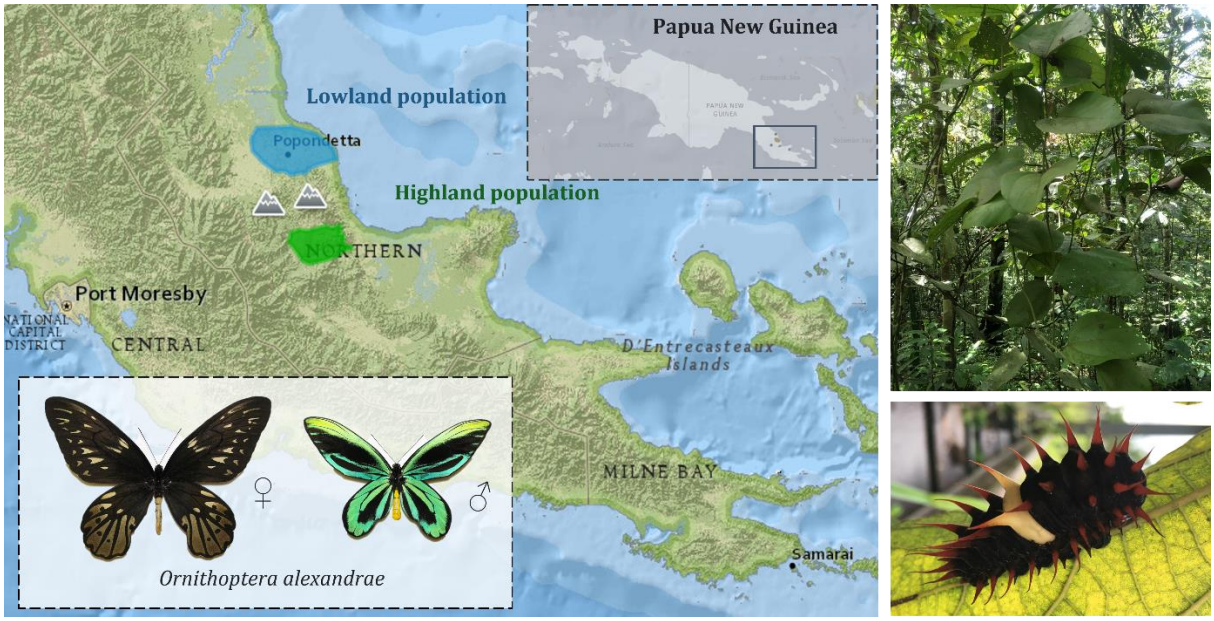


Figure 1. Left, distribution range of *Ornithoptera alexandrae*, lowland population in blue and highland population in green. Topright, *Aristolochia cf. meridionaliana* plant in its environment. Bottomright, *O. alexandrae* larva. Photos: Fabien L. Condamine. Papua New Guinea gray map is from IUCN Red List https://www.iucnredlist.org/ (Böhm, 2018), and *O. alexandrae* distribution map has been designed on MapMaker https://mapmakerclassic.nationalgeographic.org based on Mitchell et al. (2016).

As the world's largest butterfly, *Ornithoptera alexandrae* can measure up to 28-30 cm in wingspan (Mitchell et al. 2016 and references therein). *Ornithoptera alexandrae* is endemic to the Northern Province of Papua New Guinea, in a narrow range around Popondetta (Northern Province; **Figure 1**). Long-term field observations in the last decades have shown there are two recognized allopatric populations: a lowland population in Popondetta plains (≤300 m above sea level), and a highland population occurring on the relatively inaccessible Managalas Plateau about 800 m above sea level (Collins and Morris 1985; Parsons 1999; Böhm 2018). A mountain range separates the two populations, bounded in the West by Mount Lamington volcano (1700 m), and eastward to a mountain 2,140 m high. According to available data, there have been no sightings in between. The volcanic activity of Mount Lamington (last eruption in 1951 with activity until 1956; Global Volcanism Program, 2022), the flooding, drought and fires occurring in the region as well as recent logging and agricultural activities might explain today's fragmented distribution of *O. alexandrae* (Parsons 1992; Mitchell et al. 2016). The relatively small distribution range composed of two patches has been interpreted by Haugum and Low (1979) as a relict occurrence, potentially due to an evolutionary bottleneck or demographic decline. However, genetic studies are crucially lacking to assess this hypothesis.

Ornithoptera alexandrae is considered as a threatened species in the IUCN Red List (Böhm 2018). While this species is very rare over an area of occurrence of 8,710 km², its actual area of occupancy is only 128-140 km² fragmented in two populations, placing it in the Endangered category (Böhm 2018). There are doubts about the previously thought monophagy of its caterpillar due to the lack of comprehensive morphological or genetic studies on the Southeast Asian genus *Aristolochia* (Parsons 1996; Buchwalder et al. 2014; Mitchell et al. 2016). It has long been thought that there were no particular restriction due to host plant distribution to explain the local occurrence of *O. alexandrae* as there are many areas where the main larval food plants (previously thought to be *Aristolochia dielsiana*, possibly being *A. meridionaliana* in the highlands and *A. alexandriana* in the lowlands) grow prolifically. This pattern has been described in some monophagous lepidopteran species (Quinn et al. 1997) and may highlight the fact that the distribution of *O. alexandrae* species is also driven by other factors of a microclimatic, pedologic or geological nature that might limit its distribution. On the other hand, some factors represent a threat or vulnerability for the species: it is a large species with a relatively specialized ecology (larvae on a single or very few host-plants and adults in a single habitat) (Koh et al. 2004, Palash et al. 2022), it is sympatric to other *Ornithoptera* species (Koh et al. 2004), and its habitat is or has been fragmented by fires, droughts, and volcanic eruptions, and is severely affected by agriculture (Parsons 1992; Mitchell et al.

2016) and other human activities, with habitat conversion leading to a local decline in larval host vines. Accordingly, *O. alexandrae* is placed at the top of the CITES list (Appendix I). However, although its international trade is prohibited, this species is highly prized and is subject to an illegal trade that is dangerous for the demography of the species, and its survival (Mitchell et al. 2016). Thus, better monitoring of this species is recommended by the IUCN to better track population trajectories (Böhm 2018), particularly because its numbers and current trend in population dynamics are unknown.

The two allopatric populations of *O. alexandrae* are externally morphologically similar but express important biological differences, such as slight differences in size (on average 14% larger in highlands) and development time (on average 34.5% longer in highlands) (Straatman 1971; Mitchell et al. 2016). The soil and host plant eaten by the larvae might also differ in the two populations (Haugum and Low 1979; Mitchell et al. 2016), so it is unclear how divergent these two populations are. In fact, genomic research on other butterfly groups has revealed that superficial similarity in adults can hide a previously unrecognized cryptic lineage (e.g., Hebert et al. 2004; Burns et al. 2008). Knowledge of evolutionary and genetic history of the species and populations could help conservation and breeding programs to save the species. Genomics is considered as a powerful tool for studying the past and present structure and diversity of populations and brings an invaluable source of information, especially of species that are naturally rare and difficult to study (e.g., Westbury et al. 2018; Van der Valk et al. 2020; Morin et al. 2021). Genome sequencing is increasingly recognized as an important contribution to identifying, characterizing, and conserving biodiversity (Formenti et al. 2022). Reference genomes provide primary data for understanding historical demography (Morin et al. 2021), gene and trait evolution (Warren et al. 2021), or even susceptibility to inbreeding depression and accumulation deleterious mutations (Chattopadhyay et al. 2019; Van der Valk et al. 2020; Robinson et al. 2022). Genomic resources are also useful for broader studies of population structure, relatedness, and recovery potential (e.g., Garner et al. 2016; Morin et al. 2018; Tunstall et al. 2018), or for assessing correlations between current IUCN status and past demography (Nadachowska-Brzyska et al. 2015). These types of estimates (e.g. sequentially Markovian coalescent [SMC] methods, Li and Durbin 2011) have been widely used for conservation purposes for vertebrates, such as mammals (Morin et al. 2021) or birds (Nadachowska-Brzyska et al. 2015) and insect pests (Hazzouri et al. 2020; You et al. 2020). Despite the continuous increase of threatened insects (Sánchez-Bayo and Wyckhuys 2019), it has been much less used for insect conservation (but see Mikheyev et al. 2017; Podsiadlowski et al. 2021). It is indeed challenging to study the demography of invertebrates using polymorphism-based methods because the risk of violating the assumptions of SMC-type models is high. For instance, Sellinger et al. (2021) revealed that these methods of inference perform

poorly when the ratio between the recombination and mutation rates is important, therefore highlighting that the consideration of this ratio is crucial and still much too little considered in this type of analyses in the literature.

Here we perform a first genomic study of *O. alexandrae* to understand the past and present demography of this species and to bring insights into its Endangered status, which may have implications for conservation strategies. Since a local conservation program has been set up and ongoing to rear the two populations, the taxonomic status of these two populations (i.e., populations or species) may inform conservation management of this threatened species (Mitchell et al. 2016). If the two populations are too divergent, it could be complicated to breed specimens from the Managalas Plateau with specimens from the Popondetta lowlands. Given the above-mentioned biological features of this butterfly species, we aim to: (1) assemble high-quality and annotated whole genomes for the two populations, (2) assess the level of nuclear heterozygosity, (3) estimate the demographic history of the species and the two populations in relation to past environmental change and to test whether the current range of the species is relictual, and (4) provide information for the policy makers to improve their conservation strategy.

Results and Discussion

High-quality assemblies for birdwing butterflies

We collected live specimens from the two populations of *Ornithoptera alexandrae* with one adult and one caterpillar per population and sequenced the DNA combining a mean of 72.5x of long reads (Oxford Nanopore) for draft assembly, 82x of short reads (Illumina) for polishing, as well as 51.7 million cleaned RNAseq reads (8.4 Gb) for genome annotation (see *Materials and Methods*). Using Flye assembler (Kolmogorov et al. 2019) and Pilon polisher (Walker et al. 2014), we assembled the four genomes of *O. alexandrae* that range from 321 to 327 Mb, which are very contiguous with a mean of 582 contigs (ranging from 305 to 1222 contigs) and a mean N50 of 9.9 Mb (**Table 1**). Over a total of 5,286 core genes of the Lepidoptera database (odb10, Manni et al. 2021), BUSCO recovered on average 98.83% ±0.05 complete genes, 0.23% ±0.05 fragmented genes and 0.97% ±0.06 missing genes (**Table 1**). The genome size and gene completeness of our *O. alexandrae* assemblies are comparable to the genome of a related species: *Troides helena* (330 Mb, BUSCO score = 96.6%), which was assembled with similar data and methods (He et al. 2022). Furthermore, the genome size stands among the smallest within the family Papilionidae but still is 30 to 40% larger than some *Papilio* (the sister tribe of Troidini; He et al. 2022; Liu et al. 2020) illustrating the dynamic genome size evolution of the family.

Table 1. Assembly statistics for the genomes of *O. alexandrae*, *O. priamus* and *T. oblongomaculatus*. LR = long reads, SR = short reads, bp = base pairs. For BUSCO scores, S = single-copy genes, D = duplicated genes, F = fragmented genes, and M = missing genes out of 5,286 genes in odb10 lepidopteran database. Annotation was only done for *O. alexandrae* individuals.

	Ornithoptera alexandrae				*Ornithoptera priamus*	*Troides oblongomaculatus*
	Highland		Lowland			
	FC560 (♂)	FC561 (♀)	FC562 (♀)	FC563 (♀)	FC565 (♂)	FC569 (♀)
Raw data sequenced (Gb) (LR + SR)	30.98 + 33.2	31.82 + 31.0	41.18 + 33.8	29.44 + 37.6	38.99 + 36.9	38.34 + 57.4
Final mean coverage (LR + SR)	66x + 84x	59x + 79x	95x + 70x	70x + 95x	86x + 96x	87x + 137x
Assembly size (bp)	326,746,405	325,591,695	320,565,121	321,134,305	316,500,205	348,219,047
Number of contigs	333	1,222	465	305	862	465
N50 (bp)	10,703,829	7,761,365	9,990,636	11,239,331	4,951,043	5,909,187
Max length (bp)	20,033,826	12,853,001	14,559,367	13,869,660	10,532,357	13,649,974
Nucleotid assembly BUSCO score (%)	S:98.6; D:0.2; F:0.3; M:0.9	S:98.6; D:0.2; F:0.2; M:1.0	S:98.7; D:0.2; F:0.2; M:0.9	S:98.6; D:0.2; F:0.2; M:1.0	S:98.4; D:0.3; F:0.3; M:1.0	S:98.5; D:0.2; F:0.3; M:1.0
Protein coding genes (number and mean size)	17,617 6,095bp	17,449 6,082bp	16,159 6,473bp	16,508 6,325bp	-	-

After removing potential exogenous contigs of the assemblies (see *Materials and Methods*), we selected FC563 as the reference genome for further analyses, as it had the best assembly statistics and we found no evidence of contamination (i.e., bacteria, plants). We assessed the quality of this *O. alexandrae* FC563 assembly by looking at its correspondence with the reference genome of *Papilio bianor* (chromosome-level assembly, Lu et al. 2019). We found 24 contigs that match with more than 70% of the length of *P. bianor* chromosomes out of 30 chromosomes. This represents a cumulative length of 81% of the total assembly. Among those, 18 contigs have a single reciprocal match with one *P. bianor* chromosome (representing a cumulative length of 48% of the assembly). One of the most fragmented chromosomes is the Z chromosome in which 11 contigs of *O. alexandrae* FC563 assembly match to chromosome 30 (Z) of *P. bianor*. This is not surprising as FC563 is a female and, therefore, has half the coverage on the Z compared to the autosome. However, a similar analysis performed on the male FC560 led to eight contigs matching to chromosome 30 (Z) of *P. bianor*, suggesting that chromosome Z is difficult to assemble. These eight contigs linked to the Z chromosome were independently identified using coverage and heterozygosity information in FC563, representing a cumulative length of 14.1 Mb (**Table S1**). The FC563 assembly is therefore composed of 24 full-length or nearly full-length chromosomes, including 18 full-length chromosomes (**Figure S1**). Analysis of the chromosome-level synteny between *O. alexandrae* and *P. bianor* shows a high level of genomic synteny (**Figure S1**). Our results suggest that the combination of Nanopore long reads and Illumina short reads perform notably well to recover chromosome-scale assemblies as we reconstructed genome

assemblies comparable to the best assemblies of Papilionidae available so far (comparison with *Papilio bianor*, Lu et al. 2019: table 1) without relying on Hi-C techniques.

Using transcriptomic data, protein homology and *de novo* genes prediction, we annotated the genomes of the two populations (FC560 and FC653) using the MAKER pipeline (Holt and Yandell 2011). Protein predictions retrieved 17,617 genes for FC560 and 16,508 genes for FC563. We carried out BUSCO analyses with these two proteoms, and estimated 97.7% complete genes, 0.9% fragmented genes and 1.4% missing genes for FC560, and 97.2% complete genes, 1.0% fragmented genes and 1.8% missing genes for FC563. Because FC560 annotation contained more genes, it was transferred to the other two genomes of *O. alexandrae*. Overall, 33-35% of the genome was annotated as repeat sequences with mostly unclassified categories of interspersed repeats (**Table S2**). This proportion of repeats is relatively high compared to other Papilionidae genomes already available (22% in *Papilio glaucus*, 22.4% in *P. xuthus*; Cong et al. 2015; Lu et al. 2019), except for the species with larger genomes such as *Papilio bianor* (55%; Lu et al. 2019) and *Parnassius apollo* (65%; Podsiadlowski et al. 2021). This is consistent with a positive correlation between assembly size and repeat content in insects (Petersen et al. 2019; Heckenhauer et al. 2022; Sproul et al. 2022). Within *O. alexandrae*, all individuals have similar cumulative repeat size, and no difference was detected between lowland and highland individuals despite the fact that highland individuals had a slightly larger genome assembly than lowland individuals (~5 Mb, **Table S2**). These annotated genomes are available in GenBank (Bioproject PRJNA938052).

For comparison purposes, we also sequenced and assembled two other birdwing butterfly species (Troidini): *Ornithoptera priamus poseidon* and *Troides oblongomaculatus papuensis*, both ranked as Least Concern in the IUCN Red List (see *Materials and Methods*). Both genomes were very similar to the *O. alexandrae*'s ones in quality and assembly statistics (**Table 1**).

The world's largest butterfly has low levels of genomic diversity

Using short reads data and annotation with MitoFinder (Allio et al. 2020) (see *Materials and Methods*), each mitogenome was reconstructed in a single contig and we retrieved 36-37 genes (including 13 protein-coding genes and 2 rRNA), with evidence of circularization. The mitogenomic diversity (π-diversity) including coding and non-coding region (such as the D-loop) was calculated at ~0.0704% (**Table S3**). This is comparable to the low level of mitochondrial diversity for mammalian species such as the Tasmanian devil or Island fox (Westbury et al. 2018). Up to our knowledge, there is no estimation of mitochondrial diversity based on whole mitogenomes of butterflies, but there are studies estimating mitochondrial diversity relying on the cytochrome *c* oxidase subunit I (COI) DNA barcode marker (e.g., π-diversity: Mackintosh et al. 2019; haplotype diversity: Dincă et al. 2021). The mitochondrial diversity of the DNA barcode for *O. alexandrae* is ~0.076%

(**Table S3**), which ranks fifth lowest of 38 European butterflies (Mackintosh et al. 2019). However, a low mitochondrial diversity does not equate low autosomal diversity (Allio et al. 2017; Mackintosh et al. 2019).

Using mapping data on the best genome assembly of *O. alexandrae* FC563, we calculated the autosomal heterozygosity of the two populations. All four individuals of *O. alexandrae* show a heterozygosity around 0.08% (**Table S4**). Using a similar metric relying on four-fold degenerate sites from coding sequences, this heterozygosity is the lowest value observed within the butterfly studied by Mackintosh et al. (2019) (**Figure 2, Table S4**). We also performed sensitivity analyses to corroborate these results using different data types (short reads vs. long reads) and window sizes (100 kb vs. 1 Mb) and proportion of missing data (10% vs. 20%). Overall, the heterozygosity of *O. alexandrae* ranges from 0.0745 to 0.0838% (Table S4), suggesting the low estimated heterozygosity is not an artifact due to methods and/or data. Determinants of heterozygosity are still a long-term debate (Romiguier et al. 2014; Galtier and Ellegren 2016; Mackintosh et al. 2019; Buffalo 2021) and intrinsic features of *O. alexandrae* like its large size, its tropicality, or its microendemism might explain such a low

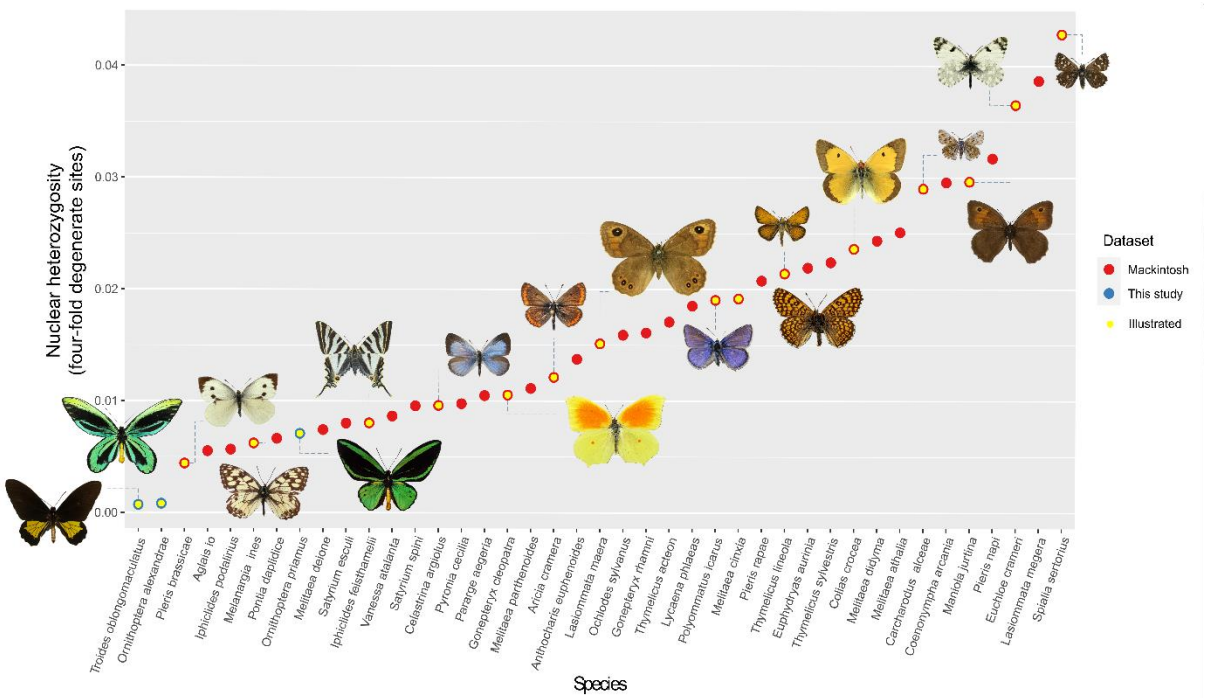


Figure 2. Level of heterozygosity for butterflies estimated with the four-fold degenerate sites (neutral diversity). Red points are values from Mackintosh et al. (2019) and blue points are from this study. Yellow point indicates illustrated species. Photos (not at scale): *T. oblongomaculatus* (CC BY 4.0, Peter Wing); *O. alexandrae* (Fabien L. Condamine); *P. brassicae* and *G. cleopatra* (CC BY-SA 3.0, Sarefo); *M. ines* (CC BY-SA 4.0, Atylotus); *O. priamus* (Eliette L. Reboud); *I. feisthamelii* and *M. cinxia* (CC BY-SA 4.0, Didier Descouens); *C. argiolus* (CC BY 3.0, Alan Cassidy); *A. cramera* and *P. icarus* (Robin Noel); *L. maera* and *C. crocea* (CC BY-SA 3.0, Vítězslav Maňák); *T. lineola* and *S. sertorius* (CC BY-NC-SA, Peter Huemer); *C. alceae* (CC BY-SA 3.0, Didier Descouens); *M. jurtina* (Public domain, Pekka Malinen); *E. crameri* (Alexander Slutsky).

heterozygosity. However, for two closely related species having similar lifestyle but larger distributional range and abundance than *O. alexandrae* (**Figure 3)**, we found that *O. priamus* has a higher heterozygosity rate (autosomal 0.433%, neutral diversity 0.708%, ca. 6 and 10 times higher) and that *T. oblongomaculatus* has an even lower heterozygosity than *O. alexandrae* (0.0737% autosomal, 0.0704% neutral). Interestingly, these levels of heterozygosity are much lower than the level estimated in the genus *Papilio* ranging from 1.0 to 2.3% although estimated with a different method (k-mer distribution; Lu et al. 2019). Accordingly, the low level of heterozygosity cannot be explained by the species' range or body size, as suggested in Mackintosh et al. (2019) (**Figure 3**).

Historical demography

Although the causes of the low genetic diversity of these birdwing butterflies are unclear, we can wonder how this translates into demographic history of the species given the threats they are experiencing. Predictions can be formulated such as a prolonged decline of effective population size such as inferred for the brown hyena and the Californian condor (Westbury et al. 2018; Robinson et al. 2021) or a low but stable effective population size such as in the vaquita porpoise (Morin et al. 2021; Robinson et al. 2022).

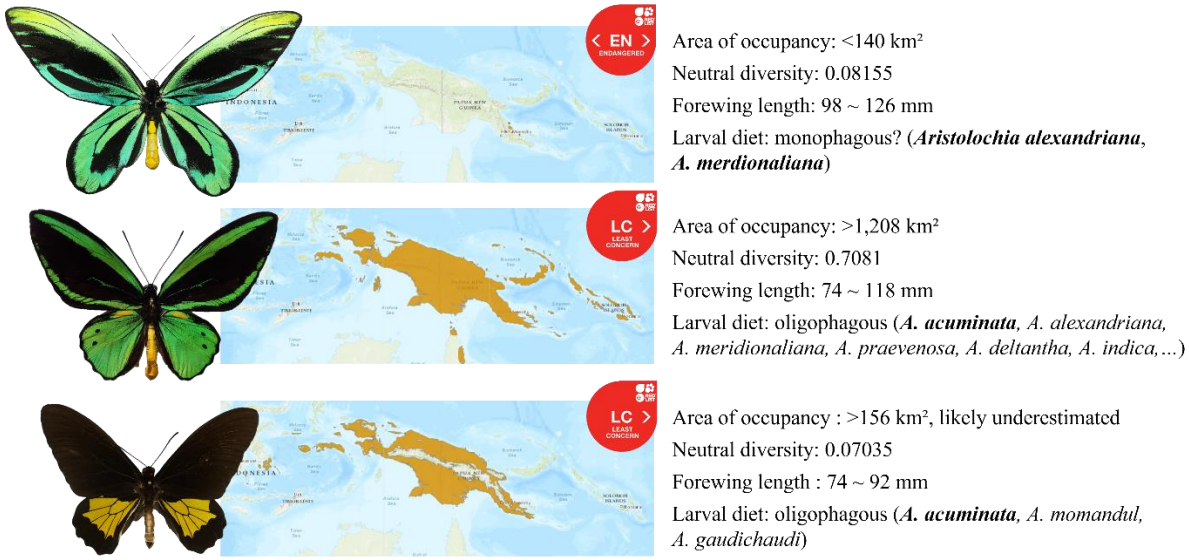


Figure 3. Comparison of *O. alexandrae*, *O. priamus* and *T. oblongomaculatus*. Distribution maps, IUCN status and areas of occupancy are from IUCN Red List https://www.iucnredlist.org/ (Böhm, 2018). Neutral diversity of *O. alexandrae* is the mean neutral diversity values of the four individuals. Forewing length is from Nakae (2021) and its larval diet is from Mitchell et al. (2016) and Böhm (2018). Larval diets should be taken with caution due to uncertainties of the *Aristolochia* taxonomy. Host plant species in bold font are known to be the main diet of the species. Photos (male specimens, relatively scaled to the mean species forewing value): *O. alexandrae* (Fabien L. Condamine), *O. priamus* (Eliette L. Reboud), *T. oblongomaculatus* (CC BY 4.0, Peter Wing NHM specimen).

To our knowledge, there are still few examples of demographic history in insects, other than pest insects (e.g., red-palm weevil, Hazzouri et al. 2020; diamondback moth, You et al. 2020; but see Walton et al. 2021; Manthey et al. 2022; García-Berro et al. 2023). Within swallowtail butterflies, the recent study of the Apollo butterfly (*Parnassius apollo*), ranked as a threatened species, showed using SNP data synchronous population declines throughout different mountain massifs despite high heterozygosity levels (Kebaïli et al. 2022).

Sequential Markovian coalescent models (e.g. PSMC, MSMC) are widely used to study the trajectory of the ancestral effective population size (*Ne*) over time. We applied the MSMC2 model to the best genome assembly of *Ornithoptera alexandrae* as well as *O. priamus* and *Troides oblongomaculatus* before focusing on the two populations of *O. alexandrae*. SMC models generally infer a *Ne* under the panmictic assumption. This strong assumption is often false in reality, and theoretical work and simulations have shown that SMC dynamics might also be caused by population structure and connectivity changes (Teixeira et al. 2021), so this should be borne in mind when interpreting these population size analyses (Bentley and Armstrong, 2021; Teixeira et al. 2021). Furthermore, it has been shown that SMC methods do not perform well when the mutation rate μ is rarer than the recombination rate r (Sellinger et al. 2021). Unfortunately, this is likely to be very common in many groups such as fungi, sea cucumbers, nematodes, corals, insects, and plants (see Wilfert et al. 2007; Stapley et al. 2017 for recombination rates and Lynch et

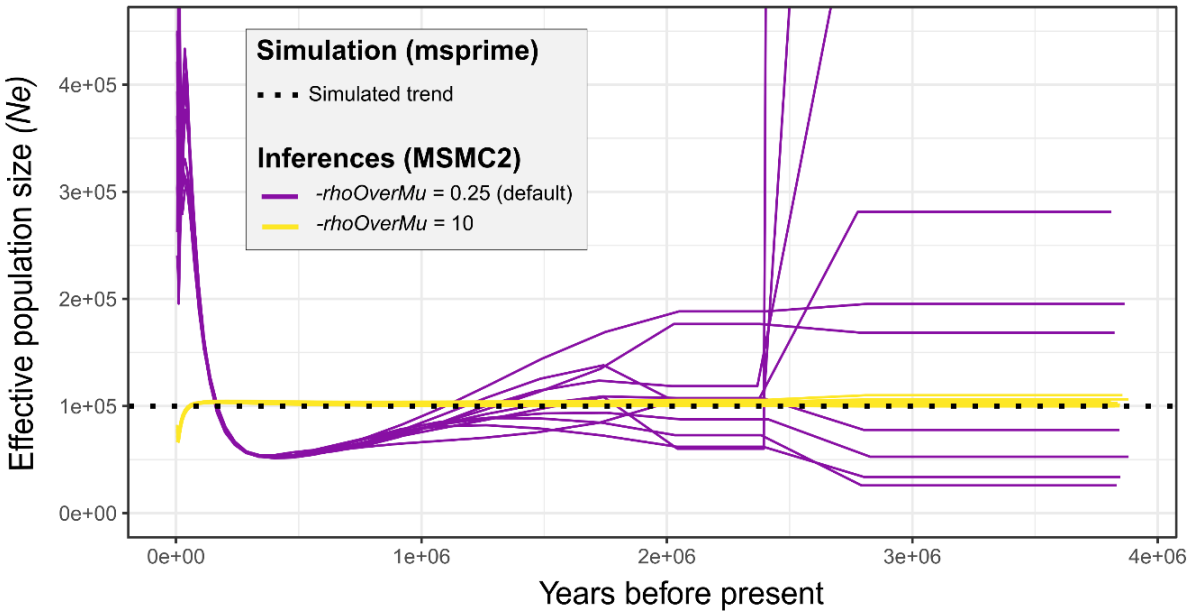


Figure 4. Performance of MSMC2 inferences on simulated data. Dotted line represents the simulated scenario produced by msprime (Stable *Ne* = 1e5, *r* =1e-8, μ = 1.319e-9). Colored lines represent the demography inferred by MSMC2 on this data with different *-rhoOverMu* settings (ten repetitions each).

al. 2016; Wang & Obbard, 2023 for mutation rates). In practice, there is concern that, at least many small non-vertebrate genomes have a much lower μ than r (Sellinger et al. 2020, Sellinger et al. 2021) whose demographic inferences would be affected and poorly addressed in the literature (e.g. coral: Fuller et al. 2020; insects: Waldvogel et al. 2017; Hazzouri et al. 2020; You et al. 2020; Wang et al. 2022; Garcia-Berro et al. 2023).

Demographic inferences on simulated data with the range of recombination and mutation rate parameters of *O. alexandrae* (i.e., r about ten times higher than μ, see *Materials and Methods*) confirmed the methodological issue, which was almost completely reduced when the model was set with an appropriate initial value of the ratio (**Fig 4**). We do not know exactly how the model behaves with an appropriate initial value on real data, but our results changed significantly when the initial value was adjusted (**Figure S2**), highlighting that this is probably a very important parameter to consider in SMC analyses and further emphasizing the importance of taking the results of any SMC analyses with caution.

Finally, MSMC2 analyses traced relatively different demographic histories for the three butterflies (**Figure 5**). The effective population size of *O. alexandrae* seems to have been at a low but continuously declining number from ~250,000 to 50,000 individuals during the last million year. However, we inferred a recent demographic change in the last 10,000 years (see below). This demography dynamic is similar to that of the vaquita porpoise, whose effective population size has never been elevated but remained stable for much of its known history and now has one of the lowest rates of heterozygosity

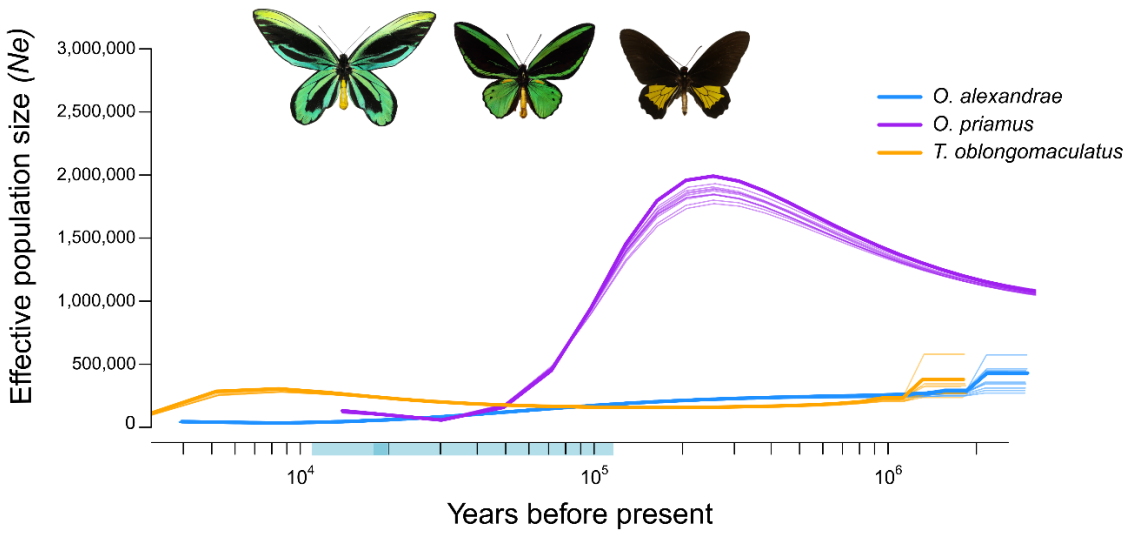


Figure 5. Estimated historical demography of three birdwing butterfly species. MSMC2 estimates of the effective population size (*Ne*) with *-rhoOverMu* = 10, for *O. alexandrae* (blue), *O. priamus* (purple) and *T. oblongomaculatus* (orange). Bootstraps are represented in clear lines. The pale blue rectangle along the time bar indicates the limits of the last glaciation period (11,700 to 115,000 years ago) with the last glacial maximum in darker blue (19-20,000 years ago). The recent present (last 4,000 years) is not represented.

known among marine mammals (Morin et al. 2021; Robinson et al. 2022). Altogether, these results suggest that the ancestral effective population of *O. alexandrae* has never been large, thus suggesting that it has been microendemic for a very long time. Interestingly, *T. oblongomaculatus* experienced a similar fairly low and regular *Ne* during most of its history with a *Ne* that culminated (at ~0.3 million individuals) around 10,000 years ago (**Figure 5**). It was followed by a decreasing trend toward the present, which seems consistent with the low rate of heterozygosity that we estimated today. In contrast, *O. priamus* shows an early peak of *Ne* (~2.0 million individuals ~250,000 years ago) followed by a severe decline until ~40,000 years ago.

In the literature, temporal variations of *Ne* are usually compared with past climatic fluctuations such as temperature and/or sea level, in particular in line with Quaternary glaciations (e.g., Nadachowska-Brzyska et al. 2015; Westbury et al. 2018; Hazzouri et al. 2020; Morin et al. 2021; Teixeira et al. 2021). Although it is tempting to associate the inferred *Ne* variations of the studied birdwing butterflies with the glaciation-interglaciation cycles, it remains difficult to extract a correlation because of climatic heterogeneity at the global scale and estimates of demographic parameters (Bentley and Armstrong, 2022). The last million years was mostly a glacial period that have also been documented in New Guinea (e.g., Chappell and Polach 1991; Barrows et al. 2011), punctuated by intermittent warming events (e.g., Eemian event ~115,000-130,000), with the last glacial period (Würm glaciation event) starting ~115,000 and ending ~11,700 years ago (last glacial maximum at 19,000-20,000 years). During the last glacial maximum, temperatures were 5°C colder (Barrows et al. 2011) and sea level was 70 m lower (Chappell and Polach 1991) than today. The last glaciation period coincides with the decrease of *Ne* for *O. priamus*. On the contrary, the scale induced by *O. priamus*' inference does not allow to discern any congruence between climatic changes and the *Ne* of *T. oblongomaculatus* and *O. alexandrae*. In addition, this result relies only on a single genome for *O. alexandrae* whereas the two populations of *O. alexandrae* might have had different demographic histories.

Within *O. alexandrae*, every individual showed a similar demographic pattern (low and steadily declining *Ne* during most of the history, **Figure 6a**). Nonetheless, the two populations of *O. alexandrae* seem to separate initially ~10,000 years ago, and have likely been followed by a relative increase of the *Ne* in the last period (after the split). These results were validated when using multiple genomes per population (see *Materials and Methods*) (**Figure 6b**) and computed the cross coalescence rate (CCR) between populations (**Figure 6c**). The CCR gives the probability that a coalescence happens between rather than within the populations and, therefore, quantifies isolation (Wang et al. 2020). Using a generation time of 0.75 years, the CCR started to decrease ~10,000 years ago (CCR = 0.95) and was below 5% at ~1,000 years ago (CCR = 0.05) (**Figure 6c**). These results suggest that two

populations have recently diverged, potentially just after the last glaciation, raising the question of the extent to which the two populations are fully isolated or whether some gene flow (migration) is still ongoing.

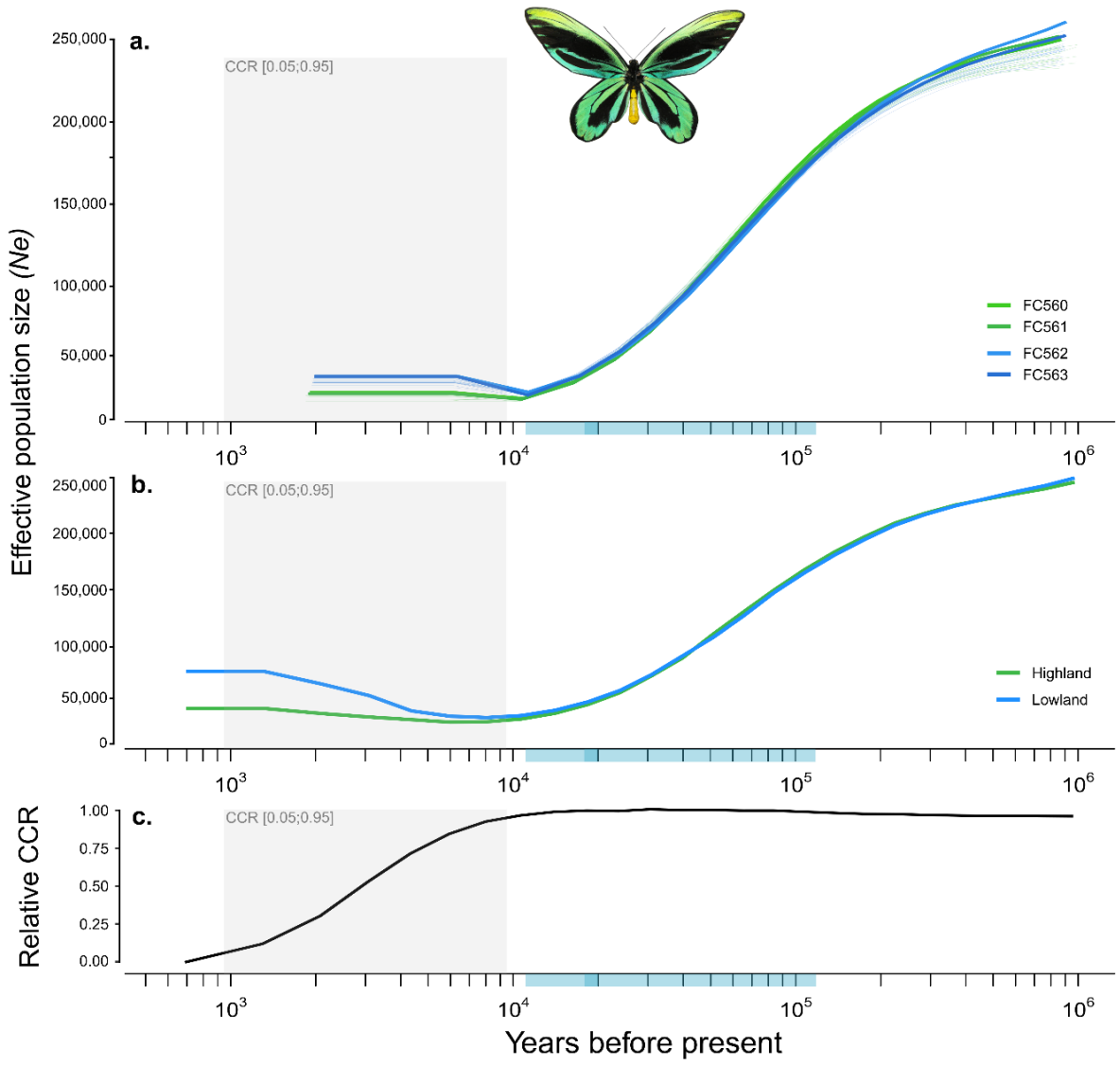


Fig. 6. Demographic inferences of *O. alexandrae* populations. **a.** MSMC estimates for each individual of *O. alexandrae*, with bootstraps represented in clear lines. **b.** MSMC estimates for both populations of *O. alexandrae*. **c.** Relative Cross Coalescent Rate (CCR) estimated with the MSMC analysis, representing the time interval of population divergence. The gray zone represents the 5%-95% CCR ([947; 9460] years ago), and has been reported on each plot. The pale blue rectangle along the time bar indicates the limits of the last glaciation period (11,700 to 115,000 years ago) with the last glacial maximum in darker blue (19-20,000 years ago). For every graph, the recent present (last 700 years) is not represented.

Population structure

To explore the recent divergence between the two populations, we computed the median F_{ST} on sliding windows of 100 kb across the whole genome (see *Materials and Methods*). We estimated a median F_{ST} of 0.03 but the F_{ST} was highly variable among genomic positions with top 5% windows having F_{ST} above 0.14 (**Figure 7**). The F_{ST} and nucleotide diversity clearly decrease in the center of the large contigs in regions (**Figure 7**) as expected along chromosomes (e.g., Tine et al. 2014; Delmore et al. 2018). This large-scale variation in F_{ST} and genetic diversity, repeated among scaffolds, suggest the action of endogenous effects such as recombination rate variation, rather than the effect of exogenous local adaptation (Burri 2017). This is consistent with reduced recombination rate in centers and extremities of the chromosomes as recently reported in the painted lady (*Vanessa cardui*; Shipilina et al. 2022). In addition, the cumulative size of 100-kb windows with less than two heterozygous sites represents less than 1% of the genome in all individuals. Moreover, not a single window of 1 Mb with one heterozygous site or less was detected in any individuals. It suggests that the genome of *O. alexandrae* shows no sign of recent inbreeding.

Using the approximate Bayesian computation (ABC) framework implemented in DILS (Fraïsse et al. 2021), we fitted the four models (secondary contact [SC], strict isolation [SI], isolation with migration [IM], ancient migration [AM]) across randomly selected genomic regions to estimate population sizes, migration rates and time of population separation according to each model (i.e., the date when gene flow stopped) between lowland and highland populations (see *Materials and Methods*).

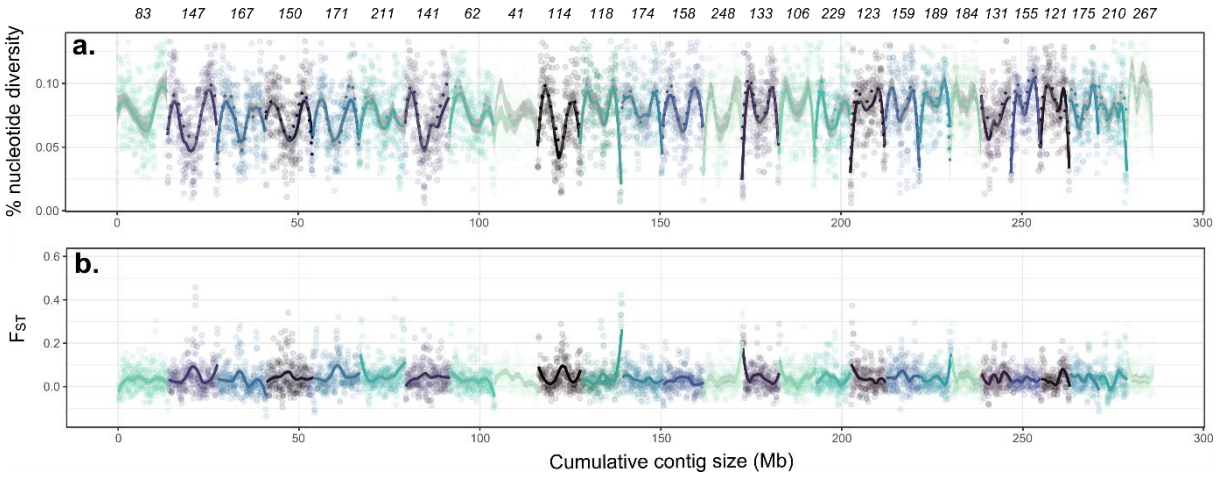


Fig. 7. Genome-wide genetic diversity and differentiation within *Ornithoptera alexandrae*. **a.** Percentage of nucleotide diversity of the genome of *O. alexandrae* for all individuals. **b.** Genetic differentiation (F_{ST}) between the highland and lowland populations. Nucleotide diversity and F_{ST} were computed on non-sex-related contigs >500 kb in length, over sliding windows of 100 kb with an overlap of 50 kb. Alternance of colors indicates the different chromosome-level contigs. Contig names are indicated at the top in italics.

On the four runs, DILS only set aside the secondary contact model, but could not identify a best-fitting model between the three others: (2 runs supported migration (IM model), and 2 runs found isolation (1 run for AM and 1 run for SI). Depending on the best fitted model, the time of split was either more recent than MSMC (isolation) or similar (migration) and varied from 5,100 generations (~3,825 years ago) to 16,850 generations (~12,637 years ago) for the median of the posterior distribution of the SI model and the IM model respectively (**Table S5**).

Despite the uncertainty of the DILS scenarios, notably on the presence or intensity of migratory flows between the two populations, all the analyses seem overall to agree on the rather consistent isolation of the two populations for a few thousand years. This is interesting because field observations tend to say that the two populations are externally morphologically similar but express biological differences. The highland population is on average larger than the lowland population (264 vs. 218 mm wingspan on average; Mitchell et al. 2016). Furthermore, *O. alexandrae* populations may feed on two different *Aristolochia* species, which could be *A. meridionaliana* in the highlands and *A. alexandriana* in the lowlands (Haugum and Low 1979; Parsons 1996; Mitchell et al. 2016). However, it is difficult to reach a conclusion on this point since little is known on the taxonomy of these two plant morphs and, as mentioned by Buchwalder et al. (2014) who started a taxonomic revision of the genus Papuan *Aristolochia*, there is still significant work to be done to update the nomenclature of this group. The two populations of *O. alexandrae* also differ slightly in their life cycle. From the egg, through larval and pupal stages to adults, the life cycle of *O. alexandrae* takes about 131 days (74 days as larva) in the lowland, while it takes 200 days (125 days as larva) in the highland of the Managalas Plateau (Straatman 1971; Mitchell et al. 2016). Such a difference may be explained by a temperature gradient between the low and highlands (up to 4°C cooler than the plain).

Overall, our results indicate a lack of evidence of a strong genomic differentiation and a genomic barrier to gene flow between the two allopatric populations of *O. alexandrae* that could justify the description of two species. Nevertheless, given the morpho-ecological differences and the fact that there is a modest population structure (F_{ST}) and divergence time probably initiated after the Last Glacial Maximum, it is possible that the populations deserve subspecies status. However, determining whether the observed morpho-ecological differences are fixed in the genome and to what extent the genitalia differ would shed more light on this proposal, although the latter is difficult to satisfy due to the lack of available individuals to dissect for this threatened and protected species.

Conservation implications

Conservation today is often approached as a triage, leaving out some "unsalvageable" species in an attempt, at best, to save others (Hayward and Castley 2018). As discussed by Wiedenfeld et al.

(2021), this approach to the sixth mass extinction should be primarily a matter of prioritizing the allocation of economic and material resources to the rescue of any species, and the current system is failing to allocate enough resources to save biodiversity. The choice to help certain species rather than others is risky since this choice is necessarily based on partial knowledge and limited data, such as their apparent rarity, sometimes interpreted as an abrupt decline, of certain species. However, it seems that comprehensive and meticulous works of some studies such as the one carried out on the kākāpō (Dussex et al. 2021), Island fox (Robinson et al. 2018) or the vaquita porpoise (Morin et al. 2021, Robinson et al. 2022) in recent years are a further argument that species are not necessarily doomed to extinction due to their rarity and low numbers (e.g., Robinson et al. 2018; Wiedenfeld et al. 2021; Robinson et al. 2022). On the contrary, these rare species may have been rare for so long that this may have led them for instance to have purged more deleterious recessive alleles, they might not suffer from inbreeding depression, and their low number may be less of a problem in recovering their population than many other species that would appear to us less endangered at first sight. Protection measures are proposed and sometimes implemented as a result of scientific studies revealing the urgency of acting to save species. Most of these measures concern vertebrates (e.g., large mammals), often omitting threatened insect diversity from conservation policies (but see Mikheyev et al. 2017). As Harvey et al. (2020) advocates in their roadmap for insect conservation, there is a range of actions to be taken, from short-term actions such as drastically or immediately stopping threats, to medium- and long-term actions such as conducting new research to fill the knowledge gap regarding insect decline or launching sustainable funding initiatives with the aim of restoring, protecting and creating new vital habitats for insects.

Ornithoptera alexandrae is one of the most charismatic invertebrate species and is receiving special attention as a flagship species for invertebrate conservation (Mitchell et al. 2016). The biological characteristics of *O. alexandrae* correspond to a wide range of traits identified as vulnerability factors, notably its large size, its host and environment specificity, the vulnerability of its host plant or the presence of inter-specific competition in its distribution range (Koh et al. 2004; Mattila et al. 2006). On the other hand, the relatively similar past evolutionary patterns and dynamics of O. alexandrae with the vaquita porpoise, showing a low heterozygosity and a low *Ne* for a long time, suggest that we can have hope about the capacity of *O. alexandrae* to avoid extinction. However, the species experienced declining trends in the past and these overall results should not be an argument to relax the protection measures. On the contrary, we take the opportunity to encourage the authorities and associations for increasing protection actions, especially since we know very little about the health of the butterfly populations for the last 4,000 years, while the destruction of its habitat has clearly not decreased since.

Logging opens up previously inaccessible areas, resulting in changes in *O. alexandrae* habitat by increasing the number and extent of clearings of primary and secondary forests (Mitchell et al. 2016). Today, oil palm exploitation likely constitutes threats through planned or ongoing deforestation, as well as the increased human population on the Popondetta Plain associated with this activity (Mitchell et al. 2016). Given the difficulty to access the Managalas Plateau, the anthropic pressure is today higher in lowlands, suggesting that the lowland population is experiencing more threats. Therefore, we urge policy makers to (1) protect the rainforest environment of *O. alexandrae*, especially in the highlands where a specific protection area could be exclusively dedicated to it, and (2) strongly support the lowland population through the ongoing rearing program with breeding and release in the wild. Both proposals are discussed and analyzed in detail in Mitchell et al. (2016) and we fully agree with their statement on the urgency of action to protect *O. alexandrae* and how to address the problem. Note that we advise avoiding interbreeding between the two populations as there is an ongoing process of natural evolution leading to genetic divergence between the two populations.

An interesting research perspective would be to perform a temporal sampling of genomic data from museum specimens. This can provide a more accurate approach to quantify genetic threats in endangered species and to estimate recent decreases in genome-wide diversity, increases in inbreeding levels, and accumulation of deleterious genetic variation (Díez-del-Molino et al. 2018) like it has been done on kākāpō (Dussex et al. 2021) and the vaquita porpoise (Robinson et al. 2022).

Materials and Methods

Sampling, DNA and RNA extractions and sequencing

With permits, a total of four individuals of Ornithoptera alexandrae were collected in Papua New Guinea (Oro Province) in November 2019 by F.L.C. and D.B. Two specimens per population were collected: one female and one caterpillar for the lowland population (near Popondetta), and one male and one caterpillar for the highland population in the Managalas Plateau (near Kawowoki village). For each specimen, head, thorax, and abdomen have been separated with scalpels, crushed with surgical scissors, and conserved separately in RNAlater, then stored in freezers at - 20°C after a few days at ambient temperature (and one day at 4°C as a transition). One individual of Ornithoptera priamus and one individual of Troides oblongomaculatus have been collected under similar conditions during the same mission.

Tissues from the thorax or abdomen were used to extract high molecular weight DNA. As part of tests that ended up being similar in terms of quality of sequencing, we used two different extraction methods. The first was the phenol-chloroform method, including a specific ratio of 0.8x AMPure beads applied to retain the longest DNA fragments (Tilak et al. 2020). The second method was the use of the Qiagen genomic DNA kit. This second solution

was ultimately applied to most samples due to a better 260/230 ratio in Nanodrop assays, as DNA purity is essential for long-read sequencing, especially for Oxford Nanopore Technology (ONT) sequencing. In addition, one of the four samples of FC561 was treated with the Short Read Eliminator Kit XS (Circulomics, PacBio, USA) to discard sequences below 10 kb long. Final DNA purity and concentrations were measured using Nanodrop (Thermo Fisher, USA) and Qubit (Thermo Fisher, USA). RNA was extracted for O. alexandrae only. Extraction and purification were conducted with the Qiagen RNeasy kit. We used part of the thorax of caterpillars (FC561/FC563) and part of the abdomen for adults (FC560/FC562), that were crushed in the lysis buffer.

Library preparations and sequencing

Whole-genome libraries were constructed using the resulting high-molecular-weight DNA as input for the Nanopore LSK-109 ligation kit (Oxford Nanopore Technologies, UK) following the manufacturer's protocol. Long-read sequencing was performed on a GridION device with two to four R9.4.1 flow cells, depending on the individuals (**Table S6**).

Remaining DNA extractions of each individual were sent to Novogene Europe (Cambridge, UK) for library preparations. Libraries were generated using NEBNext DNA Library Prep Kit following manufacturer's recommendations and indices were added to each sample. Genomic DNA was randomly fragmented to a size of 350 bp by shearing, then DNA fragments were end-polished, A-tailed, and ligated with the NEBNext adapter for Illumina sequencing, and further PCR enriched by P5 and indexed P7 oligos. The PCR products were purified (AMPure XP system) and the resulting libraries were analyzed for size distribution by Agilent 2100 BioAnalyzer and quantified using real-time PCR. Since the genome sizes for the Troidini species was estimated to be about 320 Mb (*Ornithoptera*) and 340 Mb (*Troides*), Illumina 150 bp paired-end sequencing was run on a NovaSeq 6000 instrument to obtain about 32 and 34 Gb per sample corresponding to a genome depth-of-coverage of about 100x.

The quality and quantity of all RNAs were checked using Nanodrop, Qubit, and 1.0% agarose gel electrophoresis and sent to Novogene for library preparations. Messenger RNA was purified from total RNA using poly-T oligo-attached magnetic beads. After fragmentation, the first strand cDNA was synthesized using random hexamer primers followed by the second strand cDNA synthesis. The library was ready after end repair, A-tailing, adapter ligation, size selection, amplification, and purification. The library was checked with Qubit and real-time PCR for quantification and BioAnalyzer for size distribution detection. Quantified libraries have been pooled and sequenced on Illumina platforms, according to effective library concentration to a data amount of about 8 Gb per sample.

Assembly of reference genomes

For GridION sequencing, all fast5 files were basecalled using Guppy 5.0.15 (developed by ONT) using super-high accuracy mode and a quality control of 10 (min_score 10). Sequencing adapters were trimmed using Porechop 0.2.3 (https://github.com/rrwick/Porechop).

Draft genome assemblies were performed with Flye 2.8.3 (Kolmogorov et al. 2019) with default options. Illumina reads were cleaned, filtered and paired with fastp 20.0 (Chen et al. 2018) using default options. Paired-end sequences were mapped on the Flye assembly using BWA 0.7.17 (Li 2013). Resulting SAM files were converted to sorted indexed BAM files with SAMtools (Li et al. 2009). Flye draft assemblies were polished with two rounds of Pilon 1.24 (Walker et al. 2014) using this mapping information. Assembly statistics were then assessed using the gVolante2 platform (Nishimura et al. 2017) to retrieve the number and size of contigs, the presence, completeness and duplication of BUSCO genes of the Lepidoptera odb10 database (Manni et al. 2021). More information and statistics about quality of sequencing, assembly and polishing are displayed in **Table S6**. Before submitting genomes assemblies to GenBank, we used BlobTools 1.1.1 (Laetsch and Blaxter 2017) set to the ncbi and diamond databases to check for possible contaminations. We found no evidence of artificial contamination coming from laboratory manipulation, but some contigs were clearly identified as belonging to exogenous organisms such as host plants and symbionts. We removed all contigs that were belonging to the Plant (subsequently identified as *Aristolochia* by BLAST), Microsporidia (unicellular fungal insect parasites) or Pseudomonadota (*Wolbachia*, *Enterobacter*) phylum. The FC563 assembly showed no evidence of contamination, while 83% of the total contigs removed were from FC561 (**Table S7**). We used the scaffold function of RagTag 2.1.0 (https://github.com/malonge/RagTag; Alonge et al. 2019; Alonge et al. 2021) to find the correspondence between *Papilio bianor* chromosomes and *O. alexandrae* FC563 assembly.

Genome annotation

We performed a full pipeline of annotations for the individuals FC560 and FC563. The pipeline was composed of the six following steps. First, we reconstructed the repeat sequences using RepeatModeler 2.0.1 (Flynn et al. 2020). The consensus sequences generated by RepeatModeler were blasted against the "reference transcriptome" database of UniProt (download in October 2021, https://www.uniprot.org/) using diamond blastx (Buchfink et al. 2015), and we excluded all the proteins that were not associated to repeat sequences from the consensus sequences. We then annotated the repeat sequences in the respective assemblies using RepeatMasker (Smit et al. 2013-2015) using both the Dfam libraries (Storer et al. 2021) setting the parameter "-species" on "Arthropoda" and the newly identified repeat sequences reconstructed using RepeatModeler. Second, we assembled the RNA-seq data by cleaning

the reads with fastp 20.0 (Chen et al. 2018), mapped the read using HISAT2 (Kim et al. 2019) onto the reference genome, and we annotated the cDNA using StringTie (Petrea et al. 2015) producing a GTF file. The cDNA sequences were converted in fasta using the "gtf_genome_to_cdna_fasta.pl" script of TransDecoder (https://github.com/TransDecoder/TransDecoder). The RNA-seq data of all the individuals were mapped against the reference genome of FC563. Third, we ran MAKER 2.31.11 (Holt and Yandell 2011) using the information of the annotated repeat sequences and the cDNA sequences provided as "rm_gff" and "est" option in the control file of MAKER, respectively. We also used the proteins sequences of *Heliconius melpomene, Melitaea cinxia, Papilio machaon, Papilio xuthus, Papilio glaucus* provided as "protein" option in the control file of MAKER to help identify genes using homology information. Fourth, SNAP (Korf et al. 2004) and AUGUSTUS (Stanke et al. 2006) were used to produce gene prediction models from the first round of MAKER. BUSCO 5 (Simão et al. 2015) with options "--long" and "--augustus" and the Endopterygota database was used to produce the gene prediction model of AUGUSTUS. Fifth, we ran again MAKER using the annotation from the first round and the gene models of SNAP and AUGUSTUS. Sixth, the steps 3 and 4 were repeated using the second round of MAKER annotation to produce a third and final round of annotations. Finally, for individuals FC561 and FC562, we used Liftoff 1.6.3 (Shumate and Salzberg 2021) to map the annotation of FC560 on the two other assemblies.

Mitogenomic diversity

Long reads of every individual were corrected with short reads using LoRDEC 0.9 (Salmela and Rivals 2014). For each individual of *O. alexandrae*, the corrected long reads were mapped with Minimap2 2.17 (Li 2018) on the reference mitogenome of *O. richmondia* from a previous study (NC_037869.1, Condamine et al. 2018). The reads that mapped with the references were filtered by quality via SAMtools (Li et al. 2009) ("view -q 30"). For each individual, a subset of reads from 3.6 Mb to 6.3 Mb were created so that mitogenomes would have an expected depth of coverage between 200x and 400x. We used Flye 2.8.3 to assemble the mitogenomes and the resulting assemblies were given to MitoFinder 1.4 (Allio et al. 2020) to annotate (gene, tRNA, rRNA) and extract genes (gene, rRNA). The cleaned short-read data were also directly given to MitoFinder to produce annotated mitogenomes based on the short-read data only. As these mitogenomes were of slightly better quality (better annotation, presence of the complete or nearly complete D-loop), which were those submitted to GenBank (Accession numbers OQ59006-OQ59009). A nucleotide alignment was produced with MAFFT 7.310 (Katoh and Standley 2013) after having manually adjusted the sequences due to circularization. We used Seaview 4.7 (Gouy et al. 2010) to visualize the four whole mitochondrial genome alignment and to count pairwise differences using

"Statistics" of Seaview, and the mean of this pairwise distance was calculated, and ultimately divided by the alignment length (**Table S3**). The same steps were carried out with the COI mitochondrial gene only.

Nuclear heterozygosity of Troidini

For *O. alexandrae*, we selected FC563 as the reference assembly, as it has the highest N50, mean coverage and BUSCO score, the lowest number of contigs, and has no contamination (Table 1). Genomes of *O. priamus* and *T. oblongomaculatus* were their own reference. The corrected reads (LoRDEC, see *Mitogenomic diversity* section) of every individual were mapped on their reference genome using "-a" option of Minimap2 2.17 (Li 2018). We used SAMtools to compress, sort and index these mappings. SNP calling was performed with Longshot 0.4.1 (Edge and Bansal 2019), using a threshold of 15x minimum and 150x maximum for the depth of coverage (minimal quality of 20; default quality) and applying a transition/transversion rate for genotype prior estimation (ts_tv_ratio) of 2.0 (Edge and Bansal 2019). SNPs with a quality below 200 were excluded. All positions, SNPs and homozygous, must be contained within the coverage thresholds, otherwise they were considered as ambiguous. As the quality of phasing may be important for population genomics and demographic analyses, we checked the average size of phasing from Longshot. The average length of the phased blocks is 435 kb, and a haplotype N50 of 1.7 Mb (**Table S8**, for phasing statistics per individual). In addition, to ensure that our heterozygosity estimates did not depend on the data and method (Bentley and Armstrong 2022), we also calculated the heterozygosity rate based on the short reads data to evaluate the robustness of our results. Illumina cleaned reads were mapped on references using the speedseq pipeline (Chiang et al. 2015) that relies on BWA 0.7.17 (Li 2013). We excluded the so-called discordant and splitter reads and the reads with mapping quality below 30. Genotype calling was performed using FreeBayes 1.3.1 (Garrison and Marth 2012) set with the "--use-best-n-alleles 4" option, and the same coverage threshold as for long reads data. SNPs with quality scores below 50 and out of these coverage thresholds were excluded. Homozygous positions were also selected based on the same coverage threshold and considered ambiguous otherwise. For ONT and Illumina data, heterozygosity was computed as the number of SNPs divided by the total number of sites excluding ambiguous positions. Using the above criteria, heterozygosity levels were similar between ONT and Illumina data (**Table S4**). To compare our results with the values of Mackintosh et al. (2019), heterozygosity was also computed using only four-fold degenerate sites, as identified by a custom script using BIO++ library (Guéguen et al. 2013).

Estimation of the demographic history and effective population size

We used a Sequential Markovian Coalescent (SMC) model (McVean and Cardin 2005, e.g. PSMC, Li and Durbin 2011; MSMC2, Schiffels and Wang 2020) to estimate the ancestral effective

population size (*Ne*) trajectory of the studied Troidini species.

The SMC model requires calibrations, in particular a value of mutation rate. We estimated this rate based on synonymous mutations by selecting the four-fold degenerate sites of the third codon positions of BUSCO genes from the odb10 lepidopteran database. We retrieved the set of fasta nucleotide sequences using BUSCOMP 0.13.0 (Edwards 2019) on local runs of BUSCO 5 (odb10_Lepidoptera) of the six studied individuals and we considered only the genes that contained all individuals, which corresponded to 5,127 genes (~97% of the lepidopteran gene database). Assuming that these mutations are neutral, we applied the formula $D = 2 \times T \times \mu$ where D is the genetic divergence between two species, T is the divergence in millions of years and μ is the mutation rate per million years (Kimura 1983; Birky and Walsh 1988). Here, we chose *O. priamus* as the divergent species of *O. alexandrae*, and set T=12.03166 Ma (median value of the divergence time between these two species, with 95% credibility interval = 7.9662–16.7068; *sensu* Allio et al. 2021). To estimate D, we split the 5,127 genes into six bins based on the GC-content and we estimated the branch lengths from the six corresponding trees inferred by IQ-TREE 1.6.12 (Nguyen et al. 2015) with a GTR+Γ6 substitution model. The divergence D ranged from 0.043 to 0.052 for the lowest GC-content bins to the highest, respectively. We then took an average D (= 0.0475) between the two *Ornithoptera* species to obtain a mean value of μ equal to 1.9740e-09 mutations per site per year. As *O. alexandrae* highland population produces one generation per year while the lowland population produces two (Mitchell et al. 2016), we set the generation time to one and a half generations per year, therefore the mutation rate μ was estimated at 1.3160e-09 mutations per site per generation, which is at the lowest end of the range estimated for *Heliconius* (μ = 1.3 - 5.5e-09; Keightley et al. 2015).

It has been shown that SMC models do not perform well when the ratio of recombination rate *r* over mutation rate μ becomes greater than one (Sellinger et al. 2021). Assuming *O. alexandrae* genome is 325 Mb long, distributed in 30 chromosomes and that there is a single crossover per tetrad per male meiosis, the recombination rate would be *r* = 2.7e-8. An analysis of the nymphalid *Vanessa cardui* (Shipilina et al. 2022) estimated an average *r* between 3.81e-8 and 4.05e-8 with substantial inter-chromosomal variation, meaning that the average recombination rate of *O. alexandrae* is an order of magnitude greater than its mutation rate at least.

To investigate if this parameter range was a problem in SMC analyses, we used msprime 1.2.0 (Kelleher et al. 2016) to produce ten simulated datasets of a stable demographic history scenario with *Ne* = 100,000 under a *r* = 1e-8 and a μ = 1.316e-9 on 30 chromosomes of 10 Mb each. We ran the Multiple Sequentially Markovian Coalescent (MSMC) model implemented in the MSMC2 software (https://github.com/stschiff/msmc2; Schiffels and Wang 2020) with default

options (i.e. *-rhoOverMu = 0.25*) to test whether the inferred demography was recovered stable. We ran the same data by setting an initial value *-rhoOverMu = 10*.

We applied the MSMC2 model on real data, first for each Troidini species (FC563 was selected for *O. alexandrae*). We used the VCF files generated using Longshot (as described in the *Nuclear heterozygosity of Troidini* section) and created the so-called "mask file" for each individual based on the depth of coverage thresholds of >20x and <150x using a custom python script. These files were then combined using the "generate_multihetsep.py" of MSMC2 to generate "multihetsep.txt" input files (https://github.com/stschiff/msmc-tools/blob/master/msmc-tutorial/guide.md). MSMC2 was run using default parameters (especially the initial value of the ratio [r / μ] *-rhoOverMu = 0.25)*, and with *-rhoOverMu = 10* (See **Figure S2**, for a comparison of the results with both options).

By applying the same methods, we ran MSMC2 on each individual of *O. alexandrae* and then applied the model on the two populations (both composed by two genomes). MSMC2 was run using default parameters, or with *-rhoOverMu = 10*. The "-I" option was used to consider relevant haplotypes depending on the analyses (i.e., single individuals, two populations and cross coalescence rate between populations). For each analysis on real data, we generated 10 bootstraps per individual using the multihetsep_bootstrap.py script available at: https://github.com/stschiff/msmc-tools. We generated all graphs with the R package ggplot2 (Wickham 2016) by considering a generation time of 0.5 for *O. priamus* and *T. oblongomaculatus* (two generations per year), 0.75 for *O. alexandrae* (one and a half generations per year) and μ = 1.316e-9 for every individual.Finally, we relied on the ABC framework implemented in DILS (Fraïsse et al. 2021) to test several scenarios of divergence between populations. Alternative methods like ∂a∂i (Gutenkunst et al. 2009) or FastSimCoal (Excoffier and Foll 2011) could not be implemented because of no modeling of *Ne* through time and sample size limitations of our dataset to compute site-frequency spectrum, respectively. DILS takes into account linkage information that is informative about past demography (Fraïsse et al. 2021). To fit DILS on our data, we randomly selected 5,000 windows of 4,000 bp (2 Mb in total) because the coalescence program is time-consuming to simulate large chunks of chromosomes with recombination. The analysis was replicated four times to evaluate variability and reproducibility of the ABC inferences. DILS implements a pipeline that selects the best-fitting demographic model by comparing models with variations in *Ne* and migration among loci allowing to consider linked selection and alleles that could be selected against during hybridization (Fraïsse et al. 2021). The four demographic scenarios tested include: strict isolation (SI), isolation with migration (IM), ancient migration (AM), and secondary contact (SC). DILS used a random forest method (Pudlo et al. 2016) to select the best model and estimate posterior

parameters' distributions using rejection and neural network methods implemented in the R package abc (Csilléry et al. 2012).

Population structure

The population differentiation due to genetic structure was estimated with the nucleotide diversity and fixation index (F_{ST}) that were computed using seq_stat_2pop (https://github.com/benoitnabholz/seq_stat_2pop) using Bio++ library (Guéguen et al. 2013). The seq_stat_2pop program uses fasta sequences as input such that VCF files were converted into fasta sequences using a custom python program using coverage information for the homozygous sites as explained above (see *Nuclear heterozygosity of Troidini* section).

Nucleotide diversity (π-diversity) was computed as the mean pairwise divergence between pairs of chromosomes. F_{ST} was computed using the nucleotide sequence as: $F_{ST} = 1 - \pi_{intra} / \pi_{total}$; where π_{intra} is the mean nucleotide diversity of the two populations ($\pi_{intra} = (\pi_{highland}+\pi_{lowland}) / 2$) and π_{total} is the nucleotide diversity computed using all individuals (Nei 1982). F_{ST} and nucleotide diversity were computed on non-sex-related contigs >500 kb in length, over sliding windows of 100 kb with 50 kb overlapping regions (windows of 100 kb and overlapping regions of 100 kb were tested and led to similar results). It is considered that a F_{ST} value greater than 0.15 is significant in differentiating populations (Frankham et al. 2010).

Acknowledgments

We thank two anonymous reviewers who provided very relevant and constructive comments to improve the study. This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (project GAIA, agreement no. 851188). This project has been generously supported by the Swallowtail & Birdwing Butterfly Trust, and we thank the Trust's Chair, N. Mark Collins, for his continuous advice and encouragement. We thank Henry S. Barlow and Ian Orrell who helped with the organization of the field work as well as with the export and CITES permits. We thank Conwell Nukara for field work assistance in the Managalas Plateau. This study also benefited from lab facilities near Popondetta funded by the Sime Darby Foundation. CITES permits were obtained for *Ornithoptera alexandrae* (FR1903400125-I, 019418, delivered October 23, 2019), *O. priamus* (FR1903400124-I, 019429, delivered November 1, 2019), *Troides oblongomaculatus* (FR1903400124-I, 019429, delivered November 1, 2019).

Author Contributions

F.L.C. and D.B. conceived and supervised the project. F.L.C. and D.B. collected the samples. E.C. and M.-K.T. carried out the molecular experiments. E.L.R. and B.N. performed the bioinformatic analyses. E.L.R., B.N., and F.L.C. discussed the results. E.L.R. and F.L.C. wrote the draft manuscript and B.N., M.-K.T and D.B. made comments. All authors have read and approved the final manuscript.

Data Availability

The birdwing genomes, mitogenomes and sequencing data in the present study, including Nanopore, Illumina, and RNA data are available from the Genome database and Sequence Read Archive (SRA) under the Bioproject accession number PRJNA938052, with the corresponding BioSamples accession numbers FC560: SAMN33424250; FC561: SAMN33424251; FC562: SAMN33424252; FC563: SAMN33424253; FC565: SAMN33424254; FC569: SAMN33424255.

References

Ackery PR. 1997. The Natural History Museum collection of *Ornithoptera* (birdwing) butterflies (Lepidoptera: Papilionidae). *Biol Curat.* **8**:11-17.

Allio R, Donega S, Galtier N, Nabholz B. 2017. Large variation in the ratio of mitochondrial to nuclear mutation rate across animals: implications for genetic diversity and the use of mitochondrial DNA as a molecular marker. *Mol Biol Evol.* **34**:2762-2772.

Allio R, Schomaker-Bastos A, Romiguier J, Prosdocimi F, Nabholz B, Delsuc F. 2020. MitoFinder: efficient automated large-scale extraction of mitogenomic data in target enrichment phylogenomics. *Mol Ecol Res.* **20**:892-905.

Allio R, Nabholz B, Wanke S, Chomicki G, Pérez-Escobar OA, Cotton AM, Clamens AL, Kergoat GJ, Sperling FAH, Condamine FL. 2021. Genome-wide macroevolutionary signatures of key innovations in butterflies colonizing new host plants. *Nat Commun.* **12**:354.

Alonge M, Lebeigle L, Kirsche M, Aganezov S, Wang X, Lippman ZB, Schatz MC, Soyk S. unpublished data. Automated assembly scaffolding elevates a new tomato system for high-throughput genome editing. *bioRxiv*. https://www.biorxiv.org/content/10.1101/2021.11.18.469135v1, last accessed November 9, 2022.

Alonge M, Soyk S, Ramakrishnan S, Wang X, Goodwin S, Sedlazeck FJ, Lippman ZB, Schatz MC. 2019. RaGOO: fast and accurate reference-guided scaffolding of draft genomes. *Genome Biol.* **20**:224.

Barrows TT, Hope GS, Prentice ML, Fifield LK, Tims SG. 2011. Late Pleistocene glaciation of the Mt Giluwe volcano, Papua New Guinea. *Quatern. Sci Rev.* **30**: 2676-2689.

Bentley BP, Armstrong EE. 2022. Good from far, but far from good: The impact of a reference genome on evolutionary inference. *Mol Ecol Res.* **22**:12–14.

Birky CW Jr, Walsh JB. 1988. Effects of linkage on rates of molecular evolution. *Proc Natl Acad Sci.* **85**:6414–6418.

Böhm M. 2018. *Ornithoptera alexandrae*. The IUCN Red List of Threatened Species 2018: https://dx.doi.org/10.2305/IUCN.UK.2018-1.RLTS.T15513A88565197.en.

Buchfink B, Xie C, Huson DH. 2015. Fast and sensitive protein alignment using DIAMOND. *Nat Methods* **12**:59–60.

Buchwalder K, Samain MS, Sankowsky G, Neinhuis C, Wanke S. 2014. Nomenclatural updates of *Aristolochia* subgenus *Pararistolochia* (Aristolochiaceae). *Austral Syst Bot.* **27**:48-55.

Buffalo V. 2021. Quantifying the relationship between genetic diversity and population size suggests natural selection cannot explain Lewontin's Paradox. *eLife* **10**:e67509.

Burns JM, Janzen DH, Hajibabaei M, Hallwachs, W, Hebert PDN. 2008. DNA barcodes and cryptic species of skipper butterflies in the genus *Perichares* in Area de Conservacion Guanacaste, Costa Rica. *Proc Natl Acad Sci.* **105**:6350-6355.

Burri R. 2017. Interpreting differentiation landscapes in the light of long-term linked selection. *Evol Lett.* **1**:118–131.

Chappell J, Polach H. 1991. Post-glacial sea-level rise from a coral record at Huon Peninsula, Papua New Guinea. *Nature* **349**:147-149.

Chattopadhyay B, Garg KM, Soo YJ, Low GW, Frechette JL, Rheindt FE. 2019. Conservation genomics in the fight to help the recovery of the critically endangered Siamese crocodile *Crocodylus siamensis*. *Mol Ecol.* **28**:936-950.

Chen S, Zhou Y, Chen Y, Gu J. 2018. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**:i884-i890.

Chiang C, Layer RM, Faust GG, Lindberg MR, Rose DB, Garrison EP, Marth GT, Quinlan AR, Hall IM. 2015. SpeedSeq: ultra-fast personal genome analysis and interpretation. *Nat Methods* **12**:966–968.

Collins NM, Morris MG. 1985. Threatened swallowtail butterflies of the world. The IUCN Red Data Book, Cambridge.

Condamine FL, Nabholz B, Clamens AL, Dupuis JR, Sperling FAH. 2018. 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 NV. 2015. Tiger swallowtail genome reveals mechanisms for speciation and caterpillar chemical defense. *Cell Rep*. **10**:910-919.

Csilléry K, François O, Blum MG. 2012. abc: an R package for approximate Bayesian computation (ABC). *Meth Ecol Evol.* **3**:475–479.

Delmore KE, Lugo Ramos JS, Van Doren BM, Lundberg M, Bensch S, Irwin DE, Liedvogel M. 2018. Comparative analysis examining patterns of genomic differentiation across multiple episodes of population divergence in birds. *Evol Lett.* **2**:76–87.

Díez-del-Molino D, Sánchez-Barreiro F, Barnes I, Gilbert MTP, Dalén L. 2018. Quantifying Temporal Genomic Erosion in Endangered Species. *Trends Ecol Evol.* **33**:176–185.

Dincă V, Dapporto L, Somervuo P, Vodă R, Cuvelier S, Gascoigne-Pees M, Huemer P, Mutanen M, Hebert PDN, Vila, R. 2021. High resolution DNA barcode library for European butterflies reveals continental patterns of mitochondrial genetic diversity. *Commun Biol.* **4**:315.

Dussex N, Van Der Valk T, Morales, HE, Wheat CW, Díez-del-Molino D, Von Seth J, Foster Y, Kutschera VE, Guschanski K, Rhie K, *et al*. 2021. Population genomics of the critically endangered kākāpō. *Cell Genomics* **1**:100002.

Edge P, Bansal V. 2019. Longshot enables accurate variant calling in diploid genomes from single-molecule long read sequencing. *Nat Commun.* **10**:4660.

Edwards RJ. 2019. BUSCOMP: BUSCO compilation and comparison–Assessing completeness in multiple genome assemblies. *F1000Research* **8**:995.

Ellegren H, Galtier N. 2016. Determinants of genetic diversity. *Nat Rev Genet.* **17**:422-433.

Excoffier L, Foll M. 2011. Fastsimcoal: a continuous-time coalescent simulator of genomic diversity under arbitrarily complex evolutionary scenarios. *Bioinformatics* **27**:1332-1334.

Flynn JM, Hubley R, Goubert C, Rosen J, Clark AG, Feschotte C, Smit AF. 2020. RepeatModeler2 for automated genomic discovery of transposable element families. *Proc Natl Acad Sci.* **117**:9451–9457.

Formenti G, Theissinger K, Fernandes C, Bista I, Bombarely A, Bleidorn C, Ciofi C, Crottini A, Godoy JA, Höglund J, *et al*. 2022. The era of reference genomes in conservation genomics. *Trends Ecol Evol.* **37**:197–202.

Fraïsse C, Popovic I, Mazoyer C, Spataro B, Delmotte S, Romiguier J, Loire E, Simon A, Galtier N, Duret L, *et al.* 2021. DILS: Demographic inferences with linked selection by using ABC. *Mol Ecol Res.* **21**:2629-2644.

Frankham R, Ballou JD, Briscoe DA. 2010. Introduction to conservation genetics. Cambridge University Press, UK.

Fuller ZL, Mocellin VJ, Morris LA, Cantin N, Shepherd J, Sarre L, Peng J, Liao Y, Pickrell PA, Matz M, et al. 2020. Population genetics of the coral *Acropora millepora*: Toward genomic prediction of bleaching. *Science.* **369**:eaba4674.

García-Berro A, Talla V, Vila R, Wai HK, Shipilina D, Chan KG, Pierce NE, Backström N, Talavera G. 2023. Migratory behavior is positively associated with genetic diversity in butterflies. *Mol Ecol.* **32**:560-574.

Garner BA, Hand BK, Amish SJ, Bernatchez L, Foster JT, Miller KM, Morin PA, Narum SR, O'Brien SJ, Roffler G, *et al*. 2016. Genomics in conservation: case studies and bridging the gap between data and application. *Trends Ecol Evol.* **31**:81-83.

Garrison E, Marth G. unpublished data. Haplotype-based variant detection from short-read sequencing. *arXiv*. https://doi.org/10.48550/arXiv.1207.3907 last accessed November 9, 2022.

Global Volcanism Program, 2022. Lamington (253010) in Volcanoes of the World (v. 5.0.0; 1 Nov 2022). Distributed by Smithsonian Institution, compiled by Venzke, E. https://doi.org/10.5479/si.GVP.VOTW5-2022.5.0

Gouy M, Guindon S, Gascuel O. 2010. SeaView version 4: a multiplatform graphical user interface for sequence alignment and phylogenetic tree building. *Mol Biol Evol.* **27**:221-224.

Guéguen L, Gaillard S, Boussau B, Gouy M, Groussin M, Rochette NC, Bigot T, Fournier D, Pouyet F, Cahais V, *et al.* 2013. Bio++: Efficient extensible libraries and tools for computational molecular evolution. *Mol Biol Evol.* **30**:1745–1750.

Gutenkunst RN, Hernandez RD, Williamson SH, Bustamante CD. 2009. Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS Genet.* **5**:e1000695.

Haugum J, Low AM. 1979. A monograph of the birdwing butterflies. The systematics of *Ornithoptera*, *Troides* and related genera. Vol 1, *Ornithoptera*. Scandinavian Science Press, Klampenborg, 308 pp.

Hayward MW, Castley JG. 2018. Editorial: triage in conservation. *Front Ecol Evol*. **5**:168.

Hazzouri KM, Sudalaimuthuasari N, Kundu B, Nelson D, Al-Deeb MA, Le Mansour A, Spencer JJ, Desplan C, Amiri KMA. 2020. The genome of pest *Rhynchophorus ferrugineus* reveals gene families important at the plant-beetle interface. *Commun Biol.* **3**:323.

He JW, Zhang R, Yang J, Chang Z, Zhu LX, Lu SH, Xie FA, Mao JL, Dong ZW, Liu GC, *et al*. 2022. High-quality reference genomes of swallowtail butterflies provide insights into their coloration evolution. *Zool Res.* **43**:367-379.

Hebert PDN, Penton EH, Burns JM, Janzen DH, Hallwachs W. 2004. Ten species in one: DNA barcoding reveals cryptic species in the neotropical skipper butterfly *Astraptes fulgerator*. *Proc Natl Acad Sci.* **101**:14812-14817.

Heckenhauer J, Frandsen PB, Sproul JS, Li Z, Paule J, Larracuente AM, Maughan PJ, Barker MS, Schneider JV, Stewart RJ, *et al*. 2022. Genome size evolution in the diverse insect order Trichoptera. *GigaScience.* **11**:giac011.

Holt C, Yandell M. 2011. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics* **12**:491.

Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol.* **30**:772-780.

Kebaïli C, Sherpa S, Rioux D, Després L. 2022. Demographic inferences and climatic niche modelling shed light on the evolutionary history of the emblematic cold-adapted Apollo butterfly at regional scale. *Mol Ecol.* **31**:448-466.

Keightley PD, Pinharanda A, Ness RW, Simpson F, Dasmahapatra KK, Mallet J, Davey JW, Jiggins CD. 2015. Estimation of the spontaneous mutation rate in *Heliconius melpomene*. *Mol Biol Evol.* **32**:239-243.

Kelleher J, Etheridge AM, McVean G. 2016. Efficient coalescent simulation and genealogical analysis for large sample sizes. *PLoS Comput Biol*. **12**:e1004842.

Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. 2019. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat Biotechnol.* **37**:907–915.

Kimura M. 1983. The neutral theory of molecular evolution. Cambridge University Press, UK.

Koh LP, Sodhi NS, Brook BW. 2004. Ecological correlates of extinction proneness in tropical butterflies. *Cons Biol.* **18**:1571-1578.

Kolmogorov M, Yuan J, Lin Y, Pevzner PA. 2019. Assembly of long, error-prone reads using repeat graphs. *Nat Biotech.* **37**:540-546.

Korf I. 2004 Gene finding in novel genomes. *BMC Bioinformatics* **5**:59.

Laetsch DR, Blaxter ML. 2017. BlobTools: Interrogation of genome assemblies. *F1000Research* **6**:1287.

Li H, Durbin R. 2011. Inference of human population history from individual whole-genome sequences. *Nature* **475**:493–496.

Li H. 2013. Unpublished data. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv*. https://doi.org/10.48550/arXiv.1303.3997, last accessed November 9, 2022.

Li H. 2018. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**:3094-3100.

Li H. Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. 2009. The sequence alignment/map format and SAMtools. *Bioinformatics* **25**:2078-2079.

Liu G, Chang Z, Chen L, He J, Dong Z, Yang J, Lu S, Zhao R, Wan W, Ma G. 2020. Genome size variation in butterflies (Insecta, Lepidoptera, Papilionoidea): a thorough phylogenetic comparison. *Syst Entomol.* **45**:571–582.

Lu S, Yang J, Dai X, Xie F, He J, Dong Z, Mao J, Liu G, Chang Z, Zhao R, *et al.* 2019. Chromosomal-level reference genome of Chinese peacock butterfly (*Papilio bianor*) based on third-generation DNA sequencing and Hi-C analysis. *GigaScience* **8**:giz128.

Lynch M, Ackerman MS, Gout JF, Long H, Sung W, Thomas WK, Foster PL. 2016. Genetic drift, selection and the evolution of the mutation rate. *Nat Rev Genet.* **17**:704-714.

Mackintosh A, Laetsch DR, Hayward A, Charlesworth B, Waterfall M, Vila R, Lohse K. 2019. The determinants of genetic diversity in butterflies. *Nat Commun.* **10**:3466.

Manni M, Berkeley MR, Seppey M, Simão FA, Zdobnov EM. 2021. BUSCO update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. *Mol Biol Evol.* **38**:4647-4654.

Manthey JD, Girón JC, Hruska JP. 2022. Impact of host demography and evolutionary history on endosymbiont molecular evolution: A test in carpenter ants (genus *Camponotus*) and their *Blochmannia* endosymbionts. *Ecol Evol.* **12**:e9026.

Mattila N, Kaitala V, Komonen A, Kotiaho JS, Päivinen J. 2006. Ecological determinants of distribution decline and risk of extinction in moths. *Cons Biol.* **20**:1161-1168.

McVean GA, Cardin NJ. 2005. Approximating the coalescent with recombination. *Phil Trans R Soc B.* **360**:1387-1393.

Meek AS. 1906. Letter to Karl Jordan dated 3rd Feb 1906 from "Above "Biagi" (5000 ft) Head of Mambare River", Papua New Guinea. Meek letter 155, BMNH Archives.

Mikheyev AS, Zwick A, Magrath MJ, Grau ML, Qiu L, Su YN, Yeates D. 2017. Museum genomics confirms that the Lord Howe Island stick insect survived extinction. *Curr Biol.* **27**:3157-3161.

Mitchell DK, Dewhurst CF, Tennent WJ, Page WW. 2016. Queen Alexandra's birdwing butterfly *Ornithoptera alexandrae* (Rothschild, 1907): A review and conservation proposals. Kuala Lumpur: Southdene Sdn Bhd.

Morin PA, Foote AD, Hill CM, Simon-Bouhet B, Lang AR, Louis M. 2018. SNP discovery from single and multiplex genome assemblies of non-model organisms. In: Head SR, Ordoukhanian P, Salomon DR, editors. Next Generation Sequencing. New York (NY): Springer New York. p. 113-144

Morin PA, Archer FI, Avila CD, Balacco JR, Bukhman YV, Chow W, Fedrigo O, Formenti G, Fronczek JA, Fungtammasan A, *et al.* 2021. Reference genome and demographic history of the most endangered marine mammal, the vaquita. *Mol Ecol Res.* **21**:1008-1020.

Nadachowska-Brzyska K, Li C, Smeds L, Zhang G, Ellegren H. 2015. Temporal dynamics of avian populations during Pleistocene revealed by whole-genome sequences. *Curr Biol.* **25**:1375-1380.

Nakae M. 2021. *Papilionidae of the World*. Tokyo: Roppon-Ashi Entomological Books.

Nei M. 1982. Evolution of human races at the gene level. In: Bonné-Tamir B, editor. Human Genetics, Part A: The unfolding Genome. New York (NY): Alan R. Liss. p. 167–181.

Nguyen LT, Schmidt HA, Von Haeseler A, Minh BQ. 2015. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol Biol Evol.* **32**:268-274.

Nishimura O, Hara Y, Kuraku S. 2017. gVolante for standardizing completeness assessment of genome and transcriptome assemblies. *Bioinformatics* **33**:3635-3637.

Palash A, Paul S, Resha SK, Khan MK. 2022. Body size and diet breadth drive local extinction risk in butterflies. *Heliyon* **8**:e10290.

Parsons MJ. 1992. The world's largest butterfly endangered: the ecology, status and conservation of *Ornithoptera alexandrae* (Lepidoptera: Papilionidae). *Trop Lepido Res.* **3**:33-60.

Parsons MJ. 1996. New species of *Aristolochia* and *Pararistolochia* (Aristolochiaceae) from Australia and New Guinea. *Botanical Journal of the Linnean Society*, **120**:199-238.

Parsons MJ. 1999. The butterflies of Papua New Guinea: their systematics and biology. London: Academic Press.

Pertea M, Pertea GM, Antonescu CM, Chang T-C, Mendell JT, Salzberg SL. 2015. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat Biotechnol.* **33**:290–295.

Petersen M, Armisén D, Gibbs RA, Hering L, Khila A, Mayer G, Richards S, Niehuis O, Misof B. 2019. Diversity and evolution of the transposable element repertoire in arthropods with particular reference to insects. *BMC Ecol Evol.* **19**:11.

Podsiadlowski L, Tunström K, Espeland M, Wheat CW. 2021. The genome assembly and annotation of the Apollo butterfly *Parnassius apollo*, a flagship species for conservation biology. *Genome Biol Evol.* **13**:evab122.

Pudlo P, Marin J-M, Estoup A, Cornuet J-M, Gautier M, Robert CP. 2016. Reliable ABC model choice via random forests. *Bioinformatics* **32**:859–866.

Quinn RM, Gaston KJ, Roy DB. 1997. Coincidence between consumer and host occurrence: macrolepidoptera in Britain. *Ecol Entomol*. **22**:197–208.

Robinson JA, Brown C, Kim BY, Lohmueller KE, Wayne R.K. 2018. Purging of strongly deleterious mutations explains long-term persistence and absence of inbreeding depression in island foxes. *Curr Biol.* **28**:3487-3494.

Robinson JA, Bowie RC, Dudchenko O, Aiden EL, Hendrickson SL, Steiner CC, Ryder OA, Mindell DP, Wall JD. 2021. Genome-wide diversity in the California condor tracks its prehistoric abundance and decline. *Curr Biol*. **31**:2939-2946.

Robinson JA, Kyriazis CC, Nigenda-Morales SF, Beichman AC, Rojas-Bracho L, Robertson KM, Fontaine MC, Wayne RK, Lohmueller KE, Taylor BL, *et al*. 2022. The critically endangered vaquita is not doomed to extinction by inbreeding depression. *Science* **376**:635-639.

Romiguier J, Gayral P, Ballenghien M, Bernard A, Cahais V, Chenuil A, Chiari Y, Dernat R, Duret L, Faivre N, *et al.* 2014. Comparative population genomics in animals uncovers the determinants of genetic diversity. *Nature* **515**:261-263.

Salmela L, Rivals E. 2014. LoRDEC: accurate and efficient long read error correction. *Bioinformatics* **30**:3506-3514.

Sánchez-Bayo F, Wyckhuys KA. 2019. Worldwide decline of the entomofauna: A review of its drivers. *Biol Cons.* **232**:8-27.

Schiffels S, Wang K. 2020. MSMC and MSMC2: The Multiple Sequentially Markovian Coalescent. In: Dutheil JY, editor. Statistical Population Genomics. New York (NY): Springer New York. p. 147–166.

Sellinger TPP, Abu Awad D, Moest M, Tellier A. 2020. Inference of past demography, dormancy and self-fertilization rates from whole genome sequence data. *PLoS Genet.* **16**:e1008698.

Sellinger TPP, Abu Awad D, Tellier A. 2021. Limits and convergence properties of the sequentially Markovian coalescent. *Mol Ecol Res.* **21**:2231-2248.

Shipilina D, Näsvall K, Höök L, Vila R, Talavera G, Backström N. 2022. Linkage mapping and genome annotation give novel insights into gene family expansions and regional recombination rate variation in the painted lady (*Vanessa cardui*) butterfly. *Genomics* **114**:110481.

Shumate A, Salzberg SL. 2021. Liftoff: accurate mapping of gene annotations. *Bioinformatics* **37**:1639–1643.

Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**:3210–3212.

Smit AFA, Hubley R, Green P. 2013-2015. RepeatMasker Open-4.0. http://www.repeatmasker.org.

Sproul JS, Hotaling S, Heckenhauer J, Powell A, Larracuente AM, Kelley JL, Pauls SU, Frandsen PB. unpublished data. Repetitive elements in the era of biodiversity genomics: insights from 600+ insect genomes. *bioRxiv*. https://doi.org/10.1101/2022.06.02.494618, last accessed November 9, 2022.

Stapley J, Feulner PG, Johnston SE, Santure AW, Smadja CM. 2017. Variation in recombination frequency and distribution across eukaryotes: patterns and processes. *Phil Trans R Soc B*. **372**:20160455.

Stanke M, Tzvetkova A, Morgenstern B. 2006. AUGUSTUS at EGASP: using EST, protein and genomic alignments for improved gene prediction in the human genome. *Genome Biol.* **7**:S11.

Storer J, Hubley R, Rosen J, Wheeler TJ, Smit AF. 2021. The Dfam community resource of transposable element families, sequence models, and genome annotations. *Mobile DNA* **12**:2.

Straatman R. 1971. The life history of the *Ornithoptera alexandrae* Rothschild. *J Lepido Soc.* **25**:58-64.

Teixeira H, Salmona J, Arredondo A, Mourato B, Manzi S, Rakotondravony R, Mazet O, Chikhi L, Metzger J, Radespiel, U. 2021. Impact of model assumptions on demographic inferences: the case study of two sympatric mouse lemurs in northwestern Madagascar. *BMC Ecol Evol.* **21**:1-18.

Tennent WJ. 2021. The man who shot butterflies. Oxfordshire: Storm Entomological Publications.

Tilak MK, Allio R, Delsuc F. 2020. An optimized protocol for sequencing mammalian roadkill tissues with Oxford Nanopore Technology (ONT) V1. *Protocols.io.* https://dx.doi.org/10.17504/protocols.io.6bthann

Tine M, Kuhl H, Gagnaire P-A, Louro B, Desmarais E, Martins RST, Hecht J, Knaust F, Belkhir K, Klages S, *et al.* 2014. European sea bass genome and its variation provide insights into adaptation to euryhalinity and speciation. *Nat Commun.* **5**:5770.

Tunstall T, Kock R, Vahala J, Diekhans M, Fiddes I, Armstrong J, Paten B, Ryder OA, Steiner CC. 2018. Evaluating recovery potential of the northern white rhinoceros from cryopreserved somatic cells. *Genome Res.* **28**:780-788.

Van Der Valk T, Gonda CM, Silegowa H, Almanza S, Sifuentes-Romero I, Hart TB, Hart JA, Detwiler KM, Guschanski K. 2020. The genome of the endangered dryas monkey provides new insights into the evolutionary history of the vervets. *Mol Biol Evol.* **37**:183-194.

Waldvogel AM, Wieser A, Schell T, Patel S, Schmidt H, Hankeln T, Feldmeyer B, Pfenninger M. 2018. The genomic footprint of climate adaptation in *Chironomus riparius*. *Mol Ecol.* **27**:1439-1456.

Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, Cuomo CA, Zeng Q, Wortman J, Young SK, *et al*. 2014. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One* **9**:e112963.

Walton W, Stone GN, Lohse K. 2021. Discordant Pleistocene population size histories in a guild of hymenopteran parasitoids. *Mol Ecol.* **30**:4538-4550.

Wang K, Mathieson I, O'Connell J, Schiffels S. 2020. Tracking human population structure through time from whole genome sequences. *PLoS Genet.* **16**:e1008552.

Wang S, Teng D, Li X, Yang P, Da W, Zhang Y, Zhang Y, Liu G, Zhang X, Wan W, *et al.* 2022. The evolution and diversification of oakleaf butterflies. *Cell.* **185**:3138-3152.

Wang Y, Obbard DJ. 2023. Experimental estimates of germline mutation rate in eukaryotes: a phylogenetic meta-analysis. *bioRxiv*. https://doi.org/10.1101/2023.01.24.525323, last accessed February 11, 2023.

Warren WC, Boggs TE, Borowsky R, Carlson BM, Ferrufino E, Gross JB, Hillier L, Hu Z, Keene AC, Kenzior A, *et al.* 2021. A chromosome-level genome of *Astyanax mexicanus* surface fish for comparing population-specific genetic differences contributing to trait evolution. *Nat Commun.* **12**:1447.

Westbury MV, Hartmann S, Barlow A, Wiesel I, Leo V, Welch R, Parker DM, Sicks F, Ludwig A, Dalén L, *et al.* 2018. Extended and continuous decline in effective population size results in low genomic diversity in the world's rarest hyena species, the brown hyena. *Mol Biol Evol.* **35**:1225-1237.

Wickham H. 2016. ggplot2: Elegant graphics for data analysis. Springer-Verlag New York Available from: http://ggplot2.tidyverse.org

Wilfert L, Gadau J, Schmid-Hempel P. 2007. Variation in genomic recombination rates among animal taxa and the case of social insects. *Heredity*. **98**:189-197.

You M, Ke F, You S, Wu Z, Liu Q, He W, Baxter SW, Yuchi Z, Vasseur L, Gurr GM, *et al.* 2020. Variation among 532 genomes unveils the origin and evolutionary history of a global insect herbivore. *Nat Commun.* **11**:2321.

Supplementary information

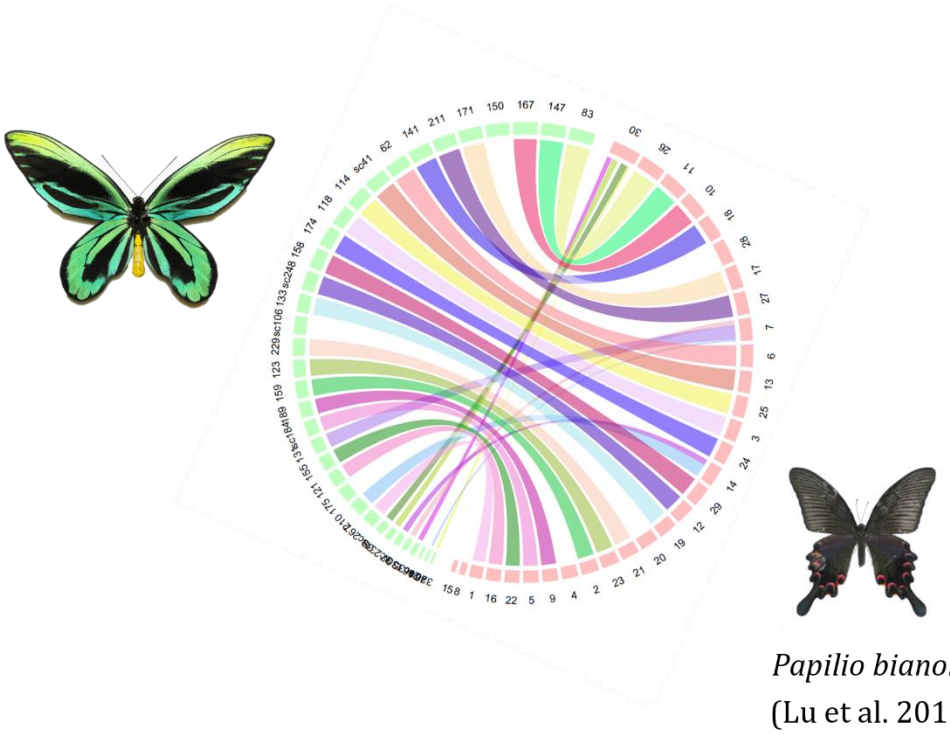


Figure S1. Chromosomes-level synteny between *Ornithoptera alexandrae* (FC563) and *Papilio bianor* (Lu et al. 2019) assemblies.

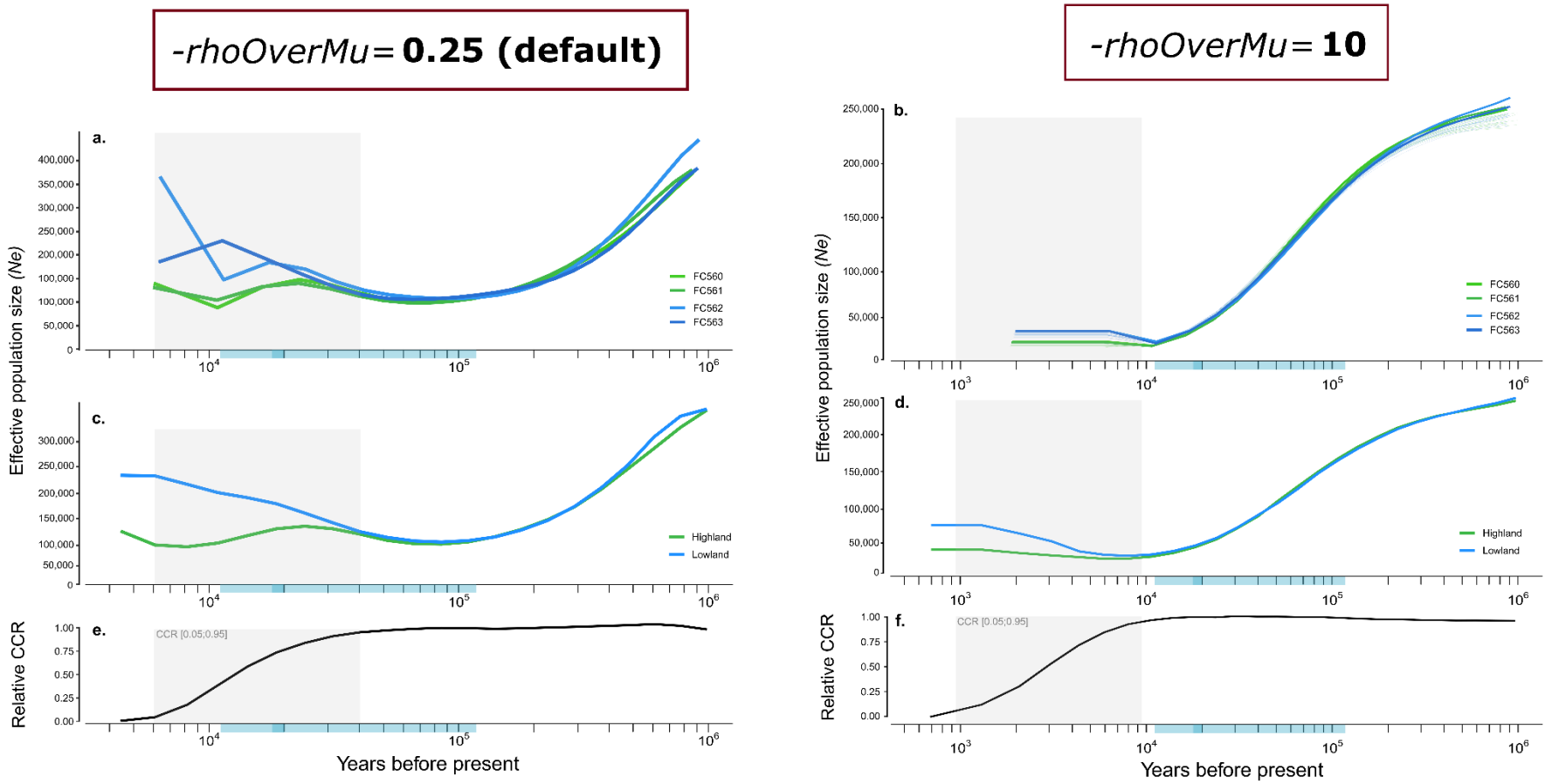


Figure S2. Comparison of MSMC2 demographic inferences with *-rhoOverMu*=0.25 (left panel) and with *-rhoOverMu*=10 (right panel). **a, b.** MSMC estimates for each individual of *O. alexandrae*. **c, d.** MSMC estimates for both populations of *O. alexandrae*. **e, f.** Relative Cross Coalescent Rate (CCR) estimated with the MSMC analysis, representing the time interval of population divergence. The gray zone represents the 5%-95% CCR has been reported on each plot. The pale blue rectangle along the time bar indicates the limits of the last glaciation period (11,700 to 115,000 years ago) with the last glacial maximum in darker blue (19-20,000 years ago).

Table S1. Name and size of the contigs and scaffolds identified as Z-chromosomes contigs for each *Ornithoptera alexandrae* assembly. For females, putative W-linked contigs were searched but none was found.

	FC560 (male)		FC561 (female)		FC562 (female)		FC563 (female)	
	Contig names	**Size**	**Contig names**	**Size**	**Contig names**	**Size**	**Contig names**	**Size**
Contigs identified as Z chromosomes	FC560_contig_305	390,746	FC561_contig_925	26,523	FC562_contig_430	23,078	FC563_contig_410	55,209
	FC560_contig_42	1,015,128	FC561_contig_941	28,322	FC562_contig_443	75,994	FC563_contig_26	540,695
	FC560_contig_4	1,133,153	FC561_contig_928	31,165	FC562_contig_533	145,148	FC563_contig_29	578,903
	FC560_contig_237	1,158,379	FC561_contig_1550	39,059	FC562_contig_433	458,309	FC563_contig_320	1,146,340
	FC560_scaffold_209	8,312,266	FC561_contig_434	55,640	FC562_contig_488	505,858	FC563_contig_296	1,447,376
			FC561_contig_807	190,270	FC562_contig_429	545,073	FC563_scaffold_35	2,797,441
			FC561_contig_199	207,446	FC562_contig_586	922,147	FC563_scaffold_231	3,442,805
			FC561_contig_825	483,740	FC562_contig_409	1,431,949	FC563_contig_39	4,093,109
			FC561_contig_202	545,398	FC562_contig_438	6,725,330		
			FC561_contig_291	565,005	FC562_contig_267	1,4559,367		
			FC561_contig_287	731,992				
			FC561_contig_819	954,722				
			FC561_contig_297	1,148,102				
			FC561_scaffold_808	3,090,565				
Total		**12,009,672**		**8,097,949**		**25,392,253**		**14,101,878**

Table S2. Proportion of annotated repeat content using RepeatModeler and RepeatMasker for each *Ornithoptera alexandrae* individual. Main repeat classes proportions are detailed in the 3rd-7th columns. LINEs = Long interspersed nuclear elements; LTRs = Long terminal repeats.

Reference genome	Total repeats	LINEs	LTRs	DNA transposons	Unclassified	Simple repeats
FC560	112 022 984 bp (34.28 %)	2.99 %	0.08 %	0.69 %	28.44 %	1.60 %
FC561	109 571 207 bp (33.65 %)	2.97 %	0.08 %	0.69 %	27.84 %	1.58 %
FC562	112 013 483 bp (34.94 %)	3.02 %	0.07 %	0.70 %	29.10 %	1.57 %
FC563	113 279 483 bp (35.27 %)	3.03 %	0.08 %	0.70 %	29.38 %	1.60 %

Table S3. Pairwise mitochondrial difference between each *Ornithoptera alexandrae* individual. A) Mitogenomic diversity for the whole mitochondrial alignment. B) Mitogenomic diversity for the COI-5P region only. The calculation of the mean pairwise distance and the corresponding π-diversity are reported at the right sides of the Supplementary tables.

A) Alignment length (Mitochondrial) 15,137 bp

	FC560	FC561	FC562
FC561	7	-	
FC562	10	13	-
FC563	9	12	1

Calculation

mean pairwise distance	(7+10+13+9+12+1) / 6	≈	8.6667
π-diversity	8.6667 / 15137	≈	**0.0005725**

B) Alignment length (COI-5P) 658 bp

	FC560	FC561	FC562
FC561	1	-	
FC562	1	0	-
FC563	1	0	0

Calculation

mean pairwise distance	(1+1+1) / 6	=	0.5
π-diversity	0.5 / 658	≈	**0.00075987**

Data type	Window sizes / filters	Populations/ Species	Number of sites	Median Nucleotide diversity (%)	Coefficient of variation of Nucleotide diversity
Illumina / freebayes	100 kb / contig > 1 MB / < 10% missing	**Highland**	200 Mb	0.0763	0.2895089
		Lowland		0.0793	0.2785817
Illumina / freebayes	1 MB / contig > 1 MB / < 10% missing	**FC560**	280 Mb	0.0748	0.1902523
		FC561		0.0745	0.2032465
		FC562		0.0777	0.2142416
		FC563		0.0780	0.1772522
ONT / Longshot	1 MB / contig > 1 MB / < 10% missing	**FC560**	280 Mb	0.0799	0.1949459
		FC561		0.0789	0.2058297
		FC562		0.0836	0.215576
		FC563		0.0836	0.1816642
		FC565	298 Mb	0.4334	0.1022
		FC569	272 Mb	0.0737	0.2547
ONT / Longshot	100 kb / contig > 1 MB / < 20% missing data	**Highland**	311 Mb	0.0770	0.3490045
		Lowland		0.0804	0.3373432
ONT / Longshot	CDS (4 fold degenerated site)	**FC560**	3 Mb	0.0793	
		FC561		0.0805	
		FC562		0.0826	
		FC563		0.0838	
		FC565	2.4 Mb	0.7081	
		FC569	1.9 Mb	0.07035	

Table S4. Calculated median autosomal diversity using different thresholds, windows sizes, methods, and data for *Ornithoptera alexandrae*, *O. priamus* and *Troides oblongomaculatus*. Diversities were estimated according to genotype caller (Longshot or freebayes) and data type (Illumina or ONT).

Table S5: Results of demographic inferences from four independent DILS runs using randomly selected 5,000 windows of 4,000 bp across the genomes of *Ornithoptera alexandrae* (highland and lowland population). For each run, the best scenario chosen by DILS is displayed (strict isolation [SI], isolation with migration [IM], ancient migration [AM]). 1M means that the migration is homogeneous among loci, 2N means that the variation of the effective population size is heterogeneous among loci. The 95% height posterior density (HPD) and median of the estimated time of split in provided.

Run	Model	Time of split (Tsplit - neural network)			Predominant migration direction
		2.5% HPD	Median	97.5% HPD	
1	SI_2N	4,445	**5,104**	5,464	NA
2	IM_1M_2N	15,502	**16,850**	18,296	Low to High
3	AM_1M_2N	5,897	**7,200**	8,522	High to Low
4	IM_1M_2N	10,319	**12,319**	14,195	High to Low

Table S6: Summary statistics about ONT sequencing, Illumina sequencing, assembly, polishing and BUSCO assessment. Abbreviation: Nb = number, seq = sequenced, QS = quality score, Comp = complete, S = single, D = duplicated, Frag = fragmented, Miss = missing.

Samples		Sequencing								Assembly											
		Oxford Nanopore Technology						Illumina		Coverage		Polishing	Assembly statistics (from BUSCO)						BUSCO v5 -gVolante n:5286			
Code	Tissue	Nb. flow cells	Raw Gb seq	Nb passed reads (QS >10)	% of the total nb of reads	Mean QS of passed reads (Guppy)	N50 of reads after Porechop	Gb seq.	Number of filtered reads (fastp)	Long reads	Short reads		Nb. Scaffolds	Sum length	N50	Median length	Max length	Mean length	Comp. S (%)	Comp. D (%)	Frag. (%)	Miss. (%)
FC560	adult (RNAlater)	2	30.98	7.99M	82.4%	14.54	6027	33.2	220.59M	66	NA	NA	333	326574548	10698373	8897	20022694	980704	98.1	0.2	0.6	1.1
											84	Pilon x2	333	326746405	10703829	8902	20033826	981220	98.6	0.2	0.3	0.9
FC561	larva (RNAlater)	4	31.82	7.46M	79.6%	14.31	4679	31.0	205.54M	59	NA	NA	1222	325401291	7756035	2155	12845025	266286	97.6	0.2	0.6	1.3
											79	Pilon x2	1222	325591695	7761365	2154	12853001	266442	98.6	0.2	0.2	1
FC562	adult (RNAlater)	2	41.18	13.68M	82.2%	14.79	4420	33.8	224.67M	95	NA	NA	465	320433737	9992309	2468	14553139	689105	98.2	0.2	0.5	1.1
											70	Pilon x2	465	320565121	9990636	2469	14559367	689387	98.7	0.2	0.2	0.9
FC563	larva (RNAlater)	2	29.44	6.73M	84.0%	14.86	6192	37.6	249.64M	70	NA	NA	305	321019789	11235072	3181	13864591	1052524	98.2	0.2	0.5	1.1
											95	Pilon x2	305	321134305	11239331	3187	13869660	1052899	98.6	0.2	0.2	1
FC565	adult (RNAlater)	3	38.99	9.21M	77.2%	14.14	4744	36.9	244.46M	86	NA	NA	862	316787646	4956205	3289	10541647	367503	98.1	0.2	0.6	1.1
											96	Pilon x2	862	316500205	4951043	3290	10532357	367170	98.4	0.3	0.3	1
FC569	adult (RNAlater)	3	38.34	7.20M	79.4%	14.27	7305	57.4	379.87M	87	NA	NA	465	348071568	5906186	4840	13643788	748541	97.8	0.3	0.8	1.1
											137	Pilon x2	465	348219047	5909187	4846	13649974	748858	98.5	0.2	0.3	1
-	**Mean**	**2.67**	**35.125**	**8.71M**	**80.8%**	**14.49**	**5561**	**38.3**	**254.13M**	**77.2**	**93.5**	-	**609**	-	**8424964**	**4140**	**14247422**	**684220**	**98.28**	**0.22**	**0.43**	**1.05**

Table S7: Number of contigs removed and their phylum rank (Plant, Microsporidia or Pseudomonadota) after Blobtools analysis. The proportion displayed for each individual is the number of contigs removed for this individual over the total number of contigs removed for the six individuals (272).

	Genus	FC560	FC561	FC562	FC563	FC565	FC569	Total of contigs	Participation in contig removal (%)
Plant	*Aristolochia*	0	2	0	0	0	0	**2**	**0.74%**
Microsporidia	*Nosema*	20	150	1	0	4	0	**175**	**64.34%**
	Enterospora	1	10	0	0	0	0	**11**	**4.04%**
	Hepatospora	1	4	0	0	0	0	**5**	**1.84%**
	Pseudoloma	1	18	0	0	0	0	**19**	**6.99%**
	Anncaliia	0	29	0	0	0	0	**29**	**10.66%**
	Hamiltosporidium	0	13	0	0	0	0	**13**	**4.78%**
Pseudomonadota	*Enterobacter*	0	0	2	0	0	5	**7**	**2.57%**
	Wolbachia	0	0	0	0	11	0	**11**	**4.04%**
	Participation in contig removal (%)	**8.46%**	**83.09%**	**1.10%**	**0.00%**	**5.51%**	**1.84%**	**272**	

Table S8: Statistics of phased-block length of each individual based on the VCF-files produced by Longshot 0.4.1 (Edge and Bansal 2019).

	Phased-block statistics		
	Median	**Mean**	**N50**
FC560	241,522.5	453,200.64	997,034
FC561	66,846.5	145,535.28	331,165
FC562	58,083.5	441,643.51	1,684,646
FC563	86,028	554,597.54	1,951,149
FC565	12,037.5	508,556.90	3,911,890
FC569	225,368	498,947.01	1,240,361
Mean	**114,981**	**433,746.81**	**1,686,040.83**
Standard dev.	**95,066.23**	**146,941.63**	**1,227,496.50**

CHAPTER 2.1



La soif de *Graphium antheus* - Eliette Reboud

Deciphering the systematics and evolutionary history of the recalcitrant tribe Leptocircini (Lepidoptera: Papilionidae) with whole-genome data

Authors

Eliette L. Reboud[1]*, Benoit Nabholz[1,2], Emmanuelle Chevalier[1], Bérénice Lafon[1], Marie-ka Tilak[1], Carlos G. C. Mielke[3], Adam M. Cotton[4] & Fabien L. Condamine[1]*

Affiliations

[1] *Institut des Sciences de l'Evolution de Montpellier (Université Montpellier | CNRS | IRD | EPHE), Place Eugène Bataillon, 34095 Montpellier, France*
[2] *Institut Universitaire de France (IUF), Paris, France*
[3] *Caixa postal 1206, 84.145-000 Carambeí, Paraná, Brazil*
[4] *86/2 Moo 5, Tambon Nong Kwai, Hang Dong, Chiang Mai, Thailand*

Corresponding authors (*)

Eliette L. Reboud: eliette.reboud@gmail.com

Fabien L. Condamine: fabien.condamine@gmail.com

Running head

Phylogenomic of the tribe Leptocircini (Papilionidae)

Abstract

Leptocircini is a dazzling tribe of Papilionidae, including dragontails, kite swallowtails, and swordtails. This tribe shows a wide distribution notably throughout the tropics of Africa, Southeast Asia, and the Neotropics, making it a fascinating subject in evolutionary biology. However, in spite of accounting for a quarter of the global swallowtail butterfly diversity, Leptocircini have been surprisingly neglected in global phylogenetic analyses. This has left unanswered questions about their taxonomy (number and validity of species) and systematics (number of subgenera, genera, and the relationships between them). Here, we present a new taxonomic working list for Leptocircini, featuring 152 valid species. Using a combination of long and short reads data, we produced five new reference genomes, and we generated highly covered and scaffolded whole genomes for 149 individuals to infer densely sampled phylogenetic hypotheses. Based on mitochondrial or thousands of nuclear genes and multiple phylogenetic approaches, a robust phylogenomic tree is recovered, representing >95% of the Leptocircini species, which allowed examining several key phylogenetic hypotheses. We found the monotypic genus *Protographium* as sister of genus *Graphium*. Additionally, we found that subgenera *Paranticopsis* and *Pathysa* are paraphyletic, which we attribute to an ancient and continuous gene flow. Our results also allowed revising the most recently proposed phylogenetic structure of *Eurytides sensu lato*, into three subgenera. Our analyses finally raised concerns about potential taxonomic inflation in two species groups of *Graphium* and within *Protesilaus*. Overall, our study sheds light on the clade's evolutionary history and paves the way for further research on this diverse and charismatic tribe of butterflies.

Keywords: *Eurytides*, gene flow, *Graphium*, incomplete lineage sorting, phylogenomics, reference genome, scaffolding.

Introduction

In Lepidoptera, swallowtail butterflies (Papilionidae) of the tribe Leptocircini Kirby, 1896, include divers-looking species such as dragontails, kite swallowtails, and swordtails. They comprise about 160 worldwide distributed species, which represents about a quarter of all Papilionidae (about 600 recognized species; Zakharov et al. 2004, Haüser et al. 2005, Condamine et al. 2012, Nakae, 2021). The phylogenetic relationships of the tribe Leptocircini remain highly uncertain and have never been the subject of comprehensive and thorough phylogenetic studies, contrary to some well-studied swallowtail tribes (Papilionini: Ae 1979, Hancock 1983, Igarashi 1984, Aubert et al. 1999, Caterino and Sperling 1999, Reed and Sperling 1999, Yagi et al. 1999, Zakharov et al. 2004, Condamine et al. 2013a, 2013b, Lewis et al. 2015, Wu et al. 2015, Owens et al. 2017, Condamine et al. 2023 = **Appendix II**; Troidini: Silva-Brandao et al. 2005, Condamine et al. 2012, 2013a, 2015; Parnassiini: Nazari et al. 2007, Michel et al. 2008, Condamine et al. 2018, He et al. 2023, Tian et al. 2023).

Nevertheless, Leptocircini were the subject of several morphological studies and revisions based mainly on genitalia and wing veins in the 1980s-2000s (e.g. Munroe 1960, Hancock 1983, Miller 1987, Smith and Van-Wright 2001, Makita et al. 2003, Hancock 2006). Recently, the genus *Graphium* Scopoli, 1777 has been the focus of molecular work in localized parts of the phylogeny to describe new species, often based on the mitochondrial COI gene or a few genes (e.g. *Pazala*: Hu et al. 2019, Zhang et al. 2020, *Graphium*: Page and Treadaway 2014, Cotton et al. 2022). Leptocircini also appeared in several attempts of global phylogeny of Papilionidae, but usually with a poor sampling fraction (e.g. Makita et al. 2003 [25%], Simonsen et al. 2011 [6.8%], Page and Treadaway 2014 [24%], Zhang et al. 2019 [29%], Zhang et al. 2021 [6.8%]) and none of these phylogenies are sufficiently robust to provide conclusive evidence about the systematics. The most recent comprehensive phylogenetic work dealing with the Papilionidae and including Leptocircini was conducted by Allio et al. (2021), and included seven mitochondrial and nuclear genes. However, the sampling of Leptocircini only included 42% of the tribe and the phylogenetic relationships within this tribe were not robust enough to discuss the recovered topology and its consistency with the existing literature. This apparent neglect is quite surprising for a group as studied and popular as the Papilionidae, but probably stem from multiple explanations. Indeed, knowing that Papilionidae are scarce compared to Nymphalidae or Pieridae, Leptocircini are even more rare and known to be particularly difficult to collect in the field due to their still largely unknown behaviors and traits, the speed and height of their flight, and their small size (Collins and Morris 1985, Smith and Vane-Wright 2001, Nakae 2021).

Leptocircini are distributed worldwide

Most of the species richness of Leptocircini is contained in the genus *Graphium*, which alone includes ~110 of the 160 species of Leptocircini. As a typical component of the Old World tropics, *Graphium* is composed of five subgenera including a species-rich African subgenus (*Arisbe* Hübner, [1819]) and four Indomalayan-Australasian subgenera (*Graphium* Scopoli, 1777, *Paranticopsis* Wood-Mason and Nicéville, [1887], *Pathysa* Reakirt, [1835], *Pazala* Moore, 1888) (**Figure 1**). The Australian monotypic genus *Protographium* Munroe, 1961 is thought to be sister to *Graphium* based on its morphology (Munroe 1960, Munroe and Erlich 1960, Nakae 2021). The genus *Eurytides*, Hübner [1821] is the second richest genus in Leptocircini, with >40 species, and exclusively found in the Americas. The systematics and taxonomy of both *Graphium* and *Eurytides* have undergone many changes and uncertainties, with numerous subgenera or genera having been proposed (see taxonomic details in the following paragraph). Here, our working hypothesis follows the latest global taxonomic assessment of Nakae (2021), which considers five subgenera for *Graphium: Arisbe*, *Graphium*, *Paranticopsis*, *Pathysa*, *Pazala* (see details hereafter) and four subgenera for *Eurytides*: *Neographium* Möhn, 2002, *Eurytides* Hübner, [1821], *Protesilaus* Swainson, [1832], *Mimoides*, Brown 1991.

External to *Protographium* and these two diverse groups are two less diverse genera including three species each: (1) the genus *Iphiclides*, Hübner [1819] is the only Leptocircini with an Eurasian distribution, and (2) the genus *Lamproptera*, Gray 1832, called dragontail butterfly, in a south-east Asian clade with a divergent and peculiar morphology (**Figure 1**).

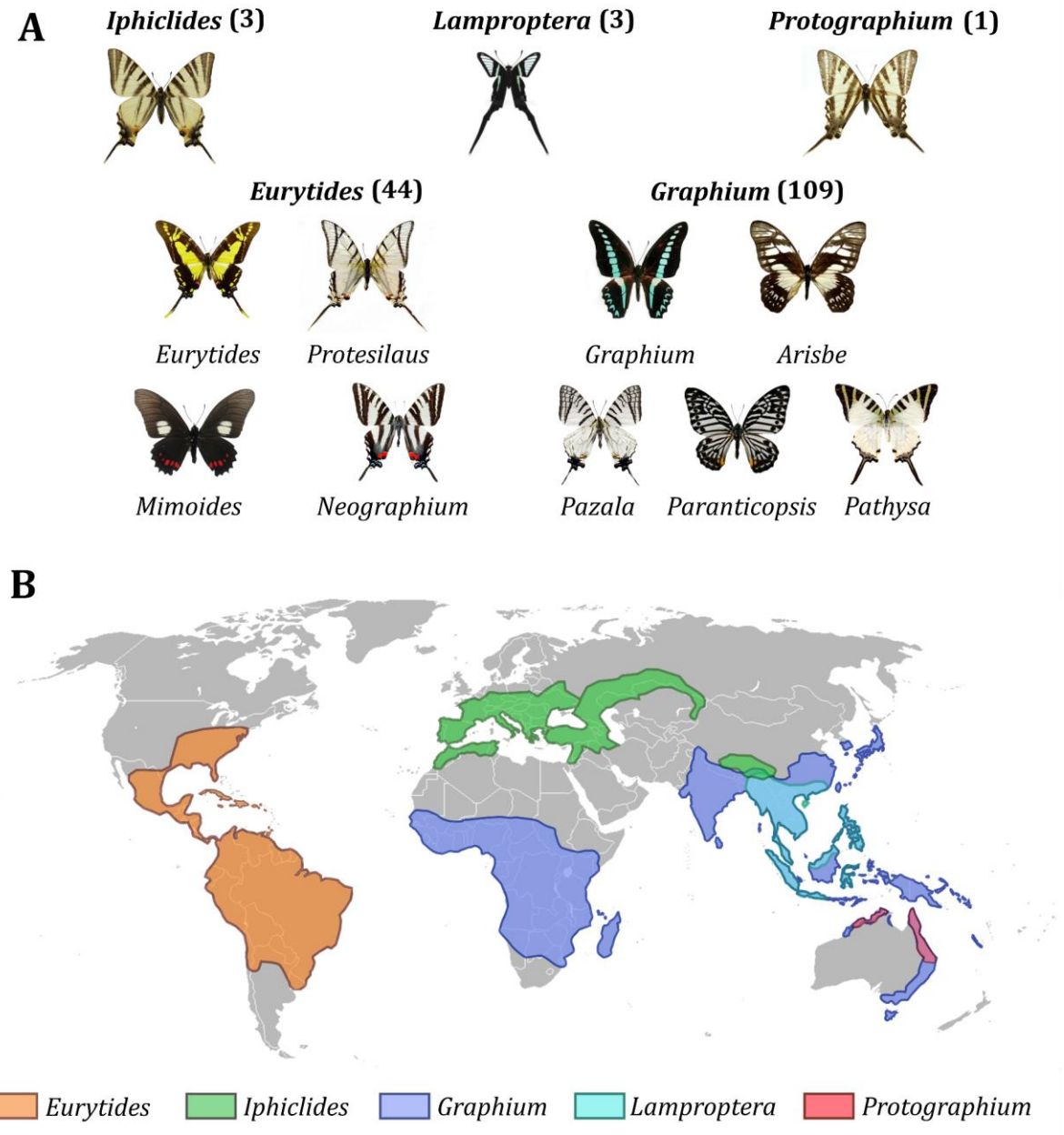


Figure 1: Overview of Leptocircini and their global distribution. A. Habitus of Leptocircini genera (bold font) and subgenera (regular font). The current species richness is indicated for each genus between parentheses. B. Distribution pattern of Leptocircini genera. The distribution map of each genus is drawn approximately according to literature and *iNaturalist* occurrences (www.inaturalist.org) and GBIF (https://www.gbif.org/fr/). Butterfly pictures: *Iphiclides podalirius*, *Lamproptera curius*, *Protographium leosthenes*, *Eurytides (E.) thyastes*, *E. (Protesilaus) telesilaus.*, *E. (Neographium) marcellus*, *E. (Mimoides) ariarthes*, *Graphium (Arisbe) adamastor*, *G. (G.) milon* , *G. (Paranticopsis) delesserti*, *G. (Pazala) wenlingae.*, *G. (Pathysa) antiphates.*

Leptocircini have a complex and still unstable taxonomic history.

The following paragraph describes in detail the taxonomic histories of the genera Graphium *and* Eurytides*, as this study follows on from considerable morphological and taxonomic work and attempts to finally resolve the remaining gray areas. It therefore seems important that these should be presented precisely in order to address them.*

The genus *Graphium* has been the subject of several morphological and molecular studies, attempting to understand its classification and evolutionary relationships. However, there is still controversy and unresolved questions despite these efforts. To our knowledge, all studies have shown the genus *Graphium* (= *Graphium s.l.*) to be monophyletic (e.g. Makita et al. 2003, Allio et al. 2021), and despite several studies that upgraded subgenera to genera (e.g. Niculescu 1977, Igarashi 1984, D' Abrera 1982, Page and Treadaway 2014), the genus *Graphium s.l.*, as including several subgenera, is now commonly accepted and used by the experts and non-experts alike. However, the classification and validity of subgenera within the genus *Graphium s.l.* remain highly controversial, despite multiple morphological and molecular analyses. The subgenus *Pazala* is usually considered as monophyletic and valid and placed as the sister group to the other *Graphium* subgenera (Hancock 1983, Miller 1987) but sometimes not (Makita et al. 2003, Allio et al. 2021). However, the subgenera *Arisbe*, *Graphium* (= *Graphium s.s.*), *Paranticopsis* and *Pathysa* are subjects of more complex taxonomic history.

Saigusa et al. (1982) considered *Graphium s.s.* as a monophyletic clade with three species groups: *sarpedon*, *agamemnon*, and *eurypylus*. However, no species of *Arisbe* were included in their analysis, which was later considered as a limitation, as Hancock (1993) found that tailed *Arisbe* species were more closely related to *Graphium s.s.* He therefore divided *Arisbe* into non-tailed species (*Arisbe*) and tailed species (*antheus* like) and included these latter in *Graphium s.s.* Smith and Vane-Wright (2001) tested this hypothesis in a cladistic framework, but were "unable to resolve African and Oriental species groups". On the other hand, several studies, instead of questioning *Arisbe*'s monophyly, have rather considered the monophyly of *Graphium s.s.* as doubtful, and in particular the placement of the *eurypylus* species group. Indeed, the monophyly of *Graphium s.s.* was not supported in Makita et al. (2003), the *eurypylus* species group being sister to *Arisbe*. Page and Treadaway (2014) suggested a narrower definition of *Graphium s.s.*, including only certain species groups (*sarpedon*, *codrus*, *weiskei*, *wallacei* and *agamemnon* groups) and established a generic status for *Arisbe* and subgeneric status for *eurypylus*, *Pazala*, and *Pathysa* (all included in this broader genus *Arisbe*). Recently, Allio et al. (2021) found a polyphyletic *Graphium s.s.* subgenus, but the *eurypylus* species group was found sister to *Pazala* and the other *Graphium s.s.* species group, and not sister to *Arisbe*. Recent studies based on

molecular data have been unable to resolve the issue and have been inconclusive, perhaps due to limited species representation.

Several studies considered *Paranticopsis* to be included in *Pathysa* (Munroe and Erlich 1960, Hancock 1983, Miller 1987). Smith and Vane-Wright (2001) used *Paranticopsis* as a valid subgenus, but the study only included *G. macareus*, which showed relatedness to the African clade *Arisbe*, rather than *Pathysa* and *Pazala*. Later, Makita et al. (2003), Page and Treadaway (2014) and Wilson et al. (2014) and Allio et al. (2021) all observed polyphyletic patterns or unresolved relationships between *Paranticopsis* and *Pathysa*, but their phylogenies were poorly supported.

Thus, classification and relationships within genus *Graphium* remain complex and controversial, with various studies showing contradictory results. Further analyses, including a denser species sampling for all the different species groups, are needed to provide clearer answers for these questions.

A sister-group relationship between *Lamproptera* and *Graphium* was proposed by Hancock (1983) and Miller (1987) following Ehrlich (1957) on the basis of larval and pupal morphological characters (Racheli and Cotton 2009). This relationship was recovered based on morphological characters in Smith and Vane-Wright (2001) (but without any *Eurytides* included) and based on morphological and molecular data in Page and Treadaway (2014). However, the relationships between *Lamproptera* and *Graphium* were described as unstable by Makita et al. (2003), as their molecular data showed that the *Eurytides* and *Iphiclides* group were more closely related to *Graphium* than *Lamproptera*. Yet they did not formally question the "traditional" *Lamproptera* + *Graphium* group that was proposed at the time. Simonsen et al. (2011) found that *Graphium* instead forms the sister group to the Neotropical Leptocircini (*Eurytides s.l.*, see after) and that the genera *Iphiclides* and *Lamproptera* were early-diverging clades of Leptocircini but with a non-robust relationship. Finally, Allio et al. (2020a, 2021), found *Lamproptera* sister to *Iphiclides*, forming a sister clade to all other Leptocircini but again, with rather weak support even with genomic data.

The Neotropical Leptocircini have never been the focus of any well sampled study. They appeared in several previously cited studies about *Graphium s.l.* (e.g. Makita et al. 2003, Page and Treadaway 2014), but as outgroups and represented by only a few species. They also appear in several general studies of Papilionidae (Simonsen et al. 2011, Allio et al. 2021), where up to 18 species (Allio et al. 2021) were sampled (44%). Based on all these studies, Neotropical Leptocircini seem to form a monophyletic group, but it still has a complex internal classification history.

The genus *Eurytides* (= *Eurytides s.l.*) was proposed by Munroe (1960), who divided it into two subgenera *Protesilaus* and *Eurytides s.s.* They were long thought to be closely related to the Australian *Protographium leosthenes* (that Munroe (1960), also erected as a monospecific genus), and some *Eurytides* species were

even placed in this genus by Brown (1991). The two subgenera *Protesilaus* and *Eurytides s.s.*, proposed by Munroe (1960), were elevated as genera by Hancock (1983), then downgraded again as subgenera by Miller (1987), and brought back to genera by Brown (1991). Brown (1991) also described *Mimoides* as a genus including the Neotropical mimetic Leptocircini species (that were previously placed by Munroe (1960) in the *Protesilaus* subgenus). Lately, Zhang et al. (2019) propose to downgrade *Mimoides* to a subgenus of *Eurytides s.l.* to account for their close relationship with *E. marcellus*.

Möhn (2002) excluded back Neotropical species from the *Protographium* genus, and proposed a new genera *Neographium* Möhn, 2002, for the north, central and antillean species with two subgenera *Neographium* Möhn, 2002 (*philolaus* species group) and *Asiographium* Möhn, 2002 (monospecific, type *E. asius*). He also described two new subgenera of *Eurytides*: *Bellerographium* Möhn, 2002 (monospecific, type *E. bellerophon*), *Eurygraphium* Möhn, 2002 (*thyastes* species group). Callaghan et al. (2004) synonymized some of these names in *Protographium* (apparently, *Neographium*, *Asiographium*, *Eurygraphium*, but I was not able to access the checklist). Yet, the type specimen of *Protographium* is the Australian *P. leosthenes*, which appears to be unrelated to *Eurytides*, and therefore this synonymization is probably invalid. Finally, Zhang et al. (2021) proposed the monospecific subgenus *Boreographium* Grishin, 2021 (monospecific, type *E. marcellus*), based on a limited sampling, to account for the proximity of *E. marcellus* (formerly included in *Neographium*) with *Mimoides*.

Nowadays experts tend to use *Eurytides* as a generic name with several subgenera (as suggested at least by Simonsen et al. 2011) and the commonly encountered subgenera are *Protesilaus*, *Mimoides*, *Eurytides s.s.* and *Neographium* (e.g. Nakae 2021). Yet, non-experts still regularly use subgenera as genus names, and employ the genus *Protographium* to identify some Neotropical species.

A well-sampled phylogenomic dataset could help to resolve the tree of Leptocircini

To date, all phylogenetic studies of Leptocircini lack a good species sampling or are limited to morphological or a few molecular markers, which have resulted in highly unstable classification and many unresolved relationships. It has long been shown that phylogenetic studies based on a handful of molecular markers result in poorly supported phylogenetic relationships due to a limited number of phylogenetically informative characters (Wiens and Penkrot 2002, Funk and Omland 2003, Wiens et al. 2010, Ross 2014, Mutanen et al. 2016). On the other hand, genomic data can alleviate some of the phylogenetic issues by expanding the number of characters from a few thousand to hundreds of thousands (Delsuc et al. 2005, Philippe and Blanchette 2007; Pennisi 2008), and help to produce better resolved phylogenetic trees and particularly robust backbones. For instance, using

genome-scale data associated with a reference genome of *Papilio xuthus* (Li et al. 2015), Allio et al. (2020a) were able to retrieve thousands of genes and inferred a robust phylogenetic backbone for swallowtail butterflies. Although this study lacked the sampling to constitute a complete species-level phylogeny, it represents a promising approach to combining reference genomes and highly fragmented whole genomes. Indeed, in the era of high-quality complete genomes that are flourishing (Formenti et al. 2022), the swallowtail butterflies are no exception. However, the genus *Papilio* remains the most studied, with 20 complete genomes published since 2015 (e.g. Lu et al. 2019 and reference therein), while the first high-quality genomes of Leptocircini were only published recently (*Lamproptera curius*, He et al. 2022; *Iphiclides podalirius*, Mackintosh et al. 2022). However, there is no reference genome for *Graphium* available yet, even though this genus represents the second most speciose genus of all Papilionidae. Furthermore, a reference genome for *Graphium* can provide crucial data for the genomics of swallowtail butterflies since the expected size of their genomes is larger than those of their sister genera (Allio et al. 2020a).

In this study, we first establish a list of valid Leptocircini species based on previous molecular and morphological studies and then infer a robust and near complete species-level phylogeny for the tribe. We sequenced and assembled five new reference genomes of Leptocircini, including four *Graphium* species, and 143 *de novo* shotgun whole genomes, which together allow us to extract thousands of nuclear genes and mitogenomes for around 93% of the total specific diversity of Leptocircini. Our study aims to provide a reference phylogenetic framework for assessing the monophyly of genera and subgenera as well as species relationships within Leptocircini. For this objective, our study includes for the first time all enigmatic and long-debated groups and species to be robustly placed in the swordtail tree of life, such as *Protographium leosthenes*, *Graphium phidias*, and other species that will be key for testing diversification and biogeographic hypotheses (e.g. *Protographium leosthenes*, *Eurytides marcellus*), the evolution of mimicry (e.g. *Mimoides*, *Paranticopsis*) and host-plant associations (e.g. *Pazala*, *Paranticopsis*, *Lamproptera*, *Iphiclides*), as well as rare and understudied species (e.g. *Arisbe*, *Eurytides s.l.*). On the basis of this unprecedented dataset, we also explored the hypothesis of gene flow and incomplete lineage sorting to explain topological discordances within *Pathysa* and *Paranticopsis*.

Materials and Methods

Species list and taxon sampling

We first established a taxonomic working list of Leptocircini species (**Table S1**). This species list takes into account some recent taxonomic list (Nakae 2021) and the results of recent studies (Hu et al. 2019, Zhang et al. 2020, Cotton et al. 2022, Huang 2023,

and this study). Due to the limited general phylogenetic studies conducted on Leptocircini butterflies recently, it is likely that they will continue to receive attention as a model clade for taxonomic studies, and we expect that many species boundaries will be tested with new genomic data, or at a populational level, which could ultimately revise this taxonomic working list.

Of the 160 species currently featured in the new taxonomic list presented here (including all "valid" and "dubious" and "probably invalid" status), we sampled 152 species representing 95% of the total diversity (**Table S1**), using a combination of fresh and museum specimens. However, the total number of species might evolve with further systematic studies, as the validity of some species is still debated (e.g. conflicts within *Pazala*, Hu et al. 2018, 2019 *versus* Huang 2023). The missing species were essentially concentrated in the *Pazala*, *Arisbe* subgenera, some of them being extremely rare or of doubtful status (**Table S1**) We added 11 outgroup species representing several swallowtail genera to root the Leptocircini tree based on large-scale phylogenetic studies of Papilionidae (Condamine et al. 2012, Allio et al. 2020a, 2021): *Baronia brevicornis*, *Papilio xuthus*, *Parnassius apollo*, *Ornithoptera alexandrae*, *Teinopalpus imperialis*, *Meandrusa payeni*, *Bhutanitis thaidina*, *Sericinus montela*, *Luehdorfia chinensis*, *Hypermnestra helios* and *Archon apollinus*. The complete taxon sampling and accession numbers are provided in **Table 1**.

Sequencing was performed in ~50x short-reads Illumina sequencing based on previous estimates of genome sizes (Allio et al. 2020b, He et al. 2022). Among these sampled species, five reference genomes were produced with a combination of long reads and short reads data: *Iphiclides podalirius* (collected in the wild), *G. antheus* for subgenus *Arisbe*, *G. agamemnon* and *Graphium doson* for subgenus *Graphium*, and *G. antiphates* for subgenus *Pathysa*, all issued from rearing.

DNA extractions and library preparations for whole genome sequencing

The extraction of most specimens was done on museum and collections specimens (taking legs, rarely abdomens). The extraction and library preparation were performed with the DNAeasy Blood and Tissue kit (QIAGEN), applying the methods of Meyer and Kirscher (2010), but with slight modifications as described in Tilak et al. (2015). DNA quality and concentration and length fragment varied markedly among specimens. Final DNA purity and concentrations were measured using both Nanodrop (Thermo Fisher, USA) and Qubit (Thermo Fisher, USA) and resulting libraries were analyzed for size distribution by Agilent 2200 TapeStation. Illumina 150 bp paired-end sequencing was run on a NovaSeq 6000 instrument to obtain a genome depth-of-coverage of about 50x (varying from 20 to 60 Gb per library depending on the genus).

For reference genomes, DNA extractions were done with butterflies killed and stored in a freezer at -20°C without any additional preservation product. Tissues from the thorax were used to extract high-molecular weight DNA.

Following Reboud et al. (2023) who tested two different extraction methods, we used the Qiagen genomic DNA kit to obtain a better 260/230 ratio as estimated with Nanodrop assays guaranteeing DNA purity for long-read sequencing with Oxford Nanopore Technology (ONT). Final DNA purity and concentrations were measured using both Nanodrop (Thermo Fisher, USA) and Qubit (Thermo Fisher, USA). Whole-genome libraries were prepared using the resulting high-molecular-weight DNA as input for the Nanopore LSK-109 ligation kit (Oxford Nanopore Technologies, UK) following the manufacturer's protocol. Long-read (LR) sequencing was performed on a GridION device with two to six R9.4.1 flow cells. Remaining DNA extractions were sent to Novogene Europe (Cambridge, UK) for two short-read Illumina library preparations. Libraries were generated using NEBNext DNA Library Prep Kit following manufacturer's recommendations and indices were added to each library. Genomic DNA was randomly fragmented to a size of 350 bp by shearing, then DNA fragments were end-polished, A-tailed, and ligated with the NEBNext adapter for Illumina sequencing, and further PCR enriched by P5 and indexed P7 oligos. The PCR products were purified (AMPure XP system) and the resulting libraries were analyzed for size distribution by Agilent 2100 BioAnalyzer and quantified using real-time PCR. Since the genome sizes for *Graphium* was estimated to be about 1 Gb and *Iphiclides* about 390 Mb (Allio et al. 2020a), Illumina 150 bp paired-end sequencing was run on a NovaSeq 6000 instrument to obtain about 100 and 40 Gb, respectively, corresponding to a genome depth-of-coverage of ~100x after combining the two libraries.

Assembling reference genomes

For each reference genome (*G. agamemnon, G. antheus, G.doson, G. antiphates, I. podalirius*), raw LR sequence data (fast5 files) were basecalled using Guppy 5.0.15 (Oxford Nanopore Technology) with the super-high accuracy mode and a quality control of 10 (min_score 10). Sequencing adapters were trimmed using Porechop 0.2.3 (https://github.com/rrwick/Porechop).
Draft genome assemblies were performed with the LR assembler Flye 2.8.3 (Kolmogorov et al. 2019) with default options. For *Graphium* genomes, duplicated haplotigs and heterozygous contig overlaps were removed from the draft assembly using purge_dups 1.2.5 (Guan et al. 2020) based on the LR depth (**Figure 2**).

The Illumina raw reads were cleaned, filtered, and paired using fastp 20.0 (Chen et al. 2018) with default options (**Figure 2**). To improve base accuracy and reduce assembly errors, the Flye draft assemblies were polished using short-reads (SR) with POLCA (Zimin and Salzberg 2020) implemented in MaSuRCA 4.0.1 (Zimin et al. 2013). Assembly statistics were then assessed using the gVolante2 platform (Nishimura et al. 2017) to retrieve the number and size of contigs, the presence, completeness and duplication of BUSCO genes of the Lepidoptera *odb10* database (Manni et al. 2021). We checked

for possible contaminations using BlobTools 1.1.1 (Laetsch and Blaxter 2017) set to the *ncbi* and *diamond* databases. We found no evidence of artificial contamination coming from laboratory manipulation, but some contigs were clearly identified as belonging to exogenous organisms such as symbionts. We removed all contigs that belonged to plants or symbionts (Wolbachia, Ascomyota, Acetobacteraceae, Bacteroidota, and Microsporidia).

Assembling shotgun whole genomes

The Illumina SR were filtered and paired with fastp 19.5-21.0 (Chen et al. 2018) using the default options (**Figure 2**). The assemblies were performed by MEGAHIT 1.2.7-1.2.9 (Li et al. 2015) and submitted to the gVolante platform to obtain assembly statistics as well as assessing the amount of genes found from the *odb10* database of Lepidoptera (BUSCO 5, Manni et al. 2021). The gVolante platform retrieves BUSCO genes classed in "single copy", "multicopy" and "fragmented" categories. To improve the contiguity and the number of BUSCO genes recovered in these draft assemblies, they were scaffolded with the scaffold option of RagTag (Alonge et al. 2021) when a close reference genome was available (same species group or same subgenus). Species of the genera *Eurytides* and *Protographium* did not have any close reference genome. In that case, only assemblies with less than 25% BUSCO completeness were scaffolded on the reference *Graphium agamemnon* to retrieve more BUSCO genes. To compare draft genome assemblies with scaffolded genome assemblies, the resulting assemblies were again submitted to the gVolante platform. The BUSCO single copy genes in fasta nucleotide file of each individual were retrieved using BUSCOMP 0.13 software (Edwards 2019).

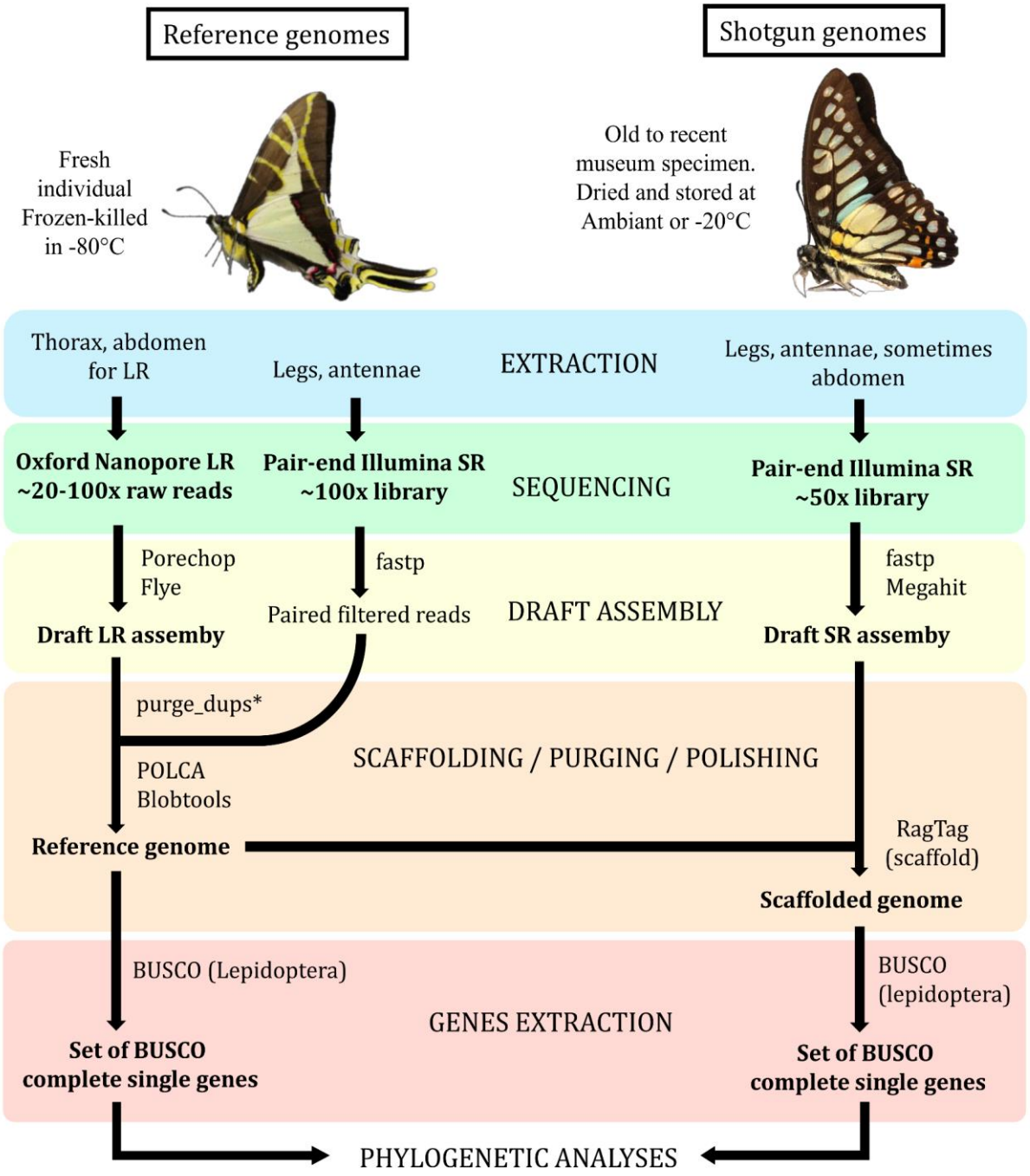


Figure 2: Conceptualization of the sequencing and assembly pipeline for reference genomes (left) and shotgun genomes (right) used to construct phylogenomic datasets of nuclear data, based on long reads (LR) and short reads (SR) data.

Assembling mitogenomes and phylogenetic reconstruction

For reference genomes, each *Graphium* individual, LR were mapped with Minimap 2.17 (Li 2018) on the mitogenome of *G. xenocles* (MZ394042) for *G. agamemnon*, *G. antiphates*, *G. doson*, and on the mitogenome of *G. leonidas* (FC1205, this study) for *G. antheus*. The reads that mapped with the reference were filtered by quality via SAMtools (Li et al. 2009) (*view -q 30*). For each individual, a subset of reads was created so that mitogenomes would have an expected depth of coverage of 200× (3.6 Mb). We used Flye 2.8.3 to assemble the mitogenomes and the resulting assemblies were given to MitoFinder 1.4 (Allio et al. 2020b) to annotate and extract thirteen protein-coding genes (ATP6, ATP8, COX1-3, CYTB, ND1-4, ND4L, ND5-6) and 2 non-coding genes (rrnL, rrnS). The complete mitogenome of *Iphiclides podalirius* was identified as a contig of the whole-genome assembly and was also selected for annotation by MitoFinder.

For shotgun assemblies, Megahit draft-assemblies of each species were given to MitoFinder (Allio et al. 2020b) to extract and annotate their mitogenome. Some assemblies only displayed fragmented mitogenomes, pseudogenes, fungi contamination or cross-contamination. When the identification of "real" mitogenomic contigs were not possible manually, GetOrganelle (Jin et al. 2020) was used with a close reference for mapping to reconstruct complete, or less fragmented, mitogenomes. This method was only used on difficult individuals as the computation time of GetOrganelle is longer and not necessary when MEGAHIT already reconstructed the full mitogenome. For every species, we used MitoFinder to annotate and extract genes. For missing species, we retrieved the COI-5P gene or complete mitochondrion from GenBank (see **Table S2** for more details).

We selected 11 outgroup species as previously described (see *Species list and taxon sampling* section), either using the mitogenome of the same individual as for the nuclear analyses (*Baronia brevicornis*, *Ornithoptera alexandrae*, *Archon apollinus*, *Hypermnestra helios*; see **Table S2**) or by taking the complete mitogenome of the same species on GenBank (*Bhutanitis thaidina*, *Parnassius apollo*, *Teinopalpus imperialis*, *Meandrusa payeni*, *Papilio xuthus*, *Sericinus montela*, *Luehdorfia chinensis*; see **Table S2**).

To account for frameshift (both artificial and biological) and stop codons, sequences were aligned with MACSE 2.07 (Ranwez et al. 2018), setting an invertebrate mitochondrial genetic code (*-gc_def 5*), a frame-shift penalty of 5 (*-fs 5*) and an internal stop codon penalized by 10 (*-stop 10*). Maximum-likelihood (ML) phylogenetic inference was implemented with IQ-TREE 2.2.0 (Minh et al. 2020a) using ModelFinder to select the best-fit partition scheme and the best-fitting substitution model for each partition (*-m MFP+MERGE* option, Chernomor et al. 2016, Kalyaanamoorthy et al. 2017) applied to an initial subset of 41 possible partitions (the three codon-position of the 13 protein-coding genes and the 2 non-coding genes). Node robustness of the ML

tree was evaluated with 1,000 ultrafast bootstraps (UFBS; *-B 1000* option, Hoang et al. 2018).

Phylogenomic analyses with nuclear data

Phylogenetic analyses of genome-scale data were based on BUSCO genes. For each of the 5,286 lepidopteran core genes, we assessed the number of represented individuals, and applied two different sampling criteria on these genes to form two datasets. *Dataset 1* corresponds to all genes that gather at least 145 out of 159 individuals (91.2% of sampling for each gene), which corresponds to 1,402 genes. The *Dataset 2* was produced with a threshold of 50% (80/159 individuals) sampling for each gene, which corresponds to 4,525 genes.

The nucleotide sequences of the 1,402 genes (*Dataset 1*) were aligned with MAFFT 7.453 (Katoh and Standley 2013). For each alignment, a gene tree was reconstructed with IQ-TREE with the best-fitting substitution model for each gene (*-m TESTNEW* option) and 1,000 UFBS. A consensus tree of these gene trees was produced with ASTRAL 5.7.7 (Zhang et al. 2018) following a tree-reconciliation (or "supertree approach").

A supermatrix approach was also performed on both Datasets. First, the nucleotide alignments of *Dataset 1* were concatenated into a single matrix. Maximum-likelihood phylogenetic inference was carried out with IQ-TREE using ModelFinder to select the best-fit partition scheme and the best-fitting substitution model for each partition (*-m MFP+MERGE* option), on an initial subset of 1,402 possible partitions (one partition per gene). Node robustness was evaluated with 1,000 ultrafast bootstraps (UFBS; -B 1000 option). *Dataset 1* was also analyzed with an amino-acid supermatrix, and the phylogeny was reconstructed with IQ-TREE using the LG protein model (*-m LG+G4+F,* Le and Gascuel 2008) and node robustness evaluated with 1,000 UFBS. These same options were also used with a concatenated matrix of 4,525 amino acid genes (*Dataset 2*).

Finally, because traditional statistical measures of nodal support are prone to inflation in phylogenomic data sets (Minh et al. 2020b), we performed gene concordance factor (gCF) and site concordance factor (sCF) analysis implemented in IQ-TREE 2 (Minh et al. 2020a), using respectively the gene trees of *Dataset 1*, and the nucleotide supermatrix of *Dataset 1*, in comparison with the consensus tree of the IQTREE_1402_NT analysis (**Table 1**).

Table 1: Summary of the five different phylogenomic analyses and their abbreviations. The asterisk corresponds to the analysis from which the matrix and the resulting topology were used in the CFs analysis.

	Mitogenome	***Dataset 1***	***Dataset 2***
Nucleotide	IQTREE_15_Mito	IQTREE_1402_NT*, ASTRAL_1402_NT	
Amino acid		IQTREE_1402_AA	IQTREE_4525_AA

Analysis of incomplete lineage sorting and gene flow

To investigate the hypotheses that incomplete lineage sorting (ILS) or gene flow (GF) have caused topological conflicts in the subgenera *Pathysa/Paranticopsis* (see *Results*), we used Aphid (Galtier 2023), a maximum-likelihood method that aims at quantifying the sources of phylogenetic conflict via topology and branch length analysis of rooted three-species gene trees. Given a triplet of species A, B, C and outgroup species, and assuming a main topology ((A, B), C), Aphid classifies all gene trees into five categories: no event detected (the gene tree follow the main topology ((A, B), C) and its branching time), non discordant ILS (same topology but branching times are higher than in the main topology), non discordant GF (same topology but branching times are smaller than in the main topology), discordant ILS (noted ILSc (c for conflict), where topologies differ from the main topology, and branching times are higher), and discordant GF (GFc, topologies differ from the main topology and branching times are smaller). For the two last cases, Aphid also gives the "discordant topology imbalance" associated with the event (predominance of ((A, C), B) or ((B, C), A)). The rationale is that GF events occur after population isolation and therefore result in shorter branch lengths, whereas ILS events are caused by ancestral polymorphisms older than population isolation. Alphid can therefore identify GF even when alternative topologies occur at similar frequencies, unlike the statistics derived from the ABBA/BABA test.

All gene trees that could be rooted with *Baronia brevicornis* were selected for the analyses (1,330 gene trees). To investigate the potential ILS and GF that could lead to the observed paraphyly of *Pathysa* and polyphyly of *Paranticopsis*, we focused our tests on five species groups: *Pathysa 1* (*dorcus-rhesus* species group), *Pathysa 2* (*androcles-euphrates*), *Paranticopsis 1* (*delesserti-ramaceus*), *Paranticopsis 2* (*encelades-thule*), and *Pathysa 3* (*G. agetes* and *G. stratiotes*). Four configurations of triplets were tested, each requiring different A, B, C species (**Figure 3**). In our case, A, B and C did not correspond to single species but to entire clades (species-groups). To account for the impact of selecting only one species as a representative of an entire species-group, the analysis of each triplet was repeated with five different A, B and C species randomly selected in each species group (**Table S3**). For each series of triplets, we calculated the mean of the five replicates tested. In all analyses, *Graphium antheus* was selected as outgroup of the triplets. Aphid was run with default options.

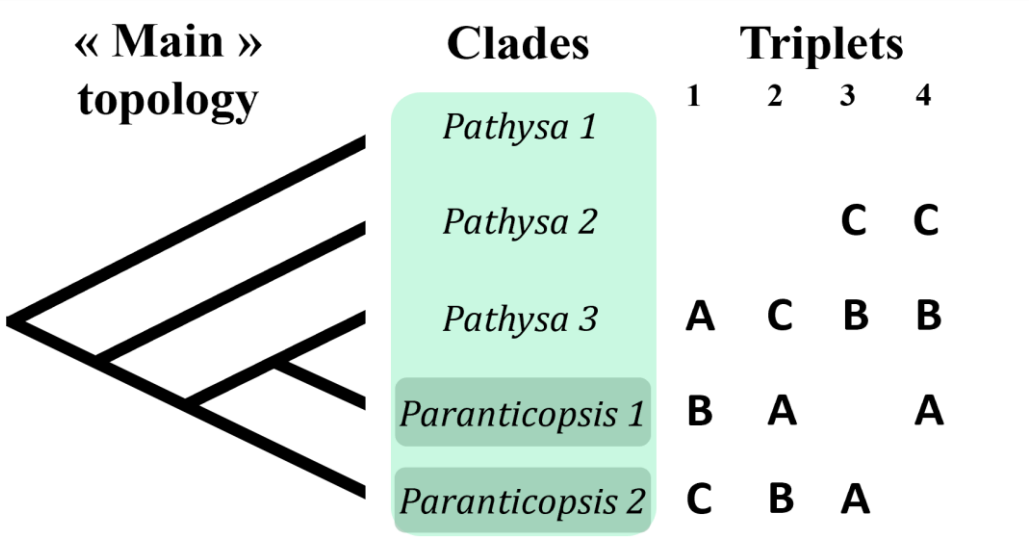


Figure 3: Definition of the four series of triplets for the Aphid analyses. On the left is the observed phylogenetic relationships between the different clades, considered as the "main" topology. On the right are the four tested series of ABC triplets for Aphid. The first conforms to the main topology, the second assumes that *Paranticopsis* 1 and 2 shared parts of their genome because of a speciation event. The third and fourth allow testing hypotheses of gene flow or ILS with *Pathysa* 2.

Results

Reference genomes

We assembled *de novo* genomes of five Leptocircini species representing the genera *Iphiclides* (*I. podalirius*), and *Graphium* with the subgenera *Arisbe* (*Graphium antheus*), *Pathysa* (*G. antiphates*) and *Graphium* (*G. agamemnon*, *G. doson*) using long and short reads data (**Table 2**). Final depth of coverage estimates of long-reads (LR) assemblies ranged from 12 to 66× coverage, and short-reads (SR) polishing (Illumina) of these genomes from ~66 to 110× coverage. Overall, all assemblies had high BUSCO scores (between 97.3% and 98.8% complete BUSCO genes) but moderate N50 (70 kb - 1.1 Mb, **Table 2**). The sizes of the assembled genomes varied from 469 Mb (*I. podalirius*) to 2,052 Mb (*G. antiphates*). Two assemblies (*G. antheus*: 1,686 Mb; *G. antiphates*: 2,052 Mb) had higher size than expected in the literature, even after having removed the detected artificial duplication in the assembly through the *purge_dup* step (see *Materials and Methods*) and both assemblies still had >12% of duplicated BUSCO genes (**Table 2**). This could be due to highly heterozygous genomes but could not be corrected with other genome assemblers either.

Table 2: Assembly statistics for the five *de novo* reference genomes of Leptocircini.

	Iphiclides podalirius	*Graphium*			
		G. agamemnon	*G. antheus*	*G. antiphates*	*G. doson*
	FC536	**FC1381**	**FC1382**	**FC1341**	**FC873**
Raw data sequenced (Gb) (LR + SR)	40.0 + 20.0	31.1 + 103.1	30.9 + 103.8	101.9 + 105.3	20.6 + 71.1
Final mean coverage (LR + SR)	66x + 96x	19x + 110x	12x + 68x	32x + 68x	14x + 105x
Assembly size (bp)	468,946,765	960,973,235	1,685,715,539	2,052,007,334	989,610,395
Number of contigs	2,755	8,469	22,326	28,705	36,745
N50 (bp)	1,107,841	236,847	149,287	173,984	70,611
Max length (bp)	13,306,297	2,403,220	1,528,824	9,291,578	1,128,770
Nucleotid assembly BUSCO score (%)	S:97.0; D:1.8; F:0.3; M:0.9	S:96.5; D:1.9; F:0.8; M:0.8	S:85.1; D:13.1; F:0.9; M:0.9	S:85.8; D:12.3; F:0.9; M:1.0	S:93.2; D:4.1; F:1.5; M:1.2

Shotgun whole-genome assemblies

We assembled 147 *de novo* whole Leptocircini genomes using Illumina pair-end short-reads data. Assembly with MEGAHIT resulted in draft assemblies with an average Complete single-copy genes BUSCO score of 57.9% (**Figure 4**). The mean of Fragmented and Missing genes was 16.7% and 24.0% respectively. Mean assemblage sizes vary greatly according to genus: 424 Mb for *Protographium*, 489 Mb for *Iphiclides*, 839 Mb for *Lamproptera*, 839 Mb for *Eurytides*, and 1,093 Mb for *Graphium*. After a scaffolding step with RagTag with a phylogenetically close reference genome when available, the average complete single-copy genes BUSCO of all assemblies improved to 73.0% (**Figure 4**) and that of Fragmented and Missing genes BUSCO score decreased to 10.8% and 15.3% respectively.

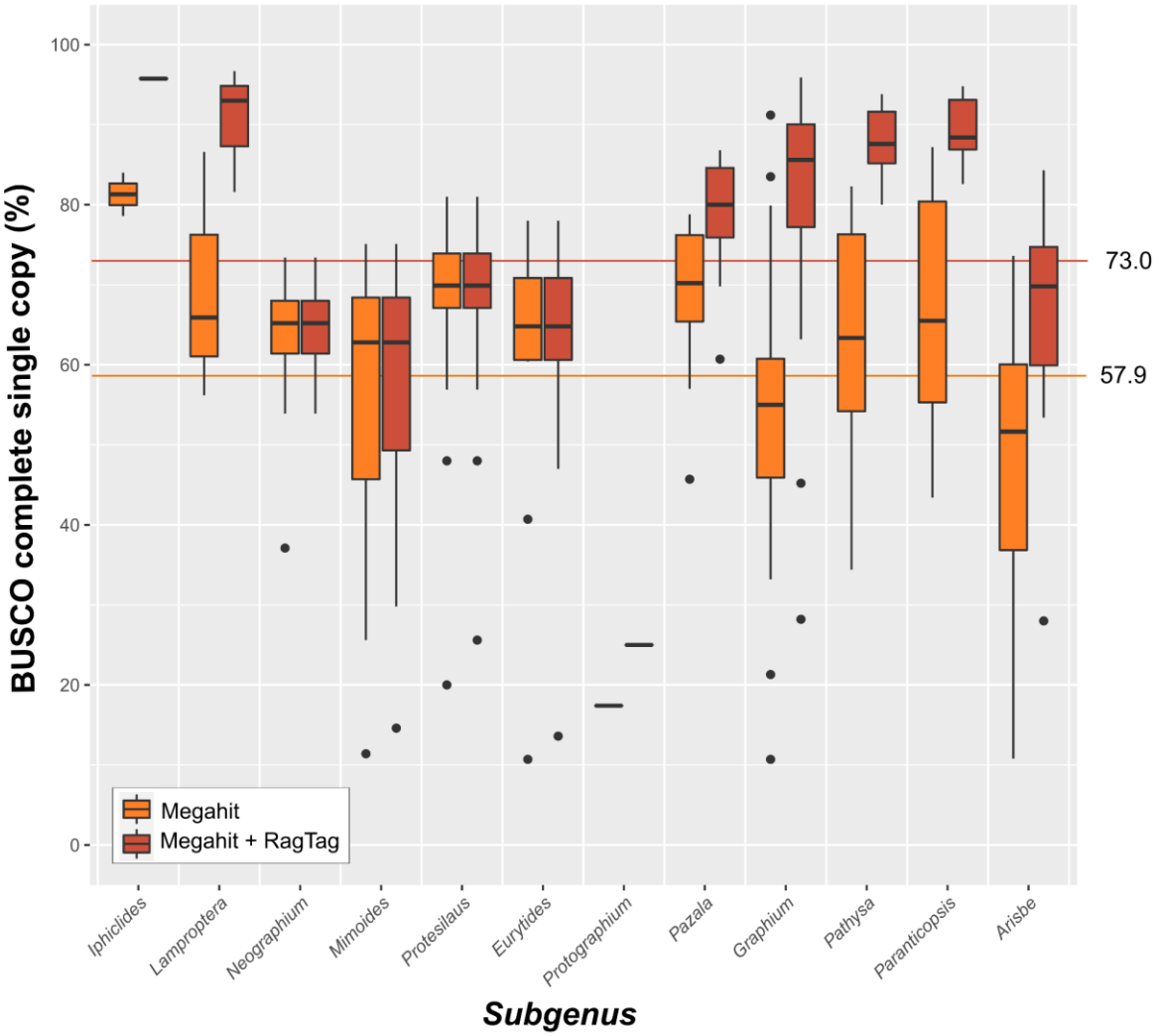


Figure 4: Effect of scaffolding fragmented shotgun whole genomes on the recovery of BUSCO genes. Boxplots of BUSCO single complete genes values of all assembly (*n*=147 specimens) within a subgenus, before (orange) and after (red) the scaffolding. Horizontal colored lines correspond to the mean BUSCO single complete score of all individuals before (orange) and after (red) the scaffolding. Draft assemblies of subgenera *Neographium*, *Mimoides*, *Protesilaus* and *Eurytides* were only scaffolded when the initial BUSCO score was under 25%.

Phylogenomics of Leptocircini

All genome-scale phylogenetic analyses provided very similar phylogenetic trees, with identical backbones, but differed in branch length estimates and relationships in some species groups (**Figure 5**, **Figures S1, S2, S3, S4**).

All analyses recovered high mean and median node support (**Figure 6**). For instance, the IQ-TREE analysis of the *Dataset 1* (supermatrix of 1,402 nucleotide genes for a total matrix of 3.5 Mb, "IQTREE_1402_NT") showed a mean node support of 99.5 and a median of 100, with 94.2% of nodes having maximal node support (**Figure 6**).

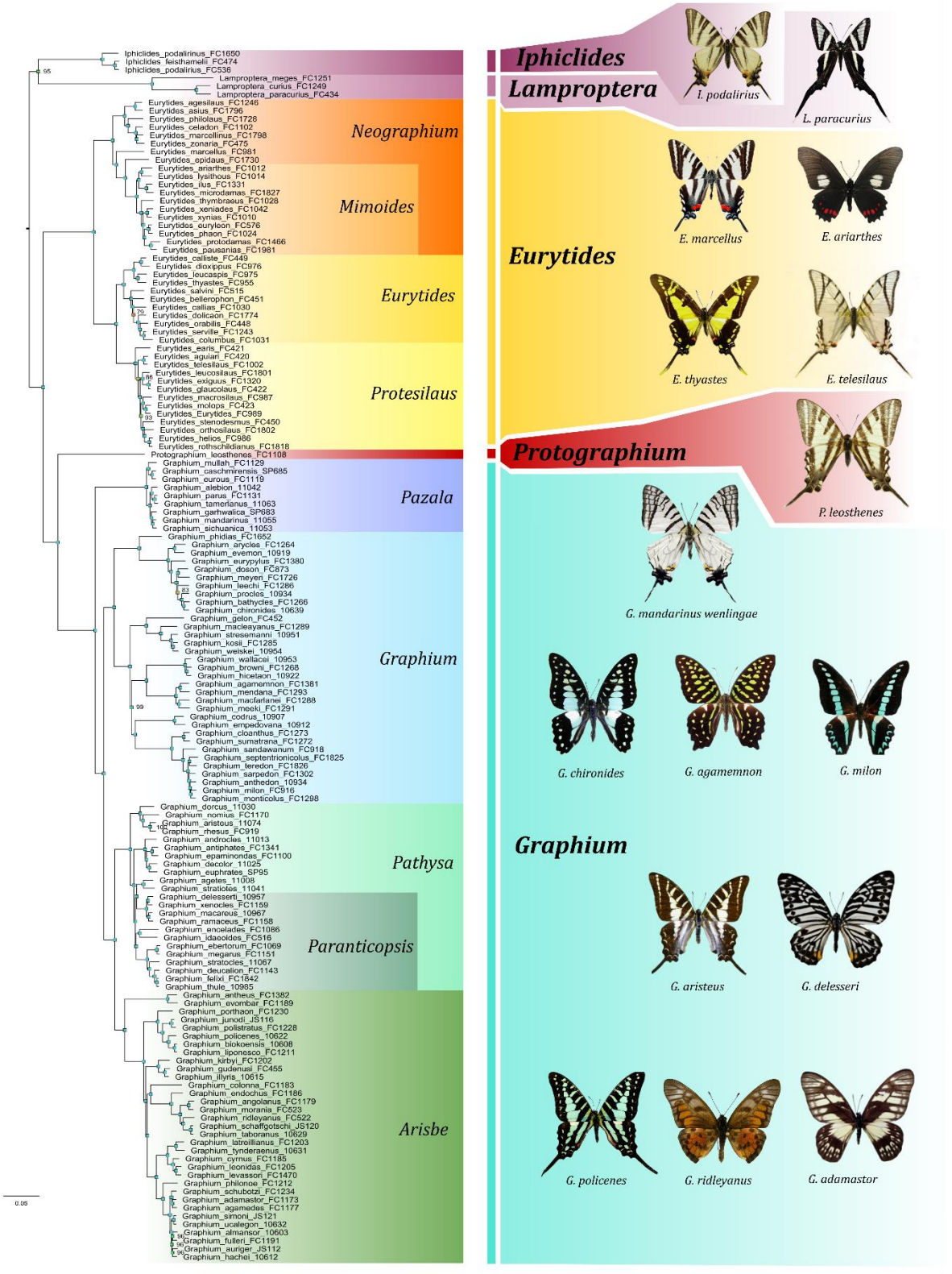


Figure 5: Phylogenomic relationships of Leptocircini (*IQTREE_1402_NT*). The phylogeny has been reconstructed with a supermatrix approach of 1,402 single-copy genes and a maximum-likelihood inference in IQ-TREE. Light blue squares at nodes correspond to a bootstrap value equal to 100. Values <100 are indicated. Genera are in italic bold font, and subgenera in regular italic. Illustration credit: *L. paracurius* (Adam Cotton), *G. mandarinus wenlingae* (Hu et al. 2019), *G. milon, G. aristeus* (Alex Dumchus), *G. ridleyanus* (CC-BY-NC Thomas Desloges).

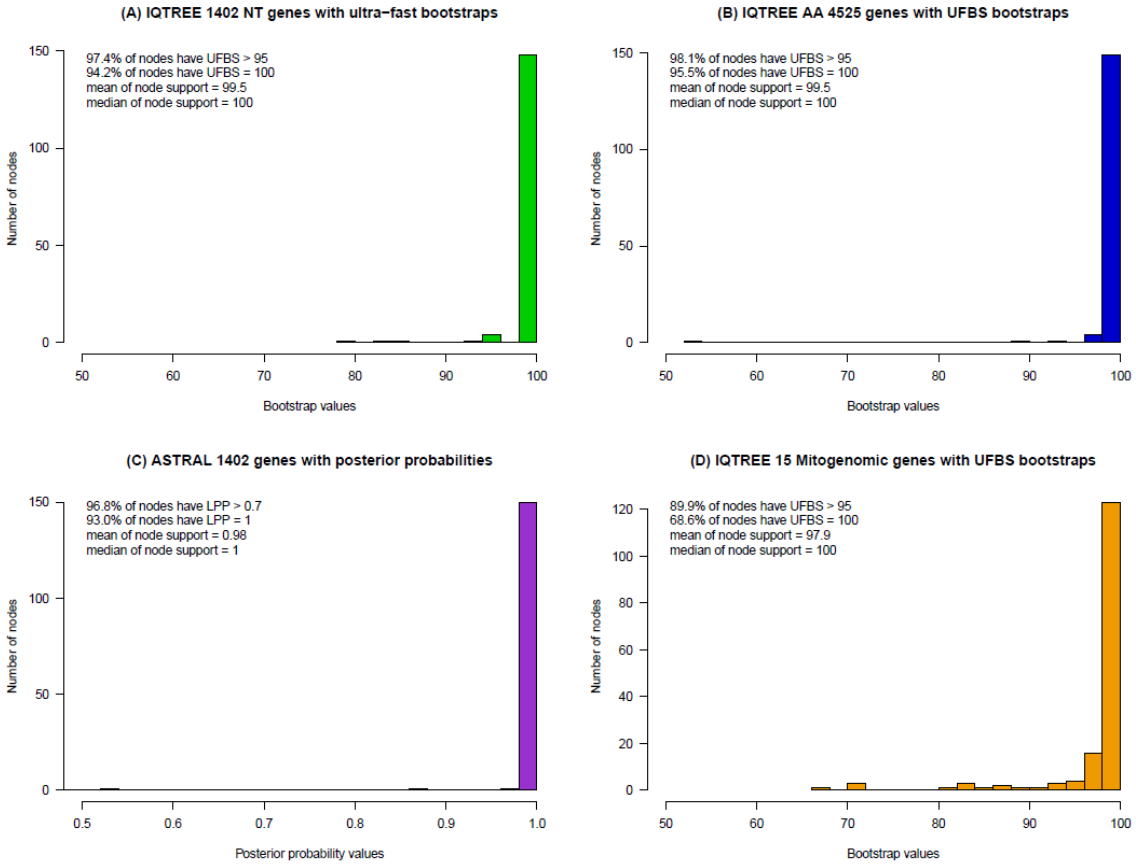


Figure 6: Node support of the different phylogenomic analyses. **A**. Histogram of UFBS values of IQ-TREE phylogeny with 1,402 genes (nucleotides), corresponding to the tree in Figure 5, **B**. Histogram of UFBS values of IQ-TREE phylogeny with 4,525 genes (amino acids), corresponding to tree of Figure S2, **C**. Histogram of LPP values of ASTRAL phylogeny with 1,402 genes (nucleotides), corresponding to the tree in Figure S1, **D**. Histogram of UFBS values of IQ-TREE phylogeny with 15 mitochondrial genes (nucleotides), corresponding to the tree in Figure S4.

The number of final partitions in IQ-TREE analysis was reduced to 183 (compared to the initial 1402 possible partitions) for the IQTREE_1402_NT analysis, and to 16 partitions (compared to the initial 41 partitions) for the mitochondrial analysis. Maximum likelihood trees of the nucleotide *Dataset 1* (IQTREE_1402_NT, **Figure 5**; ASTRAL_1402_NT, **Figure S1**) and amino-acid *Dataset 2* (IQTREE_4525_AA, **Figure S2**) recovered strong support for the monophyly of all genera: *Iphiclides* ($UFBS_{NT}$/$UFBS_{AA}$/LPP = 100/100/1), *Lamproptera* (100/100/1), *Eurytides* (100/100/1), and *Graphium* (100/100/1). *Protographium* was recovered sister of *Graphium* with maximal support (100/100/1), except in the mitochondrial analysis, where *Iphiclides* was sister of *Protographium* ($UFBS_{mito}$= 59), and both were placed as sister of *Graphium* the with low support ($UFBS_{mito}$ = 71).

Within genus *Eurytides*, the subgenera *Mimoides*, *Eurytides* and *Protesilaus* were recovered as

monophyletic with strong supports ($UFBS_{NT}$/$UFBS_{AA}$/LPP = 100/100/1 for the the crown node of each subgenus), but the subgenus *Neographium* as defined by Nakae (2021) was found paraphyletic in all analyses, with *Eurytides epidaus* sister to all *Mimoides* (100/100/1), and *E. marcellus* sister of *Mimoides* + *E. epidaus* (100/100/1).

Within genus *Graphium*, the subgenera *Pazala*, *Graphium* and *Arisbe* were recovered as monophyletic with strong supports ($UFBS_{NT}$/$UFBS_{AA}$/LPP = 100/100/1 for the the crown node of each subgenus), but the subgenus *Pathysa* as delimited by Nakae (2021) was found paraphyletic in all analyses (100/100/1, **Figure 5**, **Figure S5**), and the subgenus *Paranticopsis* was found polyphyletic in all analyses (100/97/1, see **Figure 5**), separated into two distinct clades: the *delesserti* species group, found sister to *G. agetes* (*Pathysa*) + *G. stratiotes* (*Pathysa*), and the *megarus* species group.

Overall, topological conflict between trees were mostly limited to shallow nodes with short internode lengths within well-supported clades or deeper nodes with low branch support, especially in the *Eurytides* subgenus *Protesilaus* (**Figure S6**), the *idaeoides/encelades* species within the *Graphium* (*Pathysa/Paranticopsis*) subgenera (**Figure S5**), the *adamastor* species group (*Graphium* subgenus *Arisbe*, **Figure 5, S1, S2, S3, S4**), and in the *Graphium* subgenus *Graphium* (**Figure 5, S1, S2, S3, S4**).

Concordance factors

Gene and site concordance factors (gCF and sCF) had systematically lower values than UFBS (see **Figure S7** for comparison of $UFBS_{NT}$/gCF/sCF). The median and mean values were respectively 77.42 and 68.49 for the gCF, and 55.80 and 56.48 for the sCF. Several clades had overall poor gCF and sCF values (**Figure 7**, **Figures S7, S8**), such as the sister relationship between *Lamproptera* and *Iphiclides*, or deep nodes in the subgenera *Arisbe*, *Graphium* and *Pathysa/Paranticopsis*, or recent nodes within species group, such as the *Eurytides* (*Protesilaus*) clade, the *adamastor* species group (subgenus *Arisbe*), and the *sarpedon* species group (subgenus *Graphium*).

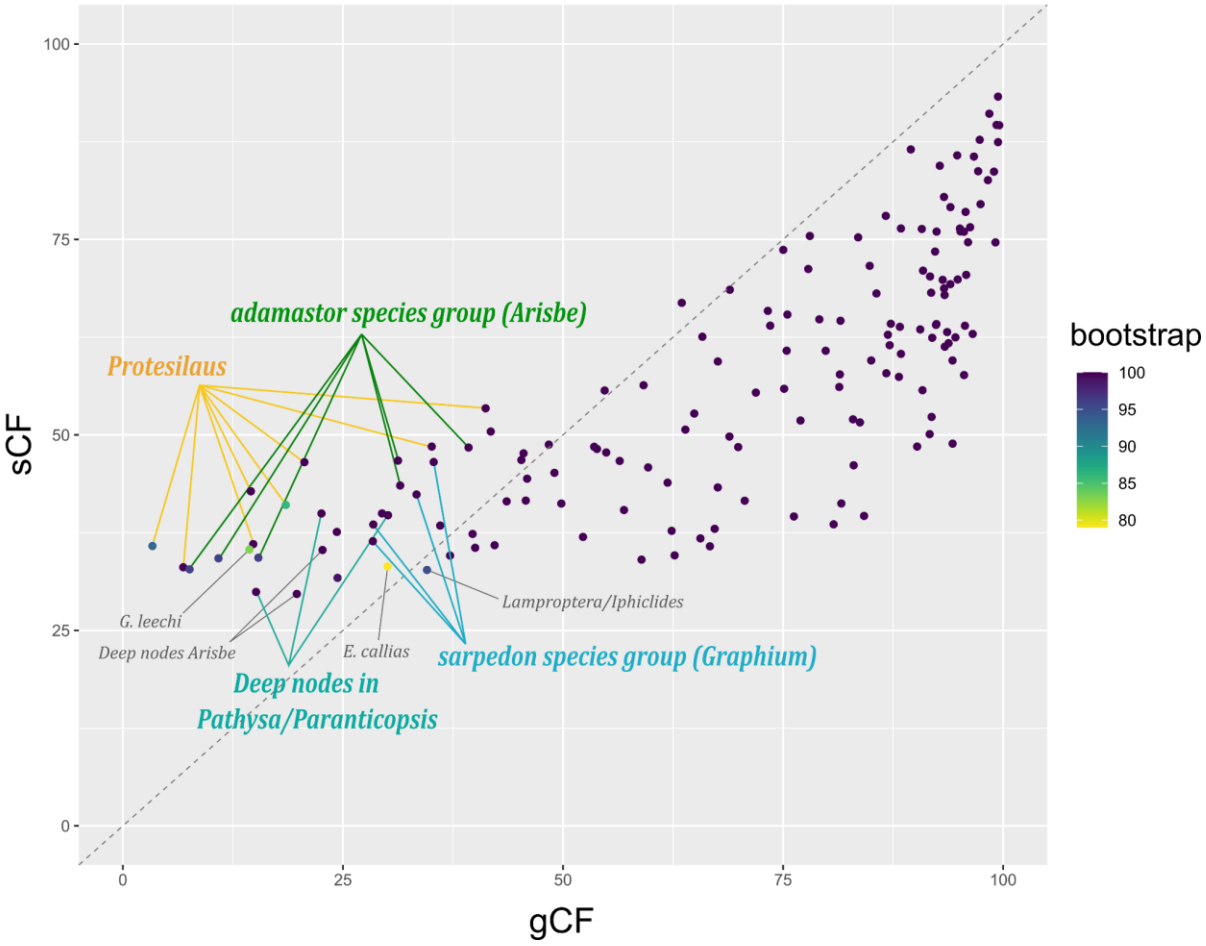


Figure 7: Values of gCF and sCF values. Each point represents a node in the phylogeny and is colored by its ultrafast bootstrap (UFBS) value. The dashed line corresponds to the identity line. Several clades or taxa with both poor gCF and sCF values are pointed by colored or gray lines. *Eurytides* is abbreviated by *E.* and *Graphium* by *G.* in species names.

Incomplete lineage sorting and gene flow in Pathysa-Paranticopsis

As described above, the subgenus *Pathysa* was found paraphyletic and *Paranticopsis* polyphyletic: the *delesserti* species group is found sister to *G. agetes* and *G. stratiotes* and they are separated from the *megarus* species group. Yet this topology is associated with a high rate of discordant gene tree topologies that could be explained by GF or ILS. Our sampling of *Pathysa* and *Paranticopsis* species is almost complete (except *Graphium euphratoides*, thought to be in the *Pathysa 2* species group) and all species possess more than 90% of the 1,402 genes. Our data thus provides a suitable framework for testing the hypotheses of GF and ILS between *Paranticopsis* and *Pathysa*.

Aphid was run on four different series of triplets. The total number of genes classified by Aphid was very similar between triplets and within replicates (>1000, **Table S3**). When the observed topology was set as the main topology, i.e. triplet n°1 ((*Pathysa 3*, *Paranticopsis 1*),

Paranticopsis 2), GF was detected between *Paranticopsis 1* and *Paranticopsis 2* (46.7% of GF, and 56.9% of imbalance toward (*Paranticopsis 1*, *Paranticopsis 2*) in these 46.7%, **Table 3**). When the tested topology was set to triplet n°2 ((*Paranticopsis 1*, *Paranticopsis 2*), *Pathysa 3*), a high proportion of GF was detected between *Paranticopsis 1* and *Pathysa 3* (which was expected as it is supported by the "main topology"). In that case, the proportion of ILS between *Paranticopsis 1* and *Pathysa 3* was not particularly higher than in the first triplet. Finally, when *Pathysa 2* was included in the triplet (triplets n°3 and n°4, **Table 3**), GF was detected between *Pathysa 2* and the two other taxa of the triplet, with a slight imbalance for GF between *Pathysa 2 and Paranticopsis* (55.8 and 58.4% of imbalance in triplet n°3 and n°4 respectively, **Table 3**). This GF was higher when the *Paranticopsis* clade was represented by *Paranticopsis 2* than by *Paranticopsis 1* (33.4% in triplet n°3 against 25.7% in triplet n°4; **Table 3**).

Table 3. Summary of Aphid results for the four series of triplets within *Paranticopsis* and *Pathysa*. Each value is the mean of five replicates for each series of triplet. Values indicated in the "Posterior imbalance GFc triplet" column correspond to the percentage of the predominant discordant topology (displayed in parenthesis) over the other one. ILS = incomplete lineage sorting, GF = gene flow. ILSc and GFc are for discordant topologies. Par = *Paranticopsis*, Pat = *Pathysa*.

Tested triplet	Number of analyzed genes	Events (%)					Posterior imbalance GFc triplet (%)
		none	ILS	GF	ILSc	GFc	
1 - A=Pat3; B=Par1; C=Par2	1034	23.3	3.4	12.6	14.1	46.7	((BC)A): 56.9
2 - A=Par1; B=Par2; C=Pat3	1038	22.5	4.5	6.4	13.7	52.9	((AC)B): 63.4
3 - A=Par2; B=Pat3; C=Pat2	1014	29.0	5.9	15.1	16.5	33.4	((AC)B): 55.8
4 - A=Par1; B=Pat3; C=Pat2	1023	31.1	5.9	20.5	16.7	25.7	((AC)B): 58.4

Discussion

de novo *assemblies of Leptocircini*

Our study provides the first robust and near complete phylogenomic framework for the tribe Leptocircini, thanks to the *de novo* assemblies of 152 individuals, including five reference genomes. The final phylogeny represents more than 93% of total Leptocircini richness. We have produced the first long-read genomes for the genus *Graphium*, and despite considerable sequencing efforts (e.g. *Graphium antiphates*, 114 Gb of long reads), the assembly of highly contiguous genomes of *Graphium* species remains a challenge. Their genomes appear to be highly heterozygous, and contain many repeats, further complicating the assembly step (see **General Discussion**). Nevertheless, these new reference genomes allowed RagTag to scaffold and recover around 20% more genes than the draft assemblies of the shotgun whole genomes, even recovering genes from the "missing" BUSCO category. The closer the reference

used was, the better the scaffolding step, and in our case, we still lack a good *Eurytides s.l.* reference genome to scaffold the Neotropical Leptocircini species. With an average of 73% complete single-copy BUSCO genes, the scaffolded genomes provided robust data for reconstructing the phylogeny and will surely be useful for further genome studies such as molecular evolution (e.g. dN/dS as in Allio et al. 2021).

Phylogeny and global taxonomic revision of Leptocircini

The following section include a proposition of taxonomic revision that will need to be reviewed and completed by expert taxonomists such as Adam M. Cotton for its form and consistency with the literature before being fully acted.

Overall, genus-level relationships are fairly consistent with previous studies. We recovered *Protographium* as sister to all *Graphium*, in agreement with the morphological results of Munroe (1960), and Zhang et al. (2019). *Lamproptera* and *Iphiclides* are found sister although both node supports and gCF indicate it is complicated node to resolve, which is consistent with Condamine et al. (2012), Allio et al. (2020a, 2021), but not consistent with Makita et al. (2003), Simonsen et al. (2011), Page and Treadaway (2014), and Kawahara et al. (2023).

Our phylogeny is the first to show a robust and well-sampled phylogenetic framework of *Eurytides s.l.* Following the recent classification of Nakae (2021), four subgenera of *Eurytides s.l.* were considered in this study: *Eurytides s.s.* Hübner, [1821]; *Protesilaus* Swainson, 1832; *Mimoides* Brown, 1991; and *Neographium* Möhn, 2002. But *Neographium* is not found monophyletic, as *Mimoides* is nested within the clade, because of *Eurytides marcellus* and *Eurytides epidaus*. This was previously discussed in Zhang et al. (2021), who proposed the monospecific subgenus *Boreographium* Grishin, 2021 (type *E. marcellus*). Yet, the issue is repeated with *E. epidaus*, which would therefore require the description of a new monospecific subgenus. To avoid the description of too many taxonomic entities (monospecific subgenera), we propose to merge all species of our clade *Noegraphium+Mimoides* into the subgenus *Mimoides*, which means that *Boreographium*, *Asiographium* and *Neographium* are synonyms of subgenus *Mimoides* (*Boreographium* **syn. nov.**, *Asiographium* **syn. nov.**, and *Neographium* **syn. nov.**). We also propose that subgenus *Eurygraphium* Möhn, 2002 (*thyastes* species group) becomes synonym of subgenus *Eurytides s.s.* (*Eurygraphium* **syn. nov.**).

The relationships within genus *Graphium* are very consistent with Miller (1987), but differ from Smith and Vane-Wright (2001), Makita et al. (2003), Page and Treadaway (2014), Allio et al. (2021), especially regarding the strong support of a monophyly of *Graphium s.s.*

Graphium phidias have had an ambiguous placement in the literature. It was long placed within *Paranticopsis*, due to its similar wing pattern (Munroe 1960, and with caution, e.g. Racheli and Cotton

2009, Page and Treadaway 2014). It was found sister to the *eurypylus* group in Makita *et al.* (2003), and sister to *Arisbe*, *Paranticopsis* and *Pathysa* in Allio et al. (2021). It was placed in *Graphium s.s.* in Nakae (2021). In our study, *Graphium phidias* is strongly supported to be sister to the *eurypylus* group, with a rather long branch.

Graphium s.s. is usually divided in three species groups: the *eurypylus*, *sarpedon* and *agamemnon* species groups (Munroe 1960, Saigusa et al. 1982, Makita et al. 2003, Racheli and Cotton 2009, Page and Treadaway 2014). However the *sarpedon* species group is not recovered monophyletic, as the *agamemnon* species group is nested within the clade, dividing it into the *sarpedon* and the *weiskei* species groups. Yet, the short branches separating these three clades (as well as their low gCF supports) suggest a rapid evolution into different species groups, and it would not be surprising if these relationships were to change.

The inclusion of *Paranticopsis* within *Pathysa* has long been suspected in the literature (e.g. Munroe and Ehrlich 1960, Hancock 1983, Miller 1987, Makita et al. 2003, Page and Treadaway 2014), but has never been clearly demonstrated or strongly supported, leaving the issue still open to doubt. Our study provides strong support for this relationship, by placing *Graphium stratiotes* and *Graphium agetes* as the sister clade of the *delesserti* species group, making *Paranticopsis* polyphyletic. An ambiguous placement of *G. agetes* and *G. stratiotes* with species of the *Paranticopsis* group had been found several times in previous studies (e.g. Makita et al. 2003, Wilson et al. 2014), but was never strongly supported. Wilson et al. (2014) found that two subspecies of *G. agetes* were either sister of *Paranticopsis* or *Pathysa* species groups, but this relied on a poorly resolved maximum parsimony tree based on 28S rRNA sequences. Interestingly, Inayoshi (2023) reports the morphology of *G. agetes* male genitalia supports its relationship with "the mimetic species of the *macareus* group." Whatever the complexity of relationships within the clade *Pathysa+Paranticopsis*, (i.e. the polyphyly of *Paranticopsis* being real or not, see below), *Pathysa* would remain in any case paraphyletic. We therefore propose to retain only *Pathysa* (Reakirt, 1865), and synonymize *Paranticopsis* Wood-Mason and de Nicéville, 1887 **syn. nov.**

Ancient gene flow could explain discordant phylogenetic relationships.

In our study, we found a strong phylogenetic discordance within the *Pathysa s.l.* clade, with low sCF and gCF values at key nodes of the different species groups. According to Minh et al. (2020), low gCF values are difficult to interpret as they may result from weak phylogenetic signals in individual loci, or from discordance between gene trees resulting from GF or ILS. Here, the small tree branches recovered suggest rapid radiation from this clade that could reinforce unresolved or erroneous genetic trees.

Yet, the distinctive distribution of traits within this clade (shift to mimicry and host plant being discordant with the

observed topology, see **Figure 8)** led us to investigate the possibility of ILS and GF in the history of this clade. Indeed, in this clade, two sets of morphological traits are observed. On one hand, *Pathysa* (as circumscribed by Nakae 2021) feed on Annonaceae and have tailed wings and a wing pattern with dark stripes on a pale background (**Figure 8**). On the other hand, *Paranticopsis* as circumscribed by Nakae (2021) mimic some Nymphalidae that display no tails and a mostly black wing pattern (*Idaeopsis* and *Parantica*, Nakae 2021) and feed on Aquifoliaceae (Hancock 1983) (**Figure 8**). The "*Pathysa* traits" are common in other distant parts of the tree. For instance, *Protographium leosthenes* and almost all *Graphium* and *Eurytides* species feed on Annonaceae, and the striped wing pattern with tail is largely distributed across the phylogeny of Papilionidae. Therefore, it is assumed that these traits are likely to be the ancestral traits, whereas the *Paranticopsis* would be considered as the "derived traits" (Hancock 1983). Our consensus phylogeny i.e. the "observed" topology (**Figure 6**, **Figure 8**, **Figure S5**) indicates that *Paranticopsis* is nested within *Pathysa*, causing *Pathysa* to be paraphyletic (as discussed and revised above, see Taxonomic section). Although this was expected based on previous works and literature (e.g. Munroe and Ehrlich 1960, Hancock 1983, Miller 1987, Smith and Vane-Wright 2001), the polyphyly of *Paranticopsis* appears surprising given the rather complex derived traits shared by both *Paranticopsis* species groups (**Figure 8**

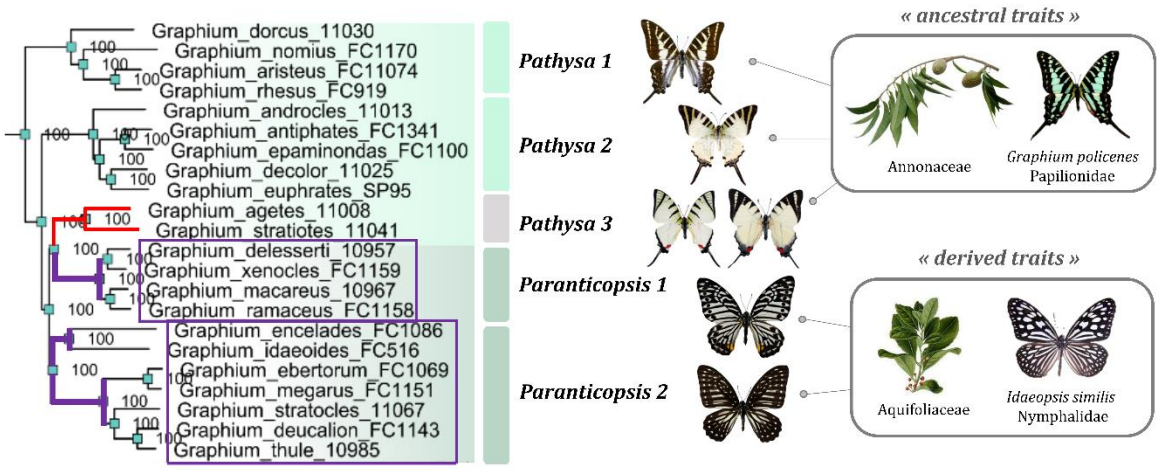


Figure 8: Phylogenetic relationships of the clade *Pathysa/Paranticopsis* (*IQTREE_1402_NT* analysis). The different species groups are delimited and named. Each species group is linked to its phenotypic traits, corresponding either to the "ancestral traits" (feeding on Annonaceae, striped wing pattern and tail), or the "derived traits" (feeding on Aquifoliaceae, and mimetic wing pattern). Branch supports (ultrafast bootstrap) are indicated at nodes

.

The questions are then: Is the distribution of traits we observe in the phylogeny the result of 1) convergence of the derived traits (or of the ancestral traits), 2) ILS from ancient polymorphism of ancestral and derived traits in the common ancestor of *Pathysa 3* and *Paranticopsis, 3*) GF between the different clades of *Paranticopsis* and/or *Pathysa*, including the gene coding for the considered traits or not. Another way of asking this problem is: Are the regions of the genome shared by *Pathysa 3* and *Paranticopsis 1* the results of a speciation event, or the result of a massive GF?

Although basic in appearance, these different scenarios were overall very difficult to validate or even to distinguish. Yet, several hypotheses could be ruled out by our analyses. First, the hypothesis that the observed topology is the "true" topology and that *Paranticopsis* derived traits are the results of a convergence gained twice in the phylogeny is unlikely because of the large number of discordant topologies in the gene trees. Second, the hypothesis that the observed distribution of traits is the result of ILS is not very likely because of the rather low ILS supports found in all triplets. Finally, the pattern of GF and their direction was the most predominant, but also the most complex to decipher.

When the "observed" topology was set as the "main" topology, i.e. the tree ((*Pathysa 3, Paranticopsis 1), Paranticopsis 2*), a high proportion of GF was detected between *Paranticopsis 1* and *Paranticopsis 2.* This is supporting the scenario of a gene flow from *Paranticopsis 2* to *Paranticopsis 1,* and that this gene flow was including the derived traits (**Figure 9C**). In that case, the last event of speciation would have been between *Pathysa 3* and *Paranticopsis 1* (as represented in the observed topology). Thereby, when the tested topology is set to (*Pathysa 3, (Paranticopsis 1, Paranticopsis 2*)), if the last speciation event was in reality between *Pathysa* 3 and *Paranticopsis 1*, we should have found a large number of trees supporting ILS between *Pathysa 3* and *Paranticopsis 1* (**Box 1**). Yet this was not clearly the case, therefore we could not totally exclude the scenario in which *Paranticopsis 1* and *Paranticopsis 2* was the last event of speciation and that massive GF happened between *Pathysa 3* and *Paranticopsis 1*, (but then, without including the gene coding for the derived traits, **Figure 9D**). These two scenarios could probably be distinguished by comparing the age (or absolute divergence) of the most recent nodes in the two types of topology (*Paranticopsis* monophyletic or not), but this would be probably very difficult because of the overall very small branch length between our three clades. Finally, we found some GF between *Pathysa 2* and *Pathysa 3,* meaning that ancestral traits could also have been transferred to *Pathysa 3* (**Figure 9E**). This could not really be excluded, but it is overall less likely because the amount of GF detected is smaller than in the two first cases.

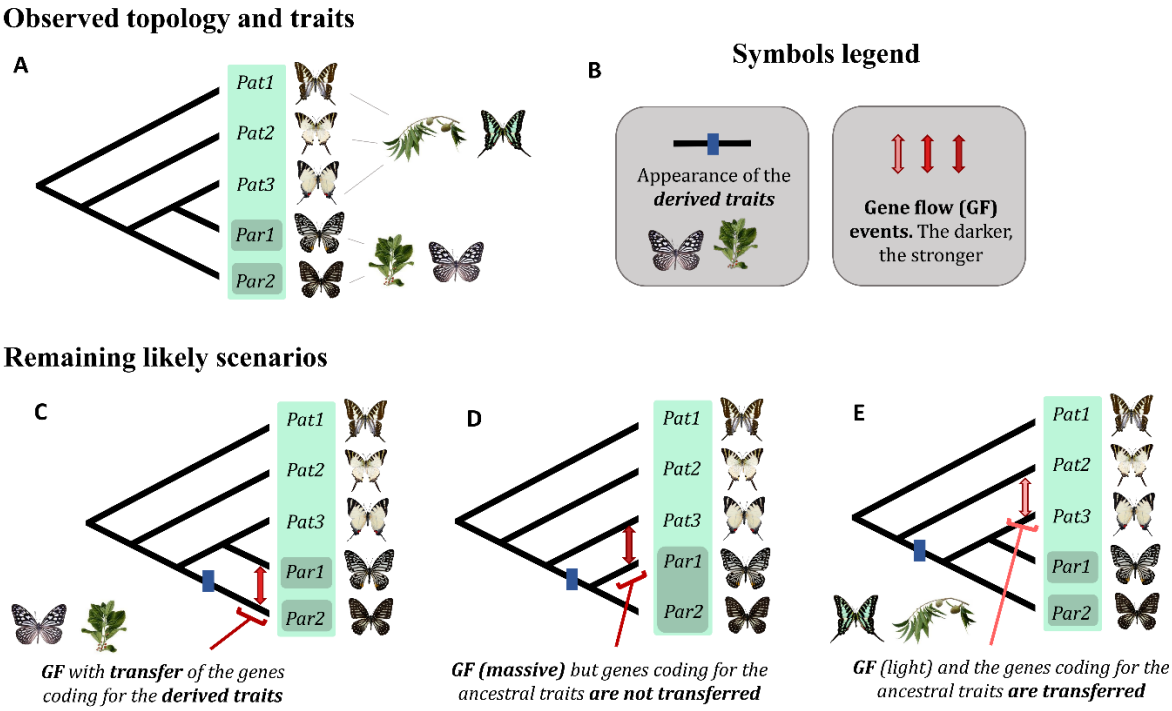
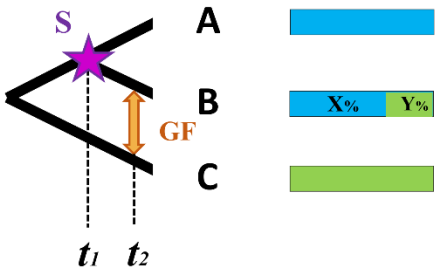


Figure 9: A. Observed topology and trait distribution between the different species groups of *Pathysa* and *Paranticopsis*. B. Significance of the symbols used in the illustration of the remaining likely scenarios. C, D and E. Scenarios that could explain the observed topology and trait distribution. *Pat* = *Pathysa*, *Par* = *Paranticopsis*. Note that in C, the appearance of derived traits could have also happened along the *Par1* branch, and be transferred to *Par2*.

Box 1. Theoretical case of GF and its implication in topological discordance analysis.

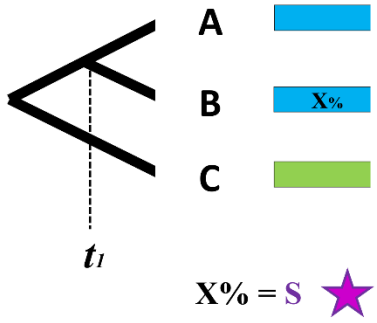
Let us consider a situation H0, implying three species A, B and C, where:

- the last event of Speciation (S) happened between A and B at t_1, and,
- an event of Gene Flow (GF) introgressed **Y**% of the genome C in B at t_2, and,
- $t_1 > t_2$:

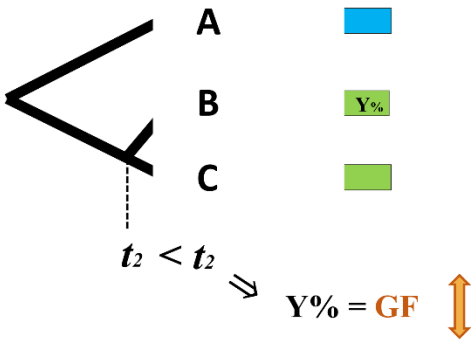


If H0 is true and the topology in Aphid is fixed as ((A, B), C):

- **X**% (blue genes) will be classified as S at t_1

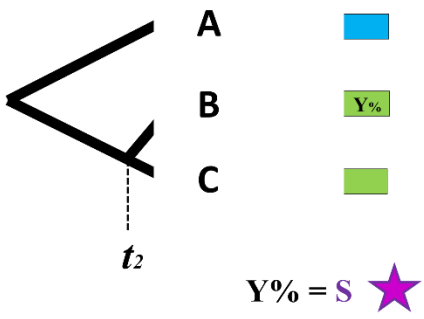


- As $t_2 < t_1$, **Y**% (green genes) will be classified as GF at t_2,

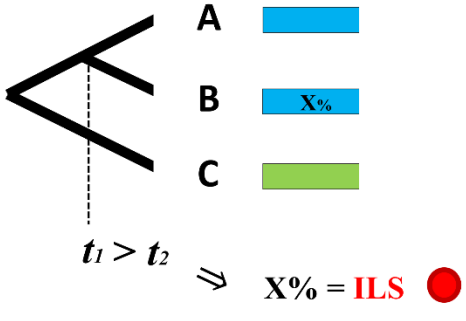


If H0 is true and the topology in Aphid is fixed as ((B, C), A):

- **Y**% (green genes) will be classified as S at t_2



- As $t_1 > t_2$, **X**% (blue genes) will be classified as ILS at t_1



In the end, the case of *Pathysa* and *Paranticopsis* is an example of the difficulty in classifying and interpreting the origin of discordant topologies in a genomic dataset. This confirms that one of the main challenges of modern phylogenomics is to manage the heterogeneity of evolutionary histories contained and revealed in the multitude of markers used (Richards et al. 2018, Scornavacca et al. 2020, Dong et al. 2022). Furthermore, our case is likely to be particularly difficult to decipher because we are testing discordances for phylogenetic scales that are deeper than simple species pairs, which probably blurs the dominant signal by the long history and possible multiplicity of evolutionary processes between the different clades (Zhang et al. 2020, Dong et al. 2022). In any case, our example includes such small branches in the recovered consensus tree that it suggests a rapid radiation of this clade and thus undoubtedly reinforces the proportion of unresolved or difficult-to-classify trees for Aphid (Galtier 2023). Here, gene ontology would likely help to understand how and when the genes encoding the derived or ancestral trait would have been transferred.

Further work in species groups is needed

Apart from *Pathysa s.l.*, concordance factor analyses (gCF and sCF) revealed poor support in several species groups, such as the *Protesilaus*, *Arisbe* and some *Graphium s.s* (**Figure 7, Figures S7, S8**). These were usually more "recent nodes" that could also reflect ongoing GF and recent ILS.

In the *adamastor* species group (*Arisbe*), most gCF values are comprised between 7% to 39% (8 nodes, gCF: mean = 36.0, median = 31.4), which means that in all these cases, the main topology is at best retrieved in only a third of the gene tree topologies, whereas the last value of the index, called gDFP (including all trees that are not either the main topology or the two alternatives of the quartet *i.e.* paraphyly of the quartet and polytomy topologies) was usually the most predominant alternative for the *adamastor* species group (8 nodes, gDFP : mean = 51.3, median = 54.3). It is difficult to draw conclusions based only on this information, but it shows us that in this species group, the 1,402 genes have generally not enough information to clearly separate species, or are not informative enough to distinguish or resolve well between the "species" of the *adamastor* species group. Yet, these same genes are informative in other parts of the tree, such as the rest of the *Arisbe* that do not have so many irresolutions (whole tree gCF: mean = 68.49, median = 77.42; gDFP: gDFP : mean = 18.618, median = 9.735, **Figure S8**). It could reflect an ongoing GF between these different "species", and could also mean that they might not all deserve a species status. The last reviews about this particular species group already warned about a potential over-splitting in this group (Smith and Vane-Wright 2001, Hancock 2006). Here we even miss some of their putative species such as *G. aurivilliusi* (only mitogenome), *G. kigoma* and *G. poggianus*, *G. ucalegonides* and *G. rileyi* (now rather considered subspecies of

fulleri), *G. olbrechtsi*, *G. abri*, *G. bouery*… (see their taxonomic status in **Table S1**)

The same pattern of low gCF values and strong prevalence of alternative topologies and polytomies were found in the subgenus *Protesilaus* (12 nodes, gCF mean = 39.8, median = 27.9; gDFP mean = 52.0, median = 58.9). Indeed, the observed topology of this clade in our different analyses was very unstable (**Figure S5**), and this clade probably needs robust population-level sampled phylogeny and taxonomic revisions (as the current species are largely sympatric cryptic species). At least three species might not be valid but were tested here: *Eurytides rothschilidianus* might be a subspecies of *E. helios*, *E. aguiari* might be a subspecies of *E. telesilaus*, and *E. exiguus* could be a subspecies of *E. glaucolaus* (**Table S1**).

Finally, our phylogenomic tree based on the nuclear genome is missing some recently described species in the subgenus *Pazala* (*G. confucius*, *G. daiyuanae*, *G. hoeneanus*, *G.wenlingae* only in COI, **Figure S4**), and our sampling was conservative in the *sarpedon* species group of *Graphium s.s.* by only selecting *G. anthedon* for the group *anthedon/isander/choredon* group from Cotton et al. (2022). In these two cases, the degree of irresolution was still high despite our conservative sampling (*Pazala*, 8 nodes: gDFP mean = 19.5, median = 18.5; *sarpedon*, 5 nodes, gDFP mean = 35.1 median = 35.5), and we are concerned that further splitting in these groups will result in the same scenario observed in the two previous cases. Huang (2023) already synonymized four of the species recently described in *Pazala*, but this back-and-forth description and synonymization increases confusion and taxonomic instability, which should be avoided (Christenhusz 2020). Similarly, there may be some structure in the *sarpedon* species, as this is clearly what is expected in a largely distributed species with several populations from different islands, but as Huang (2020) pointed out, populations can maintain strong structuring in some cases, without leading to speciation, and phylogeny as measures of population structure does not appear to be good predictors of speciation. We therefore urge taxonomists of this group to keep in mind that there may be structure in the genomics of the species, or in a few genes, but this population structure does not necessarily mean that they are different species or even that speciation is ongoing.

Overall, this tends to suggest that we should be very cautious when describing new species or changing the status of subspecies and species levels based on a few individuals and/or a few genes, and to split species on the basis of a slight genetic structure and no strong morphological differentiation. Numerous studies showed that this is not enough to describe species (nor to synonymize them), and this might inevitably lead to never-resolved phylogenies.

Acknowledgments

We would like to thank Thomas Wanchet from *La Ferme aux Papillons* for sending us fresh *Graphium* individuals from his breeding activity. We also thank Gerardo Lamas for opening the collections of Lima (Peru), and James Stewart for providing some *Arisbe* specimens. We are grateful to Daphné Navratil for her help in exploring the complexity of *Graphium* genomes, and Rémi Allio for his advice on genome assemblies. We thank Nicolas Galtier and Mathilde Barthe for discussions on the implementation and interpretation of Aphid analyses. This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (project GAIA, agreement no. 851188).

References

Adams R, DeGiorgio M. 2023. Likelihood-based tests of species tree hypotheses. *Mol. Biol. Evol.* **40:**msad159.

Ae SA. 1979. The phylogeny of some Papilio based on interspecific hybridization data. *Syst. Entomol*. **4**:1-16.

Allio R, Scornavacca C, Nabholz B, Clamens AL, Sperling FA, Condamine FL. 2020a. Whole genome shotgun phylogenomics resolves the pattern and timing of swallowtail butterfly evolution. *Syst. Biol*. **69**:38-60.

Allio R, Schomaker-Bastos A, Romiguier J, Prosdocimi F, Nabholz B, Delsuc F. 2020b. MitoFinder: efficient automated large-scale extraction of mitogenomic data in target enrichment phylogenomics. *Mol. Ecol. Res*. **20**:892–905.

Allio, R., Nabholz, B., Wanke, S., Chomicki, G., Pérez-Escobar, O.A., Cotton, A.M., Clamens, A.-L., Kergoat, G.J., Sperling, F.A.H., Condamine, F.L. 2021. Genome-wide macroevolutionary signatures of key innovations in butterflies colonizing new host plants. *Nat. Commun*. **12**:354.

Alonge M, Lebeigle L, Kirsche M, Jenike K, Ou S, Aganezov S, Wang X, Lippman ZB, Schatz MC, Soyk S. 2022. Automated assembly scaffolding using RagTag elevates a new tomato system for high-throughput genome editing. *Genome Biol*. **23**:1-19.

Aubert J, Legal L, Descimon H, Michel F. 1999. Molecular phylogeny of swallowtail butterflies of the tribe Papilionini (Papilionidae, Lepidoptera). *Mol. Phylogenet. Evol.* **12**:156–167.

Blischak PD, Chifman J, Wolfe AD, Kubatko LS. 2018. HyDe: a Python package for genome-scale hybridization detection. *Syst. Biol*. **67:**821-829.

Brown Jr KS. 1991. Neotropische Segelfalter: ihre Praeimaginalstadien nebst Beschreibung einer neuen Gattung *Mimoides* n. gen. (Papilionidae: Graphiini). *Entomol.* **101**:373-392.

Bryant D, Hahn MW. 2020. The concatenation question. In: *Phylogenetics in the Genomic Era* (Eds. Scornavacca C, Delsuc D, Galtier N). Chapter 3–4. No commercial publisher. Authors open access book.

Callaghan CJ, Casagrande MM, Lamas G, Mielke OH, Pyrcz TW, Robbins RK, Viloria AL. 2004. Atlas of Neotropical Lepidoptera. Checklist: Part 4A. Hesperoidea - Papilionoideae (Ed. Lamas G). In: *Atlas of Neotropical Lepidoptera*, (Ed. Heppner JB). **4**:439.

Caterino MS, Sperling FAH. 1999. *Papilio* phylogeny based on mitochondrial cytochrome oxidase I and II genes. *Mol. Phylogenet. Evol.* **11**:122–137.

Chen S, Zhou Y, Chen Y, Gu J. 2018. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**:i884-i890.

Chernomor O, Von Haeseler A, Minh BQ. 2016. Terrace aware data structure for phylogenomic inference from supermatrices. *Syst. Biol*. **65**:997-1008.

Christenhusz MJ. 2020. On species concepts, phylogenetics and the science of natural history—three current issues facing taxonomy. *Megataxa* **1**:67-72.

Collins NM, Morris MG. 1985. *Threatened Swallowtail Butterflies of the World: the IUCN Red Data Book*. IUCN.

Condamine FL. Allio R, Reboud EL, Dupuis JR, Toussaint EF, Mazet N, Hu S-J, Lewis DS, Kunte K, Cotton AM, *et al.* 2023. A comprehensive phylogeny and revised taxonomy illuminate the origin and diversification of the global radiation of *Papilio* (Lepidoptera: Papilionidae). *Mol. Phylogenet. Evol.* **183**:107758.

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

Condamine FL, Sperling FAH, Kergoat GJ, 2013a. Global biogeographic pattern in swallowtail diversification demonstrates alternate colonization routes in Northern and Southern Hemispheres. *J. Biogeogr*. **40**:9–23.

Condamine FL, Toussaint EFA, Cotton AM, Genson GS, Sperling FAH, Kergoat GJ. 2013b. Fine-scale biogeographic and temporal diversification processes of peacock swallowtails (*Papilio* subgenus *Achillides*) in the Indo-Australian Archipelago. *Cladistics* **29**:88–111.

Condamine FL, Sperling FAH, Wahlberg N, Rasplus J-Y, Kergoat GJ. 2012. What caused the latitudinal gradient of species diversity in swallowtail butterflies? *Ecol. Lett*. **15**:267–277.

Cotton AM, Doleck T, Zhang XIN, Inayoshi Y, Lohman DJ, Hu SJ. 2022. *Graphium septentrionicolus* Page & Treadaway, 2013 (Lepidoptera: Papilionidae) is a distinct species. *Zootaxa* **5154**:211-224.

d'Abrera B. 1982. *Butterflies of the Oriental Region, Part I Papilionidae, Pieridae & Danaidae*. Hill House, Victoria.

Degnan JH, Rosenberg NA. 2006. Discordance of species trees with their most likely gene trees. *PLoS Genet.* **2**:e68.

Delsuc F, Brinkmann H, Philippe H. 2005. Phylogenomics and the reconstruction of the tree of life. *Nat. Rev. Genet.* **6**:361–375.

DeRaad DA, McCullough JM, DeCicco LH, Hime PM, Joseph L, Andersen MJ, Moyle RG. 2023. Mitonuclear discordance results from incomplete lineage sorting, with no detectable evidence for gene flow, in a rapid radiation of *Todiramphus* kingfishers. *Mol. Ecol.* **32**:4844-4862.

Dong W, Li E, Liu Y, Xu C, Wang Y, Liu K, Cui K, Sun J, Zhang Z, *et al.* 2022. Phylogenomic approaches untangle early divergences and complex diversifications of the olive plant family. *BMC Biol.* **20**:92.

Edelman NB, Frandsen PB, Miyagi M, Clavijo B, Davey J, Dikow RB, Garcia-Accinelli G., Van Belleghem SM, Patterson N, Neafsey DE, Challis R, *et al.* 2019. Genomic architecture and introgression shape a butterfly radiation. *Science* **366**:594-599.

Edwards RJ. 2019. BUSCOMP: BUSCO compilation and comparison–Assessing completeness in multiple genome assemblies. *F1000Research* **8**:995.

Ehrlich PR. 1957. The Morphology, Phylogeny and Higher Classification of the Butterflies (Lepidoptera: Papilionoidea). Doctoral dissertation, University of Kansas.

Feng S, Bai M, Rivas-González I, Li C, Liu S, Tong Y, Yang T, Chen G, Xie D, Sears KE, *et al*. 2022. Incomplete lineage sorting and phenotypic evolution in marsupials. *Cell* **185**:1646-1660.

Formenti G, Theissinger K, Fernandes C, Bista I, Bombarely A, Bleidorn C, Ciofi C, Crottini A, Godoy JA, Höglund J, *et al*. 2022. The era of reference genomes in conservation genomics. *Trends Ecol. Evol.* **37**:197–202.

Funk DJ, Omland KE. 2003. Species-level paraphyly and polyphyly: frequency, causes, and consequences, with insights from animal mitochondrial DNA. *Annu. Rev. Ecol. Evol. Syst.* **34**:397-423.

Galtier N. 2023. Phylogenetic conflicts: distinguishing gene flow from incomplete lineage sorting. *bioRxiv* https://doi.org/10.1101/2023.07.06.547897

Guan D, McCarthy SA, Wood J, Howe K, Wang Y, Durbin R. 2020. Identifying and removing haplotypic duplication in primary genome assemblies. *Bioinformatics* **36**:2896-2898.

Hancock DL. 1993. Origins and evolution of the Afrotropical Papilionidae (Lepidoptera). *Arnoldia Zimbabwe* **9**:557-583.

Hancock DL. 2006. Notes on the afrotropical species of *Graphium* Scopoli (Lepidoptera: Papilionidae). *Metamorphosis* **17**:6-19.

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

Häuser CL, de Jong R, Lamas G, Robbins RK, Smith C, Vane-Wright RI. 2005. Papilionidae revised GloBIS⁄GART species checklist (2nd draft). Available at: http://www.globis.insects-online.de/species&tree_h=.Papilionidae&sub=yes&tree_status=plus&tree_seq=0.

He B, Zhao Y, Su C, Lin G, Wang Y, Li L, Ma J, Yang Q, Hao J. 2023. Phylogenomics reveal extensive phylogenetic discordance due to incomplete lineage sorting following the rapid radiation of alpine butterflies (Papilionidae: *Parnassius*). *Syst. Entomol.* **48**:585–599.

He JW, Zhang R, Yang J, Chang Z, Zhu LX, Lu SH, Xie F-A, Mao J-L, Dong Z-W, Li G-C. 2022. High-quality reference genomes of swallowtail butterflies provide insights into their coloration evolution. *Zool. Res*. **43**:367-379.

Hoang DT, Chernomor O, Von Haeseler A, Minh BQ, Vinh LS. 2018. UFBoot2: improving the ultrafast bootstrap approximation. *Mol. Biol. Evol.* **35**:518-522.

Hu SJ, Condamine FL, Monastyrskii A, Cotton A. 2019. A new species of the *Graphium* (*Pazala*) *mandarinus* group from Central Vietnam (Lepidoptera: Papilionidae). *Zootaxa* **4554**:286-300.

Huang H. 2023. New or little known butterflies from China – 5 (Lepidoptera: Papilionidae, Pieridae, Nymphalidae, Lycaenidae et Hesperiidae). *Neue Entomologische Nachrichten* **82**:93-214.

Huang JP. (2020). Is population subdivision different from speciation? From phylogeography to species delimitation. *Ecol. Evol.* **10**:6890-6896.

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

Inayoshi Y. 2021. A check list of butterflies in Indo-China: Chiefly from Thailand, Laos & Vietnam. Available from: http://yutaka.it-n.jp (accessed August 1, 2023)

Jin JJ, Yu WB, Yang JB, Song Y, DePamphilis CW, Yi TS, Li DZ. 2020. GetOrganelle: a fast and versatile toolkit for accurate de novo assembly of organelle genomes. *Genome Biol.* **21**:1-31.

Kalyaanamoorthy S, Minh BQ, Wong TK, Von Haeseler A, Jermiin LS. 2017. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat. Meth.* **14**:587-589.

Katoh K, Standley DM. 2013. MAFFT Multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**:772–780.

Kawahara AY, Storer C, Carvalho APS, Plotkin DM, Condamine FL, Braga MP, *et al.* 2023. A global phylogeny of butterflies reveals their evolutionary history, ancestral hosts and biogeographic origins. *Nat. Ecol. Evol.* **7**:903-913.

Kolmogorov M, Yuan J, Lin Y, Pevzner PA. 2019. Assembly of long, error-prone reads using repeat graphs. *Nat. Biotech.* **37**:540-546.

Laetsch DR, Blaxter ML. 2017. BlobTools: Interrogation of genome assemblies. *F1000Research* **6**:1287.

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

Lewis DS, Sperling FAH, Nakahara S, Cotton AM, Kawahara AY, Condamine FL. 2015. Role of Caribbean Islands in the diversification and biogeography of Neotropical *Heraclides* swallowtails. *Cladistics* **31**:291–314.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. 2009. The sequence alignment/map format and SAMtools. *Bioinformatics* **25**:2078–2079.

Li H. 2018. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**:3094–3100.

Li X, Fan D, Zhang W, Liu G, Zhang L, Zhao L, Fang X, Chen L, Dong Y, Chen Y, *et al.* 2015. Outbred genome sequencing and CRISPR/Cas9 gene editing in butterflies. *Nat. Commun.* **6**:8212.

Li D, Liu CM, Luo R, Sadakane K, Lam TW. 2015. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics* **31**:1674-1676.

Lu S, Yang J, Dai X, Xie F, He J, Dong Z, *et al.* 2019. Chromosomal-level reference genome of Chinese peacock butterfly (*Papilio bianor*) based on third-generation DNA sequencing and Hi-C analysis. *GigaScience* **8**:giz128.

Mackintosh A, Laetsch DR, Baril T, Ebdon S, Jay P, Vila R, *et al.* 2022. The genome sequence of the scarce swallowtail, *Iphiclides podalirius*. *G3:Genes Genomes Genet.* **12**:jkac193.

Makita H, Shinkawa T, Kondo K, Xing L, Nakazawa T. 2003. Phylogeny of the *Graphium* butterflies inferred from nuclear 28S rDNA and mitochondrial ND5 gene sequences. *Lepidoptera Sci*. **54**:91-110.

Mallet J, Besansky N, Hahn MW. 2016. How reticulated are species? *BioEssays* **38**:140–149.

Manni M, Berkeley MR, Seppey M, Simão FA, Zdobnov EM. 2021. BUSCO update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. *Mol. Biol. Evol.* **38**:4647-4654.

Meyer M, Kircher M. 2010. Illumina sequencing library preparation for highly multiplexed target capture and sequencing. *Cold Spring Harb Protoc.* **6**:t5448.

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. France* **44**:1-36.

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

Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, Von Haeseler A, Lanfear R. 2020a. IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. *Mol. Biol. Evol.* **37**:1530-1534.

Minh BQ, Hahn MW, Lanfear R. 2020b. New methods to calculate concordance factors for phylogenomic datasets. *Mol. Biol. Evol.* **37**:2727-2733.

Möhn E. 2002. Papilionidae VIII. *Baronia*, *Euryades*, *Protographium*, *Neographium* und *Eurytides*. Goecke & Evers, Keltern. 15 pp. vol. 14, p. ls 1-36. Schmetterlinge der Erde.Tagfalter. Teil 5.

Munroe E. 1960. The classification of the Papilionidae (Lepidoptera). *Memoirs Entomol. Soc. Canada* **92**:5-51.

Munroe E, Ehrlich PR. 1960. Harmonization of concepts of higher classification of the Papilionidae. *J. Lepid. Soc.* **14**:169-175.

Mutanen M, Kivelä SM, Vos RA, Doorenweerd C, Ratnasingham S, Hausmann A, *et al.* 2016. Species-level para-and polyphyly in DNA barcode gene trees: strong operational bias in European Lepidoptera. *Syst. Biol.* **65**:1024-1040.

Nakae M. 2021. Papilionidae of the World. Roppon-Ashi Entomological Books.

Nazari V, Zakharov EV, Sperling FAH. 2007. Phylogeny, historical biogeography, and taxonomic ranking of Parnassiinae (Lepidoptera, Papilionidae) based on morphology and seven genes. *Mol. Phylogenet. Evol.* **42**:131-156.

Niculescu EV. 1977. Un nouveau genre de Graphiini: *Klinzigia* n. gen. *Bull. Soc. Entomol. Mulhouse* **33**:51–52.

Nishimura O, Hara Y, Kuraku S. 2017. gVolante for standardizing completeness assessment of genome and transcriptome assemblies. *Bioinformatics* **33**:3635-3637.

Owens H, Lewis DS, Dupuis JR, Clamens AL, Sperling FAH, Kawahara AY, Guralnick RP, Condamine FL. 2017. The latitudinal diversity gradient in New World swallowtail butterflies is caused by contrasting patterns of out-of- and into-the-tropics dispersal. *Glob. Ecol. Biogeogr.* **26**:1447–1458.

Page MG, Treadaway CG. 2014. Revisional notes on the *Arisbe eurypylus* species group (Lepidoptera: Papilionoidea: Papilionidae). *Stuttgarter Beiträge zur Naturkunde: A, Neue Serie* **7**:253-284.

Pennisi E. 2008. Building the tree of life, genome by genome. *Science* **5884**:1716-1717.

Philippe H, Blanchette M. 2007. Overview of the first phylogenomics conference. *BMC Evol. Biol.* **7**:S1.

Racheli T, Cotton AM. 2009. Papilionidae part 1. – In: Bozano, G. C. (ed.), Guide to the Butterflies of the Palearctic Region. — Omnes Artes, Milano, 70 pp.

Rannala B, Edwards S, Leaché A. 2020. The multi-species coalescent model and species tree inference. In: *Phylogenetics in the Genomic Era* (Eds. Scornavacca C, Delsuc D, Galtier N). Chapter 3–3. No commercial publisher. Authors open access book.

Ranwez V, Douzery EJP, Cambon C, Chantret N, Delsuc F. 2018. MACSE v2: toolkit for the alignment of coding sequences accounting for frameshifts and stop codons. *Mol. Biol. Evol.* **35**:2582-2584.

Reboud EL, Nabholz B, Chevalier E, Tilak MK, Bito D, Condamine FL. 2023. Genomics, population divergence, and historical demography of the world's Largest and endangered butterfly, the Queen Alexandra's birdwing. *Genome Biol. Evol.* **15**:evad040.

Reed RD, Sperling FAH. 1999. Interaction of process partitions in phylogenetic analysis: an example from the swallowtail butterfly genus *Papilio*. *Mol. Biol. Evol.* **16**:286–297.

Richards EJ, Brown JM, Barley AJ, Chong RA, Thomson RC. 2018. Variation across mitochondrial gene trees provides evidence for systematic error: how much gene tree variation is biological? *Syst. Biol.* **67**:847-860.

Rivas-González I, Rousselle M, Li F, Zhou L, Dutheil JY, Munch K, *et al.* 2023. Pervasive incomplete lineage sorting illuminates speciation and selection in primates. *Science* **380**:eabn4409.

Ross HA. 2014. The incidence of species-level paraphyly in animals: a re-assessment. *Mol. Phylogenet. Evol.* **76**:10-17.

Saigusa T, Nakanishi A, Shima H, Yata O. 1982. Phylogeny and geographical distribution of the swallowtail subgenus *Graphium* (Lepidoptera: Papilionidae). *Entomologia Generalis* 59-69.

Scornavacca C, Delsuc F, Galtier N. 2020. *Phylogenetics in the Genomic Era*. No commercial publisher. Authors open access book. 568 pp.

Silva-Brandao KL, Freitas AVL, Brower AV, Solferini VN. 2005. Phylogenetic relationships of the New World Troidini swallowtails (Lepidoptera: Papilionidae) based on COI, COII, and EF-1α genes. *Mol. Phylogenet. Evol.* **36**:468-483.

Simion P, Delsuc F, Philippe H. 2020. To what extent current limits of phylogenomics can be overcome? In: *Phylogenetics in the Genomic Era* (Eds. Scornavacca C, Delsuc D, Galtier N). Chapter 2–1. No commercial publisher. Authors open access book.

Simonsen TJ, Zakharov EV, Djernaes M, Cotton AM, Vane-Wright RI, Sperling FAH. 2011. Phylogenetics and divergence times of Papilioninae (Lepidoptera) with special reference to the enigmatic genera *Teinopalpus* and *Meandrusa*. *Cladistics* **27**:113-137.

Smith CR, Vane-Wrigth RI. 2001. A review of the afrotropical species of the genus Graphium (Lepidoptera: Rhopalocera: Papilionidae). *Bull. Nat. Hist. Mus. Lond. (Ent.)* **70**:503-719.

Suvorov A, Kim BY, Wang J, Armstrong EE, Peede D, D'Agostino ER, *et al.* 2022. Widespread introgression across a phylogeny of 155 *Drosophila* genomes. *Curr. Biol.* **32**:111-123.

Tian X, Mo S, Liang D, Wang H, Zhang P. 2023. Amplicon capture phylogenomics provides new insights into the phylogeny and evolution of alpine *Parnassius* butterflies (Lepidoptera: Papilionidae). *Syst. Entomol.* **48**:571–584.

Tilak MK, Justy F, Debiais-Thibaud M, Botero-Castro F, Delsuc F, Douzery EJP. 2015. A cost-effective straightforward protocol for shotgun Illumina libraries designed to assemble complete mitogenomes from non-model species. *Cons. Genet. Resour.* **7**:37-40.

Wiens JJ, Penkrot TA. 2002. Delimiting species using DNA and morphological variation and discordant species limits in spiny lizards (*Sceloporus*). *Syst. Biol.* **51**:69-91.

Wiens JJ, Kuczynski CA, Stephens PR. 2010. Discordant mitochondrial and nuclear gene phylogenies in emydid turtles: implications for speciation and conservation. *Biol. J. Linn. Soc.* **99**:445-461.

Wilson JJ, Karen-Chia HM, Sing KW, Sofian-Azirun M. 2014. Towards resolving the identities of the *Graphium* butterflies (Lepidoptera: Papilionidae) of Peninsular Malaysia. *J. Asia-Pacific Entomol.* **17**:333-338.

Wu LW, Yen SH, Lees DC, Lu CC, Yang PS, Hsu YF. 2015. Phylogeny and historical biogeography of Asian *Pterourus* butterflies (Lepidoptera: Papilionidae): A case of intercontinental dispersal from North America to East Asia. *PLoS One* **10**:e0140933.

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.

Zakharov EV, Caterino MS, Sperling FAH. 2004. Molecular phylogeny, historical biogeography, and divergence time estimates for swallowtail butterflies of the genus *Papilio* (Lepidoptera: Papilionidae). *Syst. Biol.* **53**:193-215.

Zhang C, Rabiee M, Sayyari E, Mirarab S. 2018. ASTRAL-III: polynomial time species tree reconstruction from partially resolved gene trees. *BMC Bioinformatics* **19**:15-30.

Zhang C, Zhang T, Luebert F, Xiang Y, Huang CH, Hu Y, *et al.* 2020. Asterid phylogenomics/phylotranscriptomics uncover morphological evolutionary histories and support phylogenetic placement for numerous whole-genome duplications. *Mol. Biol. Evol.* **37**:3188-3210.

Zhang HH, Cotton AM, Condamine FL, Wang RJ, Hsu YF, Duan K, *et al.* 2020. Revision of *Pazala* Moore, 1888: The *Graphium* (*Pazala*) *alebion* and *G.* (*P.*) *tamerlanus* groups, with notes on taxonomic and distribution confusions (Lepidoptera: Papilionidae). *Zootaxa* **4759**:77-97.

Zhang J, Cong Q, Shen J, Opler PA, Grishin NV. 2019. Changes to North American butterfly names. *Taxonomic Rep. Intl. Lepido. Survey* **8**:2.

Zhang J, Cong Q, Shen J, Opler PA, Grishin NV. 2021. Genomics-guided refinement of butterfly taxonomy. *Taxonomic Rep. Intl. Lepido. Survey* **9**:3.

Zimin AV, Salzberg S.L. 2020. The genome polishing tool POLCA makes fast and accurate corrections in genome assemblies. *PLoS Comput. Biol.* **16**:e1007981.

Zimin AV, Marçais G, Puiu D, Roberts M, Salzberg SL, Yorke JA. 2013. The MaSuRCA genome assembler. *Bioinformatics* **29**:2669-2677.

Supplementary information

Table S1: Taxonomic working list for Leptocircini species. The list includes subgenera that are currently recovered as monophyletic in molecular phylogenies, and species that belong to each subgenus. In the comment section, species in blue were sampled in our study and species in red were unsampled.

Genus *Eurytides* Hübner, [1821]	Sampling	Taxonomic status	Comment
Subgenus *Eurytides* Hübner, [1821]			
Eurytides (Eurytides) bellerophon Dalman, 1823	Sampled	Valid	
Eurytides (Eurytides) callias Rothschild & Jordan, 1906	Sampled	Valid	
Eurytides (Eurytides) columbus Kollar, 1850	Sampled	Valid	
Eurytides (Eurytides) dolicaon Cramer, 1775	Sampled	Valid	
Eurytides (Eurytides) iphitas Hübner, 1821	Unsampled	Valid	Possibly extinct (Grice et al. 2019, Domagala and Gonzales 2021). Last specimen seen in 1977
Eurytides (Eurytides) orabilis Butler, 1872	Sampled	Valid	
Eurytides (Eurytides) salvini Bates, 1864	Sampled	Valid	
Eurytides (Eurytides) serville Godart, 1824	Sampled	Valid	
Subgenus *Mimoides* Brown, 1991			
Eurytides (Mimoides) agesilaus Guerin & Percheron, 1835	Sampled	Valid	The species *E. oberthueri* Rothschild & Jordan, 1906 is supposed natural hybrid of *E.agesilaus* x *E.philolaus*
Eurytides (Mimoides) anaxilaus Felder & Felder, 1865	Unsampled	Valid	
Eurytides (Mimoides) ariarathes Esper, 1788	Sampled	Valid	
Eurytides (Mimoides) asius Fabricius, 1781	Sampled	Valid	
Eurytides (Mimoides) calliste Bates, 1864	Sampled	Valid	
Eurytides (Mimoides) celadon Lucas, 1852	Sampled	Valid	
Eurytides (Mimoides) dioxippus Hewitson, 1856	Sampled	Valid	
Eurytides (Mimoides) epidaus Doubleday, 1846	Sampled	Valid	
Eurytides (Mimoides) euryleon Hewitson, 1856	Sampled	Valid	
Eurytides (Mimoides) ilus Fabricius, 1793	Sampled	Valid	
Eurytides (Mimoides) leucaspis Godart, 1819	Sampled	Valid	
Eurytides (Mimoides) lysithous Hübner, 1821	Sampled	Valid	
Eurytides (Mimoides) marcellinus Doubleday, 1845	Sampled	Valid	
Eurytides (Mimoides) marcellus Cramer, 1777	Sampled	Valid	
Eurytides (Mimoides) microdamas Burmeister, 1878	Sampled	Valid	
Eurytides (Mimoides) pausanias Hewitson, 1852	Sampled	Valid	

Eurytides (Mimoides) phaon Boisduval, 1836	Sampled	Valid	
Eurytides (Mimoides) philolaus Boisduval, 1836 ***stat. rev.***	Sampled	Valid	*E. p. xanticles* could be a species
Eurytides (Mimoides) protodamas Godart, 1819	Sampled	Valid	
Eurytides (Mimoides) thyastes Drury, 1782	Sampled	Valid	
Eurytides (Mimoides) thymbraeus Boisduval, 1836	Sampled	Valid	
Eurytides (Mimoides) xeniades Hewitson, 1867	Sampled	Valid	
Eurytides (Mimoides) xynias Hewitson, 1875	Sampled	Valid	
Eurytides (Mimoides) zonaria Butler, 1869	Sampled	Valid	
Subgenus *Protesilaus* Swainson, [1832]			
Eurytides (Protesilaus) aguiari D'Almeida, 1937	Sampled	Doubtful	This species could be a subspecies of *E. telesilaus*
Eurytides (Protesilaus) earis Rothschild & Jordan, 1906	Sampled	Valid	
Eurytides (Protesilaus) exiguus Winhard, 2018	Sampled	***stat. nov.***	We propose a *stat. nov.*, as a subspecies of *E. glaucolaus*
Eurytides (Protesilaus) glaucolaus Bates, 1864	Sampled	Valid	
Eurytides (Protesilaus) helios Rothschild & Jordan, 1906	Sampled	Valid	*E. rothschildianus in litt.* (Mielke, pers. comm) could be a subspecies of *E. helios* (based on results of this study).
Eurytides (Protesilaus) leucosilaus Zikán, 1937	Sampled	Valid	Sometimes conidered as subspecies of *E. molops*, but monophyly not recovered in this study. Might be a subspecies of *E. glaucolaus*
Eurytides (Protesilaus) macrosilaus Gray, 1853	Sampled	Valid	
Eurytides (Protesilaus) molops Rothschild & Jordan, 1906	Sampled	Valid	
Eurytides (Protesilaus) orthosilaus Weymer, 1899	Sampled	Valid	
Eurytides (Protesilaus) protesilaus Linnaeus, 1758	Sampled	Valid	
Eurytides (Protesilaus) stenodesmus Rothschild & Jordan, 1906	Sampled	Valid	
Eurytides (Protesilaus) telesilaus Felder & Felder, 1864	Sampled	Valid	

Genus *Graphium* Scopoli, 1777	**Sampling**	**Taxonomic status**	**Comment**
Subgenus *Arisbe* Hübner, [1819]			
Graphium (Arisbe) abri Smith & Vane-Wright, 2001	Unsampled	Probably unvalid	Described as species with two specimens. Could be a subspecies or melanic form of *G. adamastor* or *G. agamedes.*
Graphium (Arisbe) adamastor Boisduval, 1836	Sampled	Valid	
Graphium (Arisbe) agamedes Westwood, 1842	Sampled	Valid	
Graphium (Arisbe) almansor Honrath, 1884	Sampled	Valid	
Graphium (Arisbe) angolanus Goeze, 1779	Sampled	Valid	
Graphium (Arisbe) antheus Cramer, 1779	Sampled	Valid	

Graphium (Arisbe) auriger Butler, 1876	Sampled	Valid	
Graphium (Arisbe) aurivilliusi Seeldrayers, 1896	Mito only	***syn. nov.***	Only the Mitogenome could be assembled and it is identical to *G. adamastor*. Could be a natural hybrid of *G. adamastor* x *G. schubotzi* (Hancock 2006)
Graphium (Arisbe) biokoensis Gauthier, 1984	Sampled	Valid	*G. b. bouyeri* could be a separate species. Not enough data
Graphium (Arisbe) colonna Ward, 1873	Sampled	Valid	
Graphium (Arisbe) cyrnus Boisduval, 1836	Sampled	Valid	
Graphium (Arisbe) endochus Boisduval, 1836	Sampled	Valid	
Graphium (Arisbe) evombar Boisduval, 1836	Sampled	Valid	
Graphium (Arisbe) fulleri Grose-Smith, 1883	Sampled	Valid	
Graphium (Arisbe) gudenusi Rebel, 1911	Sampled	Valid	
Graphium (Arisbe) hachei Dewitz, 1881	Sampled	Valid	
Graphium (Arisbe) illyris Hewitson, 1873	Sampled	Valid	
Graphium (Arisbe) junodi Trimen, 1893	Sampled	Valid	
Graphium (Arisbe) kigoma Carcasson, 1964	Unsampled	Doubtful	A subspecies of *almansor* or *poggianus* (Hancock 2006)
Graphium (Arisbe) kirbyi Hewitson, 1872	Sampled	Valid	
Graphium (Arisbe) latreillianus Godart, 1819	Sampled	Valid	
Graphium (Arisbe) leonidas Fabricius, 1793	Sampled	Valid	
Graphium (Arisbe) levassori Oberthür, 1886	Sampled	Valid	
Graphium (Arisbe) liponesco Suffert, 1904	Sampled	Valid	
Graphium (Arisbe) morania Angas, 1849	Sampled	Valid	
Graphium (Arisbe) olbrechtsi Berger, 1950	Unsampled	Probably unvalid	Might be a subspecies of *G. schubotzi*. Not enough data
Graphium (Arisbe) philonoe Ward, 1873	Sampled	Valid	
Graphium (Arisbe) poggianus Honrath, 1884	Unsampled	Valid	
Graphium (Arisbe) policenes Cramer, 1775	Sampled	Valid	
Graphium (Arisbe) policenoides Holland, 1892	partial Mito only	Valid	
Graphium (Arisbe) polistratus Grose-Smith, 1889	Sampled	Valid	
Graphium (Arisbe) porthaon Hewitson, 1865	Sampled	Valid	
Graphium (Arisbe) ridleyanus White, 1843	Sampled	Valid	
Graphium (Arisbe) schaffgotschi Niepelt, 1927	Sampled	Valid	
Graphium (Arisbe) schubotzi Schultze, 1913	Sampled	Valid	
Graphium (Arisbe) simoni Aurivillius, 1899	Sampled	Valid	

Graphium (Arisbe) taboranus Oberthür, 1886	Sampled	Valid	
Graphium (Arisbe) tynderaeus Fabricius, 1793	Sampled	Valid	
Graphium (Arisbe) ucalegon Hewitson, 1865	Sampled	Valid	
Subgenus *Graphium* Scopoli, 1777			
Graphium (Graphium) agamemnon Linnaeus, 1758	Sampled	Valid	
Graphium (Graphium) anthedon Felder & Felder, 1864	Sampled	Valid	*G. a. isander* and *G. a choredon* were considered subspecies in this study, because of the subspecies status and paraphyly of *G. a. crudus* in Cotton et al. 2022
Graphium (Graphium) arycles Boisduval, 1836	Sampled	Valid	
Graphium (Graphium) bathycles Zincken, 1831	Sampled	Valid	
Graphium (Graphium) browni Godman & Salvin, 1879	Sampled	Valid	
Graphium (Graphium) chironides Honrath, 1884	Sampled	Valid	
Graphium (Graphium) cloanthus Westwood, 1841	Sampled	Valid	
Graphium (Graphium) codrus Cramer, 1777	Sampled	Valid	
Graphium (Graphium) doson Felder & Felder, 1864	Sampled	Valid	
Graphium (Graphium) empedovana Corbet, 1941	Sampled	Valid	
Graphium (Graphium) eurypylus Linnaeus, 1758	Sampled	Valid	
Graphium (Graphium) evemon Boisduval, 1836	Sampled	Valid	
Graphium (Graphium) gelon Boisduval, 1859	Sampled	Valid	
Graphium (Graphium) hicetaon Mathew, 1886	Sampled	Valid	
Graphium (Graphium) kosii Müller & Tennent, 1999	Sampled	Valid	
Graphium (Graphium) leechi Rothschild, 1895	Sampled	Valid	
Graphium (Graphium) macfarlanei Butler, 1877	Sampled	Valid	
Graphium (Graphium) macleayanus Leach, 1814	Sampled	Valid	
Graphium (Graphium) meeki Rothschild & Jordan, 1901	Sampled	Valid	
Graphium (Graphium) mendana Godman & Salvin, 1888	Sampled	Valid	
Graphium (Graphium) meyeri Hopffer, 1874	Sampled	Valid	
Graphium (Graphium) milon Felder & Felder, 1864	Sampled	Valid	
Graphium (Graphium) monticolus Fruhstorfer, 1896	Sampled	Valid	
Graphium (Graphium) phidias Oberthür, 1896	Sampled	Valid	
Graphium (Graphium) procles Grose-Smith, 1887	Sampled	Valid	
Graphium (Graphium) sandawanum Yamamoto, 1977	Sampled	Valid	
Graphium (Graphium) sarpedon Linnaeus, 1758	Sampled	Valid	
Graphium (Graphium) septentrionicolus Page & Treadaway, 2013	Sampled	Valid	

Graphium (Graphium) stresemanni Rothschild, 1916	Sampled	Valid	*G. s. batjanensis* could be a separate species
Graphium (Graphium) sumatrana Hagen, 1894	Sampled	Valid	
Graphium (Graphium) teredon Felder & Felder, 1865	Sampled	Valid	
Graphium (Graphium) wallacei Hewitson, 1858	Sampled	Valid	
Graphium (Graphium) weiskei Ribbe, 1900	Sampled	Valid	
Subgenus *Pathysa* Reakirt, [1835]			
Graphium (Pathysa) agetes Westwood, 1841	Sampled	Valid	
Graphium (Pathysa) androcles Boisduval, 1836	Sampled	Valid	
Graphium (Pathysa) antiphates Cramer, 1775	Sampled	Valid	
Graphium (Pathysa) aristeus Stoll, 1780	Sampled	Valid	
Graphium (Pathysa) decolor Staudinger, 1888	Sampled	Valid	
Graphium (Pathysa) delessertii Guérin-Méneville, 1839 ***stat. rev.***	Sampled	Valid	
Graphium (Pathysa) deucalion Boisduval, 1836 ***stat. rev.***	Sampled	Valid	
Graphium (Pathysa) dorcus de Haan, 1840	Sampled	Valid	
Graphium (Pathysa) ebertorum Koçak, 1983 ***stat. rev.***	Sampled	Valid	
Graphium (Pathysa) encelades Boisduval, 1836 ***stat. rev.***	Sampled	Valid	
Graphium (Pathysa) epaminondas Oberthür, 1879	Sampled	Valid	
Graphium (Pathysa) euphrates Felder & Felder, 1862	Sampled	Valid	
Graphium (Pathysa) euphratoides Eimer, 1889	Unsampled	Valid	
Graphium (Pathysa) felixi Joicey & Noakes, 1915 ***stat. rev.***	Sampled	Valid	
Graphium (Pathysa) idaeoides Hewitson, 1853 ***stat. rev.***	Sampled	Valid	
Graphium (Pathysa) macareus Godart, 1819 ***stat. rev.***	Sampled	Valid	
Graphium (Pathysa) megarus Westwood, 1841 ***stat. rev.***	Sampled	Valid	
Graphium (Pathysa) nomius Esper, 1798	Sampled	Valid	
Graphium (Pathysa) ramaceus Westwood, 1872 ***stat. rev.***	Sampled	Valid	
Graphium (Pathysa) rhesus Boisduval, 1836	Sampled	Valid	
Graphium (Pathysa) stratiotes Grose-Smith, 1887	Sampled	Valid	
Graphium (Pathysa) stratocles Felder & Felder, 1861 ***stat. rev.***	Sampled	Valid	
Graphium (Pathysa) thule Wallace, 1865 ***stat. rev.***	Sampled	Valid	
Graphium (Pathysa) xenocles Doubleday, 1842 ***stat. rev.***	Sampled	Valid	
Subgenus *Pazala* Moore, 1888			
Graphium (Pazala) alebion Gray, 1853	Sampled	Valid	
Graphium (Pazala) confucius Hu, Duan & Cotton, 2018	Mito NCBI only	Probably unvalid	Subspecies of *G. mandarinus* in Huang 2023

Graphium (Pazala) daiyuanae Hu, Zhang & Cotton, 2018	COI-5P NCBI only	Probably unvalid	Subspecies of *G. mandarinus* in Huang 2023
Graphium (Pazala) eurous Leech, 1893	Sampled	Valid	*G. e. asakurae*, *G. e. caschmirensis*, *G. e. meridionalis* are putative species. Only Mito worked for *G. asakurae*
Graphium (Pazala) garhwalica Katayama, 1988	Sampled	Valid	
Graphium (Pazala) hoeneanus Cotton & Hu, 2018	COI-5P NCBI only	Probably unvalid	Subspecies of *G. mandarinus* in Huang 2023
Graphium (Pazala) mandarinus Oberthür, 1879	Sampled	Valid	
Graphium (Pazala) mullah Alphéraky, 1897	Sampled	Valid	
Graphium (Pazala) paphus Nicéville, 1886	Mito only	Valid	
Graphium (Pazala) parus Nicéville, 1900	Sampled	Valid	
Graphium (Pazala) sichuanica Koiwaya, 1993	Sampled	Valid	
Graphium (Pazala) tamerlanus Oberthür, 1876	Sampled	Valid	
Graphium (Pazala) wenlingae Hu, Cotton & Monastyrskii, 2019	COI-5P NCBI only	Probably unvalid	Subspecies of *G. mandarinus* in Huang 2023

Genus *Iphiclides* Hübner, [1819]	**Sampling**	**Taxonomic status**	**Comment**
Iphiclides feisthamelii Duponchel, 1832	Sampled	Valid	
Iphiclides podalirinus Oberthür, 1890	Sampled	Valid	
Iphiclides podalirius Linnaeus, 1758	Sampled	Valid	

Genus *Lamproptera* Gray, 1832	**Sampling**	**Taxonomic status**	**Comment**
Lamproptera curius Fabricius, 1787	Sampled	Valid	
Lamproptera meges Zincken, 1831	Sampled	Valid	
Lamproptera paracurius Hu, Zhang & Cotton, 2014	Sampled	Valid	

Genus *Protographium* Munroe, 1961* (appear in Munroe [1960])	**Sampling**	**Taxonomic status**	**Comment**
Protographium leosthenes Doubleday, 1846	Sampled	Valid	

Table S2: Accession numbers of all Leptocircini specimens and outgroups. The third sub-column of the "Nuclear data" column indicates the reference for the scaffolding step, when the assembly was scaffolded with RagTag. The third sub-column of the "Mitochondrial data" shows if the same individual was used for nuclear and mitochondrial data.

Informations			**Nuclear data**			**Mitochondrial data**		
Genus	**Subgenus**	**species**	**Accession nuclear**	**Reference**	**Scaffolding reference if any**	**Accession number mitogenome**	**Mitogenome completness**	**Same individual**
Baronia		*brevicornis*	GCA_031353405.1	Marino et al. *under review*		OR063968.1	15	yes
Ornithoptera		*alexandrae*	GCA_029030555.1	Reboud et al. 2023		OQ590009.1	15	yes
Papilio		*xuthus*	GCF_000836235.1	Li et al. 2015		KU356933.1	15	no
Parnassius		*apollo*	GCA_907164705.1	Podsiadlowski et al. 2021		KF746065.1	15	no
Meandrusa		*payeni*	GCA_029286785.1	He et al. 2022		MN013000.1	15	no
Teinopalpus		*imperialis*	GCA_029286865.1	He et al. 2022		LS975139.1	15	no
Bhutanitis		*thaidina*	GCA_029286595.1	He et al. 2022		MN012973.1	15	no
Sericinus		*montela*	GCA_029286605.1	He et al. 2022		MN013020.1	15	no
Luehdorfia		*chinensis*	GCA_029286615.1	He et al. 2022		KU360130.1	15	no
Hypermnestra		*helios*	SRX5734328 (SRA)	Allio et al. 2020		LS975127.1	15	yes
Archon		*apollinus*	SRX5734362 (SRA)	Allio et al. 2020		LT999971.1	15	yes
Eurytides	*Eurytides*	*bellerophon*	FC451	This study	-	=	15	yes
Eurytides	*Eurytides*	*callias*	FC1030	This study	-	=	15	yes
Eurytides	*Eurytides*	*calliste*	FC449	This study	-	=	15	yes
Eurytides	*Eurytides*	*columbus*	FC1031	This study	FC1381	=	15	yes
Eurytides	*Eurytides*	*dioxippus*	FC976	This study	-	=	15	yes
Eurytides	*Eurytides*	*dolicaon*	FC1774	This study	-	=	15	yes
Eurytides	*Eurytides*	*leucaspis*	FC975	This study	-	=	15	yes
Eurytides	*Eurytides*	*orabilis*	FC448	This study	-	=	15	yes
Eurytides	*Eurytides*	*salvini*	FC515	This study	-	=	15	yes
Eurytides	*Eurytides*	*serville*	FC1243	This study	-	=	15	yes

Eurytides	*Eurytides*	*thyastes*	FC955	This study	-	=	15	yes
Eurytides	*Mimoides*	*agesilaus*	FC1246	This study	-	=	15	yes
Eurytides	*Mimoides*	*asius*	FC1796	This study	-	=	15	yes
Eurytides	*Mimoides*	*celadon*	FC1102	This study	-	=	15	yes
Eurytides	*Mimoides*	*epidaus*	FC1730	This study	-	=	15	yes
Eurytides	*Mimoides*	*marcellinus*	FC1798	This study	-	=	15	yes
Eurytides	*Mimoides*	*marcellus*	FC981	This study	-	=	15	yes
Eurytides	*Mimoides*	*philolaus*	FC1728	This study	-	=	15	yes
Eurytides	*Mimoides*	*zonaria*	FC475	This study	-	=	15	yes
Eurytides	*Mimoides*	*ariarathes*	FC1012	This study	-	=	15	yes
Eurytides	*Mimoides*	*euryleon*	FC576	This study	-	=	15	yes
Eurytides	*Mimoides*	*ilus*	FC1331	This study	-	=	15	yes
Eurytides	*Mimoides*	*lysithous*	FC1014	This study	-	=	15	yes
Eurytides	*Mimoides*	*microdamas*	FC1827	This study	-	=	15	yes
Eurytides	*Mimoides*	*pausanias*	FC1981	This study	FC1381	=	15	yes
Eurytides	*Mimoides*	*phaon*	FC1024	This study	-	=	15	yes
Eurytides	*Mimoides*	*protodamas*	FC1466	This study	-	=	15	yes
Eurytides	*Mimoides*	*thymbraeus*	FC1028	This study	-	=	15	yes
Eurytides	*Mimoides*	*xeniades*	FC1042	This study	-	=	15	yes
Eurytides	*Mimoides*	*xynias*	FC1010	This study	-	=	15	yes
Eurytides	*Protesilaus*	*aguiari*	FC420	This study	-	=	15	yes
Eurytides	*Protesilaus*	*earis*	FC421	This study	FC1381	=	15	yes
Eurytides	*Protesilaus*	*exiguus*	FC1320	This study	-	=	15	yes
Eurytides	*Protesilaus*	*glaucolaus*	FC422	This study	-	=	15	yes
Eurytides	*Protesilaus*	*helios*	FC986	This study	-	=	15	yes
Eurytides	*Protesilaus*	*leucosilaus*	FC1801	This study	-	=	15	yes
Eurytides	*Protesilaus*	*macrosilaus*	FC987	This study	-	=	15	yes
Eurytides	*Protesilaus*	*molops*	FC423	This study	-	=	15	yes
Eurytides	*Protesilaus*	*orthosilaus*	FC1802	This study	-	=	15	yes
Eurytides	*Protesilaus*	*protesilaus*	FC989	This study	-	=	15	yes
Eurytides	*Protesilaus*	*rothschildianus*	FC1818	This study	-	=	15	yes

Eurytides	*Protesilaus*	*stenodesmus*	FC450	This study	-	=	15	yes
Eurytides	*Protesilaus*	*telesilaus*	FC1002	This study	-	=	15	yes
Graphium	*Arisbe*	*adamastor*	FC1173	This study	FC1382	=	15	yes
Graphium	*Arisbe*	*agamedes*	FC1177	This study	FC1382	=	15	yes
Graphium	*Arisbe*	*almansor*	10603	This study	FC1382	=	15	yes
Graphium	*Arisbe*	*angolanus*	FC1179	This study	FC1382	=	15	yes
Graphium	*Arisbe*	*antheus*	FC1382	This study	-	=	15	yes
Graphium	*Arisbe*	*auriger*	JS112	This study	FC1382	=	15	yes
Graphium	*Arisbe*	*aurivilliusi*		This study	-	FC1485	15	
Graphium	*Arisbe*	*biokoensis*	10608	This study	FC1382	=	15	yes
Graphium	*Arisbe*	*colonna*	FC1183	This study	FC1382	=	15	yes
Graphium	*Arisbe*	*cyrnus*	FC1185	This study	FC1382	=	15	yes
Graphium	*Arisbe*	*endochus*	FC1186	This study	FC1382	=	15	yes
Graphium	*Arisbe*	*evombar*	FC1189	This study	FC1382	=	15	yes
Graphium	*Arisbe*	*fulleri*	FC1191	This study	FC1382	=	15	yes
Graphium	*Arisbe*	*gudenusi*	FC455	This study	FC1382	=	15	yes
Graphium	*Arisbe*	*hachei*	10612	This study	FC1382	=	15	yes
Graphium	*Arisbe*	*illyris*	10615	This study	FC1382	=	15	yes
Graphium	*Arisbe*	*junodi*	JS116	This study	FC1382	=	15	yes
Graphium	*Arisbe*	*kirbyi*	FC1202	This study	FC1382	=	15	yes
Graphium	*Arisbe*	*latreillianus*	FC1203	This study	FC1382	=	15	yes
Graphium	*Arisbe*	*leonidas*	FC1205	This study	FC1382	=	15	yes
Graphium	*Arisbe*	*levassori*	FC1470	This study	FC1382	=	15	yes
Graphium	*Arisbe*	*liponesco*	FC1211	This study	FC1382	=	15	yes
Graphium	*Arisbe*	*morania*	FC523	This study	FC1382	=	15	yes
Graphium	*Arisbe*	*philonoe*	FC1212	This study	FC1382	=	15	yes
Graphium	*Arisbe*	*policenes*	10622	This study	FC1382	=	15	yes
Graphium	*Arisbe*	*policenoides*		This study	-	FC1492	only COX1	
Graphium	*Arisbe*	*polistratus*	FC1228	This study	FC1382	=	15	yes
Graphium	*Arisbe*	*porthaon*	FC1230	This study	FC1382	=	15	yes
Graphium	*Arisbe*	*ridleyanus*	FC522	This study	FC1382	=	15	yes

Graphium	*Arisbe*	*schaffgotschi*	JS120	This study	FC1382	=	15	yes
Graphium	*Arisbe*	*schubotzi*	FC1234	This study	FC1382	=	15	yes
Graphium	*Arisbe*	*simoni*	JS121	This study	FC1382	=	15	yes
Graphium	*Arisbe*	*taboranus*	10629	This study	FC1382	=	15	yes
Graphium	*Arisbe*	*tynderaeus*	10631	This study	FC1382	=	15	yes
Graphium	*Arisbe*	*ucalegon*	10632	This study	FC1382	=	15	yes
Graphium	*Graphium*	*agamemnon*	FC1381	This study	-	=	15	yes
Graphium	*Graphium*	*anthedon*	10934	This study	FC1381	=	15	yes
Graphium	*Graphium*	*arycles*	FC1264	This study	FC873	=	15	yes
Graphium	*Graphium*	*bathycles*	FC1266	This study	FC873	=	15	yes
Graphium	*Graphium*	*browni*	FC1268	This study	FC1381	=	15	yes
Graphium	*Graphium*	*chironides*	10639	This study	FC873	=	15	yes
Graphium	*Graphium*	*cloanthus*	FC1273	This study	FC1381	=	15	yes
Graphium	*Graphium*	*codrus*	10907	This study	FC1381	=	15	yes
Graphium	*Graphium*	*doson*	FC873	This study	-	=	15	yes
Graphium	*Graphium*	*empedovana*	10912	This study	FC1381	=	15	yes
Graphium	*Graphium*	*eurypylus*	FC1380	This study	FC873	=	15	yes
Graphium	*Graphium*	*evemon*	10919	This study	FC873	=	15	yes
Graphium	*Graphium*	*gelon*	FC452	This study	FC1381	=	15	yes
Graphium	*Graphium*	*hicetaon*	10922	This study	FC1381	=	15	yes
Graphium	*Graphium*	*kosii*	FC1285	This study	FC1381	=	15	yes
Graphium	*Graphium*	*leechi*	FC1286	This study	FC873	=	15	yes
Graphium	*Graphium*	*macfarlanei*	FC1288	This study	FC1381	=	15	yes
Graphium	*Graphium*	*macleayanus*	FC1289	This study	FC1381	=	15	yes
Graphium	*Graphium*	*meeki*	FC1291	This study	FC1381	=	15	yes
Graphium	*Graphium*	*mendana*	FC1293	This study	FC873	=	15	yes
Graphium	*Graphium*	*meyeri*	FC1726	This study	FC1381	=	15	yes
Graphium	*Graphium*	*milon*	FC916	This study	FC1381	=	15	yes
Graphium	*Graphium*	*monticolus*	FC1298	This study	FC1381	=	15	yes
Graphium	*Graphium*	*phidias*	FC1652	This study	FC1381	=	15	yes
Graphium	*Graphium*	*procles*	10935	This study	FC873	=	15	yes

Graphium	*Graphium*	*sandawanum*	FC918	This study	FC1381	=	15	yes
Graphium	*Graphium*	*sarpedon*	FC1302	This study	FC1381	=	15	yes
Graphium	*Graphium*	*septentrionicolus*	FC1825	This study	FC1381	=	15	yes
Graphium	*Graphium*	*stresemanni*	10951	This study	FC1381	=	15	yes
Graphium	*Graphium*	*sumatrana*	FC1272	This study	FC1381	=	15	yes
Graphium	*Graphium*	*teredon*	FC1826	This study	FC1381	=	15	yes
Graphium	*Graphium*	*wallacei*	10953	This study	FC1381	=	15	yes
Graphium	*Graphium*	*weiskei*	10954	This study	FC1381	=	15	yes
Graphium	*Pathysa*	*agetes*	11008	This study	FC1341	=	15	yes
Graphium	*Pathysa*	*androcles*	11013	This study	FC1341	=	15	yes
Graphium	*Pathysa*	*antiphates*	FC1341	This study	-	=	15	yes
Graphium	*Pathysa*	*aristeus*	11074	This study	FC1341	=	15	yes
Graphium	*Pathysa*	*decolor*	11025	This study	FC1341	=	15	yes
Graphium	*Pathysa*	*delessertii*	10957	This study	FC1341	=	15	yes
Graphium	*Pathysa*	*deucalion*	FC1143	This study	FC1341	=	15	yes
Graphium	*Pathysa*	*dorcus*	11030	This study	FC1341	=	15	yes
Graphium	*Pathysa*	*ebertorum*	FC1069	This study	FC1341	=	15	yes
Graphium	*Pathysa*	*encelades*	FC1086	This study	FC1341	=	15	yes
Graphium	*Pathysa*	*epaminondas*	FC1100	This study	FC1341	=	15	yes
Graphium	*Pathysa*	*euphrates*	SP95	This study	FC1341	=	15	yes
Graphium	*Pathysa*	*felixi*	FC1842	This study	FC1341	=	15	yes
Graphium	*Pathysa*	*idaeoides*	FC516	This study	FC1341	=	15	yes
Graphium	*Pathysa*	*macareus*	10967	This study	FC1341	=	15	yes
Graphium	*Pathysa*	*megarus*	FC1151	This study	FC1341	=	15	yes
Graphium	*Pathysa*	*nomius*	FC1170	This study	FC1341	=	15	yes
Graphium	*Pathysa*	*ramaceus*	FC1158	This study	FC1341	=	15	yes
Graphium	*Pathysa*	*rhesus*	FC919	This study	FC1341	=	15	yes
Graphium	*Pathysa*	*stratiotes*	11041	This study	FC1341	=	15	yes
Graphium	*Pathysa*	*stratocles*	11067	This study	FC1341	=	15	yes
Graphium	*Pathysa*	*thule*	10985	This study	FC1341	=	15	yes
Graphium	*Pathysa*	*xenocles*	FC1159	This study	FC1341	=	15	yes

Graphium	*Pazala*	*alebion*	11042	This study	FC1381	=	15	yes
Graphium	*Pazala*	*asakurae*		This study	-	FC1496	ND4L is missing	
Graphium	*Pazala*	*caschmirensis*	SP685	This study	FC1381	=	15	yes
Graphium	*Pazala*	*confucius*				OK136253	15	
Graphium	*Pazala*	*daiyuanae*				MG197773	only CO1-5P	
Graphium	*Pazala*	*eurous*	FC1119	This study	FC1381	=	15	yes
Graphium	*Pazala*	*garhwalica*	SP683	This study	FC1381	=	15	yes
Graphium	*Pazala*	*hoeneanus*				MG197764	only CO1-5P	
Graphium	*Pazala*	*mandarinus*	11055	This study	FC1381	=	15	yes
Graphium	*Pazala*	*mullah*	FC1129	This study	FC1381	=	15	yes
Graphium	*Pazala*	*paphus*		This study	FC1381	FC1493	ND2 is missing	
Graphium	*Pazala*	*parus*	FC1131	This study	FC1381	=	15	yes
Graphium	*Pazala*	*sichuanica*	11053	This study	FC1381	=	15	yes
Graphium	*Pazala*	*tamerlanus*	11063	This study	FC1381	=	15	yes
Graphium	*Pazala*	*wenlingae*				MH892413	only CO1-5P	
Iphiclides		*feisthamelii*	FC474	This study	FC536	=	15	yes
Iphiclides		*podalirinus*	FC1650	This study	FC536	=	15	yes
Iphiclides		*podalirius*	FC536	This study	-	=	15	yes
Lamproptera		*curius*	FC1249	This study	GCA_029286875.1	=	15	yes
Lamproptera		*meges*	FC1251	This study	GCA_029286875.1	=	15	yes
Lamproptera		*paracurius*	FC434	This study	GCA_029286875.1	=	15	yes
Protographium		*leosthenes*	FC1108	This study	FC1381	=	15	yes

Table S3: Aphid triplets replicates and their results. Each series of triplet contains 5 replicates, with a random combination of species from the tested clades.

Tested triplet	Number of analyzed genes	Events (%)					Posterior imbalance GFc triplet (%)	
		none	ILS	GF	ILSc	GFc		
stratiotes_11041,xenocles_FC1159,encelades_FC1086	1027	25	4,2	12,8	15,5	42,5	((BC)A)	57,8
stratiotes_11041,macareus_10967,stratocles_11067	1047	23,6	3,6	11,4	15,3	46,1	((BC)A)	58,3
agetes_11008,xenocles_FC1159,idaeoides_FC516	1018	24,6	3,8	12,5	14,3	44,8	((BC)A)	56,3
agetes_11008,delesserti_10957,deucalion_FC1143	1034	22,1	2,8	12,6	13	49,4	((BC)A)	55,5
agetes_11008,delesserti_10957,ebertorum_FC1069	1044	21,3	2,4	13,7	12,2	50,4	((BC)A)	56,6
Mean	**1034**	**23,32**	**3,36**	**12,6**	**14,06**	**46,64**	**56,9**	
ramaceus_FC1158,stratocles_11067,agetes_11008	1042	22,5	4,2	7,4	13,3	52,7	((AC)B)	61,7
xenocles_FC1159,encelades_FC1086,stratiotes_11041	1027	23	4,4	4,7	14,1	53,9	((AC)B)	66,2
delesserti_10957,thule_10985,stratiotes_11041	1039	23,1	4,3	7,5	13,4	51,7	((AC)B)	62,2
delesserti_10957,encelades_FC1086,agetes_11008	1039	21,4	5	5,3	13,8	54,6	((AC)B)	64,8
ramaceus_FC1158,ebertorum_FC1069,agetes_11008	1045	22,7	4,4	7,2	13,8	51,8	((AC)B)	62,3
Mean	**1038,4**	**22,54**	**4,46**	**6,42**	**13,68**	**52,94**	**63,44**	
euphrates_SP95,stratiotes_11041,stratocles_11067	1038	29,4	6,1	14,6	16,6	33,3	((AC)B)	54,1
euphrates_SP95,agetes_11008,idaeoides_FC516	1032	28	6	14,9	17,3	33,8	((AC)B)	61,6
decolor_11025,stratiotes_11041,thule_10985	1020	28,7	5,9	15,3	17,4	32,8	((AC)B)	56,2
antiphates_FC1341,agetes_11008,stratocles_11067	964	29,6	5,8	15,4	15	34,1	((AC)B)	52,1
decolor_11025,agetes_11008,deucalion_FC1143	1018	29,3	5,8	15,4	16,4	33,1	((AC)B)	55
Mean	**1014,4**	**29**	**5,92**	**15,12**	**16,54**	**33,42**	**55,8**	
epaminondas_FC1100,agetes_11008,macareus_10967	992	31,3	5,4	19,6	15,3	28,3	((AC)B)	58,3
decolor_11025,agetes_11008,delesserti_10957	1043	30,6	6	21,2	17,4	24,7	((AC)B)	57,8
androcles_11013,agetes_11008,delesserti_10957	1046	31,3	5,9	21,8	16,7	24,3	((AC)B)	57,5
antiphates_FC1341,stratiotes_11041,macareus_10967	982	31	6	19,7	16,1	27,3	((AC)B)	59,7
androcles_11013,stratiotes_11041,ramaceus_FC1158	1050	31,4	6,3	20,3	18	24,1	((AC)B)	58,6
Mean	**1022,6**	**31,12**	**5,92**	**20,52**	**16,7**	**25,74**	**58,38**	

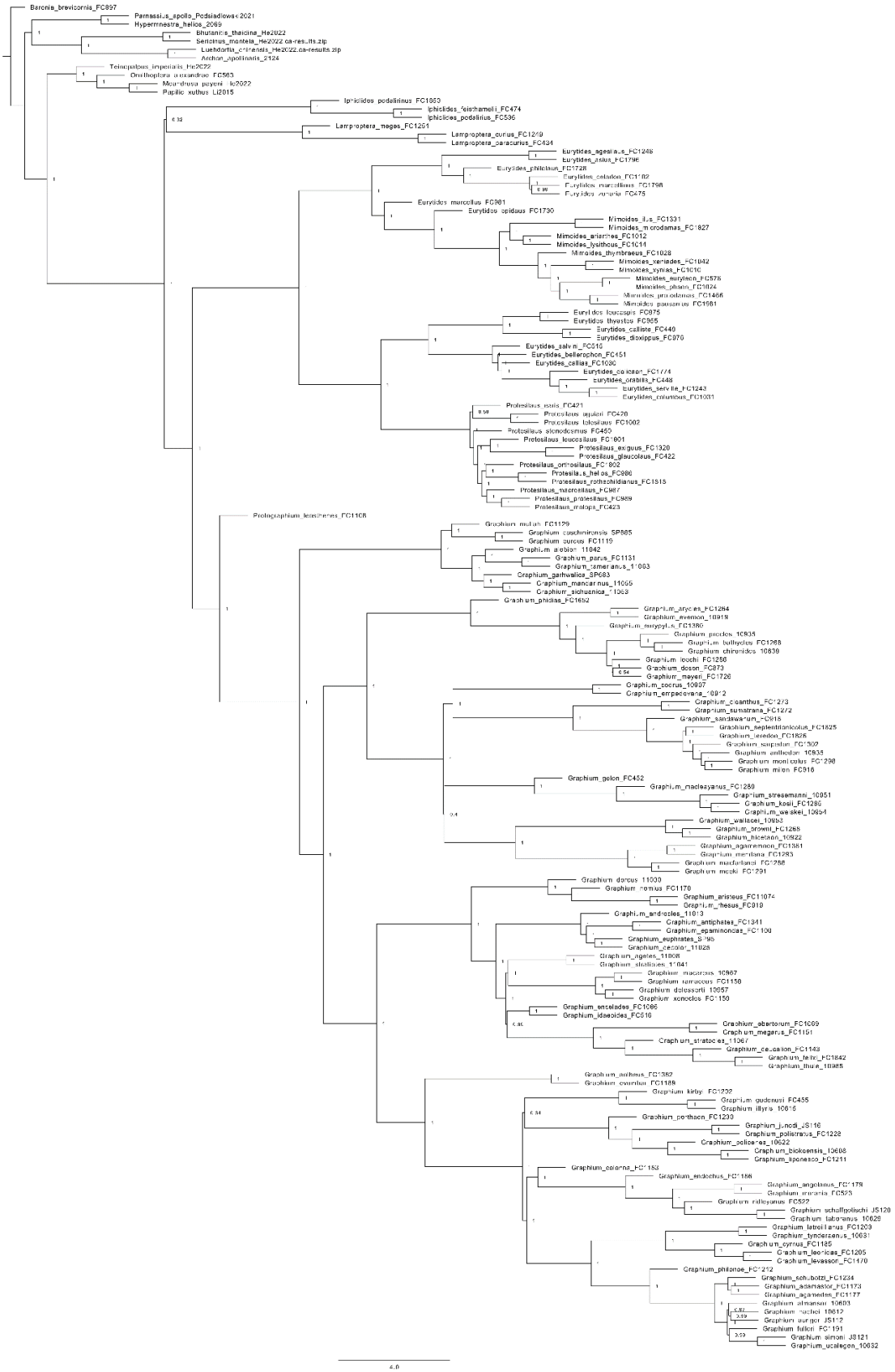


Figure S1: Phylogenomic relationships of Leptocircini with the supertree approach (*ASTRAL_1402_NT*). The phylogeny has been reconstructed with ASTRAL, a supertree approach based on 1,402 gene trees inferred with the maximum-likelihood inference in IQ-TREE.



Figure S2: Phylogenomic relationships of Leptocircini with the Dataset 2 in amino-acids (*IQTREE_4525_AA*). The phylogeny has been reconstructed with a supermatrix approach of 4,525 single-copy amino-acid genes and a maximum-likelihood inference in IQ-TREE.

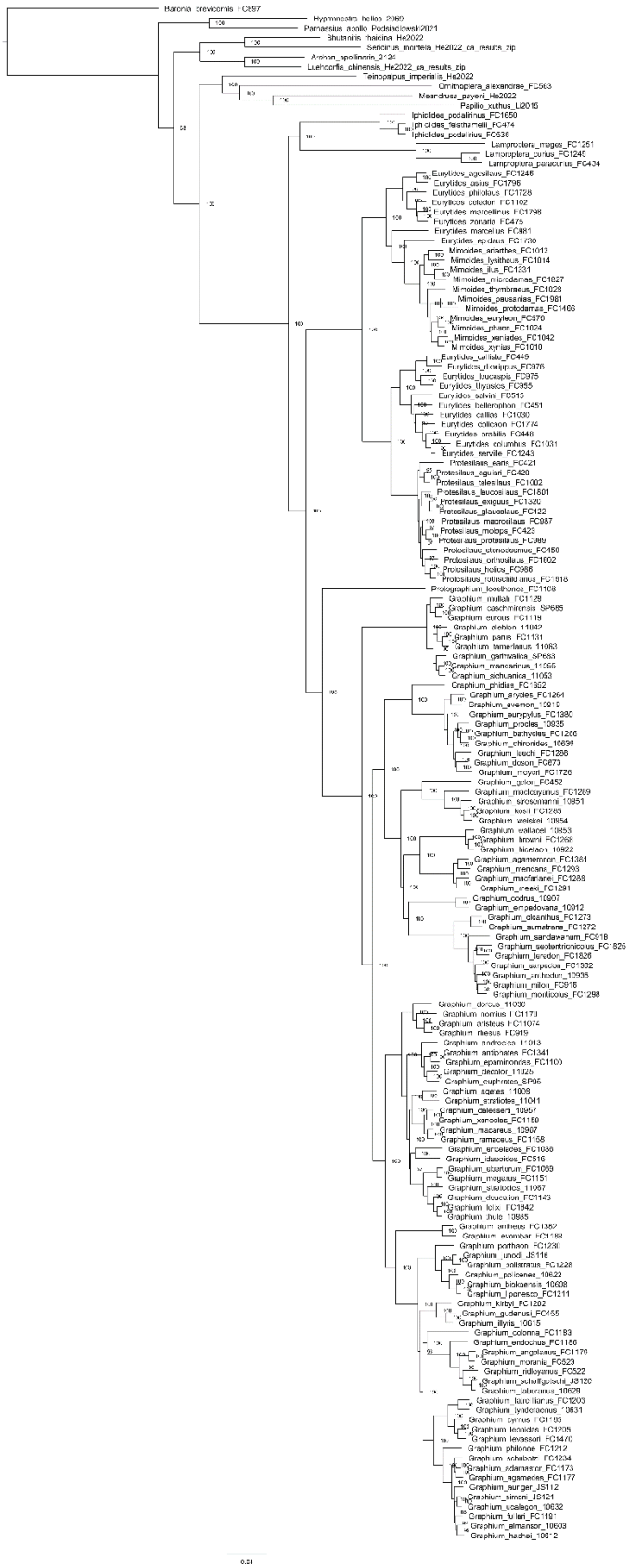


Figure S3: Phylogenomic relationships of Leptocircini (*IQTREE_1402_AA*). The phylogeny has been reconstructed with a supermatrix approach of 1,402 single-copy amino-acid genes and a maximum-likelihood inference in IQ-TREE.

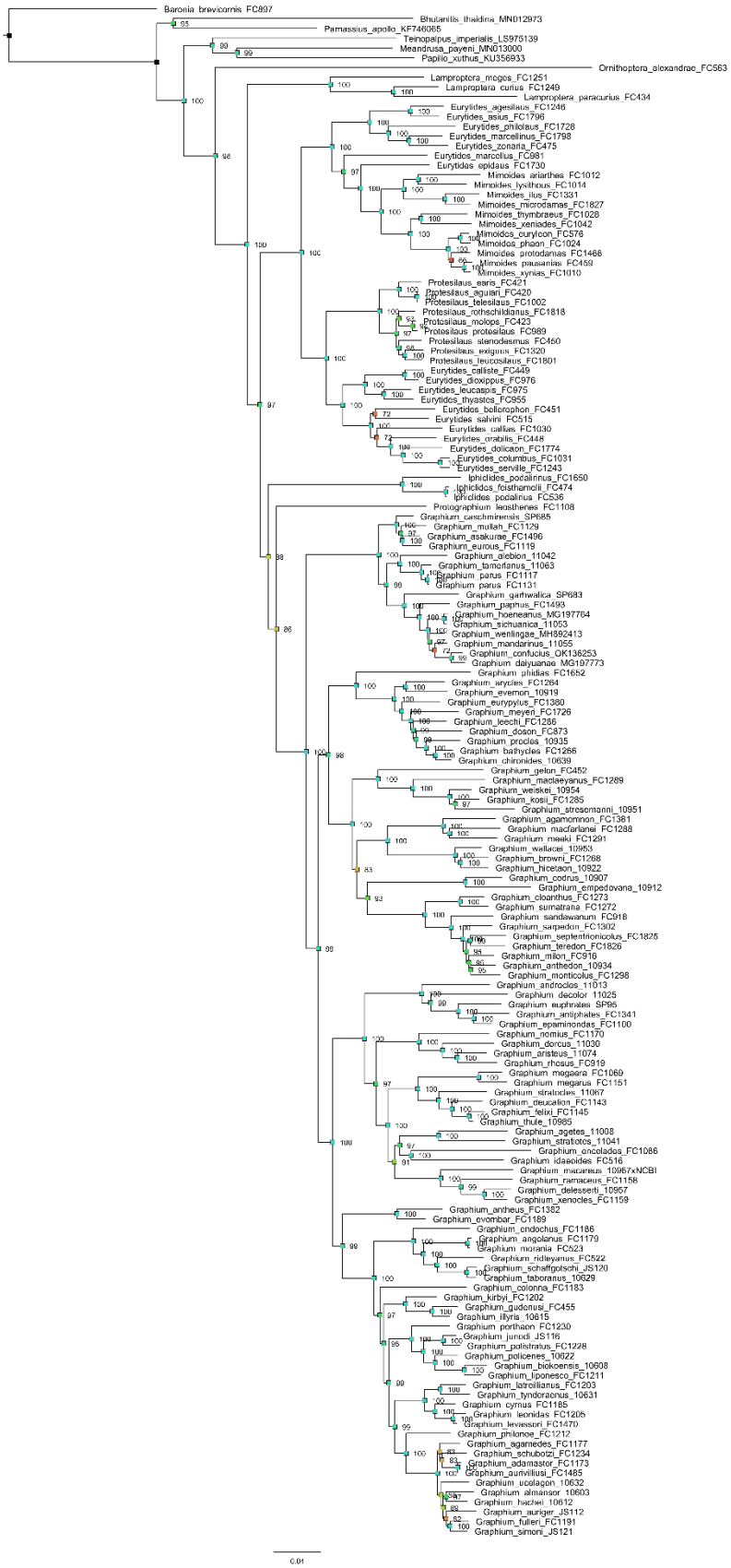


Figure S4: Phylogenomic relationships of Leptocircini using mitochondrial data (*IQTREE_15_Mito*). The phylogeny has been reconstructed with a supermatrix approach of 15 mitochondrial genes and a maximum-likelihood inference in IQ-TREE.

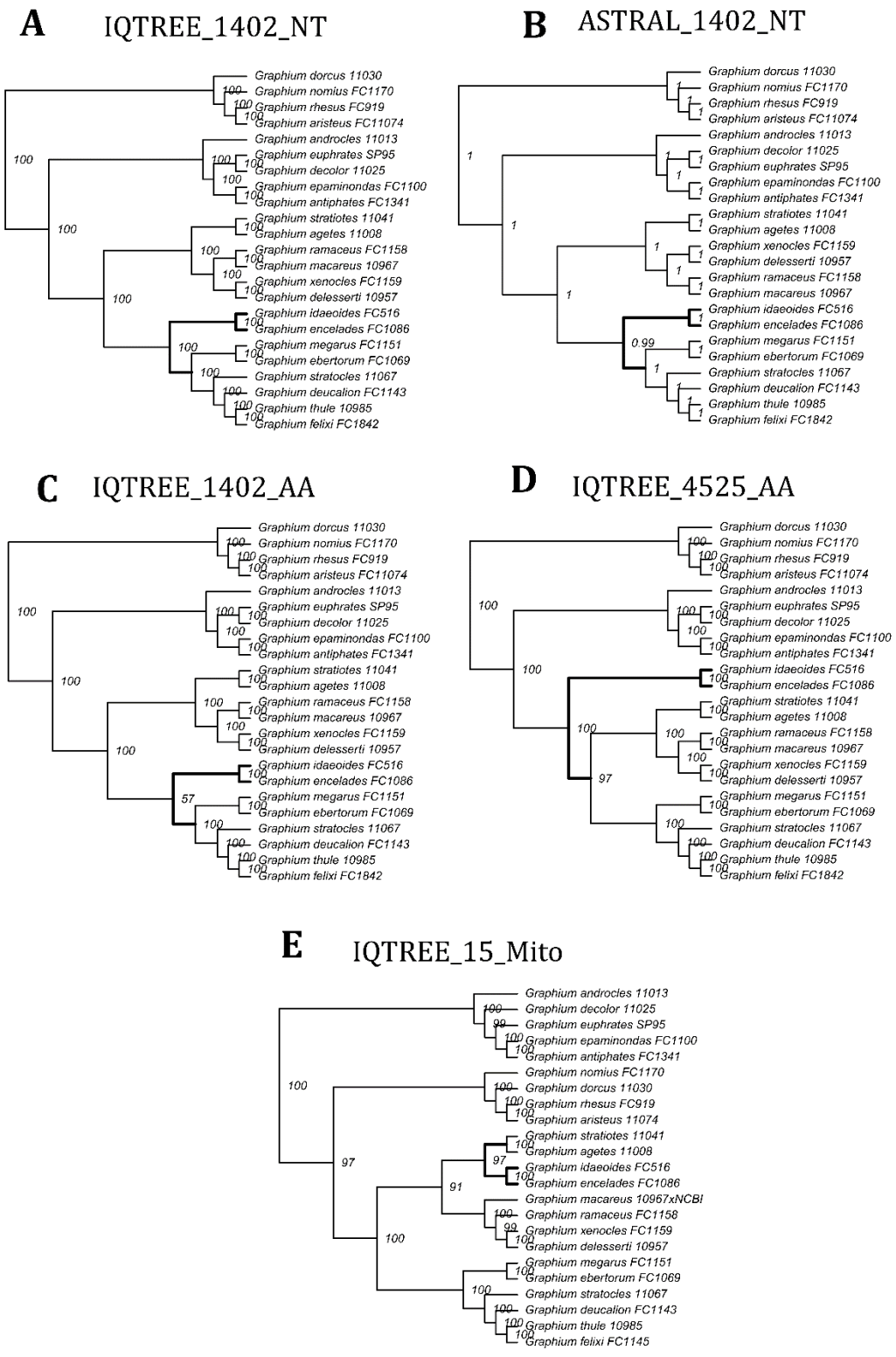


Figure S5: Discordance in phylogenomic analyses for the clade *Pathysa s.l.*. **A**. Subtree of the supermatrix using MFP+MERGE option and Dataset 1, i.e *IQTREE_1402_NT* analysis, **B**. Subtree of the supertree approach, i.e. *ASTRAL_1402_NT* analysis, **C.** Subtree of the supermatrix using amino acid Dataset 2 i.e. *IQTREE_4525_AA* analysis, **D.** Subtree of the supermatrix using amino acid Dataset 1 i.e *IQTREE_1402_AA* analysis, **E.** Subtree of the mitogenome analysis i.e. *IQTREE_15_Mito* analysis. Highlighted branches indicate those whose positions vary according to the analysis.

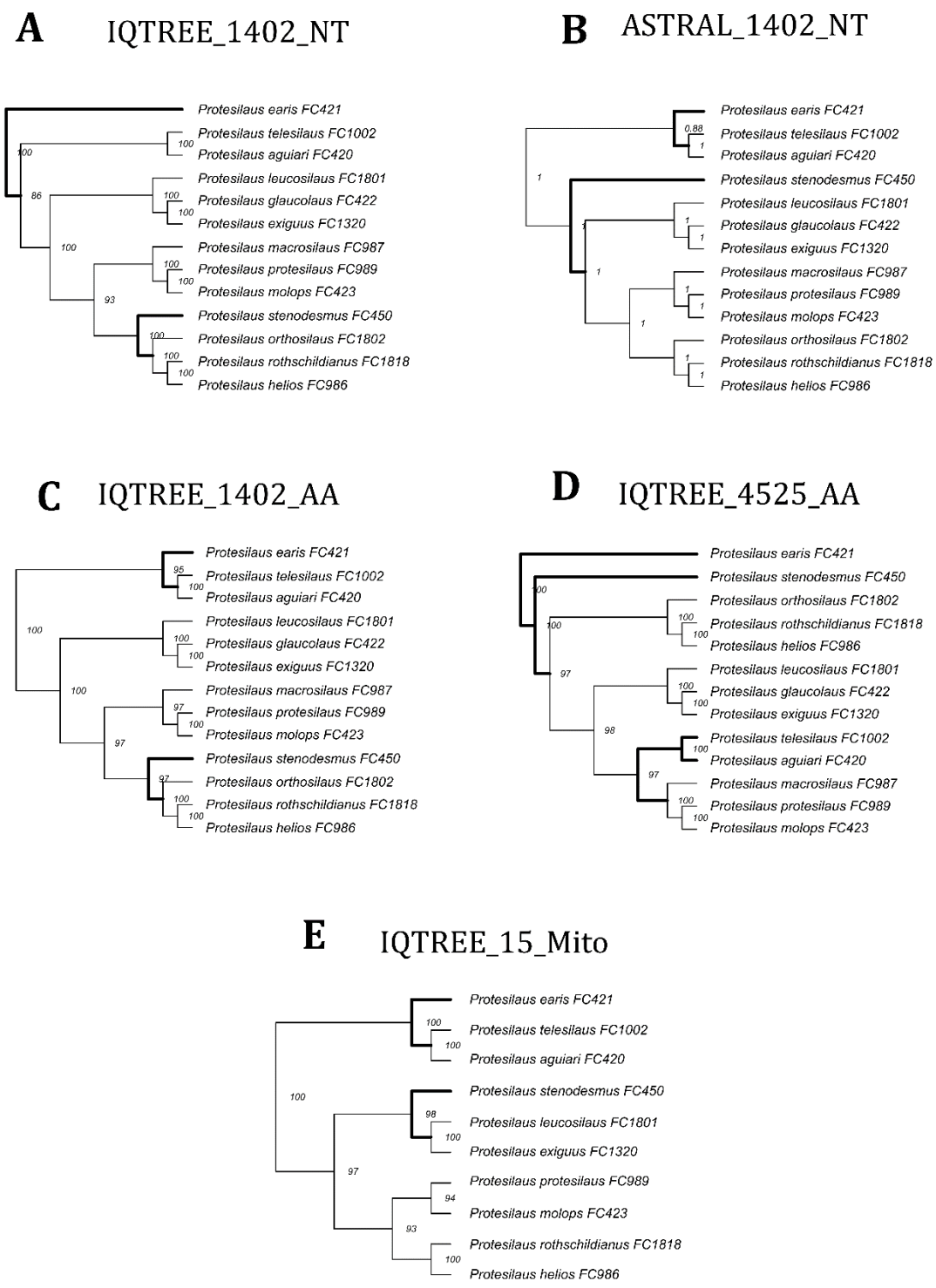


Figure S6: Discordance in phylogenomic analyses for the subgenus *Protesilaus*. **A**. Subtree of the supermatrix using MFP+MERGE option and Dataset 1, i.e analysis *IQTREE_1402_NT* analysis, **B**. Subtree of the supertree approach, i.e. *ASTRAL_1402_NT* analysis, **C.** Subtree of the supermatrix using amino acid Dataset 2 i.e. *IQTREE_4525_AA* analysis, **D.** Subtree of the supermatrix using amino acid Dataset 1 i.e *IQTREE_1402_AA* analysis, **E.** Subtree of the mitogenome analysis i.e. *IQTREE_15_Mito* analysis. Note that species are bearing the subgenus name *Protesilaus* instead of their genus name *Eurytides*. Highlighted branches indicate those whose positions vary according to the analysis.

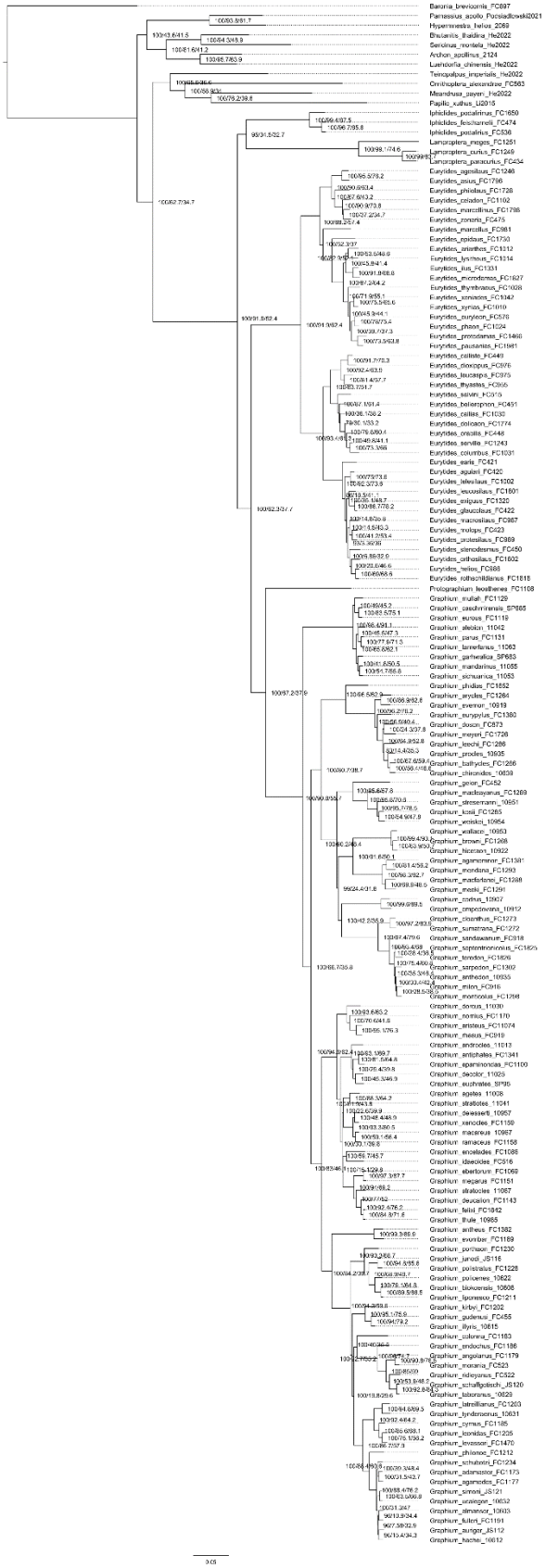


Figure S7: Results of gene and site Concordance Factor. The phylogenomic tree is from the *IQTREE_1402_NT* analysis. The UFBS/gCF/sCF values are displayed at each node. gCF values have been calculated with the 1,402 gene trees, and sCF values with the supermatrix of the same analysis.

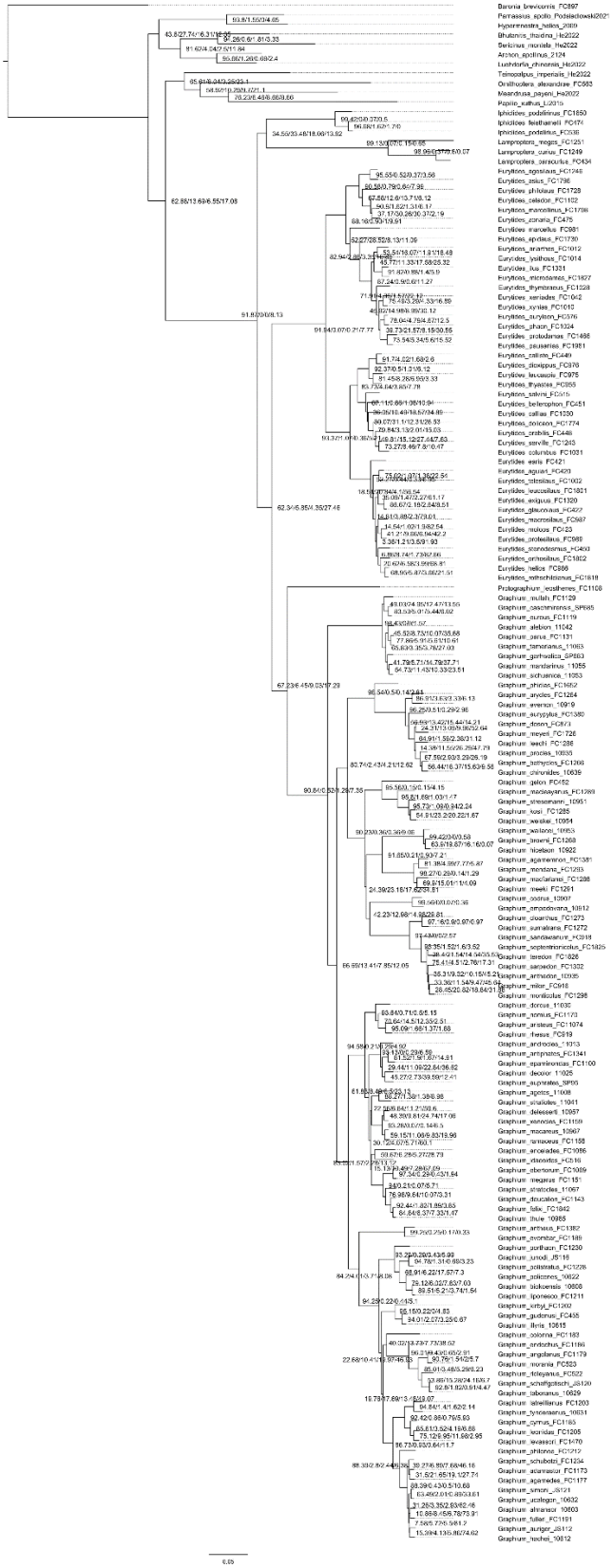
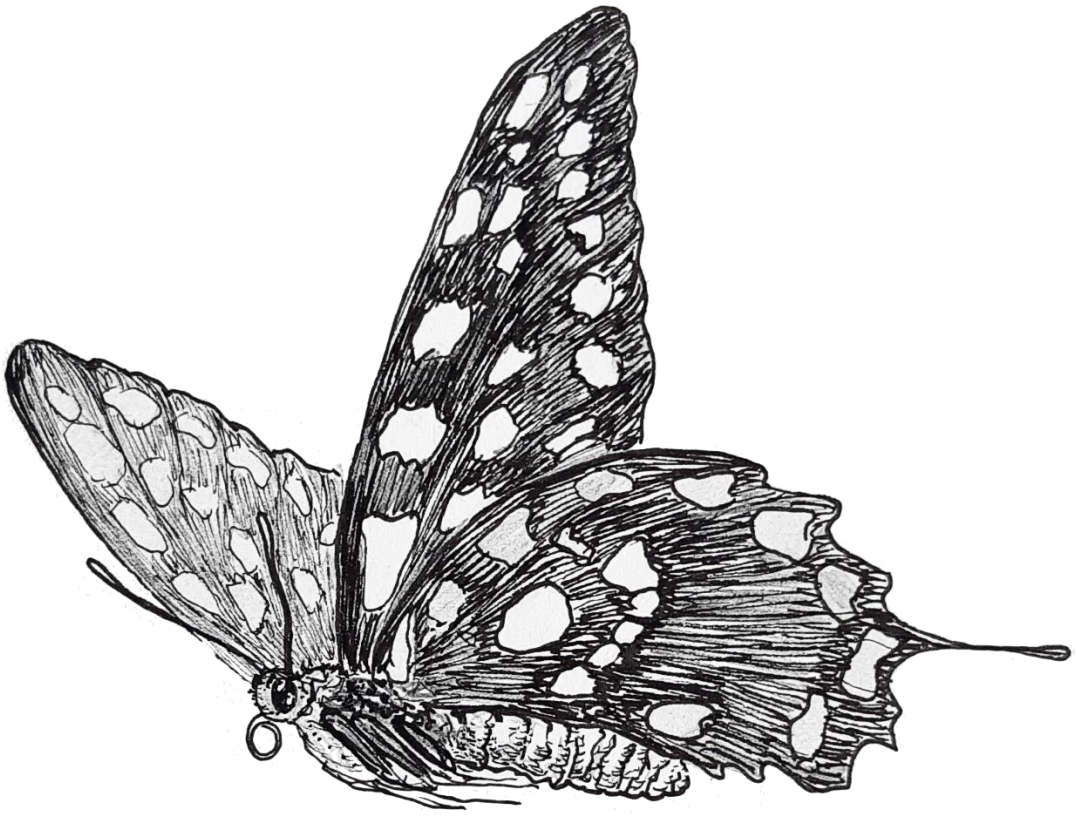


Figure S8: Gene Concordance factor results. The phylogenomic tree is from the *IQTREE_1402_NT* analysis. The gCF/gDF1/gDF2/gDFP is displayed at each node. These values have been calculated with the 1,402 gene trees. gCF is the number of gene trees in accordance with the consensus tree, gDF1 and gDF2 are the two triplet alternatives. gDFP are all the remaining trees (polytomies and paraphyly of the triplet) so that gCF + gDF1 + gDF2 + gDFP = 1.

CHAPTER 2.2



Pharmacophagus antenor en vol - Eliette Reboud

Species-level genome-wide phylogeny of the giant swallowtail butterflies of the tribe Troidini (Lepidoptera: Papilionidae)

Eliette L. Reboud[1]*, Benoit Nabholz[1,2], Emmanuelle Chevalier[1], Bérénice Lafon[1], Marie-ka Tilak[1], Carlos G. C. Mielke[3], Jovanni Fagua[4], Adam M. Cotton[5] & Fabien L. Condamine[1]*

Affiliations

[1] *Institut des Sciences de l'Evolution de Montpellier (Université Montpellier | CNRS | IRD | EPHE), Place Eugène Bataillon, 34095 Montpellier, France*
[2] *Institut Universitaire de France (IUF), Paris, France*
[3] *Caixa postal 1206, 84.145-000 Carambeí, Paraná, Brazil*
[4] *Department of Biology, Pontificia Universidad Javeriana, Bogotá D.C., Colombia*
[5] *86/2 Moo 5, Tambon Nong Kwai, Hang Dong, Chiang Mai, Thailand*

Corresponding authors (*)

Eliette L. Reboud: eliette.reboud@gmail.com

Fabien L. Condamine: fabien.condamine@gmail.com

Running head

Phylogenomic of the tribe Troidini (Papilionidae)

Abstract

Troidini is a remarkable tribe of Papilionidae distributed throughout the tropics of Southeast Asia and the Neotropics, and including some of the largest and most charming butterflies. However, despite some well-studied groups such as the Neotropical cattlehearts and birdwing butterflies, a comprehensive phylogeny remains incomplete, particularly for the genera *Atrophaneura*, *Byasa*, *Losaria*, and *Pachliopta*. As a result, several unresolved issues remain regarding their taxonomy and systematics. This paper aims to present a revised taxonomic working list for the Troidini, consisting of 139 valid species. Using a combination of long- and short-read data, we assembled five new reference genomes and generated highly covered and scaffolded whole genomes for 131 species (>94% of the total species diversity) to infer densely sampled phylogenetic hypotheses. Based on mitochondrial or thousands of nuclear genes and multiple phylogenetic approaches, we obtained robust phylogenomic trees representing >97% of Troidini species. The backbone of previous studies was generally conserved, confirming the existence of the subtribes Battina and Troidina, as well as *Cressida* being sister to *Euryades* + *Parides*. Unexpectedly, we found that *Byasa daemonius* was sister to all *Atrophaneura* and *Byasa*, while *B. crassipes* was sister to *Atrophaneura*. These findings raise questions about the validity of these genera. Therefore, we revised the status of genus *Byasa*, now considering it a subgenus of *Atrophaneura*, and proposed that *B. daemonius* should be part of a new *Atrophaneura* subgenus. Additionally, we have reclassified *Losaria* as a *Pachliopta* subgenus. Our study also highlights certain relationships between the genera *Ornithoptera* and *Troides*, and we have revised and discussed some of their species. Overall, this research will undoubtedly serve as a foundation for further study of this diverse and captivating butterfly tribe.

Keywords: Birdwing Butterflies, *Byasa*, *Parides*, *Pharmacophagus*, phylogenomic, reference genome

Introduction

Within Papilionidae, the tribe Troidini Talbot, 1939 comprises about 140 almost exclusively tropical species divided into 12 genera, with most species encountered in the lowland rainforests of the Neotropical and Indo-Australasian regions (Braby et al. 2005, Condamine et al. 2015, Nakae 2021, Gamboa et al. 2022) (Figure 1). The Troidine butterflies are famous because of their very charismatic species (e.g. *Ornithoptera sp.*) but actually presents a great morphological disparity, ranging from the discrete Amazonian dwarf black butterfly *Parides klagesi* (<5 cm wingspan) to the giant Papuan butterfly *Ornithoptera alexandrae* (up to 28 cm wingspan for the females) with the exuberant colors of its males (Mitchell et al. 2016, Reboud et al. 2023). All species larvae of Troidini feed on Aristolochiaceae, a large family of plants containing toxins such as aristolochic acids (which are known to be nephrotoxic and carcinogenic; Debelle et al. 2008, Chen et al. 2015) and labdanoic acids, and that the larvae sequester for chemical defenses against predators (Klitzke 1992, Leitão and Kaplan 1992, Silva-Brandão and Solferini 2007, Allio et al. 2021a).

Troidini have undergone major changes in composition since its introduction by Talbot in 1939. Initially, it was containing only a few of the genera currently included (especially genera in the birdwing butterflies (today *Troides*, *Trogonoptera* and *Ornithoptera*), the *Atrophaneura sensu-lato* group (today *Atrophaneura*, *Losaria*, *Pachliopta*, *Byasa*), and the pipevine swallowtails (*Battus*), with much variation in the composition according to each author (Talbot 1939, 1949 based on previous work of Horsfield and Moore 1858, Haase 1893, Rothschild and Jordan 1906, Bryk 1929, Zeuner 1943, Ford 1941, 1944, Ehrlich 1958). It was not until the work of Munroe (1960), Hancock (1978), Igarashi (1984) and Miller (1987) that the tribe acquired the identity it has today, with the inclusion of the genera *Parides*, *Euryades*, *Cressida* (the two latter were formerly in the tribe Cressidini), *Pharmacophagus* and all previously cited genera (Figure 1). But even then, and for the decades following, relationships within the tribe have been widely debated and controversial (Tyler et al. 1994, Morinaka et al. 1999, Braby et al. 2005, Silva-Brandão et al. 2005, Racheli 2006, Hancock 2007) and the backbone of Troidini remained widely unknown.

Due to their bipartite distribution, phylogenetic studies of the Troidini have often focused either on the Neotropics (*Parides*, *Battus*, *Euryades*: Silva-Brandao et al. 2005, Condamine et al. 2012a), or on the Indo-Australasian region (*Ornithoptera*, *Trogonoptera*, *Troides*: Morinaka et al. 2000, Condamine et al. 2015; *Losaria*: Xu et al. 2020). In the latter, however, the clade comprising the genera *Pachliopta*, *Byasa*, *Atrophaneura* and *Losaria* has been largely left out of phylogenetic studies, and no phylogeny at the species level has been published to date. A surprising fact, given that the composition of this group and the number of genera recognized was one of the most

hotly debated questions within the classification of Troidini and remains so today, pending new robust and complete evidence (Racheli 2006). Originally, the four previously cited genera were included into a single genera *Atrophaneura* as species groups alongside the monospecific genus *Pharmacophagus* (Ford 1944). The group *Pachliopta* was considered too distinct and ranked to the genus level by Munroe 1960. This was disputed by Hancock (1980, 1983), but nonetheless accepted by Miller (1987), who also included the *Losaria* group within the *genus Pachliopta* as a subgenus. *Pharmacophagus* was separated as an independent genus by Igarashi (1979, 1984) and Miller (1987). Finally, Tsukada and Nishiyana (1982) ranked *Losaria* at the genus level, and Racheli and Cotton (2010) elevated the tailed species of *Atrophaneura* into the separate genus *Byasa* (based on early suggestions and revisions of Chou et al. 1994, Wu and Bai 2001). These four genera are currently the most accepted in recent studies (Condamine et al. 2018, Allio et al. 2021a, Nakae 2021).

A few studies have recently focused on Troidini as a whole, yet the tribe was often integrated into a broader phylogeny or suffered from a lack of extensive sampling (Simonsen et al. 2011, Condamine et al. 2013, 2018, Allio et al. 2021a). In the global phylogeny of Papilionidae by Allio et al. (2021a), including seven mitochondrial and nuclear genes, Troidini were better represented than other clades such as Leptocircini, with a sampling of *ca.* 62% of all known species, and all genera subgenera or species group represented.

In this study, we conducted a phylogenomic analysis of Troidini following the same bioinformatic framework as applied to Leptocircini (**Chapter 2.1**). We first established a list of valid Troidini species based on previous molecular and morphological studies and then inferred a robust and nearly complete species-level phylogeny for the tribe. We sequenced and assembled new reference genomes of five genera of Troidini, which are added to the reference genomes of *Ornithoptera* and *Troides* from Reboud et al. (2023, Chapter 1), and *Byasa* from He et al. (2022). We also sequenced and assembled *de novo* shotgun whole genomes for 127 species and reassembled the whole genome of four species from Allio et al. (2020a). Altogether this approach allows extracting thousands of nuclear orthologous genes and mitogenomes for around 97% of the total specific diversity of Troidini. We aim to provide a reference phylogenetic framework for assessing the monophyly of genera and subgenera as well as fine-scale species relationships within Troidini. For this objective, our study includes for the first time all enigmatic and long-debated groups and species to be robustly placed in the swallowtail and birdwing tree of life, such as the rare *Parides ascanius*, *P. klagesi*, *P. vercingetorix*, *Ornithoptera allotei*, *Byasa crassipes*, *B. daemonius*, and other species that will be key for testing diversification and biogeographic hypotheses (e.g. *Pharmacophagus antenor*, *Cressida cressida*).

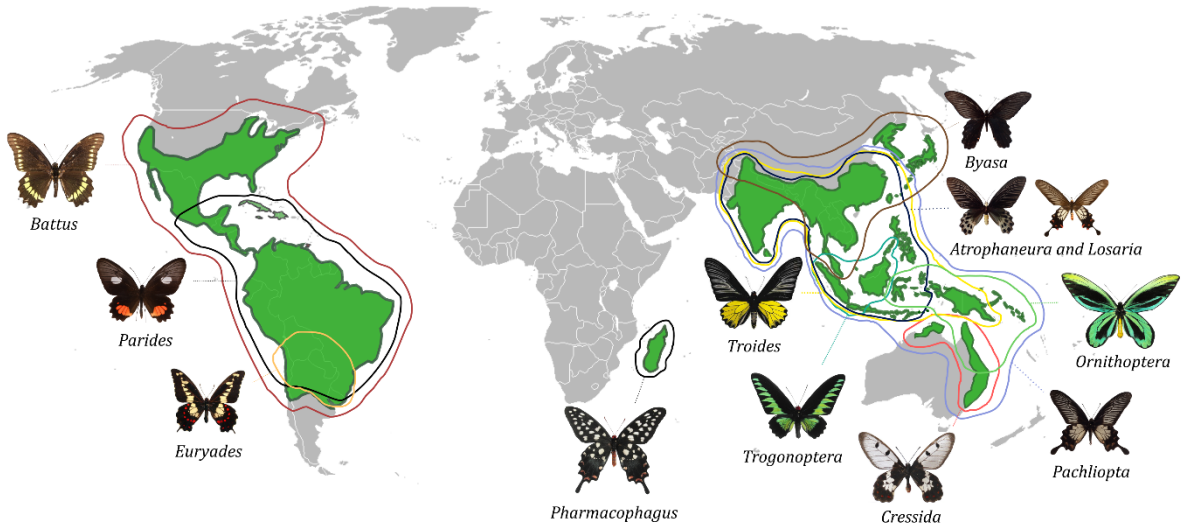


Figure 1: Overview of Troidini and their global distribution, with their habitus. The distribution map of each genus is drawn approximately according to literature and iNaturalist occurrences (www.inaturalist.org, GBIF). Note the complete absence of extant Troidini in continental Africa. Butterfly illustrations: *Battus polydamas* (Cosmin O. Manci); *Parides anchises* (CC BY-NC 3.0, Didier Descouens); *Cressida cressida* (CC BY-NC 4.0, JKDw); *L. coon, Euryades duponchelli, Pharmacophagus antenor, Trogonoptera brookiana, Troides aeacus, Ornithoptera alexandrae* (Fabien L. Condamine); *Atrophaneura sycorax, Pachliopta pandiyana, Byasa alcinous* (Eliette L. Reboud.)

Materials and Methods

The 'Materials and methods' of this Chapter is almost identical as the one described for Leptocircini in Chapter 2.1 (Material and Methods section and Figure 2), although it has been adapted to the overall smaller genome size of this tribe.

Species list and taxon sampling

We first established a taxonomic working list of Troidini species (Table S1). This species list takes into account some recent taxonomic list (Nakae 2021) and the results of recent studies (Silva-Brandao et al. 2008, Condamine et al. 2012a, 2015, Xu et al. 2020, von der Linden et al. *under review* = **Appendix III**, and this study).

Of the 139 species currently recognized in the new taxonomic list presented here, we sampled 135 species representing 97% of the total diversity (Table S1). However, the total number of species might evolve with further systematic studies, as the validity of some species is still debated (e.g. *B. polydamas*, *T. cuneifera sumatrana* and *T. miranda neomiranda*; **Appendix III**). We therefore included nine specimens of species of currently debated status (see **Table S1, S2**). Furthermore, five alive Troidine species were used to produce reference genomes for different genera using a combination of long reads and short reads data for the

following species: *Battus polydamas*, *Pachliopta aristolochiae*, *Parides eurimedes*, *P. vertumnus*, and *Pharmacophagus antenor*. As for Leptocircini, we added 11 outgroup species representing several swallowtail genera to root the Troidini tree based on large-scale phylogenetic studies of Papilionidae (Condamine et al. 2012, Allio et al. 2020a, 2021a): *Baronia brevicornis*, *Papilio xuthus*, *Parnassius apollo*, *Iphilides podalirius*, *Teinopalpus imperialis*, *Meandrusa payeni*, *Bhutanitis thaidina*, *Sericinus montela*, *Luehdorfia chinensis*, *Hypermnestra helios* and *Archon apollinus*. The complete taxon sampling and accession numbers are provided in **Table S2**.

DNA extractions and library preparations for whole genome sequencing

The genome size of Troidini was estimated between 250-450 Mb (*Losaria* - *Battus* ; Allio et al. 2020a). The extraction of most specimens was done on museum and collections specimens (taking legs, rarely abdomens). The extraction and library preparation were performed with the DNAeasy Blood and Tissue kit (QIAGEN), applying the methods of Meyer & Kirscher (2010), but with slight modifications as described in Tilak et al. (2015). DNA quality and concentration and length fragment varied markedly among specimens. Final DNA purity and concentrations were measured using both Nanodrop (Thermo Fisher, USA) and Qubit (Thermo Fisher, USA) and resulting libraries were analyzed for size distribution by Agilent 2200 TapeStation. Illumina 150 bp paired-end sequencing was run on a NovaSeq 6000 instrument to obtain a genome depth-of-coverage of about 50x.

For reference genomes, DNA extractions were done with butterflies killed and stored in a freezer at -20°C without any additional preservation product, issued from rearing. Tissues from the thorax were used to extract high-molecular weight DNA. Following Reboud et al. (2023) who tested two different extraction methods, we used the Qiagen genomic DNA kit to obtain a better 260/230 ratio as estimated with Nanodrop assays guaranteeing DNA purity for long-read sequencing with Oxford Nanopore Technology (ONT). Final DNA purity and concentrations were measured using both Nanodrop (Thermo Fisher, USA) and Qubit (Thermo Fisher, USA). Whole-genome libraries were prepared using the resulting high-molecular-weight DNA as input for the Nanopore LSK-109 ligation kit (Oxford Nanopore Technologies, UK) following the manufacturer's protocol. Long-read (LR) sequencing was performed on a GridION device with two to six R9.4.1 flow cells. Remaining DNA extractions were sent to Novogene Europe (Cambridge, UK) for two short-read Illumina library preparations. Libraries were generated using NEBNext DNA Library Prep Kit following manufacturer's recommendations and indices were added to each library. Genomic DNA was randomly fragmented to a size of 350 bp by shearing, then DNA fragments were end-polished, A-tailed, and ligated with the NEBNext adapter for Illumina sequencing, and further PCR enriched by P5 and

indexed P7 oligos. The PCR products were purified (AMPure XP system) and the resulting libraries were analyzed for size distribution by Agilent 2100 BioAnalyzer and quantified using real-time PCR. Illumina 150 bp paired-end sequencing was run on a NovaSeq 6000 instrument to obtain a genome depth-of-coverage of ~100x after combining the two libraries. The sequencing protocol for long and short reads of the reference genomes is illustrated in **Chapter 2.1**, **Figure 2**.

Assembling reference and shotgun genomes

For each reference genome (*Battus polydamas*, *Pachliopta aristolochiae*, *Parides eurimedes*, *P. vertumnus*, *Pharmacophagus antenor*), raw LR sequence data (fast5 files) were basecalled using Guppy 5.0.15 (Oxford Nanopore Technology) with the super-high accuracy mode and a quality control of 10 (min_score 10). Sequencing adapters were trimmed using Porechop 0.2.3 (https://github.com/rrwick/Porechop).

Draft genome assemblies were performed with the LR assembler Flye 2.8.3 (Kolmogorov et al. 2019) with default options. For *Parides vertumnus*, duplicated haplotigs and heterozygous contig overlaps were removed from the draft assembly using purge_dups 1.2.5 (Guan et al. 2020) based on the LR depth. The Illumina raw reads were cleaned, filtered, and paired using fastp 19.5-21.0 (Chen et al. 2018) with default options. To improve base accuracy and reduce assembly errors, the Flye draft assemblies were polished using short-reads (SR) with POLCA (Zimin and Salzberg 2020) implemented in MaSuRCA 4.0.1 (Zimin et al. 2013). Assembly statistics were then assessed using the gVolante2 platform (Nishimura et al. 2017) to retrieve the number and size of contigs, the presence, completeness, and duplication of BUSCO genes of the Lepidoptera *odb10* database (Manni et al. 2021). We checked for possible contaminations using BlobTools 1.1.1 (Laetsch and Blaxter 2017) set to the *ncbi* and *diamond* databases. We found no evidence of artificial contamination coming from laboratory manipulation, but some contigs were clearly identified as belonging to exogenous organisms such as symbionts. We removed all contigs that belonged to plants or symbionts (Wolbachia, Ascomyota, Acetobacteraceae, Bacteroidota, and Microsporidia).

For shotgun genomes, the Illumina SR were filtered and paired with fastp using the default options. The assemblies were performed by MEGAHIT 1.2.7-1.2.9 (Li et al. 2015) and submitted to the gVolante platform to obtain assembly statistics as well as assessing the number of genes found from the *odb10* database of Lepidoptera (BUSCO 5, Manni et al. 2021). The gVolante platform retrieves BUSCO genes classed in “single copy”, “multicopy” and “fragmented” categories. To improve the contiguity and the number of BUSCO genes recovered in these draft assemblies, they were scaffolded with the scaffold option of RagTag (Alonge et al. 2021) on a close reference genome. For this scaffolding step, three reference genomes were added to the five *de novo* reference genomes from previous study (*O.*

alexandrae, *T. oblongomaculatus*, Reboud et al. 2023; *Byasa hedistus*, He et al. 2022). Almost all shotgun genomes had a close reference and could therefore be scaffolded with Ragtag (except five species: *Cressida cressida*, *Trogonoptera spp.*, *Euryades spp.*, that had no close references). Because of their phylogenetic proximity, *Atrophaneura* species were scaffolded on *Byasa hedistus* reference genomes, and *Losaria* species of *Pachliopta aristolochiae* genomes. Additionally, five species were assembled and scaffolded using reads from a previous study (*Byasa alcinous*, *Euryades corethrus*, *Ornithoptera richmondia*, *Parides photinus*, *Troides plateni*, Allio et al. 2020a; see **Table S2** for SRA accession numbers). To compare draft genome assemblies with scaffolded genome assemblies, the resulting assemblies were again submitted to the gVolante platform. The BUSCO single copy genes in the fasta nucleotide file of each individual were retrieved using BUSCOMP 0.13 software (Edwards 2019).

Assembling mitogenomes and phylogenetic reconstruction

For reference genomes, we used the short reads and GetOrganelle (Jin et al. 2020) and a close reference for mapping to reconstruct complete mitogenomes (*Pachliopta aristolochiae* NC034280 for *P. aristolochiae*, *Pharmacophagus antenor* LS975119 for *P. antenor*, and *Parides photinus* LS974638 for *P. vertumnus* and *P. eurimedes*). For *Battus polydamas*, we used the mitogenome of *Battus polydamas polydamas* FC780 for the mitogenome phylogenetic reconstruction, as *Battus polydamas* is suspected to be a species complex and we were not sure about the subspecies of our specimen FC1386 of *Battus polydamas*. The resulting assemblies were given to MitoFinder 1.4 (Allio et al. 2020b) to annotate and extract thirteen protein-coding genes (ATP6, ATP8, COX1-3, CYTB, ND1-4, ND4L, ND5-6) and 2 non-coding genes (rrnL, rrnS).

For shotgun genomes, Megahit draft-assemblies of each species were given to MitoFinder (Allio et al. 2020b) to extract and annotate their mitogenome. Some assemblies only displayed fragmented mitogenomes, pseudogenes, fungi contamination or cross-contamination. When the identification of "real" mitogenomic contigs were not possible manually, GetOrganelle was used with a close reference for mapping to reconstruct complete, or less fragmented, mitogenomes.

We selected 11 outgroup species as previously described (see *Species list and taxon sampling* section), either using the mitogenome of the same individual as for the nuclear analyses (*Baronia brevicornis*, *Iphiclides podalirius*, *Archon apollinus*, *Hypermnestra helios*; see **Table S2**) or by taking the complete mitogenome of the same species on GenBank (*Bhutanitis thaidina*, *Parnassius apollo*, *Teinopalpus imperialis*, *Meandrusa payeni*, *Papilio xuthus*, *Sericinus montela*, *Luehdorfia chinensis*; see **Table S2**). Sequences were aligned with MACSE 2.07 (Ranwez et al. 2018), setting an invertebrate mitochondrial genetic code (*-gc_def 5*), a frame-shift penalty of 5 (*-fs 5*) and an internal stop codon penalized by 10 (*-stop*

10). Maximum-likelihood (ML) phylogenetic inference was implemented with IQ-TREE 2.2.0 (Minh et al., 2020a) using ModelFinder to select the best-fit partition scheme and the best-fitting substitution model for each partition (*-m MFP+MERGE* option, Chernomor et al. 2016, Kalyaanamoorthy et al. 2017) applied to an initial subset of 41 possible partitions (the three codon-position of the 13 protein-coding genes and the 2 non-coding genes). Node robustness of the ML tree was evaluated with 1,000 ultrafast bootstraps (UFBS; *-B 1000* option, Hoang et al. 2018).

Phylogenomic analyses with nuclear data

Phylogenetic analyses of genome-scale data were based on BUSCO genes. For each of the 5,286 lepidopteran core genes, we assessed the number of represented individuals, and selected all genes that gather at least 149 out of 154 individuals (96.8% of sampling for each gene), which corresponds to 1,663 genes. The nucleotide sequences of the 1,663 genes were aligned with MAFFT 7.453 (Katoh & Standley 2013). For each alignment, a gene tree was reconstructed with IQ-TREE with the best-fitting substitution model for each gene (*-m TESTNEW* option) and 1,000 UFBS. A consensus tree of these gene trees was produced with ASTRAL 5.7.7 (Zhang et al. 2018) following a tree-reconciliation or supertree approach.

A supermatrix approach was also performed with the nucleotide alignment of the 1,663 genes. Maximum-likelihood phylogenetic inference was carried out with IQ-TREE using ModelFinder to select the best-fit partition scheme and the best-fitting substitution model for each partition (*-m MFP+MERGE* option), on an initial subset of 1,663 possible partitions (one partition per gene). Node robustness was evaluated with 1,000 ultrafast bootstraps (UFBS; *-B 1000* option). To verify a potential attraction of *Byasa crassipes* by *B. daemonius*, a second topology search was conducted without *Byasa daemonius* (best partition and models retrieved from the previous analysis).

Finally, we performed gene concordance factor (gCF) and site concordance factor (sCF) analysis implemented in IQ-TREE 2, using the 1,663 gene trees and the nucleotide supermatrix in comparison with the consensus tree inferred with the supermatrix with IQ-TREE analysis.

Results and Discussion

Reference genomes

We assembled five *de novo* reference genomes of Troidini species representing the genera *Battus* (*B. polydamas*), *Pharmacophagus* (*P. antenor*), *Pachliopta* (*P. aristolochiae*), and *Parides* (*P. vertumnus*, *P. eurimedes*) using long and short reads data (**Table 1**). This study reports the first long-read genomes for the genera *Pachliopta*, *Pharmacophagus*, and *Parides*.

Final depth of coverage estimates of long-reads (Nanopore) assemblies ranged from 32 to 73× coverage, and short-reads polishing (Illumina) of these genomes from ~21 to 172× coverage. Overall, all assemblies obtained high BUSCO scores (between 96.6% and 98.3% single complete BUSCO genes, and low duplication rate: max 1.9%, *Parides eurimedes*). The N50 ranged from 0.2 to 6.0 Mb. The sizes of the assembled genomes varied from 303 Mb (*Pachliopta aristolochiae*, *Parides vertumnus*) to 478 Mb (*B. polydamas*). The latter seems to confirm the largest size of genome from the genus *Battus* compared to other Troidini (Allio et al. 2020a, Chaturvedi et al. 2023), as the genome size of *Battus polydamas* was very similar to the recently released genome of *Battus philenor* (443 Mb, Chaturvedi et al. 2023).

Table 1. Assembly statistics for the five *de novo* reference Troidini genomes. LR = long reads, SR = short reads.

	Battus polydamas	*Pachliopta aristolochiae*	*Parides eurimides*	*Parides vertumnus*	*Pharmacophagus antenor*
	FC1386	**FC539**	**FC1383**	**FC1330**	**FC1333**
Raw data sequenced (Gb) (LR + SR)	27.3 + 43.7	24.0 + 31.2	23.4 + 37.0	36.4 + 52.1	22.9 + 57.8
Final mean coverage (LR + SR)	32x + 21x	58x + 96x	57x + 119x	73x + 146x	47x + 172x
Assembly size (bp)	478,088,683	303,225,749	303,458,263	334,702,165	359,868,922
Number of contigs	7,205	1,677	4,183	6,092	650
N50 (bp)	1,763,096	1,161,693	389,348	234,896	6,004,580
Max length (bp)	15,764,897	4,634,664	3,099,472	9,214,910	11,347,106
Nucleotid assembly BUSCO score (%)	S:97.9; D:0.7; F:0.4; M:0.1	S:98.2; D:0.3; F:0.4; M:1.1	S:96.6; D:1.9; F:0.4; M:1.1	S:97.2; D:1.0; F:0.5; M:1.3	S:98.3; D:0.4; F:0.2; M:1.1

Short-reads whole-genomes assemblies

We assembled 131 *de novo* whole genomes using Illumina pair-end short-reads data for Troidini (**Table S2**). Assemblies with Megahit resulted in draft assemblies with an average of 81.3% for the single-copy complete genes in BUSCO. The mean of fragmented and missing genes was 8.8% and 8.7%, respectively. The genomes of Troidini were more homogeneous in terms of size than those of Leptocircini, with an average of 368 Mb, the genus with the longest genomes being *Battus* with on average 543 Mb, and the shortest *Atrophaneura* and *Losaria* with on average 295 and 299 Mb, respectively. After a scaffolding step with RagTag with a phylogenetically close reference genome when available, the average proportion of single-copy complete genes in BUSCO improved to 95.6% for all assemblies (**Figure 2**), and that of fragmented and missing genes decreased to 1.7% and 2.1%, respectively.

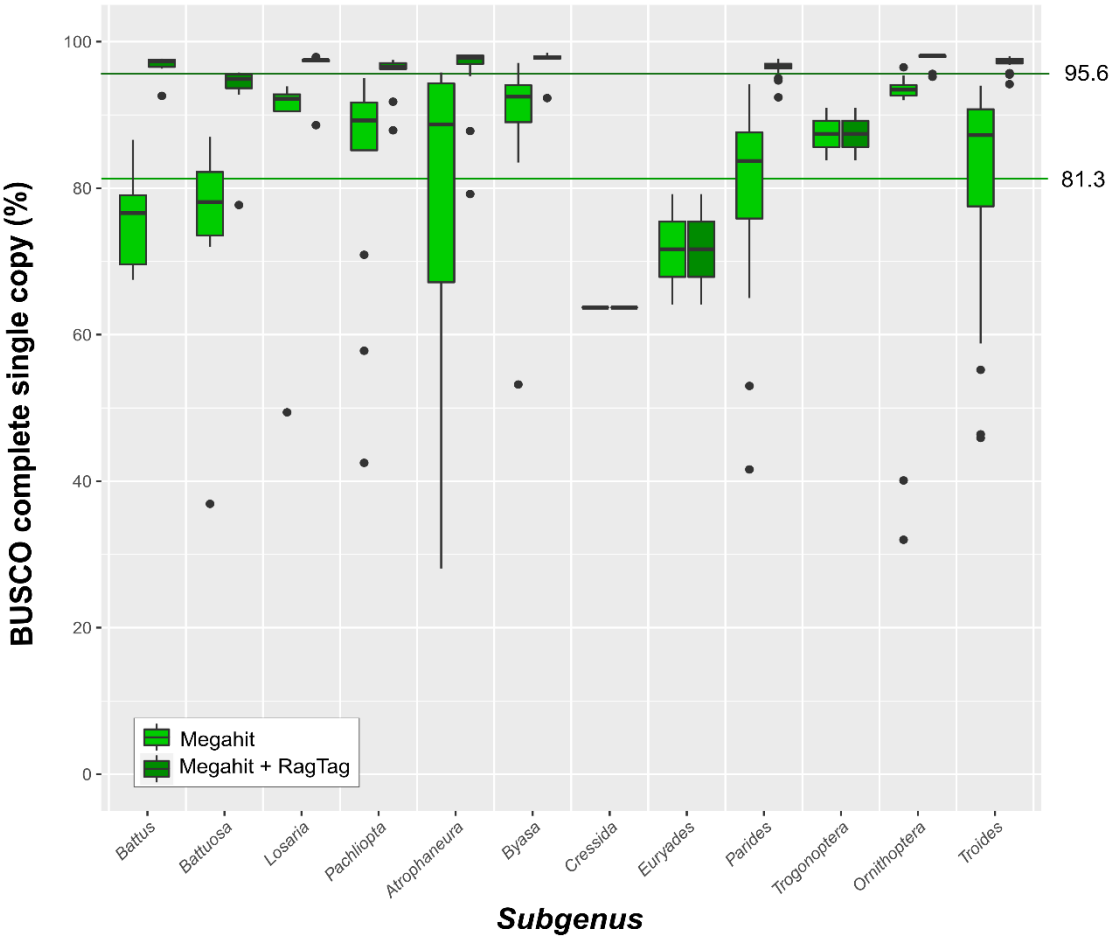


Figure 2: Effect of scaffolding fragmented shotgun whole genomes on the recovery of BUSCO genes. Boxplots of BUSCO single complete genes values of all assembly (*n* =131 specimens) within a subgenus, before (green) and after (dark green) the scaffolding with RagTag. Horizontal colored lines correspond to the mean BUSCO single complete score of all individuals before (green) and after (dark green) the scaffolding. Draft assemblies of species *Cressida cressida*, *Trogonoptera spp.* and *Euryades spp.* were not scaffolded.

The scaffolding step was overall more efficient than for Leptocircini (Figure 3). Some specimens had spectacular evolution of BUSCO score after the scaffolding step, e.g. Atrophaneura kuehni: 28.1% to 79.2%, A. varuna 33.4% to 87.8%, or Ornithoptera allotei: 32.0% to 95.6%. This efficiency is probably due to the overall better quality of the shotgun draft assemblies of Troidini (58.5% mean BUSCO score of Leptocircini draft assemblies instead of 81.7% for Troidini), but also probably because of the better quality of reference genomes and their close relatedness with each query.

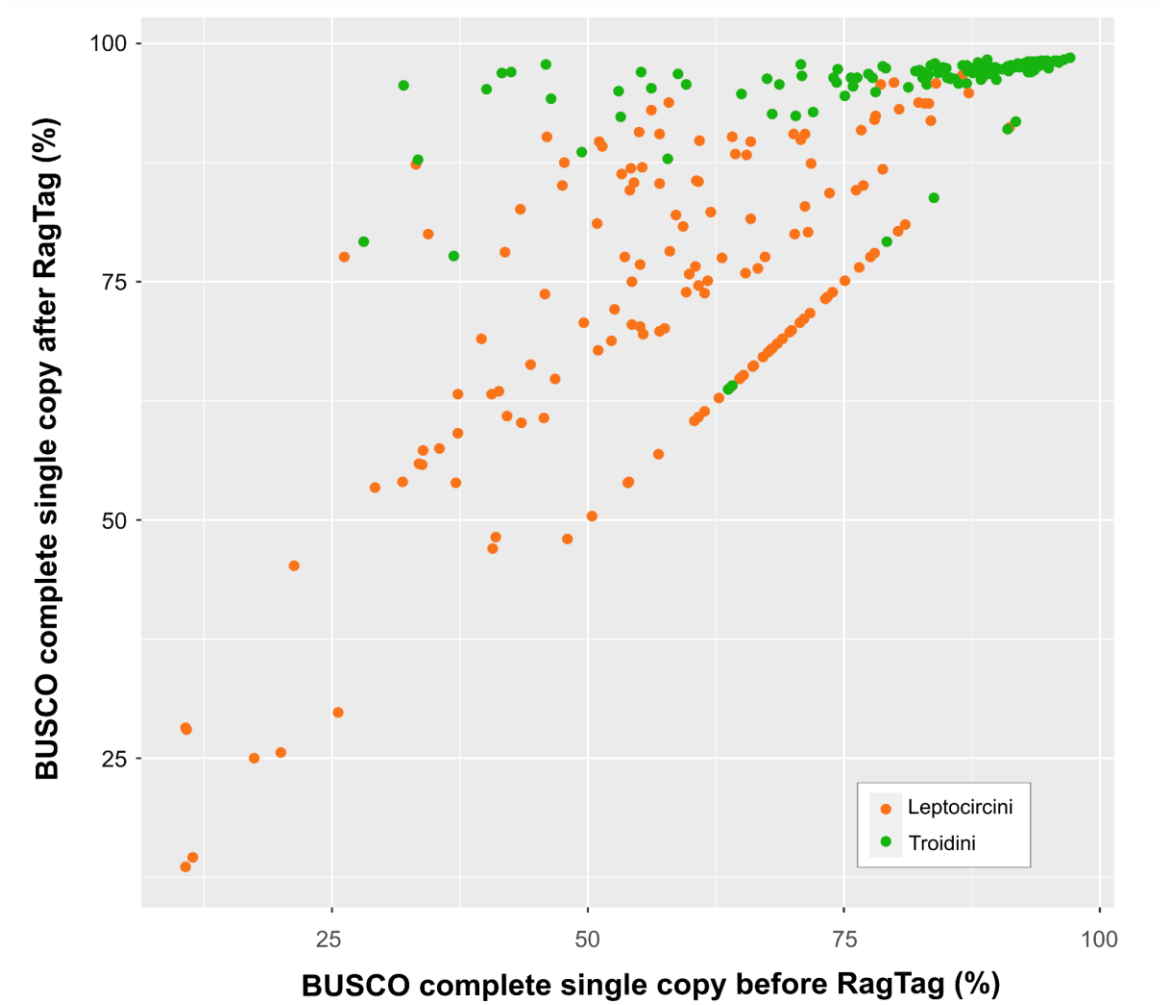


Figure 3: Comparison of BUSCO single copy gene score before and after the scaffolding step for Troidini (green) and Leptocircini (orange). Points on the first diagonal correspond to the specimens which were not scaffolded (because there was no close reference).

A species-level molecular phylogeny of Troidini

Our study provides the first robust and nearly complete phylogenomic framework for the tribe Troidini, thanks to the *de novo* assemblies of 139 individuals, including five reference genomes, and three rare species from which we were only able to extract the mitogenome (*Pachliopta pandiyana, Parides klagesi P. pizarro steinbachi*). Adding four reference genomes from previous studies (He et al. 2022, Reboud et al. 2023), and the COI-5P sequence of *Parides iphidamas,* our phylogenies represent 97% of the total species richness of Troidini.

The phylogenomic analyses with ASTRAL (tree-reconciliation or supertree approach) and IQ-TREE (supermatrix approach) provided almost identical topologies, with high mean and median node support (**Figure 4**, **Figure S1**). The mitochondrial analysis also presented a very similar topology, but with slight changes, mostly limited to the species group level (**Figure S2**, and see below). The number of final partitions in IQ-TREE analysis was reduced to 211 (compared to the initial 1663 possible partitions) for the nuclear analysis, and to 13 partitions (compared to the initial 41 partitions) for the mitochondrial analysis. All analyses recovered strong supports for the monophyly of all genera (UFBS/$UFBS_{mito}$/LPP = 100/100/1), except genus *Byasa* that was found paraphyletic (UFBS/$UFBS_{mito}$/gCF/sCF/LPP = 100/100/93.3/54.8/1, see below). The backbone of Troidini was the same as Allio et al. (2020a, 2021a), confirming *Battus* as the first genera to derive in the tribe Troidini (UFBS/gCF/sCF/LPP = 100/86.8/41.3/1), forming the sub-tribe Battina. The first genus to derive in the other sub-tribe, called Troidina, was *Pharmacophagus* (UFBS/gCF/sCF/LPP = 100/92.1/49.3/1) as proposed by Miller (1987). The genus *Cressida* was confirmed as sister to the Neotropical clade *Euryades* + *Parides* with strong support (UFBS/gCF/sCF/LPP = 100/86.6/43.5/1) as in Allio et al. (2020, 2021), and therefore bringing new independent evidence for its long debated and difficult placement (Ford 1944, Munroe 1960, Munroe and Ehrlich 1960, Condamine et al. 2012b, Condamine et al. 2018).

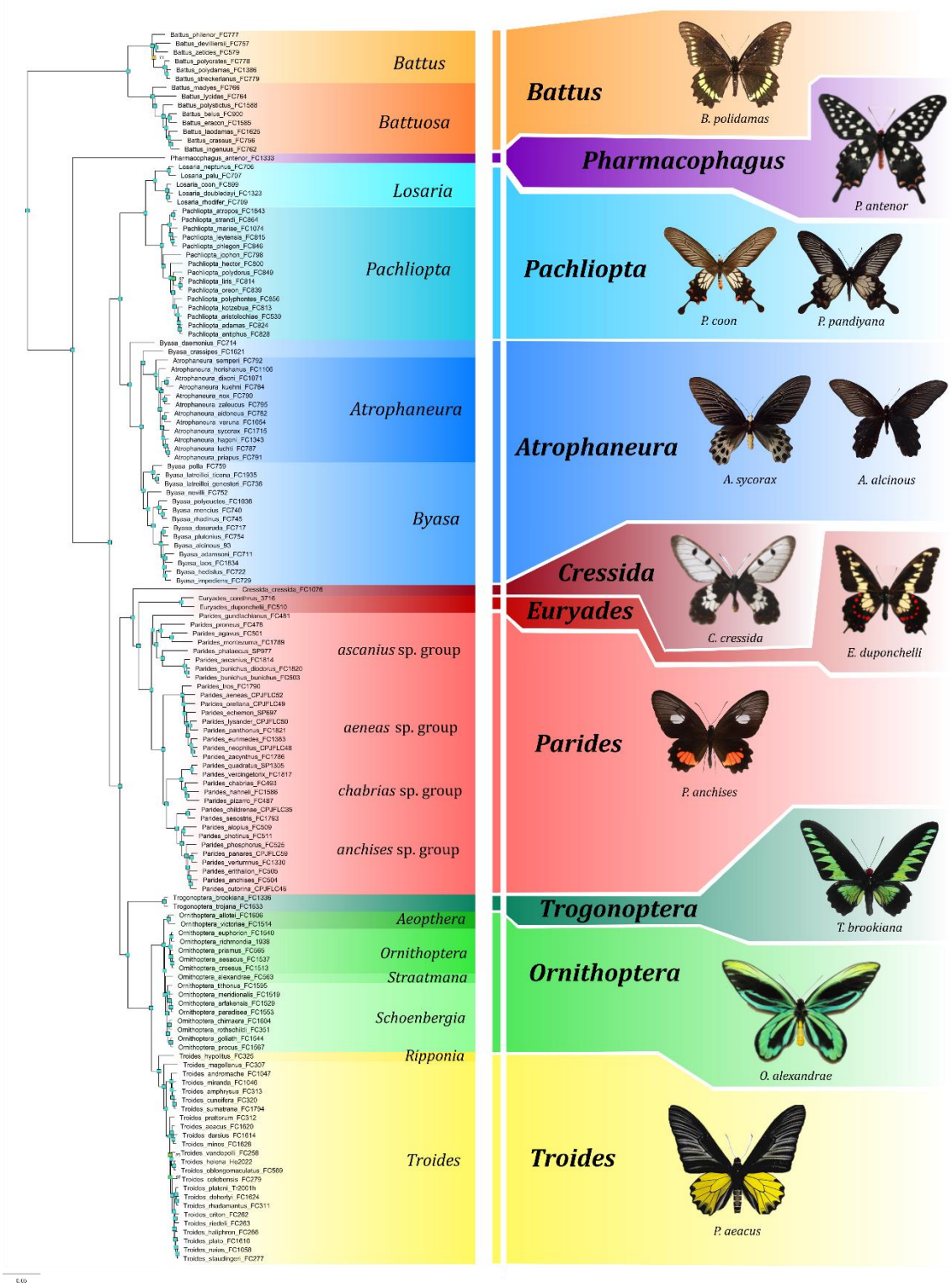


Figure 4: Phylogenomic relationships of Troidini. The phylogeny has been reconstructed with a supermatrix approach of 1,663 single-copy genes and a maximum-likelihood inference in IQ-TREE. Light blue squares at nodes correspond to a bootstrap value equal to 100. Values <100 are indicated. Genera are in italic bold font, and subgenera in regular italic. Illustration credit: *Atrophaneura sycorax*, *Byasa alcinous*, *Pachliopta pandiyana* (Eliette L. Reboud); *B. crassipes* (Robert Westphal); *B. hedistus* (CC BY-NC 4.0, yixianshuiesuan); *Parides anchises* (CC BY-NC 3.0, Didier Descouens); *B. polydamas* (Cosmin O. Manci); *Cressida cressida* (CC BY-NC 4.0, JKDw); others (Fabien L. Condamine).

The clade formed by the genera *Atrophaneura*, *Byasa*, *Pachliopta* and *Losaria*, was the least documented in terms of expected relationships. The rare pair of species *Pachliopta pandiyana* (India, only in mitochondrial analysis) and *P. jophon* (Sri Lanka) were found sister in the mitochondrial phylogenomic analysis ($UFBS_{mito}$ = 100, **Figure S2**). Surprisingly, *Byasa daemonius* was found sister to all *Byasa* and *Atrophaneura* with strong support (UFBS/$UFBS_{mito}$/gCF/sCF/LPP = 100/100/93.3/54.8/1) (**Figure 4, Figures S1, S2, S3**), making the genus *Byasa* (as currently circumscribed in Nakae 2021) paraphyletic. Moreover, *Byasa crassipes* was recovered sister to all *Atrophaneura* with nuclear data (UFBS/gCF/sCF/LPP = 100/42.9/37.5/1), and this position was maintained when *B. daemonius* was removed from the analysis (thus reducing the putative long branch attraction). However, *B. crassipes* was found sister to all *Byasa* in the mitochondrial phylogenomic analysis ($UFBS_{mito}$ = 100, **Figure 5**). Given the peculiar morphology of *B. crassipes*, its scarcity and the low gCF/sCF support of its placement, hypotheses of gene flow, incomplete lineage sorting or natural hybrids with *Atrophaneura* or *Byasa* species would be interesting to explore further. Regarding *B. daemonius*, while the ecology and life traits of this species are fairly well known thanks to the recent work of Chen et al. (2015) and Zhou et al. (2017), little has been documented about its phylogenetic proximity to other groups of species. Our study is therefore the first to document its early divergence from the remaining species of *Byasa* and *Atrophaneura*.

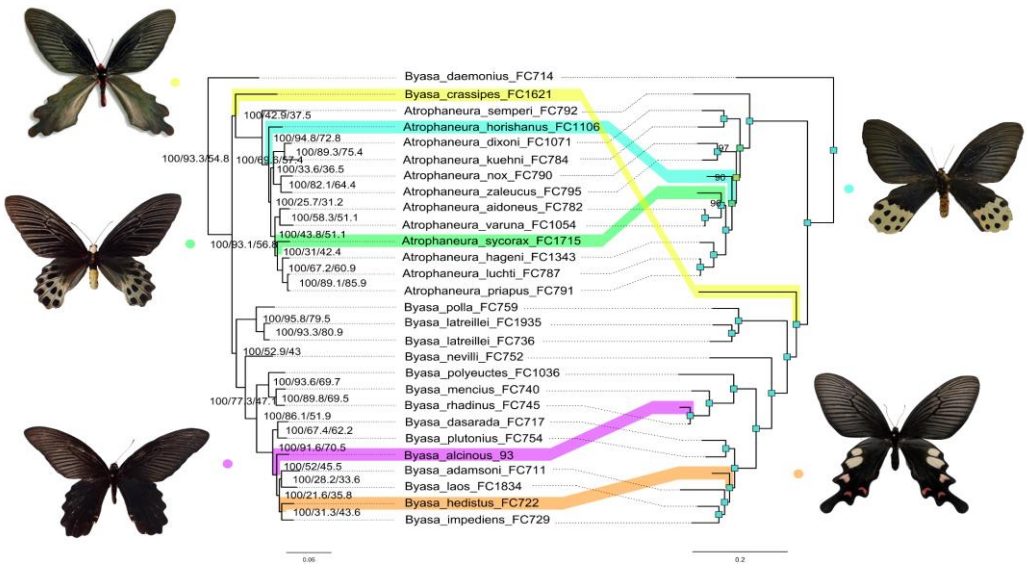


Figure 5: Topological conflict between the nuclear (left) and the mitochondrial (right) analyses for the clade *Atrophaneura/Byasa*. On the nuclear tree, node values are UFBS/gCF/sCF values. For the mitochondrial tree, nodes corresponding to $UFBS_{mito}$ = 100 are represented by blue squares, and values of $UFBS_{mito}$ <100 are indicated. Species involved in a topological conflict are illustrated and colored in the trees. Illustration: *Atrophaneura horishanus* (CC BY-NC 4.0 Tom Szpirer); *A. sycorax* and *Byasa alcinous* (Eliette L. Reboud); *B. crassipes* (Robert Westphal); *B. hedistus* (CC BY-NC 4.0, yixianshuiesuan).

The birdwing butterfly clade, formed by the genera *Trogonoptera*, *Ornithoptera* and *Troides,* was the only part of the tree where topology discordances were found between the ASTRAL and IQ-TREE inferences (**Figure 6**). *Ornithoptera allotei* was found sister to *O. victoriae* in the IQ-TREE analysis (UFBS/gCF/sCF = 100/64.3/60.4), but sister to the *priamus* clade in the ASTRAL analysis (LPP = 1). This is expected and interesting given the long recognized, albeit debated, hybrid nature of this species (Rousseau-Decelle 1939, Schmid 1970, McAlpine 1970, Straatman 1976, Deslisle and Sclavo 2015). Unfortunately, this hypothesis is not easy to verify with only one specimen of this very rare species and might require further work with Museum samples. Perhaps, more surprisingly, *O. alexandrae* was not found sister to *O. victoriae* as in Condamine et al. (2015) and Allio et al. (2021a), but rather sister to the subgenus *Schoenbergia* as in Morinaka et al. (2000). This placement was stable in all phylogenomic analyses (UFBS/$UFBS_{mito}$/gCF/sCF/LPP = 100/99/35.1/39.5/1) and is consistent with the existence of the monospecific subgenus name *Straatmana* introduced by Deslisle (2001, 2007). *Troides prattorum*, which is endemic to Buru Island only, was found sister of the *aeacus* species group in the IQ-TREE analysis, which is in accordance with Condamine et al. (2015). However, it was found sister of the clade formed by *T. vandepolli*, *T. helena* and *T. oblongomaculatus* in the ASTRAL analysis. In Allio et al. (2021a), *T. prattorum* and *T. vandepolli* were forming a clade sister to the clade *T. helena*/*oblongomaculatus*. Interestingly, the description of this species written by Haugum and Low (1982-1985) mentions that "*some characters of the* prattorum *female clearly point at* aeacus *as the closest relative and others are reminiscent of the* haliphron *group*". Given that this species is sympatric or peripatric to *T. oblongomaculatus bouruensis*, it would be interesting to investigate a potential ancient hybrid origin of this species.

A few other *Troides* species have a position worth being noted. This is the case of *T. darsius*, which was thought to be part of the *haliphron* species group (Zeuner 1943, Haugum and Low 1982-1985, Nakae 2021) and that we found sister of *T. minos*. Given that *T. darsius* is distributed in Sri Lanka, its suspected relationship with the *haliphron* group (Wallacea distribution) have long appeared puzzling to many authors (Zeuner 1943, Haugum and Low 1982-1985), who hypothesized local extinctions to explain this distribution. As *T. minos* is distributed in South India, this relationship between these two species seems more coherent in terms of biogeography, but might remain surprising for morphological characters. *Troides celebensis* is thought to be a natural hybrid of *T. helena hephaestus* x *T. haliphron* or *T. oblongomaculatus thestius* x *T. haliphron*. Our mitochondrial analysis places it close to *helena*/*oblongomaculatus* ($UFBS_{mito}$ = 98), but it is placed between the *helena* species group and *haliphron* species group in both nuclear analyses (UFBS/gCF/sCF/LPP = 100/36.2/45.3/1). This seems to support the hybrid status of this species.

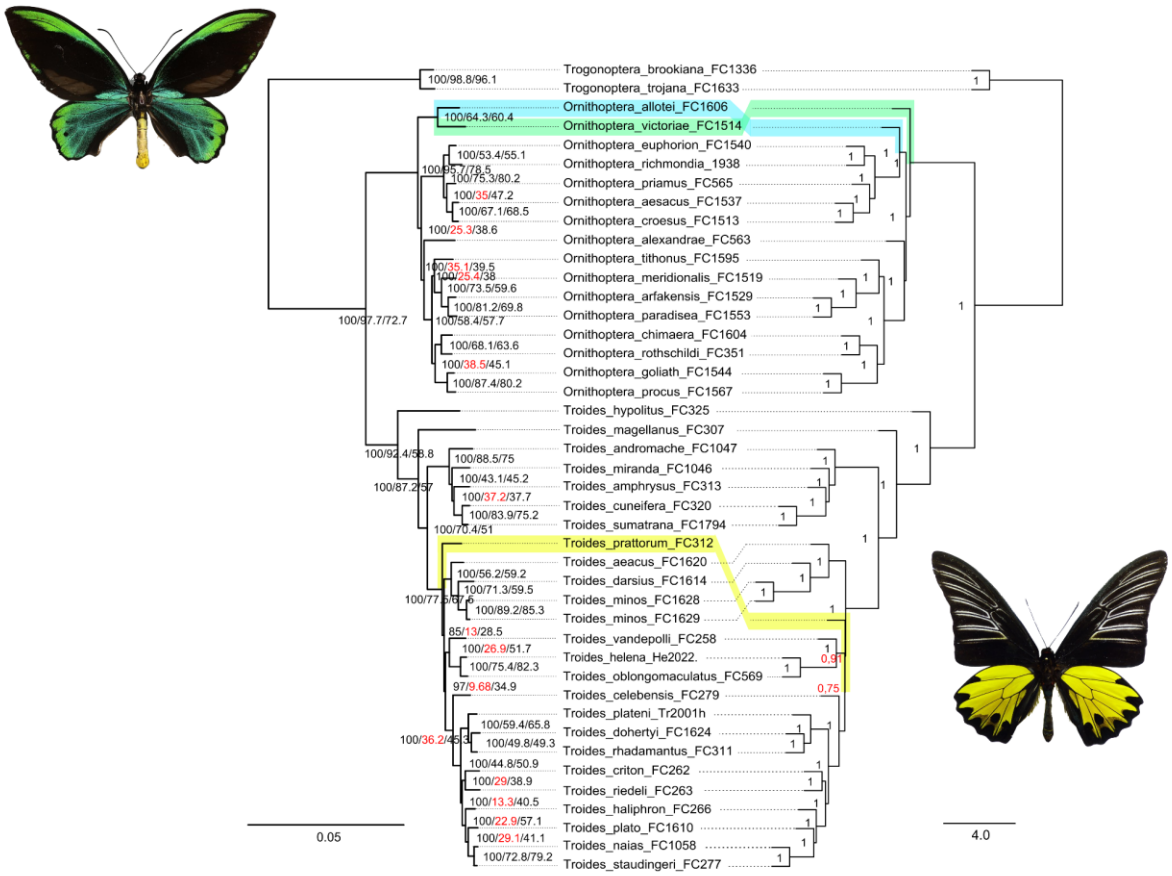


Figure 6: Discordant topologies between IQ-TREE (left) and ASTRAL (right) inferences in the birdwing butterflies. On the IQ-TREE tree, node values represent the support of UFBS/gCF/sCF values. Poor gCF values are colored in red. On the ASTRAL tree, poor LPP supports are colored in red. Illustration: male of *Ornithoptera allotei* (CC BY 4.0, Peter Wing); *Troides prattorum* (Jens Jakusch).

Placement of some rare species of Parides

The Neotropical genus *Parides* contains several rare or poorly known species. Using both nuclear and mitochondrial data, our study includes all the 33 recognized species of *Parides*, such as the rare *P. klagesi* and *P. ascanius* as well as every species of the *chabrias* species group (which was poorly represented in the last recent studies featuring *Parides*, Condamine et al. 2012a, Allio et al. 2021a). It also included some subspecies whose taxonomic level might need to be revised such as *P. bunichus diodorus* or *P. pizarro steinbachi*.

We could only assemble the mitogenome of the rare *Parides klagesi*, and found it sister to *P. montezuma* ($UFBS_{mito}$ = 100), which places it in the *ascanius* group (as circumscribed by Racheli and Olmisani 1998, Racheli 2006). The recovered branch was surprisingly long despite the apparent good quality and integrity of its mitogenome contig. This species is the smallest species of *Parides*, making it probably the smallest species of

Troidini. It was thought to be restricted to three areas in northern and eastern Venezuela (Caura, Suapure and Caroní rivers; Collins and Morris 1985, Tyler et al. 1994) and to extend into Guyana (Tyler et al. 1994). It was recently discovered in Santo Antônio do Tauá, within the state of Pará, Brazil (Mielke and Casagrande 2007) and is now listed as Least Concern in the IUCN RedList (Grice et al. 2019). Nakae (2021) did not consider this species as valid but rather as a dwarf form of *Parides anchises*, as suggested by Tyler et al. (1994) for one (but only one!) of his three displayed specimens (Tyler et al. 1994). Interestingly, this mitochondrial phylogeny places *P. klagesi* (Venezuela, Guyana, N Brazil) in a clade formed by *P. proneus* (SE Brazil), *P. agavus* (Paraguay, N. Argentina, SE Brazil) and *P. montezuma* (S Mexico to Costa Rica), thus displaying an interesting geographical connection/continuum between these four taxa.

Parides ascanius is a rare species restricted to the lowland coastal swamps of Rio de Janeiro (Brazil), and classified as Vulnerable (VU) in the IUCN Red List (Grice et al. 2019). It was found sister to *P. bunichus bunichus* and *P. b. diodorus* with high support ($UFBS/UFBS_{mito}$/gCF/sCF/LPP = 100/100/99.3/91.6/1) as suspected by Tyler et al. (1994) based on morphological characters.

Finally, the rare *P. vercingetorix*, endemic to French Guiana and possibly Suriname, was found sister to *P. quadratus* with high support ($UFBS/UFBS_{mito}$/gCF/sCF/LPP = 100/100/97.3/83.8/1). As suggested by Tyler et al. (1994) and Racheli (2006), these two species are included in the *chabrias* group, a clade formed by three other uncommon species: the mimetic *P. hahneli* and *P. chabrias* as well as *P. pizarro* (whose subspecies *P. p. steinbachi* is rare and suspected to be a separate species endemic to Bolivia).

Global revision of Troidini

The following sections include a proposition of taxonomic revision that will need to be reviewed and completed by expert taxonomists such as Adam M. Cotton for its form and consistency with the literature before being fully acted.

a. Higher-level systematic

The taxonomy of *Atrophaneura*, *Byasa*, *Pachliopta* and *Losaria* has been complicated, leading to a difficult higher-level systematics (Racheli and Cotton 2010, Nakae 2021). Our study suggests that a taxonomic revision of *Atrophaneura* and *Byasa* is needed, as *Byasa crassipes* and *B. daemonius* are not found within the genus *Byasa* as currently delimited (Nakae 2021). Based on our data and analyses, we propose to separate clade formed by *Atrophaneura*, *Byasa*, *Pachliopta* and *Losaria* into two genera, *Atrophaneura* (type species: *Atrophaneura semperi* (C. & R. Felder, 1861)) and *Pachliopta* (type species: *Pachliopta aristolochiae* (Fabricius, 1775)). Genus *Pachliopta* would then include two subgenera with *Pachliopta* and *Losaria* (type species: *Losaria coon* (Fabricius, 1793)). Genus *Atrophaneura* would be composed of subgenus

Atrophaneura, subgenus *Byasa* (type species: *Byasa polyeuctes* (Doubleday, 1842)), and a new subgenus for *A. daemonius*. Finally, after in-depth investigation, *Byasa crassipes* (Oberthür, 1893) could be transferred into the subgenus *Atrophaneura*, but doubts remain about this point because the mitochondrial tree shows a topological discordance with the nuclear tree.

The species groups as described in Racheli and Olmisani (1998) and Racheli (2006) were very well supported (see Table 1 in Racheli 2006) except the *aeneas* and *lysander* species groups, which would need to be merged into a single *aeneas* species group that is monophyletic. To date, *Parides* have no subgenus classification, which may seem rather surprising given the species richness of this group. Although Racheli (2006) did not talk about subgenus, he pointed out that a few names are available for denominating the different species groups. It may be worthwhile to consider proposing a subdivision of this Neotropical genus, whose phylogeny now appears to be fairly well resolved.

b. Species groups and possible taxonomic inflation

In this study, we have performed an analysis of concordance factors, not only to highlight putative cases of incomplete lineage sorting and gene flow in deep nodes of the phylogeny, but also to detect potential ongoing cases of gene flow within recent clades. The median and mean values of CF for the whole tree were respectively equal to 71.11 and 66.97 for the gCF and 51.84 and 55.29 for the sCF. The last value of the gCF index corresponds to the proportion of trees that are not either the main topology or the two alternatives of the quartet *i.e.* paraphyly of the quartet and polytomy topologies. For the whole tree, gDFP value was equal to 11.95 (median) and 19.32 (mean). Several clades presented lower CF values compared to these average scores, namely within some *Parides, Ornithoptera* and *Troides* species groups.

Within *Parides,* two groups presented rather low CF values and high paraphyletic or unresolved relationship in the gene trees (high gDFP): the *aeneas* species group (8 nodes, gCF: mean = 53.16, median = 50.73; gDFP: mean = 26.93, median = 30.34, **Figure S4**), and the *anchises* species group (6 nodes, gCF: mean = 49.37, median = 46.19; gDFP: mean = 33.36, median = 38.63, **Figure S4**). These two groups are both pan-Amazonian and display the ‘classical’ red mark on their hindwing and pale-blue pattern on their forwings (they might mimic one another, Puissant et al. 2023). There was no particular concern about taxonomic inflation in the two groups, but studies about sympatric and parapatric speciation and gene flow would be very interesting to conduct in the Amazon Basin.

The genera *Troides* and *Ornithoptera* form two rapid radiations in the Indo-Australasian Archipelago, which both displayed relatively low CF supports. These discordances may be due to ILS, as rapid radiation is a known source of this type of conflict (Suh et al. 2015, Meleshko et al. 2021, DeRaad et al. 2023). The insular habit of these two groups and the great interest they have generated undoubtedly

makes their taxonomy difficult and will require further work to ensure that some cryptic species are not ignored (e.g. *Troides amphrysus* group: **Appendix III**, *Ornithoptera victoriae*, *O. goliath*, *O. priamus*) or that taxonomic inflation is not present in these groups. For instance, within *Ornithoptera*, the subgenus *Ornithoptera* (4 nodes, mean = 57.72, median = 60.28, gDFP: mean = 26.18, median = 25.13, **Figure S4**) and the group *Schoenbergia* + *Straatmana* (8 nodes, mean = 58.45, median = 63.26, gDFP: mean = 28.11, median = 28.28, **Figure S4**) might be interesting to investigate both in terms of species delimitation, and origin of tree discordance.

On the other hand, within *Troides*, the apparent low genetic divergence and the very low CF support, associated with high gDFP of the subgenus *Troides* excluding *magellanus* and the *amphrysus* species group (15 nodes: mean = 40.96 median = 36.18, gDFP: mean = 45.72, median = 45.94, **Figure S4**) raised serious concern about taxonomic inflation of this group. First, *Troides dohertyi* did not seem to deserve the species status, given the genetic divergence in their nuclear and mitochondrial data, and we therefore revise the status of this taxa as a subspecies of *T. rhadamantus*: *T. r. dohertyi* ***stat. rev.*** (as suggested before, e.g. Haugum and Low 1982-1985). Our study also emphasizes a need for revision within the species *T. haliphron*. Indeed, Condamine et al. (2015) recovered *T. haliphron* paraphyletic because the subspecies *T. h. naias* was found sister of *T. staudingeri*. We independently replicated this result with genome data and therefore revise the position of the subspecies *T. haliphron naias* into the species *T. staudingeri*: *T. s. naias* ***comb. nov.*** Overall, the divergences within the *helena*, *haliphron* and *aeacus* species group, altogether with the low gCF support might indicate a need for revision in this group.

Finally, two pairs of species in the *Atrophaneura* and *Pachliopta* groups need to be discussed. First, the species *Atrophaneura priapus* and *A. luchti* are two sympatric or peripatric species of Java (Indonesia), which look very similar but present slight differences in coloration. *Atrophaneura luchti* is listed as endangered by the IUCN Red List and has a very restricted range in the Ijen Mountains of Eastern Java (medium to high altitude, Collins and Morris 1985, Rushbrooke and Moonen 2019). It is also said to be bred by local people on the nearby Mount Argopuro (Nakae 2021), but this point is not mentioned in the IUCN Red List report of Rushbrooke and Moonen (2019). On the contrary, *Atrophaneura priapus* occurs in all Java (Collins and Morris 1985, Hancock 1988, Rushbrook and Moonen 2019). In this study, our sampling included *A. priapus priapus* (West Java, Mont Wayana, Indonesia) and *A. luchti* (East Java, Mount Argopuro). Our pair of species, despite presenting distinct phenotypes (*A. luchti* have red hair on the body, whereas A. priapus is yellow), displayed little genetic divergence in the nuclear and mitochondrial data (**Figures 5**, **Figure S2**), which could either mean that our specimen of *A. luchti*, that would have been reared, is not exactly representative of the the Ijen

population, that the genome of this species have been introgressed by *A. priapus* in a secondary contact, or that *A. luchti* does not deserve the species status. A further investigation with Museum specimens of *A. luchti* that were collected in the Ijen Mountains might help to determine if a subgenus or local form status would be more appropriate for this species. The second doubtful pair in our study was *Pachliopta liris* and *P. oreon* that are both distributed in the Lesser Sunda Islands (East Indonesia): in Timor for *P. liris* and in Sumba for *P. oreon*. In our study, this pair was represented by *P. liris pallidus* (Moa, Indonesia) and *Pachliopta oreon batuataensis* (Batuatas, Wacuala), and they displayed little divergence in the nuclear and mitochondrial data (**Figure 4**, **Figure S2**). This could either mean that the subspecies collected were misattributed to the species they are included in (like *T. staudingeri naias*, which was previously attributed to *T. haliphron*), or that a subspecies status for *Pachliopta oreon* might be more appropriate than a species status. A better sampling of each subspecies that compose these two species would be needed for further taxonomic revision.

To conclude, several species pairs or species groups may need taxonomic revision and further study with intraspecific taxon sampling of specimens for several key species. Galtier (2019) proposed an approach aimed at identifying pairs of species that are well recognized by the community in the group of species studied, and calculating for each pair indices of population polymorphism structure, indices of genetic differentiation, and analyses of population structure. These analyses provide points of comparison and may help in species delimitation. This method has been used in the case of mammals (Allio et al. 2021b, Barthe et al. *under review*) and could prove useful in the genus *Troides*, for example.

Acknowledgments

We would like to thank Thomas Wanchet from *La Ferme aux Papillons* for sending us fresh *Battus* and *Parides* individuals from his breeding activity. We thank Anne-Laure Clamens for some DNA extractions of *Parides*. This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (project GAIA, agreement no. 851188).

References

Allio R, Schomaker-Bastos A, Romiguier J, Prosdocimi F, Nabholz B, Delsuc F. 2020b. MitoFinder: efficient automated large-scale extraction of mitogenomic data in target enrichment phylogenomics. *Mol. Ecol. Res*. **20**:892–905.

Allio R, Scornavacca C, Nabholz B, Clamens AL, Sperling FA, Condamine FL. 2020a. Whole genome shotgun phylogenomics resolves the pattern and timing of swallowtail butterfly evolution. *Syst. Biol*. **69**:38-60.

Allio R, Tilak MK, Scornavacca C, Avenant NL, Kitchener AC, Corre E, *et al.* 2021b. High-quality carnivoran genomes from roadkill samples enable comparative species delineation in aardwolf and bat-eared fox. *eLife* **10**:e63167.

Allio R., Nabholz B, Wanke S, Chomicki G, Pérez-Escobar OA, Cotton AM, Clamens, AL, Kergoat GJ, Sperling FAH, Condamine FL. 2021a. Genome-wide macroevolutionary signatures of key innovations in butterflies colonizing new host plants. *Nat. Commun.* **12**:354.

Alonge M, Lebeigle L, Kirsche M, Jenike K, Ou S, Aganezov S, Wang X, Lippman ZB, Schatz MC, Soyk S. 2022. Automated assembly scaffolding using RagTag elevates a new tomato system for high-throughput genome editing. *Genome Biol*. **23**:1-19.

Barthe M, Rancilhac L, Arteaga MC, Feijó A, Tilak MK, Justy F, Loughry WJ, McDonough CM, de Thoisy B, Catzeflis, F. *et al*. under review. Exon capture museomics deciphers the continent-wide nine-banded armadillo species complex and identifies a new species endemic to the Guiana Shield. Submitted to *Syst. Biol.*

Braby MF, Trueman JW, Eastwood R. 2005. When and where did troidine butterflies (Lepidoptera: Papilionidae) evolve? Phylogenetic and biogeographic evidence suggests an origin in remnant Gondwana in the Late Cretaceous. *Invertebrate Syst.* **19**:113-143.

Bryk 1929. Papilionidae I-III. In E. Strand (ed.), Lepidopterorum catalogus, 35, 37, and 39.

Chaturvedi S, Escalona M, Marimuthu MP, Nguyen O, Chumchim N, Fairbairn CW, *et al.* 2023. A draft reference genome assembly of the pipevine swallowtail butterfly, *Battus philenor hirsuta*. *J. Heredity* https://doi.org/10.1093/jhered/esad043.

Chen G, Luo S, Mei N, Shen D, Sun W. 2015. Case study of building of conservation coalitions to conserve ecological interactions. *Cons. Biol.* **29**:1527-1536.

Chen S, Zhou Y, Chen Y, Gu J. 2018. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**:i884-i890.

Chernomor O, Von Haeseler A, Minh BQ. 2016. Terrace aware data structure for phylogenomic inference from supermatrices. *Syst. Biol*. **65**:997-1008.

Chou I. 1994. *Monographia Rhopalocerorum Sinensium*. Henan Scientific and Technological Publishing House, Zhengzhou, 2 vol., 854 pp.

Collins NM, Morris MG. 1985. *Threatened Swallowtail Butterflies of the World: the IUCN Red Data Book*. IUCN.

Condamine FL, Nabholz B, Clamens AL, Dupuis JR, Sperling FAH. 2018. Mitochondrial phylogenomics, the origin of swallowtail butterflies, and the impact of the number of clocks in Bayesian molecular dating. *Syst. Entomol.* **43**:460-480.

Condamine FL, Silva-Brandão KL, Kergoat GJ, Sperling FAH. 2012a. Biogeographic and diversification patterns of Neotropical Troidini butterflies (Papilionidae) support a museum model of diversity dynamics for Amazonia. *BMC Evol. Biol.* **12**:82.

Condamine FL, Sperling FAH, Kergoat GJ. 2013a. Global biogeographic pattern in swallowtail diversification demonstrates alternate colonization routes in Northern and Southern Hemispheres. *J. Biogeogr*. **40**: 9–23.

Condamine FL, Sperling FAH, Wahlberg N, Rasplus J-Y, Kergoat GJ. 2012b. What caused the latitudinal gradient of species diversity in swallowtail butterflies? *Ecol. Lett*. **15**:267–277.

Condamine FL, Toussaint EFA, Clamens AL, Genson G, Sperling FAH, Kergoat GJ. 2015. Deciphering the evolution of birdwing butterflies 150 years after Alfred Russel Wallace. *Sci. Rep.* **5**:11860.

Debelle FD, Vanherweghem JL, Nortier JL. 2008. Aristolochic acid nephropathy: a worldwide problem. *Kidney International* **74**:158-169.

DeRaad DA, McCullough JM, DeCicco LH, Hime PM, Joseph L, Andersen MJ, Moyle RG. 2023. Mitonuclear discordance results from incomplete lineage sorting, with no detectable evidence for gene flow, in a rapid radiation of *Todiramphus* kingfishers. *Mol. Ecol.* **32**:4844-4862.

Deslisle G, Sclavo JP. 2015. *Outstanding Birdwing Butterflies (Papilionidae: Genus Ornithoptera), 2 vol.* Nice, Ornitho Press, 1642 pp.

Deslisle G. 2001 *Zeunera*, a new subgenus in genus *Ornithoptera* Boisduval 1832 for species *alexandrae* Rothschild 1907. *Lambillionea* **101**:288-321.

Deslisle G. 2007. *Straatmania*: A replacement name for the subgenus *Zeunera* Deslisle 2001 (Lepidoptera: Papilionidae). *Lambillionea* **107**:238.

Edwards RJ. 2019. BUSCOMP: BUSCO compilation and comparison–Assessing completeness in multiple genome assemblies. *F1000Research* **8**:995.

Ehrlich PR. 1958. The comparative morphology, phylogeny and higher classification of the butterflies (Lepidoptera: Papilionoidea). *Kans. Univ. Sci. Bull.* **39**:305-370.

Ford EB. 1941. Studies on the chemistry of pigments in the Lepidoptera, with reference to their bearing on systematics. 1. The anthoxanthins. *Proc. R. Entomol. Soc. Lond. Series A, General Entomol.* **16**:65-90.

Ford EB. 1944. Studies on the chemistry of pigments in the Lepidoptera, with reference to their bearing on systematics. 3. The red pigments of the Papilionidae. *Proc. R. Entomol. Soc. Lond. Series A, General Entomol.* **19**:92-106.

Galtier N. 2019. Delineating species in the speciation continuum: A proposal. *Evol. Appli.* **12**:657-663.

Gamboa S, Condamine FL, Cantalapiedra JL, Varela S, Pelegrín JS, Menéndez I, *et al.* 2022. A phylogenetic study to assess the link between biome specialization and diversification in swallowtail butterflies. *Glob. Change Biol.* **28**:5901-5913.

Grice H, Freitas AVL, Rosa A, Dias FMS, Mega N, Marini-Filho O, Mielke O, Casagrande M. 2019. *Parides klagesi* (amended version of 2018 assessment). The IUCN Red List of Threatened Species: http://dx.doi.org/10.2305/IUCN.UK.2019-1.RLTS.T16243A145166241.en

Guan D, McCarthy SA, Wood J, Howe K, Wang Y, Durbin R. 2020. Identifying and removing haplotypic duplication in primary genome assemblies. *Bioinformatics* **36**:2896-2898.

Haase E. 1893. *Untersuchungen über die Mimicry auf Grundlage eines natürlichen Systems der Papilioniden* (No. 8). E. Nägele.

Hancock DL. 1978. *Phylogeny and biogeography of Papilionidae (Lepidoptera)*. University of Queensland.

Hancock DL. 1980. The status of the genera *'Atrophaneura'* Reakirt and *'Pachliopta'* Reakirt (Lepidoptera: Papilionidae). *Austral. Entomol*. **7**: 27-32.

Hancock DL. 1983. Classification of the Papilionidae (Lep.): a phylogenetic approach. *Smithersia* **2**:1-48.

Hancock DL. 1988. A revised classification of the genus' Atrophaneura'Reakirt (Lepidoptera: Papilionidae). *Austral. Entomol.* **15**:7-16.

Hancock DL. 2007. Phylogeny of the Troidine butterflies (Lepidoptera: Papilionidae) revisited: Are the red-bodied swallowtails monophyletic? *Austral. Entomol.* **34**:33-42.

Haugum J, Low AM. 1982-1985. *A monograph of the birdwing butterflies. The systematics of Ornithoptera, Troides and related genera*. Vol 2, Trogonoptera, Troides. Scandinavian Science Press, Klampenborg, 308 pp.

He JW, Zhang R, Yang J, Chang Z, Zhu LX, Lu SH, Xie F-A, Mao J-L, Dong Z-W, Li G-C. 2022. High-quality reference genomes of swallowtail butterflies provide insights into their coloration evolution. *Zool. Res*. **43**:367-379.

Hoang DT, Chernomor O, Von Haeseler A, Minh BQ, Vinh LS. 2018. UFBoot2: improving the ultrafast bootstrap approximation. *Mol. Biol. Evol.* **35**:518-522.

Horsfield T, Moore F. 1858. *A Catalogue of the Lepidopterous Insects in the Museum of Natural History at the East-India House*. London, W.H. Allen and Co.

Igarashi S. 1979. *Papilionidae and Their Early Stages, vols.1 and 2.* Kodansha, Tokyo [in Japanese].

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

Jin JJ, Yu WB, Yang JB, Song Y, DePamphilis CW, Yi TS, Li DZ. 2020. GetOrganelle: a fast and versatile toolkit for accurate de novo assembly of organelle genomes. *Genome Biol.* **21**:1-31.

Kalyaanamoorthy S, Minh BQ, Wong TK, Von Haeseler A, Jermiin LS. 2017. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat. Meth.* **14**:587-589.

Katoh K, Standley DM. 2013. MAFFT Multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**:772–780.

Klitzke CF. 1992. Ecologia química e coevolução na interface Troidini (Papilionidae)/Aristolochia (Aristolochiaceae). Master Thesis, Instituto de Biologia, Universidade Estadual de Campinas, Campinas.

Kolmogorov M, Yuan J, Lin Y, Pevzner PA. 2019. Assembly of long, error-prone reads using repeat graphs. *Nat. Biotech.* **37**:540-546.

Laetsch DR, Blaxter ML. 2017. BlobTools: Interrogation of genome assemblies. *F1000Research* **6**:1287.

Leitão GG, Kaplan MAC. 1992. Química do gênero *Aristolochia*. *Revista Brasileira de Farmácia* **73**:65–75.

Li D, Liu CM, Luo R, Sadakane K, Lam TW. 2015. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics* **31**:1674-1676.

Manni M, Berkeley MR, Seppey M, Simão FA, Zdobnov EM. 2021. BUSCO update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. *Mol. Biol. Evol.* **38**:4647-4654.

McAlpine DK. 1970. A note on the status of *Ornithoptera allotei* Rothschild (Lepidoptera, Papilionidae). *J. Aust. Entomol. Soc.* **9**:233-234.

Meleshko O, Martin MD, Korneliussen TS, Schröck C, Lamkowski P, Schmutz J, Healey A, Piatkowski BT, Shaw AJ, Weston DJ, *et al*. 2021. Extensive genome-wide phylogenetic discordance is due to incomplete lineage sorting and not ongoing introgression in a rapidly radiated bryophyte genus. *Mol. Biol. Evol*. **38**:2750-2766.

Mielke OH, Casagrande MM. 2007. *Parides klagesi* (Ehrmann) rediscovered in Brazil (Lepidoptera, Papilionidae, Troidini). *Revista Brasileira Entomologia* **51**:187-189.

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

Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, Von Haeseler A, Lanfear R. 2020. IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. *Mol. Biol. Evol.* **37**:1530-1534.

Mitchell DK, Dewhurst CF, Tennent WJ, Page WW. 2016. Queen Alexandra's birdwing butterfly *Ornithoptera alexandrae* (Rothschild, 1907): A review and conservation proposals. Kuala Lumpur: Southdene Sdn Bhd.

Morinaka S, Maeyama T, Maekawa K. 1999. Molecular phylogeny of birdwing butterflies based on the representatives in most genera of the tribe Troidini (Lepidoptera: Papilionidae). *Entomol. Sci*. **2**:347-358.

Morinaka S, Minaka N, Sekiguchi M, Erniwati D, Prijono SN, Ginarsa IK, *et al.* 2000. Molecular phylogeny of birdwing butterflies of the tribe Troidini (Lepidoptera: Papilionidae) – Using all species of the genus *Ornithoptera*. *Biogeography* **2**: 103-111.

Munroe E, Ehrlich PR. 1960. Harmonization of concepts of higher classification of the Papilionidae. *J. Lepid. Soc.* **14**:169-175.

Munroe E. 1960. The classification of the Papilionidae (Lepidoptera). *Memoirs Entomol. Soc. Canada* **92**:5-51.

Nakae M. 2021. *Papilionidae of the World*. Roppon-Ashi Entomological Books.

Nishimura O, Hara Y, Kuraku S. 2017. gVolante for standardizing completeness assessment of genome and transcriptome assemblies. *Bioinformatics* **33**:3635-3637.

Racheli T, Cotton AM. 2010. *Guide to the Butterflies of the Palearctic Region: Papilionidae part II, Subfamily Papilioninae, Tribe Troidini*. Omnes Artes, Milano.

Racheli T, Olmisani L. 1998. A cladistic analysis of the genus *Parides* Hübner, [1819], based on androconial structures. *Neue Entomologische Nachrichten*, 119-130.

Racheli T. 2006. *The Genus Parides*. Butterfly of the World, Supplement 13. Bauer and Frankenbach / Goeke and Evers, Keltern.

Ranwez V, Douzery EJP, Cambon C, Chantret N, Delsuc F. 2018. MACSE v2: toolkit for the alignment of coding sequences accounting for frameshifts and stop codons. *Mol. Biol. Evol.* **35**:2582-2584.

Reboud EL, Nabholz B, Chevalier E, Tilak MK, Bito D, Condamine FL. 2023. Genomics, population divergence, and historical demography of the world's Largest and endangered butterfly, the Queen Alexandra's birdwing. *Genome Biol. Evol.* **15**:evad040.

Rothschild W, Jordan K. 1906. A revision of the American Papilios. *Novitates Zool*. **13**:412-752.

Rousseau-Decelle G. 1939. Note sur un nouvel exemplaire de *Troides Allottei* Rothsch. (Hybride d'*Ornithoptera Victoria regis* Rothsch et d'*Ornithoptera Urvilleanus* Guér.)[Lep. Papilionidae]. *Bull. Soc. Entomol. Fr.* 44:97-101.

Rushbrooke M, Moonen J. 2019. *Atrophaneura luchti*. The IUCN Red List of Threatened Species, https://dx.doi.org/10.2305/IUCN.UK.2019-3.RLTS.T2377A122599164.en Accessed on 21 September 2023.

Schmid F. 1970. Considérations sur le mâle d'*Ornithoptera allotei* Rothschild et sur la phylogénie des Ornithoptères. *J. Lepidopter. Soc.* **24**:88-105.

Silva-Brandao KL, Azeredo-Espin AML, Freitas AVL. 2008. New evidence on the systematic and phylogenetic position of *Parides burchellanus* (Lepidoptera: Papilionidae). *Mol. Ecol. Res.* **8**:502-511.

Silva-Brandao KL, Freitas AVL, Brower AV, Solferini VN. 2005. Phylogenetic relationships of the New World Troidini swallowtails (Lepidoptera: Papilionidae) based on COI, COII, and EF-1α genes. *Mol. Phylogenet. Evol.* **36**:468-483.

Silva-Brandao KL, Solferini VN. 2007. Use of host plants by Troidini butterflies (Papilionidae, Papilioninae): constraints on host shift. *Biol. J. Linn. Soc.* **90**:247-261.

Simonsen TJ, Zakharov EV, Djernaes M, Cotton AM, Vane-Wright RI, Sperling FAH. 2011. Phylogenetics and divergence times of Papilioninae (Lepidoptera) with special reference to the enigmatic genera *Teinopalpus* and *Meandrusa*. *Cladistics* **27**:113-137.

Straatman R. 1976. Hybridisation of birdwing butterflies (Lepidoptera: Papilionidae) in Papua New Guinea. *Lepidoptera Science* **27**:156-162.

Suh A, Smeds L, Ellegren H. 2015. The dynamics of incomplete lineage sorting across the ancient adaptive radiation of neoavian birds. *PLoS Biol*. **13**:e1002224.

Talbot G. 1939. *The Fauna of British India, including Ceylon and Burma*. Butterflies, Vol. 1. London, Taylor and Francis.

Talbot G. 1949. *Fauna of British India. Butterflies, vol. 1*. London, Taylor and Francis.

Tilak MK, Justy F, Debiais-Thibaud M, Botero-Castro F, Delsuc F, Douzery EJP. 2015. A cost-effective straightforward protocol for shotgun Illumina libraries designed to assemble complete mitogenomes from non-model species. *Cons. Genet. Res.* **7**:37-40.

Tsukada E, Nishiyama Y. 1982. *Butterflies of the South East Asian Islands*. Plapac Co. Ltd., Tokyo.

Tyler HA, Brown Jr. KS, Wilson KH. 1994. *Swallowtail Butterflies of the Americas. A study in Biological Dynamics, Ecological Diversity, Biosystematics, and Conservation.* Gainesville, Scientific Publishers. 376 pp.

Wu CS, Bai JW. 2001. A review of the genus *Byasa* Moore in China (Lepidoptera: Papilionidae). *Oriental Insects* **35**:67-82.

Xu ZB, Wang YY, Condamine FL, Cotton AM, Hu SJ. 2020. Are the yellow and red marked club-tail Losaria coon the same species? *Insects* **11**:392.

Zeuner FE. 1943. Studies in the systematics of *Troides* Hübner (Lepidoptera Papilionidae) and its allies; distribution and phylogeny in relation to the geological history of the Australasian Archipelago. *Trans. Zool. Soc. Lond.* **25**:107-184.

Zhang C, Rabiee M, Sayyari E, Mirarab S. 2018. ASTRAL-III: polynomial time species tree reconstruction from partially resolved gene trees. *BMC Bioinformatics* **19**:15-30.

Zhou S, Chen G, Xiao H. 2017. Description of a new species of Neochalcis (Hymenoptera: Chalcidoidea) reared from pupa of *Byasa daemonius* (Alpheraky) in Yunnan, China. *Entomotaxonomia* **39**:135-138.

Zimin AV, Marçais G, Puiu D, Roberts M, Salzberg SL, Yorke JA. 2013. The MaSuRCA genome assembler. *Bioinformatics* **29**:2669-2677.

Zimin AV, Salzberg S.L. 2020. The genome polishing tool POLCA makes fast and accurate corrections in genome assemblies. *PLoS Comput. Biol.* **16**:e1007981.

Supplementary information

Table S1: Taxonomic working list for Troidini species. The list includes subgenera that are currently recovered as monophyletic in molecular phylogenies, and species that belong to each subgenus. In the comment section, species in blue were sampled in our study and species in red were unsampled.

Genus *Atrophaneura* Reakirt, 1865	Sampling	Taxonomic status	Comment
Subgenus *Atrophaneura Reakirt, 1865*			
Atrophaneura (Atrophaneura) aidoneus Doubleday, 1845	Sampled	Valid	
Atrophaneura (Atrophaneura) crassipes Oberthür, 1879	Sampled	Valid	
Atrophaneura (Atrophaneura) dixoni Grose-Smith, 1901	Sampled	Valid	
Atrophaneura (Atrophaneura) hageni Rogenhofer, 1889	Sampled	Valid	
Atrophaneura (Atrophaneura) horishanus Matsumura, 1910	Sampled	Valid	
Atrophaneura (Atrophaneura) kuehni Honrath, 1887	Sampled	Valid	
Atrophaneura (Atrophaneura) luchti Roepke, 1935	Sampled	Doubtful	Given the nuclear and mitochondrial divergence found in this study, this could be a subspecies of *A. priapus*
Atrophaneura (Atrophaneura) matsudai Morita, 2003	Unsampled	Valid	
Atrophaneura (Atrophaneura) nox Swainson, 1822	Sampled	Valid	
Atrophaneura (Atrophaneura) priapus Boisduval, 1836	Sampled	Valid	
Atrophaneura (Atrophaneura) semperi Felder & Felder, 1861	Sampled	Valid	
Atrophaneura (Atrophaneura) sycorax Grose-Smith, 1885	Sampled	Valid	
Atrophaneura (Atrophaneura) varuna White, 1842	Sampled	Valid	
Atrophaneura (Atrophaneura) zaleucus Hewitson, 1865	Sampled	Valid	
Subgenus *Byasa* Moore, 1882			
Atrophaneura (Byasa) adamsoni Grose-Smith, 1886	Sampled	Valid	
Atrophaneura (Byasa) alcinous Klug, 1836	Sampled	Valid	
Atrophaneura (Byasa) confusus Rothschild, 1895	Sampled	Valid	
Atrophaneura (Byasa) crassipes Oberthür, 1879	Sampled	Valid	might be in the *Atrophaneura* subgenus
Atrophaneura (Byasa) dasarada Moore, 1857	Sampled	Valid	
Atrophaneura (Byasa) genestieri Oberthür, 1918	Sampled	Doubtful	Considered species *in litt.* (Hu et al. *in prep*), but might be a subspecies of *A. latreillei*
Atrophaneura (Byasa) hedistus Jordan, 1928	Sampled	Valid	
Atrophaneura (Byasa) impediens Seitz, 1907	Sampled	Valid	
Atrophaneura (Byasa) laos Riley & Godfrey, 1921	Sampled	Valid	
Atrophaneura (Byasa) latreillei Donovan, 1826	Sampled	Valid	

Atrophaneura (Byasa) mencius Felder & Felder, 1862	Sampled	Valid	
Atrophaneura (Byasa) nevilli Wood-Mason, 1882	Sampled	Valid	
Atrophaneura (Byasa) plutonius Oberthür, 1876	Sampled	Valid	
Atrophaneura (Byasa) polla de Niceville, 1897	Sampled	Valid	
Atrophaneura (Byasa) polyeuctes Doubleday, 1842	Sampled	Valid	
Atrophaneura (Byasa) rhadinus Jordan, 1928	Sampled	Valid	
Subgenus *subgen. nov.* 2023			
Atrophaneura (subgen. nov.) daemonius Alpheraky, 1895	Sampled	Valid	

Genus *Battus* Scopoli, 1777	**Sampling**	**Taxonomic status**	**Comment**
Subgenus *Battuosa* Möhn, 1999			
Battus (Battuosa) belus Cramer, 1777	Sampled	Valid	
Battus (Battuosa) crassus Cramer, 1777	Sampled	Valid	
Battus (Battuosa) eracon Godman & Salvin, 1897	Sampled	Valid	
Battus (Battuosa) ingenuus Dyar, 1907	Sampled	Valid	
Battus (Battuosa) laodamas Felder & Felder, 1859	Sampled	Valid	
Battus (Battuosa) lycidas Cramer, 1777	Sampled	Valid	
Battus (Battuosa) madyes Doubleday, 1846	Sampled	Valid	
Battus (Battuosa) polystictus Butler, 1874	Sampled	Valid	
Subgenus *Battus* Scopoli, 1777			
Battus (Battus) devilliersii Godart, 1823	Sampled	Valid	
Battus (Battus) philenor Linnaues, 1771	Sampled	Valid	
Battus (Battus) polydamas Linnaeus, 1758	Sampled	Valid	*B. p. psittacus*, *B. p. polycrates* and *B. p.streckerianus* could be species
Battus (Battus) zetides Munroe, 1971	Sampled	Valid	

Genus *Cressida* Swainson, 1832	**Sampling**	**Taxonomic status**	**Comment**
Cressida cressida Fabricius, 1775	Sampled	Valid	

Genus *Euryades* C. & R. Felder, 1864	**Sampling**	**Taxonomic status**	**Comment**
Euryades corethrus Boisduval, 1836	Sampled	Valid	
Euryades duponchelii Lucas, 1836	Sampled	Valid	

Genus *Ornithoptera* Boisduval, 1832	Sampling	Taxonomic status	Comment
Subgenus *Aetheoptera* Rippon, 1890			
Ornithoptera victoriae Gray, 1856	Sampled	Valid	*O. allotei* is considered as a natural hybrid of *O. priamus urvilleanus* x *O. victoriae regis*. Little is known about fertility and maitenance of the hybrids
Subgenus *Ornithoptera* Boisduval, 1832			
Ornithoptera aesacus Ney, 1903	Sampled	Valid	
Ornithoptera croesus Wallace, 1859	Sampled	Valid	
Ornithoptera euphorion Gray, 1853	Sampled	Valid	
Ornithoptera priamus Linnaeus, 1758	Sampled	Valid	
Ornithoptera richmondia Gray, 1853	Sampled	Valid	
Subgenus *Schoenbergia* Pagenstecher, 1892			
Ornithoptera arfakensis Joicey & Noakes, 1916	Sampled	Valid	
Ornithoptera chimaera Rothschild, 1904	Sampled	Valid	
Ornithoptera goliath Oberthür, 1888	Sampled	Valid	*O. g. procus* could be a species
Ornithoptera meridionalis Rothschild, 1897	Sampled	Valid	
Ornithoptera paradisea Staudinger, 1893	Sampled	Valid	
Ornithoptera rothschildi Kenrick, 1911	Sampled	Valid	
Ornithoptera tithonus de Haan, 1840	Sampled	Valid	
Subgenus *Straatmana* Deslisle, 2007			
Ornithoptera alexandrae Rothschild, 1907	Sampled	Valid	

Genus *Pachliopta* Reakirt, [1865]	Sampling	Taxonomic status	Comment
Subgenus *Losaria* Moore, 1902			
Pachliopta (Losaria) coon Fabricius, 1793	Sampled	Valid	
Pachliopta (Losaria) doubledayi Wallace, 1865	Sampled	Valid	
Pachliopta (Losaria) neptunus Guerin-Meneville, 1840	Sampled	Valid	
Pachliopta (Losaria) palu Martin, 1912	Sampled	Valid	
Pachliopta (Losaria) rhodifer Butler, 1876	Sampled	Valid	
Subgenus *Pachliopta* Reakirt, [1865]			
Pachliopta (Pachliopta) adamas Zincken, 1831	Sampled	Valid	
Pachliopta (Pachliopta) antiphus Fabricius, 1793	Sampled	Valid	

Pachliopta (Pachliopta) aristolochiae Fabricius, 1775	Sampled	Valid	
Pachliopta (Pachliopta) atropos Staudinger, 1888	Sampled	Valid	
Pachliopta (Pachliopta) hector Linnaeus, 1758	Sampled	Valid	
Pachliopta (Pachliopta) jophon Gray, 1853	Sampled	Valid	
Pachliopta (Pachliopta) kotzebuea Eschscholtz, 1821	Sampled	Valid	
Pachliopta (Pachliopta) leytensis Murayama, 1978	Sampled	Valid	
Pachliopta (Pachliopta) liris Godart, 1819	Sampled	Valid	
Pachliopta (Pachliopta) mariae Semper, 1874	Sampled	Valid	
Pachliopta (Pachliopta) oreon Doherty, 1891	Sampled	Doubtful	Given the nuclear and mitochondrial divergence found in this study, this could be a subspecies of *P. p. liris*
Pachliopta (Pachliopta) pandiyana Moore, 1881	Mito only	Valid	
Pachliopta (Pachliopta) phlegon Felder & Felder, 1864	Sampled	Valid	
Pachliopta (Pachliopta) polydorus Linnaeus, 1763	Sampled	Valid	
Pachliopta (Pachliopta) polyphontes Boisduval, 1836	Sampled	Valid	
Pachliopta (Pachliopta) schadenbergi Semper, 1886	Unsampled	Valid	
Pachliopta (Pachliopta) strandi Bryk, 1930	Sampled	Valid	

Genus *Parides* Hübner, [1819]	**Sampling**	**Taxonomic status**	**Comment**
Parides aeneas Linnaeus, 1758	Sampled	Valid	
Parides agavus Drury, 1782	Sampled	Valid	
Parides alopius Godman & Salvin, 1890	Sampled	Valid	
Parides anchises Linnaeus, 1758	Sampled	Valid	
Parides ascanius Cramer, 1775	Sampled	Valid	
Parides bunichus Hübner, 1821	Sampled	Valid	*P. b. almas* and *P. b. diodorus* could be separate species
Parides chabrias Hewitson, 1852	Sampled	Valid	
Parides childrenae Gray, 1832	Sampled	Valid	
Parides cutorina Staudinger, 1898	Sampled	Valid	
Parides echemon Hübner, 1813	Sampled	Valid	
Parides erithalion Boisduval, 1836	Sampled	Valid	
Parides eurimedes Stoll, 1782	Sampled	Valid	
Parides gundlachianus Felder & Felder, 1864	Sampled	Valid	
Parides hahneli Staudinger, 1882	Sampled	Valid	
Parides iphidamas Fabricius, 1793	COI-5P NCBI only	Valid	
Parides klagesi Ehrmann, 1904	Mito only	Valid	

Parides lysander Cramer, 1775	Sampled	Valid	
Parides montezuma Westwood, 1842	Sampled	Valid	
Parides neophilus Geyer, 1837	Sampled	Valid	
Parides orellana Hewitson, 1852	Sampled	Valid	
Parides panares Gray, 1853	Sampled	Valid	
Parides panthonus Cramer, 1780	Sampled	Valid	
Parides phalaecus Hewitson, 1869	Sampled	Valid	
Parides phosphorus Bates, 1861	Sampled	Valid	
Parides photinus Doubleday, 1844	Sampled	Valid	
Parides pizarro Staudinger, 1884	Sampled	Valid	*P. p. kuhlmanni* and *P. p. steinbachi* could be separate species
Parides proneus Hübner, 1831	Sampled	Valid	
Parides quadratus Staudinger, 1890	Sampled	Valid	
Parides sesostris Cramer, 1779	Sampled	Valid	
Parides tros Fabricius, 1793	Sampled	Valid	
Parides vercingetorix Oberthür, 1888	Sampled	Valid	Was first described as *Papilio coelus* Boisduval, 1836 (homonym *P. coelus* Cramer, 1782)
Parides vertumnus Cramer, 1779	Sampled	Valid	
Parides zacynthus Fabricius, 1793	Sampled	Valid	

Genus *Pharmacophagus* Haase, 1891	**Sampling**	**Taxonomic status**	**Comment**
Pharmacophagus antenor Drury, 1773	Sampled	Valid	

Genus *Trogonoptera* Rippon, [1890]	**Sampling**	**Taxonomic status**	**Comment**
Trogonoptera brookiana Wallace, 1855	Sampled	Valid	
Trogonoptera trojana Honrath, 1886	Sampled	Valid	

Genus *Troides* Hübner, [1819]	**Sampling**	**Taxonomic status**	**Comment**
Subgenus *Ripponia* Haugum, 1975			
Troides (Ripponia) hypolitus Cramer, 1775	Sampled	Valid	
Subgenus *Troides* Hübner, [1819]			
Troides (Troides) aeacus Felder & Felder, 1860	Sampled	Valid	
Troides (Troides) amphrysus Cramer, 1779	Sampled	Valid	

Troides (Troides) andromache Staudinger, 1892	Sampled	Valid	
Troides (Troides) criton Felder & Felder, 1860	Sampled	Valid	
Troides (Troides) cuneifera Oberthür, 1879	Sampled	Valid	*T. c. sumatrana* could be a separate species
Troides (Troides) darsius Gray, 1853	Sampled	Valid	
Troides (Troides) haliphron Boisduval, 1836	Sampled	Valid	
Troides (Troides) helena Linnaeus, 1758	Sampled	Valid	*T. celebensis* could be a natural hybrid of *T. helena* x *T. haliphron* but little is known about this point and *T. celebensis* could be a separate species.
Troides (Troides) magellanus Felder & Felder, 1862	Sampled	Valid	
Troides (Troides) minos Cramer, 1779	Sampled	Valid	
Troides (Troides) miranda Butler, 1869	Sampled	Valid	*T. m. neomiranda* could be a species
Troides (Troides) oblongomaculatus Goeze, 1779	Sampled	Valid	
Troides (Troides) plateni Staudinger, 1888	Sampled	Valid	
Troides (Troides) plato Wallace, 1865	Sampled	Valid	
Troides (Troides) prattorum Joicey & Talbot, 1922	Sampled	Valid	
Troides (Troides) rhadamantus Lucas, 1835	Sampled	Valid	*T. r. dohertyi* Rippon, 1903 ***stat. rev*** is ranked as a subspecies (and melanic form)
Troides (Troides) riedeli Kirsch, 1885	Sampled	Valid	
Troides (Troides) staudingeri Röber, 1888	Sampled	Valid	*T. s. naias* ***comb. nov.*** is placed as subspecies of *T. staudingeri* instead of *T. haliphron*
Troides (Troides) vandepolli Snellen, 1890	Sampled	Valid	

Table S2: Accession numbers of all Troidini specimens and outgroups. The third sub-column of the "Nuclear data" column shows the reference for the scaffolding step, when the assembly was scaffolded with RagTag. The third sub-column of the "Mitochondrial data" indicates whether the same individual was used for nuclear and mitochondrial data.

Informations			**Nuclear data**			**Mitochondrial data**		
Genus	***Subgenus***	***species***	**Accession nuclear**	**Reference**	**Scaffolding reference if any**	**Accession number mitogenome**	**Mitogenome completness**	**Same individual**
Baronia		*brevicornis*	GCA_031353405.1	Marino et al. *under review*		OR063968.1*	15	yes
Papilio		*xuthus*	GCF_000836235.1	Li et al. 2015		KU356933.1	15	no
Parnassius		*apollo*	GCA_907164705.1	Podsiadlowski et al. 2021		KF746065.1	15	no
Meandrusa		*payeni*	GCA_029286785.1	He et al. 2022		MN013000.1	15	no
Teinopalpus		*imperialis*	GCA_029286865.1	He et al. 2022		LS975139.1	15	no
Bhutanitis		*thaidina*	GCA_029286595.1	He et al. 2022		MN012973.1	15	no
Sericinus		*montela*	GCA_029286605.1	He et al. 2022		MN013020.1	15	no
Luehdorfia		*chinensis*	GCA_029286615.1	He et al. 2022		KU360130.1	15	no
Hypermnestra		*helios*	SRX5734328 (SRA)	Allio et al. 2020		LS975127.1	15	yes
Archon		*apollinus*	SRX5734362 (SRA)	Allio et al. 2020		LT999971.1	15	yes
Iphiclides		*podalirius*	FC536	Chapter 2.1		=	15	yes
Atrophaneura	*Atrophaneura*	*aidoneus*	FC782	This study	GCA_029286795.1	=	15	yes
Atrophaneura	*Atrophaneura*	*dixoni*	FC1071	This study	GCA_029286795.1	=	15	yes
Atrophaneura	*Atrophaneura*	*hageni*	FC1343	This study	GCA_029286795.1	=	15	yes
Atrophaneura	*Atrophaneura*	*horishanus*	FC1106	This study	GCA_029286795.1	=	15	yes
Atrophaneura	*Atrophaneura*	*kuehni*	FC784	This study	GCA_029286795.1	=	15	yes
Atrophaneura	*Atrophaneura*	*luchti*	FC787	This study	GCA_029286795.1	=	15	yes
Atrophaneura	*Atrophaneura*	*nox*	FC790	This study	GCA_029286795.1	=	15	yes
Atrophaneura	*Atrophaneura*	*priapus*	FC791	This study	GCA_029286795.1	=	15	yes
Atrophaneura	*Atrophaneura*	*semperi*	FC792	This study	GCA_029286795.1	=	15	yes
Atrophaneura	*Atrophaneura*	*sycorax*	FC1715	This study	GCA_029286795.1	=	15	yes

Atrophaneura	*Atrophaneura*	*varuna*	FC1054	This study	GCA_029286795.1	=	15	yes
Atrophaneura	*Atrophaneura*	*zaleucus*	FC795	This study	GCA_029286795.1	=	15	yes
Atrophaneura	*Byasa*	*adamsoni*	FC711	This study	GCA_029286795.1	=	15	yes
Atrophaneura	*Byasa*	*alcinous*	93	This study	GCA_029286795.1	=	15	yes
Atrophaneura	*Byasa*	*crassipes*	FC1621	This study	GCA_029286795.1	=	15	yes
Atrophaneura	*Byasa*	*dasarada*	FC717	This study	GCA_029286795.1	=	15	yes
Atrophaneura	*Byasa*	*genesteri*	FC736	This study	GCA_029286795.1	=	15	yes
Atrophaneura	*Byasa*	*hedistus*	FC722	This study	GCA_029286795.1	=	15	yes
Atrophaneura	*Byasa*	*impediens*	FC729	This study	GCA_029286795.1	=	15	yes
Atrophaneura	*Byasa*	*laos*	FC1834	This study	GCA_029286795.1	=	15	yes
Atrophaneura	*Byasa*	*latreillei*	FC1935	This study	GCA_029286795.1	=	15	yes
Atrophaneura	*Byasa*	*mencius*	FC740	This study	GCA_029286795.1	=	15	yes
Atrophaneura	*Byasa*	*nevilli*	FC752	This study	GCA_029286795.1	=	15	yes
Atrophaneura	*Byasa*	*plutonius*	FC754	This study	GCA_029286795.1	=	15	yes
Atrophaneura	*Byasa*	*polla*	FC759	This study	GCA_029286795.1	=	15	yes
Atrophaneura	*Byasa*	*polyeuctes*	FC1036	This study	GCA_029286795.1	=	15	yes
Atrophaneura	*Byasa*	*rhadinus*	FC745	This study	GCA_029286795.1	=	15	yes
Atrophaneura	*subgen. nov*	*daemonius*	FC714	This study	GCA_029286795.1	=	15	yes
Battus	*Battuosa*	*belus*	FC900	This study	FC1386	=	15	yes
Battus	*Battuosa*	*crassus*	FC756	This study	FC1386	=	15	yes
Battus	*Battuosa*	*eracon*	FC1585	This study	FC1386	=	15	yes
Battus	*Battuosa*	*ingenuus*	FC762	This study	FC1386	=	15	yes
Battus	*Battuosa*	*laodamas*	FC1625	This study	FC1386	=	15	yes
Battus	*Battuosa*	*lycidas*	FC764	This study	FC1386	=	15	yes
Battus	*Battuosa*	*madyes*	FC766	This study	FC1386	=	15	yes
Battus	*Battuosa*	*polystictus*	FC1588	This study	FC1386	=	15	yes
Battus	*Battus*	*devilliersii*	FC757	This study	FC1386	=	15	yes
Battus	*Battus*	*philenor*	FC777	This study	FC1386	=	15	yes
Battus	*Battus*	*polycrates*	FC778	This study	FC1386	=	15	yes
Battus	*Battus*	*polydamas*	FC1386	This study	-	FC780	15	yes
Battus	*Battus*	*streckerianus*	FC779	This study	FC1386	=	15	yes

Battus	*Battus*	*zetides*	FC579	This study	FC1386	=	15	yes
Cressida		*cressida*	FC1076	This study	-	=	15	yes
Euryades		*corethrus*	3716	This study	-	NC037862.1	15	yes
Euryades		*duponchelii*	FC510	This study	-	=	15	yes
Ornithoptera	*Aethoptera*	*allotei*	FC1606	This study	GCA_029030555.1	=	15	yes
Ornithoptera	*Aethoptera*	*victoriae*	FC1514	This study	GCA_029030555.1	=	15	yes
Ornithoptera	*Ornithoptera*	*aesacus*	FC1537	This study	GCA_029034505.1	=	15	yes
Ornithoptera	*Ornithoptera*	*croesus*	FC1513	This study	GCA_029034505.1	=	15	yes
Ornithoptera	*Ornithoptera*	*euphorion*	FC1540	This study	GCA_029034505.1	=	15	yes
Ornithoptera	*Ornithoptera*	*priamus*	GCA_029034505.1	Reboud et al. 2023	-	= (FC565)	15	yes
Ornithoptera	*Ornithoptera*	*richmondia*	1938	This study	GCA_029034505.1	=	15	yes
Ornithoptera	*Schoenbergia*	*arfakensis*	FC1529	This study	GCA_029030555.1	=	15	yes
Ornithoptera	*Schoenbergia*	*chimaera*	FC1604	This study	GCA_029030555.1	=	15	yes
Ornithoptera	*Schoenbergia*	*goliath*	FC1544	This study	GCA_029030555.1	=	15	yes
Ornithoptera	*Schoenbergia*	*meridionalis*	FC1519	This study	GCA_029030555.1	=	15	yes
Ornithoptera	*Schoenbergia*	*paradisea*	FC1553	This study	GCA_029030555.1	=	15	yes
Ornithoptera	*Schoenbergia*	*procus*	FC1567	This study	GCA_029030555.1	=	15	yes
Ornithoptera	*Schoenbergia*	*rothschildi*	FC351	This study	GCA_029030555.1	=	15	yes
Ornithoptera	*Schoenbergia*	*tithonus*	FC1595	This study	GCA_029030555.1	=	15	yes
Ornithoptera	*Straatmana*	*alexandrae*	GCA_029030555.1	Reboud et al. 2023	-	OQ590009.1	15	yes
Pachliopta	*Losaria*	*coon*	FC899	This study	FC539	=	15	yes
Pachliopta	*Losaria*	*doubledayi*	FC1323	This study	FC539	=	15	yes
Pachliopta	*Losaria*	*neptunus*	FC706	This study	FC539	=	15	yes
Pachliopta	*Losaria*	*palu*	FC707	This study	FC539	=	15	yes
Pachliopta	*Losaria*	*rhodifer*	FC709	This study	FC539	=	15	yes
Pachliopta	*Pachliopta*	*adamas*	FC824	This study	FC539	=	partial (4)	yes
Pachliopta	*Pachliopta*	*antiphus*	FC828	This study	FC539	=	15	yes
Pachliopta	*Pachliopta*	*aristolochiae*	FC539	This study	-	=	15	yes
Pachliopta	*Pachliopta*	*atropos*	FC1843	This study	FC539	=	15	yes
Pachliopta	*Pachliopta*	*hector*	FC800	This study	FC539	=	15	yes
Pachliopta	*Pachliopta*	*jophon*	FC798	This study	FC539	=	15	yes

Pachliopta	*Pachliopta*	*kotzebua*	FC813	This study	FC539	=	15	yes
Pachliopta	*Pachliopta*	*leytensis*	FC815	This study	FC539	=	15	yes
Pachliopta	*Pachliopta*	*liris*	FC814	This study	FC539	=	15	yes
Pachliopta	*Pachliopta*	*mariae*	FC1074	This study	FC539	=	15	yes
Pachliopta	*Pachliopta*	*oreon*	FC839	This study	FC539	=	15	yes
Pachliopta	*Pachliopta*	*pandiyana*				FC1666	partial (11)	
Pachliopta	*Pachliopta*	*phlegon*	FC846	This study	FC539	=	15	yes
Pachliopta	*Pachliopta*	*polydorus*	FC849	This study	FC539	=	15	yes
Pachliopta	*Pachliopta*	*polyphontes*	FC856	This study	FC539	=	15	yes
Pachliopta	*Pachliopta*	*strandi*	FC864	This study	FC539	=	15	yes
Parides		*aeneas*	CPJFLC52	This study	FC1383	=	15	yes
Parides		*agavus*	FC501	This study	FC1383	=	15	yes
Parides		*alopius*	FC509	This study	FC1330	=	15	yes
Parides		*anchises*	FC504	This study	FC1330	=	15	yes
Parides		*ascanius*	FC1814	This study	FC1383	=	15	yes
Parides		*bunichus*	FC503	This study	FC1383	=	15	yes
Parides		*chabrias*	FC493	This study	FC1330	=	15	yes
Parides		*childrenae*	CPJFLC35	This study	FC1330	=	15	yes
Parides		*cutorina*	CPJFLC46	This study	FC1330	=	15	yes
Parides		*diodorus*	FC1820	This study	FC1383	=	15	yes
Parides		*echemon*	SP697	This study	FC1383	=	15	yes
Parides		*erithalion*	FC505	This study	FC1330	=	15	yes
Parides		*eurimedes*	FC1383	This study	-	=	15	yes
Parides		*gundlachianus*	FC481	This study	FC1383	=	15	yes
Parides		*hahneli*	FC1586	This study	FC1330	=	15	yes
Parides		*iphidamas*				KP849215.1	only CO1-5P	
Parides		*klagesi*				FC1816	15	
Parides		*lysander*	CPJFLC50	This study	FC1383	=	15	yes
Parides		*montezuma*	FC1789	This study	FC1383	=	15	yes
Parides		*neophilus*	CPJFLC48	This study	FC1383	=	15	yes
Parides		*orellana*	CPJFLC49	This study	FC1383	=	15	yes

Parides		*panares*	CPJFLC59	This study	FC1330	=	15	yes
Parides		*panthonus burcellanus*	FC1821	This study	FC1383	=	15	yes
Parides		*phalaecus*	SP977	This study	FC1383	=	15	yes
Parides		*phosphorus*	FC525	This study	FC1330	=	15	yes
Parides		*photinus*	FC511	This study	FC1330	=	15	yes
Parides		*pizarro*	FC487	This study	FC1383	=	15	yes
Parides		*proneus*	FC478	This study	FC1383	=	15	yes
Parides		*quadratus*	SP1305	This study	FC1383	=	15	yes
Parides		*sesostris*	FC1793	This study	FC1330	=	15	yes
Parides		*steinbachi*				FC1667	15	yes
Parides		*tros*	FC1790	This study	FC1383	=	15	yes
Parides		*vercingetorix*	FC1817	This study	FC1383	=	15	yes
Parides		*vertumnus*	FC1330	This study	-	=	15	yes
Parides		*zacynthus*	FC1786	This study	FC1383	=	15	yes
Pharmacophagus		*antenor*	FC1333	This study	-	=	15	yes
Trogonoptera		*brookiana*	FC1336	This study	-	=	15	yes
Trogonoptera		*trojana*	FC1633	This study	-	=	15	yes
Troides	*Ripponia*	*hypolitus*	FC325	This study	GCA_029032895.1	=	15	yes
Troides	*Troides*	*aeacus*	FC1620	This study	GCA_029032895.1	=	15	yes
Troides	*Troides*	*amphrysus*	FC313	This study	GCA_029032895.1	=	15	yes
Troides	*Troides*	*andromache*	FC1047	This study	GCA_029032895.1	=	15	yes
Troides	*Troides*	*celebensis*	FC279	This study	GCA_029032895.1	=	15	yes
Troides	*Troides*	*criton*	FC262	This study	GCA_029032895.1	=	15	yes
Troides	*Troides*	*cuneifera*	FC320	This study	GCA_029032895.1	=	15	yes
Troides	*Troides*	*darsius*	FC1614	This study	GCA_029032895.1	=	15	yes
Troides	*Troides*	*dohertyi*	FC1624	This study	GCA_029032895.1	=	15	yes
Troides	*Troides*	*haliphron*	FC266	This study	GCA_029032895.1	=	15	yes
Troides	*Troides*	*helena*	GCA_029286815.1	This study	-	KT179907.1	15	yes
Troides	*Troides*	*magellanus*	FC307	This study	GCA_029032895.1	=	15	yes
Troides	*Troides*	*minos*	FC1628	This study	GCA_029032895.1	=	15	yes

Troides	*Troides*	*miranda*	FC1046	This study	GCA_029032895.1	=	15	yes
Troides	*Troides*	*naias*	FC1058	This study	GCA_029032895.1	=	15	yes
Troides	*Troides*	*oblongomaculatus*	GCA_029032895.1	Reboud et al. 2023	-	= (FC563)	15	yes
Troides	*Troides*	*plateni*	Tr2001h	This study	GCA_029032895.1	=	15	yes
Troides	*Troides*	*plato*	FC1610	This study	GCA_029032895.1	=	15	yes
Troides	*Troides*	*prattorum*	FC312	This study	GCA_029032895.1	=	15	yes
Troides	*Troides*	*rhadamantus*	FC311	This study	GCA_029032895.1	=	15	yes
Troides	*Troides*	*riedeli*	FC263	This study	GCA_029032895.1	=	15	yes
Troides	*Troides*	*staudingeri*	FC277	This study	GCA_029032895.1	=	15	yes
Troides	*Troides*	*sumatrana*	FC1794	This study	GCA_029032895.1	=	15	yes
Troides	*Troides*	*vandepolli*	FC258	This study	GCA_029032895.1	=	15	yes



Figure S1: Phylogenomic relationships of Troidni with the supertree approach. The phylogeny has been reconstructed with ASTRAL, a supertree approach based on 1,663 gene trees inferred with the maximum-likelihood inference in IQ-TREE.

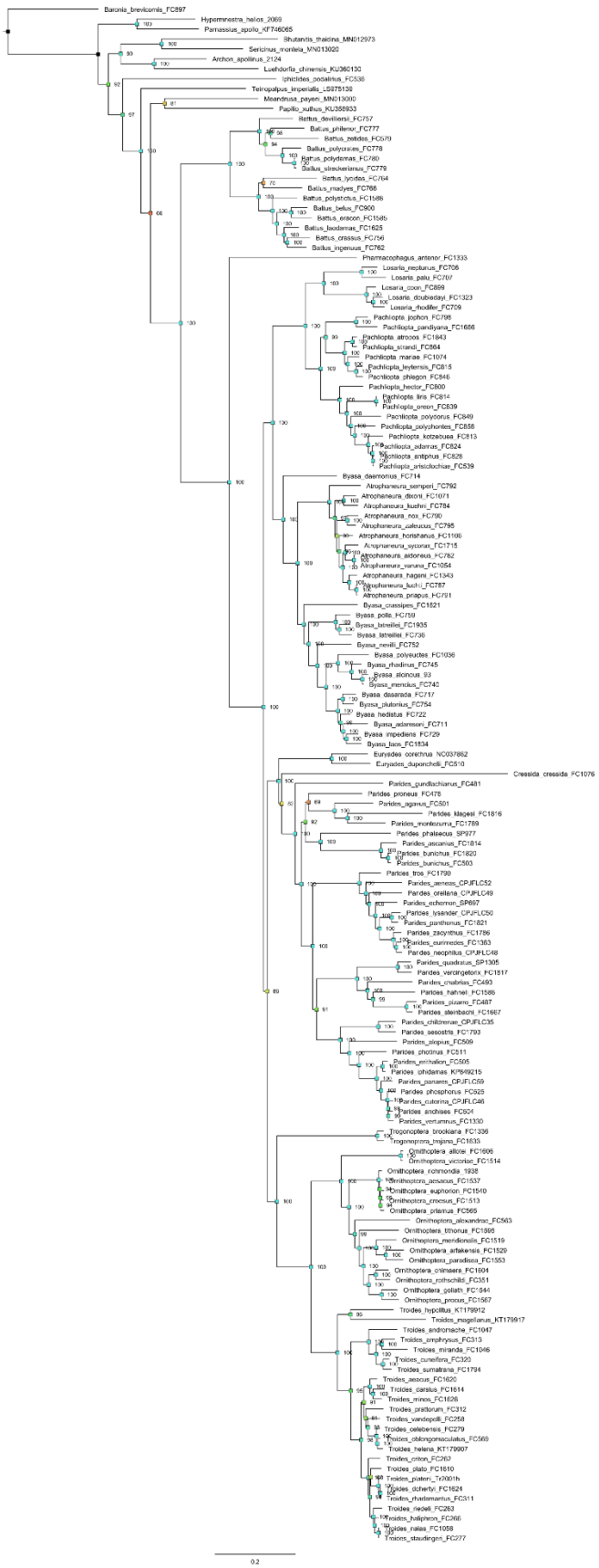


Figure S2: Phylogenomic relationships of Troidini using mitochondrial data (*IQTREE_15_Mito*). The phylogeny has been reconstructed with a supermatrix approach of 15 mitochondrial genes and a maximum-likelihood inference in IQ-TREE.

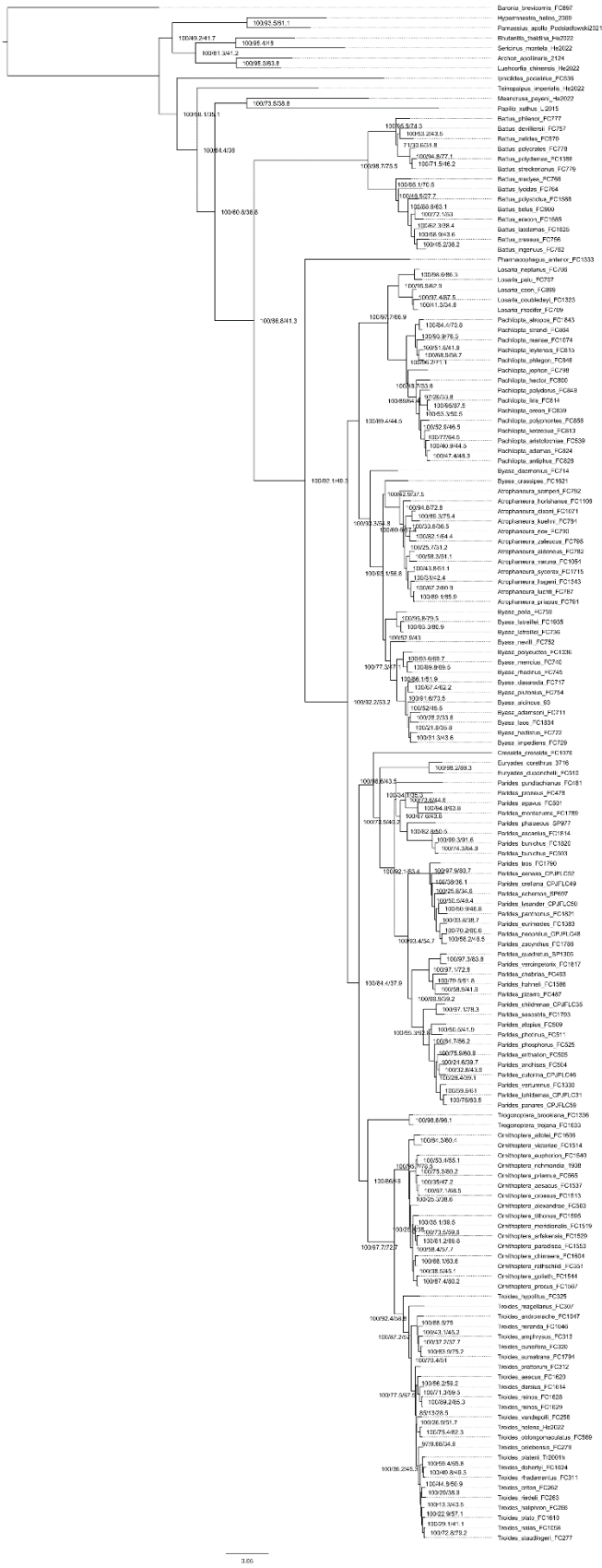


Figure S3: Results of gene and site Concordance Factor. The phylogenomic tree is from the IQ-TREE analysis. The UFBS/gCF/sCF values are displayed at each node. gCF values have been calculated with the 1,663 gene trees, and sCF values with the supermatrix of the same analysis.

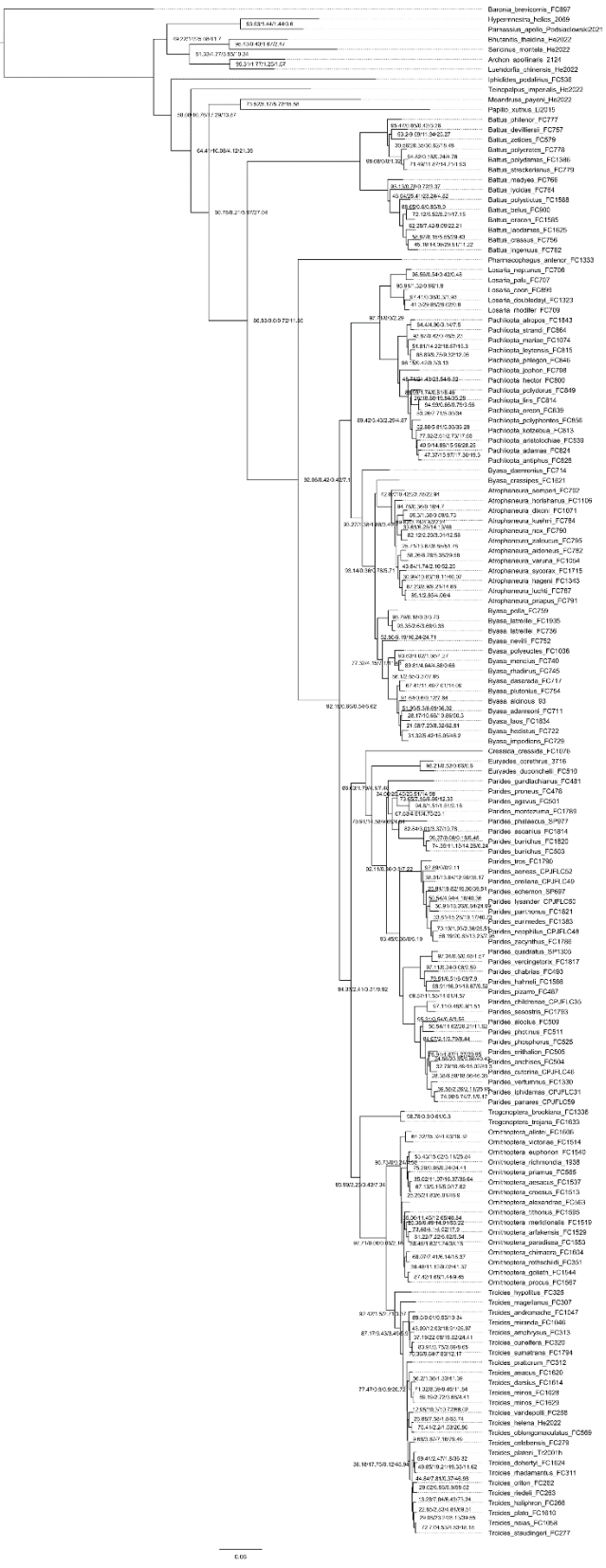


Figure S4: Gene Concordance factor results. The phylogenomic tree is from the IQ-TREE analysis. The gCF/gDF1/gDF2/gDFP is displayed at each node. These values have been calculated with the 1,663 gene trees. gCF is the number of gene trees in accordance with the consensus tree, gDF1 and gDF2 are the two triplet alternatives. gDFP are all the remaining trees (polytomies and paraphyly of the triplet) so that gCF + gDF1 + gDF2 + gDFP = 1.

CHAPTER 3



Chenille de Troidini - Eliette Reboud

Comparative macroevolutionary history of two tribes of swallowtail butterflies: did host-plant shifts play a dominant role in diversification?

Eliette L. Reboud[1]*, Eleanor Sheridan[1], Nathan Mazet[1], Benoit Nabholz[1,2], Fabien L. Condamine[1]*

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

Corresponding authors (*)

Eliette L. Reboud: eliette.reboud@gmail.com

Fabien L. Condamine: fabien.condamine@gmail.com

Running head

Diversification of tropical Papilioninae

Abstract

Among swallowtail butterflies, the Troidini and Leptocircini tribes are striking for their similarities (age, origin, distribution, specific richness), but also for their differences (host breadth, morphological disparity, aposematism). In this study, we take advantage of these characteristics to investigate the impact of plant-insect interactions on the evolutionary history of Papilionidae, as well as the impact of biogeographic events as confounding effects of host-plant shifts. We inferred time-calibrated phylogenies using a Bayesian relaxed-clock approach calibrated with secondary calibrations, and several independent runs with random sampling of genes to account for computational constraints and methodological difficulties. We then estimated ancestral traits and branch-specific as well as clade-specific diversification dynamics, expecting that host-plant shift impacted butterfly diversification, while studying the modes of host-plant change (evolutionary radiation, oscillation, musical chairs). We found that the two tribes originated on the Laurasian continent during the Eocene and mainly diversified southwards into the Neotropics or the Indo-Australasian Archipelago. The colonization of Africa by *Graphium* was inferred to have taken place in the middle Oligocene. The enigmatic colonization of Madagascar by *Pharmacophagus antenor* was suggested to have taken an African continental route with subsequent local extinction. Overall, we found higher diversification rates in Troidini than in Leptocircini, and very low extinction rates. For Leptocircini, we inferred an ancestral host-plant diet consisting of Annonaceae, Hernandiaceae and Rosaceae, as well as several host broadenings in various clades during their evolutionary history. Our analyses suggest a diversification shift in the *Arisbe*, *Protesilaus* and *Pathysa* clades, but host-plant changes did not appear as a clear driver of these diversification patterns. In the Troidini, diversification shifts were found mainly in the *Ornithoptera* and *Troides*, but also in the *Pachliopta* and *Atrophaneura*, suggesting an effect of island-driven speciation.

Keywords: birth-death models, escape and radiate, historical biogeography, Leptocircini, molecular dating, Troidini.

Introduction

Within the swallowtail butterflies, the tribes Leptocircini and Troidini are as striking for their similarities as they are for their differences. Both appeared at the same age (~40 Ma) and the same continent (Laurasia; Condamine et al. 2012, Allio et al. 2021). Their richness, which is currently quite equivalent, is concentrated in the tropics (96% and 99% for Leptocircini and Troidini, respectively), particularly in the Indo-Australasian Archipelago (46% and 64%, respectively) and in the Neotropics (27% and 35%, respectively). They both include monospecific or species-poor genera (e.g. the Australian *Protographium leosthenes* [Leptocircini] and *Cressida cressida* [Troidini]) but also species-rich genera (e.g. the Neotropical *Eurytides* [Leptocircini] and *Parides* [Troidini]). Nonetheless, these two phylogenetically close tribes have some key differences. Firstly, Troidini are absent from the African continent (except for the troidine *Pharmacophagus antenor* in Madagascar). Secondly, Leptocircini feed on different host plants (e.g. Annonaceae, Lauraceae, Hernandiaceae, Rosaceae, and Aquifoliaceae), mostly as specialist but sometimes displaying a broader host breadth including two or three families, whereas Troidini feed exclusively on Aristolochiaceae, a plant family thought to the ancestral host plant to the Papilionidae (Condamine et al. 2012, Allio et al. 2021). Accordingly, Leptocircni have likely experienced several host-plant shifts during their evolutionary history, while the Troidini never have any. Finally, Leptocircini and Troidini show a great disparity in their morphologies and heterogeneity of these morphologies (Chotard 2022, Puissant et al. 2023, **Figure 1**). While Leptocircini have evolved at least twice (and for whole clades now) towards mimicking other butterflies (the mimetic *Mimoides* imitate *Parides* or *Heliconius*, and the mimetic *Pathysa* (ex-*Paranticopsis*) imitate *Parantica* and *Idaeopsis*), it seems to be uncommon in Troidini (but see some Neotropical species such as *Parides hahneli* for example of Danainae mimicry). On the other hand, Troidini show greater disparities in shape and size than Leptocircini, with marked sexual dimorphism in several clades (the case of the *Ornithoptera* being the most striking, Wallace 1865). In addition, they often present aposematic colors (e.g. *Pharmacophagus*, *Byasa* and *Pachliopta* species all have red bodies and red patches on their wings, **Figure 1**) or very bright colors (males of *Ornithoptera* and *Trogonoptera*). On the contrary, Leptocircini often display a combination of striped pattern and tailed wings (e.g. *Iphiclides*, *Protographium*, *Protesilaus*, *Pazala*), that been shown to deflect predation in *Iphiclides podalirius* (Chotard et al. 2022, Chotard 2022, **Figure 1**).

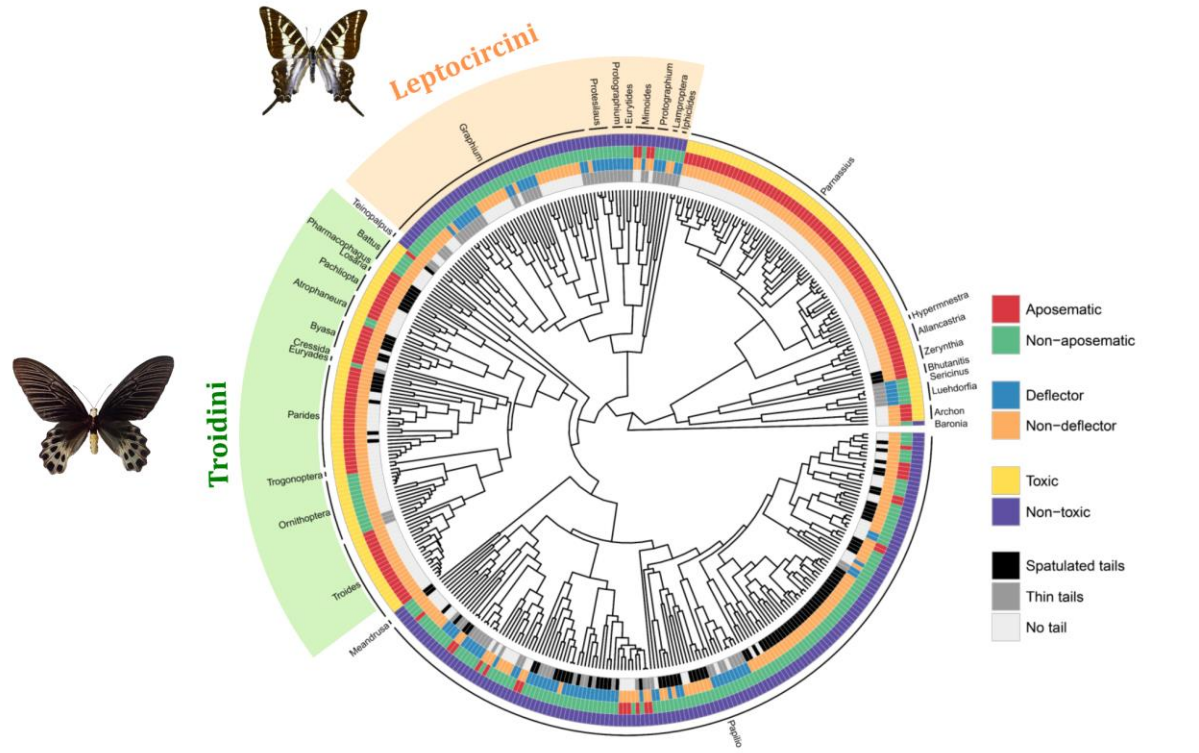


Figure 1: Phylogenetic tree of 408 Papilionidae species with associated traits. The Leptocircini and Troidini tribes are indicated in orange and green respectively). Illustrations: Troidini - *Atrophaneura sycorax* (Eliette Reboud); Leptocircini - *Graphium aristeus* (Alex Dumchus). Figure modified from Chotard (2022).

These similarities and differences revive one of the most puzzling questions in evolutionary biology: understanding how ecological, geological and evolutionary processes have influenced extant and extinct biodiversity patterns (Benton 2016). The heterogeneity of biodiversity distribution, both among lineages and geographical areas (Myers et al. 2000, Barthlott et al. 2007), is the outcome of millions of years of evolution of Earth's organisms, driven by natural selection and further shaped by environmental conditions, and stochastic processes (Benton and Emerson 2007). These processes are classically divided into abiotic and biotic factors (Barnosky 2001, Benton 2009), both personified in the entity of the Red Queen and the Court Jester, respectively (see **General Introduction of this Thesis**). Both factors could act on the spatial distribution of organisms (e.g. abiotic: land masses movements, oceanic currents, winds; biotic: individual dispersal capacity, phoresy, host plant association and niche preference) and on their fitness, their reproductive isolation or success (e.g. abiotic: temperature, positions of landmasses and climatic barriers; biotic: sexual selection).

Indeed, environmental changes and geological events are important candidates

for explaining changes in the rates of diversification (Condamine et al. 2019), and insects are no exception (Labandeira and Sepkoski 1993, Jouault et al. 2022). Indeed several insect groups have been the subject of biogeographic studies to study the diversification of insects as a function of extrinsic factors such as paleogeographic or paleoclimatic events (e.g. Coleoptera: Condamine et al. 2013a, Toussaint et al. 2016, 2017; Lepidoptera: Condamine et al. 2012, Chazot et al. 2019, Rougerie et al. 2022; bees: Almeida et al. 2019, 2023, Bossert et al. 2022).

On the other hand, one of the most famous theories supporting the impact of biotic factors on a macroevolutionary scale has been proposed by Ehrlich and Raven (1964). This theory, which was later called the "escape-and-radiate" model (Thompson 1989), hypothesizes an arms race between host plants and phytophagous insects that would induce species diversification through the creation of new adaptive zones (Mitter et al. 1988). The coevolution of phytophagous insects and their specific host plants is an antagonistic relationship, which would trigger physicochemical defenses of the plants and circumvention of these defenses by the insects (Simon et al. 2015). These defenses could range from mechanical characteristics of the plants to chemical defenses with toxins, that would trigger behavioral modification (Fordyce 2003, Pinto et al. 2009, Dimarco and Fordyce 2013) and genetic adaptation (Cohen et al. 1992, Berenbaum et al. 1996, Li et al. 2003, Heidel-Fischer et al. 2019, Steiner et al. 2019, Okamura et al. 2019, Griese et al. 2021) from the feeders. The ability to sequester toxins could even become an additional benefit as larvae become unpalatable themselves (Nishida 2002). This theory is a particular case of the concept of key innovation, i.e. a novel phenotypic trait leading to the subsequent radiation and success of a taxonomic group, by increasing specialization, allowing the access to new niches, or increasing fitness compared to other taxonomic groups (Heard and Hauser 1995). In insects, acquisition of flight (De Queiroz 1998, Nicholson et al. 2014), sociality (Legendre and Condamine 2018, Chouvenc et al. 2021), and parasitoidism (Blaimer et al. 2023), or apparition of metamorphosis (Nicholson et al. 2014, Rainford et al. 2014) are examples of key innovations that could have helped their diversification. In phytophagous insects, the escape and radiate theory have received various empirical support at several taxonomic scales (e.g. Cardinal and Danforth 2013, Arhens et al. 2014, Sahoo et al. 2017, Letsch et al. 2018, Shin et al. 2018, Allio et al. 2021).

While the debate of biotic and abiotic drivers has long been polarized as to which of these factors played the greater role in the diversification of organisms, many studies now seem to agree that the two could have played, or even interacted, in shaping clades' diversification. As a consequence, the conclusions of a given study may depend on which factors were tested (e.g. Silvestro et al. 2015, Liow et al. 2015, Condamine et al. 2019b, 2020, Guo et al. 2023). That is why it is increasingly proposed to integrate the two types of

factors (Lagomarsino et al. 2016, Condamine et al. 2019a, 2019b, 2021).

However, addressing this question comes with major methodological difficulties. The first is and has always been the accuracy of time-calibrated trees. Many insect groups, such as Lepidoptera (Sohn et al. 2015), usually do not have a well-enough furnished fossil to provide fossil calibrations. Although a good sampling fraction and many genes are known to reduce the risk of errors in dating (Linder et al. 2005), genomic datasets generated today represent computational bottlenecks that make phylogenomic dating unfeasible in reasonable timeframes (dos Reis et al. 2016, Collins and Hrbek 2018, Smith et al. 2018, Allio et al. 2020a). Bayesian methods, which are today considered the most reliable (such as BEAST or PhyloBayes), are still constrained by the number of taxa or the amount of data to work within a reasonable time (Smith et al. 2018, To et al. 2016).

The second major issue lies in the way to disentangle effects of an host-plant shift with the impact of the colonization of a new biome or environment that could have occurred simultaneously to a given clade. For instance, in Papilionidae, the genus *Parnassius*, the only swallowtail genus to feed on Papaveraceae (Hancock 1983) and to thrive at high elevations (Weiss 1991), is a good example of the possibility that host-plant preference might go hand in hand with geographical or biome occupation (Condamine et al. 2018). Understanding the potential confounding effects of biogeographic context with host-plant preference is important for accurately assessing the true contribution of host-plant shifts to species diversification (Nyman et al. 2010, 2012, 2019, Sahoo et al. 2017). Becerra and Venable (1999) review that two main scenarios could explain macroevolutionary patterns of host-plant use. Either shifts by herbivorous insects are mediated by the similarities of secondary compounds of their hosts (Ehrlich and Raven 1964, Futuyma and McCafferty 1990, Becerra 1997, Köpf et al. 1998), such as the furanocoumarins-containing plant families similarities could explain colonization of both Rutaceae and Apiaceae for *Papilio* species (Berenbaum 1983). Or the patterns of evolution of host-plant use could be the result of parallel cladogenesis. If insect lineages remain associated with their hosts over a long time, events that isolate host populations might also isolate populations of their associated insects, which may eventually result in allopatric cospeciation (e.g. chrysomelid genus *Phyllobrotica*, Farrell and Mitter 1990; *Ficus* and fig wasps co-evolution, Cruaud et al. 2012). Interestingly in their empirical case, Becerra and Venable (1999) found that chemistry-driven vs. parallel cladogenesis may vary within the same group depending on the nature and intensity of the (antagonistic) plant-insect association. Whereas the distribution of their polyphagous beetle species was more driven by the geographic distribution of host plants, the evolution of monophagous species was mostly explained by host-plant chemistry (Becerra and Venable 1999).

This brings us to the third challenge, which tackles the concept of "trait shift" itself: how exactly is a host-plant shift

defined? Could a host expansion be considered as a shift? Although the escape and radiate model has been instrumental in insect-plant interactions (Janz 2011, Jousselin and Elias 2019), two alternative hypotheses of host-plant shifting at macroevolutionary scales have been presented: the "oscillation" versus "musical chair" models. The musical chair model hypothesizes that speciation is driven by a complete host-plant switching without niche expansion (Hardy and Otto 2014). On the contrary, the oscillation hypothesis predicts that insect herbivore lineages undergo alternating phases of host expansion and contraction, and speciation would therefore be driven by host-plant generalists giving rise to specialist daughter species (Janz and Nylin 2008). Since both models are plausible, it is rational to incorporate any increase in the host-plant repertoire in the definition of host-plant shift, regardless of whether there is a reduction or expansion of that repertoire. But again, defining a host-plant repertoire is not easy, as it could refer to plant genera, plant families, or even chemical compounds categories (Wahlberg 2001, Kergoat et al. 2005). To be consistent with the concept of key innovation, we may define shifts as gain of traits that allow to colonize or maintain a niche (Schluter 2000, Terekhanova et al. 2014, Marques et al. 2018), and this could apply both to colonization of new biomes or geographical regions, and to gain of new host plants, and acquisition of adaptation to this new host-plant defenses. Yet, having information on host-plant chemistry, degree of toxicity and similarity of toxins remains difficult at a macroevolutionary scale and for entire plant families. Many of these toxins are known to interact with the cytochrome P450 monooxygenase superfamily (Schuler 1996), a gene family with highly divergent functions, but which has been shown to play a major role in plant-insect interactions, and possibly, in the diversification of the two antagonists (Berenbaum 1996, Jun et al. 2015).

As reviewed by Jousselin and Elias (2019), numerous predictions can be made and may be tested regarding the impact of a host-plant change on species diversification with phylogenies. These predictions relate in particular to patterns of host-plant shifts within a group of species and the dynamics of diversification rates following these shifts. These predictions differ depending on whether the dynamics of the shifts are based on "oscillation" or "musical chair" models, for example. The general case of speciation driven by specialization on a host plant tends to predict patterns of adaptive radiation (Mitter et al. 1998, Janz 2001), in which saturation of the colonized niche (new host plant) is progressively reached after an early burst, and therefore slowing down the speciation rates over time after these shifts. In practice, these macroevolutionary models and predictions remain relatively difficult to test (Jousselin and Elias 2019), but can be explored by reconstructing ancestral traits and studying diversification rates with insect phylogenies (e.g. Sahoo et al. 2017, Condamine et al. 2018, Letsch et al. 2018).

Recently, Allio et al. (2021) demonstrated that a large proportion of genomes display evolutionary signatures

involved in plant-insect coevolutionary innovations in Papilionidae, with more genes subject to positive selection than initially suspected. These genome-wide adaptations are associated with increases in diversification rates during host-plant shifts. For instance, the genus *Parnassius* shifted to Crassulaceae and subsequently to Papaveraceae with associated bursts of diversification and colonizations of mountain ranges (Condamine et al. 2018). This naturally raised the question of possible confounding effects between plant shifts and other species traits. Although Allio et al. (2021) tested whether the concomitance of host-plant shifts and changes in climatic preferences could have contributed to increased adaptive evolution, they did not detect any particular effects due to environmental changes (Allio et al. 2021). Yet, the phylogenetic framework was not ideal as it allowed only a few control branches to be tested.

In this study, we exploit the similarity in distribution and tropicality and the differences in host-plant breadth of Leptocircini and Troidini, along with their new phylogenomic framework, to bring new insights in our understanding of the effect of biotic and abiotic factors on diversification. First, we inferred time-calibrated phylogenies using a Bayesian relaxed-clock approach calibrated using secondary calibrations, and several independent runs to account for computational constraints and methodological difficulties. Second, we investigated the macroevolutionary history of each tribe using the dated trees by comparing their biogeographical history and diversification dynamics over time. Given the natural history of the two focal tribes, we hypothesized that changes in host plant in Leptocircini should have a detectable effect on the diversification of the shifting clades, and that their relatively similar age and distribution with Troidini might partly help to disentangle confounding effects related to biogeographic history.

Materials and Methods

Molecular dataset and phylogenomic inference

The species-level phylogenies of both tribes were inferred by combining nuclear genome data (orthologous genes) with the mitochondrial gene cytochrome c oxidase subunit 1 (COI). The CO1 gene of all individuals was retrieved from the same genomic data through mitogenome assemblies in MitoFinder (Allio et al. 2020b, see **Chapters 2.1** and **2.2**). We only kept the COI gene as it allows including the species that have not been sampled or failed at whole-genome sequencing through available data on GenBank. We were able to gather 156 species of Leptocircini (95% of the total species diversity) and 144 species of Troidini (97% of the total species diversity). To account for frameshift and stop codons in the COI, sequences were aligned with MACSE 2.07 (Ranwez et al. 2018), setting an invertebrate mitochondrial genetic code (*-gc_def 5*), a frame-shift penalty of 5 (*-fs 5*), and an internal stop codon penalized by 10 (*-stop 10*). The best partition scheme of the alignment was recovered with IQ-TREE

2.2.0 (Minh et al. 2020) using ModelFinder to select the best-fit partition scheme and the best-fitting substitution model for each partition (*-m MFP+MERGE* option, Chernomor et al. 2016, Kalyaanamoorthy et al. 2017) applied to an initial subset of 3 possible partitions (the three codon positions).

The resulting CO1 alignment was concatenated to the nuclear genes supermatrix composed of orthologous genes obtained from BUSCO (1,402 genes for Leptocircini, 1,663 genes for Troidini; see **Chapters 2.1** and **2.2**). The best-fitting partition scheme associated with the best-fitting substitution models of each partition were added to the best partition scheme and best-fitting substitution models of the COI analysis. We then performed a maximum-likelihood (ML) phylogenetic inference with IQ-TREE using default parameters. Node robustness of the ML tree was evaluated with 1,000 ultrafast bootstraps (UFBS; *-B 1000* option).

Bayesian relaxed-clock inference

Divergence times were estimated using a Bayesian approach with Markov Chain Monte Carlo (MCMC) as implemented in the MCMCtree included in PAML 4.9.j (Yang 1997). Analyses were performed on a fixed topology resulting from the tree combining COI and BUSCO data. Because no fossils are known in the two studied groups, the molecular clocks were calibrated using secondary calibration points (Ho and Phillips 2009). Following the study of Condamine et al. (2023) on the genus *Papilio*, we relied on secondary calibrations with uniform priors applied to the nodes where we set the calibrations to the nodes that are in common between our phylogeny and the global phylogeny of Papilionidae. We retrieved the 95% credibility interval (CI) from recent estimates of divergence times of the family (Allio et al. 2021) as follows: (1) the crown of Papilionidae (root of the tree) between 47.8 and 70.9 Ma, (2) the crown of Parnassiinae between 29.9 and 58.9 Ma, (3) the crown of Luehdorfiini between 13.9 and 35.3 Ma, (4) the crown of Papilioninae between 34.4 and 62.9 Ma, (5) the crown of Teinopalpini + Papilionini between 32.2 and 58.6 Ma (Allio et al. 2021), and (6) the crown of Papilionini between 27.5 and 50.9 Ma (Allio et al. 2021). For Leptocircini, we added the calibration (7) to the crown of Papilionini + Troidini between 30.8 and 56.1 Ma (Allio et al. 2021).

The priors used for node age distribution are, by default in MCMCtree, uniform density distributions with soft bounds, which allows ages to exceed the predefined bounds by 5% (Yang and Rannala 2006). For each tribe, we randomly created 10 molecular datasets including 50 nuclear orthologous genes and each data does not include a gene overlapping with another dataset (i.e. 500 genes are sampled in total). We also added the CO1 gene to every nuclear dataset. Each dataset was analyzed twice independently with MCMCtree to ensure MCMC convergence, leading to a total of 20 runs (10 datasets run twice independently). For each dataset, we first approximated branch lengths with a maximum likelihood step developed by Thorne et al. (1998) and implemented in PAML (gradient [g] and Hessian matrix

[H], *usedata = 3*), in order to reduce the computational time of the likelihood function calculation during the MCMC (dos Reis and Yang 2011, dos Reis et al. 2017). Then, we used the approximate likelihood method (*usedata = 2*) to estimate divergence times, setting an uncorrelated rate model (Drummond et al. 2006; option *clock = 2*). We used the most complex substitution model implemented in MCMCtree (HKY85, *model = 4*). We followed MCMCtree tutorials to set the alpha and gamma prior parameters (dos Reis et al. 2017). Specifically, the alpha parameter for gamma rates at sites was set to 2 (*alpha = 2*) and the number of categories in discrete gamma was set to 5 (*ncatG = 5*). The branching tree process was set as a Birth-Death model, with default parameters. The Bayesian inference was set to 5 million MCMC generations, with a 10% of burn-in (500,000 generations). Convergence of each MCMCtree run was verified using Tracer 1.7.1 (Rambaut et al. 2018) ensuring that effective sample sizes reached 200 for most of the parameters. Finally, a consensus dated tree of each tribe was obtained with TreeAnnotator 1.10 (Suchard et al. 2018) by giving the 10 approximations (for each run, one over the two independent approximations made to check for convergence) of each MCMCtree analysis. For the macroevolutionary analyses (biogeography and diversification) of each tribe, we used this summary tree, from which we removed outgroups species.

Estimating historical biogeography

We estimated the ancestral ranges of origin and geographic range evolution for Leptocircini and Troidini using the Dispersal-Extinction-Cladogenesis (DEC) model (Ree et al. 2005, Ree and Smith 2008, Ronquist and Sanmartín 2011). To infer the biogeographic history of a clade, DEC requires a time-calibrated tree, the current distribution of each species for a set of geographic areas, and a time-stratified geographic model that is represented by connectivity and dispersal rates matrices for specified time intervals spanning the entire evolutionary history of the group. Here we used an extended DEC model (DECX, Beeravolu and Condamine 2016, available at: https://github.com/champost/DECX) that allows classical vicariance as a cladogenetic event by using temporally flexible constraints on the connectivity between any two given areas following the movement of landmasses and dispersal opportunity over time.

The geographic distribution for all Leptocircini and Troidini species was categorized by coding the presence or the absence of each species in the following 13 areas: (1) West Palearctic, defined as Europe west of the Urals and the part of Asia west of this north–south line, (2) East Palearctic, defined as everywhere considered Palearctic east of the Urals, above 3,000 m in the Himalayas and north of Sichuan in China, (3) West Nearctic, defined as North America west of the Rocky Mountains, (4) East Nearctic, defined as North America east of the Rocky Mountains, (5) Central America, defined as

from the northern border of Mexico southwards to the border between Panama and Colombia, (6) the Caribbean Islands except Trinidad and Tobago, (7) South America, defined as all countries from Colombia to Argentina and including Trinidad and Tobago, (8) Africa, defined as the whole of the African continent and Arabian Peninsula but excluding the islands in the Indian Ocean, (9) Madagascar, defined as the island of Madagascar and all other Indian Ocean islands in the vicinity, (10) India, defined as the area below 3,000 m from Northwest Pakistan to the border with Myanmar, (11) continental Southeast Asia from Myanmar, southern China, to Taiwan and Ithmus of Kra in Indonesia, (12) insular Southeast Asia including Indonesia and Wallacea, defined as western Indonesia to Lydekker's Line; including the Lesser Sunda Islands but excluding Timor, Wetar and associated islands, which are Australasian in origin, and (13) Australasia, defined as everywhere east of Lydekker's Line but including Timor, Wetar and small nearby islands.

Species which only marginally enter an area were excluded (e.g. *Eurytides philolaus* barely occurs in South America, or *Byasa polyeuctes* barely occurs in East Palearctic). We used data available in the literature (e.g. Collins and Morris 1985, Tyler et al. 1994, Scriber et al. 1995, Allio et al. 2021, Nakae 2021) and corroborated with iNaturalist (www.inaturalist.org) and GBIF (www.gbif.org) occurrences (last accessed September 27, 2023).

A time-stratified geographic model was built using connectivity matrices that consider paleogeographic changes through time with time slices indicating the possibility or not for a species to colonize a new area (Beeravolu and Condamine 2016). Based on paleogeographic reconstructions (e.g. Scotese 2004, Blakey 2008, Seton et al. 2012), we created a connectivity matrix to represent major changes in tectonic conditions that may have affected the distribution of these butterflies.

We specified constraints on area connectivity by coding 0 if any two areas are not connected or 1 if they are connected at a given period (**Table S1**). We did not add dispersal matrices because setting the values for dispersal rates between regions through time is highly subjective, and it has been shown that dispersal probability categories had minor effects on ancestral state estimation (Chacon and Renner 2014). Therefore, we assumed a dispersal scalar matrix with equal dispersal rates between areas through time. Biogeographic ranges larger than four areas in size were disallowed as valid biogeographic states if they were not subsets of the terminal species ranges; widespread ranges comprising areas that have never been geographically connected were also removed. We estimated the most likely ancestral states at each node by performing a DEC analysis under an ML procedure as implemented in DECX (Beeravolu and Condamine 2016) using both species distribution matrix and connectivity matrices.

Estimating ancestral host-plant preferences

We inferred the temporal evolution of host–plant association up to the ancestral host plant(s) at the root of Leptocircini using also the ML approach as implemented in the DECX model (Beeravolu and Condamine 2016). As for the historical biogeography, this approach requires a time-calibrated tree and a matrix of character states (current host–plant preferences) for each species in the tree. An extensive bibliographic survey was conducted to obtain primary larval host plants for each species (e.g. Hancock 1983, Collins and Morris 1985, Chou 1994, Möhn 2002, Beccaloni et al. 2008, Suvak 2015, Hu et al. 2019, Zhang et al. 2020, Williams 2022). Some plants found in the literature review had only one occurrence (usually for very generalist *Graphium s.s.* species), or were considered as dubious observations by some authors and were therefore not considered for the analyses. The host associations of Leptocircini species were categorized using the following 11 character states: (1) Annonaceae, (2) Magnoliaceae, (3) Monimiaceae, (4) Lauraceae, (5) Hernandiaceae, (6) Malpighiaceae, (7) Rosaceae; (8) Rutaceae, (9) Anacardiaceae, (10) Apocynaceae, and (11) Aquifoliaceae.

As all extant Troidini feed on Aristolochiaceae, it was considered that Aristolochiaceae was the ancestral state and that no host-plant shift happened within the plant family.

Estimating diversification rates

To test whether changes in diversification rate occurred across the evolutionary history of both tribes, we used two different methods to estimate branch-specific speciation and extinction rates, and one method to estimate clade-specific speciation and extinction rates. Because of taxonomic uncertainties in the two tribes (see Chapters 2.1 and 2.2), we performed these analyses on two datasets, with a 'lump' (dubious species are excluded) and 'complete' (dubious species are included) taxonomy.

- **Branch-specific diversification rates**

We used the cladogenetic diversification rate shift (ClaDS) model to estimate branch-specific diversification (speciation and extinction) rates using Bayesian inferences (Maliet et al. 2019). This model is a birth-death model where speciation rates are inherited at a speciation event, but with a shift drawn from a probability distribution parameterized with the parental rates (Maliet et al. 2019). We applied the ClaDS model with constant turnover (ClaDS2) as implemented in the PANDA Julia package (Maliet and Morlon 2022), i.e. the speciation and extinction rates can vary across branches, but the turnover (that is speciation divided by extinction) is constant across the branches. The number of MCMC was set to 3 (default parameters) and convergence was checked every 1,000 iterations. At these checkpoints, if the Gelman parameter value (MCMC statistic) is below a limit value (default 1.05) the convergence is considered reached and the run is stopped (Maliet et al. 2019). For each

tribe, the sampling fraction of each genus (or subgenus when possible) was provided to the model depending on the taxonomy.

To cross-validate ClaDS results, we ran the Bayesian approach for estimating branch-specific speciation and extinction rates (Höhna et al. 2019) as implemented in RevBayes (Höhna et al. 2016). This model is a birth-death model that breaks time into small intervals and for which speciation and extinction rates can change at each time interval. RevBayes is able to model rate shifts on extinct lineages although there is no information of rate shifts for extinct lineages in a time tree including extant species only, but these events must be accounted for in computing the likelihood. The number of rate categories is fixed in the analysis but RevBayes allows any number to be specified, thus allowing direct comparison of different macroevolutionary regimes. RevBayes allows both speciation and extinction to be variable. The total number of species in the tribe was given to the model for a global sampling fraction (Leptocircini: 156/164 = 95.12% [complete], 146/150 = 97.33% [lump]; and Troidini: 144/147 = 97.96% [complete], 131/134 = 97.76% [lump]), with the sampling fractions depending on the complete or the lump taxonomy. We set the discrete rate categories prior to 10, and tested the initial value of the number of expected rate shifts to 10. The parameters of the model were estimated using two reversible-jump Markov Chain Monte Carlo (rjMCMC) of 10,000 iterations each, sampling every 200 iterations, and a burn-in of 2,000. The convergence of the runs was checked with ESS above 200 using Tracer 1.7 (Rambaut et al. 2018).

- **Clade-specific diversification rates**

We applied the ML approach of Morlon et al. (2011), automated and implemented in the package RPANDA 2.0 (Morlon et al. 2016) by Mazet et al. (2023). This method aims at capturing heterogeneity of diversification by allowing preselected subclades to follow different birth–death models from the rest of the tree (called the backbone). For both the subclades and the backbones (different backbones are tested because of the different combination of shifts), speciation and/or extinction rates can change through time. Extinction rate can exceed speciation, meaning that diversification rates can be negative and suggesting a period of declining diversity (Morlon et al. 2011).

To infer the trends of diversification rates of clades after a detected shift (e.g. to test for adaptive radiation pattern, Jousselin and Elias 2019), we defined *a priori* shifts to be tested based on the clades that presented shifts in the branch-specific analyses as recovered by ClaDS and RevBayes in the complete or the lump datasets. Thus, for Leptocircini we specified the following subclades: (1) mimetic *Mimoides*, (2) *Protesilaus*, (3) *Pazala*, (4) all *Pathysa*, (5) mimetic *Pathysa*, including *Graphium stratiotes* and *G. agetes*, and (6) *Arisbe*, excluding *G. antheus* and *G. evombar*. For Troidini, the preselected subclades were: (1) subgenera *Atrophaneura* + *Byasa* (i.e. genus *Atrophaneura* excluding *A. daemonius*), (2) subgenus *Atrophaneura*, including *Byasa*

crassipes, (3) subgenus *Byasa*, (4) genera *Ornithoptera* + *Troides*, (5) genus *Ornithoptera*, and (6) genus *Troides*.

Following Mazet et al. (2023), we first computed the clades' sampling fractions of each tribe with the *get.sampling.fraction* function, based on the complete or the lump taxonomy. We then estimated the number of possible shifts to be tested in combination with corresponding backbones using the *get.comb.shift* function. To be tested, a clade must be monophyletic and contain more than five species. All other clades are included in the backbone. The selection of subclades created combinations of subclade(s)/backbone that were compared to the clade-homogeneous birth–death model. We designed and fitted five diversification models to each clade and backbone with the *shift.estimate* function that determines the best fitting diversification model and most likely combination of shifts. The five diversification models were: (1) a pure-birth model, with constant speciation rate and no extinction (BCST); (2) a birth-death model, with both constant speciation and extinction rates (BCST_DCST); (3) a model with variable speciation through time and no extinction (BVAR); (4) a model with variable speciation through time with constant extinction (BVAR_DCST); and (5) a model with constant speciation and variable extinction through time (BCST_DVAR).

Results and Discussion

Origin and divergence times of both swallowtail tribes

For Leptocircini, the ten MCMCtree runs of the two approximation analyses converged well (ESS≫200 for most of the nodes and parameters). The combined tracefile of the ten first approximations of each analysis also yielded ESS>200 (or near 200, smallest node ESS = 140) for most of the nodes and parameters, with almost identical estimates of divergence times with <0.5 million years of difference for all nodes (Figure S1-S2). The dating analyses estimated that Leptocircini originated in the early Eocene *ca.* 50.60 Ma (95% CI = 49.39–52.22 Ma, **Figure S2**). Comparing our new results with previous age estimates the crown age of Leptocircini have always been inferred younger, ranging between 34 and 45 Ma (Condamine et al. 2012, 2018, Allio et al. 2020a, 2021, Kawahara et al. 2023), but these studies usually had a poor taxonomic sampling of Leptocircini, which might explain this difference (Garzón-Orduña et al. 2015).

Surprisingly, the convergence of each MCMCtree run was less pronounced for Troidini, despite a similar approach and dataset size, similar calibrations and overall better representation of each specimen per locus (the mean BUSCO score of Troidini is higher as for Leptocircini). Nonetheless, all ten approximations had their two independent estimations being almost identical, and the combined tracefile of the ten first approximations of each analysis yielded ESS≫200 for all nodes and parameters. The dating analyses estimated

that Troidini originated in the middle Eocene *ca.* 45.52 Ma (95% CI = 45.06–46.15 Ma).

For Leptocircini, all genera appeared before the end of Oligocene (23 Ma) except the genus *Iphiclides* whose crown age is dated *ca.* 8.65 Ma (95% CI = 7.40–11.59 Ma) (**Figure 2, Figure S2**) . We also estimated that all subgenera of *Eurytides* and *Graphium* diverged before the end of Miocene (5.33 Ma). It is worth noting that greater incertitude was for *Pathysa s.l.* (**Figure S2, node 5**), which is in line with the taxonomic incertitude described in **Chapter 2.1**. For Troidini all genera appeared during the Oligocene or early/mid-Miocene (**Figure 4**, **Figure S4**).

Ancestral host-plant estimations for Leptocircini

DEC analyses inferred a broad ancestral host breadth composed of Annonaceae, Rosaceae and Hernandiaceae for Leptocircini (relative probability = 0.905). This was followed by two successive events of host breath splitting (**Figure 2**): a first specialization event led the common ancestor of *Eurytides* + *Protographium* + *Graphium* to specialize on Annonaceae, while the common ancestor of *Iphiclides* + *Lamproptera* got specialized on Rosaceae + Hernandiaceae. Further specialization on both sides led *Lamproptera* to specialize on Hernandiaceae, while *Iphiclides* specialized on Rosaceae. This differs from Allio et al. (2021), who found that only Hernandiaceae and Annonaceae were ancestral to Leptocircini, and that the gain of Rosaceae only appeared at the next node.

The internal nodes of the *Eurytides* are mainly dominated by host conservatism of the Annonaceae, but host expansion is inferred with the addition of the Lauraceae within the subgenera *Eurytides* and *Mimoides*, as well as generalization and successive specialization on Magnoliaceae within the subgenera *Eurytides* and *Protesilaus*. The common ancestor of *Graphium* is inferred with a generalist Lauraceae-Annonaceae diet, followed by successive specialization towards Lauraceae for *Pazala*, and Annonaceae for the rest of *Graphium s.l.* Within subgenus *Graphium*, the deep nodes are inferred as solely Annonaceae, but broadening of the diet towards Magnoliaceae, Hernandiaceae, Monimiaceae and Lauraceae is inferred around the Miocene-Pliocene transition (**Figure 2**).

All the plants cited above (except Rosaceae) belong to the Magnoliids (orders Laurales and Magnoliales, **Figure 2**). It is only in the *Pathysa* + *Arisbe* clade that diet extensions appeared in more distant parts of the angiosperm phylogeny, with shifts on Rosids and Asterids. Firstly, a shift to the Aquifoliaceae (Aquifoliales) is inferred for the crown node of mimetic *Pathysa* + *G. agetes* + *G. stratiotes*. As these two last taxa do not feed on Aquifoliaceae, a return to specialization on Annonaceae is inferred for their ancestor. We here refer to Chapter 2.1 for a discussion about this topology and distribution of traits. Within *Arisbe*, nodes are largely dominated by Annonaceae, but some nodes are inferred as a diet broadening with the addition of Apocynaceae (Gentiales), Malpighiaceae (Malpighiales) and Anacardiaceae

(Sapindales). From a methodological point of view, the inferences in this part of the tree seem surprising. While a large proportion of the *Arisbe* in the *adamastor* group feeds on Anacardiaceae, this plant family is not inferred in the ancestral host of their common ancestors (except for *G.adamastor* - *G. aurivilliusi*). The same counterintuitive case can be observed in the *policenes* and *leonidas* species group with Apocynaceae, for example. It is difficult to understand why the model behaved like this. We hypothesize that the vast dominance of Annoncaeae in the current Leptocircini host preference (>80%), or the dominance of specialist species (>70%), may lead to biases in the analysis towards a strong attraction to Annonaceae specialists in the ancestral states.

Globally, our analysis seems to be more in line with the oscillation hypothesis than the musical chair hypothesis (Hardy and Otto 2014, Jousselin and Elias 2019), as almost all host shifts on a new plant family was composed of a succession of generalization of the ancestor and specialization of the daughter lineages (e.g. shift on Lauraceae for *Pazala* is preceded by a generalist ancestor on Annonaceae + Lauraceae). It is quite possible that other statistical approaches would have favored an alternative evolutionary hypothesis. Indeed, DECX does not incorporate the founder-event speciation (+J parameter) because of concerns with statistical validity of model choice among DEC-derived models (Ree and Sanmartín 2018). In a biogeographical context, founder-event speciation often leads to inferences that are decoupled from time, with null or extremely low extinction rates, an effect of the model favoring cladogenetic events over anagenetic events (Ree and Sanmartín 2018), which makes it inadequate for reconstructing the history of ancient groups with widespread distributions. In the case of ancestral host-plant reconstruction, both "founder event" (i.e. host shift) and "classical dispersion" (i.e. hist expansion) might make sense, and depend on the mechanism underlying the shift, and probably in the evolutionary cost of being generalist for the defense of several plant chemical compounds.

In that regard, the inference of Rosaceae as part of the ancestral repertoire of Leptocircini along with Hernandiaceae and Annonaceae could be surprising. These two latter are Magnoliids and may share common chemical compounds (and therefore similar pathway of adaptation and chemical defenses for larvae) such as similar alkaloids and lignans (Coulerie and Poullain 2016). For instance, cytotoxic aporphine alkaloids are found in both Annonaceae (Pena-Hidalgo et al. 2021) and Hernandiaceae (Hussain et al. 1986; Chen et al. 2011, Wang et al. 2019). Interestingly, Aristolochiaceae, which are also Magnoliid plants and ancestral host plants of Papilionidae (Condamine et al. 2012, Allio et al. 2021), also contain aporphine alkaloid along with aristolochic acids (benzylisoquinoline alkaloid, magnoflorine, Feeny 1991, Brown et al. 1995, Bentley 2006, Cui et al. 2022). On the other hand, Rosaceae is part of Rosids, a phylogenetically distant clade that may contain very different chemical compounds. For example, *Prunus spinosa*,

one of *Iphiclides* host plant, is known to contain coumarins (Irizar et al., 1992), flavonoids (Irizar et al., 1992, Olszewska and Wolbis 2002), norisoprenoid glycosides (Humpf and Schreier, 1992), and cyanogenic glycosides (prunasin; Gleadow and Woodrow 2002, Kumarasamy et al. 2003, Calatayud et al. 2013), which are all known to be relatively toxic (Zagrobelny et al. 2004). Some of these compounds may require peculiar detoxifying pathways, as shown by the example of co-evolution between the burnet moth (*Zygaena*) and their cyanogenic-glycosides-producing host plant Fabaceae (Rosid) (Zagrobelny et al. 2004, Zagrobelny et al. 2007, Takos et al. 2011).

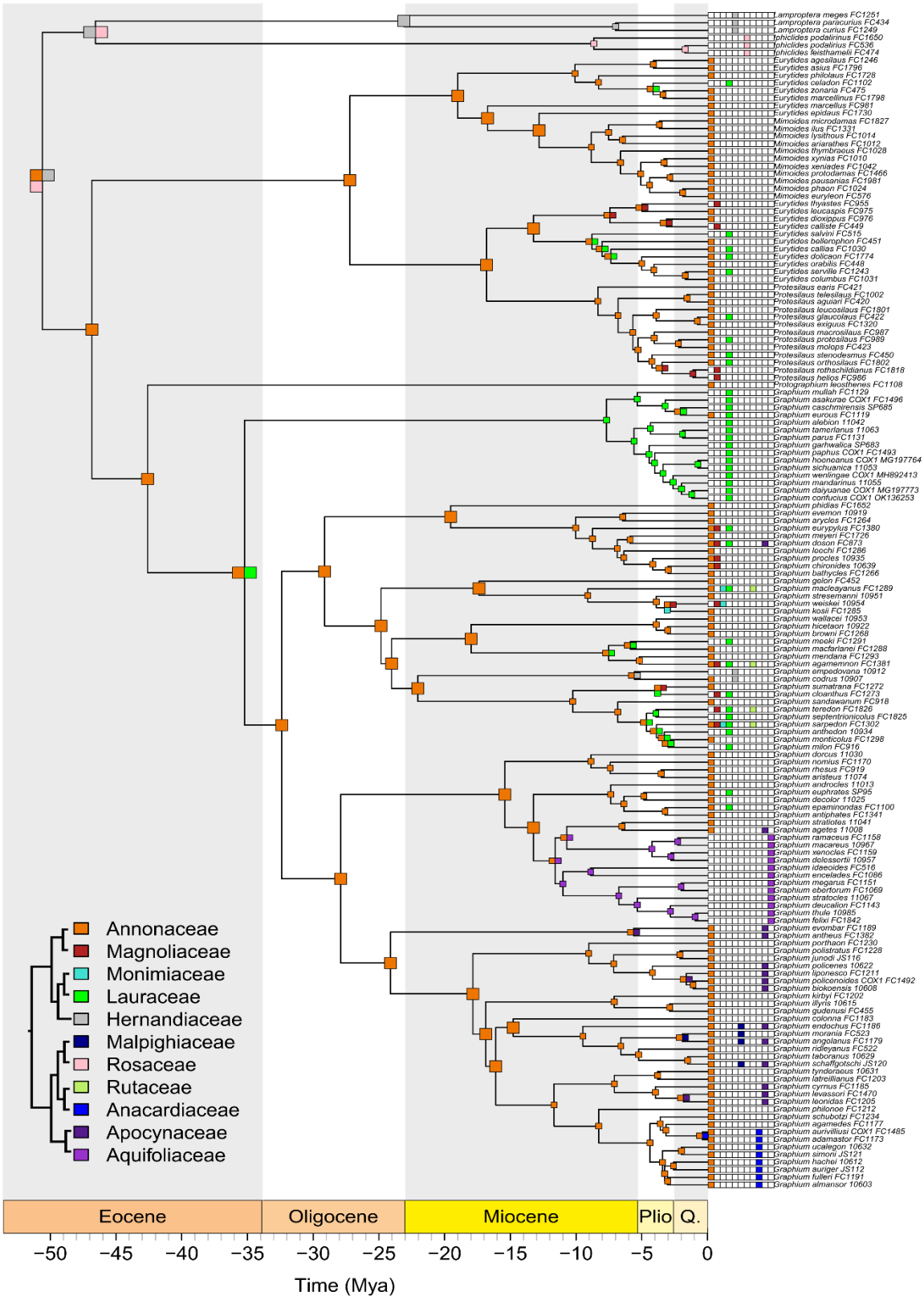


Figure 2: Dated phylogeny and ancestral state reconstruction of host–plant association for Leptcorcini. The outgroups are removed. The bottom left legend presents the 11 host-plants families and includes their phylogenetic relationships. The two distinct clades are Magnoliids (Annonaceae, Magnioliaceae, Monimiaceae, Lauraceae, Hernandiaceae) and Eudicots (Malpighiaceae, Rosaceae, Rutaceae, Anacardiaceae [Rosids] and Apocynaceae, Aquifoliaceae [Asterids])

Historical biogeography of Leptocircini and Troidini

For Leptocircini, the DECX analyses recovered an ancestral range including West Nearctic, Central America, East Palearctic, and continental Southeast Asia as the most likely ancestral geographic origin (relative probability = 0.210), when Asia and North America were connected by the Bering land bridge in the Eocene (**Figure 3**). The following best ancestral ranges differ only in the direction and extension around this Bering land bridge, as the second-best ancestral area was composed of West Nearctic, East Palearctic, and continental Southeast Asia and insular Southeast Asia (relative probability = 0.203), followed by West Nearctic, East Nearctic, East Palearctic and continental Southeast Asia (relative probability = 0.191). Hence, in all cases a Northern (Laurasian) origin is preferred over a Southern (Gondwanan) origin (as proposed by Hancock 1983). Similarly, the DECX analyses recovered a range including West Nearctic, Central America, East Palearctic, and Caribbean Islands as the most likely ancestral geographic origin for Troidini (relative probability = 0.1610) (**Figure 4**). West Nearctic, Central America, East Palearctic, and continental Southeast Asia was the most second likely ancestral geographic origin (relative probability = 0.150), followed by West Nearctic, East Nearctic, East Palearctic and continental Southeast Asia (relative probability = 0.191). This Laurasian origin (including continental Southeast Asia and Central America) had already been inferred for both tribes in previous studies (Condamine et al. 2012, Allio et al. 2021), or with sister clades such as Papilionini (Condamine et al. 2023), but contradicts the Gondwanian origin of Troidini inferred by Braby et al. (2005) and supported by Hancock (2007).

From this Laurasian region, both tribes disperse into southern regions (**Figure 3, Figure 4**). With their Laurasian origin extending as far as continental Southeast Asia, Leptocircini were inferred to first colonize insular Southeast Asia. In a second phase, the West Nearctic edge of their distribution extended to the East Nearctic, and the Neotropics (through the three Neotropical regions: Central America, South America and the Caribbean Islands) with the *Eurytides* clade. This colonization of the Neotropics is interesting because it displays uniform patterns of South American occupation for the *Eurytides* and *Protesilaus* subgenera, while numerous vicariance, extinction and dispersal events were inferred for the *Mimoides* subgenus (which today contains several species in Central America and the Caribbean). The origin of *Graphium* + *Protographium* is inferred in an area comprising Eastern Palearctic, Continental and Insular Southeast Asia. The dispersal of *Protographium* to Australia between 40 Ma and today probably happened through the Southeast Asian insular route, which would have involved a number of local extinctions. The Indian Peninsula was colonized by the ancestor of *Graphium*, in addition to their maintenance in continental and insular Southeast Asia. The *Pazala* would have lost their insularity to diversify

solely in the continental Asia, while *Graphium s.s.* and *Pathysa* colonized and mostly diversified into insular Southeast Asia, Australia and surrounding islands. Finally, the presence of *Graphium* in the Indian Peninsula may subsequently have constituted an opportunity of passageway for the colonization of Africa by the ancestors of *Arisbe* during the Oligocene. Some forest corridors have indeed been documented between India, the Arabian Peninsula and continental Africa at this period (Boureau et al. 1983, Toussaint et al. 2019), although massive interchanges of mammalian faunas, for instance, are usually documented around the Oligocene-Miocene transition (Antoine et al. 2003, Antoine et al. 2010, Gernelle et al. 2023).

According to the best DECX estimate for Troidini, the ancestral range of the tribe extended as far as Central America and the Caribbean Islands. According to our inference, it is from this region that subtribe Battina colonized South America and diversified within the Neotropics. The Troidina, on the other hand, have spread throughout the northern part of the world, occupying the whole Paleactic and West Nearctic. The colonization of Madagascar from this vast northern region by *Pharmacophagus* between 40 Ma and today remains enigmatic. It could have involved India, as hypothesized by several authors (Braby et al. 2005, Hancock 2007, Condamine et al. 2013b, Nakae 2021), but we could also hypothesize a colonization route via continental africa, although this has never been proposed to our knowledge. Indeed, none of these regions were inferred as part of the ancestral ranges to explain the colonization of Madagascar, but the expansion in West Palearctic for the ancestor of all Troidina in our analysis probably means that the model likely inferred a colonization route via continental Africa. This scenario would not only require connections between Eurasia and Africa since 40 Ma (which is plausible geologically and biologically at least *ca.* 20-30 Ma, see above), but also the presence of Aristolochiaceae in continental Africa within this timeframe. The latter information is scarcely documented. A dated phylogeny of Aristolochiaceae has recently been carried out, but the sample size was small, particularly for the African taxa (Allio et al. 2021). The timing of the appearance of Aristolochiaceae in Africa is therefore still poorly known and makes it difficult to support the colonization route of a troidine taxon to Madagascar via continental Africa.

The common ancestor of the rest of the subtribe Troidina was inferred to be in a range that still includes the West Nearctic and the East Palearctic, as well as continental and insular Southeast Asia. From this large ancestral range, the ancestor of the genera *Atrophaneura* and *Pachliopta* would have initially reduced its range to continental Southeast Asia only, before diversifying, through a large number of dispersal and vicariance events, into the Indian Peninsula, Continental and Insular Southeast Asia (and as far as Australia for the subgenus *Pachliopta*). At the same time, a vicariance event separated the birdwing butterflies' ancestor, which remained in the islands of Southeast Asia, from the common ancestor of *Parides*,

Euryades and *Cressida*, which would have also extended their range to Central America at the same time as this vicariance event. It was from this area that a new colonization of Central America took place for the common ancestor of *Euryades* and *Parides*, while *Cressida* dispersed towards Australia within a time span of 25 million years, probably passing again through insular Southeast Asia. It is interesting to note that the two colonisations of South America by the Troidini, as well as the one of Leptocircini, all took place in a short period between the middle Oligocene and the early Miocene, which corresponds to the onset of temporary land bridges between North and South America (Cornée et al. 2021, Garrocq et al. 2021). Finally, the diversification of the birdwings is similar to the analysis by Condamine et al. (2015), although less finely divided, with a vicariance between the *Ornithoptera*, which diversified in Australia, Papua New Guinea and the surrounding islands, and the *Troides*, which diversified in insular Southeast Asia and, in a second time, extended their distribution northwards, returning to Continental Asia and India.

Overall, we note that the biogeographic patterns of the two tribes have been quite similar, resulting in sympatric evolution of the two tribes to a large extent, as far as our regional divisions can tell. Rather than a particular effect of the tribes, we found that the biogeographic patterns of diversification were more area-driven than clade-driven, as both tribes presented similar percentages in each region. Furthermore, more than half of the vicariance events (24/36 = 67%, **Table 1**) in the two tribes took place in the Indo-Australasian Archipelago and particularly between the continental and insular regions of Southeast Asia (e.g. *Ornithoptera* and *Troides*). On the other hand, northern regions have witnessed numerous local extinctions, both in the deep nodes of the two phylogenies and in more recent clades. The *Pazala*, which diversified in the region at the boundary of East Paleartic, India, and south-east Asia, was inferred to have had 8 local extinctions for 28 branches (**Figure 3**). The backbone included many local extinction events (16/73 = 22%) whereas it represents only 4% of the total number of branches.

Table 1: Biogeographic processes of Leptocircini and Troidini, which occurred in their backbone or in the different tropical regions (Neotropics, Indo Australasian Archipelago (APP) and Continental Africa. The lines correspond to the number of branches involved in the region, for Leptocircini, Troidini, or both, and the number of vicariance, dispersal and local extinction events along the branches. For each column, the first sub-column is the number of events (or number of branches) and the second sub-column corresponds to the proportion of this number compared to the total number of the same type of events, totalling the two tribes. For example, 55% of the total number of branches (330/605) took place in the IAA, for 67% of the total number of vicariance events (24/36). Note that for Leptocini, the clade Iphiclides was included in the backbone.

		Backbone		**Neotropics**		**Indo Australasian Archipelago**		**Continental Africa**		**Total**	
Leptocircni	#branch	14	2%	84	14%	148	24%	68	11%	**314**	52%
	vicariance	1	3%	4	11%	8	22%	3	8%	**16**	44%
	dispersion	9	12%	8	11%	23	32%	1	1%	**41**	56%
	local extinction	9	15%	7	11%	19	31%	1	2%	**36**	59%
Troidini	#branch	9	1%	100	17%	182	30%			**291**	48%
	vicariance	1	3%	3	8%	16	44%			**20**	56%
	dispersion	7	10%	14	19%	11	15%			**32**	44%
	local extinction	7	11%	8	13%	10	16%			**25**	41%
Total	**#branch**	**23**	4%	**184**	30%	**330**	55%	**68**	11%	**605**	100%
	vicariance	**2**	6%	**7**	19%	**24**	67%	**3**	8%	**36**	100%
	dispersion	**16**	22%	**22**	30%	**34**	47%	**1**	1%	**73**	100%
	local extinction	**16**	26%	**15**	25%	**29**	48%	**1**	2%	**61**	100%

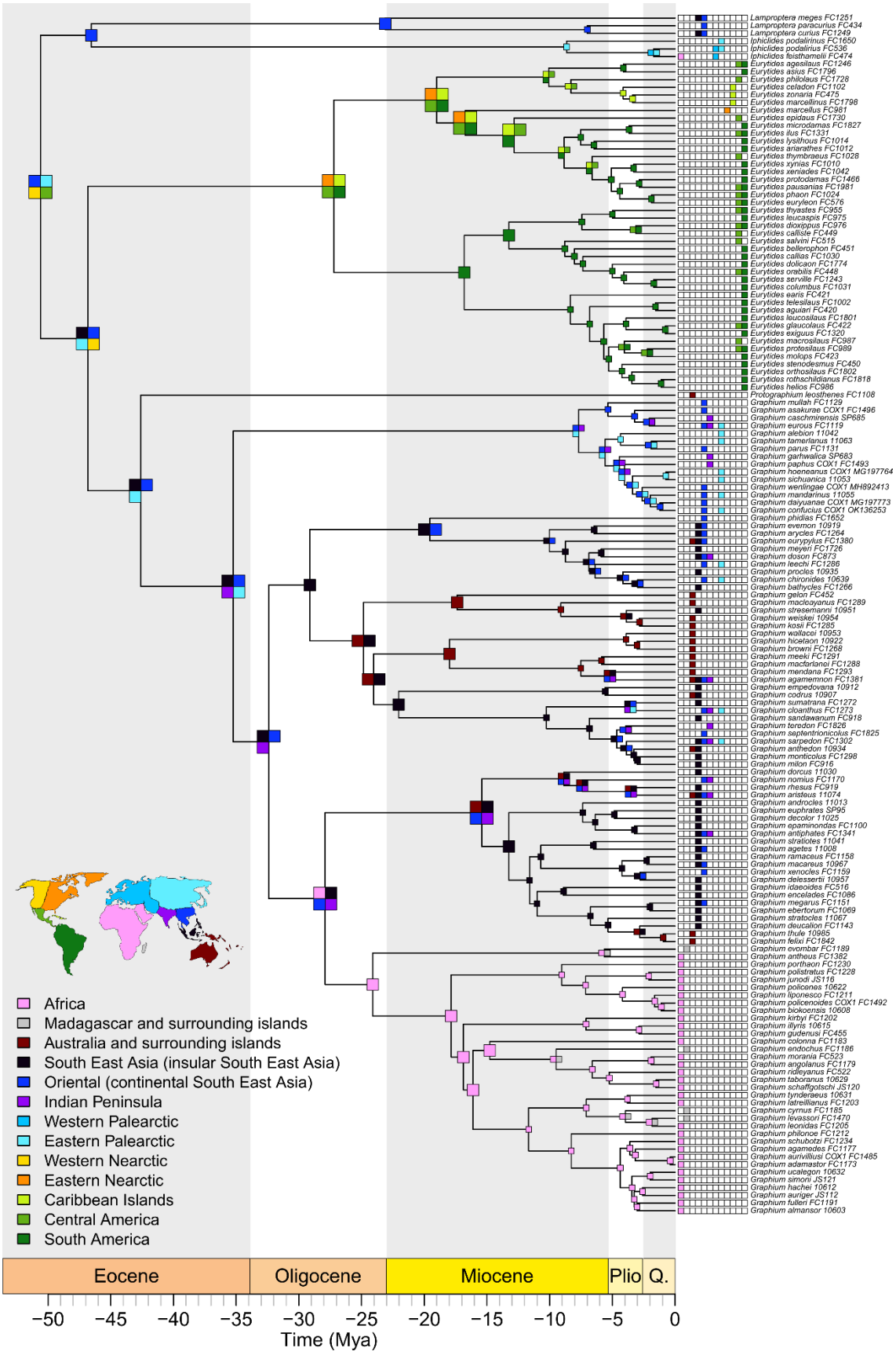


Figure 3: Dated phylogeny and historical biogeography of Leptocircini. The bottom-left corner map represents present-day delimitation of the world in 13 areas. Current distribution of each species is displayed at the tips. The outgroups were removed before the DECX analysis.

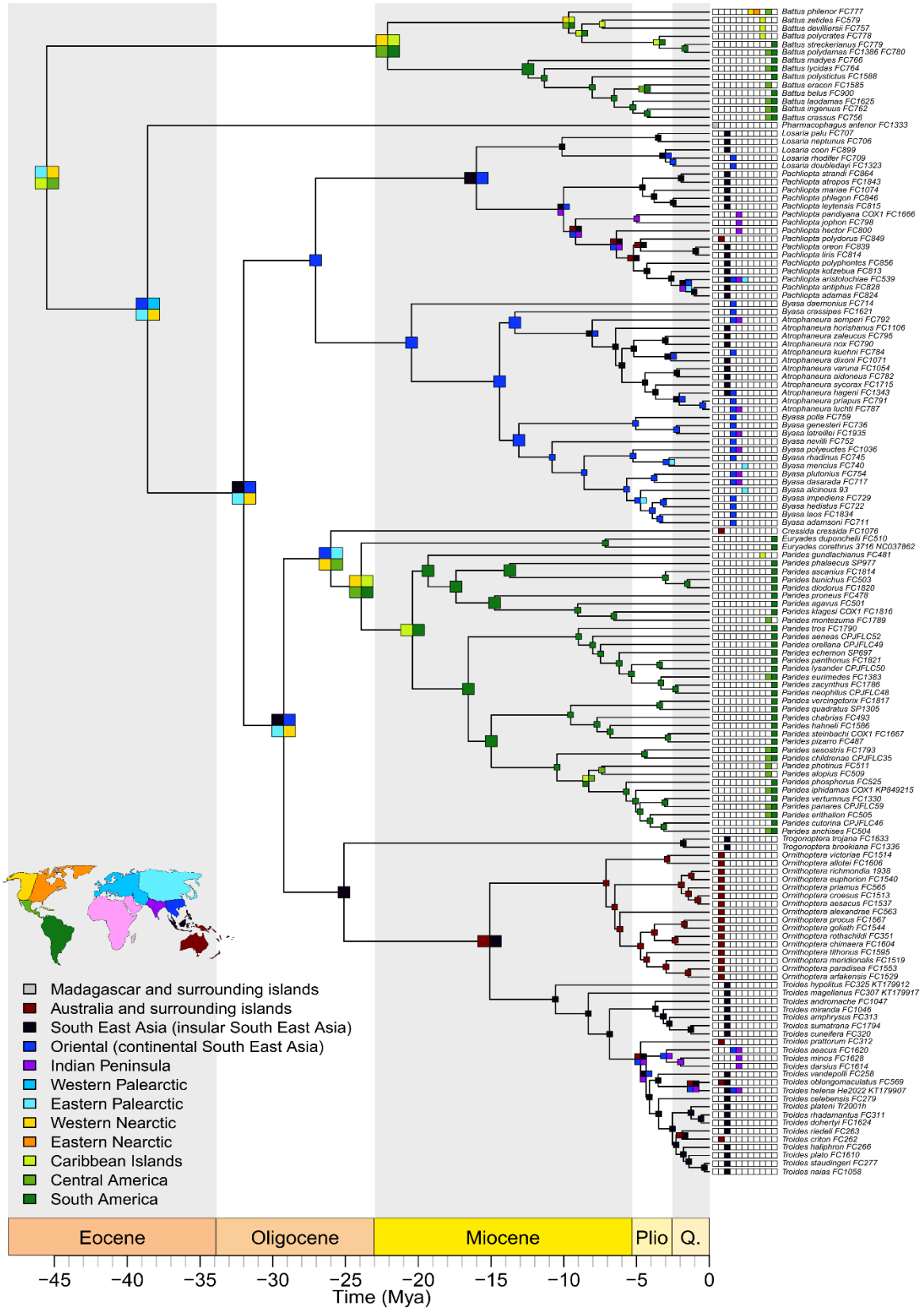


Figure 4: Dated phylogeny and historical biogeography of Troidini. The bottom-left corner map represents present-day delimitation of the world in 13 areas. Current distribution of each species is displayed at the tips. The outgroups were removed before the DECX analysis. Note that Africa was present in DECX matrices but was not present in the present-day distribution of species and was not inferred at any node.

Diversification of Leptocircini and Troidini

In all analyses, ClaDS and RevBayes converged well as informed by the ESS scores. We also note that ClaDS inferred higher speciation rates than RevBayes in both tribes, but the pattern of rate variation among branches was very consistent between the two methods (**Figure 5, Figure 6**).

For Leptocircini, using the complete taxonomy dataset, the subgenera *Protesilaus*, *Pazala* and *Arisbe* (*adamastor* group especially), and the mimetic *Pathysa* (the former *Paranticopsis*, including *stratiotes* and *agetes*, see **Chapter 2.1**) had the highest speciation rates (*ca.* 0.17-0.20 events per million years per lineage [events/Myr/lineage]) in the ClaDS analysis (**Figure 5**). The species-poor clades (*Lamproptera*, *Iphiclides* and *Protographium*) had very low speciation rates inferred (*ca.* 0.06 events/Myr/lineage). This trend was the same for the RevBayes analysis of the same dataset, although the speciation rates range was narrower, and overall lower (*ca.* 0.07-0.08 for the lowest, *ca.* 0.13-0.14 for the highest; **Figure 5**). Interestingly, the extinction rate was relatively low and equal across the tree (*ca.* 0.0001-0.0002 events/Myr/lineage). When the same analyses were applied on the lump taxonomy dataset, the range of speciation rate was particularly reduced in RevBayes (0.10-0.13), but remains approximately the same in ClaDS (0.06-0.18) (**Figure S5**). The speciation rates of *Pazala* were reduced compared to the complete taxonomy (*ca.* 0.10-0.12 events/Myr/lineage in ClaDS; 0.12 events/Myr/lineage in RevBayes), and not particularly detected as a shift compared to the backbone, for both ClaDS and RevBayes (**Figure S5)**. With this taxonomy, *Arisbe* (*adamastor* group especially) had the highest rate (*ca.* 0.18 events/Myr/lineage in ClaDS; *ca.* 0.13 events/Myr/lineage in RevBayes). The increase in speciation rate was still a bit higher than the backbone for *Protesilaus*, but it seems even with the mimetic *Mimoides* (*ca.* 0.15 events/Myr/lineage in ClaDS, *ca.* 0.12 events/Myr/lineage in RevBayes, for both clades). Finally, contrary to the complete taxonomy analyses, the increase in speciation was not concentrated on the mimetic *Pathysa* only but rather on all *Pathysa* (**Figure S5**).

According to the ClaDS and RevBayes results, the following clades were specified as *a priori* shifts of diversification to be tested with RPANDA: (1) mimetic *Mimoides*, (2) *Protesilaus*, (3) *Pazala*, (4) all *Pathysa*, (5) mimetic *Pathysa*, including *Graphium stratiotes* and *G. agetes*, and (6) *Arisbe*, excluding *G. antheus* and *G. evombar*. With the complete taxonomy, the best combination of diversification shifts in RPANDA recovered a shift for all subclades except for mimetic *Pathysa* (as the shift was inferred to all *Pathysa*, **Figure S6**). The clades that experienced a shift in diversification were best explained either by a pure-birth model (BCST; *Protesilaus* and *Arisbe*) or by a model with a decreasing speciation rate through time (BVAR; mimetic *Mimoides*, *Pazala*, all *Pathysa*). The backbone tree

followed a constant birth-death model (BCST-DCST). When applied to the lump taxonomy, the shift for the *Pazala* clade was not retrieved, and the backbone followed a pure-birth model (**Figure S7**).

In Troidini, a wider range of speciation rates is inferred than in Leptocircini for both ClaDS and RevBayes. The extinction rate inferred by RevBayes is low and consistent with that of Leptocircini (*ca.* 0.0001-0.0002 events/Myr/lineage). Using the complete taxonomy dataset, the genus *Troides* had the highest rates (*ca.* 0.36-0.39 events/Myr/lineage in ClaDS; *ca.* 0.21-0.22 events/Myr/lineage in RevBayes), followed by *Ornithoptera* (*ca.* 0.30 events/Myr/lineage in ClaDS; *ca.* 0.20 events/Myr/lineage in RevBayes), and the subgenera *Atrophaneura, Byasa*, and *Pachliopta* (*ca.* 0.20 events/Myr/lineage in ClaDS; *ca.* 0.15-0.17 events/Myr/lineage in RevBayes). With the lump taxonomy, the rates for *Troides* and *Ornithoptera* were reduced but still higher than the other clades (*ca.* 0.24-0.27 events/Myr/lineage in ClaDS; *ca.* 0.17-0.18 events/Myr/lineage in RevBayes; **Figure S8**).

Accordingly, the following troidine clades were specified to RPANDA to be tested as significant diversification shifts: (1) subgenera *Atrophaneura* + *Byasa*, (2) subgenus *Atrophaneura*, including *Byasa crassipes*, (3) subgenus *Byasa*, (4) genera *Ornithoptera* + *Troides*, (5) genus *Ornithoptera*, and (6) genus *Troides*. With the complete taxonomy, the best combination in RPANDA inferred a shift for the subclades *Pachliopta* (BCST model), *Atrophaneura* + *Byasa* (BVAR model with a decreasing speciation rate through time), and *Ornithoptera* + *Troides* (BCST model). The backbone tree followed a pure-birth model (BCST model) (**Figure S9**). The result was the same for the lump taxonomy (**Figure S10**).

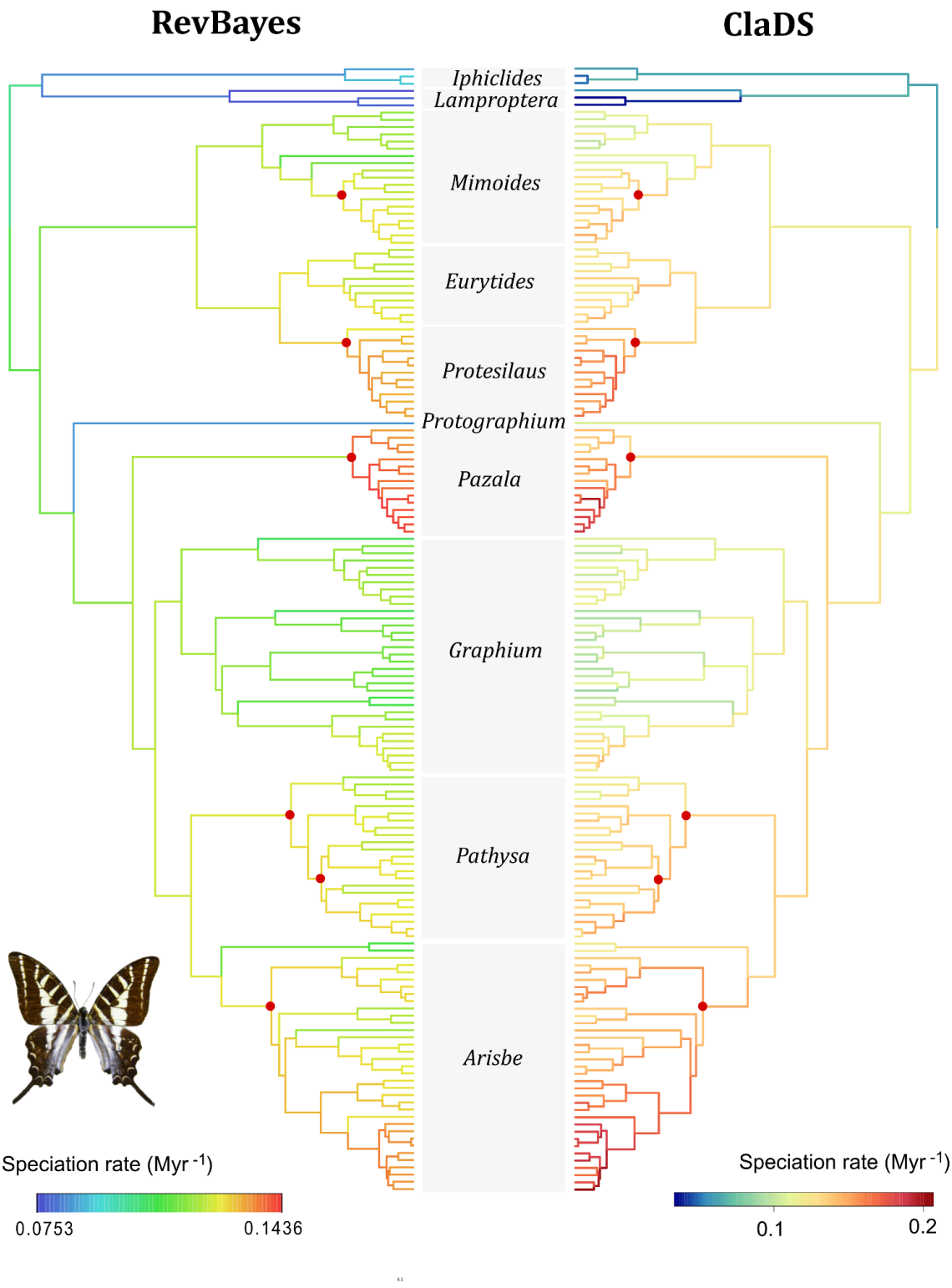


Figure 5: RevBayes (left) and ClaDS (right) results on the complete taxonomy dataset of Leptocircini. Each analysis has a different scale, reported below the tree. Red points indicate the tested subclade given to RPANDA. Illustration: *Graphium aristeus*, Alex Dumchus.

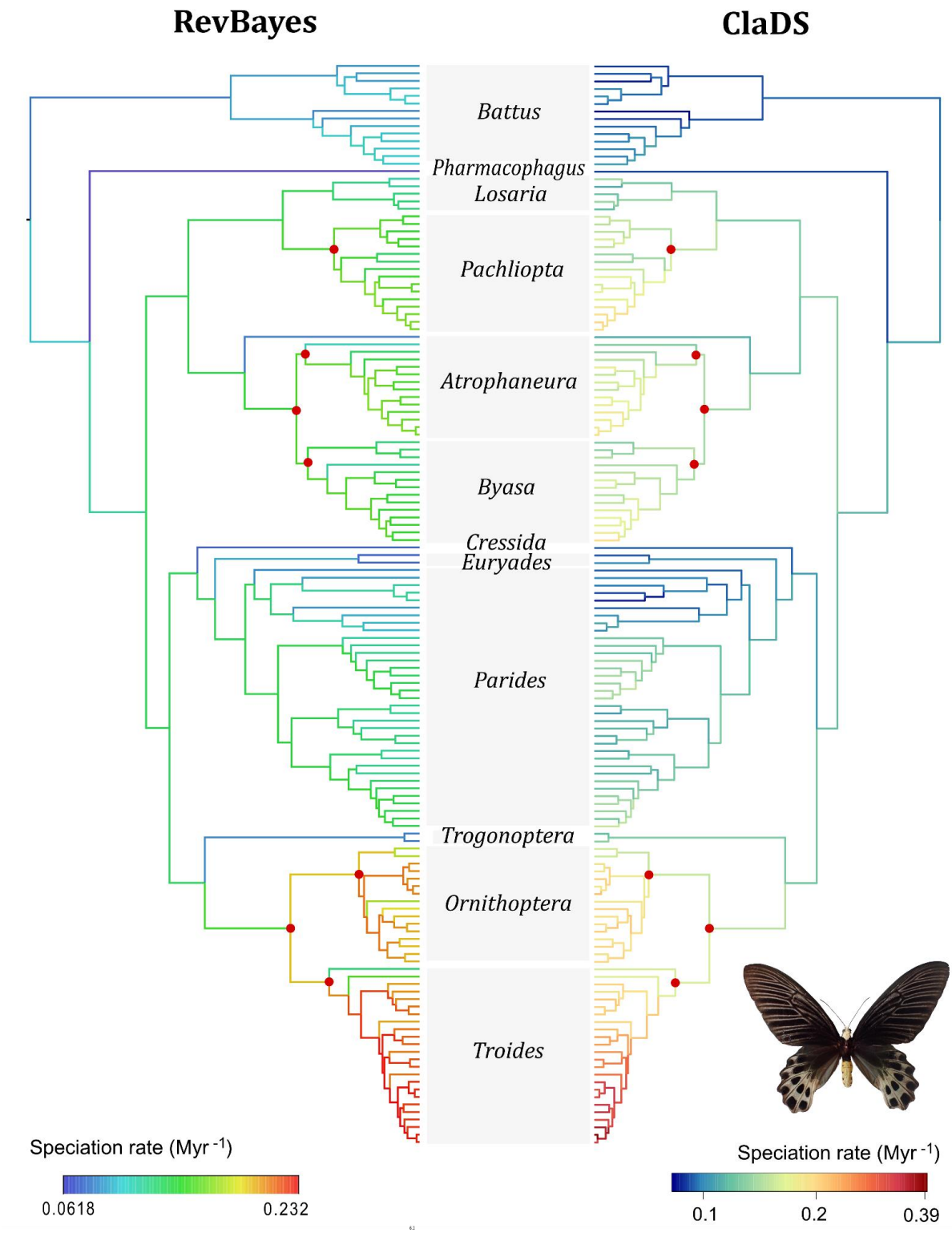


Figure 6: RevBayes (left) and ClaDS (right) results on the complete taxonomy dataset of Troidini. Each analysis has a different scale, reported below the tree. Red points indicate the tested subclade given to RPANDA. Illustration: *Atrophaneura sycorax*, Eliette Reboud.

Synthesis of the biogeographic and diversification history

One of the aims of the macroevolutionary comparison of the phylogenies of Leptocircini and Troidini was to be able to disentangle some confounding effects of the host-plant shift, particularly that of biogeography.

Overall, no clear pattern of the effect of host-plant shift on the diversification has been observed in the studied Papilionidae. Leptocircini have undergone several changes of host plant, some changing their specialization, such as the *Pazala* on Lauraceae or the mimetic *Pathysa* on Aquifoliaceae, or others broadening their host breadth, such as the subgenera *Graphium* and the *Arisbe*. The *Pazala* presented a shift of speciation rates in the complete-taxonomy analysis, but this effect was almost nonexistent when a stringent taxonomy was applied to this clade. Furthermore, the mimetic *Pathysa* clade does appear to have increased diversity in the branch-specific diversification analyses, but the clade-specific analysis seems to support a shift at the crown of all *Pathysa* rather than at the crown of mimetic *Pathysa* (and Aquifoliaceae-feeder) alone. Therefore we do not exactly verify the "musical chair" model prediction which states that speciation rates are positively correlated with host switching rate (Hardy and Otto 2014, Jousselin and Elias 2019). On the other hand, host expansion could have been put forward as a driver for the diversification of the *Arisbe* and in particular the *adamastor* group, but this does not seem to have been particularly the case for the subgenus *Graphium*, which also acquired a broader host breadth than their ancestor. Thus, on the scale of the Leptocircini, the change in host plant did not seem to be a predominant driver of their diversification. That said, the Lauraceae and Annonaceae were among the lowest net diversifiers in the analysis by Allio et al. (2021) at the global scale of the Papilionidae (compared with the shift toward Apiaceae, Papaveraceae and Crassulaceae, for example).

Interestingly, the two tribes showed very similar geographic patterns. Diversification in the islands, particularly in Wallacea, seems to have played an important role in the diversification of the Troidini, since the highest rates of speciation were found in all the island clades such as the birdwing butterflies and *Atrophaneura* and *Pachliopta sensu lato*. This same pattern has been highlighted with *Papilio* (Condamine et al. 2013c, Joshi and Kunte 2022, Condamine et al. 2023). Yet, some clades that presented similar biogeographic patterns, such as the subgenus *Graphium*, did not particularly have such an increase in speciation rates. Some studies suggest that the *sarpedon* group comprises more species than the one included in this study (Cotton et al. 2022). But even so, this would only concern very recent nodes of the *Graphium*, and this clade would still present very long branches from its origin in insular South east Asia in the middle-late Oligocene.

On another note, Leptocircini contains two clades which present batesian mimicry (of other toxic groups), and it could have been hypothesized that this mimicry had been a driver of their

diversification (Jiggins et al. 2006, Elias et al. 2008). Here, the mimicry of *Pathysa* and *Mimoides* does not seem to have clearly led to a burst of speciation for these particular clades, although the groups in which each of these clades have overall higher speciation rates than the backbone tree. Yet here, the uncertainty around rates of speciation could come from a lack of statistical power to detect a significant shift of diversification (because these clades only include 13 and 11 species, respectively) or from taxonomic uncertainties in the case of *Pathysa* (see **Chapter 2.1**).

Finally, the analyses carried out on the two tribes have shown again how complex these issues remain to assess. Several points seem to emerge from this study. Firstly, the importance of taxonomic delimitation seems crucial. The diversification analysis of Leptocircini with the complete taxonomy indicated changes in speciation in the *adamastor* group (*Arisbe*), in *Pazala* and in *Protesilaus*. These three clades are precisely the clades for which taxonomy is particularly delicate and subject to taxonomic inflation (as discussed in **Chapter 2.1**; see also Hancock 2006, Huang 2023). By applying a more stringent "lump" taxonomy, we found much less marked shifts for the last three clades mentioned. The discussion about the dependence of macroevolutionary analyses on taxonomy and species delimitation is growing increasingly important in the literature in the recent years as it shows some drastic effect of taxonomy on diversification analyses (Ruane et al. 2014, Faurby et al. 2016, Utami et al. 2022).

Acknowledgments

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (project GAIA, agreement no. 851188).

References

Ahrens D, Schwarzer J, Vogler AP. 2014. The evolution of scarab beetles tracks the sequential rise of angiosperms and mammals. *Proc. R. Soc. B: Biol. Sci.* **281**:20141470.

Allio R, Nabholz B, Wanke S, Chomicki G, Pérez-Escobar OA, Cotton AM, Clamens AL, Kergoat GJ, Sperling FAH, Condamine FL. 2021. Genome-wide macroevolutionary signatures of key innovations in butterflies colonizing new host plants. *Nat. Commun.* **12**:354.

Allio R, Schomaker-Bastos A, Romiguier J, Prosdocimi F, Nabholz B, Delsuc F. 2020b. MitoFinder: efficient automated large-scale extraction of mitogenomic data in target enrichment phylogenomics. *Mol. Ecol. Res*. **20**:892–905.

Allio R, Scornavacca C, Nabholz B, Clamens AL, Sperling FA, Condamine FL. 2020a. Whole genome shotgun phylogenomics resolves the pattern and timing of swallowtail butterfly evolution. *Syst. Biol*. **69**:38-60.

Almeida EA, Bossert S, Danforth BN, Porto DS, Freitas FV, Davis CC, *et al.* 2023. The evolutionary history of bees in time and space. *Curr. Biol.* **33**:3409-3422.

Almeida EA, Packer L, Melo GA, Danforth BN, Cardinal SC, Quinteiro FB, Pie MR. 2019. The diversification of neopasiphaeine bees during the Cenozoic (Hymenoptera: Colletidae). *Zoologica Scripta* **48**:226-242.

Antoine PO, Downing KF, Crochet JY, Duranthon F, Flynn LJ, Marivaux L, *et al.* 2010. A revision of *Aceratherium blanfordi* Lydekker, 1884 (Mammalia: Rhinocerotidae) from the Early Miocene of Pakistan: postcranials as a key. *Zool. J. Linn. Soc.* **160**:139-194.

Antoine PO, Welcomme JL, Marivaux L, Baloch I, Benammi M, Tassy P. 2003. First record of Paleogene Elephantoidea (Mammalia, Proboscidea) from the Bugti Hills of Pakistan. *J. Vert. Paleontol.* **23**:977-980.

Barnosky AD. 2001. Distinguishing the effects of the Red Queen and Court Jester on Miocene mammal evolution in the northern Rocky Mountains. *J. Vert. Paleontol.* **21**:172-185.

Barthlott W, Hostert A, Kier G, Kueper W, Kreft H, Mutke J, Daud Rafiqpoor M, Sommer JH. 2007. Geographic patterns of vascular plant diversity at continental to global scales (Geographische Muster der Gefäßpflanzenvielfalt im kontinentalen und globalen Maßstab). *Erdkunde* 305-315.

Beccaloni GW, Viloria AL, Hall SK, Robinson GS. 2008. *Catalogue of the Hostplants of the Neotropical Butterflies*. Sociedad Entomológica Aragonesa.

Becerra JX, Venable DL. 1999. Macroevolution of insect–plant associations: the relevance of host biogeography to host affiliation. *Proc. Natl Acad. Sci. USA* **96**:12626-12631.

Beeravolu CR, Condamine FL. 2016. An extended maximum likelihood inference of geographic range evolution by dispersal, local extinction and cladogenesis. *bioRxiv* https://doi.org/10.1101/038695

Bentley KW. 2006. β-Phenylethylamines and the isoquinoline alkaloids. *Natural Prod. Rep.* 23:444-463.

Benton MJ, Emerson BC. 2007. How did life become so diverse? The dynamics of diversification according to the fossil record and molecular phylogenetics. *Palaeontol.* **50**:23-40.

Benton MJ. 2009. The Red Queen and the Court Jester: species diversity and the role of biotic and abiotic factors through time. *Science* **323**:728–732.

Benton MJ. 2016. Origins of biodiversity. *PLoS Biol.* **14**:e2000724.

Berenbaum M. 1983. Coumarins and caterpillars: a case for coevolution. *Evolution* **37**:163-179.

Berenbaum MR, Favret C, Schuler MA. 1996. On defining "key innovations" in an adaptive radiation: Cytochrome P450s and papilionidae. *Am. Nat.* **148**:S139-S155.

Blaimer BB, Santos BF, Cruaud A, Gates MW, Kula RR, Mikó I, Rasplus JY, Smith DR, Talamas EJ, Brady SG, *et al.* 2023. Key innovations and the diversification of Hymenoptera. *Nat. Commun.* **14**:1212.

Blakey RC. 2008. Gondwana paleogeography from assembly to breakup—A 500 my odyssey. *Geol. Soc. Am. Special Papers* **441**:1-28.

Bossert S, Wood TJ, Patiny S, Michez D, Almeida EA, Minckley RL, Packer L, Neff JL, Copeland RS, Straka J, *et al.* 2022. Phylogeny, biogeography and diversification of the mining bee family Andrenidae. *Syst. Entomol.* **47**:283-302.

Boureau E, Cheboldaeff-Salard M, Koeniguer JC, Louvet P. 1983. Evolution des flores et de la végétation Tertiaires en Afrique, au nord de l'Equateur. *Bothalia* **14**:355-367.

Braby MF, Trueman JW, Eastwood R. 2005. When and where did troidine butterflies (Lepidoptera: Papilionidae) evolve? Phylogenetic and biogeographic evidence suggests an origin in remnant Gondwana in the Late Cretaceous. *Invertebrate Syst.* **19**:113-143.

Brown KS, Klitzke CF, Berlingeri C, Dos Santos PER, Scriber JM, Tsubaki Y, Lederhouse RC. 1995. Neotropical swallowtails: chemistry of food plant relationships, population ecology, and biosystematics. In: *Swallowtail Butterflies: Their Ecology and Evolutionary Biology* (Eds. Scriber JM, Tsubaki Y, Lederhouse RC). Scientific Publishers, Gainesville, Florida. pp. 405-446.

Calatayud PA, Desneu, N, Le Gall P. 2013. Caractéristiques chimiques des plantes. *Interactions insectes-plantes*, 217-228.

Cardinal S, Danforth BN 2013. Bees diversified in the age of eudicots. *Proc. R. Soc. B: Biol. Sci.* **280**:20122686.

Chacón J, Renner SS. 2014. Assessing model sensitivity in ancestral area reconstruction using Lagrange: a case study using the Colchicaceae family. *J. Biogeogr.* **41**:1414-1427.

Chazot N, Willmott KR, Lamas G, Freitas AVL, Piron-Prunier F, Arias CF, Mallet J, De-Silva DL, Elias M. 2019. Renewed diversification following Miocene landscape turnover in a Neotropical butterfly radiation. *Global Ecol. Biogeogr.* **28**:1118-1132.

Chen JJ, Hung HC, Sung PJ, Chen IS, Kuo WL. 2011. Aporphine alkaloids and cytotoxic lignans from the roots of *Illigera luzonensis*. *Phytochem.* **72**:523-532.

Chernomor O, Von Haeseler A, Minh BQ. 2016. Terrace aware data structure for phylogenomic inference from supermatrices. *Syst. Biol*. **65**:997-1008.

Chotard A, Ledamoisel J, Decamps T, Herrel A, Chaine AS, Llaurens V, Debat V. 2022. Evidence of attack deflection suggests adaptive evolution of wing tails in butterflies. *Proc. R. Soc. B: Biol. Sci.* **289**:20220562.

Chotard A. 2022. Evolution of tails in swallowtail butterflies (Papilionidae, Lepidoptera): macro-evolutionary and experimental approaches. Doctoral dissertation, Muséum National d'Histoire Naturelle.

Chou I. 1994. *Monographia Rhopalocerorum Sinensium*. Henan Scientific and Technological Publishing House, Zhengzhou, 2 vol., 854 pp.

Chouvenc T, Šobotník J, Engel MS, Bourguignon T. 2021. Termite evolution: mutualistic associations, key innovations, and the rise of Termitidae. *Cell. Mol. Life Sci.* **78**:2749-2769.

Collins NM, Morris MG. 1985. *Threatened Swallowtail Butterflies of the World*. The IUCN Red Data Book, Cambridge.

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

Condamine FL, Allio R, Reboud EL, Dupuis JR, Toussaint EFA, Mazet N, *et al.* 2023. A comprehensive phylogeny and revised taxonomy illuminate the origin and diversification of the global radiation of *Papilio* (Lepidoptera: Papilionidae). *Mol. Phylogenet. Evol.* **183**:107758.

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

Condamine FL, Rolland J, Morlon H. 2019a. Assessing the causes of diversification slowdowns: temperature-dependent and diversity-dependent models receive equivalent support. *Ecol. Lett.* **22**:1900-1912.

Condamine FL, Romieu J, Guinot G. 2019b. Climate cooling and clade competition likely drove the decline of lamniform sharks. *Proc. Natl Acad. Sci. USA* **116**:20584-20590.

Condamine FL, Silvestro D, Koppelhus EB, Antonelli A. 2020. The rise of angiosperms pushed conifers to decline during global cooling. *Proc. Natl Acad. Sci. USA* **117**:28867-28875.

Condamine FL, Soldati L, Clamens AL, Rasplus JY, Kergoat GJ. 2013a. Diversification patterns and processes of wingless endemic insects in the Mediterranean Basin: historical biogeography of the genus *Blaps* (Coleoptera: Tenebrionidae). *J. Biogeogr.* **40**:1899-1913.

Condamine FL, Sperling FAH, Kergoat GJ. 2013b. Global biogeographic pattern in swallowtail diversification demonstrates alternate colonization routes in Northern and Southern Hemispheres. *J. Biogeogr*. **40**: 9–23.

Condamine FL, Sperling FAH, Wahlberg N, Rasplus JY, Kergoat GJ. 2012. What caused the latitudinal gradient of species diversity in swallowtail butterflies? *Ecol. Lett.* **15**:267–277.

Condamine FL, Toussaint EFA, Clamens AL, Genson G, Sperling FAH, Kergoat GJ. 2015. Deciphering the evolution of birdwing butterflies 150 years after Alfred Russel Wallace. *Sci. Rep.* **5**:11860.

Condamine FL, Toussaint EFA, Cotton AM, Genson GS, Sperling FAH, Kergoat GJ. 2013c. Fine-scale biogeographic and temporal diversification processes of peacock swallowtails (*Papilio* subgenus *Achillides*) in the Indo-Australian Archipelago. *Cladistics* **29**:88–111.

Cornée JJ, Münch P, Philippon M, Boudagher-Fadel M, Quillévéré F, Melinte-Dobrinescu M, *et al.* 2021. Lost islands in the northern Lesser Antilles: possible milestones in the Cenozoic dispersal of terrestrial organisms between South-America and the Greater Antilles. *Earth-Sci. Rev.* **217**:103617.

Cotton AM, Doleck T, Zhang XIN, Inayoshi Y, Lohman DJ, Hu SJ. 2022. *Graphium septentrionicolus* Page & Treadaway, 2013 (Lepidoptera: Papilionidae) is a distinct species. *Zootaxa* **5154**:211-224.

Coulerie P, Poullain C. 2016. New Caledonia: A 'Hot Spot' for valuable chemodiversity: Part 2: basal angiosperms and eudicot rosids. *Chem. Biodiv.* **13**:18-36.

Cruaud A, Rønsted N, Chantarasuwan B, Chou LS, Clement WL, Couloux A, *et al.* 2012. An extreme case of plant–insect codiversification: figs and fig-pollinating wasps. *Syst. Biol.* **61**:1029-1047.

Cui X, Meng F, Pan X, Qiu X, Zhang S, Li C, Lu S. 2022. Chromosome-level genome assembly of *Aristolochia contorta* provides insights into the biosynthesis of benzylisoquinoline alkaloids and aristolochic acids. *Horticulture Res.* **9**:uhac005.

De Queiroz A. 1998. Interpreting sister-group tests of key innovation hypotheses. *Syst. Biol.* **47**:710-718.

Dimarco RD, Fordyce JA. 2013. Larger clutches of chemically defended butterflies reduce egg mortality: evidence from *Battus philenor*. *Ecological Entomol.* **38**:535-538.

dos Reis M, Álvarez-Carretero S, Yang Z. 2017. MCMCtree tutorials. Available from: http://abacus.gene.ucl.ac.uk/software/MCMCtree.Tutorials.pdf

dos Reis M, Donoghue PC, Yang Z. 2016. Bayesian molecular clock dating of species divergences in the genomics era. *Nat. Rev. Genet.* **17**:71-80.

dos Reis M, Yang Z. 2011. Approximate likelihood calculation on a phylogeny for Bayesian estimation of divergence times. *Mol. Biol. Evol*. **28**:2161–2172.

Drummond AJ, Ho SYW, Phillips MJ, Rambaut A. 2006. Relaxed phylogenetics and dating with confidence. *PLoS Biol.* **4**:e88.

Ehrlich PR, Raven PH. 1964. Butterflies and plants: a study in coevolution. *Evolution* **18**:586-608.

Elias M, Gompert Z, Jiggins C, Willmott K. 2008. Mutualistic interactions drive ecological niche convergence in a diverse butterfly community. *PLoS Biol.* **6**:e300.

Farrell B, Mitter C. 1990. Phylogenesis of insect/plant interactions: have *Phyllobrotica* leaf beetles (Chrysomelidae) and the Lamiales diversified in parallel? *Evolution* **44**:1389-1403.

Faurby S, Eiserhardt WL, Svenning JC. 2016. Strong effects of variation in taxonomic opinion on diversification analyses. *Meth. Ecol. Evol.* **7**:4-13.

Feeny P. 1991. Chemical constraints on the evolution of swallowtail butterflies. In: *Plant-Animal Interactions: Evolutionary Ecology in Tropical and Temperate Regions* (Eds. Price PW, Lewinsohn TM, Fernandes GW, Benson WW). New York, Wiley, pp. 315-340.

Fordyce JA. 2003. Aggregative feeding of pipevine swallowtail larvae enhances hostplant suitability. *Oecologia* **135**:250-257.

Fordyce JA. 2003. Aggregative feeding of pipevine swallowtail larvae enhances hostplant suitability. *Oecologia*, **135**:250-257.

Futuyma DJ, McCafferty SS. 1990. Phylogeny and the evolution of host plant associations in the leaf beetle genus *Ophraella* (Coleoptera, Chrysomelidae). *Evolution* **44**:1885-1913.

Garrocq C, Lallemand S, Marcaillou B, Lebrun JF, Padron C, Klingelhoefer F, *et al.* 2021. Genetic relations between the Aves Ridge and the Grenada back-arc basin, East Caribbean Sea. *J. Geophys. Res.: Solid Earth* **126**:e2020JB020466.

Garzón-Orduña IJ, Silva-Brandão KL, Willmott KR, Freitas AV, Brower AV. 2015. Incompatible ages for clearwing butterflies based on alternative secondary calibrations. *Syst. Biol.* **64**:752-767.

Gernelle K, Lihoreau F, Boisserie JR, Marivaux L, Métais G, Antoine PO. 2023. New material of *Parabrachyodus hyopotamoides* from Samane Nala, Bugti Hills (Pakistan) and the origin of Merycopotamini (Mammalia: Hippopotamoidea). *Zool. J. Linn. Soc.* **198**:278–309.

Gleadow RM, Woodrow IE. 2002. Mini-review: constraints on effectiveness of cyanogenic glycosides in herbivore defense. *J. Chem. Ecol.* **28**:1301-1313.

Griese E, Caarls L, Bassetti N, Mohammadin S, Verbaarschot P, Bukovinszkine'Kiss G, *et al.* 2021. Insect egg-killing: a new front on the evolutionary arms-race between brassicaceous plants and pierid butterflies. *New Phytol.* **230**:341-353.

Guo Z, Flannery-Sutherland JT, Benton MJ, Chen ZQ. 2023. Bayesian analyses indicate bivalves did not drive the downfall of brachiopods following the Permian-Triassic mass extinction. *Nat. Commun.* **14**:5566.

Hancock DL. 1983. Classification of the Papilionidae (Lep.): a phylogenetic approach. *Smithersia* **2**:1-48.

Hancock DL. 2007. Phylogeny of the Troidine butterflies (Lepidoptera: Papilionidae) revisited: Are the red-bodied swallowtails monophyletic? *Austral. Entomol.* **34**:33-42.

Hardy NB, Otto SP. 2014. Specialization and generalization in the diversification of phytophagous insects: tests of the musical chairs and oscillation hypotheses. *Proc. R. Soc. B: Biol. Sci.* **281**:20132960.

Heard SB, Hauser DL. 1995. Key evolutionary innovations and their ecological mechanisms. *Hist. Biol.* **10**:151-173.

Heidel-Fischer HM, Kirsch R, Reichelt M, Ahn SJ, Wielsch N, Baxter SW, *et al.* 2019. An insect counteradaptation against host plant defenses evolved through concerted neofunctionalization. *Mol. Biol. Evol.* **36**:930-941.

Ho SYW, Phillips MJ. 2009. Accounting for calibration uncertainty in phylogenetic estimation of evolutionary divergence times. *Syst. Biol.* **58**:367–380.

Höhna S, Freyman WA, Nolen Z, Huelsenbeck JP, May MR, Moore BR. 2019. A Bayesian approach for estimating branch-specific speciation and extinction rates. *bioRxiv* https://doi.org/10.1101/555805

Höhna S, Landis MJ, Heath TA, Boussau B, Lartillot N, Moore BR, Huelsenbeck JP, Ronquist F. 2016. RevBayes: Bayesian phylogenetic inference using graphical models and an interactive model-specification language. *Syst. Biol.* **65**:726–736.

Hu SJ, Condamine FL, Monastyrskii A, Cotton A. 2019. A new species of the *Graphium* (*Pazala*) *mandarinus* group from Central Vietnam (Lepidoptera: Papilionidae). *Zootaxa* **4554**:286-300.

Humpf HU, Schreier P. 1992. 3-Hydroxy-5, 6-epoxy-. beta.-ionol. beta.-D-glucopyranoside and 3-hydroxy-7, 8-dihydro-. beta.-ionol. beta.-D-glucopyranoside: New C13

norisoprenoid glucoconjugates from sloe tree (*Prunus spinosa* L.) leaves. *J. Agricultural Food Chem.* **40**:1898-1901.

Hussain SF, Freyer AJ, Guinaudeau H, Shamma M, Siddiqui MT. 1986. Seven new aporphine-benzylisoquinoline alkaloids from *Thalictrum cultratum*. *J. Natural Prod.* **49**:494-499.

Irizar AC, Fernandez MF, González AG, Ravelo AG. 1992. Constituents of *Prunus spinosa*. *J. Natural Prod.* **55**:450-454.

Janz N, Nylin S. 2008. The oscillation hypothesis of host-plant range and speciation. In: *Specialization, Speciation, and Radiation: The Evolutionary Biology of Herbivorous Insects* (Eds. Tilmon K). University of California Press, pp. 203-215.

Janz N. 2011. Ehrlich and Raven revisited: mechanisms underlying codiversification of plants and enemies. *Annu. Rev. Ecol. Evol. Syst.* **42**:71-89.

Jiggins CD, Mallarino R, Willmott KR, Bermingham E. 2006. The phylogenetic pattern of speciation and wing pattern change in Neotropical *Ithomia* butterflies (Lepidoptera: Nymphalidae). *Evolution* **60**:1454–1466.

Joshi J, Kunte K. 2022. Polytypy and systematics: diversification of *Papilio* swallowtail butterflies in the biogeographically complex Indo-Australian Region. *bioRxiv*, 2022-03.

Jouault C, Nel A, Perrichot V, Legendre F, Condamine FL. 2022. Multiple drivers and lineage-specific insect extinctions during the Permo–Triassic. *Nat. Commun.* **13**:7512.

Jousselin E, Elias M. 2019. Testing host-plant driven speciation in phytophagous insects: a phylogenetic perspective. *PCi Evol. Biol.* https://hal.science/hal-02400722.

Jun XU, Wang XY, Guo WZ. 2015. The cytochrome P450 superfamily: key players in plant development and defense. *J. Integrative Agri.* **14**:1673-1686.

Kalyaanamoorthy S, Minh BQ, Wong TK, Von Haeseler A, Jermiin LS. 2017. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat. Meth.* **14**:587-589.

Kawahara AY, Storer C, Carvalho APS, Plotkin DM, Condamine FL, Braga MP, *et al.* 2023. A global phylogeny of butterflies reveals their evolutionary history, ancestral hosts and biogeographic origins. *Nat. Ecol. Evol.* **7**:903-913.

Kergoat GJ, Delobel A, Fédière G, Le Rü B, Silvain JF. 2005. Both host-plant phylogeny and chemistry have shaped the African seed-beetle radiation. *Mol. Phylogenet. Evol.* **35**:602-611.

Köpf A, Rank NE, Roininen H, Julkunen-Tiitto R, Pasteels JM, Tahvanainen J. 1998. The evolution of host-plant use and sequestration in the leaf beetle genus *Phratora* (Coleoptera: Chrysomelidae). *Evolution* **52**:517-528.

Kumarasamy Y, Cox PJ, Jaspars M, Nahar L, Sarker SD. 2003. Cyanogenic glycosids from *Prunus spinosa* (Rosaceae). *Biochem. Syst. Ecol.* **31**:1063-1065.

Labandeira CC, Sepkoski Jr, JJ. 1993. Insect diversity in the fossil record. *Science* **261**:310-315.

Lagomarsino LP, Condamine FL, Antonelli A, Mulch A, Davis CC. 2016 The abiotic and biotic drivers of rapid diversification in Andean bellflowers (Campanulaceae). *New Phytol.* **210**:1430–1442.

Legendre F, Condamine FL. 2018. When Darwin's special difficulty promotes diversification in insects. *Syst. Biol.* **67**:873-887.

Letsch H, Gottsberger B, Metzl C, Astrin J, Friedman AL, McKenna DD, Fiedler K. 2018. Climate and host-plant associations shaped the evolution of ceutorhynch weevils throughout the Cenozoic. *Evolution* **72**:1815-1828.

Li W, Schuler MA, Berenbaum MR. 2003. Diversification of furanocoumarin-metabolizing cytochrome P450 monooxygenases in two papilionids: specificity and substrate encounter rate. *Proc. Natl Acad. Sci. USA* **100**:14593-14598.

Linder HP, Hardy CR, Rutschmann F. 2005. Taxon sampling effects in molecular clock dating: an example from the African Restionaceae. *Mol. Phylogenet. Evol.* **35**:569-582.

Liow LH, Reitan T, Harnik PG. 2015. Ecological interactions on macroevolutionary time scales: clams and brachiopods are more than ships that pass in the night. *Ecol. Lett.* **18**:1030-1039.

Maliet O, Hartig F, Morlon H. 2019. Amodel with many small shifts for estimating species-specific diversification rates. *Nat. Ecol. Evol.* **3**:1086–1092.

Maliet O, Morlon H. 2022. Fast and accurate estimation of species-specific diversification rates using data augmentation. *Syst. Biol.* **71**:353-366.

Marques DA, Jones FC, Di Palma F, Kingsley DM, Reimchen TE. 2018. Experimental evidence for rapid genomic adaptation to a new niche in an adaptive radiation. *Nat. Ecol. Evol.* **2**:1128–1138.

Mazet N, Fabre P-H, Morlon H, Condamine FL. 2023. Estimating cladespecific diversification rates and palaeodiversity dynamics from reconstructed phylogenies. *Meth. Ecol. Evol.* doi.org/10.1101/2022.05.10.490920.

Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, Von Haeseler A, Lanfear R. 2020. IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. *Mol. Biol. Evol.* **37**:1530-1534.

Mitter C, Farrell B, Wiegmann B. 1988. The phylogenetic study of adaptive zones: has phytophagy promoted insect diversification? *Am. Nat.* **132**:107-128.

Möhn E. 2002. Papilionidae VIII. *Baronia*, *Euryades*, *Protographium*, *Neographium* und *Eurytides*. Goecke & Evers, Keltern. 15 pp. vol. 14, p. ls 1-36. Schmetterlinge der Erde.Tagfalter. Teil 5.

Morlon H, Lewitus E, Condamine FL, Manceau M, Clavel J, Drury J. 2016. RPANDA: an R package for macroevolutionary analyses on phylogenetic trees. *Meth. Ecol. Evol.* **7**:589-597.

Morlon H, Parsons TL, Plotkin JB. 2011. Reconciling molecular phylogenies with the fossil record. *Proc. Natl. Acad. Sci. USA* **108**:16327–16332.

Myers N, Mittermeier RA, Mittermeier CG, da Fonseca GAB, Kent J. 2000 Biodiversity hotspots for conservation priorities. *Nature* **403**:853–858.

Nakae M. 2021. Papilionidae of the World. Tokyo: Roppon-Ashi Entomological Books.

Nicholson DB, Ross AJ, Mayhew PJ. 2014. Fossil evidence for key innovations in the evolution of insect diversity. *Proc. R. Soc. B: Biol. Sci.* **281**:20141823.

Nishida R. 2002. Sequestration of defensive substances from plants by Lepidoptera. *Annu. Rev. Entomol.* **47**:57-92.

Nyman T, Linder HP, Peña C, Malm T, Wahlberg N. 2012. Climate-driven diversity dynamics in plants and plant-feeding insects. *Ecol. Lett.* **15**:889-898.

Nyman T, Onstein RE, Silvestro D, Wutke S, Taeger A, Wahlberg N, *et al.* 2019. The early wasp plucks the flower: disparate extant diversity of sawfly superfamilies (Hymenoptera: 'Symphyta') may reflect asynchronous switching to angiosperm hosts. *Biol J. Linn. Soc.* **128**:1-19.

Nyman T, Vikberg V, Smith DR, Boevé JL. 2010. How common is ecological speciation in plant-feeding insects? A 'Higher' Nematinae perspective. *BMC Evol. Biol.* **10**:266.

Okamura YU, Sato AI, Tsuzuki N, Sawada Y, Hirai MY, Heidel-Fischer H, *et al.* 2019. Differential regulation of host plant adaptive genes in Pieris butterflies exposed to a range of glucosinolate profiles in their host plants. *Sci. Rep.* **9**:7256.

Olszewska M. 2002. Flavonoids from the leaves of *Prunus spinosa* L. *Polish J. Chem.* **76**:967-974.

Pena-Hidalgo M, Furtado LC, Costa-Lotufo LV, Ferreira MJ, Santos DY. 2021. Alkaloids from the leaves of *Annona crassiflora* and their cytotoxic activity. *Revista Brasileira Farmacognosia* **31**:244-248.

Pinto CF, Troncoso AJ, Urzua A, Niemeyer HM. 2009. Aristolochic acids affect the feeding behaviour and development of *Battus polydamas archidamas* larvae (Lepidoptera: Papilionidae: Troidini). *Eur. J. Entomol.* **106**:357-361.

Puissant A, Chotard A, Condamine FL, Llaurens V. 2023. Convergence in sympatric swallowtail butterflies reveals ecological interactions as a key driver of worldwide trait diversification. *Proc. Natl Acad. Sci. USA* **120**:e2303060120.

Rainford JL, Hofreiter M, Nicholson DB, Mayhew PJ. 2014. Phylogenetic distribution of extant richness suggests metamorphosis is a key innovation driving diversification in insects. *PLoS One* **9**:e109085.

Rambaut A, Drummond AJ, Xie D, Baele G, Suchard MA. 2018. Posterior summarization in Bayesian phylogenetics using Tracer 1.7. *Syst. Biol.* **67**:901-904.

Ranwez V, Douzery EJP, Cambon C, Chantret N, Delsuc F. 2018. MACSE v2: toolkit for the alignment of coding sequences accounting for frameshifts and stop codons. *Mol. Biol. Evol.* **35**:2582-2584.

Ree RH, Moore BR, Webb CO, Donoghue MJ. 2005. A likelihood framework for inferring the evolution of geographic range on phylogenetic trees. *Evolution* **59**:2299-2311.

Ree RH, Sanmartín I. 2018. Conceptual and statistical problems with the DEC+ J model of founder-event speciation and its comparison with DEC via model selection. *J. Biogeogr.* **45**:741-749.

Ree RH, Smith SA. 2008. Maximum likelihood inference of geographic range evolution by dispersal, local extinction, and cladogenesis. *Syst. Biol.* **57**:4–14.

Ronquist F, Sanmartín I. 2011. Phylogenetic methods in biogeography. *Annu. Rev. Ecol. Evol. Syst.* **42**:441-464.

Rougerie R, Cruaud A, Arnal P, Ballesteros-Mejia L, Condamine FL, Decaëns T, Elias M, Grey D, Hebert PDN, Kitching IJ, *et al.* 2022. Phylogenomics illuminates the evolutionary history of wild silkmoths in space and time (Lepidoptera: Saturniidae). *bioRxiv* https://doi.org/10.1101/2022.03.29.486224.

Ruane S, Bryson Jr, RW, Pyron RA, Burbrink FT. 2014. Coalescent species delimitation in milksnakes (genus *Lampropeltis*) and impacts on phylogenetic comparative analyses. *Syst. Biol.* **63**:231-250.

Sahoo RK, Warren AD, Collins SC, Kodandaramaiah U. 2017. Hostplant change and paleoclimatic events explain diversification shifts in skipper butterflies (Family: Hesperiidae). *BMC Evol. Biol.* **17**:1-9.

Schluter D. 2000. *The Ecology of Adaptive Radiation*. Oxford University Press.

Schuler MA. 1996. The role of cytochrome P450 monooxygenases in plant-insect interactions. *Plant Physiol.* **112**:1411-1419.

Scotese CR. 2004. A continental drift flipbook. *J. Geol.* **112**:729-741.

Scriber JM, Tsubaki Y, Lederhouse RC. 1995. *Swallowtail Butterflies: Their Ecology and Evolutionary Biology*. Scientific Publishers, Gainesville, Florida.

Seton M, Muller R, Zahirovic S, Gaina C, Torsvik T, Shephard G, Talsma A, Gurnis M, Turner M, Chandler M. 2012. Global continental and ocean basin reconstructions since 200 Ma. *Earth-Sci. Rev.* **113**:212–270.

Shin S, Clarke DJ, Lemmon AR, Moriarty Lemmon E, Aitken AL, Haddad S, *et al.* 2018. Phylogenomic data yield new and robust insights into the phylogeny and evolution of weevils. *Mol. Biol. Evol.* **35**:823-836.

Silvestro D, Antonelli A, Salamin N, Quental TB. 2015. The role of clade competition in the diversification of North American canids. *Proc. Natl Acad. Sci. USA* **112**:8684-8689.

Simon JC, d'Alencon E, Guy E, Jacquin-Joly E, Jaquiery J, Nouhaud P, *et al.* 2015. Genomics of adaptation to host-plants in herbivorous insects. *Briefings Functional Genom.* **14**:413-423.

Smith SA, Brown JW, Walker JF. 2018. So many genes, so little time: a practical approach to divergence-time estimation in the genomic era. *PLoS One* **13**:e0197433.

Sohn JC, Labandeira CC, Davis DR. 2015. The fossil record and taphonomy of butterflies and moths (Insecta, Lepidoptera): implications for evolutionary diversity and divergence-time estimates. *BMC Evol. Biol.* **15**:1-15.

Steiner AA, Romanovsky AA. 2019. Energy trade-offs in host defense: immunology meets physiology. *Trends Endocrinol. Metabol.* **30**:875-878.

Suchard MA, Lemey P, Baele G, Ayres DL, Drummond AJ, Rambaut A. 2018. Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. *Virus Evol.* **4**:vey016.

Suvák M. 2015. Exotic butterflies and moths (Lepidoptera) in botanical gardens-potential for education and research. *Thaiszia J. Bot.* **25**:81-147.

Takos, AM, Knudsen C, Lai D, Kannangara R, Mikkelsen L, Motawia MS, *et al.* 2011. Genomic clustering of cyanogenic glucoside biosynthetic genes aids their identification in *Lotus japonicus* and suggests the repeated evolution of this chemical defence pathway. *Plant J.* **68**:273-286.

Terekhanova NV, Logacheva MD, Penin AA, Neretina TV, Barmintseva AE, Bazykin GA, Kondrashov AS, Mugue NS. 2014. Fast evolution from precast bricks: genomics of young freshwater populations of threespine stickleback *Gasterosteus aculeatus*. *PLoS Genet.* **10**:e1004696.

Thompson JN. 1989. Concepts of coevolution. *Trends Ecol. Evol.* **4**:179-183.

Thorne JL, Kishino H, Painter IS. 1998. Estimating the rate of evolution of the rate of molecular evolution. *Mol. Biol. Evol.* **15**:1647-1657.

To TH, Jung M, Lycett S, Gascuel O. 2016. Fast dating using least-squares criteria and algorithms. *Syst. Biol.* **65**:82-97.

Toussaint EFA, Beutel RG, Morinière J, Jia F, Xu S, Michat MC, *et al.* 2016. Molecular phylogeny of the highly disjunct cliff water beetles from South Africa and China (Coleoptera: Aspidytidae). *Zool. J. Linn. Soc.* **176**:537-546.

Toussaint EFA, Seidel M, Arriaga-Varela E, Hájek J, Kral D, Sekerka L, *et al.* 2017. The peril of dating beetles. *Syst. Entomol.* **42**:1-10.

Toussaint EFA, Vila R, Yago M, Chiba H, Warren AD, Aduse-Poku K, *et al.* 2019. Out of the Orient: Post-Tethyan transoceanic and trans-Arabian routes fostered the spread of Baorini skippers in the Afrotropics. *Syst. Entomol.* **44**:926-938.

Tyler HA, Brown KS, Wilson K. 1994. *Swallowtail Butterflies of the Americas: A Study in Biological Dynamics, Ecological Diversity, Biosystematics, and Conservation*. Scientific Publishers, Gainesville, Florida.

Utami CY, Sholihah A, Condamine FL, Thébaud C, Hubert N. 2022. Cryptic diversity impacts model selection and macroevolutionary inferences in diversification analyses. *Proc. R. Soc. B: Biol. Sci.* **289**:20221335.

Wahlberg N. 2001. The phylogenetics and biochemistry of host-plant specialization in Melitaeine butterflies (Lepidoptera: Nymphalidae). *Evolution* **55**:522-537.

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

Wang K, Wang X, Zhang H, Ge Y. 2019. A new aporphine alkaloid from *Illigera aromatic*. *J. Nat. Prod.* **13**:440-445.

Weiss JC. 1991. *The Parnassiinae of the World*. Sciences Nat Hillside Books.

Williams MC. 2022. Afrotropical butterflies and skippers – a digital encyclopaedia. https://www.metamorphosis.org.za/?p=articles&s=atb. Accessed 9 December 2022.

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.

Yang Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. *Computer App. Biosci.* **13**:555-556.

Zagrobelny M, Bak S, Ekstrøm CT, Olsen CE, Møller L. 2007. The cyanogenic glucoside composition of *Zygaena filipendulae* (Lepidoptera: Zygaenidae) as effected by feeding on wild-type and transgenic lotus populations with variable cyanogenic glucoside profiles. *Insect Biochem. Mol. Biol.* **37**:10-18.

Zagrobelny M, Bak S, Rasmussen AV, Jørgensen B, Naumann CM, Møller BL. 2004. Cyanogenic glucosides and plant–insect interactions. *Phytochem.* **65**:293-306.

Zhang HH, Cotton AM, Condamine FL, Wang RJ, Hsu YF, Duan K, *et al.* 2020. Revision of *Pazala* Moore, 1888: The *Graphium* (*Pazala*) *alebion* and *G.* (*P.*) *tamerlanus* groups, with notes on taxonomic and distribution confusions (Lepidoptera: Papilionidae). *Zootaxa* **4759**:77-97.

Supplementary information

Table S1: Connectivity matrices that represent major changes in paleogeographic conditions for five time periods. Area connectivity is coded 0 if any two areas are not physically connected or 1 if they are connected at a given period. Abbreviations: (WP) West Palearctic, (EP) East Palearctic, (WN) West Nearctic, (EN) East Nearctic, (CA) Central America, (WI) the Caribbean Islands, (SA) South America, (AF) Africa, (MD) Madagascar, (IN) India, (OR) continental Southeast Asia, (SEA) insular Southeast Asia, (AU) Australasia.

0 - 5.33 Ma													
	WP	EP	WN	EN	CA	WI	SA	AF	MD	IN	OR	SEA	AU
WP	1												
EP	1	1											
WN	0	1	1										
EN	0	0	1	1									
CA	0	0	1	1	1								
WI	0	0	0	1	1	1							
SA	0	0	0	0	1	1	1						
AF	1	0	0	0	0	0	0	1					
MD	0	0	0	0	0	0	0	1	1				
IN	0	1	0	0	0	0	0	0	0	1			
OR	0	1	0	0	0	0	0	0	0	1	1		
SEA	0	0	0	0	0	0	0	0	0	0	1	1	
AU	0	0	0	0	0	0	0	0	0	0	0	1	1

5.33 - 23.03 Ma													
	WP	EP	WN	EN	CA	WI	SA	AF	MD	IN	OR	SEA	AU
WP	1												
EP	1	1											
WN	0	1	1										
EN	0	0	1	1									
CA	0	0	1	1	1								
WI	0	0	0	1	1	1							
SA	0	0	0	0	0	1	1						
AF	1	0	0	0	0	0	0	1					
MD	0	0	0	0	0	0	0	1	1				
IN	0	1	0	0	0	0	0	1	0	1			
OR	0	1	0	0	0	0	0	0	0	1	1		
SEA	0	0	0	0	0	0	0	0	0	0	1	1	
AU	0	0	0	0	0	0	0	0	0	0	0	1	1

23.03 - 33.9 Ma													
	WP	EP	WN	EN	CA	WI	SA	AF	MD	IN	OR	SEA	AU
WP	1												
EP	1	1											
WN	0	1	1										
EN	0	0	1	1									
CA	0	0	1	1	1								
WI	0	0	0	1	1	1							
SA	0	0	0	0	0	1	1						
AF	0	0	0	0	0	0	0	1					
MD	0	0	0	0	0	0	0	1	1				
IN	0	1	0	0	0	0	0	0	0	1			
OR	0	1	0	0	0	0	0	0	0	1	1		
SEA	0	0	0	0	0	0	0	0	0	0	1	1	
AU	0	0	0	0	0	0	0	0	0	0	0	0	1

33.9 - 56 Ma													
	WP	EP	WN	EN	CA	WI	SA	AF	MD	IN	OR	SEA	AU
WP	1												
EP	0	1											
WN	0	1	1										
EN	1	0	1	1									
CA	0	0	1	1	1								
WI	0	0	0	1	1	1							
SA	0	0	0	0	0	1	1						
AF	0	0	0	0	0	0	0	1					
MD	0	0	0	0	0	0	0	1	1				
IN	0	1	0	0	0	0	0	0	0	1			
OR	0	1	0	0	0	0	0	0	0	1	1		
SEA	0	0	0	0	0	0	0	0	0	0	1	1	
AU	0	0	0	0	0	0	1	0	0	0	0	0	1

56 - 66 Ma													
	WP	EP	WN	EN	CA	WI	SA	AF	MD	IN	OR	SEA	AU
WP	1												
EP	0	1											
WN	0	1	1										
EN	1	0	1	1									
CA	0	0	1	1	1								
WI	0	0	0	0	1	1							
SA	0	0	0	0	0	0	1						
AF	0	0	0	0	0	0	0	1					
MD	0	0	0	0	0	0	0	1	1				
IN	0	0	0	0	0	0	0	0	1	1			
OR	0	1	0	0	0	0	0	0	0	1	1		
SEA	0	0	0	0	0	0	0	0	0	0	1	1	
AU	0	0	0	0	0	0	1	0	0	0	0	0	1

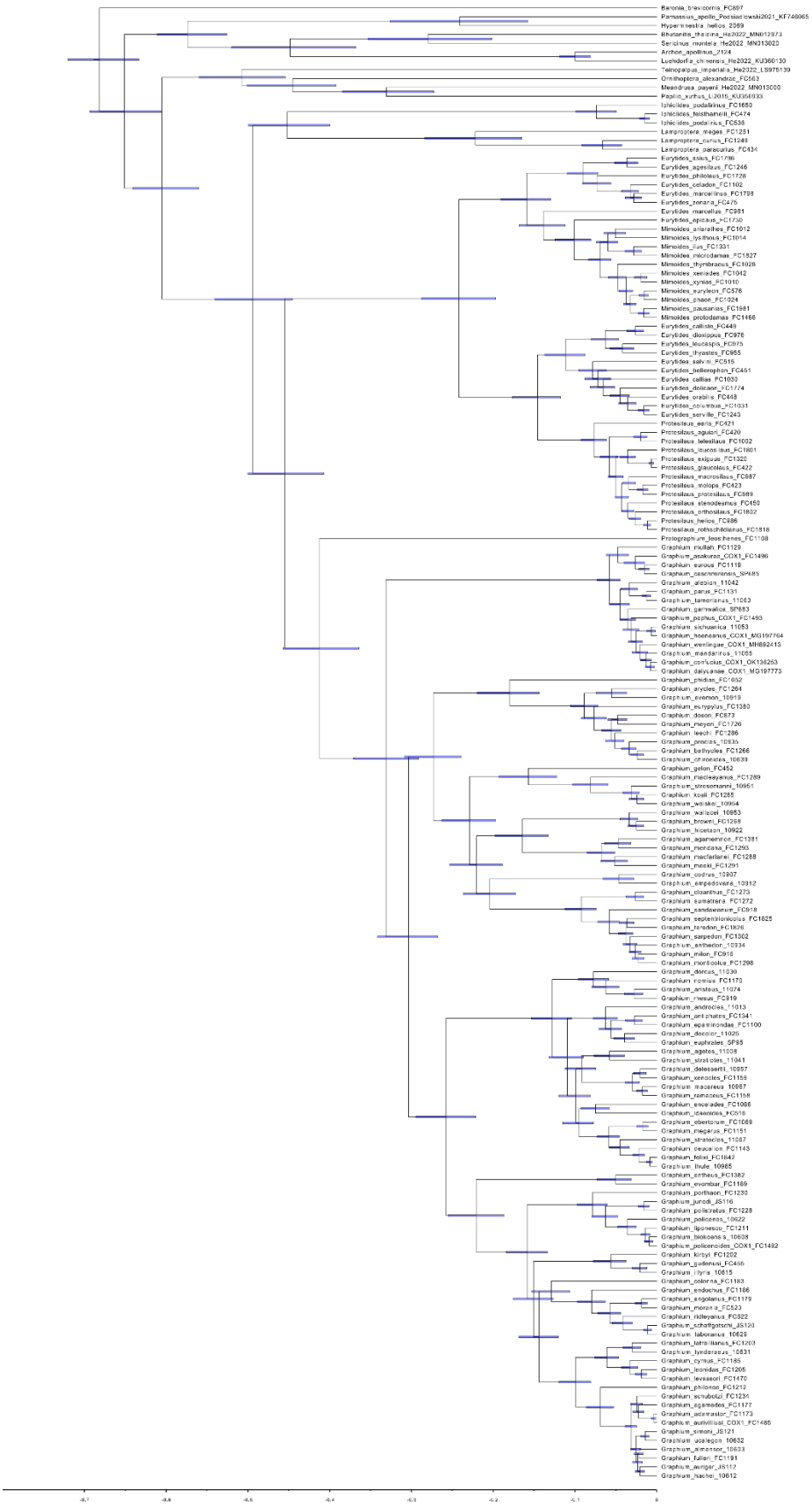


Figure S1: Bayesian molecular dating of Leptocircini. The chronogram is one of the ten MCMC approximations, showing age uncertainties with 95% CI bars at the nodes.

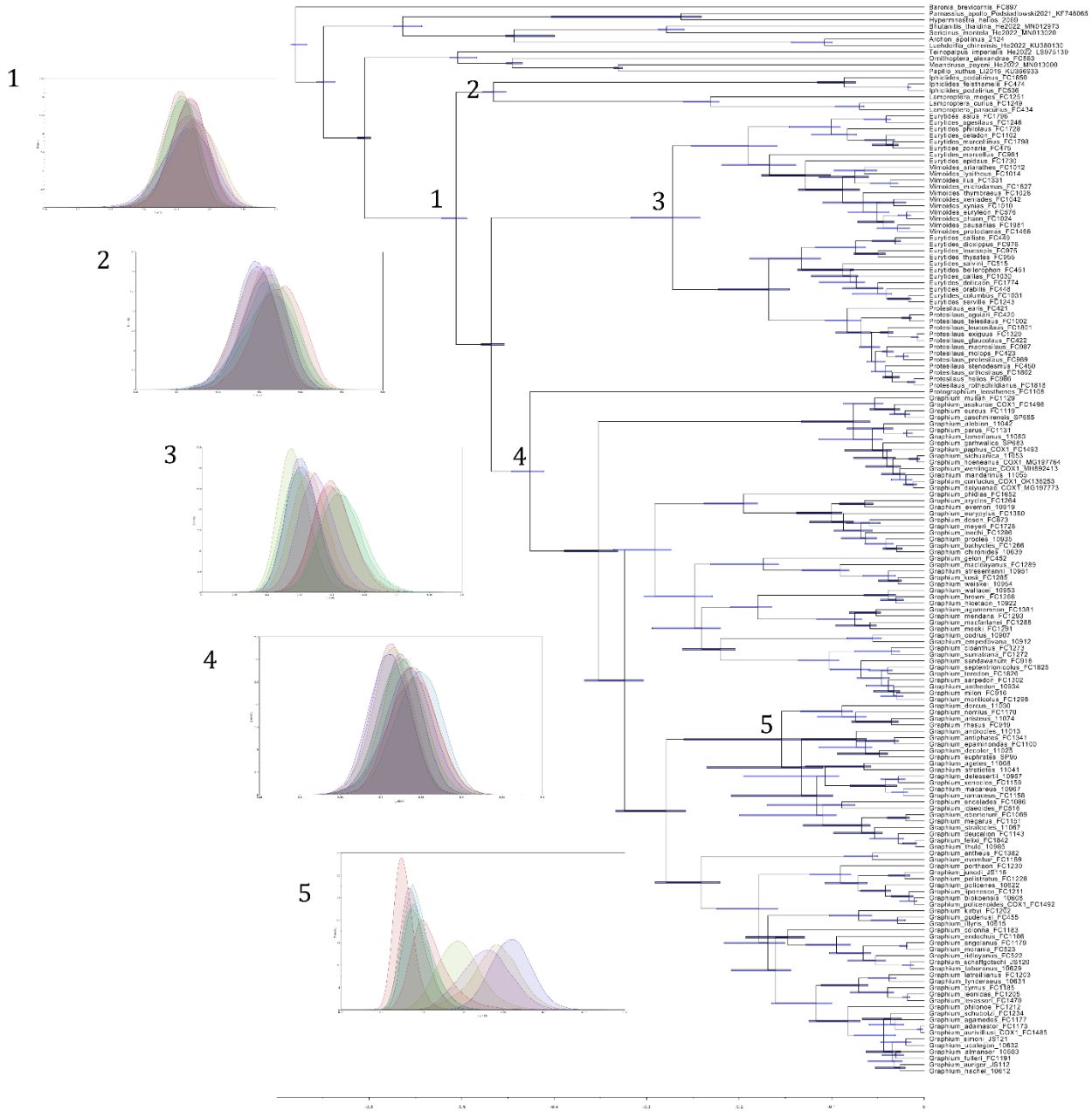


Figure S2: Consensus time-calibrated tree of Leptocircini. The chronogram represents a summary of the ten MCMC runs. Node bars represent 95% CI of the consensus estimated with TreeAnnotator. Inset plots 1-5 show posterior density of the ten runs for key nodes as follows: 1= crown Leptocircini, 2 = *Lamproptera* + *Iphiclides*, 3 = crown *Eurytides s.l.*, 4 = *Graphium* + *Protographium*, and 5 = crown *Pathysa s.l.*

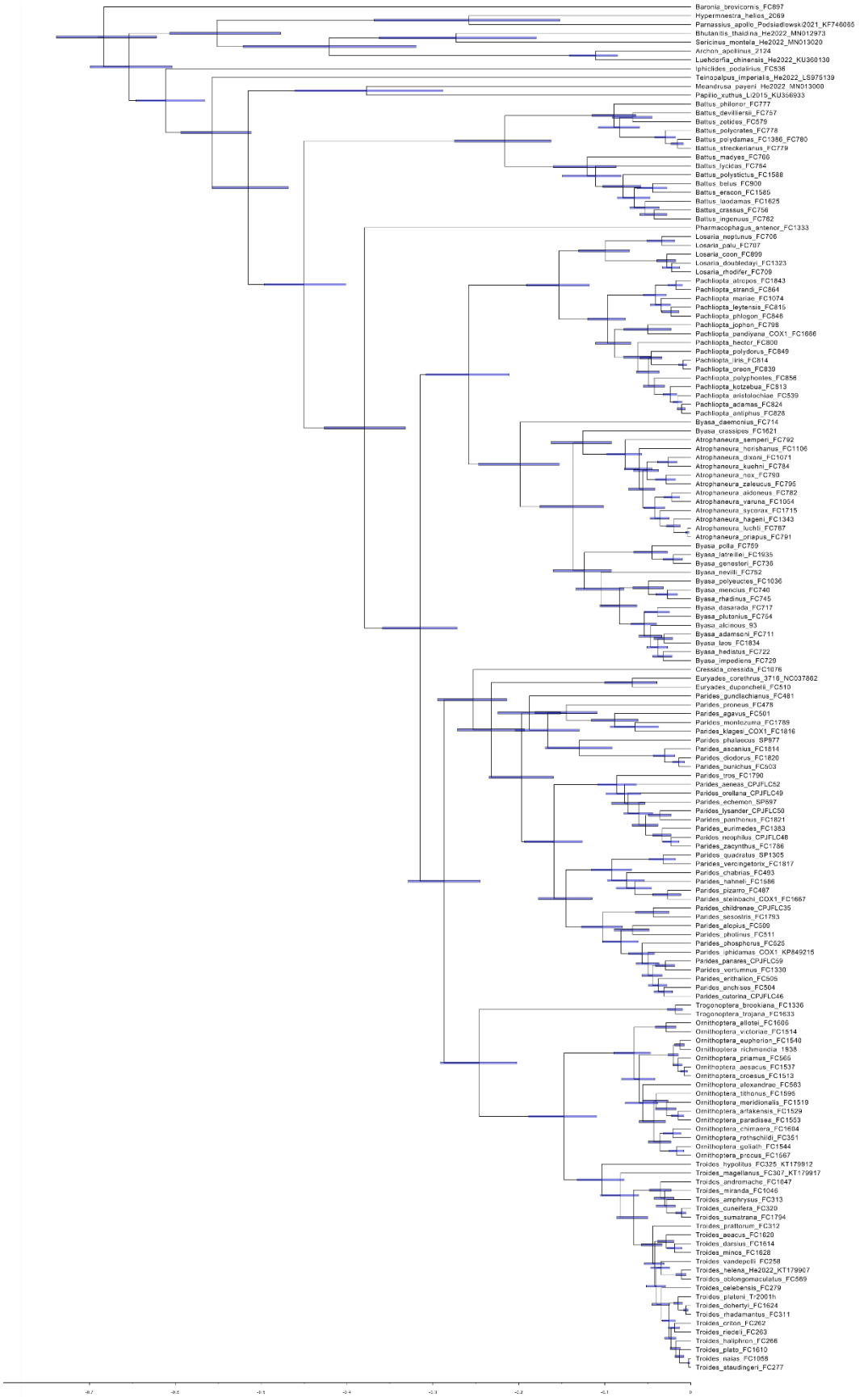


Figure S3: Bayesian molecular dating of Troidini. The chronogram is one of the ten MCMC approximations, showing age uncertainties with 95% CI bars at the nodes.

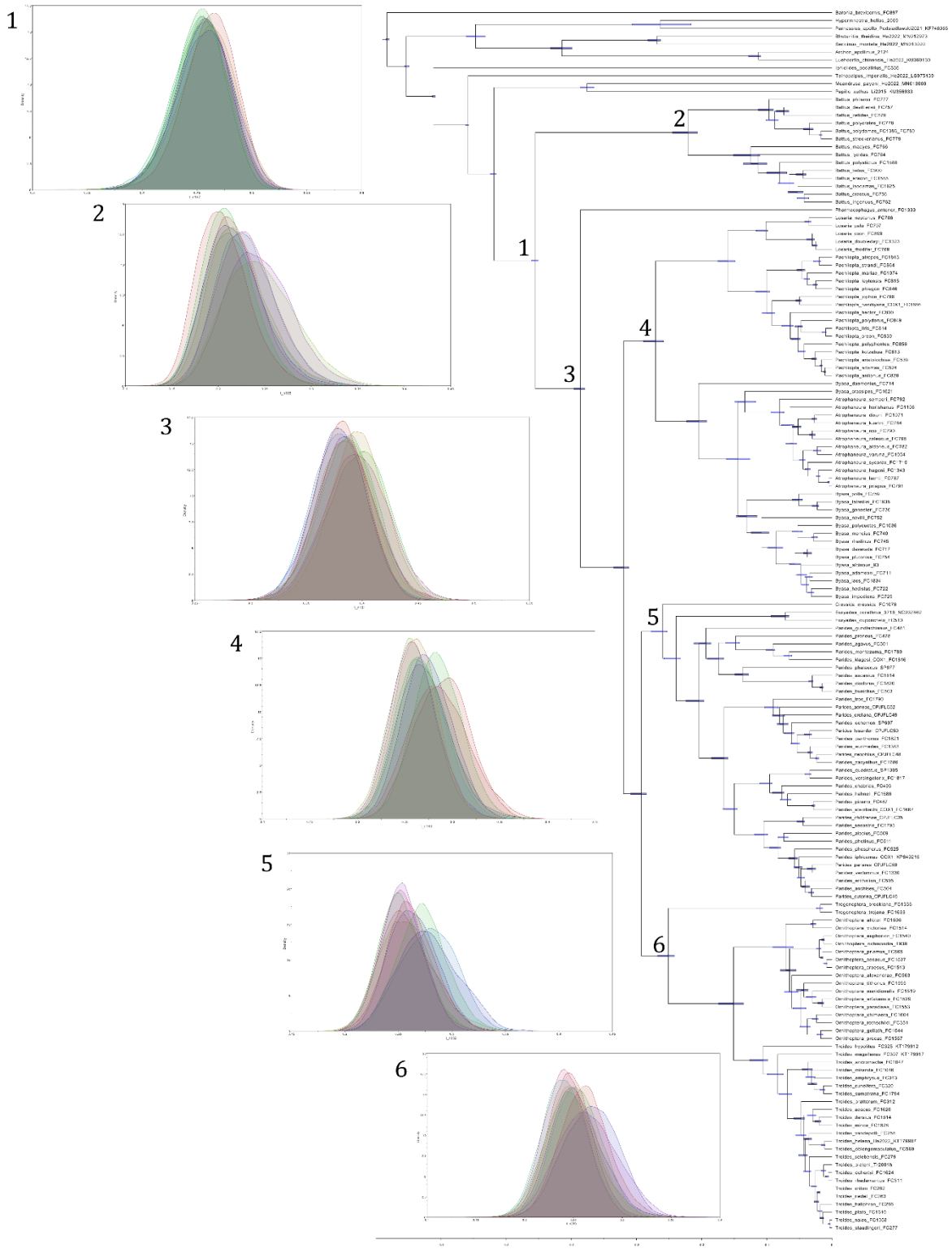


Figure S4: Consensus time-calibrated tree of Troidini. The chronogram represents a summary of the ten MCMC runs. Node bars represent 95% CI of the consensus tree by TreeAnnotator. Inset plots 1-6 show posterior densities of the ten runs for key nodes as follows: 1= crown Troidini, 2 = crown Battina, 3 = crown Troidina, 4 = *Atrophaneura* + *Pachliopta*, 5 = crown *Cressida* + *Euryades* + *Parides*, and 6 = crown birdwings.

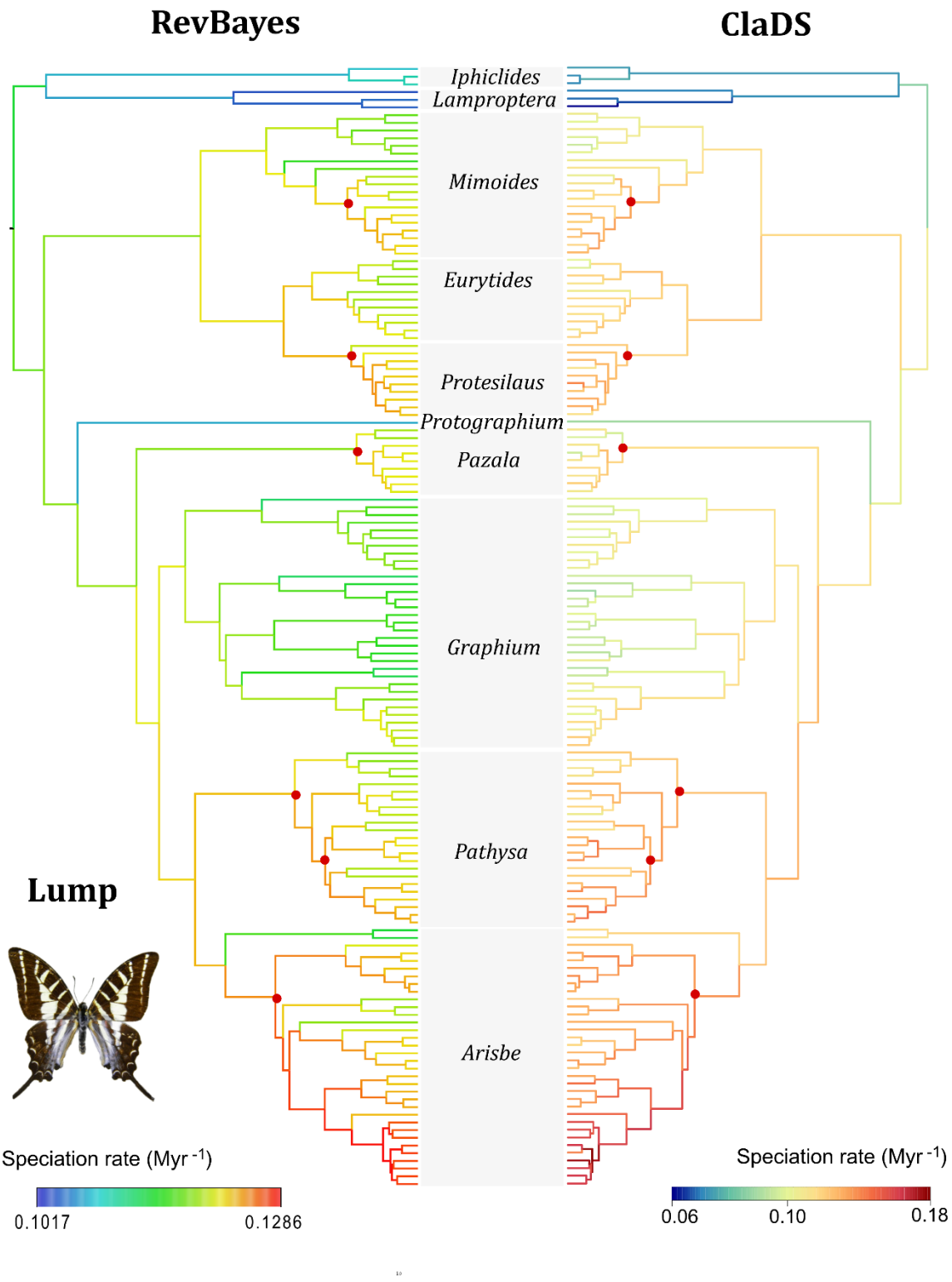


Figure S5: Global diversification dynamics of Leptocircini. The trees display speciation rates as estimated with RevBayes (left) and ClaDS (right) using the *lump* taxonomy. Each analysis has a different scale, reported below the tree. Red points are the shifts tested in RPANDA. Illustration: *Graphium aristeus* (Alex Dumchus).

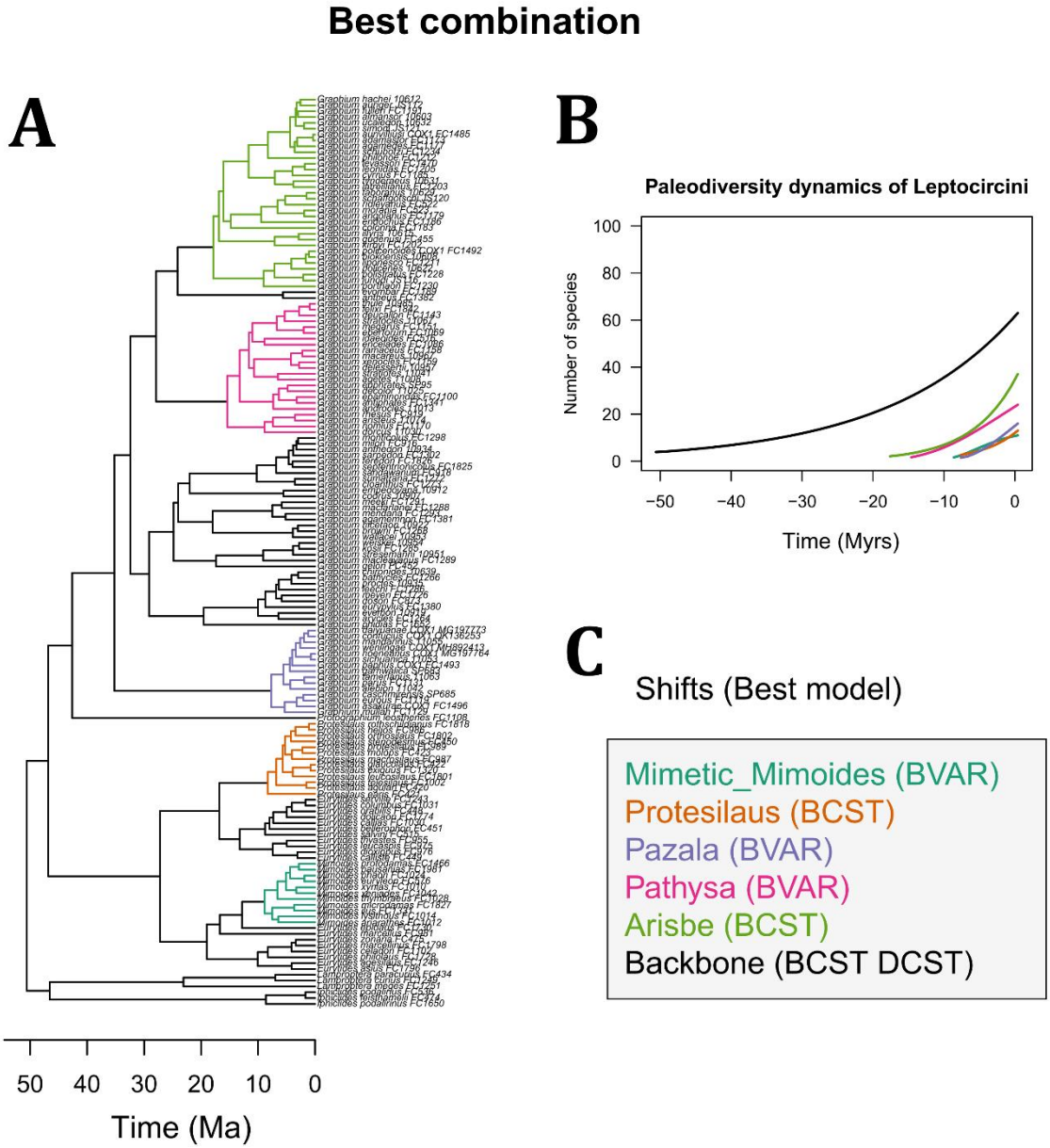


Figure S6: Diversification shifts and diversity dynamics estimated for Leptocircini (complete taxonomy) as estimated with the best-fit combination of shifts by RPANDA. **A**. Phylogeny of Leptocircini with recovered shifts highlighted in colors. **B**. Diversity dynamics for the backbone and subclade trees as estimated with the diversification model, knowing the clade age and the number of extant species. **C**. Legend of colors and best models of shifts (BCST = constant speciation rate, BVAR = variable speciation rate through time, and BCST DCST = constant speciation and extinction rates). Note that the second-best combination (deltaic<2) does not include the shift for "Mimetic_*Mimoides*", and the Backbone model becomes a BCST.

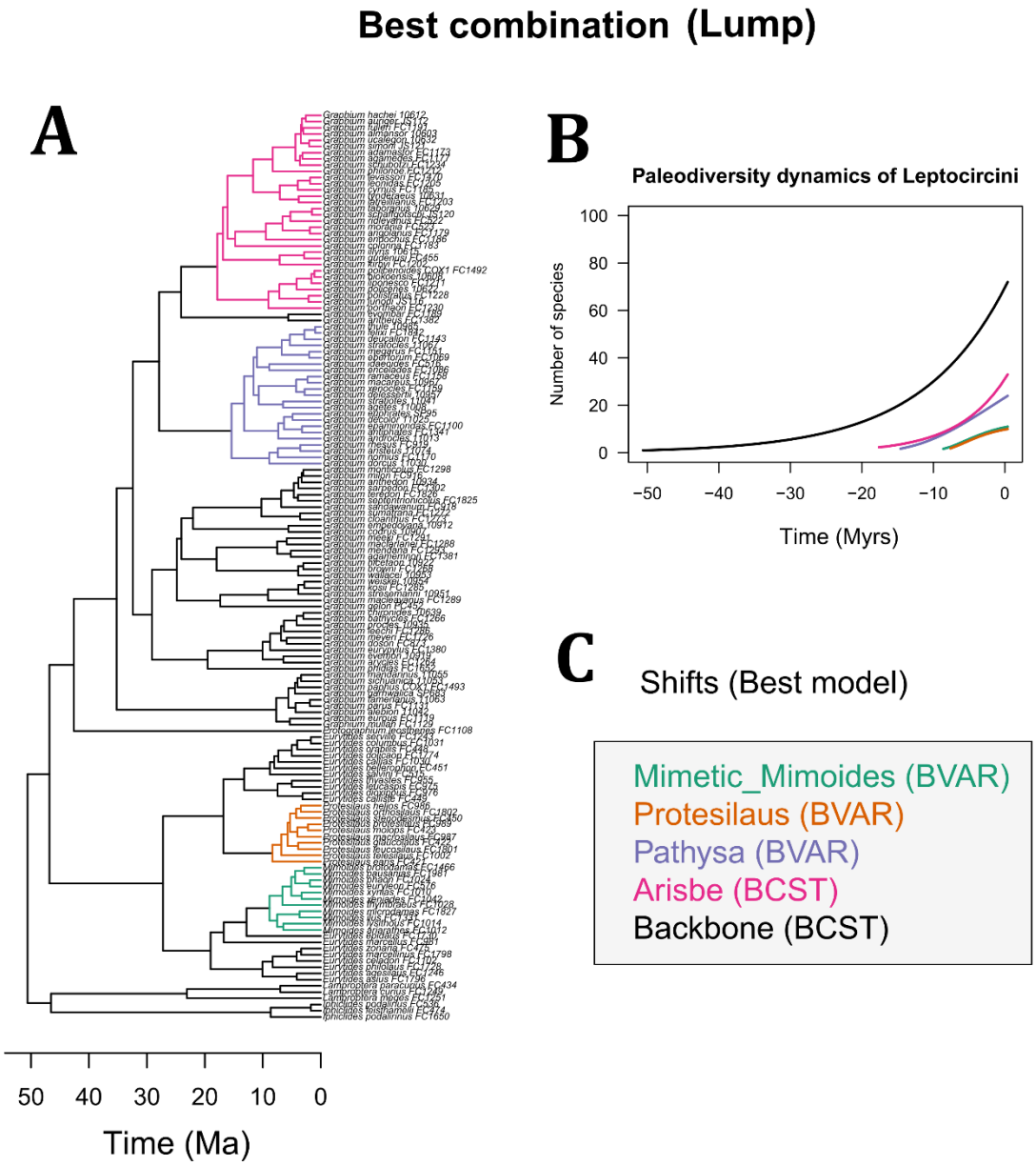


Figure S7: Diversification shifts and diversity dynamics estimated for Leptocircini (lump taxonomy) as estimated with the best-fit combination of shifts by RPANDA. **A**. Phylogeny of Leptocircini with recovered shifts highlighted in colors. **B**. Diversity dynamics for the backbone and subclade trees as estimated with the diversification model, knowing the clade age and the number of extant species. **C**. Legend of colors and best models of shifts (BCST = constant speciation rate, and BVAR = variable speciation rate through time). Note that the second-best combination (deltaAIC = 0.086) includes an additional shift for *Pazala* (BVAR), and the Backbone model becomes BCST DCST (constant speciation and extinction rates). Compared to the best combination, the third-best combination (deltaAIC = 0.395) does not include the shift for *Protesilaus*.

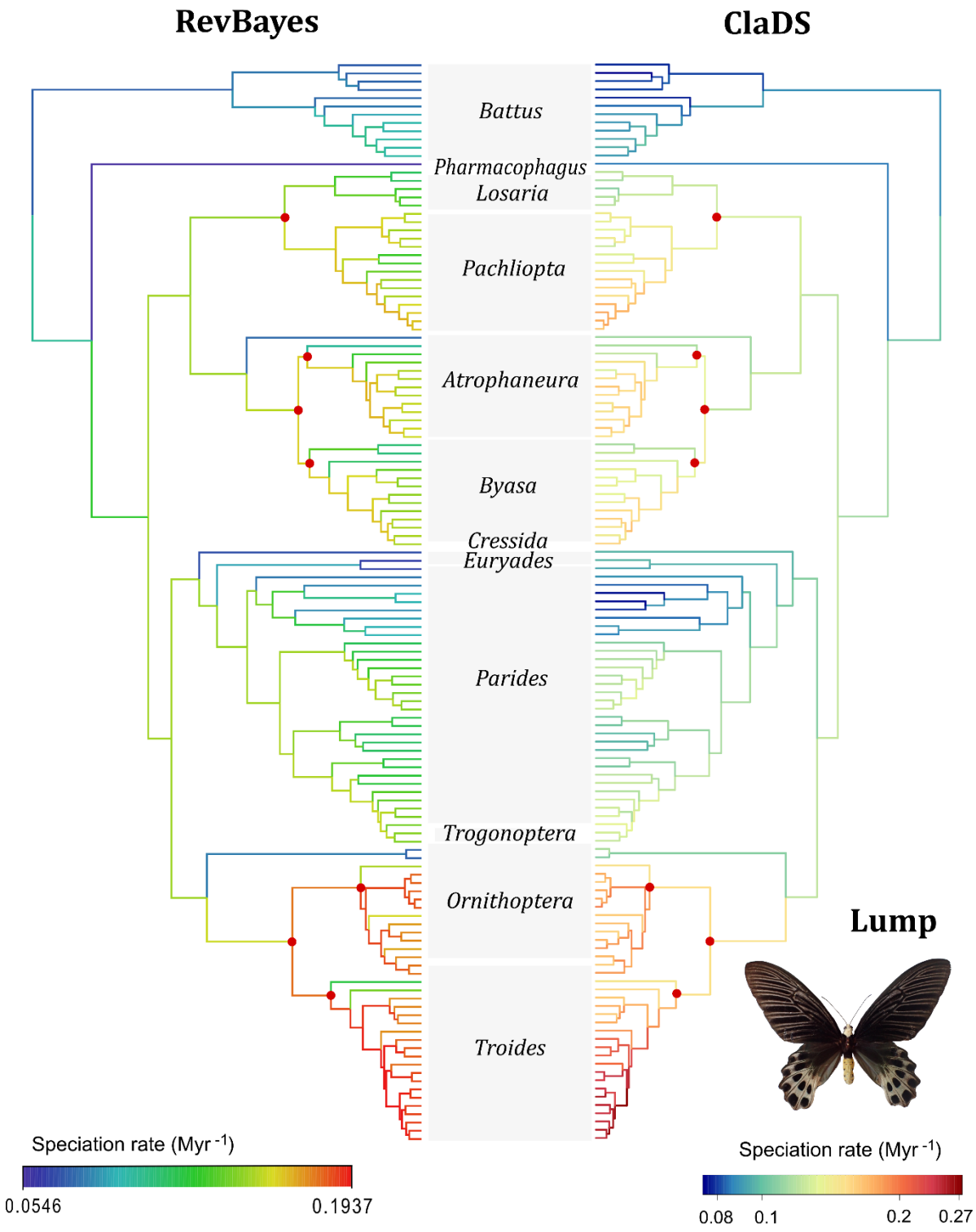


Figure S8: Global diversification dynamics of Troidini. The trees display speciation rates as estimated with RevBayes (left) and ClaDS (right) using the *lump* taxonomy. Each analysis has a different scale, reported below the tree. Red points are the shifts tested in RPANDA.

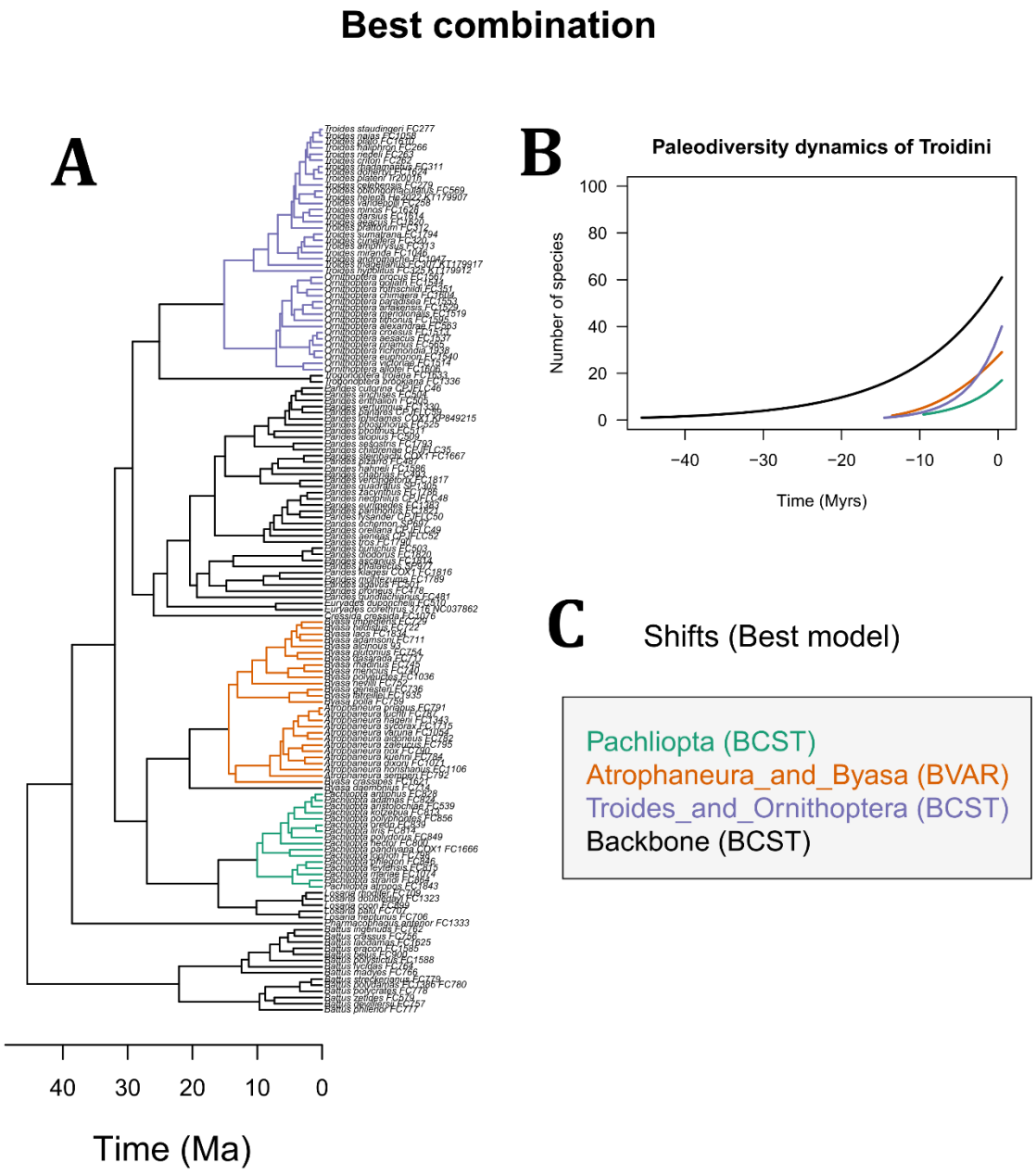


Figure S9: Diversification shifts and diversity dynamics estimated for Troidini (complete taxonomy) as estimated with the best-fit combination of shifts by RPANDA. **A**. Phylogeny of Leptocircini with recovered shifts highlighted in colors. **B**. Diversity dynamics for the backbone and subclade trees as estimated with the diversification model, knowing the clade age and the number of extant species. **C**. Legend of colors and best models of shifts (BCST = constant speciation rate, and BVAR = variable speciation rate through time).

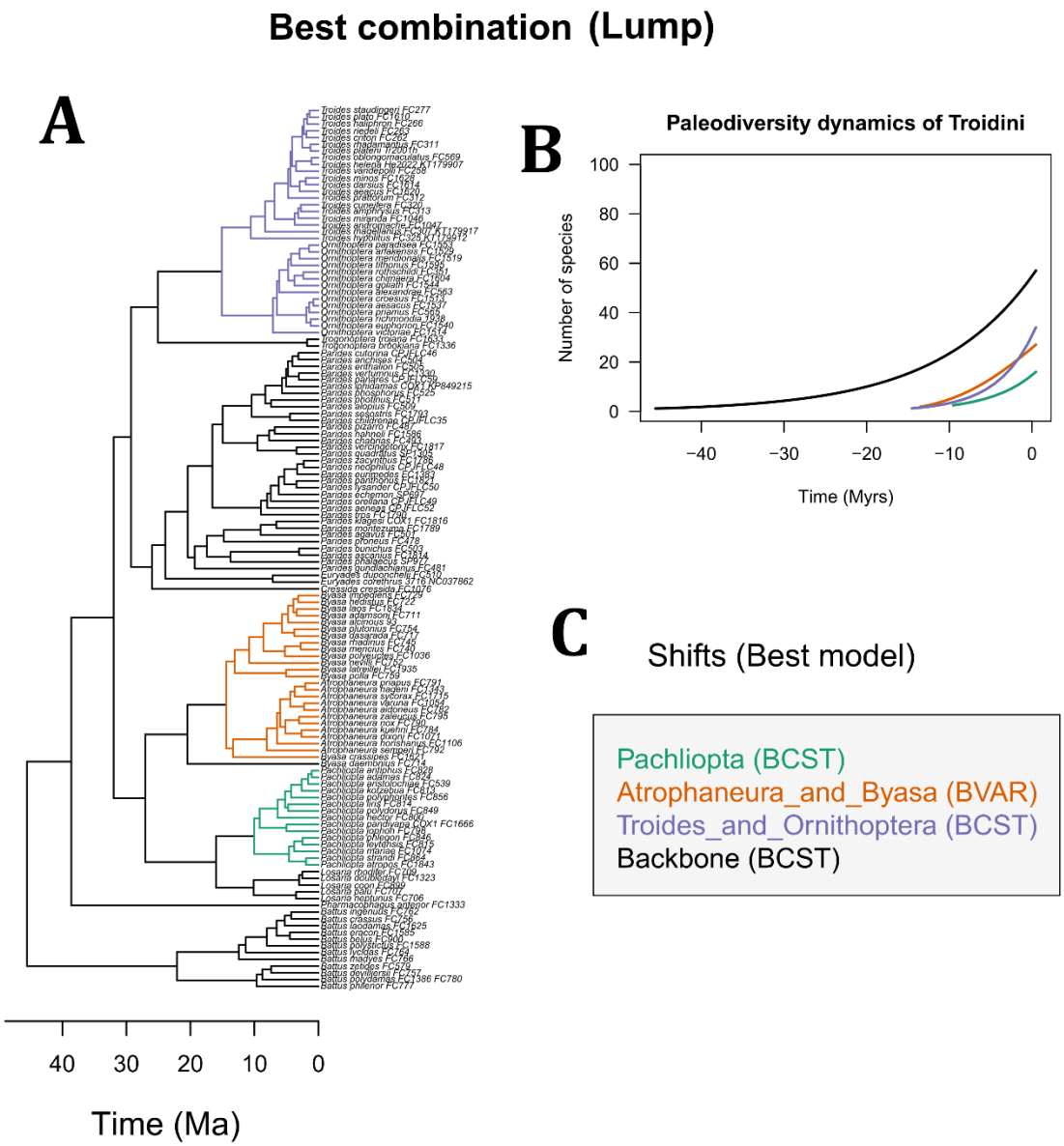


Figure S10: Diversification shifts and diversity dynamics estimated for Troidini (lump taxonomy) as estimated with the best-fit combination of shifts by RPANDA. **A**. Phylogeny of Leptocircini with recovered shifts highlighted in colors. **B**. Diversity dynamics for the backbone and subclade trees as estimated with the diversification model, knowing the clade age and the number of extant species. **C**. Legend of colors and best models of shifts (BCST = constant speciation rate, and BVAR = variable speciation rate through time). Note that the second-best combination (deltaic<2) does not include the shift for *Pachliopta*.

General Discussion



Ornithoptera priamus - Eliette Reboud

0. Foreword

This thesis is an example of the tumultuous journey undertaken from data collection to the assembly and analysis of over 300 complete genomes, to finally exploring the complex conundrum of phylogenetic and macroevolutionary results and interpretations. This type of journey has become so complex over the last twenty years that it is worth stepping back to reflect on the implications of such an undertaking, but also to ensure that this complexity has not led to blindness or haste in the use of the tools, or the interpretation of the myriad of results and information thereby obtained. The data produced during this thesis is such a source of information that I am aware it will take longer than an entire thesis to have a comprehensive, intelligently directed, and concrete view of it.

In this discussion, I will begin by reviewing the power, but also the limits, of the tools used in the current methodological framework of phylogenomics. Some aspects, such as sequencing and genomics, have evolved considerably in the last years, bringing new challenges to light, while others still constitute bottlenecks and hindrances to a scientific and rigorous understanding of some questions. Secondly, I will discuss the contribution that my study of these tribes or genomes of Papilionidae has made to the understanding of some of the major questions in conservation genomics, phylogeny, biogeography, and diversification. Finally, I will discuss the prospects that the results of this thesis promise for future studies of the Papilionidae.

1. Studying whole genome data

Studying the recent past of non-model organisms

The primary material for this thesis has been whole-genome data, whether from short-read or long-read sequencing. Genomes, and especially "reference genomes" (Formenti et al. 2022) are useful tools to study various characteristics such as gene and trait evolution (Warren et al. 2021), inbreeding depression (Robinson et al. 2022), or past demographic history (Morin et al. 2021). The study on *Ornithoptera alexandrae* in **Chapter 1** was not only a case study to determine the divergence level between the two allopatric populations, but it was also intended to be a window into the relatively recent past of this rare and endangered butterfly. In this study, we produced six high-coverage and contiguous annotated genomes of Troidini, including four for *O. alexandrae*, and applied polymorphism-based methods to infer their "historical

demography". These methods, such as the sequentially Markovian coalescent (SMC) methods (and their variants PSMC, SMC++, MSMC, eSMC, Sellinger et al. 2021), were developed in the 2000s and applied on model groups such as human genomes (McVean and Cardin 2005, Li and Durbin 2011). However, it has recently been shown that SMC methods do not perform well when the mutation rate of organisms is much lower than their recombination rate (Sellinger et al. 2021), which, unfortunately, is likely to be very common in many non-model groups such as fungi, sea cucumbers, nematodes, corals, insects, and plants (see Wilfert et al. 2007, Stapley et al. 2017 for recombination rates, and Lynch et al. 2016, Wang and Obbard 2023 for mutation rates). Indeed, by applying these methods on *O. alexandrae* (and thanks to the concern of one anonymous reviewer), we were able to observe and propose a method to address this methodological issue. In practice, we realized that this concern was still poorly addressed, even in recent studies (e.g. coral: Fuller et al. 2020; insects: Waldvogel et al. 2017, Hazzouri et al. 2020, You et al. 2020, Wang et al. 2022, Garcia-Berro et al. 2023) despite the fact that at least many small non-vertebrate genomes have a much lower mutation rate than recombination rate (Sellinger et al. 2020, 2021). Thus, my **Chapter 1** was also a plea for caution when it comes to parameterizing these methods to suit the biological characteristics of the studied organisms. We hope that this advice will be heeded by the community and that further studies will provide data on the mutation and recombination rates of species, which is currently relatively scarce for many non-model organisms.

On another note, despite the low level of genetic divergence found within this species, and the relatively recent separation of the two populations, differences appeared in the genomes of the two populations. Indeed, although the quantity of repeated elements in the genomes was very similar, a difference of 5 Mb (1.5% of the total size) was found between the assemblies of the highland and the lowland population, which is as much as between the lowest value of *O. alexandrae* and *O. priamus*. The size of the genomes, and deeper studies in the non-coding parts in particular, are clearly in the outlook and could perhaps reveal, in *O. alexandrae*, differences that we have ignored until now.

Evolution of genome size among the Papilionidae family

During this thesis, I assembled 14 long-readbased genomes of Papilionidae, which have been used to carry out both population-scale studies and species-level phylogenomic studies, covering conservation genomics, genome structure and systematics. The study of genome evolution has long been closely dependent on advances in available sequencing technologies (Formenti et al. 2022). At first, we had access to a few well-known coding regions (Sanger), or

to thousands of enzyme restriction sites (RAD sequencing), but we had to wait for massive whole genome sequencing data (Shotgun Illumina) or long reads sequencing (PacBio, ONT) to study genome architectures of non-model organism. For instance, reference genomes enable the access of repetitive regions such as centromeres, microsatellites, transposable elements and segmental duplications (Bergman and Quesneville 2007, Ewing 2015, Miga 2020, Formenti et al, 2022). Furthermore, the widespread use of whole genome sequencing has made it possible to refine estimates of the size of species genomes, and to discover, for example, the staggering size of the genome of certain organisms such as the marbled lungfish (*Protopterus aethiopicus*, 130 Gb, Hidalgo et al. 2017) or the Japanese canopy plant (*Paris japonica*, 148 Gb, Pellicer et al. 2010).

The family Papilionidae, despite its relatively modest number of species (>600, about 10 times less than Nymphalidae), has the particularity of having a high variance in genome size (Liu et al. 2020). From the very small genomes of *Papilio* (~300 Mb, Lu et al. 2019) to the large genomes of *Parnassius* (> 1.2 Gb, Podsiadlowski et al. 2021, He et al. 2022), one of the long-standing questions is how to understand the origin of these size differences, which has been hypothesized as results of duplication, repeated elements, or insertion (Petrov 2001, Liu et al. 2020). During my thesis, I sequenced and assembled the genome of the endangered *Baronia brevicornis* (Papilionidae), sister species to all other swallowtails, and oldest lineage of all extant butterflies (Condamine et al. 2018, Allio et al. 2020a). As well as being of great importance for the conservation genomics of this species, the genome of this taxon standed out as a key lineage to study the genome size evolution across the family Papilionidae. In collaboration with Alba Marino (PhD student at ISEM, Montpellier), we explored the dynamics of repeated elements within Papilionidae. We have found that genome size variations among Papilionidae were mostly explained by transposable element (TE) activities, suggesting that large genomes are a derived feature in swallowtail butterflies. Overall, the TE activity was recent and involved different TE classes depending on the species (Figure 1; Figure 4 of **Appendix IV**), which is in accordance with other genomic studies (Petersen et al. 2019, Wu and Lu 2019) including *Heliconius* butterflies (Cicconardi et al. 2023). However, the non-repeat fraction of the genome was also variable among species and not closely following overall genome size. For example, the genome of *Lamproptera curius* (tribe Leptocircini) contains 32% of annotated TEs compared with 50% of TEs for *Sericinus montela* (tribe Parnassiini), which have a similar genome size (500 Mb). This work highlights that further research is needed to fully retrace the evolution of the Papilionidae genome, maybe by the addition of missing taxa of this study.

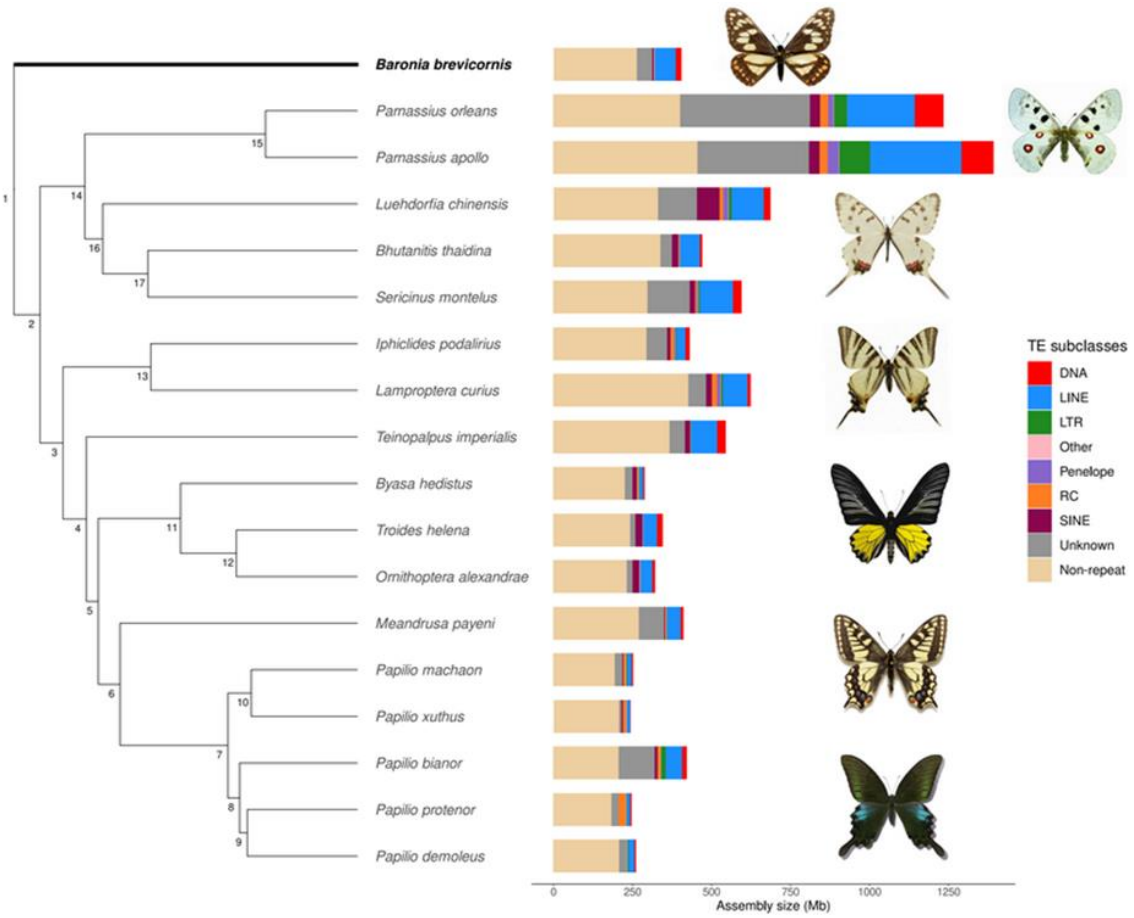


Figure 1: Variation in genome sizes (Mb) and transposable elements content across Papilioninae, Parnassiinae, and the monotypic Baroniinae represented by *Baronia brevicornis*. LINE, Long Interspersed Nuclear Elements; LTR, Long Terminal Repeats; RC, Rolling Circles; SINE, Short Interspersed Nuclear Elements. 'Other' includes simple repeats, microsatellites, and RNA sequences. Photos: Fabien L. Condamine.

The Graphium genomes conundrum

This high proportion of repeated elements has been shown to be one of the main factors in the large size of the swallowtail butterfly of the genus *Parnassius* (Podsialdowski et al. 2021, Marino et al. *under review* = **Appendix IV**) (**Figure 1**). On the other hand, genus *Graphium*, which also presents similar genome sizes, was not represented in our previous work (**Appendix IV**) so that the massive presence of repeated elements in this genus could only be hypothesized. At the beginning of my thesis, only a few shotgun fragmented *Graphium* genomes existed on NCBI (Allio et al. 2020a, Triant and Pirro 2023) or few size estimation (Liu et al. 2020), and they show a wide disparity in size estimates for this genus, ranging from 250 Mb to 1,265 Mb (**Table 1**). Furthermore, the average size of our draft shotgun assemblies was estimated around 1,000 Mb (**Table 1**).

Table 1: Sizes of *Graphium* genome assemblies in the literature (Allio et al. 2020a, Liu et al. 2020, Triant and Pirro 2023), and the mean assembly size of each subgenus based on all constituent species from our shotgun assemblies.

Genus	Subgenus	species	Assembly size (Mb)	Reference
Graphium	*(Arisbe)*	*colonna*	1264.678	Triant and Pirro 2023
Graphium	*(Graphium)*	*doson*	590.915	Allio et al. 2020a
Graphium	*(Graphium)*	*eurypylus*	1151.907	Liu et al. 2020
Graphium	*(Graphium)*	*sarpedon*	951.901	Liu et al. 2020
Graphium	*(Graphium)*	*doson*	413.025	Triant and Pirro 2023
Graphium	*(Graphium)*	*evemon*	788.461	Triant and Pirro 2023
Graphium	*(Pathysa)*	*antiphates*	333.008	Liu et al. 2020
Graphium	*(Pathysa)*	*nomius*	748.356	Liu et al. 2020
Graphium	*(Pathysa)*	*xenocles*	259.228	Liu et al. 2020
Graphium	*(Pathysa)*	*androcles*	933.096	Allio et al. 2020a
Graphium	*(Pazala)*	*confucius*	346.4429	Liu et al. 2020
Mean Megahit assembly size, before RagTag				
Graphium	*(Arisbe)*	n = 32	1,249 ± 544	This study
Graphium	*(Graphium)*	n = 31	968 ± 272	This study
Graphium	*(Pathysa)*	n = 23	1,093 ± 254	This study
Graphium	*(Pazala)*	n = 9	972 ± 151	This study

Surprisingly, our first draft assemblies of reference genomes (after the Flye assembly step) also showed a large disparity in size, ranging from 1,230 Mb to 2,602 Mb (**Table 2**). Although *Arisbe* and *Pathysa* subgenera assembly sizes were the biggest, as for the shotgun assemblies (**Table 1**), their sizes were almost as twice as expected. Moreover, the percentage of duplicated genes was high, reaching almost 40% for *G. antheus* (**Table 2**).

Table 2: Statistics of *Graphium* genome draft assemblies, using long-reads data and Flye assembler. For BUSCO scores, abbreviations stand for S: Simple, D: Duplicated, F: Fragmented, M: Missing.

Genus	Subgenus	species	Genome size (Mb)	Nb of contigs	BUSCO Score (%)			
					S	D	F	M
Graphium	*(Arisbe)*	*antheus*	2,293.699	49197	58.7	39.3	1.2	0.8
Graphium	*(Graphium)*	*agamemnon*	1,427.189	25705	82.3	15.9	1	0.8
Graphium	*(Graphium)*	*doson*	1,230.063	78046	89.1	7.7	1.7	1.5
Graphium	*(Pathysa)*	*antiphates*	2,602.652	64018	78.5	19.5	1.1	0.9

As the production of four new reference genomes for this genus appeared challenging, this provided an opportunity to supervise Daphnée Navratil, a volunteer L3 trainee, to explore this topic. We speculated that the large size of our assemblies and the associated duplicated genes could come from either an ancient whole-genome duplication shared by a part or all

Graphium (and could have been reduced in part of them), or, more likely, from transposable elements, and/or artificial issues in our genome assemblies (for example, homologous sequences kept by the assembler due to their high heterozygosity). We used several genomic approaches to determine the nature of these duplications, to estimate the "real" size of *Graphium* genomes and to improve our draft assemblies. First, we found that most duplicated genes were not shared by multiple *Graphium* assemblies, suggesting that most of the observed duplications were probably not the result of an ancient shared duplication event. Secondly, we found that these duplications could be greatly reduced with purge_dups 1.2.5 (Guan et al. 2020), a method using coverage and mapping information (**Table 3**), but remained consequent for the two largest assemblies (*G. antheus* and *G. antiphates*).

Table 3: Statistics of *Graphium* genome draft assemblies, using long-reads data and Flye assembler, purge_dups, and a polishing step with POLCA (Zimin and Salzberg 2020). For BUSCO scores, abbreviations stand for S: Simple, D: Duplicated, F: Fragmented, M: Missing.

Genus	Subgenus	species	Genome size (Mb)	Nb of contigs	BUSCO Score (%)			
					S	D	F	M
Graphium	*(Arisbe)*	*antheus*	1,686.625	22329	84.8	12.7	1.5	1.0
Graphium	*(Graphium)*	*agamemnon*	961.917	8471	96.1	1.7	1.1	1.1
Graphium	*(Graphium)*	*doson*	990.610	36748	92.7	3.9	1.8	1.6
Graphium	*(Pathysa)*	*antiphates*	2,052.517	28713	85.8	12	1.2	1.0

We conducted K-mer spectra approaches with short reads (GenomeScope, Ranallo-Benavidez et al. 2020; KAT, Mapleson et al. 2017) to estimate the level of heterozygosity of the genomes, and visualize the proportion of repeated k-mer in the long-reads assembly. But to our surprise, these two approaches revealed for at least three of the four individuals that the first peak of k-mer frequency (heterozygous) was less than half of the second peak (homozygous), which could hardly be due to a biological process, unless the ratio is 1:4 and it is coming from tetraploidy.

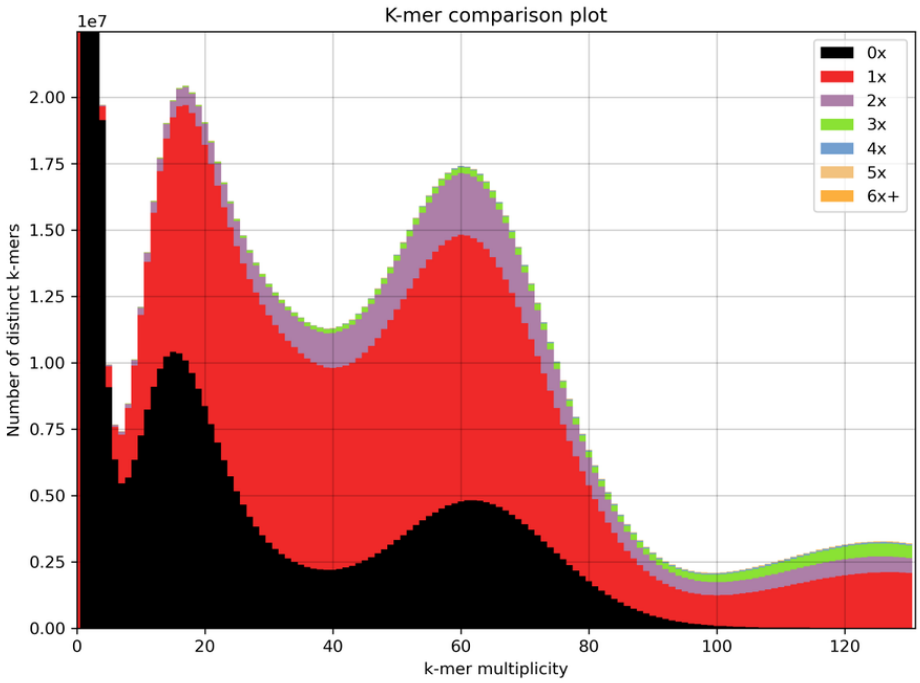


Figure 2: Example of KAT results of *G. agamemnon*. These are the frequency of k-mer (31 pb long) from the short-reads data. The colors correspond to the multiplicity of the k-mer on the (long-read-assembled) reference genome of the same individual. The black proportion represents all the kmer that are present in the short reads but not in the assembly. Red are the k-mer appearing only once in the assembly, other colors are for the k-mer appearing multiple times. Note that the first peak (heterozygous coverage) is *ca.* 18x, whereas the second peak (homozygous coverage) is 60x. This does not correspond to either a 1:2 ratio (diploid) nor to a 1:4 (tetraploid).

This lower-than-expected peak heterozygosity compared to the homozygous peak in short reads data could come from contamination of the short reads data. However, neither Blobtools (Laetsch and Blaxter 2017) nor MitoFinder (Allio et al. 2020b) revealed any contamination or cross-contamination in our samples that could explain such a pattern.

One of the surprises of this investigation was the apparent difference in information between the short-read data and the long-read data. This shows that more work is needed to understand the causes of these biases and their impact on genome assemblies. In any case, we have shown throughout this thesis and in particular in **Chapters 2.1** and **2.2** that reference genomes are a powerful tool in phylogenomics. Genome scaffolding was a crucial step in maximizing the amount of genes recovered from the genomes with, as expected, an increase in recovered genes with better reference genomes. However, this scaffolding step needs to be carried out with caution, since it assumes that the architecture of the reference and target genomes is similar, which might not be true in the case of many species pairs due to

rearrangement, duplication, or inversion (Edelman et al. 2019). However, it does seem to be an interesting tool for enhancing fragmented genomes obtained from museum specimens.

2. Interpreting and using phylogenies derived from a multitude of markers

Impact of topological uncertainties for diversification

In my thesis, we produced two almost complete phylogenies of two tribes that had no very robust phylogeny in the literature. The novelty and completeness of the sampling yielded unexpected results regarding the placements of some taxa in the two tribes, such as the relationship among the *Paranticopsis/Pathysa* clade, and the positions of *Byasa daemonius* and *B. crassipes*. Some of these cases could probably have gone unnoticed if they were not the subject of strong discordance in the multitude of gene trees, or did not represent a surprising story of the evolution of morphological traits and other life-history traits for example. The case of *Pathysa* and *Paranticopsis* is particularly interesting, although also particularly complex, and represents another example of the growing importance in the literature of taking into account of ILS and GF cases in analyses (Funk and Omland 2003, Jennings and Edwards 2005, Rheindt et al. 2009, Mallet et al. 2016, Galtier 2023). It was in this perspective that I carried out preliminary investigations with the data and tools at my disposal. However, the multiplicity of plausible scenarios at the end of the Aphid analysis clearly shows that additional data (e.g. non-coding regions) and methods (e.g. gene ontology), will be needed to better understand the evolutionary history of this group. Nonetheless, we can already wonder to what extent a slightly different topology might have had an impact on the macroevolutionary inferences of **Chapter 3**. In fact, as it was difficult to determine the location of the host-plant shift from Annonaceae to Aquifoliaceae, and as the current topology might not seem the most parsimonious in terms of trait reconstruction, we could have decided to constrain the monophyly of mimetic *Pathysa* and at the same time the host shift to Aquifoliaceae, as suggested by scenario D in Figure 9 of **Chapter 2.1**. Yet, this is not the only possible scenario inferred by our Aphid analyses, and it is clear that the genes would continue to provide discordant information for the molecular dating. The uncertainty of the node age estimated at the crown of this clade (**Chapter 3**, Figure S2) seems to illustrate this heterogeneity of evolutionary stories present in the genes and the difficulty for the molecular clock model to converge towards a unimodal node age.

Impact of taxonomic uncertainties for diversification

A particularly striking point of my thesis, and of the macroevolutionary studies I was able to carry out, is the importance of relying on taxonomy in phylogenetic inferences. The question: *How many valid species are in my group?* seems to be one of the most basic yet incredibly fundamental and complicated issues in phylogenetic and diversification studies (Utami et al. 2022, Dufresnes et al. 2023). I am not going to discuss again the origin of the difficulty in answering this question (which I detailed in my **General Introduction**), but I would like to come back to the impact that this could have on diversification studies. In my studies of both Troidini and Leptocircini, certain groups of species remain particularly disputed taxonomically, and could be the subject of a current taxonomic inflation, such as the subgenera *Protesilaus* and *Pazala*, or the *adamastor* species group in the subgenus *Arisbe*, or the genus *Troides*. The reason for the taxonomic uncertainties lies either in a general lack of knowledge of the group of species in question (e.g. many species are doubtful in the *adamastor* group but delicate to refute because of the difficulty of accessing specimens; Smith and Vane-Wright 2001) or in their apparent rapid radiation due to more or less recent and porous isolations (the *Troides* in the Sundaland Islands; Condamine et al. 2015). However, the diversification results in **Chapter 3**, particularly with the CLaDS and RevBayes analyses, have almost exactly pinpointed these groups as having the highest speciation rates, casting doubts on the taxonomic bias of these results. These results are not surprising since it has already been shown in a few previous studies that taxonomy has an impact on diversification analyses (Ruane et al. 2014, Faurby et al. 2016, Utami et al. 2022).

In my analyses, I carried out the diversification analyses on two taxonomic datasets (more or less stringent: “complete” and “lump” datasets), with the idea of limiting this impact. While the detected shifts changed for Leptocircini, the results for Troidini were very similar, although the magnitude of speciation rates of certain clades varied between taxonomic sets (e.g. from up to 0.39 events/Myr/lineage or to 0.27 events/Myr/lineage for *Troides*, with the complete or lump taxonomy respectively). This is reassuring, and also not very surprising given that these groups are indeed probably currently expanding or have undergone recent radiation (hence the difficulty in clarifying the taxonomy for some species/subspecies, which are in the “gray zone of speciation”).

These results show how important taxonomic revisions are for these still under-studied parts of the tree family, and that they must be made on solid foundations, and, as far as possible, on complete genome data to best represent the heterogeneity of evolutionary histories underlying a single genome. Defining what species are will probably always be difficult

because of the “speciation continuum”, but the whole genome data that we have produced, in addition to resequencing of individuals in predefined portion of the genome (capture sequencing), could open up the possibility of carrying out intraspecific and population-scale analyses. Subsequently, if our data allows, I would like to concentrate on the few clades that have been highlighted as doubtful in our analyses (typically the subgenera *Protesilaus* and the *Pazala*), in an attempt to redefine species boundaries. Galtier (2019) proposed an approach to pairwise calculation of indices of population polymorphism structure (such as F_{ST}), indices of genetic differentiation (D_a, D_{xy}; Nei 1987, Fraisse et al. 2021, De Jode et al. 2023), as well as analyses of population structure (PCA, admixture, structure), and to compare the value of these indices for pairs of species to be delimited with pairs of species well recognised by the community as a basis for comparison (Allio et al. 2021a). This approach could help determine these species complexes as objectively and integratively as possible, and finally gain more confidence in the results of diversification analyses.

3. Diversification of phytophagous insects

The role of host plant shifts in diversification

The number of empirical studies listed in the review by Jousselin and Elias (2019) from 1998 until today shows the extent to which the question of clade’s diversification in plant-insect interactions has not yet been resolved. Moreover, the diversity of the conclusions reached in these studies suggests that the answer is likely both complex and clade-specific. We have seen in this thesis that the hypotheses tested are somehow based on shaky and fragile grounds: the number of *Pazala* species modifies the conclusions about the diversification of the clade, the topological and age uncertainties within the *Pathysa* clade could have changed the evolutionary histories of traits (mimicry, host plant shift) and thus modified our conclusions.

Jousselin and Elias (2019) reviewed a series of comparative phylogenetic tools to test host-plant mediated speciation at a macroevolutionary scale, which are 1) comparisons of the phylogenetic histories of insects and their associated plants (Brown 1995), 2) ancestral character state estimations, and 3) diversification analyses. In this thesis we used the two last tools to test hypotheses of host-plant mediated diversification by inferring ancestral host-plant preference and correlated it with diversification dynamics of shifting and non-shifting clades. In the end, our results were not conclusive on the impact of host-plant shifts on the diversification of the Leptocircini. From these results, I have the feeling that we have either 1) lacked finesse in our divisions, by not detailing which *genus* or subdivision was eaten within

the families Annonaceae or Aristolochiaceae, for example, and that the "niche conservatism" we assumed in both tribes might have masked interesting shifts to be tested, or 2) lacked of generality, and that families of toxins would have been more relevant to test than families of plants (Kergoat et al. 2005). The taxonomic “family” level of plants surely reflects the human-based taxonomic divisions based on the morphology and genetic divergence of species, but perhaps reflect poorly the differences in plant defense mechanisms, which is the real selection pressure applied to insects.

In our study, we did not undertake a comparison of the phylogenetic history of insects and associated plants (cospeciation comparative analyses), because this would have required 1) having the genetic resources needed to infer and date the phylogeny of several plant clades, 2) to know more precisely the association of the insects with the species (or at least the genera) of the plants eaten, and 3) to be able to carry out the joint biogeographical reconstruction of the two phylogenies with similar parameters, to avoid methodological biases as much as possible (e.g. Cruaud et al. 2012, Allio et al. 2021b). A complete and well-sampled phylogeny of the Annonaceae was still lacking during my thesis (but see Couvreur et al. 2011, Li et al. 2023), but the recent work in the ERC GLOBAL (led by Thomas Couvreur) will make it possible to better understand the evolutionary history of Annonaceae. The recent study on the African rainforest tribe Monodoreae (Dagallier et al. 2023) provides an interesting source of phylogenetic data to link with the African Leptocircini.

The second perspective on the impact of host-plant shifts in swallowtail diversification is based on the production of the 300 whole genomes that we generated. In the study by Allio et al. (2021b), the genomic analyses of molecular evolution following host-plant shifts, by calculating dN/dS ratio, was based on a phylogenetic framework including all genera and a total of 44 complete genomes across Papilionidae, resulting in 28 testable nodes (of which 14 were control branches and only 5 were for climate-preference shifts). Here, the drastic increase of whole genome data gained for these two tribes, as well as the robustness of the new phylogenetic frameworks, would make it possible to reproduce this genome-wide analysis for the Leptocircini and eventually, in comparison with the Troidini. This perspective is one of the long-term prospects of the ERC GAIA (led by Fabien Condamine), in collaboration with Noémie Hévin's PhD thesis on the subfamily Parnassiinae, the addition of Papilionini to the subfamily Papilioninae by myself for the phylogenomic framework, and Gwenaelle Vigo's PhD thesis working on designing a pipeline for dN/dS analyses implementing a new selection of orthologous genes.

The role of biogeography

Similarly to the delimitation of host plants in large family levels, the regional delimitation we applied in the biogeography analyses was at global scale, and represented broadly all the major regions of the world, and in particular the three major tropical regions in which our swallowtail butterflies diversified: the Afrotropics (two regions with Madagascar), the Indo-Australasian Archipelago (IAA, four regions), and the Neotropics (three regions). These biogeographical analyses revealed interesting global patterns of biogeographical events (vicariance, dispersion, and local extinction), which seemed to be common to both tribes, but quite specific to each region (**Table 1** of **Chapter 3**). For example, the elevated proportion of vicariance found in the fragmented insular region that is the IAA seems very consistent with general patterns of speciation in insular regions (Paulay and Meyer 2002, de Bruyn et al. 2014) and has been found in many groups (Condamine et al. 2013, Condamine et al. 2015).

In fact, dividing the entire world into 13 regions, with such a large number of taxa, is quite a technical achievement thanks to the DECX model (Beeravolu and Condamine 2016, Kawahara et al. 2023). However, if we consider that the Neotropical region alone can be divided into 10 distinct regions (e.g. Antonelli et al. 2018, Chazot et al. 2019), and Wallacea into 9 regions (e.g. de Bruyn et al. 2014, Condamine et al. 2015), we could hypothesize that further refinement of these regional delimitations for our datasets would reveal probably interesting fine-scale patterns of biogeographic mode of speciation and colonization, possibly common to both tribes. We could thus clarify these vicariance patterns that we observed at a global scale, especially between insular Southeast Asia (Sunda) and Australasia (Sahul), encompassing Papua New Guinea, Australia and the surrounding islands. In addition, by delimiting the Neotropics more precisely, we might also observe more vicariance or dispersal events that would have taken place within this region (Antonelli et al. 2018), on the same principle of "terrestrial islands" (patches of forest or savannah surrounded by other types of vegetation, Prance 1996) or "sky islands" represented by the Andes (Hughes and Eastwood 2006), in the same way as mountain ranges have similar "functional" island effect for *Parnassius* in the Alps (Kebaïli et al. 2021). It would therefore be easy to study regional biogeography by focussing on specific clades in the Neotropics and compare diversification, vicariance processes and dispersal routes of *Parides* (Troidini) and *Eurytides* (Leptcoricini) for example. Likewise, we can focus on the IAA region with the *Atrophaneura*, *Pachliopta*, birdwing butterflies (Troidini) and *Graphium s.l.* (Leptocircini). Studying and comparing these groups could provide interesting information on the diversification of organisms in these two major diversification hotspots.

Disentangling with confounding effects

In this thesis, I used the differences and similarities of two tribes of Papilionidae with the idea of adding controls that would enable distinguishing confounding factors from host-plant shifts. We limited our analyses of **Chapter 3** to biogeographical changes (and mimicry in a superficial way), but the list of possible confounding effects could easily be extended to changes in biome (Gamboa et al. 2022), wing morphology (e.g. size, presence of tail; Laurila et al. 2008, Sheard et al. 2020, Chotard 2022), or wing patterns (e.g. mimicry, color, aposematism, sexual dimorphism; Bishop et al. 2016; Xing et al. 2016; Pinkert et al. 2017; Stelbrink et al. 2019, Puissant et al. 2023). And even more than the confounding effects of the coincidence of synchrony between these changes and the host-plant shifts, it would not be surprising if one of these changes were sufficient to trigger the others on a macroevolutionary scale. If host-plant shifts constitute a major ecological change by opening up new exploitation niches (Ehrlich and Raven 1964, Mitter et al. 1988), they may then constitute the starting point for a cascade of changes in other traits such as those mentioned above (Allio et al. 2021b). For example, Chazot et al. (2021) attributed the phenotypic diversification of the Neotropical nymphalid genus *Morpho* as a result of a major host-plant shift in this group.

During the first year of my thesis, I supervised a preliminary study carried out by Eleanor Sheridan (first year Master student of the Erasmus Mundus Master Programme in Evolutionary Biology) at the scale of the family Papilionidae, incorporating data on host plants, as well as biome preferences, regional distributions, and morphological trait (size and color), to explore these confounding or sequentially triggered diversification (**Appendix V** for her complete Master report). In this study, Eleanor re-assessed and assembled a comprehensive database of host-plant preference, wingspan, coloration, and distribution across regions and biomes for all extant Papilionidae (**Figure 3**). Morphological data consisted of wingspan and coloration. Wingspan - the distance between the tips of the two forewings - was obtained through a literature search. Coloration data was generated in the form of pairwise comparisons between species in collaboration with Stefan Pinkert (unpublished data) following the methods of Weller and Westneat (2019), based on mean RGB (Red Green Blue) values of pixels on calibrated Papilionidae pictures. This resulted in a dataset containing pairwise comparisons between every Papilionidae species (See **Appendix V** for more details). Using the time-calibrated phylogeny of 421 Papilionidae species inferred with Sanger data from Allio et al. (2021b), we estimated the ancestral states of traits and determined the timing of changes in these traits. Firstly, the idea was to assess the frequency with which changes in biome or geographic region occurred in parallel with host-plant shifts. Secondly, we wanted to examine

whether the colonization of new host plants increased the rate of change in wing coloration or wingspan.

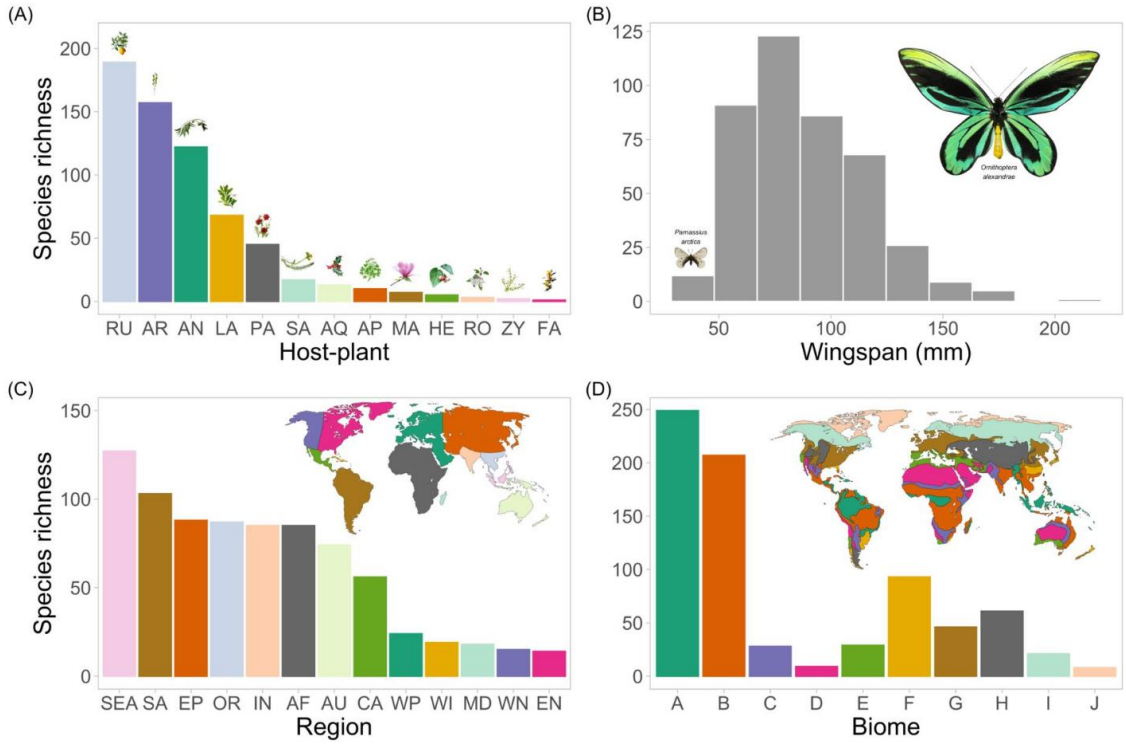


Figure 3: Diversity in the extant Papilionidae species. **(A)** The frequency of species feeding on each of the principal host-plant families: Annonaceae (AN), Apiaceae (AP), Aquifoliaceae (AQ), Aristolochiaceae (AR), Crassulaceae (SA), Fabaceae (FA), Hernandiaceae (HE), Lauraceae (LA), Magnoliaceae (MA), Papaveraceae (PA), Rosaceae (RO), Rutaceae (RU), and Zygophyllaceae (ZY). **(B)** A histogram of wingspans of all extant species. Images of butterflies are sized relative to each other: Parnassius arctica is the smallest Papilionidae species and Ornithoptera alexandrae the largest. **(C)** The frequency of species in each geographic region: Western Palearctic (WP), Eastern Palearctic (EP), Western Nearctic (WN), Eastern Nearctic (EN), Central America (CA), Caribbean Islands (WI), South America (SA), Africa (AF), Madagascar and surrounding islands (MA), India (IN), continental South East Asia (OR), insular Southeast Asia (SEA), and Australia and surrounding islands (AU). The inset map shows how we defined regions, with colors on the map corresponding to the colors of the bars. **(D)** The frequency of species in each biome: evergreen equatorial rainforest (A), tropical deciduous woodland (B), savanna (C), subtropical desert (D), sclerophyllous woodland and shrubland (E), temperate evergreen forest (F), broadleaf deciduous forest (G), steppe (H), boreal coniferous forest (I), and tundra (J). The inset map shows how the location of each biome was defined (Olson et al. 2001). Species feeding on multiple plant families or found in multiple regions or biomes are represented in multiple bars.

This preliminary study showed that colonization of new host plants often occurred in tandem with expansion into new regions (**Figure 4**). This suggests that some of the signals of selection present around host-plant shifts (Allio et al. 2021b) may also have been a result of adaptation to novel selection pressures imposed by a new region. Whether this association was causative or purely correlational is not something that can be easily evaluated. However, it is plausible that there was a causative link between at least some of the co-occurrences of host-

plant gains and regional expansions. Perhaps expansion into a new region enabled colonization of a host plant that was not present in the ancestral region. For instance, the colonization of Hernandiaceae and Annonaceae at the crown of Leptocircini coincided with expansion into the Oriental region from the Eastern Palearctic. With neither plant family present in this ancestral region (Michalak et al. 2010, Couvreur et al. 2011), the expansion into the Oriental region likely facilitated colonization of these novel host plants. On the other hand, species were found to be more similar in coloration if they had a shared history of host-plant preference (**Figure 5**).

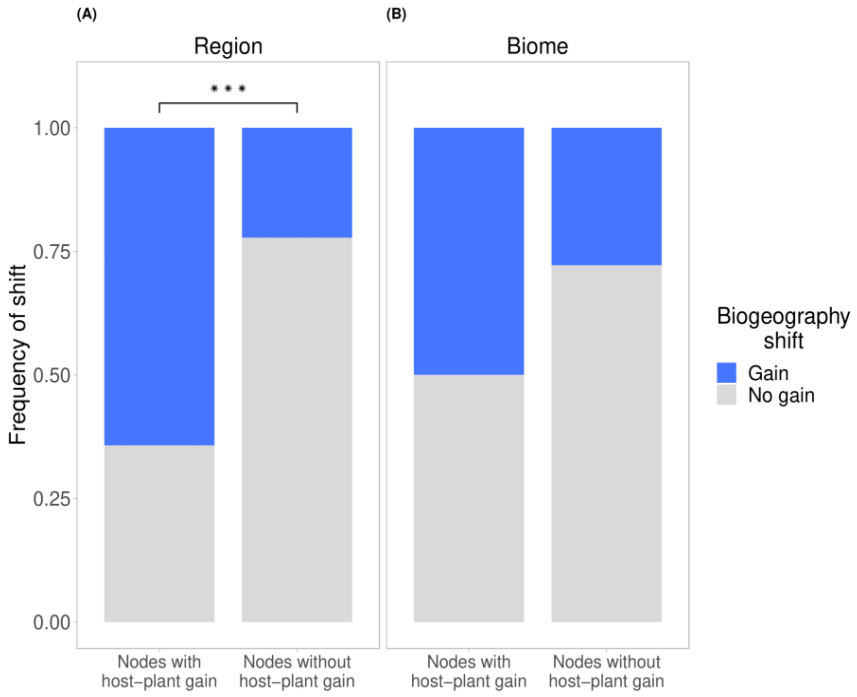


Figure 4: Figure 4: Expansion into new regions and biomes. The frequency of expansion into new regions (A) and biomes (B) for host-plant-gain nodes (n = 13) was compared to that for all other nodes (n = 828) using a chi-squared test.

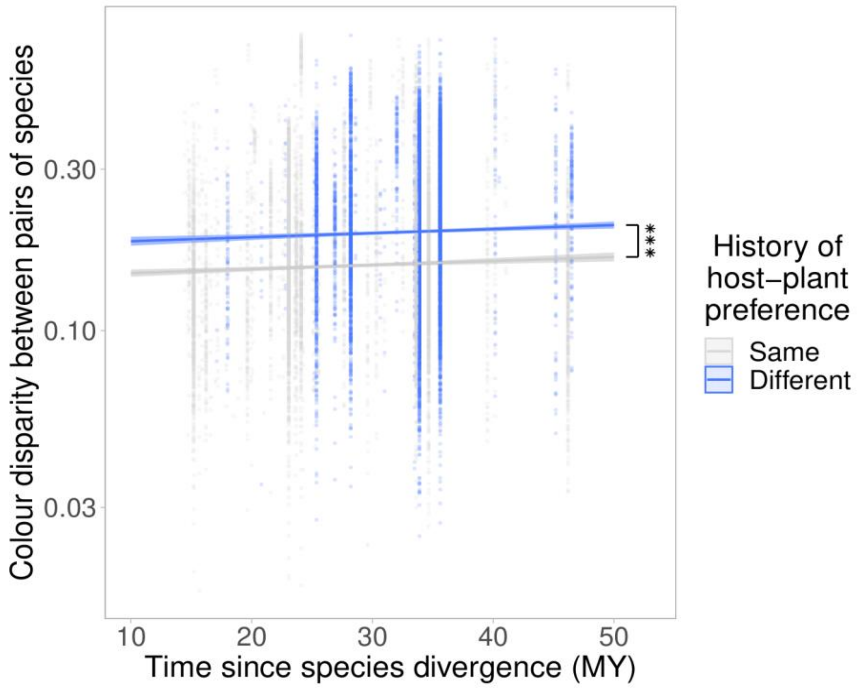


Figure 5: The effect of historical host-plant preference on modern day color disparity. Species pairs separated by at least one host-plant shift (blue, *n* = 8384) were compared to those pairs where no host-plant shifts have occurred for either since divergence (grey, *n* = 7526), accounting for the time since the species diverged.

It has been suggested that toxic species evolve resemblance to each other (Mullerian mimicry: Müller 1878, Sherratt 2008) to send consistent signals to predators. Perhaps, at least for species feeding on toxic host plants, this has contributed to those species sharing a host plant more closely resembling each other (ecological interactions may promote convergence in sympatric species; Puissant et al. 2023). Additionally, this resemblance might be indicative of rapid color evolution following a host-plant shift, resulting in lineages that diverged in host-plant preference also diverging in color. This could come about if different host-plants provided different selective pressures. For instance, shifting to a more toxic plant could have led to selection for aposematic coloration (Mallet and Joron 1999), and thus a divergence in color from a sister lineage that fed on non-toxic plants. However, this does seem to be at odds with the high levels of Batesian mimicry within the family, with non-toxic species mimicking toxic ones (Kunte 2009). Interestingly, the exact opposite was found for wingspan, with slower rates of wingspan evolution directly following colonization of a new host plant. The reasons for this are not clear. Moreover, it is the opposite pattern to that found in *Morpho* butterflies, where the most pronounced changes in wingspan followed host-plant shifts (Chazot et al. 2021). Perhaps host-plant shifts acted as bottlenecks (Munday et al. 2004), with only very few species

switching their preference. The resulting lack of genetic diversity might have slowed the rate of evolution (Barrett and Schluter 2008). If this was the case, it would imply that there must have been strong selective pressures on coloration to enable it to overcome this constraint and rapidly evolve. Regardless of the underlying reasons, these findings suggest that some of the genetic signals of selection seen in swallowtails around host-plant shifts (Allio et al. 2021b) may be in genes involved in controlling coloration but not in genes underpinning size.

Nevertheless, this study was based on a non-optimal dated phylogeny and sometimes incomplete trait data, giving results that were only preliminary. Since then, compilations of additional species-specific data on host-plants, biogeographic occurrences and biome preferences by us, and data from museum specimens of the Papilionidae by Ariane Chotard for wing morphology, and Agathe Puissant for color and reflectance and mimicry would constitute invaluable data to be put in relation with new complete phylogenies to repeat and complete these macroevolutionary analyses.

References

Allio R, Nabholz B, Wanke S, Chomicki G, Pérez-Escobar OA, Cotton AM, Clamens AL, Kergoat GJ, Sperling FAH, Condamine FL. 2021b. Genome-wide macroevolutionary signatures of key innovations in butterflies colonizing new host plants. *Nat. Commun.* **12**:354.

Allio R, Schomaker-Bastos A, Romiguier J, Prosdocimi F, Nabholz B, Delsuc F. 2020b. MitoFinder: efficient automated large-scale extraction of mitogenomic data in target enrichment phylogenomics. *Mol. Ecol. Res*. **20**:892–905.

Allio R, Scornavacca C, Nabholz B, Clamens AL, Sperling FA, Condamine FL. 2020a. Whole genome shotgun phylogenomics resolves the pattern and timing of swallowtail butterfly evolution. *Syst. Biol.* **69**:38-60.

Allio R, Tilak MK, Scornavacca C, Avenant NL, Kitchener AC, Corre E, *et al.* 2021a. High-quality carnivoran genomes from roadkill samples enable comparative species delineation in aardwolf and bat-eared fox. *eLife* **10**:e63167.

Antonelli A, Zizka A, Carvalho FA, Scharn R, Bacon CD, Silvestro D, Condamine FL. 2018. Amazonia is the primary source of Neotropical biodiversity. *Proc. Natl Acad. Sci. USA.* **115**:6034-6039.

Barrett RDH, Schluter D. 2008. Adaptation from standing genetic variation. *Trends Ecol. Evol.* **23**:38–44.

Beeravolu CR, Condamine FL. 2016. An extended maximum likelihood inference of geographic range evolution by dispersal, local extinction and cladogenesis. *bioRxiv* https://doi.org/10.1101/038695

Bergman CM, Quesneville H. 2007. Discovering and detecting transposable elements in genome sequences. *Brief. Bioinform.* **8**:382-392.

Bishop TR, Robertson MP, Gibb H, Van Rensburg BJ, Braschler B, Chown SL, Foord SH, Munyai TC, Okey I, Tshivhandekano PG. 2016. Ant assemblages have darker and larger members in cold environments. *Global Ecol. Biogeogr*. **25**:1489–1499.

Brown KS, Klitzke CF, Berlingeri C, Dos Santos PER, Scriber JM, Tsubaki Y, Lederhouse RC. 1995. Neotropical swallowtails: chemistry of food plant relationships, population ecology, and biosystematics. In: *Swallowtail Butterflies: Their Ecology and Evolutionary Biology* (Eds. Scriber JM, Tsubaki Y, Lederhouse RC). Scientific Publishers, Gainesville, Florida. pp. 405-446.

Chazot N, Blandin P, Debat V, Elias M, Condamine FL. 2021. Punctuational ecological changes rather than global factors drive species diversification and the evolution of wing phenotypes in Morpho butterflies. *J. Evol. Biol.* **34**:1592-1607.

Chazot N, Willmott KR, Lamas G, Freitas AVL, Piron-Prunier F, Arias CF, Mallet J, De-Silva DL, Elias M. 2019. Renewed diversification following Miocene landscape turnover in a Neotropical butterfly radiation. *Global Ecol. Biogeogr.* **28**:1118-1132.

Chotard A. 2022. Evolution of tails in swallowtail butterflies (Papilionidae, Lepidoptera): macro-evolutionary and experimental approaches. Doctoral dissertation, Muséum National d'Histoire Naturelle.

Cicconardi F, Milanetti E, Pinheiro de Castro EC, Mazo-Vargas A, Van Belleghem SM, Ruggieri AA, Rastas P, Hanly J, Evans E, Jiggins CD, *et al.* 2023. Evolutionary dynamics

of genome size and content during the adaptive radiation of Heliconiini butterflies. *Nat. Commun.* **14**:5620.

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

Condamine FL, Toussaint EFA, Clamens AL, Genson G, Sperling FAH, Kergoat GJ. 2015. Deciphering the evolution of birdwing butterflies 150 years after Alfred Russel Wallace. *Sci. Rep.* **5**:11860.

Condamine FL, Toussaint EFA, Cotton AM, Genson GS, Sperling FAH, Kergoat GJ. 2013. Fine-scale biogeographic and temporal diversification processes of peacock swallowtails (*Papilio* subgenus *Achillides*) in the Indo-Australian Archipelago. *Cladistics* **29**:88–111.

Couvreur TL, Pirie MD, Chatrou LW, Saunders RM, Su YC, Richardson JE, Erkens RH. 2011. Early evolutionary history of the flowering plant family Annonaceae: steady diversification and boreotropical geodispersal. *J. Biogeogr.* **38**:664-680.

Cruaud A, Rønsted N, Chantarasuwan B, Chou LS, Clement WL, Couloux A, *et al.* 2012. An extreme case of plant–insect codiversification: figs and fig-pollinating wasps. *Syst. Biol.* **61**:1029-1047.

Dagallier LPM, Condamine FL, Couvreur TL. 2023. Sequential diversification with Miocene extinction and Pliocene speciation linked to mountain uplift explains the diversity of the African rain forest clade Monodoreae (Annonaceae). *Ann. Bot.* mcad130.

de Bruyn M, Stelbrink B, Morley RJ, Hall R, Carvalho GR, Cannon CH, van den Bergh G, Meijaard E, Metcalfe I, Boitani L, *et al.* 2014. Borneo and Indochina are major evolutionary hotspots for Southeast Asian biodiversity. *Syst. Biol.* **63**:879-901.

de Jode A, Le Moan A, Johannesson K, Faria R, Stankowski S, Westram AM, Butlin RK, Rafajlović M, Fraïsse C. 2023. Ten years of demographic modelling of divergence and speciation in the sea. *Evol. Appli.* **16**:542-559

Dufresnes C, Poyarkov N, Jablonski D. 2023. Acknowledging more biodiversity without more species. *Proc. Natl Acad. Sci. USA.* **120**:e2302424120.

Edelman NB, Frandsen PB, Miyagi M, Clavijo B, Davey J, Dikow RB, García-Accinelli G, Van Belleghem SM, Neafsey DE, *et al*. 2019. Genomic architecture and introgression shape a butterfly radiation. *Science* **366**:594-599.

Ehrlich PR, Raven PH. 1964. Butterflies and plants: a study in coevolution. *Evolution* **18**:586-608.

Ewing, A. D. 2015. Transposable element detection from whole genome sequence data. *Mob. DNA.* **6**:1-9.

Faurby S, Eiserhardt WL, Svenning JC. 2016. Strong effects of variation in taxonomic opinion on diversification analyses. *Meth. Ecol. Evol.* **7**:4-13.

Formenti G, Theissinger K, Fernandes C, Bista I, Bombarely A, Bleidorn C, Ciofi C, Crottini A, Godoy JA, Höglund J, *et al*. 2022. The era of reference genomes in conservation genomics. *Trends Ecol Evol.* **37**:197–202.

Fraïsse C, Popovic I, Mazoyer C, Spataro B, Delmotte S, Romiguier J, Loire E, Simon A, Galtier N, Duret L, *et al.* 2021. DILS: Demographic inferences with linked selection by using ABC. *Mol. Ecol. Res.* **21**:2629-2644.

Fuller ZL, Mocellin VJ, Morris LA, Cantin N, Shepherd J, Sarre L, Peng J, Liao Y, Pickrell PA, Matz M, et al. 2020. Population genetics of the coral *Acropora millepora*: Toward genomic prediction of bleaching. *Science* **369**:eaba4674.

Funk DJ, Omland KE. 2003. Species-level paraphyly and polyphyly: frequency, causes, and consequences, with insights from animal mitochondrial DNA. *Annu. Rev. Ecol. Evol. Syst.* **34**:397-423.

Galtier N. 2019. Delineating species in the speciation continuum: A proposal. *Evol. Appli.* **12**:657-663.

Galtier N. 2023. Phylogenetic conflicts: distinguishing gene flow from incomplete lineage sorting. *bioRxiv*, https://doi.org/10.1101/2023.07.06.547897

Gamboa S, Condamine FL, Cantalapiedra JL, Varela S, Pelegrín JS, Menéndez I, Blanco F, Hernández Fernández M. 2022. A phylogenetic study to assess the link between biome specialization and diversification in swallowtail butterflies. *Glob. Chang. Biol.* **28**:5901-5913.

García-Berro A, Talla V, Vila R, Wai HK, Shipilina D, Chan KG, Pierce NE, Backström N, Talavera G. 2023. Migratory behavior is positively associated with genetic diversity in butterflies. *Mol. Ecol.* **32**:560-574.

Guan D, McCarthy SA, Wood J, Howe K, Wang Y, Durbin R. 2020. Identifying and removing haplotypic duplication in primary genome assemblies. *Bioinformatics* **36**:2896-2898.

Hazzouri KM, Sudalaimuthuasari N, Kundu B, Nelson D, Al-Deeb MA, Le Mansour A, Spencer JJ, Desplan C, Amiri KMA. 2020. The genome of pest *Rhynchophorus ferrugineus* reveals gene families important at the plant-beetle interface. *Commun. Biol.* **3**:323.

He JW, Zhang R, Yang J, Chang Z, Zhu LX, Lu SH, Xie FA, Mao JL, Dong ZW, Liu GC, *et al*. 2022. High-quality reference genomes of swallowtail butterflies provide insights into their coloration evolution. *Zool. Res.* **43**:367-379.

Hidalgo O, Pellicer J, Christenhusz M, Schneider H, Leitch AR, Leitch IJ. 2017. Is there an upper limit to genome size?. *Trends Plant Sci.* **22**:567-573.

Hughes C, Eastwood R. 2006. Island radiation on a continental scale: exceptional rates of plant diversification after uplift of the Andes. *Proc. Natl Acad. Sci. USA* **103**:10334-10339.

Jennings WB, Edwards SV. 2005. Speciational history of Australian grass finches (*Poephila*) inferred from thirty gene trees. *Evolution* **59**:2033-2047.

Jousselin E, Elias M. 2019. Testing host-plant driven speciation in phytophagous insects: a phylogenetic perspective. *PCi Evol. Biol.* https://hal.science/hal-02400722.

Kawahara AY, Storer C, Carvalho APS, Plotkin DM, Condamine FL, Braga MP, *et al.* 2023. A global phylogeny of butterflies reveals their evolutionary history, ancestral hosts and biogeographic origins. *Nat. Ecol. Evol.* **7**:903-913.

Kebaïli C, Sherpa S, Rioux D, Després L. 2022. Demographic inferences and climatic niche modelling shed light on the evolutionary history of the emblematic cold-adapted Apollo butterfly at regional scale. *Mol. Ecol.* **31**:448-466.

Kergoat GJ, Delobel A, Fédière G, Le Rü B, Silvain JF. 2005. Both host-plant phylogeny and chemistry have shaped the African seed-beetle radiation. *Mol. Phylogenet. Evol.* **35**:602-611.

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

Laetsch DR, Blaxter ML. 2017. BlobTools: Interrogation of genome assemblies. *F1000Research* **6**:1287.

Laurila A, Lindgren B, Laugen AT. 2008. Antipredator defenses along a latitudinal gradient in *Rana Temporaria*. *Ecology* **89**:1399–1413.

Li H, Durbin R. 2011. Inference of human population history from individual whole-genome sequences. *Nature* **475**:493–496.

Li W, Wang R, Liu MF, Folk RA, Xue B, Saunders RM. 2023. Climatic and biogeographic processes underlying the diversification of the pantropical and early divergent angiosperm family Annonaceae. *bioRxiv*. https://doi.org/10.1101/2023.08.08.549183.

Liu G, Chang Z, Chen L, He J, Dong Z, Yang J, Lu S, Zhao R, Wan W, Ma G. 2020. Genome size variation in butterflies (Insecta, Lepidoptera, Papilionoidea): a thorough phylogenetic comparison. *Syst. Entomol.* **45**:571–582.

Lu S, Yang J, Dai X, Xie F, He J, Dong Z, Mao J, Liu G, Chang Z, Zhao R, *et al.* 2019. Chromosomal-level reference genome of Chinese peacock butterfly (*Papilio bianor*) based on third-generation DNA sequencing and Hi-C analysis. *GigaScience* **8**:giz128.

Lynch M, Ackerman MS, Gout JF, Long H, Sung W, Thomas WK, Foster PL. 2016. Genetic drift, selection and the evolution of the mutation rate. *Nat. Rev. Genet.* **17**:704-714.

Mallet J, Besansky N, Hahn MW. 2016. How reticulated are species? *BioEssays* **38**:140–149.

Mallet J, Joron M. 1999. Evolution of diversity in warning color and mimicry: polymorphisms, shifting balance, and speciation. *Annu. Rev. Ecol. Evol. Syst.* **30**:201–233.

Mapleson D, Garcia Accinelli G, Kettleborough G, Wright J, Clavijo BJ. 2017. KAT: a K-mer analysis toolkit to quality control NGS datasets and genome assemblies. *Bioinform.* **33**:574-576.

McVean GA, Cardin NJ. 2005. Approximating the coalescent with recombination. *Proc. R. Soc. B: Biol. Sci*. **360**:1387-1393.

Michalak I, Zhang LB, Renner SS. 2010. Trans-Atlantic, trans-Pacific and trans-Indian Ocean dispersal in the small Gondwanan Laurales family Hernandiaceae. *J. Biogeogr.* **37**:1214–1226.

Miga KH. 2020. Centromere studies in the era of 'telomere-to-telomere'genomics. *Exp. Cell Res.* **394**:112127.

Mitter C, Farrell B, Wiegmann B. 1988. The phylogenetic study of adaptive zones: has phytophagy promoted insect diversification? *Am. Nat.* **132**:107-128.

Morin PA, Archer FI, Avila CD, Balacco JR, Bukhman YV, Chow W, Fedrigo O, Formenti G, Fronczek JA, Fungtammasan A, *et al.* 2021. Reference genome and demographic history of the most endangered marine mammal, the vaquita. *Mol. Ecol. Res.* **21**:1008-1020.

Müller F. 1878. Über die vortheile der mimicry bei schmetterlingen. *Zool. Anz*. **1**:54–55.

Munday PL, van Herwerden L, Dudgeon CL. 2004. Evidence for sympatric speciation by host shift in the sea. *Curr. Biol.* **14**:1498–1504.

Nei M. 1987. *Molecular Evolutionary Genetics*. Columbia University Press.

Paulay G, Meyer C. 2002. Diversification in the tropical Pacific: comparisons between marine and terrestrial systems and the importance of founder speciation. *Int. Comp. Biol*. **42**:922-934.

Pellicer J, Fay MF, Leitch IJ. 2010. The largest eukaryotic genome of them all? *Bot. J. Linn. Soc*. **164**:10-15.

Petersen M, Armisén D, Gibbs RA, Hering L, Khila A, Mayer G, Richards S, Niehuis O, Misof B. 2019. Diversity and evolution of the transposable element repertoire in arthropods with particular reference to insects. *BMC Ecol. Evol.* **19**:11.

Petrov DA. 2001. Evolution of genome size: new approaches to an old problem. *Trends Genet*. **17**:23-28.

Pinkert S, Brandl R, Zeuss D. 2017. Colour lightness of dragonfly assemblages across North America and Europe. *Ecography* **40**:1110–1117.

Podsiadlowski L, Tunström K, Espeland M, Wheat CW. 2021. The genome assembly and annotation of the Apollo butterfly *Parnassius apollo*, a flagship species for conservation biology. *Genome Biol. Evol.* **13**:evab122.

Prance GT. 1996. Islands in Amazonia. *Proc. R. Soc. B: Biol. Sci.* **351**:823-833.

Puissant A, Chotard A, Condamine FL, Llaurens V. 2023. Convergence in sympatric swallowtail butterflies reveals ecological interactions as a key driver of worldwide trait diversification. *Proc. Natl Acad. Sci. USA* **120**:e2303060120.

Ranallo-Benavidez TR, Jaron KS, Schatz MC. 2020. GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. *Nat. Commun.* **11**:1432.

Rheindt FE, Christidis L, Norman JA. 2009. Genetic introgression, incomplete lineage sorting and faulty taxonomy create multiple cases of polyphyly in a montane clade of tyrant-flycatchers (*Elaenia*, Tyrannidae). *Zoologica Scripta* **38**:143-153.

Robinson JA, Kyriazis CC, Nigenda-Morales SF, Beichman AC, Rojas-Bracho L, Robertson KM, Fontaine MC, Wayne RK, Lohmueller KE, Taylor BL, *et al*. 2022. The critically endangered vaquita is not doomed to extinction by inbreeding depression. *Science* **376**:635-639.

Ruane S, Bryson Jr, RW, Pyron RA, Burbrink FT. 2014. Coalescent species delimitation in milksnakes (genus *Lampropeltis*) and impacts on phylogenetic comparative analyses. *Syst. Biol.* **63**:231-250.

Sellinger TPP, Abu Awad D, Moest M, Tellier A. 2020. Inference of past demography, dormancy and self-fertilization rates from whole genome sequence data. *PLoS Genet.* **16**:e1008698.

Sellinger TPP, Abu Awad D, Tellier A. 2021. Limits and convergence properties of the sequentially Markovian coalescent. *Mol. Ecol. Res.* **21**:2231-2248.

Sheard C, Neate-Clegg MHC, Alioravainen N, Jones SEI, Vincent C, MacGregor HEA, Bregman TP, Claramunt S, Tobias JA. 2020. Ecological drivers of global gradients in avian dispersal inferred from wing morphology. *Nat. Commun.* **11**:2463.

Smith CR, Vane-Wrigth RI. 2001. A review of the Afrotropical species of the genus *Graphium* (Lepidoptera: Rhopalocera: Papilionidae). *Bull. Nat. Hist. Mus. Lond. (Ent.)* **70**:503-719.

Stapley J, Feulner PG, Johnston SE, Santure AW, Smadja CM. 2017. Variation in recombination frequency and distribution across eukaryotes: patterns and processes. *Phil. Trans. R. Soc. B* **372**:20160455.

Stelbrink P, Pinkert S, Brunzel S, Kerr J, Wheat CW, Brandl R, Zeuss D. 2019. Colour lightness of butterfly assemblages across North America and Europe. *Sci. Rep.* **9**:1760.

Triant D, Pirro S. 2023. The complete genome sequences of 9 species of swallowtail butterflies (Papilionidae, Lepidoptera). *Biodiv. Gen.* https://doi.org/10.56179/001c.73927

Utami CY, Sholihah A, Condamine FL, Thébaud C, Hubert N. 2022. Cryptic diversity impacts model selection and macroevolutionary inferences in diversification analyses. *Proc. R. Soc. B: Biol. Sci.* **289**:20221335.

Waldvogel AM, Wieser A, Schell T, Patel S, Schmidt H, Hankeln T, Feldmeyer B, Pfenninger M. 2018. The genomic footprint of climate adaptation in *Chironomus riparius*. *Mol. Ecol.* **27**:1439-1456.

Wang S, Teng D, Li X, Yang P, Da W, Zhang Y, Zhang Y, Liu G, Zhang X, Wan W, *et al.* 2022. The evolution and diversification of oakleaf butterflies. *Cell* **185**:3138-3152.

Wang Y, Obbard DJ. 2023. Experimental estimates of germline mutation rate in eukaryotes: a phylogenetic meta-analysis. *bioRxiv*, https://doi.org/10.1101/2023.01.24.525323

Warren WC, Boggs TE, Borowsky R, Carlson BM, Ferrufino E, Gross JB, Hillier L, Hu Z, Keene AC, Kenzior A, *et al.* 2021. A chromosome-level genome of *Astyanax mexicanus* surface fish for comparing population-specific genetic differences contributing to trait evolution. *Nat. Commun.* **12**:1447.

Weller HI, Westneat MW. 2019. Quantitative color profiling of digital images with earth mover's distance using the R package colordistance. *PeerJ* **7**:e6398.

Wilfert L, Gadau J, Schmid-Hempel P. 2007. Variation in genomic recombination rates among animal taxa and the case of social insects. *Heredity* **98**:189-197.

Wu C, Lu J. 2019. Diversification of transposable elements in arthropods and its impact on genome evolution. *Genes* **10**:338.

Xing S, Bonebrake TC, Tang CC, Pickett EJ, Cheng W, Greenspan SE, Williams SE, Scheffers BR. 2016. Cool habitats support darker and bigger butterflies in Australian tropical forests. *Ecol. Evol.* **6**:8062–8074.

You M, Ke F, You S, Wu Z, Liu Q, He W, Baxter SW, Yuchi Z, Vasseur L, Gurr GM, *et al.* 2020. Variation among 532 genomes unveils the origin and evolutionary history of a global insect herbivore. *Nat. Commun.* **11**:2321.

Zimin AV, Salzberg SL. 2020. The genome polishing tool POLCA makes fast and accurate corrections in genome assemblies. *PLoS Comput. Biol.* **16**:e1007981.

FRENCH SUMMARY

Contexte

Les arbres phylogénétiques, aussi appelés « phylogénies », sont des représentations graphiques de l'histoire évolutive et des relations entre ou au sein de groupes d'organismes (Kjer et al. 2016). Ils sont utilisés depuis plus de deux siècles pour illustrer et comprendre l'immense diversité des êtres vivants. Avec l'avènement de la théorie de l'évolution de Darwin et Wallace, et plus tard, les travaux d'Hennig (1950, 1966), les phylogénies sont devenues des modèles pour étudier un large éventail de questions biologiques, et elles sont utilisées dans une variété de domaines aussi divers que la médecine, l'écologie, ou encore la microévolution (étude au niveau des espèces et des populations, sur des courtes périodes de temps) et la macroévolution (étude de l'évolution espèces, genres, ou grands embranchements du vivant sur de longues périodes de temps).

L'étude de groupes entiers d'organismes pour comprendre les schémas globaux de diversité (c'est-à-dire traiter des questions macroévolutives) implique la reconstitution des phylogénies au niveau des espèces. Cependant, définir ce que sont les espèces, en particulier les différencier clairement des sous-espèces ou des populations de la même espèce, représente une question épineuse (parfois appelée "problème des espèces") et les biologistes ne sont toujours pas d'accord sur un concept unique et exact de l'espèce. En fait, ce problème du concept d'espèce - c'est-à-dire les désaccords actuels sur le concept théorique d'espèce - est étroitement lié à la question de la délimitation des espèces - c'est-à-dire la manière de déterminer les limites et le nombre d'espèces sur la base de données empiriques.

Les espèces sont difficiles à définir en raison de ce qu'on appelle un "continuum de spéciation", c'est-à-dire un gradient d'isolement reproductif plus ou moins prononcé et continu qui se produit avant que les espèces ne soient complètement isolées (elles échangent de moins en moins de matériel génétique à mesure qu'elles se séparent). Comme les séparations d'espèce prennent parfois un grand nombre de génération avant d'être finalisées, cela signifie que la délimitation des espèces à notre échelle humaine nécessite de se positionner sur le continuum de spéciation, et cette position est susceptible de dépendre de la question scientifique et du groupe étudié. Étant donné la complexité et la continuité du concept d'espèce, la description ou la révision d'une espèce doit être faite avec soin et avec beaucoup de preuves. Pour cette thèse, j'ai essayé de suivre deux principes pour tenter de limiter les difficultés inhérentes au concept d'espèce : 1) intégrer un maximum de données accessibles pour définir les espèces (c'est le concept de « taxonomie intégrative »), et 2) j'ai essayé de justifier au maximum chacun de mes

choix et de rendre mes données le plus accessible possible à d'autres afin que les études futures puissent facilement réviser mes choix taxonomiques si nécessaire.

Bien définir les briques élémentaires d'un groupe (c'est-à-dire bien définir les espèces, pour les questions macroévolutives), permet ensuite de réaliser des analyses phylogénétiques. Celles-ci cherchent à répondre à la question suivante : *étant donné un groupe contenant plus de trois espèces, quels sont les espèces les plus étroitement apparentées ?* La possibilité de répondre à cette question repose sur la notion d'héritabilité des caractères (c'est-à-dire la transmission de caractéristiques des parents à leurs descendants, Mendel 1866), sur l'homologie des caractères considérés au sein d'un groupe d'organismes (c'est-à-dire que ces caractéristiques sont bien comparables entre organismes), mais aussi sur la variabilité de ces caractères au sein du groupe étudié. Si les caractères sélectionnés sont informatifs (variabilité suffisante) et hérités (non issus de la convergence), il est possible de déterminer la parenté entre les taxons en identifiant le partage d'un ou plusieurs caractères dérivés uniques que d'autres taxons hors du groupe ne possèdent pas.

Depuis deux siècles, les études phylogénétiques se sont d'abords basées sur les caractères morphologiques des espèces. Toutefois, la découverte de l'ADN comme étant un support des caractéristiques et un témoin de l'évolution des organismes a largement révolutionné les analyses phylogénétiques. Dernièrement, les progrès de séquençage ont été tels que le nombre de caractères pouvant être utilisés dans l'analyse phylogénétique de quelques milliers à des dizaines de milliers, permettant de produire des arbres phylogénétiques de mieux en mieux résolus et robustes. Cependant, cette avancée s'accompagne de défis informatiques et statistiques. Augmenter toujours plus le nombre de données peut certes réduire le nombre d'erreurs stochastiques (dû au hasard de jeux de données particuliers) mais aussi augmenter le nombre d'erreurs systématiques (des erreurs dans l'analyse même de ces données), et ainsi donner une grande confiance dans des phylogénies erronées. Par exemple, les mesures traditionnelles de soutien aux nœuds d'une phylogénie ne sont pas conçues pour capturer l'hétérogénéité des histoires évolutive que chaque gène d'une espèce a pu subir. Les ensembles de données génomiques présentent une saturation de ces mesures, ce qui conduit à des hypothèses phylogénétiques contradictoires bénéficiant d'un soutien statistique important. Les causes de ce conflit, qu'elles soient biologiques ou systématiques, doivent être étudiées plus avant. Des facteurs biologiques tels que l'introgression génétique et le tri incomplet des lignées peuvent également créer des arbres génétiques incongrus. Lors de ma thèse, je me suis appliqué à intégrer des approches méthodologiques tenant compte de l'histoire évolutive distincte de

chaque marqueur pour tenter de rendre compte de l'évolution indépendante de chaque marqueur et ainsi continuer d'améliorer les inférences phylogénétiques.

Les arbres phylogénétiques représentent les relations de parenté entre espèces. Bien que cela représente en soi une énorme quantité d'informations, cela ne fournit pas directement d'informations sur le temps absolu de cette histoire évolutive. Or, certaines questions doivent être traduites en temps absolu (géologique) pour être explorées. C'est le cas de la biogéographie historique, par exemple, qui compare l'histoire évolutive d'un groupe avec les processus globaux de la Terre. Pour étudier ces questions, le concept de "datation moléculaire" a été pensé les années 1960, sous l'impulsion de plusieurs auteurs tels que Zuckerkandl et Pauling. Ils proposaient que le temps de divergence entre les espèces puisse être quantifié par les différences de séquences moléculaires, en supposant une vitesse d'évolution constante, appelé « horloge moléculaire », et que cette horloge pourrait être calibrée à l'aide de données extérieurs tels que les archives fossiles. Cette première idée relativement simpliste a été raffinée pour prendre en compte l'hétérogénéité des jeux de données, et aujourd'hui, la datation moléculaire est largement utilisée par les biologistes de l'évolution dans différents domaines. Toutefois, la datation de phylogénie reste donc une tâche cruciale mais délicate, du fais de la pauvreté de certains registre fossile (tel que celui de certains insectes), et de la puissance informatique nécessaire à l'analyse des données génomiques actuelles.

En étudiant le rythme de la diversification des êtres vivants, la macroévolution tente de comprendre trois choses principales : quels sont les facteurs qui font varier le nombre d'espèces dans le temps, dans l'espace et entre les groupes d'espèces ? Au fil des siècles, les scientifiques ont étudié l'hétérogénéité temporelle et spatiale de la biodiversité, en s'appuyant sur des données paléontologiques (les dinosaures par exemple). Alors que les archives fossiles fournissent des preuves directes de la richesse passée, l'hétérogénéité spatiale peut être observée dans les distributions contemporaines des espèces. Les données de distribution des espèces, associées à leurs relations phylogénétiques et aux temps de divergence, permettent de déduire l'histoire « biogéographique » d'un groupe d'espèces. Ces analyses visent à découvrir les origines des espèces ou des clades et les facteurs qui influencent les distributions actuelles de la biodiversité, et ont pu être menée à l'échelle mondiale (par exemple en essayant de comprendre pourquoi les régions tropicales sont plus riche en espèces que les régions tempérés), ou à des échelles plus fines, comme dans la région tropicale incluant l'Amazonie et ses alentours seulement (appelé « Néotropiques »). Dans cette région, des années de recherches ont permis de montrer que des évènements géologiques majeurs, tels que l'orogenèse des Andes ou l'évolution du climat régional ont eu un impact important sur la diversification des groupes

d’espèces néotropicaux. Ainsi, les changements environnementaux et les événements géologiques, qui sont des facteurs extrinsèques, sont des candidats pour expliquer l’hétérogénéité de la biodiversité actuelle, et on regroupe le rôle qu’ont eu ces facteurs extrinsèques sous l’hypothèse dite du « bouffon de Roi ». Cela s’oppose à l'hypothèse de la « Reine rouge », qui s’intéresse à l'impact des facteurs biotiques (intrinsèques), tels que des caractéristiques écologiques propre à un groupe particulier d’organisme. Bien que le débat ait souvent été polarisé autour de l’une ou l’autre de ces hypothèses, de plus en plus de preuves s’accumulent pour dire que les deux types de facteurs ont influencé la diversification des organismes, et que leur importance relative dépend du contexte spécifique. Cette interaction simultanée des facteurs complique évidemment les analyses de corrélation entre eux, faisant de l'étude de l'hétérogénéité de la biodiversité une question complexe.

En tant que groupe eucaryote le plus diversifié sur Terre, les insectes sont particulièrement intéressant pour étudier ces questions macroévolutives. Etant un groupe d’espèce ancien, les insectes ont été influencés par des changements paléoclimatiques majeurs, qui ont eu un impact sur leur diversification et leur distribution spatiale actuelle. D’un autre côté, des caractéristiques intrinsèques à ces insectes tel que des traits morphologiques, physiologiques ou comportementales comme l’acquisition du vol, la métamorphose, la socialité ou le parasitisme, ont pu jouer un rôle important dans leur succès évolutif. Ce sont ce qu’on appelle des « innovations clés », c’est-à-dire ces nouveaux traits phénotypiques qui ont pu conduire au succès d'un groupe taxonomique, en améliorant sa capacité de survie, de reproduction ou en garantissant un accès à de nouvelles niches écologique. Depuis des décennies, on cherche à savoir si l’émergence de la phytophagie (se nourrir de plantes) peut être considérée comme une innovation clé potentielle chez les insectes. En effet, les clades (groupe d’espèces apparentés) phytophages présentent souvent une grande diversité, parfois supérieur aux groupes non phytophages. Un modèle introduit par Ehrlich et Raven en 1964, appelé « escape-and-radiate », suggère qu'une course aux armements entre les insectes herbivores et les défenses des plantes hôtes pour éviter d’être mangé, entraîne une diversification des deux antagonistes, en impliquant diverses stratégies adaptatives de la part des deux acteurs (présence d’épines ou sécrétion de toxines pour les plantes, et détoxification et adaptations génétiques à ces toxines pour les insectes par exemple).

Ce modèle a été appliqué aux papillons, un groupe diversifié comptant de nombreuses espèces. La famille des Papilionidae, par exemple, présente une préférence exclusive de ses chenilles à certaines plante-hôte particulière, certaines d’entre elles étant connues pour être très toxiques. Cette spécialisation suggère que ces papillons ont des adaptations particulières à leur

plantes hôtes, et ont pu en tirer des avantages évolutifs. Effectivement, une étude de co-évolution (étude parallèle de l'histoire évolutive de deux groupes en interactions) récente sur les Papilionidae et leur plantes hôtes ont montré des traces de ces adaptations dans le génome de ces espèces, et des taux de diversification plus élevés à la suite de changements de plantes hôtes (Allio et al. 2021). Cette étude a également pris en compte l'impact potentiel des changements de préférences climatiques sur les taux de diversification, mais n'a pas permis de les mettre en evidence. Cependant, l'analyse a manqué d'un cadre phylogénétique robuste et bien échantilloné (peu d'espèces manquante) et d'autres études semblent nécessaires pour explorer l'impact d'autres facteurs tels que les biomes, l'altitude, la région et les changements morphologiques dans la diversification de ces insectes.

Ma thèse a visé à apporter un cadre phylogéomique solide pour explorer ces hypothèses de diversification, et notamment de résoudre certaines zones d'ombre de la phylogénie des Papilionidae. En particulier, la tribu des Leptocircini, qui représente un quart de la diversité du groupe, a été moins échantillonnée que le reste de l'arbre et a donné lieu à un clade mal résolu. Ma thèse a constitué les premières étapes de l'exploration de cette question, en produisant à la fois de nouvelles données sur le génome entier et une phylogénie robuste et presque exhaustive de deux tribus dont l'âge d'émergence et la richesse au sein des Papilionidae sont similaires : Leptocircini et Troidini. Alors que les Leptocircini ont connu plusieurs changements de plantes hôtes au cours de leur histoire évolutive (Figure 1), les Troidini se nourrissent exclusivement d'Aristolochiaceae, plante considérée comme la plante hôte ancestrale au sein des Papilionidae. Afin d'atteindre cet objectif, cette thèse a impliqué différents aspects tels que l'acquisition de données de terrain, le séquençage à lecture longue et courte, les assemblages de génomes de référence, les analyses génomiques et démographiques historiques à l'échelle des populations, la systématique, les analyses phylogénomiques et les analyses macroévolutives (biogéographie et diversification).

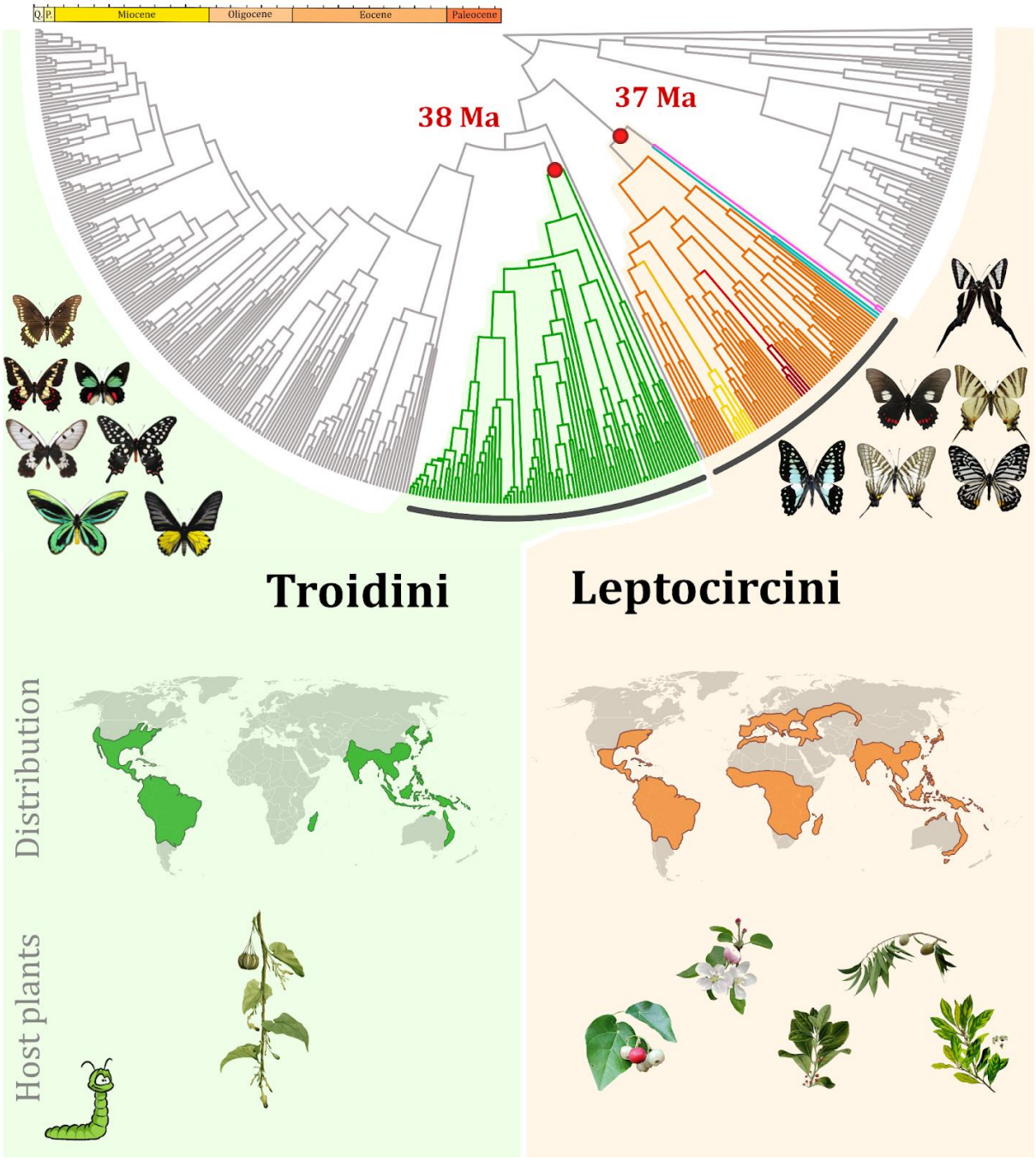


Figure 1 : Comparaison des Leptocircini et des Troidini. La phylogénie représentée est tirée de Allio et al. (2021), avec l'estimation de l'âge de la couronne des deux tribus. La tribu des Troidini est colorée en vert et la tribu des Leptocircini en orange. La répartition des Troidini et des Leptocircini est indiquée sur les cartes, ainsi que les principales plantes hôtes mangées. Pour les Leptocircini, les familles de plantes hôtes représentées sont les Annonaceae, Rosaceae, Hernandiaceae, Lauraceae et Aquifoliaceae. Illustration : *Graphium doson* (Alex Dumchus) ; *Lamproptera paracurius* (Adam M. Cotton) ; *Battus polydamas* (Cosmin O. Manci) ; *Parides childrenae* (Andrey Zheludev, CC BY-NC 2.5) ; *Cressida cressida* (CC BY-NC 4. 0, JKDw) ; *Euryades duponchelli, Pharmacophagus antenor, Trogonoptera brookiana, Troides aeacus, Ornithoptera alexandrae* (Fabien L. Condamine) ; *Atrophaneura sycorax, Pachliopta pandiyana, Byasa alcinous* (Eliette L. Reboud.)

La structure de la thèse s'articule autour des défis et des différentes étapes cruciales aux études macroévolutive, comme présentés dans cette introduction. Dans le **Chapitre 1**, j'ai étudié deux populations géographiquement isolées de l'espèce *Ornithoptera alexandrae* de Papouasie Nouvelle Guinée, pour déterminer si ces populations correspondaient à des espèces ou à des sous-espèces. Ce chapitre a aussi été l'occasion d'étudier la diversité génétique et la démographie récente de cette espèce en danger d'extinction, pour apporter des informations cruciales à l'élaboration de stratégies de conservation adaptées. Dans le **Chapitre 2**, j'ai développé un pipeline bio-informatique pour produire des génomes de référence et des données de génomes entiers inédites pour les deux tribus de Papilionidae étudiées. Ces données ont permis de reconstruire des phylogénies presque complètes de ces deux groupes, et ainsi de mieux comprendre leurs relations de parentés. En outre, ces nouvelles données génomiques ont permis d'étudier un cas intéressant d'introgression au sein de la tribu des Leptocircini. Enfin, en s'appuyant sur les données et les analyses du **Chapitre 2**, le **Chapitre 3** est une étude macroévolutive comparative des deux tribus de Papilionidae. Après une datation moléculaire des deux phylogénies reconstruites, ce chapitre a permis d'explorer l'histoire biogéographique et les dynamiques de diversification des deux tribus. L'hypothèse principale étudiée a été l'influence des changements de plantes hôtes sur la diversification des groupes, mais d'autres hypothèses telles que le mimétisme, le dimorphisme sexuel et les effets environnementaux ou géographiques ont également été examinés ou discutés.

Chapitre 1 – Etude sur la génomique, la démographie historique et la divergence des populations du plus grand papillon du monde, l'Ornithoptère de la Reine Alexandra.

Ornithoptera alexandrae, le plus grand papillon du monde, peut mesurer jusqu'à 28-30 cm d'envergure (Mitchell et al. 2016 et références). Cette espèce est endémique à la Papouasie-Nouvelle-Guinée, et des observations de terrain au cours des dernières décennies ont montré qu'il existe deux populations allopatriques reconnues de cette espèce, séparées par une chaîne de montagnes. *Ornithoptera alexandrae* est considérée comme une espèce menacée dans la liste rouge de l'UICN en raison de son aire d'occupation très petite et fragmentée. Les deux populations allopatriques sont morphologiquement similaires mais expriment des différences biologiques significatives, telles que de légères différences de taille et de temps de développement, de sorte que le degré de divergence de ces deux populations n'est pas clair.

Pourtant, la connaissance de l'histoire évolutive et génétique de l'espèce et des populations pourrait aider les programmes de conservation à protéger cette espèce. Nous avons donc réalisé la première étude génomique d'*O. alexandrae* afin de comprendre la démographie passée et présente de cette espèce et d'apporter des éclaircissements sur son statut d'espèce, ce qui pourrait avoir des implications pour les stratégies de conservation. Étant donné qu'un programme de conservation local a été mis en place pour élever ces papillons, le statut taxonomique des deux populations (c'est-à-dire deux populations d'une même espèce ou deux espèces différentes) peut informer la gestion de la conservation de cette espèce menacée.

Nous avons donc collecté des spécimens vivants des deux populations (un adulte et une chenille par population) et séquencé leur ADN en combinant des données génomiques dits de « long reads » et « short reads », complémentaires par leur longueur pour l'un, et précision de séquençage pour l'autre. Nous avons ainsi assemblé quatre génomes de haute qualité *d'O. alexandrae*. À des fins de comparaison, nous avons également séquencé et assemblé deux autres espèces de Troidini: *Ornithoptera priamus* et *Troides oblongomaculatus*, toutes deux classées dans la catégorie "préoccupation mineure" de la liste rouge de l'UICN, et dont les deux génomes produits ont très similaires en termes de qualité. Nous avons ensuite estimé les niveaux de diversité génomique des trois espèces, à la fois dans les données nucléaires et mitochondriales, et nous les avons comparés aux données de la littérature. Nous avons trouvé de très faibles niveaux d'hétérozygotie nucléaire et mitochondriale, comme le montre la **Figure 2**, où les trois espèces étudiées ont un faible niveau de diversité génétique par rapport à d'autres espèces de papillons.

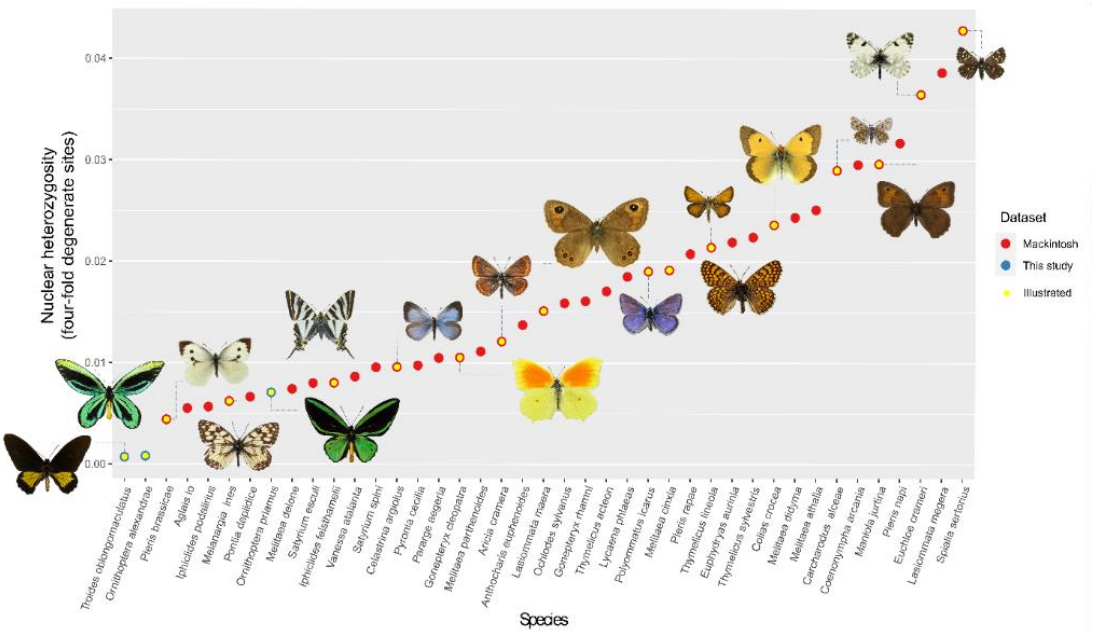


Figure 2 (page précédente) : Niveau d'hétérozygotie estimé sur des données nucléaires. Les points rouges correspondent aux valeurs de Mackintosh et al. (2019) et les points bleus à celles de la présente étude. Le point jaune indique les espèces illustrées. Photos (pas à l'échelle) : *T. oblongomaculatus* (CC BY 4.0, Peter Wing); *O. alexandrae* (Fabien L. Condamine); *P. brassicae* and *G. cleopatra* (CC BY-SA 3.0, Sarefo); *M. ines* (CC BY-SA 4.0, Atylotus); *O. priamus* (Eliette L. Reboud); *I. feisthamelii* and *M. cinxia* (CC BY-SA 4.0, Didier Descouens); *C. argiolus* (CC BY 3.0, Alan Cassidy); *A. cramera* and *P. icarus* (Robin Noel); *L. maera* and *C. crocea* (CC BY-SA 3.0, Vítězslav Maňák); *T. lineola* and *S. sertorius* (CC BY-NC-SA, Peter Huemer); *C. alceae* (CC BY-SA 3.0, Didier Descouens); *M. jurtina* (Public domain, Pekka Malinen); *E. crameri* (Alexander Slutsky).

Bien que les causes de la faible diversité génétique de ces papillons ne soient pas evidente aux premiers abords, nous nous sommes demandé comment cela se traduisait dans l'histoire démographique des espèces, compte tenu des menaces qui pèsent sur elles. Pour étudier la trajectoire de la taille effective ancestrale de la population (Ne) au fil du temps, nous avons utilisé des modèles de coalescence markovienne séquentielle (par exemple PSMC, MSMC) sur nos trois espèces. Mais avant d'appliquer ces méthodes à nos espèces, nous devions vérifier si nous ne violions pas certaines des hypothèses du modèle, car il a été démontré que cela pouvait avoir un impact considérable sur la précision des analyses. En utilisant des données simulées, nous avons montré que certains ajustements étaient nécessaires avant d'utiliser le modèle et nous avons souligné l'importance de ces ajustements, pourtant aujourd'hui largement ignorés dans de nombreuses études empiriques récentes. Nous avons ensuite appliqué ce modèle ajusté à nos données empiriques et constaté qu'*O. alexandrae* avait une faible population effective, en déclin constant au cours du dernier million d'années. Nous avons également constaté qu'elle a divergé en deux populations distinctes il y a environ 10 000 ans, probablement après la dernière glaciation. Enfin, nous avons exploré la structure génétique des deux populations et, dans l'ensemble, nos résultats indiquent l'absence de preuves d'une forte différenciation génomique et d'une barrière génomique au flux génétique entre les deux populations allopatriques *d'O. alexandrae*, ne semblant ainsi pas justifier la nécessité de description de deux espèces. En conclusion, nous demandons aux décideurs politiques de protéger l'environnement de la forêt tropicale d'*O. alexandrae*, et de soutenir ses deux populations par l'implémentation de zones protégées et le soutien au programme d'élevage, qui reproduit et relâchement des individus dans la nature.

Chapitre 2.1 - Décryptage de la systématique et de l'histoire évolutive de la tribu Leptocircini (Lepidoptera : Papilionidae) à l'aide de données génomiques complètes.

Les Leptocircini sont une tribu au sein des Papilionidae, qui comprennent environ 160 espèces réparties dans le monde entier, et représente environ un quart de tous les Papilionidae. Les relations phylogénétiques au sein de la tribu sont assez incertaines et n'avait jamais fait l'objet d'études phylogénétiques complètes et approfondies, contrairement à d'autres tribus de Papilionidae. Les Leptocircini ont tout de même fait l'objet de plusieurs études morphologiques et ont été également reprsentés dans des phylogénies moléculaires globales des Papilionidae. Cependant, leur fraction d'échantillonnage a toujours été assez pauvre.

La majeure partie de la richesse en espèces de Leptocircini est contenue dans le genre *Graphium*, qui comprend à lui seul ~110 des 160 espèces de Leptocircini. Le genre monotypique australien *Protographium* est considéré comme frère de *Graphium*. Le genre *Eurytides* est le deuxième genre le plus riche chez les Leptocircini, avec >40 espèces, et se trouve exclusivement en Amérique. A l'extérieur de ces groupes diversifiés, on trouve deux genres moins diversifiés comprenant chacun trois espèces : (1) le genre Iphiclides, qui est le seul groupe avec une distribution eurasienne, et (2) le genre Lamproptera, qui présente une morphologie divergente et particulière.

Dans cette étude, nous avons échantillonné 95 % de la diversité totale en utilisant une combinaison de papillons fraichement tué et de spécimens de musée. Nous avons assemblé 147 nouveau génomes complets de Leptocircini. Par ailleurs, nous avons assemblé *de novo* cinq génomes de référence (c'est-à-dire ayant une qualité très importante) pour cinq espèces de Leptocircini, représentant le genre *Iphiclides*, et trois des cinq sous-genres de *Graphium*. Ces génomes de référence nous ont permis d'améliorer la contiguité des 147 génomes complets et ainsi, de récupérer environ 20 % de gènes en plus que les assemblages provisoires. Dans ces génomes complets, nous avons répertorié et réunis dans un large jeu de donnée plus de 1400 gènes. Ces genes ont tous été analysés indémendamment (pour réaliser des « arbres de gènes », mais ils ont également été réunis dans une large matrice de caractère pour des analyses phylogenetiques complètes. Toutes les analyses phylogénétiques à l'échelle du génome ont fourni des arbres phylogénétiques très similaires, Ces analyses ont apporté un soutien solide à la monophylie de tous les genres. Au sein du genre *Eurytides*, les sous-genres *Mimoides*, *Eurytides* et *Protesilaus* ont été considérés comme monophylétiques avec un fort soutien, mais le sous-genre *Neographium* s'est révélé paraphylétique dans toutes les analyses. Au sein du

genre *Graphium*, les sous-genres *Pazala*, *Graphium* et *Arisbe* ont été retrouvé comme monophylétiques avec des soutiens solides, mais le sous-genre *Pathysa* s'est révélé paraphylétique dans toutes les analyses, et le sous-genre *Paranticopsis* s'est révélé polyphylétique dans toutes les analyses. De manière générale, plusieurs clades présentaient des hétérogeneité forte entre les différents arbres de gènes indépendants, suggérant une hétérogénéité des histoires évolutives au sein de leur génomes. Parmi ces clades, nous avons étudié le cas des clades paraphylétiques *Pathysa* et *Paranticopsis*. Ce clade présente des caractéristiques de spécialisation de plante hôte et des patrons alaires qui ne sont pas conformes à la topologie majoritaire. Nous avons donc testé si cela était le résultat d'une convergence, d'un flux de gènes massif entre espèces après leur séparation, ou d'un tri de lignées incomplet (polymorphisme de la population héritée par hasard de manière discordante à la topologie des espèces). Nous avons pu écarter l'hypothèse selon laquelle la topologie observée était le résultat d'une convergence acquise deux fois dans la phylogénie, ainsi que l'hypothèse selon laquelle la distribution observée des caractères est le résultat d'un tri de lignées incomplet. Le cas de flux de gènes entre lignée était le plus probable, mais aussi le plus complexe à déchiffrer. Finalement, nous n'avons pas pu conclure à la direction exacte du flux de gènes, et il est probable que tous les clades ont continué à échanger certains gènes après leur séparation, dans plusieurs directions. Il faudrait sans doute ajouter des informations d'ontologie des gènes pour préciser plus finement le scénario.

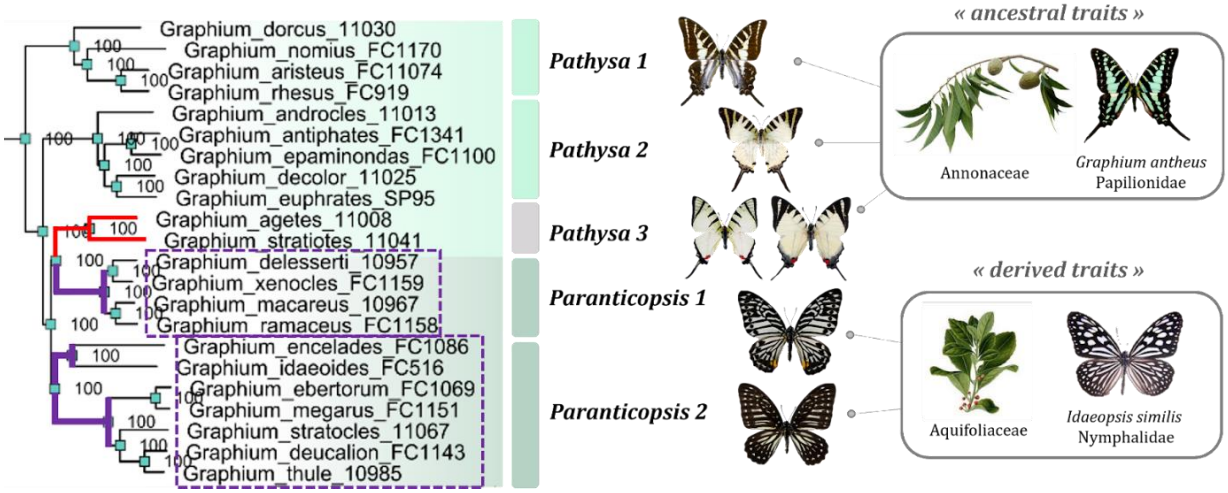


Figure 3 : Relations phylogénétiques du clade *Pathysa/Paranticopsis*. Les différents groupes d'espèces sont délimités et nommés. Chaque groupe d'espèces est lié à ses traits phénotypiques, correspondant soit aux "traits ancestraux" (alimentation sur les Annonaceae, motif rayé des ailes et présence de queues), soit aux "traits dérivés" (alimentation sur les Aquifoliaceae, et motif mimétique des ailes). Les soutiens des branches sont indiqués aux nœuds.

Chapitre 2.2 - Phylogénie à l'échelle de l'espèce des Papillons géants de la tribu Troidini (Lepidoptera : Papilionidae)

Les Troidini constitue une remarquable tribu de Papilionidae répartie dans les tropiques de l'Asie du Sud-Est et dans la région néotropicale. Malgré quelques groupes bien étudiés, il manque encore une phylogénie robuste et bien échantillonnée pour tous les genres. Nous avons donc réalisé une étude phylogénétique sur des génomes complète avec les mêmes techniques que dans le Chapitre 2.1. En utilisant une combinaison une combinaison de données de séquençage de courts et longs fragments, nous avons assemblé cinq nouveaux génomes de référence et généré des génomes entiers hautement couverts et échafaudés pour 131 espèces. Sur la base de gènes mitochondriaux ou de milliers de gènes nucléaires et d'approches phylogénétiques multiples, nous avons obtenu des arbres phylogénomiques robustes représentant >97% des espèces de Troidini. Nos analyses phylogénétiques ont été assez cohérentes avec la littérature déjà existante pour divers clades. En revanche, de manière inattendue, nous avons constaté que Byasa daemonius était sœur de toutes les Atrophaneura et Byasa, tandis que B. crassipes était sœur d'Atrophaneura. Ces résultats soulèvent des questions quant à la validité de ces genres. Par conséquent, nous avons révisé le statut du genre Byasa, le considérant désormais comme un sous-genre d'Atrophaneura, et nous avons proposé que B. daemonius fasse partie d'un nouveau sous-genre d'Atrophaneura. En outre, nous avons reclassé Losaria en tant que sous-genre de Pachliopta. Nous avons proposé la révision de plusieurs espèces, qui semblait présentée une inflation taxonomique, comme dans le genre de Troides par exemple.

Chapitre 3 – Etude macroévolutive comparée de deux tribus de Papilionidae : les changements de plantes hôtes ont-ils joué un rôle dominant dans leur diversification ?

Parmi les papillons de la famille des Papilionidae, les tribus Troidini et Leptocircini sont frappantes par leurs similitudes (âge, origine, distribution, richesse spécifique), mais aussi par leurs différences (taille du répertoire de plante hôtes, hétérogénéité des morphologies, aposématisme). Dans cette étude, nous avons tiré parti de ces caractéristiques pour étudier l'impact des interactions plante-insecte sur l'histoire évolutive des Papilionidae, ainsi que l'impact des événements biogéographiques en tant qu'effets confondants des changements de plantes hôtes. Nous avons daté les phylogénies des **Chapitres 2.1** et **2.2** à l'aide d'une approche bayésienne à horloge relâchées, calibrée avec des calibrages secondaires. Nous avons réalisé

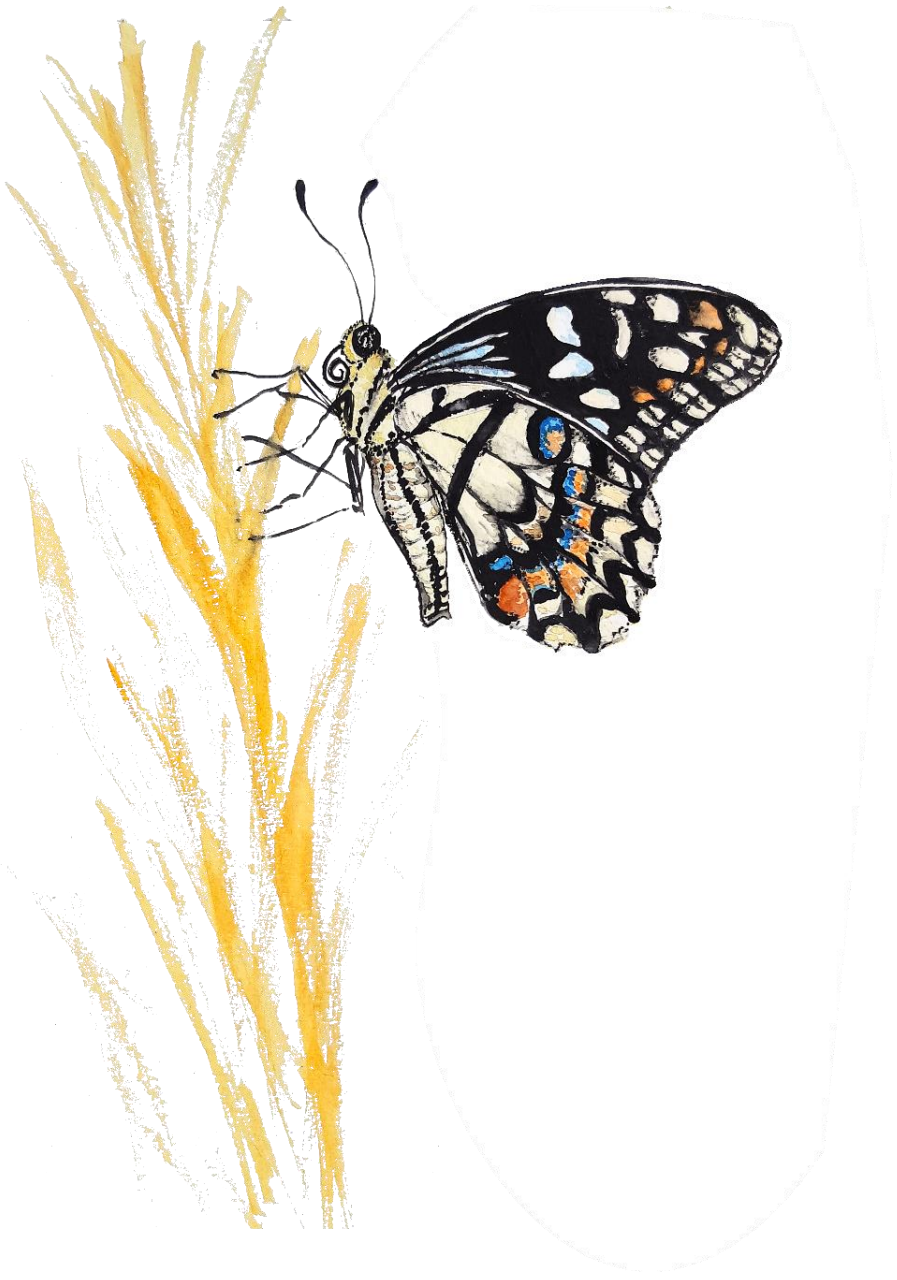
plusieurs estimations indépendantes, en prenant des jeux de gènes différents à chaque analyse, pour tenir compte de l'hétérogénéité des histoires de gènes tout en respectant les contraintes méthodologiques et les temps de calculs. Nous avons ensuite appliqué des modèles de reconstruction des traits ancestraux, ainsi que des modèles de diversification spécifiques, qui cherchent des changements de taux de diversification le long des branches, ou pour des clades entiers. Nous avons ainsi testé l'impact des changements de la plante hôte sur la diversification des papillons, tout en étudiant les modes de changement de la plante hôte (radiation évolutive, « oscillation », « chaises musicales »).

D'après nos analyses, les deux tribus sont apparues sur le continent Laurasiatique au cours de l'Eocène et se sont principalement diversifiées vers le sud dans les Néotropiques ou l'Archipel Indo-Australasien. La colonisation de l'Afrique par les *Graphium* a été déduite comme ayant eu lieu au milieu de l'Oligocène. Pour la première fois, nous envisageons que la colonisation énigmatique de Madagascar par *Pharmacophagus antenor* ait pu se dérouler en empruntant une route Africaine continentale. Dans l'ensemble, nous avons trouvé des taux de diversification plus élevés chez les Troidini que chez les Leptocircini, et des taux d'extinction extrêmement faibles pour les deux tribus. Pour les Leptocircini, les analyses suggèrent un changement de diversification dans les clades *Arisbe*, *Protesilaus* et *Pathysa*, mais le changement de plante hôte n'est pas apparu comme un moteur clair de ces patrons de diversification. Chez les Troidini, les changements de diversification ont été observés principalement dans le clade des *Ornithoptera* et le clade des *Troides*, mais aussi au sein des clades *Pachliopta* et *Atrophaneura*, ce qui suggère un effet de spéciation insulaire.

Perspectives

L'un des points particulièrement importants que nous avons pu soulever au travers de cette étude synthétique des deux tribus est qu'en appliquant une taxonomie plus ou moins stricte (pour prendre en compte les incertitudes taxonomiques des deux groupes), les résultats de la diversification s'en trouvaient particulièrement affectés. Cette dépendance des analyses de diversification à une taxonomie robuste et digne de confiance montre à quel point il est nécessaire de prendre en compte correctement ce paramètre, notamment pour des groupes encore peu étudiés, comme l'était les Leptocircini. Des études à l'échelle populationnelle dans les zones encore douteuses de l'arbre des Troidini et des Leptocircini permettrait d'affiner les résultats du **Chapitre 3**.

APPENDICES



Papilio demoleus - Eliette Reboud

Table of contents

APPENDICES 298

APPENDIX I. PULLED DIVERSIFICATION RATES, LINEAGES-THROUGH-TIME PLOTS, AND MODERN MACROEVOLUTIONARY MODELING 300
APPENDIX II A COMPREHENSIVE PHYLOGENY AND REVISED TAXONOMY ILLUMINATE THE ORIGIN AND DIVERSIFICATION OF THE GLOBAL RADIATION OF PAPILIO (LEPIDOPTERA: PAPILIONIDAE) 317
APPENDIX III. MUSEOMICS REVEALS PATTERNS OF WITHIN-SPECIES DIVERGENCE AND DIVERSIFICATION IN TROIDES BIRDWINGS (LEPIDOPTERA: PAPILIONIDAE) 341
APPENDIX IV. GENOMICS OF THE RELICT SPECIES BARONIA BREVICORNIS SHEDS LIGHT ON ITS DEMOGRAPHIC HISTORY AND GENOME SIZE EVOLUTION ACROSS SWALLOWTAIL BUTTERFLIES 388
APPENDIX V. HOST-PLANT SHIFTS IN SWALLOWTAIL BUTTERFLIES DID NOT HAPPEN IN ISOLATION. 427

Appendix I

Syst. Biol. 71(3):758–773, 2022
© The Author(s) 2021. Published by Oxford University Press on behalf of the Society of Systematic Biologists. This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contactjournals.permissions@oup.com
DOI:10.1093/sysbio/syab083
Advance Access publication October 6, 2021

Pulled Diversification Rates, Lineages-Through-Time Plots, and Modern Macroevolutionary Modeling

ANDREW J. HELMSTETTER[1,*], SYLVAIN GLEMIN[2], JOS KÄFER[3], ROSANA ZENIL-FERGUSON[4], HERVÉ SAUQUET[5,6], HUGO DE BOER[7], LÉO-PAUL M. J. DAGALLIER[8], NATHAN MAZET[9], ELIETTE L. REBOUD[9], THOMAS L. P. COUVREUR[8], AND FABIEN L. CONDAMINE[9]

[1]*Fondation pour la Recherche sur la Biodiversité - Centre for the Synthesis and Analysis of Biodiversity, Institut Bouisson Bertrand, 5 Rue de l'École de Médecine, 34000 Montpellier, France;* [2]*CNRS, UMR 6553 Ecosystèmes Biodiversité Evolution (Université de Rennes), 35042 Rennes, France;* [3]*Université de Lyon, Université Lyon 1, CNRS, Laboratoire de Biométrie et Biologie Evolutive UMR 5558, F-69622 Villeurbanne, France;* [4]*School of Life Sciences, University of Hawaii Manoa, Honolulu, HI 96822, USA;* [5]*National Herbarium of New South Wales, Royal Botanic Gardens and Domain Trust, Sydney, NSW 2000, Australia;* [6]*Evolution and Ecology Research Centre, School of Biological, Earth and Environmental Sciences, University of New South Wales, Sydney, NSW, Australia;* [7] *Natural History Museum, University of Oslo, 0318 Oslo, Norway;* [8]*IRD, DIADE, Univ Montpellier, Montpellier, France and;* [9]*CNRS, Institut des Sciences de l'Evolution de Montpellier (Université de Montpellier), Place Eugène Bataillon, 34095 Montpellier, France*
**Correspondence to be sent to: FRB-CESAB, Institut Bouisson Bertrand, 5 Rue de l'École de Médecine, Montpellier, 34000, France*
E-mail: andrew.j.helmstetter@gmail.com

Received 15 February 2021; reviews returned 29 September 2021; accepted 30 September 2021
Associate Editor: Mark Holder

Abstract.—Estimating time-dependent rates of speciation and extinction from dated phylogenetic trees of extant species (timetrees), and determining how and why they vary, is key to understanding how ecological and evolutionary processes shape biodiversity. Due to an increasing availability of phylogenetic trees, a growing number of process-based methods relying on the birth–death model have been developed in the last decade to address a variety of questions in macroevolution. However, this methodological progress has regularly been criticized such that one may wonder how reliable the estimations of speciation and extinction rates are. In particular, using lineages-through-time (LTT) plots, a recent study has shown that there are an infinite number of equally likely diversification scenarios that can generate any timetree. This has led to questioning whether or not diversification rates should be estimated at all. Here, we summarize, clarify, and highlight technical considerations on recent findings regarding the capacity of models to disentangle diversification histories. Using simulations, we illustrate the characteristics of newly proposed "pulled rates" and their utility. We recognize that the recent findings are a step forward in understanding the behavior of macroevolutionary modeling, but they in no way suggest we should abandon diversification modeling altogether. On the contrary, the study of macroevolution using phylogenetic trees has never been more exciting and promising than today. We still face important limitations in regard to data availability and methods, but by acknowledging them we can better target our joint efforts as a scientific community. [Birth–death models; extinction; phylogenetics; speciation.]

A major goal in evolutionary biology is to understand the large-scale processes that have shaped biodiversity patterns. One important way to investigate this is by modeling species diversification using speciation and extinction, which can vary over time and among groups. It is commonplace to find regions, or clades, in phylogenetic trees that accumulate lineages faster than others. Macroevolutionary models often aim to explain this variation in diversification patterns by associating bursts of speciation or extinction with factors such as time (May et al. 2016), lineages (Rabosky et al. 2013; Rabosky 2014), character traits (Maddison et al. 2007), or the environment (Condamine et al. 2013).

The growing number of large phylogenetic trees that capture a substantial proportion of living species provide increasing power and resolution for such studies (Jetz et al. 2012; Smith and Brown 2018; Stein et al. 2018; Upham et al. 2019). Furthermore, the availability of a wide variety of methods and software (e.g., BAMM (Rabosky 2014), state-speciation and extinction (SSE) models (Maddison et al. 2007; FitzJohn 2012), RPANDA (Morlon et al. 2016), MEDUSA (Alfaro et al. 2009)) has made diversification studies increasingly popular in the last decade. Approaches that can link diversification to a particular process or trait are among the most appealing to researchers because they enable us to test long-standing hypotheses in evolutionary biology and ecology. Examples include those related to the evolution of key innovations (Silvestro et al. 2014), the colonization of new areas (McGuire et al. 2014), the effect of elevation (Lagomarsino et al. 2016; Quintero and Jetz 2018), and the latitudinal diversity gradient (Rolland et al. 2014; Pulido-Santacruz and Weir 2016; Rabosky et al. 2018; Igea and Tanentzap 2020).

A typical workflow for diversification rate modeling using molecular phylogenetic trees starts with obtaining DNA sequence data for species in a study group, which are then used to estimate species relationships in the form of a phylogenetic tree. Usually, this phylogenetic tree contains only extant species, and it is time-calibrated using ages derived from different sources including fossils (Ho and Phillips 2009; Parham et al. 2012; Sauquet 2013). The output of this process is referred to as an extant timetree. Once a tree has been generated, a birth–death model is fitted to explain patterns of diversification in the tree. Note that fossils are generally

Downloaded from https://academic.oup.com/sysbio/article/71/3/758/6382322 by Ecole Francaise D'Athenes user on 23 April 2022

used for node calibration and tree shape estimation but are rarely incorporated in any subsequent estimation of diversification rates, although it is now possible to incorporate fossils as tips in phylogenetic trees (Ronquist et al. 2012; Heath et al. 2014) and in birth–death models for diversification rate estimation (Mitchell et al. 2019). Some of the most recent developments have now also allowed researchers to coestimate the phylogenetic tree and diversification model (Höhna et al. 2016; Barido-Sottani et al. 2020).

The simplest birth–death models assume that each branch of a phylogenetic tree shares the same rate of "birth" (speciation) events, as well as "death" (extinction) events (Nee et al. 1994; Nee 2006; Ricklefs 2007; Morlon et al. 2011). There are two principal parameters in the birth–death model, the speciation rate (λ)—the rate at which lineages arise, and the extinction rate (μ)—the rate at which lineages disappear. In addition, a third parameter known as sampling fraction (or ρ) is often included to account for the fact that not all extant taxa are included in the phylogenetic tree. Sampling fraction can be defined as the ratio of sampled species over the total extant species diversity for a given clade. Under the simplest framework, λ and μ are constant over time (time-independent) and the same across all clades (clade-independent). However, we can begin to relax these assumptions by allowing models to be time-dependent (Morlon et al. 2011; Louca and Pennell 2020), clade-dependent (Maddison et al. 2007; Maliet et al. 2019; Barido-Sottani et al. 2020), or both (Rabosky et al. 2013). By making use of these different parameters and assumptions, birth–death models allow us to investigate whether the net diversification rate, defined as $r = \lambda - \mu$, has varied over time or among clades and ultimately uncover the processes that have given rise to extant biodiversity across the tree of life.

A recent study (Louca and Pennell 2020 abbreviated to LP) demonstrates, using an approach based on lineages-through-time (LTT) plots, that rates of speciation and extinction over time cannot be reliably estimated using extant timetrees (see also Kubo and Iwasa 1995). LP show how results of this approach can be misleading and provide potential solutions to the issues raised by proposing new summary statistics. This study has already provoked a response from the community (e.g., Morlon et al. 2020) and stimulated considerable discussion, with some going so far as to suggest that speciation and extinction rates cannot be estimated using phylogenetic trees (Pagel 2020). As a result, this study has called into question the meaning of diversification rate estimates generated from any analytical framework. Here, we try to render this discussion accessible to the broader audience of evolutionary biologists, who might use the available methods to study diversification but do not themselves develop macroevolutionary models. Thus, we aim to outline the major concepts discussed in LP in an accessible way, put the results and conclusions of LP into historical context, and explore how the implications of this study apply to macroevolutionary modeling today.

A Summary of the Main Concepts and Findings in Louca and Pennell (2020)

An LTT plot shows how extant lineages accumulated over time, using a phylogenetic tree (Nee et al. 1992) (Fig. 1). Each point in an LTT corresponds to a change in the number of lineages from the root of a phylogenetic

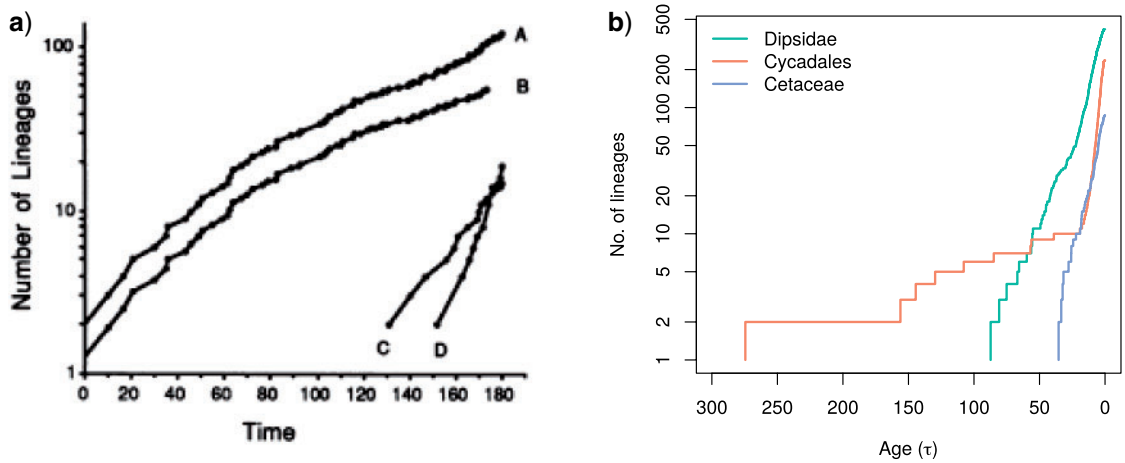


Figure 1. a) The first example of a lineages-through-time plot (LTT), taken from Nee et al. (1992) and based on a phylogenetic tree of birds. On the y-axis is the number of lineages (log scale) and the x-axis is time since origin (present on the right hand side of the graph). *"Each point corresponds to a change in the number of lineages. Line A, the pattern of origination of all 122 lineages; line B, same as A, but without the Passeri (line C) and the Ciconiiformes (line D). Line B has been shifted downward to aid visual comparison. The diversification rate is quantified by the steepness of the slope."* In this panel, time is displayed from past to present as time since origin (t). b) Three LTTs from modern phylogenetic trees of Dipsidae (campanulids) (Beaulieu and Donoghue 2013; Beaulieu and O'Meara 2016), Cycadales (Condamine et al. 2015) and Cetacea (Slater et al. 2010). In this panel, time is shown from present to past as an age (τ).

Downloaded from https://academic.oup.com/sysbio/article/71/3/758/6382322 by Ecole Francaise D'Athenes user on 23 April 2022

tree at $t=0$ to the present day at $t=T$ (Fig. 1a). Alternatively, as in LP, time can be considered as an age ($\tau=T-t$), where $\tau=0$ at the present and $\tau=T$ at the origin of the clade, or the root age (Fig. 1b). For consistency with LP, we will generally consider timescale as age (τ) in the equations we use throughout this article.

Simply put, when a clade diversifies faster, the slope of the LTT becomes steeper, but when diversification slows, the slope of the LTT levels off. When only extant lineages are considered, as in LP, LTT plots will never exhibit a drop in total lineage diversity over time (Ricklefs 2007). Regardless of whether time is τ or t in the equations, the LTT is usually plotted with the present on the right, thus its slope will never be negative. However, this does not mean that extinction does not have an effect on the shape of the LTT (Nee 2006). Previous work from almost 30 years ago (Nee et al. 1992, 1994; Harvey et al. 1994) clearly demonstrated how an LTT may change when extinction is present alongside speciation (birth–death), as opposed to speciation alone (pure-birth) in the context of models where diversification rates were constant over time. If the growth of an extant timetree is represented as an LTT on a semilog scale (i.e., lineage number is logarithmic, time is not, see Fig. 1), we would expect the trend to be linear under a pure-birth process (with constant speciation and no extinction). If extinction is introduced, the LTT would deviate from this linearity. When both rates are constant and greater than 0, the curve is expected to be linear over most of its history, but as time reaches the present the rate of lineage accumulation will increase (i.e., the LTT slope will become steeper) as the effect of extinction diminishes (Nee et al. 1994; see also Fig. 4 in Nee (2006)). By examining other deviations from linearity in the LTT plot we can begin to understand how diversification rates fluctuated over the history of a clade (Ricklefs 2007) and develop evolutionary hypotheses on why these fluctuations occurred.

The Deterministic Lineages-Through-Time Plot

The birth–death model is often implemented as a continuous-time Markov chain where at any given age (τ) we can calculate the probability of speciation (birth of a lineage) or extinction (death of a lineage) happening. Because the birth–death process is stochastic, each run (i.e., realization of the process) will result in a different history of diversification, even if the probabilities for speciation and extinction are the same.

For such models, we can approximate their expected value by conducting many realizations of the stochastic model. Alternatively, we can calculate the expected values using a set of continuous equations, which is known as a deterministic model. Although the stochastic and deterministic models both rely on the same birth–death process their behavior can be different, notably when using small phylogenetic trees (see Box 1). The deterministic model produces the expected value one would get by averaging over an infinite number of simulations, thus it is

BOX 1: HOW FAST CAN SPECIATION AND EXTINCTION RATES VARY?

Although it is standard practice, it should be noted that the approach of LP involves using and solving differential equations and therefore considers speciation and extinction to be continuous processes: at any infinitesimal time interval, the species number changes infinitesimally through speciation and extinction. In the birth–death process, however, the smallest amount of change in the number of species is one, and this happens only at particular moments in time.

Even if speciation, in reality, is a complex process that takes time (Etienne and Rosindell 2012), the branching events that represent speciation in phylogenetic trees are instantaneous. This means that the number of lineages will immediately jump from n to $n+1$, rather than gradually, in a continuous process (i.e., the durations of speciation and extinction are much shorter than the time between two events). An empirical LTT plot will thus show discrete events, appearing step-wise rather than being a continuous function, as is the dLTT. To measure pulled rates, LP propose calculating the slope and curvature of the LTT plot (though other ways of doing this now exist; Louca et al. 2021). For the dLTT, where the number of lineages is a continuous function of time, these are the first and second derivative of this function. For empirical LTT plots, one has to calculate the slope and curvature using some time interval. When working with a large phylogenetic tree and many species (as in the examples discussed by LP and Louca et al. 2018), the LTT is smooth and the slope and curvature, which are necessary for the estimation of the pulled rates, can be reliably estimated. However, many studies attempt to estimate diversification rates with relatively small numbers of species (i.e., <1000, e.g., Hutter et al. 2017 or even <100, e.g., Duan et al. 2018). Thus, as the number of species diminishes, one has to be aware that overparameterization might occur, and it would be wiser to stick to simple functions of diversification (or their pulled variants) through time, the simplest being time-independent rates, and even these are associated with large uncertainty (Aldous et al. 2011). Such problems are not unique to the estimation of pulled rates, and may be found in various macroevolutionary methods that rely on small samples of discrete variables. Yet, even in large trees, rates will always be estimated using a time interval that contains a sufficient number of speciation and extinction events. The consequences of this are that rapid changes in diversification rates might be missed due to the resolution of the chosen interval, and that capturing macroevolutionary dynamics at deep timescales will require very wide intervals, at the cost of reduced temporal resolution.

Downloaded from https://academic.oup.com/sysbio/article/71/3/758/6382322 by Ecole Francaise D'Athenes user on 23 April 2022

deterministic because it is fully defined by the parameters—no uncertainty from stochasticity is involved. This kind of approach is widely used and is also taken by LP who model the birth–death process as a set of differential equations, which is advantageous because these equations can be solved analytically.

LP refer to an LTT obtained from such models as a deterministic LTT (dLTT), which corresponds here to the expected LTT generated by trees with given speciation and extinction rates. Empirical LTTs generated using extant timetrees can be compared to model-generated dLTTs (where λ and μ are known) to disentangle how speciation and extinction have influenced patterns of diversity over time. To do this, the probability of the data given the model, or the likelihood, is calculated. Importantly, LP showed that, when λ and μ are clade-independent across the tree, the likelihood can be fully written as a function of the observed LTT and the dLTT generated by the model, even when rates vary over time (see also Lambert and Stadler 2013). Typically, by changing the parameters in the model, its dLTT resembles the empirical LTT to a greater or lesser extent, and the model is more or less likely. The best-fitting model can then be selected, representing our best hypothesis for how and to what extent speciation and extinction rates varied over time.

Model Congruence and Congruence Classes

Consider a simple model where λ and μ are constant over time and among clades and all lineages have been sampled ($\rho = 1$). In this case, the slope of the LTT plot is $r = \lambda - \mu$, except at times close to the present, where the effect of extinction diminishes and the slope becomes λ (Nee et al. 1994). Thus, with these slopes yielding λ and r we can calculate $\mu = \lambda - r$ (in practice, all parameters can be inferred at once within a likelihood framework using equations in Nee et al. (1994)). LP develop upon this classical knowledge to show that if rates vary over time (τ) it is no longer possible to estimate $\lambda(\tau)$, as the value of $\lambda(\tau)$ at present does not yield any information about its past dynamics. In other words, it is possible to choose almost any historical scenario for $\lambda(\tau)$ and obtain a complementary scenario of $\mu(\tau)$ that produces the same dLTT. If different models produce the same dLTT then they will also share the same likelihood for any given LTT.

LP call the set of models that generate the same dLTT a "congruence class." These congruence classes contain an infinite number of models with different parameter values that all produce the same dLTT. LP explain that when trying to select the best model we often start with a relatively small set of allowed models that we test. For example, a set of two models where speciation rate is fixed and extinction rate is allowed to vary over time, or *vice versa*. This would produce two equally likely models when trying to explain a slowdown in diversification, one indicating the case was an increase in extinction rate, the other a drop in the speciation rate—there is no way of distinguishing between them (Rabosky and Lovette 2008; Crisp and Cook 2009; Burin et al. 2019).

LP propose, and have recently developed upon (Louca et al. 2021; Louca and Pennell 2021), the idea that instead of selecting the model closest to the true process, we are selecting the model closest to the congruence class that includes the true process (see Fig. 3 in LP). In extreme cases, the best fitting model could thus be further from the true process than a more correct model, just because the former is included in the congruence class and the latter is not. However, LP also state that because we only assess a limited set of models, it is unlikely that we encounter models belonging to the same congruence class, but it is nevertheless possible. The consequence of multiple, equally likely models with different speciation and extinction rates is that these rates cannot be determined. This is a statistical phenomenon known as unidentifiability—the likelihood is the same for multiple parameter values making it impossible to choose one over another.

Unidentifiability

In macroevolutionary modeling, we might be interested to know how both λ and μ have changed over time (Rabosky 2006; Morlon et al. 2010). The unidentifiability issue outlined above means that we would not be able to ascertain the true parameter values of the models that generate our dLTTs. Another well-known example of this in evolutionary biology is the unidentifiability of substitution rates and divergence times when estimating phylogenetic trees (Rannala 2002). The molecular character data we use as input gives us the estimate of their product, which could be explained by different combinations of rates and times. In this case, we usually try to overcome the unidentifiability issue through the use of informative priors, such as those often placed on node ages in Bayesian approaches. This problem is not unique to macroevolutionary models, and, in fact, stems from a basic mathematical issue (Rannala 2002; Ponciano et al. 2012).

Consider a simple example in which we want to determine the parameter values for x and y:

$$2x + y = 7 \quad (1)$$

$$3x + 2y = 12. \quad (2)$$

For each value of x in equation 1, we can find a y that satisfies this equation—and there are an infinite number of equally likely possibilities. It is only when we add more information (in the form of equation 2) that we can determine the unique pair of values for x and y. Put simply, a solution can be found only if you possess at least the same number of equations as unknowns. If there are fewer equations, the unidentifiability is caused by overparameterization—there is an excess of parameters such that the model cannot estimate the values of any of them.

Though the LTT is generated through the use of many different observations and elements (DNA, fossils for

Downloaded from https://academic.oup.com/sysbio/article/71/3/758/6382322 by Ecole Francaise D'Athenes user on 23 April 2022

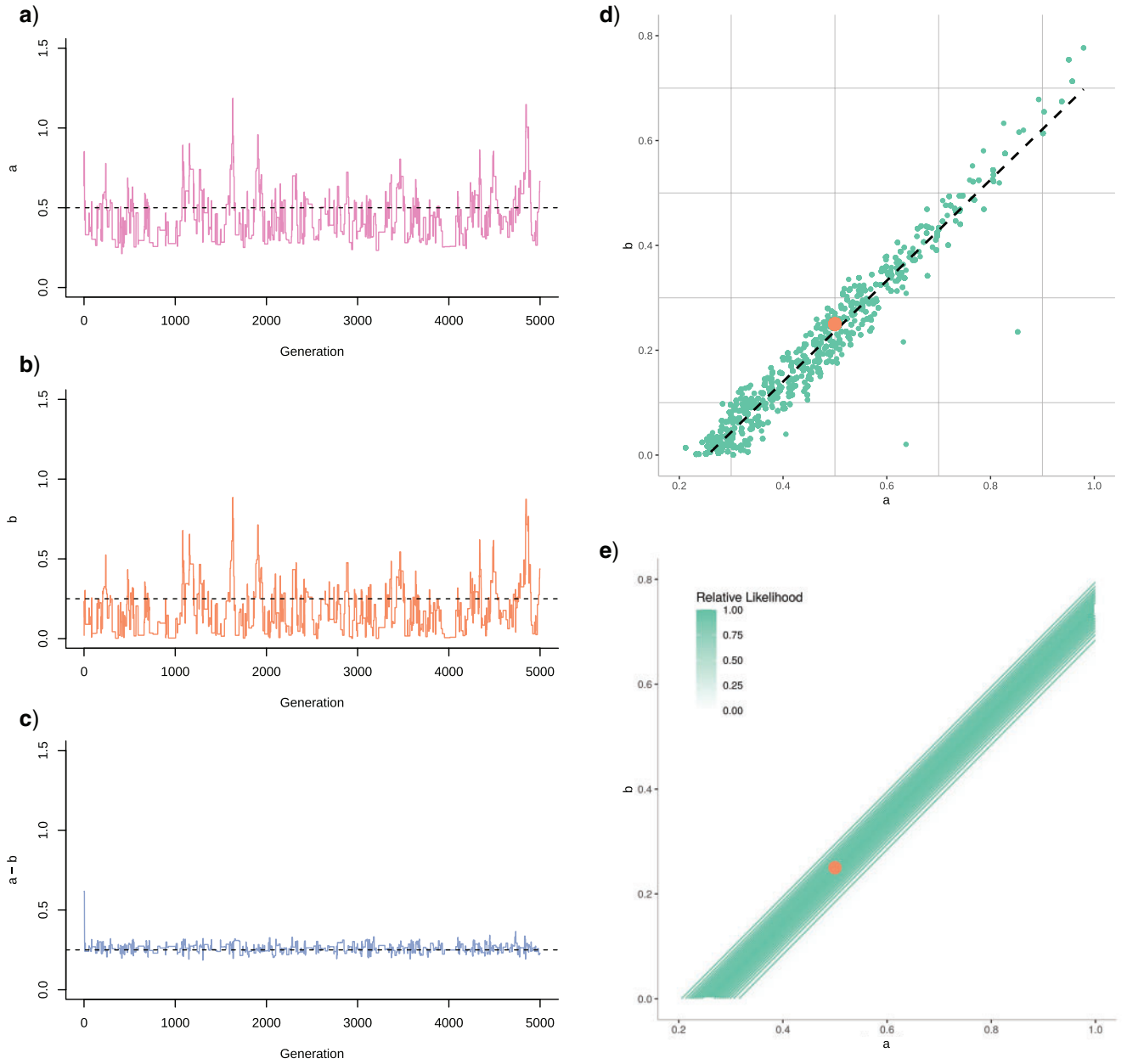


FIGURE 2. An example of unidentifiability issues encountered when trying to estimate values of two parameters (a and b) for the slope $a-b$. We used a Bayesian Monte-Carlo Markov Chain approach to try to estimate the values of a and b separately. We ran a chain for 5000 generations, sampling each generation. The traces for a) a and b) b show a great deal of uncertainty in the parameter estimates compared to the estimates for c) $a-b$. True values are shown as black dashed lines in panels (a–c) and as large circles in panels (d,e). We plotted a against b and found that the two parameters were highly correlated (d). We then calculated the relative likelihood over a range of parameters values and found a flat ridge in the likelihood where different pairs of values for a & b are equally likely—or unidentifiable (e). Further details and code showing how these analyses, and all others in this study were conducted can be found in the Supplementary material available on Dryad at https://doi.org/10.5061/dryad.5tb2rbp4g and in the associated rcompendium (Casajus 2021) at https://github.com/ajhelmstetter/pulledRates.

time-calibration, extant species sampling) it is represented by a single curve made up of one observation at any given point in time that represents the number of lineages in a clade (Fig. 1). Fitting a model to an LTT is like fitting two parameters (a and b) for a single value, the slope ($a-b$). If we try to estimate a and b separately we find it impossible (Fig. 2a,b). However, we can estimate $a-b$ (Fig. 2c) with much greater accuracy. Estimates of a and b are fully correlated (Fig. 2d) and we find a flat surface in the likelihood where different pairs of values for a and b are equally likely (Fig. 2e), signifying unidentifiability.

This problem has been highlighted previously in a macroevolutionary context (Nee 2006; Kubo and Iwasa 1995), where $a-b$ is akin to the net diversification rate ($r=\lambda-\mu$). However, the birth–death model is more

Downloaded from https://academic.oup.com/sysbio/article/71/3/758/6382322 by Ecole Francaise D'Athenes user on 23 April 2022

TABLE 1. A table detailing the parameters we can estimate with the lineages-through-time plot (LTT) approach used in Louca and Pennell (2020) when rates are either constant or time-dependent. When speciation and extinction rate are constant we are able to infer all traditional (r, λ, μ) and pulled rates (r_p, λ_p, μ_p). If sampling fraction (ρ) is unknown, we lose the ability to infer λ and μ. If λ and μ vary over time only pulled rates remain identifiable.

	Speciation	Extinction	Sampling fraction	Ref	Identifiable parameters
Constant rates	$\lambda = a$	$\mu = b$	$\rho = 1$	Nee et al. (1994)	r, λ, μ, r_p, λ_p, μ_p
	$\lambda = a$	$\mu = b$	$\rho < 1$ (known)	Nee et al. (1994)	r, λ, μ, r_p, λ_p, μ_p
	$\lambda = a$	$\mu = b$	$\rho < 1$ (unknown)	Stadler (2013)	r, r_p, λ_p, μ_p
Time-dependent rates	$\lambda = f(\tau)$	$\mu = g(\tau)$	$\rho = 1$	Louca and Pennell (2020)	r_p, λ_p, μ_p
	$\lambda = f(\tau)$	$\mu = g(\tau)$	$\rho < 1$ (known)	Louca and Pennell (2020)	r_p, λ_p, μ_p
	$\lambda = f(\tau)$	$\mu = g(\tau)$	$\rho < 1$ (unknown)	Louca and Pennell (2020)	r_p, λ_p, μ_p

complex than the example illustrated in Figure 2. As explained above, speciation and extinction rates are actually identifiable when time-independent because the slope of the LTT reaches λ at the present. Our ability to reliably estimate these traditional diversification rates (λ, μ, and r) depends on the amount of information we have available, and the assumptions we make in our model. For example, if the sampling fraction (ρ) is not known (or assumed) we can no longer reliably estimate λ and μ because this third unknown parameter in the model leads to unidentifiability (Table 1). However, in this case r remains identifiable, as the system is reduced to two parameters only (r and ρ). Likewise, as LP show, if we relax the assumption of constant rates (i.e., time-independent rates) and allow λ and μ to vary over time, then all traditional parameters become unidentifiable, including $r(\tau)$, even if ρ is known or assumed (Table 1). To exemplify the problem, LP used a very large phylogenetic tree of seed plants (Smith and Brown 2018) to show that the observed LTT is congruent with two opposing (but not symmetrical) scenarios (Fig. 2 in LP): either a continuous increase or a continuous decline in both $\lambda(\tau)$ and $\mu(\tau)$ (though the resulting diversification rates ($r(\tau)$) of these two scenarios are very similar). If we want to infer something from LTTs, then traditional model-based variables are inadequate, and we must look towards other possible solutions.

Pulled Rates

The solution proposed by LP is to use the approach described in Louca et al. (2018), namely not to estimate $\lambda(\tau)$, $\mu(\tau)$, and ρ, but "pulled" rates that can be directly measured from the shape of the LTT. There are three pulled rates (λ_p, μ_p, r_p) in Louca et al. (2018). These pulled rates are based directly on the dLTT—they can be calculated using the slope at a given time (τ) and the change in the slope, or curvature of the plot. Thus, any dLTT yields a unique set of pulled rates that summarize a congruence class, thereby eliminating the unidentifiability issue. However, these rates are not the speciation and extinction rates everyone knows—so what are they and how are they different from traditional rates?

An important consequence of using extant timetrees when investigating patterns of diversification is that LTT plots will likely underestimate the number of lineages at any given time because our trees are missing species (Ricklefs 2007; Silvestro et al. 2018). Species can be missing for two reasons: i) they went extinct or ii) they are extant but were not sampled. These two factors will have different effects on the LTT plot and our estimates of diversification rates. Extinction must have occurred in the past. Lineages that originated recently have had less time to go extinct (Nee et al. 1994; Ricklefs 2007), so the effect of extinction on our estimates using only extant species is reduced towards the present. As mentioned above, this leads to an increase in the rate of lineage accumulation towards the present, as the effect of extinction decreases, which occurs even when rates are constant (Fig. 3). Conversely, incomplete sampling of a group occurs up to the present day and more strongly affects estimates of recent history (Heath et al. 2008; Phillimore and Price 2008; Cusimano and Renner 2010), as the deeper nodes in the phylogeny can be reconstructed with only a few species. The relative importance of extinction and sampling fraction will influence whether λ_p departs from λ more in the past or in the present. To summarize, the presence of extinction will cause us to underestimate speciation rate further in the past, because the number of extinct species increases as we consider more time, while incomplete sampling will lead to underestimation of speciation rates that are more recent (Heath et al. 2008; Cusimano and Renner 2010; Brock et al. 2011).

Formally, at a given time (τ), λ_p is the estimated speciation rate multiplied by 1 minus the probability that a lineage is missing from the tree due to extinction or incomplete sampling, E. We will not go into details regarding the calculation of E here, but further information can be found in S1.1 of the supplementary material of LP. λ_p is calculated using the following equation:

$$\lambda_p(\tau) = \lambda(\tau)(1 - E(\tau)). \quad (3)$$

So, if all species are in the tree and there is no extinction (i.e., the probability of missing lineages, or E, is 0) then the λ_p is equal to the (unpulled) speciation rate, λ. Any increase in extinction rate or the number of unsampled lineages (i.e., $E > 0$) will cause λ_p to drop, or be "pulled," below speciation rate (Fig. 3). The lower the extinction rate and the greater the sampling fraction, the closer the estimate of λ_p will be to λ.

Downloaded from https://academic.oup.com/sysbio/article/71/3/758/6382322 by Ecole Francaise D'Athenes user on 23 April 2022

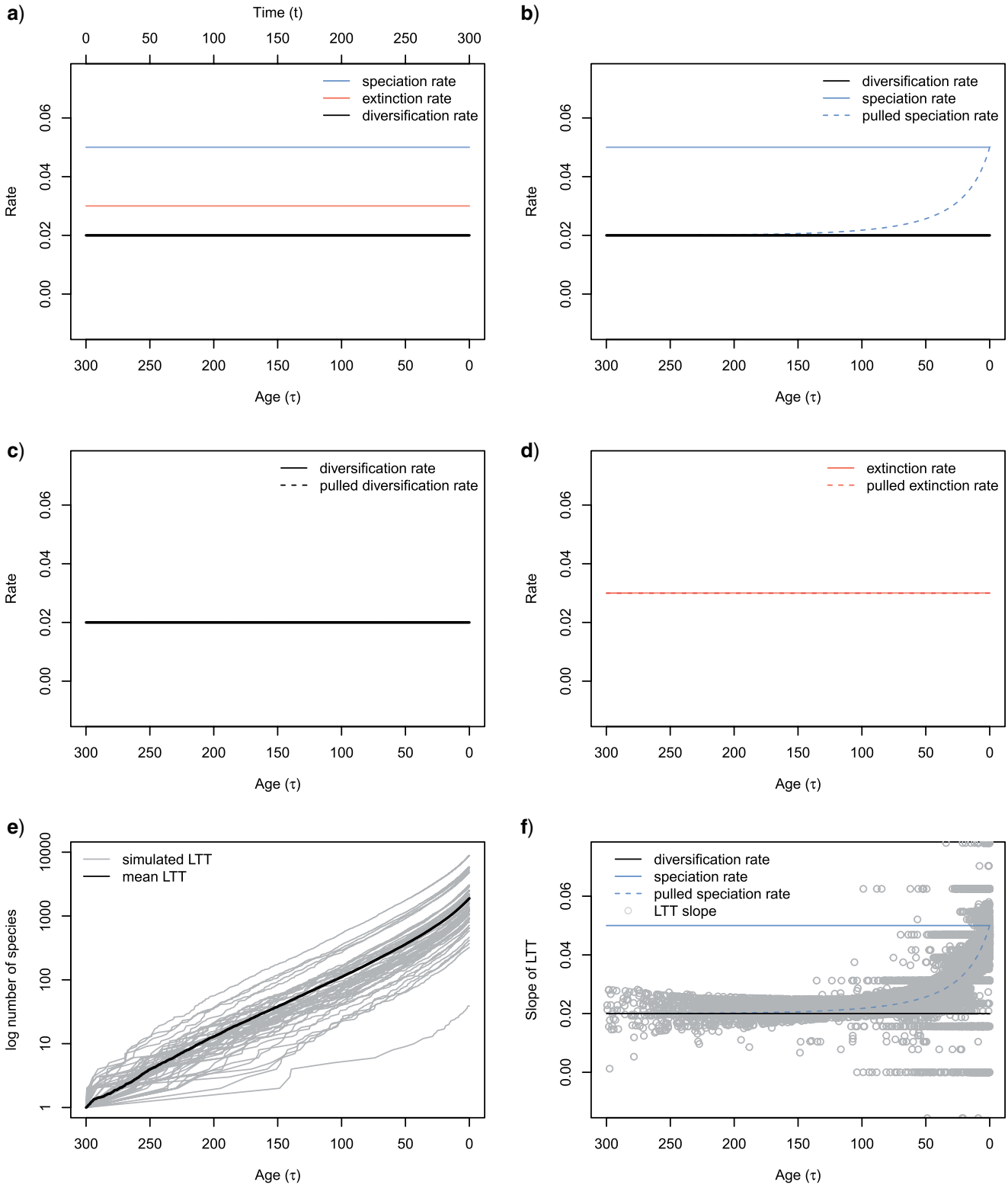


FIGURE 3. A simple example of the relationship between constant diversification rates and corresponding pulled rates. Panel a) shows values of speciation rate (λ), extinction rate (μ), and diversification rate (r) over time (τ). An additional axis, at the top of panel a) shows time since origin (t). Panel b) shows how in the past, pulled speciation rate (λ_p) is identical to the diversification rate (if sampling fraction = 100%) while closer to the present λ_p approaches speciation rate. The following two panels compare c) r and pulled diversification rate (r_p) and d) compares μ and pulled extinction rate (μ_p). In these two cases the pulled rates are identical to the traditional rates. Panel e) shows 50 LTT plots (grey lines) simulated with the parameters used in panels (a–d) and the mean LTT (black line). Panel f) shows the slopes of the LTTs in panel e) over time, matching λ_p and depicting the expected increase towards the present caused by the lack of effect of extinction—lineages do not have enough time to go extinct towards the present. Further details on how LTTs were simulated can be found in the Supplementary material available on Dryad. An interactive version of this plot, in which parameters can be modified, can be found at https://ajhelmstetter.shinyapps.io/pulled_rates/.

Downloaded from https://academic.oup.com/sysbio/article/71/3/758/6382322 by Ecole Francaise D'Athenes user on 23 April 2022

LP also use pulled diversification rate (r_p), a parameter that is similar to the net diversification rate ($r = \lambda - \mu$) but is again modified by another term. This new term is the relative ($\frac{1}{\lambda}$) rate of change in speciation rate over time ($\frac{d\lambda}{d\tau}$). This causes the pulled diversification rate to lag behind the unpulled rate. The "pull" of r_p is actually a delay in the response of this parameter when compared to diversification rate. This is in contrast to the "pull" of λ_p, which refers to a reduction in the estimated value of λ_p relative to λ. Pulled diversification rate is given by the following equation:

$$r_p = \lambda - \mu + \frac{1}{\lambda} \cdot \frac{d\lambda}{d\tau}. \quad (4)$$

Consider an example where we have an increase in speciation rate at around 100 Ma in a clade (Fig. 4). When considering time as an age (using τ), speciation rate increases as τ decreases from the origin of the clade ($\tau = 300$ Ma) to the present ($\tau = 0$). This means that when speciation accelerates, $\frac{d\lambda}{d\tau}$ is negative. This leads to a "drop" in r_p (Fig. 4c) before it stabilizes at a new value of r_p that is higher than the previous value, reflecting the increase in λ. However, the change in the slope of the LTT plot (Fig. 4e,f) is very small, so this is not precisely measurable from a realistically sized phylogenetic tree (see Box 1 for additional discussion). We note that LP also defined a pulled extinction rate, (μ_p), which can be calculated from λ_p and r_p but does not add any new information, so we do not discuss it here (see LP, Louca et al. (2018) for further details).

With these new metrics we can revisit questions such as: has diversification been constant over time? Pulled rates can be estimated with many commonly used models of diversification (Louca and Pennell 2020). For example, λ_p is the speciation rate one would get by constraining extinction to be 0 and assuming complete species sampling. For r_p, this involves estimating r by making λ time-independent. In summary, λ_p provides information about how λ changes over time while taking into account past extinction and the proportion of lineages sampled. r_p provides a slightly delayed estimate of r with extreme responses to rapid changes in λ. While λ_p can be very different from the underlying speciation and extinction rates, r_p is close to the net diversification rate as long as λ does not change too rapidly. To illustrate this, we performed simulations to identify some extreme scenarios in which the pulled rates differ markedly from their nonpulled counterparts (see Supplementary material available on Dryad), but such cases correspond to specific and unrealistic scenarios of rate variation.

Implications for Analyses of Species Diversification

Pulled Rates Can Be Difficult to Interpret

LP compared the usefulness of pulled rates to effective population size (N_e) in population genetics. N_e can be broadly defined as the number of breeding individuals in an idealized population (e.g., constant size, random mating etc.) that would be able to explain the summary statistics (e.g., amount of polymorphism, level of inbreeding) of an observed population. N_e is fairly intuitive and will react to biological phenomena in expected ways (e.g., under population structure; Whitlock and Barton 1997 or nonrandom mating; Caballero and Hill 1992). LP state that the variables they introduce are "easily interpretable." Their terminology, however, is not completely consistent nor coherent with more traditional uses, which can cause confusion. Given that $r = \lambda - \mu$ one might intuitively think that $r_p = \lambda_p - \mu_p$, but this is not the case. Pulled rates are simply different ways of summarizing congruence classes and each one is calculated using both speciation and extinction rates. λ_p is reasonably intuitive, though given that extinction is also included it is more similar to a diversification rate than a speciation rate. Indeed, λ_p is defined as the slope of the LTT plot (Louca et al. 2018) (see Figs. 3f and 4f), which corresponds to the past diversification rate, and to the speciation rate at present in the case that all extant species are included (Nee et al. 1994).

Pulled diversification rate, however, is much more difficult to interpret, perhaps initially because the "pull" of r_p is not the same as the "pull" of λ_p (see Box 2). Whereas λ_p decreases in value relative to λ, r_p is delayed in time relative to r (Fig. 5) and could therefore better be termed as "delayed" rather than "pulled." We simulated a variety of commonly investigated diversification scenarios, from simple to more complex (Fig. 5), and show that r_p and r are similar in most cases. However, r_p is at first sight not as intuitive as r or N_e. For example, drastic increases in r can lead to sharp decreases in r_p (Fig. 5a). The inverted pattern r_p presents in this case would make it challenging to present in a clear and concise way. Even so, compared to other pulled rates r_p could be particularly useful, not as an effective parameter like N_e, but as a reasonable approximation of the value of r. To investigate this further, we conducted an exploration of many different scenarios (see Supplementary material available on Dryad) and find that in a large majority of cases r_p is close to r and therefore can be used as a reasonable approximation of r for biological interpretation. The worst case arises when there are parallel patterns of variation in λ and μ. In this case r remains constant while r_p can fluctuate wildly (see Fig. 5d for an example). Yet, we found that if we introduce slight differences in the patterns of variation of the two rates r_p immediately begins to resemble r. But how would we know if λ is varying too rapidly to produce viable estimates? A general guideline would be to be cautious with inferences when λ_p and r_p are very different, that is, in cases where the slope of the LTT changes a lot. Conversely, when these two pulled rates are close it means that biological interpretations can reasonably be done. Attempting to biologically interpret fine-scale variations in r_p would certainly lead to spurious conclusions. However, changes in r_p at a large scale are good proxies for large scale variation in r. This is clearly illustrated in Figure 5a where the main trend of the r_p is a recent increase in diversification. Given that we are

Downloaded from https://academic.oup.com/sysbio/article/71/3/758/6382322 by Ecole Francaise D'Athenes user on 23 April 2022

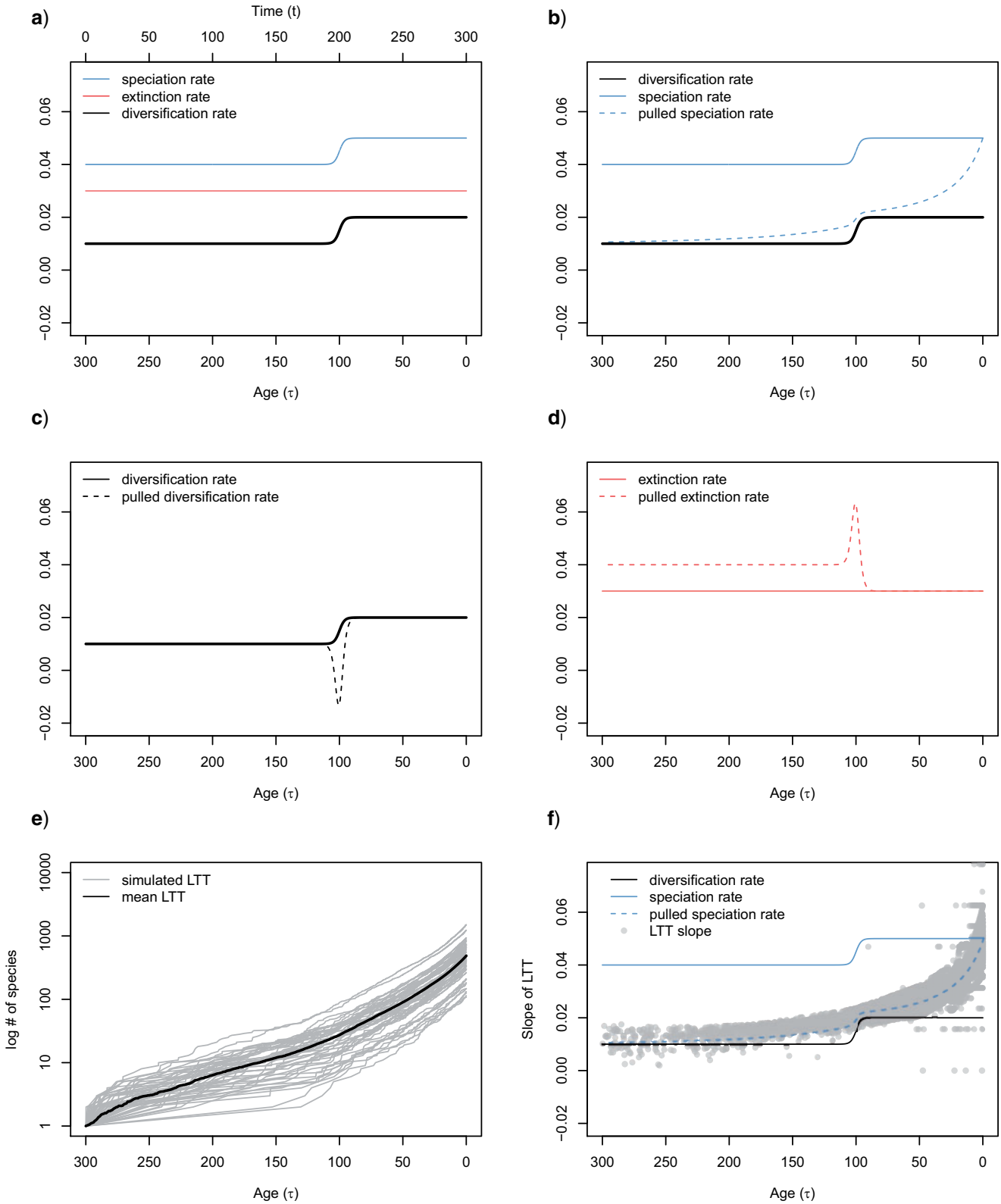


FIGURE 4. A slightly more complex example of the relationship between diversification rates and corresponding pulled rates where a single shift—an increase in speciation rate—has taken place. Panel a) shows values of speciation rate (λ), extinction rate (μ), and diversification rate (r) over time. An additional axis, at the top of panel a) shows time since origin (t). Panel b) shows the gradual change in pulled speciation rate (λ_p) during the shift in λ. Panel c) compares r and pulled diversification rate (r_p). The rapid increase in λ causes r_p to decrease suddenly before recovering to the new r. Panel d) compares μ and pulled extinction rate (μ_p) and shows an inverse pattern to panel (c). Panel e) shows 50 LTT plots (grey lines) simulated with the parameters used in panels a–d) and the mean LTT (black line). Panel f) shows the slopes of the LTTs in panel e) over time, matching λ_p and again depicting the expected increase towards the present caused by the lack of effect of extinction.

Downloaded from https://academic.oup.com/sysbio/article/71/3/758/6382322 by Ecole Francaise D'Athenes user on 23 April 2022

BOX 2: THE "PULL" IN r_p IS A RESULT OF THE LAG TIME BETWEEN EXTINCTION AND SPECIATION

Consider a simple case with no extinction ($\mu = 0$) so that changes in r only come from changes in λ. If so, $r = \lambda$ but r_p is not exactly λ because of temporal variations in λ (the term $\frac{1}{\lambda} \cdot \frac{d\lambda}{d\tau}$ in equation 4). LP suggest that "*the pulled diversification rate can be interpreted as the effective net diversification rate if* λ *was time-independent*." In our example, this means replacing a scenario, where μ is constant (at 0) and λ varies with a scenario where λ is constant and μ varies, as in LP. The difficulty of using μ to explain variation is that there is a slight delay between the effect of speciation and the effect of extinction. It is necessary to wait for species to arise before they can go extinct.

As mentioned previously, lineages that originated more recently have had less time to go extinct. In a constant birth–death process, this is only visible in recent history: the slope of the LTT plot is $r = \lambda - \mu$ during most of the past but increases to λ for very recent times where the stationary behavior has not yet been reached (Fig. 3). However, this phenomenon is not unique to very recent times—it will also occur whenever there is a change in speciation rate. Ultimately, this is the cause of the difference between r_p and r.

To illustrate this, imagine a massive increase in the number of lineages caused by a burst of speciation (Fig. 4). Over a short time period many new lineages have become available for potential extinction but they have yet to go extinct because not enough time has passed since they appeared for extinction to take place. There is now a disequilibrium between speciation and extinction, manifested as a lag in the time extinction takes to affect all of the new lineages. As time continues, these numerous new lineages will begin to go extinct, meaning that frequency of species extinction will increase to "catch up" to speciation and reach a new stationary point. This effect is stronger when λ varies rapidly (i.e., high $\frac{1}{\lambda} \cdot \frac{d\lambda}{d\tau}$). Conversely, speciation cannot occur in a lineage after it has gone extinct, so there is no similar lag caused by changes in extinction rate. This is also why variation in extinction rate would not cause r_p to deviate from r (Fig. 5b).

often interested in variation in r (rather than in both λ and μ) at a large-scale we can be less pessimistic than the identifiability result of LP would suggest. LP clearly issue a warning on overinterpretation of phylogenetic data, but our exploration of pulled rates suggests that it is still possible to get modest (in term of precision) but biologically meaningful information about variation in diversification rate. Given the difficulty of the interpretation of r_p, further work should be done to better understand the behavior of this composite parameter under different scenarios.

In summary, pulled rates are advantageous because they are nonparametric estimates that do not suffer from unidentifiability issues and they can be estimated using only information contained in the LTT plot. However, they cannot be directly interpreted in biologically meaningful terms. To estimate rates that are meaningful (e.g., λ, μ, and r) we need to make further assumptions such as constant rates of speciation and extinction over time. Likewise, if we wish to test hypotheses about species diversification using pulled rates it will be important to remember that such parameters are complicated transformations of the traditional rates we are familiar with. Framing and testing hypotheses with pulled rates is currently possible but will be challenging until we get a better grasp of their nature and utility. To alleviate this issue, we call for a more thorough investigation of the relationship between traditional and pulled rates during scenarios of interest (e.g., Fig. 5) with extensive simulations to make them more tractable and easier to interpret. Furthermore, we need studies on their inference and performance using phylogenetic trees based on empirical data.

Diversification Rates Vary among Clades

An LTT plot is a simplified way to visualize and summarize a time-calibrated phylogenetic tree, which relies on the assumption that all lineages diversify at equal rates (Morlon et al. 2011). Thus, under the assumption that λ and μ are clade-independent, LP showed that the LTT plot contains the complete information about the underlying branching process (see also Lambert and Stadler 2013). This provided the opportunity for LP to show mathematically how LTTs can lead to misinterpretation. Though the LTT plot is appropriate for a clade-independent framework, such as in LP, it lacks information that is critical for much of the macroevolutionary modeling done today. For example, useful information from extinct species such as fossils can play no role. More importantly, it does not take rate variation among lineages into account, so it is impossible to identify such variation on different parts of the tree from which the LTT was plotted.

LP show how we can test hypotheses about whether diversification rates deviate from constancy over time using pulled rates. We would be unable to pin this on changes in speciation or extinction rates, but would

Downloaded from https://academic.oup.com/sysbio/article/71/3/758/6382322 by Ecole Francaise D'Athenes user on 23 April 2022

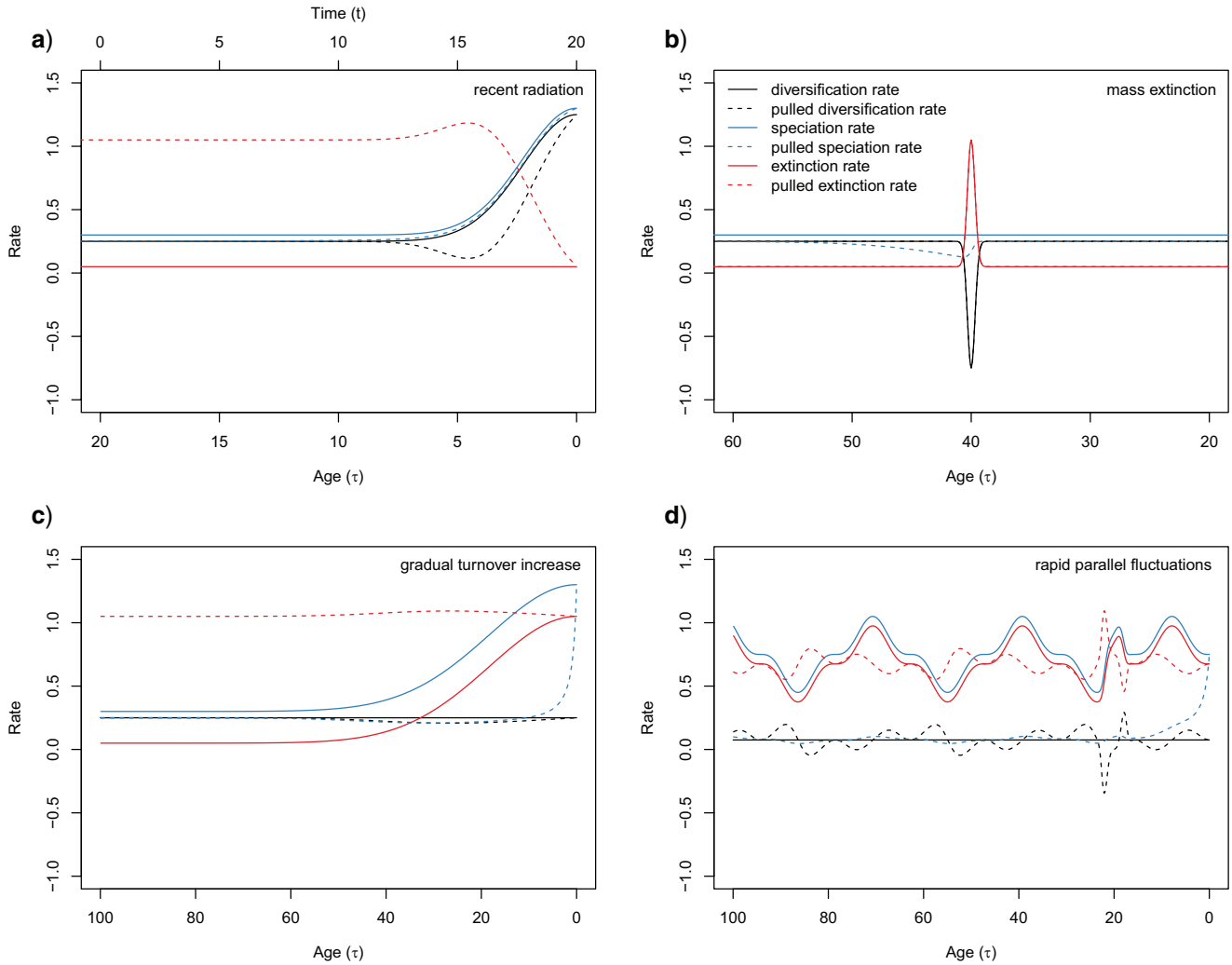


FIGURE 5. Comparison of traditional and pulled rates under three simulated diversification scenarios that are commonly investigated (a–c) and a final, more complex scenario. Panel a) shows a recent radiation where diversification rate and speciation rate sharply increase towards the present. An additional axis, at the top of panel a) shows time going forward (t). Panel b) shows a mass extinction event at 40 Ma in which extinction rate briefly but rapidly increases and then falls back to previous levels. Panel c) shows a gradual increase in species turnover rate (both speciation and extinction rate increase slowly over time). Panel d) shows a scenario where speciation and extinction rates are rapidly varying in parallel to each other. This results in a constant diversification rate (r) and a fluctuating pulled diversification rate (r_p).

get a sense of how variable diversification has been (Burin et al. 2019). This would be useful for testing whether diversification in particular clades has remained constant or been subject to large shifts in diversification (e.g., mass extinctions) but not when diversification rate has shifted in a subclade (e.g. due to the evolution of a key innovation).

The first use of pulled rates was in Louca et al. (2018), where they studied bacterial diversification, stating "*Our findings suggest that, during the past 1 billion years, global bacterial speciation and extinction rates were not substantially affected during the mass extinction events seen in eukaryotic fossil records.*" This might suggest that nothing particularly extraordinary happened in the macroevolutionary dynamics of bacteria in the last billion years. However, it is important to note that to be able to infer how rates vary over time, the framework used in Louca et al. (2018) and Louca and Pennell (2020) makes the assumption that rates do not vary among clades. Louca et al. (2018) touch on this point themselves: "*It is possible that diversification within individual bacterial clades may have been influenced by eukaryotic radiations and extinctions, and that these cases are overshadowed when considering all bacteria together.*" The rates estimated using such clade-independent models will correspond to the average rates over time in the entire study group, therefore missing out on any variation among clades—for example any difference in diversification rates between those species that use terrestrial *versus* marine environments (Louca et al. 2018). Subclades are important in driving inferred diversification patterns (see Morlon et al. 2011; Maliet et al. 2019; Rabosky 2020), so ignoring this variation may mean that we miss out on influential and interesting dynamics when using pulled rates. Such an assumption

Downloaded from https://academic.oup.com/sysbio/article/71/3/758/6382322 by Ecole Francaise D'Athenes user on 23 April 2022

enables us to investigate patterns of diversification through time but prevents us from investigating some of the most interesting and fundamental questions in macroevolutionary biology, notably, why are some clades more diverse than others?

To demonstrate how contrasting patterns of speciation and extinction can produce the same LTT plot, LP used a large phylogenetic tree of seed plants (Smith and Brown 2018) that contained 79,874 species. These range from small ephemeral plants like *Arabidopsis thaliana* to gigantic, long-lived trees such as *Eucalyptus regnans*. A large amount of research has shown that diversification rates have varied significantly among seed plant clades (e.g., O'Meara et al. 2016; Igea et al. 2017; Vamosi et al. 2018; Onstein 2020; Soltis et al. 2019; Magallón et al. 2019; Zenil-Ferguson et al. 2019). By assuming that rates are clade-independent we can use pulled rates to get a global view of what has happened in seed plants, but this goes little way towards understanding the evolutionary processes that have shaped their diversity.

Fortunately, the assumption of clade-independent rates is not common in modern approaches. Some models assume that rates are both clade-dependent and time-dependent such as Bayesian Analysis of Macroevolutionary Mixtures (BAMM) (Rabosky et al. 2013; Rabosky 2014), relaxing the assumption that all lineages share the same evolutionary rates at a given point in time (Rabosky 2017). This is a key difference from the models used in LP because it allows lineages to differ in their rates of speciation and extinction. With BAMM, the entire phylogeny could be described using a model similar to what is used in Louca and Pennell (2020), or alternatively, it could be described using multiple processes that explain rates of diversification on different parts of the tree. Within each of these processes, λ and μ are still susceptible to the same unidentifiablity issues outlined in LP because they are usually estimated from extant timetrees. However, BAMM makes use of the full topology that includes information (e.g., branch lengths) that an LTT lacks and also allows for clade-dependent rates. Often, the goal of using such approaches is to determine whether one clade is diversifying faster than another. While LP are concerned with estimating exact, identifiable estimates of diversification rate parameters this may not be important if what we are actually interested in is the difference between diversification rates in two clades e.g. are diversification rates always higher in clade A than clade B? These are two fundamentally different goals that, at the most basic level, require different models and assumptions to assess. Understanding the interaction between clade-dependent and time-dependent rates is an important avenue of future research and will make issues raised in LP more relevant to the richer, more modern modeling approaches used today.

Time-Independent, Clade-Dependent Models Are Not Implicated

Another class of models are those that allow rates to vary among clades but keep them fixed over time. Among them is the Binary-State Speciation and Extinction (BiSSE) model (Maddison et al. 2007), part of a family of models known as the state-dependent models of diversification (SSE models; Ng and Smith 2014; Beaulieu and O'Meara 2016; O'Meara and Beaulieu 2016; Caetano et al. 2018). These models are extensions of the birth–death model that also include information about character states of extant species. SSE models jointly estimate ancestral states at each node of the phylogenetic tree, rates of transition between character states, and state-dependent diversification rates that remain constant through time. As in BAMM, BiSSE makes use of evidence not included in the LTT, such as the full tree topology (Maddison et al. 2007). It also considers character state evolution, rather than just the timing of branching events as in the LTT plot. LP suggest that it remains unclear how the dependence on character states (which, if removed, collapses equations in BiSSE to those shown in Nee et al. (1994)) affects the unidentifiability issues they highlight. They go on to state that the likelihood functions of SSE models are too complex to be addressed in their study, but suggest that the same problems they uncover probably still apply.

Here, we argue that this may not be the case. Unlike BAMM, SSE models such as BiSSE are not richer versions of the framework used in LP, they are members of a different class of models. In this case, the assumed time-independence and the clade-dependence of rates are the inverse of the assumptions at the base of LP. SSE models examine rate variation that is impossible to detect when summarizing a tree with just one LTT, so it does not follow that issues found in the single LTT-based framework in LP apply. The assumption of time-independence in SSE models is important because it allows traditional rate parameters to be identifiable. Depending on our knowledge of sampling fraction, this may be all traditional rates, or net diversification rate only (Table 1). As an example, recall that LP show that λ equals λ_p when i) sampling fraction is 100% and ii) $\mu = 0$. Complete sampling was assumed in the original BiSSE model (Maddison et al. 2007), though it has since been relaxed (FitzJohn et al. 2009) so that a given sampling fraction can be assumed instead. Extinction can be set to 0 in SSE models, as this is a constant rate, and enables the estimation of λ (e.g., Joly and Schoen 2021). Therefore, under the conditions outlined in LP, SSE models can estimate λ, and therefore μ and r as well. Similarly, r_p equals r when λ is constant, which is also an assumption in BiSSE (and any other time-independent model) and suggests that BiSSE produces an identifiable estimation of r. It is clear from these examples that using time-independent, clade-dependent approaches like BiSSE can overcome (or pre-empt) the unidentifiability issues raised in LP because they allow the congruence class to collapse to a single model. However, this does not mean that BiSSE actually estimates the different diversification rates of a clade correctly; LP's results show we should take these BiSSE estimates as proxies. Users of SSE-like models should therefore take care when interpreting values of

Downloaded from https://academic.oup.com/sysbio/article/71/3/758/6382322 by Ecole Francaise D'Athenes user on 23 April 2022

rate parameters that are time-independent as we know that diversification rates can vary over time. Moreover, many studies report whether diversification rates among groups are significantly different, rather than speciation and extinction rates (cf Onstein 2020), as it is generally acknowledged that it is difficult to disentangle speciation and extinction (see also Burin et al. 2019; Louca and Pennell 2021). It is indeed possible to extend SSE models to estimate time-dependent speciation and extinction rates (Rolland et al. 2014; Condamine et al. 2017; Meseguer and Condamine 2020) but such models likely suffer from the unidentifiability issues formalized in LP. However, we stress that the likelihood of time-dependent diversification models (as in LP) is not the same as the likelihood of clade-dependent diversification models (e.g., typical SSE models), and what is unidentifiable in the former does not tell us about identifiability in the latter.

Overall, the issues raised in LP cannot yet be readily applied to commonly used macroevolutionary approaches without further work to show that criticisms related to approaches that assume rates are time-dependent and clade-independent (i.e., the LTT) are applicable to these richer or altogether different models. Yet, even if unidentifiability issues remain in such models they may not be relevant to the questions the models were built to answer, for example when the objective is to determine whether one clade has higher diversification rates than another. In cases like these, it is not the precise values of rates that are important but instead the rate difference that is, whether rates in one group of lineages are higher than another. Perhaps most importantly, this means that we should not forego building models that estimate diversification rates because one, relatively simplistic approach has long-standing problems (Kubo and Iwasa 1995) but instead continue to improve them and build upon the work done in LP. A case in point is the issue of false positives when using SSE models that was raised by Rabosky and Goldberg (2015). This criticism spurred on innovation that led to the development of models with hidden states (Beaulieu and O'Meara 2016), which are now present in various new incarnations of the SSE approach (e.g., Caetano et al. 2018; Herrera-Alsina et al. 2019).

On the Use of Models

The discussion sparked by LP highlights an important issue: evolutionary biologists should be interested in the actual history of diversification of the clades they study but, at the same time, must keep in mind that this history is difficult to infer. The framework developed by Louca et al. (2018) shows how to do this using the shape of the LTT plot, without making strong assumptions about past speciation and extinction rates. Indeed, the slope and curvature of the LTT plot contain useful information about the diversification history of the clade. Much of the debate, however, focuses on the ability to recover a "true" history of diversification. The goal of a scientific study should be to find out what really happened, but it becomes confusing if one considers a simulated birth–death process as the "true" history. This birth–death process is determined by two parameters (λ and μ) that can vary over time, and these parameters are supposed to correspond to the rate that a lineage splits into two lineages, or goes extinct. In reality, however, a species does not have a speciation and an extinction rate in the same way it has a geographic distribution and a population size. These rates only make sense when they are aggregated over a number of species and a certain amount of evolutionary time. That is, they are descriptive statistics summarizing much more complex processes that are acting at the microevolutionary level, and that would eventually lead to speciation or extinction. LP convincingly show that one cannot estimate these statistics reliably from LTT plots, and thus propose statistics that can be estimated more reliably. That these alternative statistics do not exactly correspond to the parameters of the birth–death process is not a problem; the birth–death process is only a model of diversification, and not the truth about diversification itself. The framework built by Louca et al. (2018) and LP allows us to use the LTT plot to test whether diversification rates were constant or not. If researchers want to understand how speciation and extinction actually changed to give rise to a diversification history, they will have to use other methods and data.

Conclusion

Louca and Pennell (2020) have pointed out key issues with how we approach macroevolutionary modeling, namely the inability to distinguish historical diversification scenarios under certain circumstances. Their formalization of the unidentifiability issues in models with time-dependent, clade-independent diversification rates is an important step forward that provides the mathematical tools to study the associated issues further. LP highlight the avenues we must consider and develop upon to ensure we do not make similar mistakes in the future. Whether variation in diversification rate is due to changes in speciation or extinction is certainly an interesting avenue of research, but LP have shown that exploring this would require much more than just fitting a model with speciation and extinction rates to an LTT plot. Indeed, more recent diversification models go beyond this by making use of additional information that is ignored by models relying only on an LTT. Awareness and consideration of potential unidentifiability issues is important for macroevolutionary biologists going forward when they employ such models of diversification. However, it is important to note that LP do not show that speciation and extinction rates cannot be estimated with evolutionary trees (Pagel 2020). Instead, they show that when using extant timetrees under the assumptions of an LTT-based approach, unidentifiability issues are encountered in the estimation of speciation and extinction rates, and that these problems can be

Downloaded from https://academic.oup.com/sysbio/article/71/3/758/6382322 by Ecole Francaise D'Athenes user on 23 April 2022

circumvented by making use of pulled rates, or by making other assumptions. Further work is needed to identify the extent to which the issues raised in LP apply to the more complex models of diversification used today. Comparisons should be made in empirical studies that use both traditional and pulled rates, to see if differences in results exist between these approaches in practice. In the meantime, it is important that the field continues to grow by using and building upon modern macroevolutionary methods, albeit with a critical eye.

Code availability

Code associated with this manuscript is available from the following rcompendium: https://github.com/ajhelmstetter/pulledRates.

Supplementary material

Supplementary text and data are available through the Dryad Digital Repository: https://doi.org/10.5061/dryad.5tb2rbp4g.

Funding

This work was supported by an "Investissements d'Avenir" grant managed by the "Agence Nationale de la Recherche" (CEBA, ref. ANR-10-LABX-25-01 to N.M.); funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (Project GLOBAL, grant agreement No. 865787) to T.L.P.C.; funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (project GAIA, grant agreement No. 851188) to E.L.R. and F.L.C.

Acknowledgments

We thank Mark Holder, Matt Pennell and an anonymous reviewer who provided insightful comments to improve the manuscript. This article was initially conceived from discussions held with the Macroevolution in Montpellier (MiM) group (https://fabiencondamine.org/mim/). We thank Stilianos Louca as well as Hélène Morlon and her research group for comments on an earlier version of this manuscript. We also thank Marcos Mendez, Sally Otto, Dan Schoen, Jürg Schönenberger, and the rest of the DiveRS group for comments and discussion throughout. For R.Z.F., this is publication #147 from the School of Life Sciences, University of Hawai'i at Mānoa.

References

Aldous D.J., Krikun M.A., Popovic L. 2011. Five statistical questions about the tree of life. Syst. Biol. 60(3):318–328.

Alfaro M.E., Santini F., Brock C., Alamillo H., Dornburg A., Rabosky D.L., Carnevale G., Harmon L.J. 2009. Nine exceptional radiations plus high turnover explain species diversity in jawed vertebrates. Proc. Natl. Acad. Sci. USA 106(32):13410–13414.

Barido-Sottani J., Vaughan T.G., Stadler, T. 2020. A multitype birth–death model for Bayesian inference of lineage-specific birth and death rates. Syst. Biol. 69(5):973–986.

Beaulieu J.M., Donoghue M.J. 2013. Fruit evolution and diversification in campanulid angiosperms. Evolution 67(11):3132–3144.

Beaulieu J.M., O'Meara B.C. 2016. Detecting hidden diversification shifts in models of trait-dependent speciation and extinction. Syst. Biol. 65(4):583–601.

Brock C.D., Harmon L.J., Alfaro M.E. 2011. Testing for temporal variation in diversification rates when sampling is incomplete and nonrandom. Syst. Biol. 60(4):410–419.

Burin G., Alencar L.R.V., Chang J., Alfaro M.E., Quental T.B. 2019. How well can we estimate diversity dynamics for clades in diversity decline? Syst. Biol. 68(1):47–62.

Caballero A., Hill W.G. 1992. Effective size of nonrandom mating populations. Genetics 130(4):909–916.

Caetano D.S., O'Meara B.C., Beaulieu J.M. 2018. Hidden state models improve state-dependent diversification approaches, including biogeographical models: HMM and the adequacy of SSE models. Evolution 72(11):2308–2324.

Casajus N. 2021. rcompendium: an R package to create a package or research compendium structure. R package version 0.5.1.

Condamine F.L., Leslie A.B., Antonelli A. 2017. Ancient islands acted as refugia and pumps for conifer diversity. Cladistics 33(1): 69–92.

Condamine F.L., Nagalingum N.S., Marshall C.R., Morlon H. 2015. Origin and diversification of living cycads: a cautionary tale on the impact of the branching process prior in Bayesian molecular dating. BMC Evol. Biol. 15(1):65.

Condamine F.L., Rolland J., Morlon H. 2013. Macroevolutionary perspectives to environmental change. Ecol. Lett. 16(s1):72–85.

Crisp M.D., Cook L.G. 2009. Explosive radiation or cryptic mass extinction? Interpreting signatures in molecular phylogenies. Evolution 63(9):2257–2265.

Cusimano N., Renner S.S. 2010. Slowdowns in diversification rates from real phylogenies may not be real. Syst. Biol. 59(4):458–464.

Duan T., Deng X., Chen S., Luo Z., Zhao Z., Tu T., Khang N.S., Razafimandimbison, S.G., Zhang, D. 2018. Evolution of sexual systems and growth habit in Mussaenda (Rubiaceae): insights into the evolutionary pathways of dioecy. Mol. Phylogenet. Evol. 123:113–122.

Etienne R.S., Rosindell J. 2012. Prolonging the past counteracts the pull of the present: protracted speciation can explain observed slowdowns in diversification. Syst. Biol. 61(2):204–204.

FitzJohn R.G. 2012. Diversitree : comparative phylogenetic analyses of diversification in R: *Diversitree*. Methods Ecol. Evol. 3(6):1084–1092.

FitzJohn R.G., Maddison W.P., Otto S.P. 2009. Estimating trait-dependent speciation and extinction rates from incompletely resolved phylogenies. Syst. Biol. 58(6):595–611.

Harvey P.H., May R.M., Nee S. (1994). Phylogenies without fossils. Evolution 48(3):523–529.

Heath T.A., Huelsenbeck J.P., Stadler T. 2014. The fossilized birth–death process for coherent calibration of divergence-time estimates. Proc. Natl. Acad. Sci. USA 111(29):E2957–E2966.

Heath T.A., Zwickl D.J., Kim J., Hillis D.M. 2008. Taxon sampling affects inferences of macroevolutionary processes from phylogenetic trees. Syst. Biol. 57(1):160–166.

Herrera-Alsina L., van Els P., Etienne R.S. 2019. Detecting the dependence of diversification on multiple traits from phylogenetic trees and trait data. Syst. Biol. 68(2):317–328.

Ho S.Y.W., Phillips M.J. 2009. Accounting for calibration uncertainty in phylogenetic estimation of evolutionary divergence times. Syst. Biol. 58(3):367–380.

Hutter C.R., Lambert S.M., Wiens J.J. 2017. Rapid diversification and time explain amphibian richness at different scales in the Tropical Andes, Earth's most biodiverse hotspot. Am. Nat. 190(6):828–843.

Höhna S., Landis M.J., Heath T.A., Boussau B., Lartillot N., Moore B.R., Huelsenbeck J.P., Ronquist F. 2016. RevBayes: Bayesian phylogenetic inference using graphical models and an interactive model-specification language. Syst. Biol. 65(4):726–736.

Downloaded from https://academic.oup.com/sysbio/article/71/3/758/6382322 by Ecole Francaise D'Athenes user on 23 April 2022

Igea J., Miller E.F., Papadopulos A.S.T., Tanentzap A.J. 2017. Seed size and its rate of evolution correlate with species diversification across angiosperms. PLoS Biol. 15(7):e2002792.
Igea J., Tanentzap A.J. 2020. Angiosperm speciation cools down in the tropics. Ecol. Lett. 23(4):692–700.
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(7424):444–448.
Joly S., Schoen D.J. 2021. Repeated evolution of a reproductive polyphenism in plants is strongly associated with bilateral flower symmetry. Curr. Biol. 31(7):1515–1520.e3.
Kubo T., Iwasa Y. (1995). Inferring the rates of branching and extinction from molecular phylogenies. Evolution 49(4):694–704.
Lagomarsino L.P., Condamine F.L., Antonelli A., Mulch A., Davis C.C. 2016. The abiotic and biotic drivers of rapid diversification in Andean bellflowers (Campanulaceae). New Phytol. 210(4):1430–1442.
Lambert A., Stadler T. 2013. Birth–death models and coalescent point processes: the shape and probability of reconstructed phylogenies. Theor. Popul. Biol. 90:113–128.
Louca S., McLaughlin A., MacPherson A., Joy J.B., Pennell M.W. 2021. Fundamental identifiability limits in molecular epidemiology. Mol. Biol. Evol. 38(9):4010—4024.
Louca S., Pennell M.W. 2020. Extant timetrees are consistent with a myriad of diversification histories. Nature 580(7804):502–505.
Louca S., Pennell M.W. 2021. Why extinction estimates from extant phylogenies are so often zero. Curr. Biol. 31(14):3168—3173.e4
Louca S., Shih P.M., Pennell M.W., Fischer W.W., Parfrey L.W., Doebeli M. 2018. Bacterial diversification through geological time. Nat. Ecol. Evol. 2(9):1458–1467.
Maddison, W. P., Midford, P. E., and Otto, S. P. (2007). Estimating a binary character's effect on speciation and extinction. Syst. Biol. 56(5):701–710.
Magallón S., Sánchez-Reyes L.L., Gómez-Acevedo S.L. 2019. Thirty clues to the exceptional diversification of flowering plants. Ann. Bot. 123(3):491–503.
Maliet O., Hartig F., Morlon H. (2019). A model with many small shifts for estimating species-specific diversification rates. Nat. Ecol. Evol. 3(7):1086–1092.
May M.R., Höhna S., Moore B.R. 2016. A Bayesian approach for detecting the impact of massâŁextinction events on molecular phylogenies when rates of lineage diversification may vary. Methods Ecol. Evol. 7(8):947–959.
McGuire J., Witt C., Remsen J.V., Corl A., Rabosky D., Altshuler D., Dudley R. 2014. Molecular phylogenetics and the diversification of hummingbirds. Curr. Biol. 24(8):910–916.
Meseguer A.S., Condamine F.L. 2020. Ancient tropical extinctions at high latitudes contributed to the latitudinal diversity gradient*. Evolution 74(9):1966–1987.
Mitchell J.S., Etienne R.S., Rabosky D.L. 2019. Inferring diversification rate variation from phylogenies with fossils. Syst. Biol. 68(1):1–18.
Morlon H., Hartig F., Robin S. 2020. Prior hypotheses or regularization allow inference of diversification histories from extant timetrees. bioRxiv, page 2020.07.03.185074.
Morlon H., Lewitus E., Condamine F.L., Manceau M., Clavel J., Drury J. 2016. RPANDA: an R package for macroevolutionary analyses on phylogenetic trees. Methods Ecol. Evol. 7(5):589–597.
Morlon H., Parsons T.L., Plotkin J.B. 2011. Reconciling molecular phylogenies with the fossil record. Proc. Natl. Acad. Sci. USA 108(39):16327–16332.
Morlon H., Potts M.D., Plotkin J.B. 2010. Inferring the dynamics of diversification: a coalescent approach. PLoS Biol. 8(9):e1000493.
Nee S. 2006. Birth-death models in macroevolution. Annu. Rev. Ecol. Evol. Syst. 37(1):1–17.
Nee S., May R.M., Harvey P.H. 1994. The reconstructed evolutionary process. Philos. Trans. 344(1309):305–311.
Nee S., Mooers A.O., Harvey P.H. 1992. Tempo and mode of evolution revealed from molecular phylogenies. Proc. Natl. Acad. Sci. USA 89(17):8322–8326.
Ng J. and Smith, S. D. (2014). How traits shape trees: new approaches for detecting character state-dependent lineage diversification. J. Evol. Biol. 27(10):2035–2045.
O'Meara B.C., Beaulieu J.M. 2016. Past, future, and present of state-dependent models of diversification. Am. J. Bot. 103(5):792–795.
O'Meara B.C., Smith S.D., Armbruster W.S., Harder L.D., Hardy C.R., Hileman L.C., Hufford L., Litt A., Magallón. S., Smith S.A., Stevens P.F., Fenster C.B., Diggle P.K. 2016. Non-equilibrium dynamics and floral trait interactions shape extant angiosperm diversity. Proc. R. Soc. B 283(1830):20152304.
Onstein R.E. 2020. Darwin's second 'abominable mystery': Trait flexibility as the innovation leading to angiosperm diversity. New Phytol. 228(6):1741–1747.
Pagel M. 2020. Evolutionary trees can't reveal speciation and extinction rates. Nature 580(7804):461–462.
Parham J.F., Donoghue P.C.J., Bell C.J., Calway T.D., Head J.J., Holroyd P.A., Inoue J.G., Irmis R.B., Joyce W.G., Ksepka D.T., Patan© J.S.L., Smith N.D., Tarver J.E., van Tuinen M., Yang Z., Angielczyk K.D., Greenwood J.M., Hipsley C.A., Jacobs L., Makovicky P.J., MÃ¼ller J., Smith K.T., Theodor J.M., Warnock R.C.M., Benton M.J. 2012. Best practices for justifying fossil calibrations. Syst. Biol. 61(2):346–359.
Phillimore A.B., Price T.D. 2008. Density-dependent cladogenesis in birds. PLoS Biol. 6(3):e71.
Ponciano J.M., Burleigh J.G., Braun E.L., Taper M.L. 2012. Assessing parameter identifiability in phylogenetic models using data cloning. Syst. Biol. 61(6):955–972.
Pulido-Santacruz P., Weir J.T. 2016. Extinction as a driver of avian latitudinal diversity gradients: extinction as a driver of diversity gradients. Evolution 70(4):860–872.
Quintero I., Jetz W. (2018). Global elevational diversity and diversification of birds. Nature 555(7695):246–250.
Rabosky D.L. 2006. Likelihood methods for detecting temporal shifts in diversification rates. Evolution 60(6):1152–1164.
Rabosky D.L. 2014. Automatic detection of key innovations, rate shifts, and diversity-dependence on phylogenetic trees. PLoS One 9(2):e89543.
Rabosky D.L. 2017. Phylogenetic tests for evolutionary innovation: the problematic link between key innovations and exceptional diversification. Philos. Trans. R. Soc. B 372(1735):20160417.
Rabosky D.L. 2020. Speciation rate and the diversity of fishes in freshwaters and the oceans. J. Biogeogr. 47(6):1207–1217.
Rabosky D.L., Chang J., Title P.O., Cowman P.F., Sallan L., Friedman M., Kaschner K., Garilao C., Near T.J., Coll M., Alfaro M.E. 2018. An inverse latitudinal gradient in speciation rate for marine fishes. Nature 559(7714):392–395.
Rabosky D.L., Goldberg E.E. 2015. Model inadequacy and mistaken inferences of trait-dependent speciation. Syst. Biol. 64(2):340–355.
Rabosky D.L., Lovette I.J. 2008. Explosive evolutionary radiations: decreasing speciation or increasing extinction through time? Evolution 62(8):1866–1875.
Rabosky D.L., Santini F., Eastman J., Smith S.A., Sidlauskas B., Chang J., Alfaro M.E. 2013. Rates of speciation and morphological evolution are correlated across the largest vertebrate radiation. Nat. Commun. 4(1):1958.
Rannala B. 2002. Identifiability of parameters in MCMC Bayesian inference of phylogeny. Syst. Biol. 51(5):754–760.
Ricklefs R.E. 2007. Estimating diversification rates from phylogenetic information. Trends Ecol. Evol. 22(11):601–610.
Rolland J., Condamine F.L., Jiguet F., Morlon H. (2014). Faster speciation and reduced extinction in the tropics contribute to the mammalian latitudinal diversity gradient. PLoS Biol. 12(1):e1001775.
Ronquist F., Klopfstein S., Vilhelmsen L., Schulmeister S., Murray D.L., Rasnitsyn A.P. 2012. A total-evidence approach to dating with fossils, applied to the early radiation of the Hymenoptera. Syst. Biol. 61(6):973–999.
Sauquet H. 2013. A practical guide to molecular dating. Comptes Rendus Palevol 12(6):355–367.
Silvestro D., Warnock R.C.M., Gavryushkina A., Stadler T. 2018. Closing the gap between palaeontological and neontological speciation and extinction rate estimates. Nat. Commun. 9(1):5237.
Silvestro D., Zizka G., Schulte K. 2014. Disentangling the effects of key innovations on the diversification of Bromelioideae (bromeliaceae).

Downloaded from https://academic.oup.com/sysbio/article/71/3/758/6382322 by Ecole Francaise D'Athenes user on 23 April 2022

Evolution 68(1):163–175.
Slater G.J., Price S.A., Santini F., Alfaro M.E. 2010. Diversity versus disparity and the radiation of modern cetaceans. Proc. R. Soc. B 277(1697):3097–3104.
Smith S.A., Brown J.W. 2018. Constructing a broadly inclusive seed plant phylogeny. Am. J. Bot. 105(3):302–314.
Soltis P.S., Folk R.A., Soltis D.E. 2019. Darwin review: angiosperm phylogeny and evolutionary radiations. Proc. R. Soc. B 286(1899):20190099.
Stadler T. 2013. How can we improve accuracy of macroevolutionary rate estimates? Syst. Biol. 62(2):321–329.
Stein R.W., Mull C.G., Kuhn T.S., Aschliman N.C., Davidson L.N.K., Joy J.B., Smith G.J., Dulvy N.K., Mooers A.O. 2018. Global priorities for conserving the evolutionary history of sharks, rays and chimaeras. Nat. Ecol. Evol. 2(2):288–298.
Upham N.S., Esselstyn J.A., Jetz W. 2019. Inferring the mammal tree: species-level sets of phylogenies for questions in ecology, evolution, and conservation. PLoS Biol. 17(12):e3000494.
Vamosi J.C., Magallón S., Mayrose I., Otto S.P., Sauquet H. 2018. Macroevolutionary patterns of flowering plant speciation and extinction. Annu. Rev. Plant Biol. 69(1):685–706.
Whitlock M.C., Barton N.H. 1997. The effective size of a subdivided population. Genetics 146(1):427–441.
Zenil-Ferguson R., Burleigh J.G., Freyman W.A., Igić B., Mayrose I., Goldberg E.E. 2019. Interaction among ploidy, breeding system and lineage diversification. New Phytol. 224(3):1252–1265.

Downloaded from https://academic.oup.com/sysbio/article/71/3/758/6382322 by Ecole Francaise D'Athenes user on 23 April 2022

Appendix II

Molecular Phylogenetics and Evolution 183 (2023) 107758

Contents lists available at ScienceDirect

Molecular Phylogenetics and Evolution

journal homepage: www.elsevier.com/locate/ympev

A comprehensive phylogeny and revised taxonomy illuminate the origin and diversification of the global radiation of *Papilio* (Lepidoptera: Papilionidae)

Fabien L. Condamine [a,*], Rémi Allio [a,b], Eliette L. Reboud [a], Julian R. Dupuis [c], Emmanuel F. A. Toussaint [d], Nathan Mazet [a], Shao-Ji Hu [e,f], Delano S. Lewis [g], Krushnamegh Kunte [h,1], Adam M. Cotton [i,1], Felix A.H. Sperling [j,1]

[a] *Institut des Sciences de l'Evolution de Montpellier (Université de Montpellier EPHE), Place Eugène Bataillon, 34095 Montpellier, France*
[b] *Centre de Biologie pour la Gestion des Populations, INRAE, CIRAD, IRD, Montpellier SupAgro, Université de Montpellier, Montpellier, France*
[c] *Department of Entomology, University of Kentucky, S-225 Ag Science Center North, 1100 South Limestone, Lexington, KY 40546, USA*
[d] *Natural History Museum of Geneva, CP 6434, 1211, Geneva 6, Switzerland*
[e] *Yunnan Key Laboratory of International Rivers and Transboundary Eco-security, Yunnan University, Kunming 650500, China*
[f] *Institute of International Rivers and Eco-security, Yunnan University, Kunming 650500, China*
[g] *Burman University, 6730 University Drive, Lacombe T4L 2E5, AB, Canada*
[h] *National Centre for Biological Sciences, Tata Institute of Fundamental Research, GKVK Campus, Bellary Road, Bangalore 560065, India*
[i] *86/2 Moo 5, Tambon Nong Kwai, Hang Dong, Chiang Mai, Thailand*
[j] *University of Alberta, Department of Biological Sciences, Edmonton T6G 2E9, AB, Canada*

ARTICLE INFO

Keywords:
Butterflies
Historical biogeography
Macroevolution
Paleotropics
Systematics
Taxonomy

ABSTRACT

The swallowtail genus *Papilio* (Lepidoptera: Papilionidae) is species rich, distributed worldwide, and has broad morphological habits and ecological niches. Because of its elevated species richness, it has been historically difficult to reconstruct a densely sampled phylogeny for this clade. Here we provide a taxonomic working list for the genus, resulting in 235 *Papilio* species, and assemble a molecular dataset of seven gene fragments representing ca. 80% of the currently described diversity. Phylogenetic analyses reconstructed a robust tree with highly supported relationships within subgenera, although a few nodes in the early history of the Old World *Papilio* remain unresolved. Contrasting with previous results, we found that *Papilio alexanor* is sister to all Old World *Papilio* and that the subgenus *Eleppone* is no longer monotypic. The latter includes the recently described Fijian *Papilio natewa* with the Australian *Papilio anactus* and is sister to subgenus *Araminta* (formerly included in subgenus *Menelaides*) occurring in Southeast Asia. Our phylogeny also includes rarely studied (*P. antimachus*, *P. benguetana*) or endangered species (*P. buddha*, *P. chikae*). Taxonomic changes resulting from this study are elucidated. Molecular dating and biogeographic analyses indicate that *Papilio* originated ca. 30 million years ago (Oligocene), in a northern region centered on Beringia. A rapid early Miocene radiation in the Paleotropics is revealed within Old World *Papilio*, potentially explaining their low early branch support. Most subgenera originated in the early to middle Miocene followed by synchronous southward biogeographic dispersals and repeated local extirpations in northern latitudes. This study provides a comprehensive phylogenetic framework for *Papilio* with clarification of subgeneric systematics and species taxonomic changes enumerated, which will facilitate further studies to address questions on their ecology and evolutionary biology using this model clade.

1. Introduction

Insects are indisputably the most diverse terrestrial eukaryotic clade, with over a million insect species described and many more awaiting discovery or formal description (Stork, 2018). This staggering species diversity is unevenly distributed across the insect tree of life, with many clades comprising substantially more species than their sister clades (Grimaldi and Engel, 2005; Engel, 2015). For instance, within ants (Formicidae), some genera such as *Camponotus* and *Pheidole* have over 1,500 and 1,100 species, respectively (Hölldobler and Wilson, 1990),

* Corresponding author.
E-mail address: fabien.condamine@gmail.com (F.L. Condamine).
[1] Co-senior authors.

https://doi.org/10.1016/j.ympev.2023.107758
Received 28 November 2022; Received in revised form 28 February 2023; Accepted 8 March 2023
Available online 11 March 2023
1055-7903/© 2023 Elsevier Inc. All rights reserved.

while their corresponding sister clades have less than 700 and 200 species, respectively (Economo et al., 2018). Within flies (Diptera), the most well-known case is the genus *Drosophila,* which has more than 1600 species (Brake and Bächli, 2013; O'Grady and DeSalle, 2018), which is sister to a clade composed of seven genera including about 600 species (Wiegmann et al., 2011; O'Grady and DeSalle, 2018). Such genera are difficult to resolve phylogenetically and manage taxonomically. The genus rank is widely used in molecular, morphological and paleontological systematics. However, species-rich genera that are difficult to delimit can also be treated using the subgenus rank, rendering their relationships less complicated while allowing flexibility to taxonomic ranks that are less frequently employed (Winston, 1999; Teta, 2019).

In Lepidoptera, swallowtail butterflies (Papilionidae) of the genus *Papilio* Linnaeus, 1758 include more than 200 species and represent more than one third of all Papilionidae (Fig. 1), which has about 600 recognized species (Zakharov et al., 2004a; Haüser et al., 2005; Condamine et al., 2012; Nakae, 2021a). As one of the most well-known and broadly studied groups of insects, *Papilio* swallowtails are recognized as model organisms in evolutionary biology, ecology, genomics, and

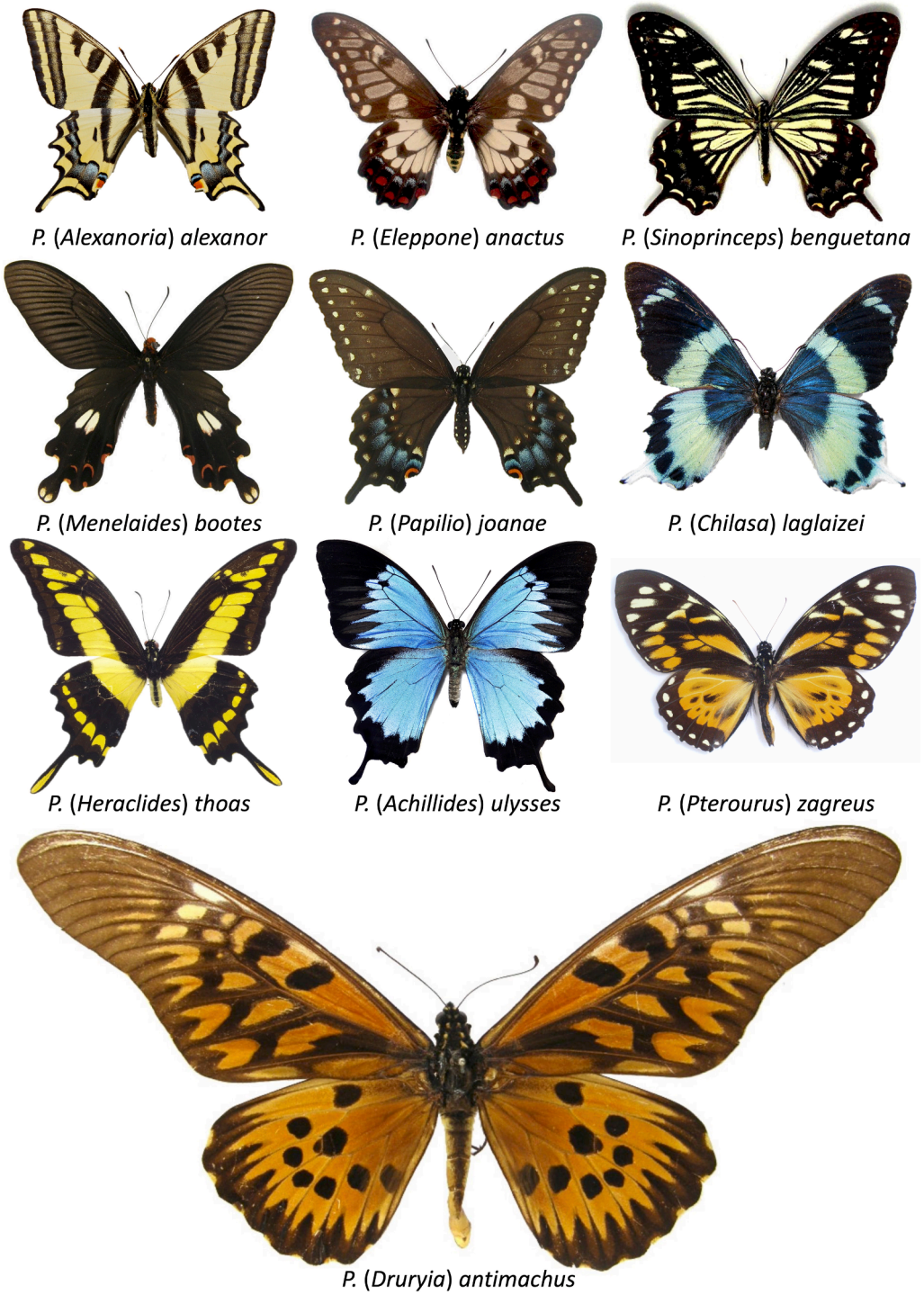


Fig. 1. Illustration of swallowtail butterfly diversity in the genus *Papilio,* showing several subgenera that have been widely used since Munroe (1961). Specimens are not uniformly scaled. Pictures from Fabien L. Condamine.

conservation biology (e.g. Collins and Morris, 1985; Scriber et al., 1995; Kunte, 2009; Kunte et al., 2014; Dupuis and Sperling, 2015). Yet the phylogeny of *Papilio* is far from being resolved despite numerous studies (Ae, 1979; Hancock, 1983; Igarashi, 1984; Miller, 1987; Tyler et al., 1994; Scriber et al., 1995; Aubert et al., 1999; Caterino and Sperling, 1999; Reed and Sperling, 1999; Yagi et al., 1999; Caterino et al., 2001; Zakharov et al., 2004a; Condamine et al., 2013a; Lewis et al., 2015; Wu et al., 2015; Owens et al., 2017, 2020). Ecological and evolutionary hypotheses on their diversification rely on a comprehensive and strongly supported phylogeny as well as accurate age estimates for significant phylogenetic events like host-plant shifts or mimicry evolution.

The first classifications of *Papilio* species relied on morphological characters (summarized in Zakharov et al., 2004a: table 1). Munroe (1961) divided *Papilio* into five sections but did not designate them as subgenera because they lacked a simple diagnosis with adult characters. Using an explicitly cladistic estimation of relationships within *Papilio*, Hancock (1983) recognized six genera (*Chilasa*, *Eleppone*, *Heraclides*, *Papilio*, *Princeps*, and *Pterourus*) based on phylogenetic evidence and inferred evolutionary antiquity, but this phylogeny also suffered from lack of character justification. Another classification was proposed by Igarashi (1984) based on the morphology of immature stages, but this work did not represent all of Hancock's genera. Igarashi (1984) nonetheless recognized seven genera (*Achillides*, *Agehana*, *Chilasa*, *Euchenor*, *Menelaides*, *Papilio*, and *Pterourus*), with numerous discrepancies between his treatment and that of Hancock (1983). Hancock's classification was criticized by Miller (1987), who did not consider elevation of *Papilio* subdivisions to the genus level to be justified. However, one group within *Papilio* was elevated in a widely available checklist of swallowtail butterflies (Haüser et al., 2005), where *Chilasa* was treated as a distinct genus. It has been challenging to apply the rules of the PhyloCode in designating subgroups within *Papilio* to generic status as it has proven difficult to find morphological synapomorphies as required by the PhyloCode (Cantino and de Queiroz, 2020). The sole use of genetic data to delimit taxa has not been adequately addressed by the PhyloCode.

Due to the limitations of traditional morphological approaches and with the development of new molecular systematic approaches, the classification of *Papilio* has received significant attention in the last three decades. Relationships among species within the *P. machaon* and *P. glaucus-troilus* species groups were studied based on allozyme variation (Sperling, 1987; Hagen and Scriber, 1991). Restriction fragment length polymorphism of mitochondrial DNA was used to compare taxa within the same species groups in later studies (Sperling, 1991, 1993a, 1993b; Sperling and Harrison, 1994, Tyler et al., 1994). Phylogenetic relationships within *Papilio* have also been analyzed using DNA sequences of a variety of genes, but these studies have been confined to single species groups or local geographic areas (e.g. Vane-Wright et al., 1999; Yagi et al., 1999; Zakharov et al., 2004b; Condamine *et al.*, 2013b; Lewis et al., 2015; Owens et al., 2017, 2020; Dupuis and Sperling, 2020; Joshi and Kunte, 2022) or have included limited sampling across *Papilio* subdivisions (e.g. Aubert et al., 1999; Caterino and Sperling, 1999; Reed and Sperling, 1999; Zakharov et al., 2004a). Although the two latest molecular phylogenetic studies of Papilionidae (Condamine et al., 2012; Allio et al., 2021) substantially increased the fraction of sampled species, they did not provide a sufficient resolution to assess the fine-scale taxonomic delimitation within the genus. Yet, a number of systematic enigmas remain in *Papilio,* starting with the number of valid species to consider in the genus.

Here we establish a list of valid *Papilio* species based on previous molecular studies and build a reliable and comprehensive time-calibrated species-level phylogeny for the genus *Papilio* using seven gene fragments for ca. 80% of the total species diversity. Our study aims at establishing a reference phylogenetic framework to evaluate both subgeneric monophyly and species relationships within *Papilio*, for the enigmatic and long-debated placement of several groups and species. The phylogeny includes species that are key for testing biogeographic hypotheses (e.g. *P. anactus*, *P. benguetana*), evolution of mimicry (e.g. *Chilasa*, *P. nobilis*), and host-plant associations (e.g. *P. alexanor*), as well as species that are rare and insufficiently studied (e.g. *P. antimachus*, *P. himeros*), endangered (*P. buddha*, *P. chikae*) or recently described (e.g. the Fijian *P. natewa*). The recent discovery of *P. natewa* (Tennent et al., 2018) may clarify the systematic position of *P. anactus*, presently placed in a monotypic subgenus (*Eleppone*) but with an unstable phylogenetic position (Hancock, 1979; Zakharov et al., 2004a; Condamine et al., 2012). Indeed, external morphology and genitalia of *P. natewa* suggest a phylogenetic affinity with *P. anactus* (Tennent et al., 2018), but this remains to be tested. Without knowledge of its larva, the enigmatic African giant swallowtail, *P. antimachus*, has often been placed in species groups of two distinct *Papilio* subgenera (*Druryia* or *Princeps*), and the subgenus *Druryia* has never been studied with a molecular approach. The position of the *nobilis* species-group also remains poorly resolved, wavering between the iconic *phorcas* and *hesperus* species-groups (Munroe, 1961, and Hancock, 1983, respectively). Yet, its placement has important implications for the understanding of mimicry evolution in the *phorcas* group (Vane-Wright et al., 1999) and evolution of iridescent wings (common origin or convergence with the subgenus *Achillides*). Including the giant blue swallowtail, *P. zalmoxis*, in the phylogeny could help solve this puzzle. Moreover, the monophyly and rank of the danaine- and moth-mimicking *Chilasa* are uncertain. Munroe (1961) split its members among two *Papilio* subunits, but Hancock (1983) placed them together in a single genus, considered to be the sister taxon of *Eleppone*. Finally, we also revisited the placement of *P. alexanor*. The relationships of this odd European Apiaceae feeder have been examined several times (Aubert et al., 1999; Caterino and Sperling, 1999; Reed and Sperling, 1999; Zakharov et al., 2004a), nearly reaching a consensus that *P. alexanor* is sister to a clade comprising *Pterourus* and *Chilasa* (Zakharov et al., 2004a; Condamine et al., 2012). However, no strong resolution has been obtained and this position has been questioned with phylogenomic analyses (Allio et al., 2020), albeit with less-than-ideal sampling density. We not only resolve these systematic conflicts and uncertainties, but also provide a phylogenetic framework to infer an evolutionary timescale for *Papilio*, estimate historical biogeography, and investigate heterogeneity of diversification dynamics.

2. Materials and methods

2.1. Taxon sampling and molecular data

We first established a taxonomic working list of *Papilio* species (Table 1). This species list combines several previous studies (Zakharov et al., 2004a, 2004b; Condamine et al., 2012, 2013a, 2013b; Shiraiwa et al., 2014; Lewis et al., 2015; Wu et al., 2015; Owens et al., 2017, 2020; Allio et al., 2021; Joshi and Kunte, 2022). Like all other working lists (Garnett et al., 2020), we acknowledge that this species list remains provisional and is likely to evolve through time (i.e. species are hypotheses, Pante et al., 2015). Indeed, *Papilio* butterflies will inevitably continue to receive attention as a model clade, and we expect many of the species' boundaries to be tested with new genomic data and evolutionary models in further studies, which are very likely to eventually modify the taxonomic list (e.g. Kunte et al., 2011; Dupuis and Sperling, 2020). From the 235 species that are currently recognized in the new taxonomic list presented here for the genus *Papilio*, we sampled 184 species representing 78.3% of the total diversity (Table 1). However, the total number of species will evolve with further systematic studies, especially of the African clades that are currently poorly represented in our analysis (Fig. 2). We added 18 outgroup species representing several swallowtail genera to root the *Papilio* tree and provide deeper relationships for secondary calibration of nodes based on previous dated studies (Condamine et al., 2012; Allio et al., 2021). The outgroups included: (1) two species of the genus *Meandrusa* (*M. payeni* and *M. sciron*) to represent the sister genus of *Papilio* and define the crown of tribe Papilionini, (2) five species of the tribe Troidini (*Battus philenor,*

Table 1
Taxonomic working list for *Papilio* species. The list includes subgenera that are currently recovered as monophyletic in molecular phylogenies, and species that belong to each subgenus. We also indicate the ratio of species sampled in the phylogeny presented in this study, which highlights subgenera that are well sampled versus poorly sampled. In total, genus *Papilio* Linnaeus, 1758 includes 235 species, and we sampled 184 species in this study (=78.3%).

***Achillides* Hübner, [1819] [30 spp.]: 29 spp. sampled in the current study (=96.6%)**	**Status**	**In the tree**
Papilio (*Achillides*) *arcturus* Westwood, 1842	Valid	Sampled
Papilio (*Achillides*) *arjuna* Horsfield, 1828	Valid	Sampled
Papilio (*Achillides*) *bianor* Cramer, 1777	Valid	Sampled
Papilio (*Achillides*) *blumei* Boisduval, 1836	Valid	Sampled
Papilio (*Achillides*) *buddha* Westwood, 1872	Valid	Sampled
Papilio (*Achillides*) *chikae* Igarashi, 1965	Valid	Sampled
Papilio (*Achillides*) *crino* Fabricius, 1793	Valid	Sampled
Papilio (*Achillides*) *daedalus* Felder & Felder, 1861	**stat. rev.** (formerly a subspecies of *palinurus*)	Sampled
Papilio (*Achillides*) *dehaanii* Felder & Felder, 1864	Valid	Sampled
Papilio (*Achillides*) *dialis* Leech, 1893	Valid	Sampled
Papilio (*Achillides*) *elephenor* Doubleday, 1845	Valid	**Unsampled**
Papilio (*Achillides*) *hermeli* Nuyda, 1992	**stat. rev.** (formerly a subspecies of *chikae*)	Sampled
Papilio (*Achillides*) *hermosanus* Rebel, 1906	Valid	Sampled
Papilio (*Achillides*) *hopponis* Matsumura, 1907	Valid	Sampled
Papilio (*Achillides*) *karna* Felder & Felder, 1865	Valid	Sampled
Papilio (*Achillides*) *krishna* Moore, [1858]	Valid	Sampled
Papilio (*Achillides*) *lorquinianus* Felder & Felder, 1865	Valid	Sampled
Papilio (*Achillides*) *maackii* Ménétriés, 1858	Valid	Sampled
Papilio (*Achillides*) *montrouzieri* Boisduval, 1859	Valid	Sampled
Papilio (*Achillides*) *neumoegeni* Honrath, 1890	Valid	Sampled
Papilio (*Achillides*) *orsippus* Godman & Salvin, 1888	**stat. rev.** (formerly a subspecies of *ulysses*)	Sampled
Papilio (*Achillides*) *palinurus* Fabricius, 1787	Valid	Sampled
Papilio (*Achillides*) *paris* Linnaeus, 1758	Valid	Sampled
Papilio (*Achillides*) *peranthus* Fabricius, 1787	Valid	Sampled
Papilio (*Achillides*) *pericles* Wallace, 1865	Valid	Sampled
Papilio (*Achillides*) *polyctor* Boisduval, 1836	**stat. rev.** (formerly a subspecies of *bianor*)	Sampled
Papilio (*Achillides*) *ryukyuensis* Fujioka, 1975	Valid	Sampled
Papilio (*Achillides*) *syfanius* Oberthür, 1886	Valid	Sampled
Papilio (*Achillides*) *telegonus* Felder & Felder, 1860	**stat. rev.** (formerly a subspecies of *ulysses*)	Sampled
Papilio (*Achillides*) *ulysses* Linnaeus, 1758	Valid	Sampled
***Alexanoria* Koçak and Kemal, 2002 [1 sp.]: 1 sampled in the current study (=100%)**	**Status**	**In the tree**
Papilio (*Alexanoria*) *alexanor* Esper, [1800]	Valid	Sampled
***Araminta* Moore, 1886 [5 spp.]: 5 sampled in the current study (=100%)**	**Status**	**In the tree**
Papilio (*Araminta*) *antonio* Hewitson, 1875	Valid	Sampled
Papilio (*Araminta*) *demolion* Cramer, 1776	Valid	Sampled
Papilio (*Araminta*) *gigon* Felder & Felder, 1865	Valid	Sampled
Papilio (*Araminta*) *liomedon* Moore, [1875]	Valid	Sampled
Papilio (*Araminta*) *noblei* Nicéville, [1889]	Valid	Sampled

Table 1 (*continued*)

***Achillides* Hübner, [1819] [30 spp.]: 29 spp. sampled in the current study (=96.6%)**	**Status**	**In the tree**
***Chilasa* Moore, [1881] [11 spp.]: 8 sampled in the current study (=72.7%)**	**Status**	**In the tree**
Papilio (*Chilasa*) *agestor* Gray, 1831	Valid	Sampled
Papilio (*Chilasa*) *carolinensis* (Jumalon, 1967)	Valid	**Unsampled**
Papilio (*Chilasa*) *clytia* Linnaeus, 1758	Valid	Sampled
Papilio (*Chilasa*) *epycides* Hewitson, 1864	Valid	Sampled
Papilio (*Chilasa*) *laglaizei* Depuiset, 1877	Valid	Sampled
Papilio (*Chilasa*) *moerneri* Aurivillius, 1919	Valid	**Unsampled**
Papilio (*Chilasa*) *osmana* (Jumalon, 1967)	Valid	Sampled
Papilio (*Chilasa*) *paradoxa* Zincken, 1831	Valid	Sampled
Papilio (*Chilasa*) *slateri* Hewitson, 1859	Valid	Sampled
Papilio (*Chilasa*) *toboroi* Ribbe, 1907	Valid	**Unsampled**
Papilio (*Chilasa*) *veiovis* Hewitson, 1865	Valid	Sampled
***Druryia* Aurivillius, 1881 [30 spp.]: 9 sampled in the current study (=30%)**	**Status**	**In the tree**
Papilio (*Druryia*) *andronicus* Ward, 1871	Valid	**Unsampled**
Papilio (*Druryia*) *antimachus* Drury, [1782]	Valid	Sampled
Papilio (*Druryia*) *arnoldiana* Vane-Wright, 1995	Valid	**Unsampled**
Papilio (*Druryia*) *bacelarae* Bivar de Sousa & Mendes, 2009	Valid	**Unsampled**
Papilio (*Druryia*) *constantinus* Ward, 1871	Valid	Sampled
Papilio (*Druryia*) *cynorta* Fabricius, 1793	Valid	**Unsampled**
Papilio (*Druryia*) *cyproeofila* Butler, 1868	Valid	**Unsampled**
Papilio (*Druryia*) *dardanus* Brown, 1776	Valid	Sampled
Papilio (*Druryia*) *delalandei* Godart, 1823	Valid	Sampled
Papilio (*Druryia*) *echerioides* Trimen, 1868	Valid	**Unsampled**
Papilio (*Druryia*) *fernandus* Fruhstorfer, 1903	Valid	**Unsampled**
Papilio (*Druryia*) *filaprae* Süffert, 1904	Valid	**Unsampled**
Papilio (*Druryia*) *fuelleborni* Karsch, 1900	Valid	**Unsampled**
Papilio (*Druryia*) *gallienus* Distant, 1879	Valid	**Unsampled**
Papilio (*Druryia*) *humbloti* Oberthür, 1888	**stat. rev.** (formerly a subspecies of *dardanus*)	Sampled
Papilio (*Druryia*) *jacksoni* Sharpe, 1891	Valid	**Unsampled**
Papilio (*Druryia*) *leucotaenia* Rothschild, 1908	Valid	**Unsampled**
Papilio (*Druryia*) *lormieri* Distant, 1874	Valid	Sampled
Papilio (*Druryia*) *mangoura* Hewitson, 1875	Valid	**Unsampled**
Papilio (*Druryia*) *mechowi* Dewitz, 1881	Valid	**Unsampled**
Papilio (*Druryia*) *mechowianus* Dewitz, 1885	Valid	**Unsampled**
Papilio (*Druryia*) *menestheus* Drury, [1773]	Valid	**Unsampled**
Papilio (*Druryia*) *meriones* Felder & Felder, 1864	**stat. rev.** (formerly a subspecies of *dardanus*)	Sampled
Papilio (*Druryia*) *nobicea* Suffert, 1904	Valid	**Unsampled**
Papilio (*Druryia*) *ophidicephalus* Oberthür, 1878	Valid	**Unsampled**
Papilio (*Druryia*) *phorcas* Cramer, 1775	Valid	Sampled
Papilio (*Druryia*) *plagiatus* Aurivillius, 1898	Valid	**Unsampled**
Papilio (*Druryia*) *rex* Oberthür, 1886	Valid	Sampled
Papilio (*Druryia*) *sjoestedti* Aurivillius, 1908	Valid	**Unsampled**
Papilio (*Druryia*) *zenobia* Fabricius, 1775	Valid	**Unsampled**
***Eleppone* Hancock, 1979 [2 spp.]: 2 sampled in the current study (=100%)**	**Status**	**In the tree**
Papilio (*Eleppone*) *anactus* Macleay, 1826	Valid	Sampled
Papilio (*Eleppone*) *natewa* Tennent, Chandra & Müller, 2018	Valid	Sampled
***Euchenor* Igarashi, 1979 [2 spp.]: 2 sampled in the current study (=100%)**	**Status**	**In the tree**
Papilio (*Euchenor*) *depilis* Rothschild, 1895	**stat. rev.** (formerly a subspecies of *euchenor*)	Sampled
Papilio (*Euchenor*) *euchenor* Guérin-Méneville, 1830	Valid	Sampled

(*continued on next page*)

Table 1 (*continued*)

***Achillides* Hübner, [1819] [30 spp.]: 29 spp. sampled in the current study (=96.6%)**	**Status**	**In the tree**
***Heraclides* Hübner, [1819] [32 spp.]: 32 sampled in the current study (=100%)**	**Status**	**In the tree**
Papilio (*Heraclides*) *anchicayaensis* Constantino, Le Crom & Salazar, 2002	Valid	Sampled
Papilio (*Heraclides*) *anchisiades* Esper, [1788]	Valid	Sampled
Papilio (*Heraclides*) *andraemon* Hübner, [1823]	Valid	Sampled
Papilio (*Heraclides*) *androgeus* Cramer, 1775	Valid	Sampled
Papilio (*Heraclides*) *aristodemus* Esper, 1794	Valid	Sampled
Papilio (*Heraclides*) *aristor* Godart, 1819	Valid	Sampled
Papilio (*Heraclides*) *astyalus* Godart, 1819	Valid	Sampled
Papilio (*Heraclides*) *caiguanabus* Poey, 1852	Valid	Sampled
Papilio (*Heraclides*) *chiansiades* Westwood, 1872	Valid	Sampled
Papilio (*Heraclides*) *cresphontes* Cramer, 1777	Valid	Sampled
Papilio (*Heraclides*) *epenetus* Hewitson, 1861	Valid	Sampled
Papilio (*Heraclides*) *erostratus* Westwood, 1847	Valid	Sampled
Papilio (*Heraclides*) *garleppi* Staudinger, 1892	Valid	Sampled
Papilio (*Heraclides*) *hectorides* Esper, 1794	Valid	Sampled
Papilio (*Heraclides*) *himeros* Hopffer, 1865	Valid	Sampled
Papilio (*Heraclides*) *homothoas* Rothschild & Jordan, 1906	Valid	Sampled
Papilio (*Heraclides*) *hyppason* Cramer, 1775	Valid	Sampled
Papilio (*Heraclides*) *isidorus* Doubleday, 1846	Valid	Sampled
Papilio (*Heraclides*) *lamarchei* Staudinger, 1892	Valid	Sampled
Papilio (*Heraclides*) *machaonides* Esper, 1796	Valid	Sampled
Papilio (*Heraclides*) *melonius* Rothschild & Jordan, 1906	Valid	Sampled
Papilio (*Heraclides*) *ornythion* Boisduval, 1836	Valid	Sampled
Papilio (*Heraclides*) *oviedo* Gundlach, 1866	Valid	Sampled
Papilio (*Heraclides*) *oxynius* (Geyer, 1827)	Valid	Sampled
Papilio (*Heraclides*) *paeon* Boisduval, 1836	Valid	Sampled
Papilio (*Heraclides*) *pallas* Gray, [1853]	Valid	Sampled
Papilio (*Heraclides*) *pelaus* Fabricius, 1775	Valid	Sampled
Papilio (*Heraclides*) *rogeri* Boisduval, 1836	Valid	Sampled
Papilio (*Heraclides*) *rumiko* (Shiraiwa & Grishin, 2014)	Valid	Sampled
Papilio (*Heraclides*) *thersites* Fabricius, 1775	Valid	Sampled
Papilio (*Heraclides*) *thoas* Linnaeus, 1771	Valid	Sampled
Papilio (*Heraclides*) *torquatus* Cramer, 1777	Valid	Sampled
***Menelaides* Hübner, [1819] [54 spp.]: 52 sampled in the current study (=96.5%)**	**Status**	**In the tree**
Papilio (*Menelaides*) *acheron* Grose-Smith, 1887	Valid	Sampled
Papilio (*Menelaides*) *aegeus* Donovan, 1805	Valid	Sampled
Papilio (*Menelaides*) *agenor* Linnaeus, 1758	**stat. rev.** (formerly a subspecies of *memnon*)	Sampled
Papilio (*Menelaides*) *albinus* Wallace, 1865	Valid	Sampled
Papilio (*Menelaides*) *alcmenor* Felder & Felder, 1865	Valid	Sampled
Papilio (*Menelaides*) *alphenor* Cramer, 1776	**stat. rev.** (formerly a subspecies of *polytes*)	Sampled
Papilio (*Menelaides*) *ambrax* Boisduval, 1832	**stat. rev.** (formerly a subspecies of *phestus*)	Sampled
Papilio (*Menelaides*) *amynthor* Boisduval, 1859	Valid	Sampled

Table 1 (*continued*)

***Achillides* Hübner, [1819] [30 spp.]: 29 spp. sampled in the current study (=96.6%)**	**Status**	**In the tree**
Papilio (*Menelaides*) *ascalaphus* Boisduval, 1836	Valid	Sampled
Papilio (*Menelaides*) *biseriatus* Rothschild, 1895	**stat. rev.** (formerly a subspecies of *helenus*)	Sampled
Papilio (*Menelaides*) *bootes* Westwood, 1842	Valid	Sampled
Papilio (*Menelaides*) *bridgei* Mathew, 1886	Valid	Sampled
Papilio (*Menelaides*) *canopus* Westwood, 1842	**stat. rev.** (formerly a subspecies of *fuscus*)	Sampled
Papilio (*Menelaides*) *capaneus* Westwood, 1843	**stat. rev.** (formerly a subspecies of *fuscus*)	Sampled
Papilio (*Menelaides*) *castor* Westwood, 1842	Valid	Sampled
Papilio (*Menelaides*) *chaon* Westwood, 1844	**stat. rev.** (formerly a subspecies of *nephelus*)	Sampled
Papilio (*Menelaides*) *cilix* Godman & Salvin, 1879	**stat. rev.** (formerly a subspecies of *fuscus*)	Sampled
Papilio (*Menelaides*) *daksha* Moore, [1889]	**stat. rev.** (formerly a subspecies of *helenus*)	Sampled
Papilio (*Menelaides*) *deiphobus* Linnaeus, 1758	Valid	Sampled
Papilio (*Menelaides*) *diophantus* Grose-Smith, 1883	Valid	Sampled
Papilio (*Menelaides*) *dravidarum* Wood-Mason, 1880	Valid	Sampled
Papilio (*Menelaides*) *erskinei* Mathew, 1886	Valid	**Unsampled**
Papilio (*Menelaides*) *forbesi* Grose-Smith, 1883	Valid	Sampled
Papilio (*Menelaides*) *fuscus* Goeze, 1779	Valid	Sampled
Papilio (*Menelaides*) *gambrisius* Cramer, 1777	Valid	Sampled
Papilio (*Menelaides*) *godeffroyi* Semper, 1866	Valid	Sampled
Papilio (*Menelaides*) *helenus* Linnaeus, 1758	Valid	Sampled
Papilio (*Menelaides*) *heringi* Niepelt, 1924	Valid	Sampled
Papilio (*Menelaides*) *hipponous* Felder & Felder, 1862	Valid	Sampled
Papilio (*Menelaides*) *hypsicles* Hewitson, 1868	**stat. rev.** (formerly a subspecies of *fuscus*)	Sampled
Papilio (*Menelaides*) *hystaspes* Felder & Felder, 1862	**stat. rev.** (formerly a subspecies of *helenus*)	Sampled
Papilio (*Menelaides*) *iswara* White, 1842	Valid	Sampled
Papilio (*Menelaides*) *iswaroides* Fruhstorfer, 1898	Valid	Sampled
Papilio (*Menelaides*) *janaka* Moore, 1857	**stat. rev.** (formerly a subspecies of *bootes*)	Sampled
Papilio (*Menelaides*) *javanus* Felder, 1862	**stat. nov.** (formerly a subspecies of *polytes*)	Sampled
Papilio (*Menelaides*) *jordani* Fruhstorfer, 1902	Valid	Sampled
Papilio (*Menelaides*) *lampsacus* Boisduval, 1836	Valid	**Unsampled**
Papilio (*Menelaides*) *macilentus* Janson, 1877	Valid	Sampled
Papilio (*Menelaides*) *mangarinus* Rothschild, 1908	**stat. nov.** (formerly a subspecies of *helenus*)	Sampled
Papilio (*Menelaides*) *memnon* Linnaeus, 1758	Valid	Sampled
Papilio (*Menelaides*) *nephelus* Boisduval, 1836	Valid	Sampled
Papilio (*Menelaides*) *oenomaus* Godart, 1819	Valid	Sampled
Papilio (*Menelaides*) *oritas* Godman & Salvin, 1879	**stat. rev.** (formerly a subspecies of *aegeus*)	Sampled
Papilio (*Menelaides*) *palawanicus* Staundinger 1888	**stat. nov.** (formerly a subspecies of *helenus*)	Sampled
Papilio (*Menelaides*) *phestus* Guérin-Méneville, 1830	Valid	Sampled
Papilio (*Menelaides*) *polytes* Linnaeus, 1758	Valid	Sampled
Papilio (*Menelaides*) *prexaspes* Felder & Felder, 1865	Valid	Sampled

(*continued on next page*)

Table 1 (*continued*)

***Achillides* Hübner, [1819] [30 spp.]: 29 spp. sampled in the current study (=96.6%)**	**Status**	**In the tree**
Papilio (*Menelaides*) *protenor* Cramer, 1775	Valid	Sampled
Papilio (*Menelaides*) *sataspes* Felder & Felder, 1865	Valid	Sampled
Papilio (*Menelaides*) *schmeltzi* Herrich-Schäffer, 1869	Valid	Sampled
Papilio (*Menelaides*) *thaiwanus* Rothschild, 1898	Valid	Sampled
Papilio (*Menelaides*) *tydeus* Felder & Felder, 1860	Valid	Sampled
Papilio (*Menelaides*) *weymeri* Niepelt, 1914	Valid	Sampled
Papilio (*Menelaides*) *woodfordi* Godman & Salvin, 1888	Valid	Sampled
Nireopapilio Cotton & Nakae, 2020 **[24 spp.]: 5 sampled in the current study (=20.8%)**	**Status**	**In the tree**
Papilio (*Nireopapilio*) *aristophontes* Oberthür, 1897	Valid	**Unsampled**
Papilio (*Nireopapilio*) *charopus* Westwood, 1843	Valid	**Unsampled**
Papilio (*Nireopapilio*) *chitondensis* Bivar de Sousa & Fernandes, 1966	Valid	**Unsampled**
Papilio (*Nireopapilio*) *chrapkowskii* Suffert, 1904	Valid	**Unsampled**
Papilio (*Nireopapilio*) *chrapkowskoides* Storace, 1952	Valid	**Unsampled**
Papilio (*Nireopapilio*) *desmondi* van Someren, 1939	Valid	**Unsampled**
Papilio (*Nireopapilio*) *epiphorbas* Boisduval, 1833	Valid	Sampled
Papilio (*Nireopapilio*) *euphranor* Trimen, 1868	Valid	**Unsampled**
Papilio (*Nireopapilio*) *hesperus* Westwood, 1843	Valid	**Unsampled**
Papilio (*Nireopapilio*) *hornimani* Distant, 1879	Valid	**Unsampled**
Papilio (*Nireopapilio*) *horribilis* Butler, [1872]	Valid	**Unsampled**
Papilio (*Nireopapilio*) *interjectana* Vane-Wright, 1995	Valid	**Unsampled**
Papilio (*Nireopapilio*) *mackinnoni* Sharpe, 1891	Valid	**Unsampled**
Papilio (*Nireopapilio*) *manlius* Fabricius, 1798	Valid	**Unsampled**
Papilio (*Nireopapilio*) *microps* Storace, 1951	Valid	**Unsampled**
Papilio (*Nireopapilio*) *nireus* Linnaeus, 1758	Valid	Sampled
Papilio (*Nireopapilio*) *nobilis* Rogenhofer, 1891	Valid	Sampled
Papilio (*Nireopapilio*) *oribazus* Boisduval, 1836	Valid	Sampled
Papilio (*Nireopapilio*) *pelodurus* Butler, [1896]	Valid	**Unsampled**
Papilio (*Nireopapilio*) *phorbanta* Linnaeus, 1771	Valid	**Unsampled**
Papilio (*Nireopapilio*) *sosia* Rothschild & Jordan, 1903	Valid	**Unsampled**
Papilio (*Nireopapilio*) *thuraui* Karsch, 1900	Valid	**Unsampled**
Papilio (*Nireopapilio*) *ufipa* Carcasson, 1961	Valid	**Unsampled**
Papilio (*Nireopapilio*) *zalmoxis* Hewitson, [1864]	Valid	Sampled
***Papilio* Linnaeus, 1758 [9 spp.]: 9 sampled in the current study (=100%)**	**Status**	**In the tree**
Papilio (*Papilio*) *bairdii* Edwards, 1866	**stat. rev.** (formerly a subspecies of *machaon*)	Sampled
Papilio (*Papilio*) *brevicauda* Saunders, 1868	Valid	Sampled
Papilio (*Papilio*) *hospiton* Géné, 1839	Valid	Sampled
Papilio (*Papilio*) *indra* Reakirt, 1866	Valid	Sampled
Papilio (*Papilio*) *joanae* Heitzman, 1974	Valid	Sampled

Table 1 (*continued*)

***Achillides* Hübner, [1819] [30 spp.]: 29 spp. sampled in the current study (=96.6%)**	**Status**	**In the tree**
Papilio (*Papilio*) *kahli* Chermock & Chermock, 1937	**stat. nov.** (formerly a subspecies of *machaon*)	Sampled
Papilio (*Papilio*) *machaon* Linnaeus, 1758	Valid	Sampled
Papilio (*Papilio*) *polyxenes* Fabricius, 1775	Valid	Sampled
Papilio (*Papilio*) *zelicaon* Lucas, 1852	Valid	Sampled
***Princeps* Hübner, [1807] [5 spp.]: 5 sampled in the current study (=100%)**	**Status**	**In the tree**
Papilio (*Princeps*) *demodocus* Esper, 1799	Valid	Sampled
Papilio (*Princeps*) *demoleus* Linnaeus, 1758	Valid	Sampled
Papilio (*Princeps*) *erithonioides* Grose-Smith, 1891	Valid	Sampled
Papilio (*Princeps*) *grosesmithi* Rothschild, 1926	Valid	Sampled
Papilio (*Princeps*) *morondavana* Grose-Smith, 1891	Valid	Sampled
***Pterourus* Scopoli, 1777 [28 spp.]: 23 sampled in the current study (=82.1%)**	**Status**	**In the tree**
Papilio (*Pterourus*) *alexiares* Hopffer, 1865	Valid	**Unsampled**
Papilio (*Pterourus*) *appalachiensis* (Pavulaan & Wright, 2002)	Valid	Sampled
Papilio (*Pterourus*) *ascolius* Felder & Felder, 1864	Valid	**Unsampled**
Papilio (*Pterourus*) *bachus* Felder & Felder, 1865	Valid	**Unsampled**
Papilio (*Pterourus*) *birchallii* Hewitson, 1863	Valid	Sampled
Papilio (*Pterourus*) *cacicus* Lucas, 1852	Valid	Sampled
Papilio (*Pterourus*) *canadensis* Rothschild & Jordan, 1906	Valid	Sampled
Papilio (*Pterourus*) *elwesi* Leech, 1889	Valid	Sampled
Papilio (*Pterourus*) *esperanza* Beutelspacher, 1975	Valid	Sampled
Papilio (*Pterourus*) *eurymedon* Lucas, 1852	Valid	Sampled
Papilio (*Pterourus*) *euterpinus* Salvin & Godman, 1868	Valid	Sampled
Papilio (*Pterourus*) *garamas* (Geyer, [1829])	Valid	Sampled
Papilio (*Pterourus*) *glaucus* Linnaeus, 1758	Valid	Sampled
Papilio (*Pterourus*) *hellanichus* Hewitson, 1868	Valid	**Unsampled**
Papilio (*Pterourus*) *homerus* Fabricius, 1793	Valid	Sampled
Papilio (*Pterourus*) *maraho* Shiraki & Sonan, 1934	Valid	Sampled
Papilio (*Pterourus*) *menatius* (Hübner, [1819])	Valid	Sampled
Papilio (*Pterourus*) *multicaudata* Kirby, 1884	Valid	Sampled
Papilio (*Pterourus*) *neyi* Niepelt, 1909	Valid	Sampled
Papilio (*Pterourus*) *palamedes* Drury, 1773	Valid	Sampled
Papilio (*Pterourus*) *pilumnus* Boisduval, 1836	Valid	Sampled
Papilio (*Pterourus*) *rutulus* Lucas, 1852	Valid	Sampled
Papilio (*Pterourus*) *scamander* Boisduval, 1836	Valid	Sampled
Papilio (*Pterourus*) *troilus* Linnaeus, 1758	Valid	Sampled
Papilio (*Pterourus*) *victorinus* Doubleday, 1844	**stat. rev.** (formerly a subspecies of *menatius*)	Sampled
Papilio (*Pterourus*) *warscewiczii* Hopffer, 1865	Valid	Sampled
Papilio (*Pterourus*) *xanthopleura* Salvin & Godman, 1868	Valid	**Unsampled**
Papilio (*Pterourus*) *zagreus* Doubleday, 1847	Valid	Sampled
Sinoprinceps Hancock, 1983 **[2 spp.]: 2 sampled in the current study (=100%)**	**Status**	**In the tree**
Papilio (*Sinoprinceps*) *benguetana* Joicey & Talbot, 1923	**stat. rev.** (formerly a subspecies of *xuthus*)	Sampled
Papilio (*Sinoprinceps*) *xuthus* Linnaeus, 1767	Valid	Sampled

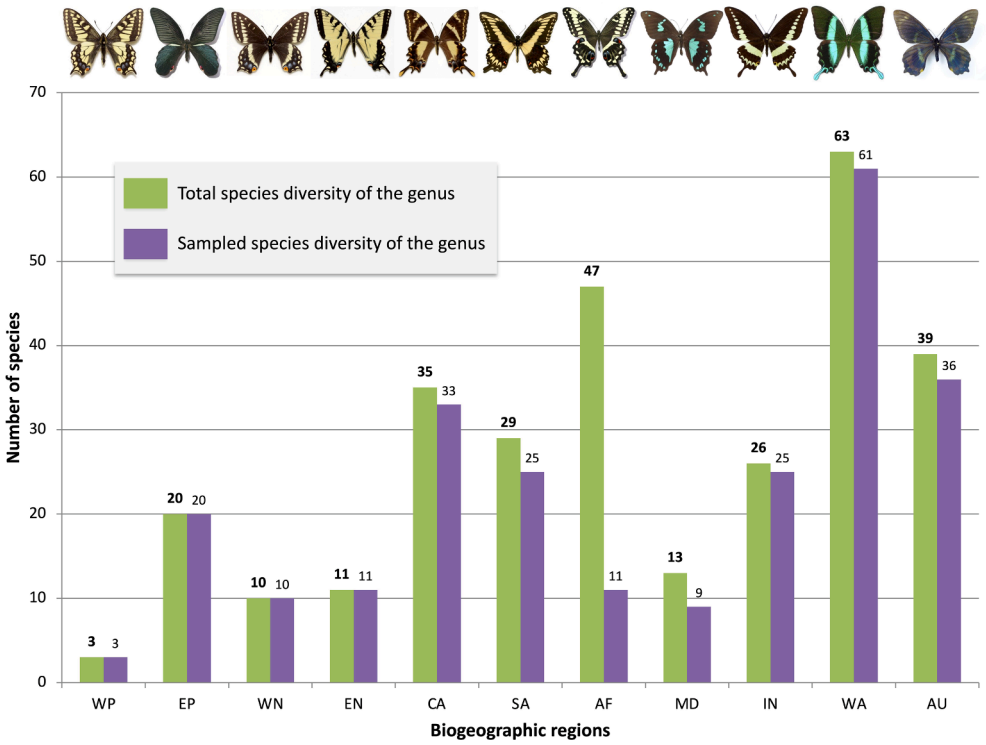


Fig. 2. Geographic distribution and sampling of *Papilio*. Histogram shows the number of species occurring in 11 biogeographic regions and sampled in the current phylogenetic tree; African species diversity is clearly under sampled. WP: Western Palearctic, EP: Eastern Palearctic, WN: Western Nearctic, EN: Eastern Nearctic, CA: Central America and Caribbean Islands, SA: South America, AF: Africa, MD: Madagascar, IN: India and Himalayan foothills, WA: Southeast Asia and Wallacea, and AU: Australasia. Pictures from Fabien L. Condamine.

Ornithoptera priamus, *Parides photinus*, *Pharmacophagus antenor* and *Troides helena*), which is sister to Papilionini, (3) two species of the tribe Leptocircini (*Graphium sarpedon* and *Lamproptera meges*), which is sister to all other Papilioninae, (4) two species of the tribe Parnassiini (*Hypermnestra helios* and *Parnassius apollo*), (5) four species of the tribe Zerynthiini (*Allancastria louristana*, *Bhutanitis mansfieldi*, *Sericinus montela* and *Zerynthia polyxena*), (6) two species of the tribe Luehdorfiini (*Archon apollinaris* and *Luehdorfia puziloi*), and (7) the single species of the subfamily Baroniinae, *Baronia brevicornis*, which is the sister lineage to all Papilionidae and was used as root of the phylogenetic tree (Condamine et al., 2012; Allio et al., 2020). Overall, the molecular dataset comprised 202 species (184 ingroups and 18 outgroups).

We assembled a supermatrix dataset with data extracted from Sanger sequencing and genome sequencing available on GenBank as of September 2021. The molecular data mainly came from previous studies (e.g. Zakharov et al., 2004a, 2004b; Condamine et al., 2012, 2013a, 2013b; Lewis et al., 2015; Wu et al., 2015; Owens et al., 2017, 2020; Allio et al., 2021; Joshi and Kunte, 2022). We used five mitochondrial gene fragments (*COI*, *COII*, *ND1*, *ND5* and *rRNA 16S*) and two nuclear gene fragments (*EF-1α* and *Wg*), chosen based on their availability among *Papilio* species. We aligned the DNA sequences for the rRNA 16S using MAFFT 7.110 (Katoh and Standley, 2013) with the E-INS-i algorithm, while we aligned the coding genes using MACSE 2.00 (Ranwez et al., 2011) with the *alignSequences* subprogram and default options. All the resulting alignments were checked for codon stops and eventually refined by eye with Mesquite 3.7 (Maddison and Maddison, 2021). All gene alignments were concatenated into a nucleotide supermatrix, which is available in FigShare (Data 1: https://doi.org/10.6084/m9.figshare.13318124.v1).

2.2. Inferring phylogenetic relationships

We performed both maximum likelihood (ML) and Bayesian Inference (BI) to reconstruct phylogenetic relationships. Although phylogenomic approaches have used amino acids on Papilionidae at the genus level (Allio et al., 2020), all our analyses relied on nucleotides for the tree inference of *Papilio* because the molecular dataset is limited to seven gene fragments and converting it to amino acids will result in a smaller dataset with limited phylogenetic information given the species diversity of the focal clade. ML inference was implemented with IQ-TREE 2.1.2 (Minh et al., 2020) using ModelFinder to select the best-fit partition scheme and the best-fitting substitution model for each partition (*-m MFP + MERGE* option, Chernomor et al., 2016; Kalyaanamoorthy et al., 2017). For IQ-TREE analyses, we estimated the most likely tree with 100 separate ML searches, which after initial model optimization on a parsimony tree used 100 random tree topologies as starting trees for each search. As recommended, we optimized ML searches to avoid local optima by (1) increasing the number of unsuccessful iterations before stopping tree optimization to 500 (*-nstop 500* option), and (2) decreasing the perturbation strength for randomized NNI to 0.2 (*-pers 0.2* option). Statistical reliability of the ML tree was evaluated with 100 non-parametric bootstraps under the optimal partitioned model to obtain ML bootstrap percentages (BP_{PART}). To compare branch supports, a second ML analysis with IQ-TREE was carried out under the same conditions but with 2,000 ultrafast bootstraps ($UFBS_{PART}$; Hoang et al., 2018). BS values and UFBS values were considered strong when higher than 70% and 95%, respectively. We also assessed branch support using Shimodaira-Hasegawa approximate likelihood ratio test (SH-aLRT; Guindon et al., 2010), with SH-aLRT values above 80% considered as strong support for a clade.

BI analyses were performed using MrBayes 3.2.7a (Ronquist et al., 2012). Because ModelFinder includes more substitution models than BEAST or MrBayes, the most likely partitioning scheme and substitution models can be altered, the molecular dataset was analyzed with PartitionFinder 2.1.1 (Lanfear et al., 2017) to estimate the best partition scheme with an initial subset of 19 possible partitions (all three codon positions for coding sequences and non-coding genes treated separately). Partitions and corresponding optimal substitution models were searched using the *greedy* algorithm, the *mrbayes* set of models, and the Bayesian Information Criterion (BIC) to compare the fit of different models. We used reversible-jump Markov chain Monte Carlo (rjMCMC) to sample the entire space of possible models instead of using the *a priori* substitution models recovered by PartitionFinder (Huelsenbeck et al., 2004). MrBayes analyses were performed with two separate runs, a

random starting tree, and eight rjMCMC (one cold and seven incrementally heated, *temp = 0.1*) running for 30 million generations with tree sampling occurring every 3,000 generations (resulting in 10,000 trees) to calculate the clade posterior probabilities (PP_{PART}). We also specified (1) a uniform prior probability of phylogenies (i.e. all possible trees are considered *a priori* equally probable), and (2) a uniform prior probability distribution on branch lengths. Convergence of the Bayesian runs was ensured by checking the average deviation of split frequencies (ADSF), the potential scale reduction factor (PSRF) values, the effective sample size (ESS) of all parameters, and by plotting the log-likelihood of the samples against the number of generations in Tracer 1.7.1 (Rambaut et al., 2018). The runs had to have values of ADSF approaching zero, PSRF close to 1.0 and ESS above 200 to assume convergence. All trees that predated the time needed to reach a log-likelihood plateau were discarded as burn-in, and the remaining samples were used to generate a 50% majority rule consensus tree (option *halfcompat*). Branch support was estimated and $PP_{PART} \geq 0.95$ was considered to indicate strong support for a given clade (Douady et al., 2003). All MrBayes analyses were performed on the computer cluster CIPRES Science Gateway (Miller et al., 2015), using BEAGLE (Ayres et al., 2012) with default parameters.

BI was also conducted using PhyloBayes MPI 1.8 (Lartillot et al., 2013) under the CAT-GTR-Γ_4 mixture model (Lartillot and Philippe, 2004), which has proven to perform well on large molecular datasets (e.g. Allio et al., 2020). The analyses were conducted on the nucleotide dataset. For each analysis, two independent MCMC starting from a random tree were run at least 10,000 cycles, with trees and associated model parameters being sampled every 10 cycles. The initial 2,000 trees sampled in each MCMC run were discarded as burn-in after checking for convergence in both likelihood and model parameters (*tracecomp* subprogram), and clade posterior probability (*bpcomp* subprogram). We checked the mean (*meandiff*) discrepancy observed across all bipartitions, considering that *meandiff* < 0.01 indicates convergence. The 50% majority-rule Bayesian consensus tree and the associated posterior probabilities (PP_{CAT}) were then computed from the remaining trees using *bpcomp*.

2.3. Testing topology hypotheses

We conducted tests of topology hypotheses with MrBayes when phylogenetic analyses did not recover the same relationships as those from previous studies (see in Zakharov et al., 2004a; Condamine et al., 2012; Lewis et al., 2015). To do so, we conducted topology hypotheses by enforcing the clade (species, genera) to be sister to another clade or to be monophyletic in a case of non-monophyly. Bayesian inferences were run with the exact same priors and parameters (see above) but including constraints and a model to compute the marginal likelihood estimate (MLE) to obtain the likelihood score of each constrained topology, which was then compared to the score of an unconstrained topology (Table 2).

We used stepping-stone sampling (Xie et al., 2011) to estimate the marginal likelihood of the specified topology and the unconstrained topology. Stepping-stone sampling (SS) is considerably more accurate than the harmonic mean of the likelihoods from a standard MCMC run, and has been shown to be more efficient than the thermodynamic integration (Baele et al., 2013). SS estimates the model likelihood by sampling a series of distributions that represent different mixtures of posterior distribution and prior distribution (Xie et al., 2011). To obtain an adequate sample from most of the steps in the algorithm, we used 100 steps with 300,000 generations each, for a total of 30 million generations. To monitor convergence during each step, we set the diagnostics frequency to once every 1,000 generations. The MLE obtained with the SS in MrBayes were used to calculate the Bayes factors (BF). The logarithm of the BF is the difference in the logarithms of the marginal model likelihoods. The BF is calculated by subtracting the MLE of the constrained topology and the MLE of the unconstrained topology. We considered BF values > 10 favoring one model over another as very significant (Nylander *et al.*, 2004; Brown and Lemmon, 2007).

2.4. Estimation of divergence times

We first tested the hypothesis of a molecular clock with PATHd8 (Britton et al., 2007). Since a strict molecular clock was not supported for 72.7% of the nodes in this dataset at $P < 0.05$, a Bayesian relaxed-clock approach considering rate variations across lineages was employed to estimate divergence times (Drummond et al., 2006). Bayesian MCMC analyses implemented in BEAST 1.10.4 (Suchard et al., 2018) were performed to approximate the posterior distribution of rates and divergence times and infer their credibility intervals.

Molecular dating analyses can be sensitive to several parameters or priors (Alfaro and Holder, 2006; Drummond and Bouckaert, 2015). For instance the prior governing the branching process (Condamine et al., 2015) or the number of molecular clocks (Angelis et al., 2018; Condamine et al., 2018a) may impact estimates of divergence times. Accordingly, we conducted analyses with the following non-default settings and priors: (1) the molecular dataset was partitioned according to the best-fit scheme recovered by PartitionFinder, (2) substitution models were set based on ModelFinder estimation in IQ-TREE (see above), (3) a birth–death tree prior was set for the branching process, and (4) one clock model was set for the mitochondrial partitions, and one clock for the nuclear partitions. Finally, we let the tree topology be estimated by BEAST (classic operator mix) but enforced the monophyly of subfamilies, tribes, and genus *Papilio* following the results of phylogenomic analyses (Allio et al., 2020).

Bayesian relaxed clocks were set up with an uncorrelated lognormal distribution clock model, with the mean set to a uniform prior between 0 and 0.5 (starting value = 0.1), and an exponential prior (lambda = 0.333) for the standard deviation. The birth–death process (Gernhard, 2008) is a more realistic branching tree prior than a Yule model for species-level divergence times analysis. The birth–death process was set using the following uniform priors: the mean growth rate ranged between 0 and 1 with a starting value at 0.1 births per lineage per million years, and the relative death rate ranged between 0 and 1 deaths per lineage per million years (starting value = 0.5). We performed four independent BEAST runs (different seeds) for 50 million generations of

Table 2
Bayesian tests of topology hypotheses using the stepping-stone sampling (SS). The topology was constrained with previously inferred relationships (Zakharov *et al.*, 2004) that were not recovered in our best topology. The SS estimated the marginal likelihood for each topology and was compared to the marginal likelihood of the best topology (unconstrained). The marginal likelihood is used to calculate Bayes factors to select the topology for the dataset. Results for BF are as follows: non-significant (0 < BF < 2), positive support (2 < BF < 6), strong support (6 < BF < 10), and very strong support (BF > 10).

Hypothesis tests on various systematic positions	Marginal likelihood (SS)	Bayes factor
Unconstrained MrBayes phylogenetic analysis	−97,079.39	–
Monophyly of the subgenus *Pterourus* (i.e. excluding *Agehana*)	−97,127.68	**48.29**
Monophyly of the subgenus *Druryia* (i.e. *Druryia* + *Nireopapilio*)	−97,269.70	**190.31**
Monophyly of the subgenus *Princeps* (i.e. *Princeps* + *dardanus* sp. gr.)	−97,277.63	**198.24**
Monophyly of the subgenus *Menelaides* (i.e. *Menelaides* + *Araminta*)	−97,145.29	**65.9**
Monophyly of the New World *Papilio*	−97,094.08	**14.69**
P. alexanor sister to *Agehana* + *Chilasa* + *Pterourus*	−97,141.38	**61.99**

MCMC each and with a sampling frequency of 5,000 generations. We discarded the first 10% of generations as burn-in and checked for convergence using Tracer (ESS > 200). We combined the four runs using LogCombiner. Using TreeAnnotator, we computed the maximum clade credibility tree with median ages and the 95% credibility intervals (CI) at each node. The xml files are available in FigShare (Data 2: https://doi.org/10.6084/m9.figshare.13318124.v1).

2.5. Fossil and secondary calibrations

To explore the effect of fossil calibrations versus secondary calibrations on estimating divergence times, we designed two calibration sets. All node calibrations were assigned a uniform prior distribution with hard bounds (Yang and Rannala, 2006).

First, we used three unambiguous fossil calibrations assigned to Papilionidae, two of which are Parnassiinae (de Jong, 2017). The first is †*Thaites ruminiana* (Scudder, 1875), a compression fossil from limestone in the Niveau du gypse d'Aix Formation of France (Aix-en-Provence, Bouches-du-Rhône) within the Chattian (23.03–28.1 Ma) of the late Oligocene (Sohn et al., 2012). †*Thaites* is sister to Parnassiini, and occasionally sister to Luehdorfiini + Zerynthiini (Condamine et al., 2018b). 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* is sister to *Archon* (Luehdorfiini, Condamine et al., 2018b), 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) found in the Green River Formation (Colorado, U.S.A.). The age of †*Praepapilio* was used to constrain the crown age of Papilionidae with a uniform distribution bounded by a minimum age of 47.8 Ma since the Green River Formation encompasses a 5 million-year period between ~ 48.5 and 53.5 Ma, which falls within the Ypresian (47.8–56 Ma) in the early Eocene (Smith *et al.*, 2003; de Jong, 2007). These three fossil calibrations have been used in previous studies (e.g. Condamine et al., 2012, 2013a, 2018a, 2018b; Allio et al., 2020). Uniform distributions of the fossil calibrations were bounded with a maximum age of 150 Ma, a conservative age congruent with the major radiation of angiosperms (Magallón et al., 2015; Foster et al., 2017; Li et al., 2019).

Second, we relied on secondary calibrations with uniform priors applied to the nodes where we set the fossil calibrations plus nodes that are in common between our phylogeny and the phylogeny of Papilionidae. We retrieved the 95% CI from recent estimations of divergence times of the family (Allio et al., 2021) as follows: (1) the crown of Papilionidae (root of the tree) between 47.8 and 70.9 Ma, (2) the crown of Parnassiinae between 29.9 and 58.9 Ma, (3) the crown of Luehdorfiini between 13.9 and 35.3 Ma, (4) the crown of Papilioninae between 34.4 and 62.9 Ma, (5) the crown of Leptocircini between 26.6 and 49.9 Ma, (6) the crown of Papilionini + Troidini between 30.8 and 56.1 Ma, (7) the crown of Troidini between 26.9 and 50.4 Ma, and (8) the crown of Papilionini between 27.5 and 50.9 Ma (Allio et al., 2021).

2.6. Inference of historical biogeography

We estimated the ancestral ranges of origin and geographic range evolution for *Papilio* using the ML approach of dispersal-extinction-cladogenesis (DEC, Ree and Smith, 2008) as implemented in the DEC eXtended version (DECX, Beeravolu and Condamine, 2016; available at: https://github.com/champost/DECX). To infer the biogeographic history of a clade, DEC requires a time-calibrated tree, the current distribution of each species for a set of geographic areas, and a time-stratified geographic model that is represented by connectivity and dispersal scalar matrices for specified time intervals spanning the entire evolutionary history of the group. DECX allows classical vicariance as a cladogenetic event by using temporally flexible constraints on the connectivity between any two given areas following the movement of landmasses and dispersal opportunity over time. DECX can analyze phylogenies on the order of several thousand tips with numerous areas (Rolland and Condamine, 2019). DECX does not incorporate the founder-event speciation (+J parameter) because of concerns with statistical validity of model choice among DEC-derived models (Ree and Sanmartín, 2018). Also, founder-event speciation often leads to inferences that are decoupled from time, with null or extremely low extinction rates, an effect of the model favoring cladogenetic events over anagenetic events (Ree and Sanmartín, 2018), which makes it inadequate for reconstructing the history of ancient groups with widespread distributions.

The geographic distribution for all 235 *Papilio* species was categorized by coding the presence or the absence of each species in the following areas: (1) West Palearctic, defined as Europe west of the Urals and the part of Asia west of this north–south line, (2) East Palearctic, defined as everywhere considered Palearctic east of the Urals, above 3,000 m in the Himalayas and north of Sichuan in China, (3) West Nearctic, defined as North America west of the Rocky Mountains, (4) East Nearctic, defined as North America east of the Rocky Mountains, (5) Central America, defined as from the northern border of Mexico southwards to the border between Panama and Colombia, also including the Caribbean islands except Trinidad and Tobago, (6) South America, defined as all countries from Colombia to Argentina and including Trinidad and Tobago, (7) Africa, defined as the whole of the African continent and Arabian Peninsula but excluding the islands in the Indian Ocean, (8) Madagascar, defined as the island of Madagascar and all other Indian Ocean islands in the vicinity, (9) India, defined as the area below 3,000 m from NW Pakistan to the border with Myanmar, (10) Indonesia and Wallacea, defined as Myanmar, SE Asia, southern China, western Indonesia to Lydekker's Line; including the Lesser Sunda Islands but excluding Timor, Wetar and associated islands, which are Australasian in origin, and (11) Australasia, defined as everywhere east of Lydekker's Line but including Timor, Wetar and small nearby islands. Species which only marginally enter an area were excluded. We used data available in the literature (e.g. Collins and Morris, 1985; Tyler et al., 1994; Scriber et al., 1995). The geographic distribution of all species and the species sampled in this study is presented in Fig. 2. The resulting matrix of species distribution for *Papilio* is available in Table S1.

A time-stratified geographic model was built using connectivity matrices that consider paleogeographic changes through time with time slices indicating the possibility or not for a species to colonize a new area (Beeravolu and Condamine, 2016). Based on paleogeographic reconstructions (e.g. Scotese, 2004; Blakey, 2008; Seton *et al.*, 2012), we created a connectivity matrix to represent major changes in tectonic conditions that may have affected the distribution of these butterflies. We specified constraints on area connectivity by coding 0 if any two areas are not connected or 1 if they are connected at a given period. We did not add dispersal matrices because setting the values for dispersal rates between regions through time is highly subjective, and it has been shown that dispersal probability categories had minor effects on ancestral state estimation (Chacon and Renner, 2014). Therefore, we assumed a dispersal scalar matrix with equal dispersal rates between areas through time. Biogeographic ranges larger than four areas in size were disallowed as valid biogeographic states if they were not subsets of the terminal species ranges; widespread ranges comprising areas that have never been geographically connected were also removed.

We estimated the most likely ancestral states at each node by performing a DEC analysis under an ML procedure as implemented in DECX (Beeravolu and Condamine, 2016) using both species distribution matrix and connectivity matrices. The files for reproducing the analyses are available in FigShare (Data 3: https://doi.org/10.6084/m9.figshare.13318124.v1).

2.7. Investigating heterogeneity of diversification rates

To provide an assessment of diversification rates through time, we used the ML approach of Morlon et al. (2011), tested and automated in Mazet et al. (2023) and implemented in the R-package *RPANDA* 2.0 (Morlon et al., 2016). This method aims at capturing heterogeneity of diversification by allowing preselected subclades to follow different birth–death models from the deeper pruned tree (the backbone). For both subclades and backbones (different backbones are tested because of the different combination of shifts), speciation and/or extinction rates can change exponentially through time: $\lambda(t) = \lambda_0 \times e^{\alpha t}$ with λ_0 denoting speciation at present, α the trend of rate variation of speciation through time *t*. Extinction rate can exceed speciation, meaning that diversification rates can be negative (Morlon et al., 2011), which results in a declining paleodiversity dynamic that can be expected for the backbone because of the higher proportion of long branches after isolating recently-originated subclades.

Following Mazet et al. (2023), for the analysis of *Papilio*, we first computed the clades' sampling fractions based on our revised taxonomy with the *get.sampling.fraction* function. We specified the nine following subclades: the subgenera *Achillides*, *Chilasa*, *Druryia*, *Heraclides*, *Menelaides*, *Nireopapilio*, *Papilio*, *Pterourus*, and the monophyletic group of Lauraceae feeders regrouping *Chilasa* and *Pterourus* (Apiaceae and Rutaceae feeders are not monophyletic). We then estimated the number of possible shifts to be tested in combination with corresponding backbones using the *get.comb.shift* function. The selection of subclades created 319 combinations of subclade(s)/backbone that were compared to the clade-homogeneous birth–death model. We designed and fitted five diversification models to each clade and backbone with the *shift.estimate* function that determines the best fitting model and most likely combination of shifts. The five models are: (1) a Yule model, where speciation is constant and extinction is null (BCST); (2) a constant birth–death model, where speciation and extinction rates are constant (BCST_DCST); (3) a variable speciation rate model without extinction (BVAR); (4) a variable speciation rate model with constant extinction (BVAR_BCST); and (5) a rate-constant speciation and variable extinction rate model (BCST_BVAR). Diversification rates are defined backward in time such as a positive dependency parameter (α for speciation and β for extinction) reflects a slowdown of rates towards the present. Finally, based on the most likely rate estimates and best combination of shifts, we estimated the diversity dynamics of each clade and resulting backbone recovered in the best shift combination using the *apply_prob_dtt* and *paleodiv* function. The files for reproducing the analyses are available in FigShare (Data 4: https://doi.org/10.6084/m9.figshare.13318124.v1).

3. Results and discussion

3.1. Global phylogeny of Papilio

Partitioned phylogenetic analyses with IQ-TREE and MrBayes provided almost identical phylogenetic trees, differing in branch length estimates (Fig. 3; Figs S1, S2 for IQ-TREE). The genus *Papilio* was always recovered as monophyletic with maximal (PP_{CAT} = 1, PP_{PART} = 1, BP_{PART} = 100, $UFBS_{PART}$ = 100) branch support (Table 3), which was expected since molecular data have never found it non-monophyletic (e.g. Aubert et al., 1999; Zakharov et al., 2004a; Condamine et al., 2012; Lewis et al., 2015; Owens et al., 2017, 2020; Allio et al., 2020). The Bayesian inference with MrBayes converged well (ADSF = 0.007293, average PSRF for parameter values = 1.000, and ESS≫200 for all parameters). Bayesian analyses reconstructed a robust phylogeny with 69.4% of the nodes recovered with strong support within *Papilio* (nodes with $PP_{PART} \geq 0.95$, Fig. 3). Robustness was slightly lower for ML analysis with non-parametric bootstrap: 67.4% of branches in the tree were strongly supported (nodes with $BP_{PART} \geq 70$, Fig. S1). In comparison, IQ-TREE analysis with $UFBS_{PART}$ recovered the most robust phylogeny with 82.5% of branches supported by $UFBS_{PART} \geq 95$ (Fig. S2). PhyloBayes analysis under the mixture (CAT-GTR-Γ_4) model ran during 11,350 cycles and converged well (*meandiff* = 0.00814). The topology (Fig. S3) was very similar to traditional partitioned analyses with IQ-TREE and MrBayes, except within the Old World *Papilio* (see below). Overall, branch support in PhyloBayes was also like IQ-TREE and MrBayes with 66.7% of branches having $PP_{CAT} \geq 0.95$ (see Fig. S4 for a comparison of branch support across all analyses).

All phylogenetic analyses reconstructed a backbone topology with three main clades (Clades 1 to 3 on Fig. 3, Table 3). Subgenus *Heraclides* (Clade 1) was sister to all remaining *Papilio*, which is composed of a clade mostly including the New World *Papilio* (Clade 2, subgenera *Chilasa* and *Pterourus*) and another clade comprising the Old World *Papilio* (Clade 3, the remaining subgenera). Although this topology was consistent across methods, branch support for the large clade of *Papilio* excluding the subgenus *Heraclides* was not robust (PP_{CAT} = 0.75, PP_{PART} = 0.98, BP_{PART} = 55, and $UFBS_{PART}$ = 90). Subgenus *Heraclides* was always monophyletic with high branch support (PP_{CAT} = 1, PP_{PART} = 0.96, BP_{PART} = 75, and $UFBS_{PART}$ = 100). The species relationships within *Heraclides* largely correspond to previous works (Lewis et al., 2015; Owens et al., 2017, 2020).

New World *Papilio sensu stricto* (i.e. subgenera *Heraclides* and *Pterourus*) were not inferred as monophyletic in all analyses. We always recovered the American subgenus *Pterourus* within a clade including Asian subgenera *Chilasa* and '*Agehana*' (Clade 2 on Fig. 3) with strong branch support (PP_{CAT} = 0.99, PP_{PART} = 1, BP_{PART} = 91, and $UFBS_{PART}$ = 100). In addition, '*Agehana*' was internal to subgenus *Pterourus* in all analyses (Table 3). This topology agrees with some phylogenetic work (Wu et al., 2015) but contrasts with others (Zakharov et al., 2004a; Condamine et al., 2012; Lewis et al., 2015), which proposed that *Heraclides* is sister to *Pterourus* + *Chilasa* + *Alexanoria*. The substantial increase of taxon sampling may account for the difference between topologies: we sampled 64 species belonging to these subgenera, while Zakharov et al. (2004a) and Condamine et al. (2012) analyzed only 21 species (including *P. alexanor*). Bayesian topology tests provided strong support (BF > 6) for the non-monophyly of the New World *Papilio* clade (*sensu* Zakharov et al., 2004a; Condamine et al., 2012; Lewis et al., 2015), and decisive support (BF > 10) for the non-monophyly of *Pterourus sensu lato* (Table 2). These new phylogenetic arrangements have important biogeographic implications (see below).

Papilio alexanor, a mysterious Eurasian lineage that is notoriously difficult to place within *Papilio* (Zakharov et al., 2004a; Wu et al., 2015), was sister to the Old World clade including all remaining species, with moderate to high branch support (PP_{CAT} = 1, PP_{PART} = 0.99, BP_{PART} = 58, and $UFBS_{PART}$ = 85). The Bayesian topology test constraining *P. alexanor* to be sister to the clade *Pterourus* + *Chilasa* + '*Agehana*' (Zakharov et al., 2004a; Condamine et al., 2012) yielded a worse MLE than the unconstrained topology with strong support (BF > 10, Table 2). Our unconstrained topology agrees with a recent but sparsely sampled phylogenomic study (Allio et al., 2020), suggesting that the phylogenetic position of *P. alexanor* is reaching a consensus.

The Old World *Papilio* clade (Clade 3 on Fig. 3) includes the highest species diversity of the genus. After the divergence with *P. alexanor*, partitioned phylogenetic analyses recovered a series of five main clades (Clades 3a to 3e on Fig. 3) sustained by short internal branches leading to different subgenera with moderate to high branch support (Table 3). These internal branches were the source of discrepancy between the Bayesian analyses (under both the mixture model and the partitioned analyses) and ML analyses. The main discrepancy was that Clades 3a and 3b were recovered as sisters in ML analyses and weakly supported (BP_{PART} = 23, and $UFBS_{PART}$ = 58; Fig. S1-2) but not recovered in Bayesian analyses (Fig. 3, Fig. S3). Despite a substantially increased taxon sampling compared to previous studies (37% of total species diversity in Condamine et al., 2012 versus 78% in this study), these results suggest that the systematic backbone of *Papilio* is still not definitely resolved, which impedes a higher-level systematic revision of the genus. We also acknowledge the limited size of the Sanger-based dataset and

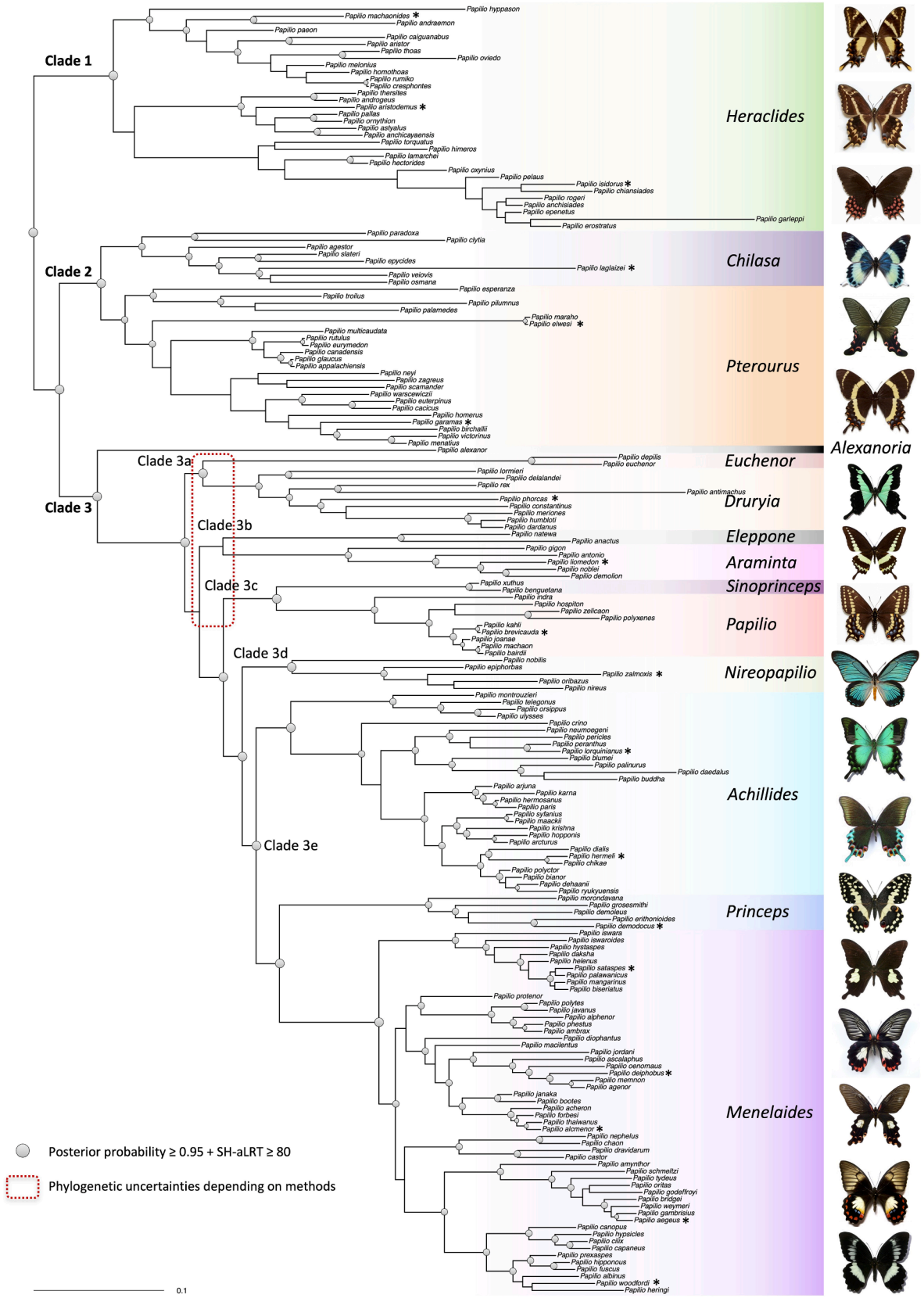


Fig. 3. Bayesian molecular phylogeny of *Papilio*. The phylogeny was inferred with MrBayes using a traditional partitioning strategy and a reversible-jump MCMC approach for selecting the best fitting substitution models. Posterior probabilities ≥ 0.95 are indicated at nodes with filled circles. The red dashed rectangle indicates phylogenetic uncertainties between methods. Existing subgeneric classification is shown with colored rectangles delineating subgenera. The outgroups are removed. Asterisks indicate species illustrated on the right. Pictures from Fabien L. Condamine. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 3
Phylogenetic, dating, and biogeographic results for the main nodes of *Papilio*. Clades 1 to 3 are shown in Fig. 3.

Clades	Node supports				Median node ages, Ma (95% CI)		Ancestral area estimates	Biogeographic interpretation
	PP_{CAT}	PP_{PART}	BS_{PART}	$UFBS_{PART}$	Fossil calibrations	Secondary calibrations		
Papilio sensu lato	1	1	100	100	30.06 (20.66 – 47.42)	29.95 (23.97 – 37.34)	EP + WN + CA + WA	Origin centred on Beringia
Clade 1: *Heraclides*	1	0.96	75	100	21.22 (13.95 – 33.86)	21.2 (16.07 – 27.48)	CA	Central America
Clade 2: *Papilio sensu lato* excl. *Heraclides*	0.75	0.98	55	90	28.57 (19.98 – 45.45)	28.5 (22.89 – 35.55)	EP + WN + CA + WA	Origin centred on Beringia
Chilasa + *Pterourus* + *'Agehana'*	0.99	1	91	100	22.73 (15.36 – 35.9)	22.64 (17.33 – 28.74)	EP + WN + CA + WA	Origin centred on Beringia
Chilasa	1	1	99	100	16.68 (12.49 – 29.71)	18.63 (14.02 – 24.44)	WA	Indonesia
Pterourus + *'Agehana'*	1	1	79	99	20.38 (13.8 – 32.28)	20.3 (15.3 – 25.92)	EP + WN + CA	Origin centred on Beringia
'Agehana'	1	1	100	100	0.28 (0.09 – 0.63)	0.27 (0.09 – 0.53)	EP + WA	Asia and Indonesia
Clade 3: Old World *Papilio*	1	0.99	58	85	26.93 (18.62 – 42.62)	26.86 (21.52 – 33.51)	EP	East Palearctic
Old World *Papilio* excl. *Alexanoria*	1	1	86	100	23.74 (16.45 – 37.49)	23.67 (18.96 – 29.58)	WP + EP + AF + WA	Paleotropics
Clade 3a: *Druryia* + *Euchenor*	0.82	0.99	74	99	22.32 (15.3 – 35.23)	22.23 (17.46 – 27.95)	WP + EP + AF + WA	Paleotropics
Euchenor	1	1	100	100	5.49 (2.74 – 9.71)	5.33 (2.84 – 8.42)	AU	Australasia
Druryia	1	1	99	100	17.95 (12.07 – 28.62)	17.88 (13.61 – 22.95)	AF	Afrotropics
Clade 3b: *Eleppone* + *Araminta*	–	0.72	50	93	21.09 (14.82 – 33.9)	20.99 (16.4 – 26.47)	WA	Indonesia
Eleppone	1	1	97	100	9.72 (4.41 – 17.0)	9.54 (4.77 – 15.03)	AU	Australasia
Araminta	1	1	100	100	14.07 (8.87 – 22.59)	13.9 (9.89 – 18.61)	WA	Indonesia
Clade 3c + Clade 3d + Clade 3e	0.99	1	72	99	21.35 (15.08 – 33.97)	21.25 (16.97 – 26.59)	WP + EP + AF + WA	Paleotropics
Clade 3c: *Sinoprinceps* + *Papilio sensu stricto*	1	1	97	100	17.39 (11.36 – 27.85)	17.27 (13.1 – 22.34)	EP	East Palearctic
Sinoprinceps	1	1	100	100	1.5 (0.63 – 2.82)	1.44 (0.67 – 2.45)	EP + WA	Asia and Indonesia
Papilio sensu stricto	1	1	100	100	10.3 (6.26 – 16.69)	10.2 (7.01 – 13.85)	EP + WN	Origin centred on Beringia
Clade 3d + Clade 3e	0.99	1	60	99	20.17 (14.01 – 31.92)	20.09 (16.09 – 25.23)	WP + EP + AF + WA	Paleotropics
Clade 3d: *Nireopapilio*	1	1	76	98	16.1 (10.21 – 25.51)	15.98 (11.78 – 20.65)	AF	Afrotropics
Clade 3e: *Achillides* + *Princeps* + *Menelaides*	–	0.91	54	96	19.52 (13.5 – 30.98)	19.41 (15.57 – 24.39)	WP + EP + AF + WA	Paleotropics
Achillides	1	1	85	99	17.1 (11.45 – 26.88)	17.0 (13.31 – 21.62)	WA + AU	Indonesia and Australasia
Princeps + *Menelaides*	0.93	1	70	99	18.21 (12.6 – 29.05)	18.15 (14.37 – 22.86)	WP + EP + AF + WA	Paleotropics
Princeps	1	1	100	100	11.29 (6.93 – 18.01)	11.21 (8.0 – 14.97)	WP + EP + AF + MD	Paleotropics
Menelaides	1	1	99	100	13.94 (9.4 – 21.98)	13.94 (10.9 – 17.65)	WA	Indonesia

phylogenomic studies may illuminate the early evolutionary history of *Papilio* in future studies to provide a definitive resolution of subgeneric classification.

Clade 3a was always composed of the subgenus *Euchenor* (Joshi and Kunte, 2022) including two New Guinean species (*Papilio euchenor* and *P. depilis*), subtended by a long branch, which was systematically recovered as sister to a clade comprising species of subgenus *Druryia* (including the *antimachus* and *dardanus* species-groups) with moderate to strong branch support (Clade 3a on Fig. 3; $PP_{CAT} = 0.82$, $PP_{PART} = 0.99$, $BP_{PART} = 74$, and $UFBS_{PART} = 99$). We still have poor knowledge on subgenus *Druryia* because existing phylogenies only included a handful of species. Here we found the *nireus*, *oribazus*, and *zalmoxis* species groups in a separate clade sister to *Papilio nobilis* with strong branch support ($PP_{CAT} = 1$, $PP_{PART} = 1$, $BP_{PART} = 76$, $UFBS_{PART} = 98$). This clade has recently been ranked as the subgenus *Nireopapilio* (Cotton and Nakae, 2020). A Bayesian topology test provided decisive support ($BF > 10$) for non-monophyly of subgenus *Druryia* (Table 2), therefore lending support to the validity of subgenus *Nireopapilio*. However, it is important to remain cautious because future studies with more comprehensive sampling could recover subgenus *Druryia* as monophyletic, notably through the inclusion of the *zenobia* species-group not sampled in this or previous studies.

Papilio natewa, the latest described *Papilio* species (Tennent et al., 2018), was always sister to the monotypic subgenus *Eleppone*, with maximal branch support in all analyses (Fig. 3, Table 3). This indicates that *Papilio natewa* can be placed in the subgenus *Eleppone,* together with *P. anactus* from eastern Australia. The subgenus *Eleppone* was always found to be sister to a strongly supported clade of five Indonesian species, the *demolion* species group, which was previously considered to belong to the subgenus *Menelaides* but now assigned to the subgenus *Araminta*. Although often recovered, this sister relationship was not highly supported (Clade 3b on Fig. 3; $PP_{PART} = 0.72$, $BP_{PART} = 50$, and $UFBS_{PART} = 93$).

The next clade (Clades 3c + 3d + 3e on Fig. 3) included species classified in six subgenera and was generally strongly supported in all analyses ($PP_{CAT} = 0.99$, $PP_{PART} = 1$, $BP_{PART} = 72$, and $UFBS_{PART} = 99$). The first to branch off this clade were subgenus *Sinoprinceps* (*Papilio xuthus* and *P. benguetana*) and subgenus *Papilio* (*machaon* species-group)

found to be sisters (Clade 3c on Fig. 3) in all analyses with maximal branch support (except $BP_{PART} = 97$). These two subgenera were both recovered as monophyletic with maximal branch supports. The remaining subgenera comprised *Nireopapilio* + *Achillides* + *Princeps* + *Menelaides* (Clade 3d + Clade 3e on Fig. 3) and formed a strongly supported clade ($PP_{CAT} = 0.99$, $PP_{PART} = 1$, $BP_{PART} = 60$, and $UFBS_{PART} = 99$). The species *Papilio nobilis* (usually ranked in subgenus *Princeps*, *nobilis* group) was always found as sister to *Nireopapilio* ($PP_{CAT} = 1$, $PP_{PART} = 1$, $BP_{PART} = 76$, and $UFBS_{PART} = 98$; Clade 3d on Fig. 3), and the latter contained the species of the *nireus*, *oribazus* and *zalmoxis* groups. The subgenus *Achillides* was strongly supported as monophyletic in all analyses ($PP_{CAT} = 1$, $PP_{PART} = 1$, $BP_{PART} = 85$, and $UFBS_{PART} = 99$), and was often found as sister to the clade *Princeps* + *Menelaides* with moderate nodal support (Clade 3e on Fig. 3). Within *Achillides*, the species relationships largely agree with the study of Condamine et al. (2013b), but we added two endangered species (*P. buddha* and *P. chikae*; Collins and Morris, 1985). Endemic to the Western Ghats biodiversity hotspot of Southern India, *P. buddha* was always recovered in the *palinurus* group (*P. daedalus* and *P. palinurus*) from Southeast Asia with strong support ($PP_{CAT} = 1$, $PP_{PART} = 1$, $BP_{PART} = 100$, and $UFBS_{PART} = 100$), but its sister relationship with *P. daedalus* was weakly supported ($PP_{CAT} = 0.8$, $PP_{PART} = 0.72$, $BP_{PART} = 64$, and $UFBS_{PART} = 62$). Such a relationship is interesting to study in terms of wing morphological evolution between the *palinurus* group and the other *Achillides* endemic to India, *P. crino*. *Papilio chikae* is endemic to the Philippines (North Luzon) and was always sister to *P. hermeli* as expected (Cabusas et al., 2020), also endemic to the Philippines (North Mindoro) with maximal support in all analyses. They were together nested in the *bianor* group with maximal branch support as previously found (Condamine et al., 2013b). Subgenus *Princeps*, containing five species of the *demoleus* species group, constituted a strongly supported clade with maximal branch support, which was found to be sister to subgenus *Menelaides* ($PP_{CAT} = 0.93$, $PP_{PART} = 1$, $BP_{PART} = 70$, and $UFBS_{PART} = 99$). A Bayesian topology test rejected the hypothesis of a monophyletic origin for all species usually placed in subgenus *Princeps* like the *dardanus* species-group ($BF > 10$, Table 2). Finally, the species comprising the subgenus *Menelaides* formed a solid monophyletic group with robust branch support in all analyses ($PP_{CAT} = 1$, $PP_{PART} = 0.99$, $BP_{PART} = 99$, and $UFBS_{PART} = 100$). The species relationships within *Menelaides* largely agree with the recent phylogenetic study of Joshi and Kunte (2022). We constrained subgenera *Menelaides* and *Araminta* (species previously included in *Menelaides*) to form a single clade, but the Bayesian analysis confirmed the non-monophyly of such an artificially inclusive *Menelaides* with decisive support ($BF > 10$, Table 2).

3.2. Origin of Papilio and subgeneric diversification

The four independent Bayesian runs of the two dating analyses converged well (ESS≫200 for most of the parameters) and were therefore combined. The analyses yielded almost identical estimates of divergence times with<0.5 million years of difference for all nodes, regardless of the calibrations used: fossils or secondary (Fig. 4, Table 3, see Figs S5-6 for the chronograms resulting from the BEAST analyses). The dating analyses estimated that *Papilio* originated in the Oligocene ca. 30.06 Ma (95% CI = 20.66–47.42 Ma) with the three-fossil-calibrations analysis, and ca. 29.95 Ma (95% CI = 23.97–37.34 Ma) with the eight-secondary-calibrations analysis. These results indicate that fossil and secondary calibrations can provide similar and consistent results in age estimates, which has not always been the case (Sauquet et al., 2012). For the subsequent analyses (biogeography and diversification), we selected the maximum clade credibility tree with median ages estimated with a Bayesian uncorrelated lognormal method calibrated with fossils (Fig. 4).

We estimated that the origin of subgenera ranges from the early Miocene (e.g. *Chilasa*, *Heraclides*, *Pterourus*), middle Miocene (e.g. *Achillides*), to the late Miocene (e.g. *Araminta*, *Menelaides*, *Papilio*) (Fig. 4, Table 3). These age estimates are slightly younger than previous analyses with age differences ranging from ca. 1 to 2 million years younger, but it is important to note that 95% CIs for these nodes strongly overlap with earlier studies. For instance, we estimated the origin of the clade '*Agehana*' + *Chilasa* + *Pterourus* at 22.73 Ma (95% CI = 15.36–35.9 Ma), whereas we previously estimated this age at 20.9 Ma (95% CI = 16.9–25.6 Ma; Lewis et al., 2015) or 19.7 Ma (95% CI = 17.2–22.9 Ma, Owens et al., 2017, 2020) but we have a similar estimation to Wu et al. (2015) who found an age of 22.63 Ma (95% CI = 18.93–26.58 Ma). Slight differences are found in (1) subgenus *Heraclides* recovered at 21.22 Ma (95% CI = 13.95–33.86 Ma) compared to our previous estimation of 22.1 Ma (95% CI = 18.0–26.9 Ma; Lewis et al., 2015) or 21.9 Ma (95% CI = 17.5–26.4 Ma; Wu et al., 2015); and (2) subgenus *Achillides* estimated at 17.1 Ma (95% CI = 11.45–26.88 Ma) whereas we previously recovered a 2-million-year older age at 19.3 Ma (95% CI = 16.4–21.8 Ma; (Condamine et al., 2013b). We argue that these slight age discrepancies mostly come from the differences in taxon sampling resulting in new phylogenetic placements (e.g. *P. alexanor*) and potentially from practices in dating techniques both influencing divergence time estimates.

3.3. Northern origin and dynamic dispersal into the tropics

DEC analyses recovered a region including West Nearctic, Central America, East Palearctic, and Sundaland as the most likely ancestral geographic origin (relative probability = 0.467), when Asia and North America were connected by the Bering land bridge in the Oligocene (Fig. 4). The second best ancestral area was composed of West Nearctic, Central America, and East Palearctic (relative probability = 0.215). Hence, a Northern (Laurasian) origin is preferred over a Southern (Gondwanan) origin. This is not unexpected given the age of the genus and the results of previous biogeographic analyses of the genus (e.g. Condamine et al., 2012, 2013; Lewis et al., 2015; Wu et al., 2015), although some studies estimated an older age and Gondwanan origin (Zakharov et al., 2004a). The DEC analyses indicate a dynamic biogeographic history with numerous dispersal events inferred ($n = 64$) in comparison to vicariance events ($n = 22$). Dispersal events were mostly southward ($n = 32$) from a northern origin (or dispersal *into the tropics*, Condamine et al., 2012; Rolland et al., 2015) than northward ($n = 17$, when including northwestward dispersals) from the equator (or dispersal *out of the tropics*, Jablonski et al., 2006). Northward dispersals are only estimated in the last 10 million years. Excluding vicariance events involving areas around the Bering Strait ($n = 2$), we found more vicariance in the Old World ($n = 13$) than in the New World ($n = 7$). We also estimated numerous local extirpation events ($n = 51$), which tend to be more numerous in the Old World subgenera ($n = 27$) than in the New World subgenera ($n = 15$), likely due to the Miocene fragmentation of the Boreotropical forest in the Holarctic (Pound et al., 2012).

3.4. The Paleotropics as a biogeographic crossroad

In the non-monophyletic New World *Papilio*, subgenus *Heraclides* originated in Central America (very likely including Caribbean Islands; Lewis et al., 2015) and clade *Chilasa* + *Pterourus* originated in a region comprising the West Nearctic, Central America, East Palearctic, and Sundaland. The clade of Old World *Papilio* originated in East Palearctic with *P. alexanor* being sister to all remaining Old World *Papilio*, which soon after their divergence colonized the Paleotropics through West and East Palearctic + Sundaland and extended to Africa (see shaded area on Fig. 4).

We found the Paleotropics to be an important ancestral area for the historical biogeography of *Papilio*. It is striking that, within the Old World clade (Clade 3), all the main nodes of the backbone (seven in total out of 26, Table 3) are estimated to originate within the Paleotropics in the early Miocene (23 to 15 Ma, Fig. 4). During this time period, Africa, West and East Palearctic, and Sundaland were mostly covered by

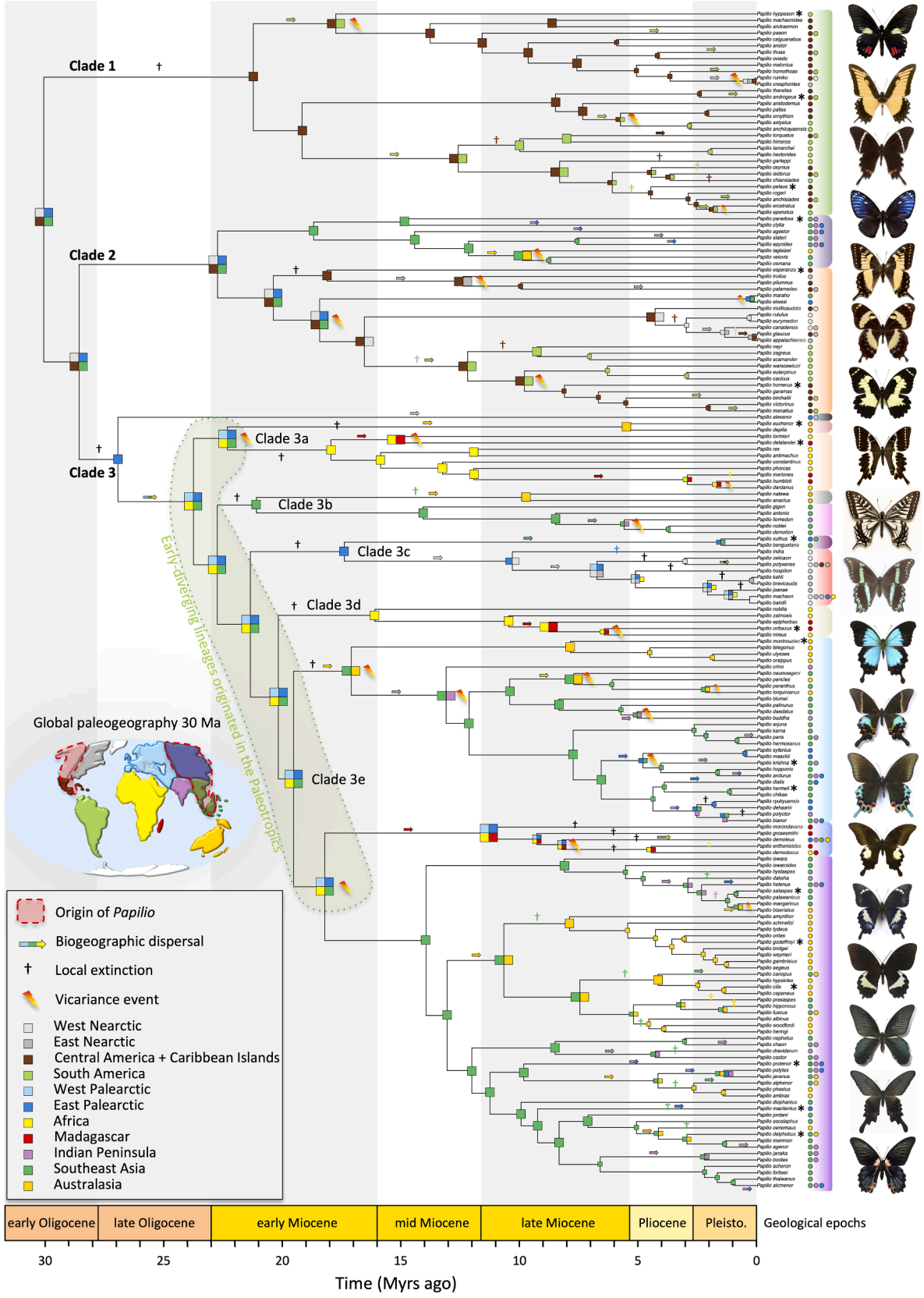


Fig. 4. Dated phylogeny and historical biogeography of *Papilio*. The maximum clade credibility tree shows median ages estimated with a Bayesian uncorrelated lognormal method calibrated with fossils (see Figs S5-7 for additional results). The bottom-left corner map represents the global paleogeography about 25 Ma with continents delimited into 11 areas. Colored areas on the map correspond to colored squares for each node, representing inferred ancestral area(s) with the DEC model, and colored circles at tips, representing present-day distributions. The outgroups are removed. Asterisks indicate species illustrated on the right. Pictures from Fabien L. Condamine.

tropical-adapted or paratropical forests providing an almost continuous tropical habitat between these large regions, which likely facilitated biotic movements within the Paleotropics (e.g. Ziegler et al., 2003; Morley, 2011). However, starting after the middle Miocene climatic optimum (Steinthorsdottir et al., 2021), the global cooling of the Cenozoic fostered the geographic contraction of the tropical belt toward the equator that disappeared from higher latitudes (Pound et al., 2012). Our biogeographic estimates indicated widespread geographic extinctions in the Paleotropics, especially in the Holarctic, and at least seven different lineages were able to disperse southwards to track their preferred macroclimatic conditions, supposedly tropical climates (Condamine et al., 2012). Like crematogastrine ants (Blaimer et al., 2018), we show that the Paleotropics had a central role in the origin and evolution of *Papilio*. The vast and stable ecological opportunity offered by the tropical rainforests likely explained the *into-the-tropics* dispersal trend to track tropical contractions as climate cooled down toward the

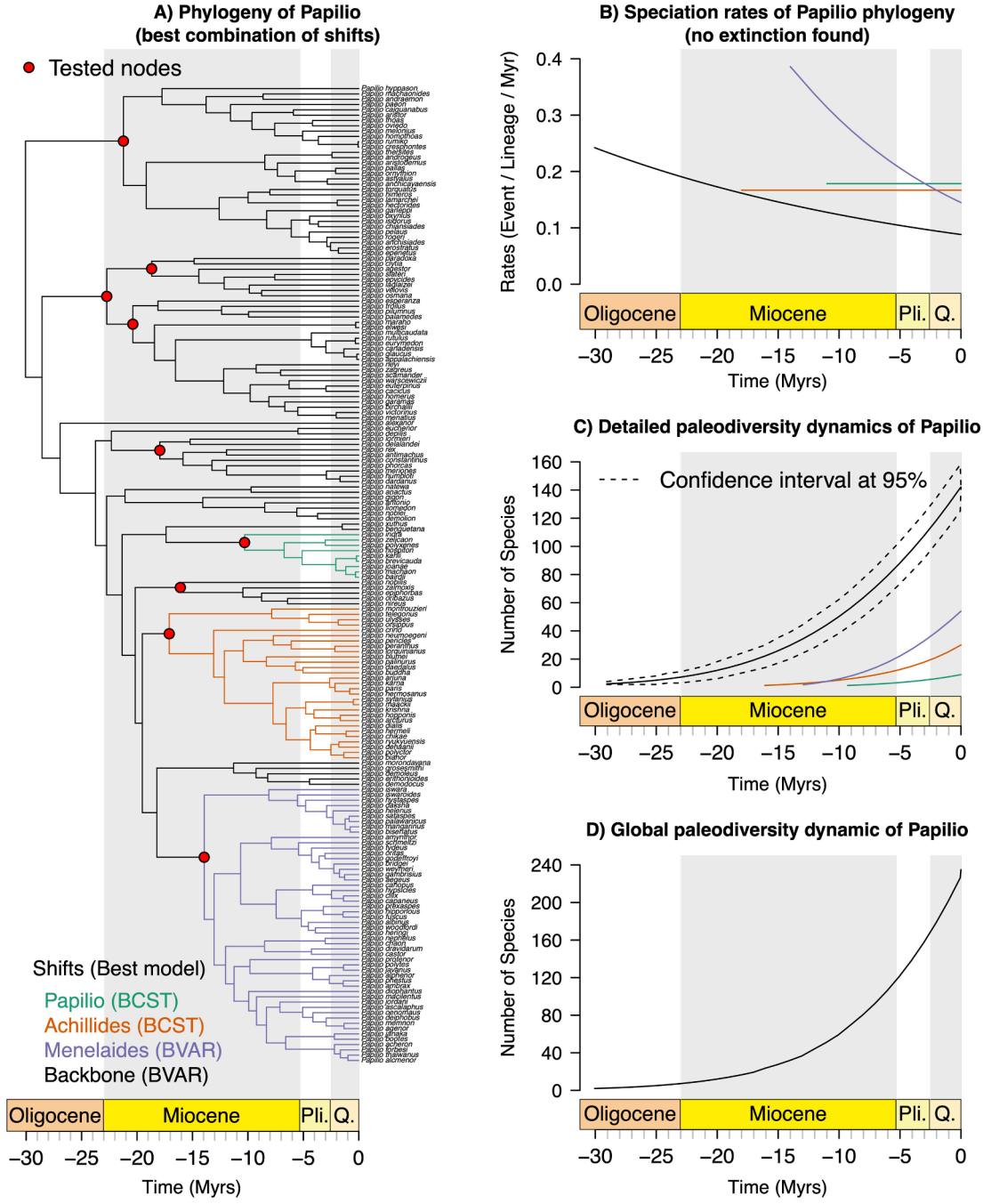


Fig. 5. Diversification shifts and diversity dynamics estimated for *Papilio* as estimated by the best-fitting combination of shifts. A) The phylogeny of *Papilio* with shifts highlighted in colors and best models in parenthesis. Red dots correspond to all tested nodes. B) The evolution of diversification rates through time for the backbone and all subclades that are found as significant shifts. C) Diversity dynamics for the backbone and subclade trees as estimated with the probabilistic approach (dotted line represents the confidence interval of diversity estimates for the backbone). For the sake of clarity, confidence intervals of diversity estimates for subclades are not represented. D) Global diversity dynamics of *Papilio* obtained by summing all the diversity dynamics for the backbone and subclade trees. Pli. = Pliocene, Q = Quaternary, Myrs = million years. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

present. This supports the hypothesis that potentially many clades, particularly inhabitants of Boreotropical floras, were likely extirpated from the Holarctic and persist today in more southern tropical locations (Meseguer et al., 2018).

3.5. Tempo and mode of Papilio diversification

The diversification analyses investigating rate heterogeneity across *Papilio* revealed four best equally-likely scenarios of diversification (ΔAICc $\leq$ 2, Table S2), which refutes the hypothesis of a single diversification rate for the whole genus (ΔAICc = 14.06; Table S2). These shift-configuration scenarios include three shifts located at subgenera *Achillides*, *Menelaides* and *Papilio* for the first best scenario (Fig. 5**A**), two shifts (*Achillides* and *Menelaides*) for the second best scenario, five shifts (*Achillides*, *Heraclides*, *Menelaides*, *Papilio* and *Pterourus*) for the third best scenario (Fig. S7A), and four shifts (*Achillides*, *Heraclides*, *Menelaides* and *Papilio*) for the fourth best scenario. Subgenera *Achillides* and *Menelaides* are always found as significant shifts. All subclades are better explained by a pure-birth model (BCST), except subgenus *Menelaides* that follows a model with a decreasing speciation rate through time (BVAR; Fig. 5**B**, Fig. S7B, Table S3).

Interestingly, after isolating the subclades that significantly diversify at different rates, two patterns emerge from the backbone (remaining lineages) of best combinations. For the two first best combinations, the backbone is explained by a model with a decreasing speciation rate over time without extinction (Fig. 5**B**), while the backbone in the two other scenarios follows a model with the same decreasing speciation rate and includes a constant extinction rate (Fig. S7B). In the latter case, net diversification rate becomes negative in the Pliocene onward (ca. 5 Ma), leading to a waxing-waning pattern of diversification (Fig. S7C). This short diversity decline near the present did not affect the trend of the global diversity dynamics because it happens when subclades are already diversifying, thus compensating the decline (Fig. S7B, C, D). The diversity dynamics of these two different patterns in the backbone (decline or not) are similar except with a faster accumulation of lineages during the early Miocene for the scenarios with a decline (Fig. 5, Fig. S7).

These four scenarios highlight common features related to adaptive radiation. First, all scenarios agree to show a two-step diversification history. The first phase of the *Papilio* radiation proceeded at a high net diversification (ranging from 0.224 to 0.426 events/Myr/lineage depending on the scenario), while the second period of diversification is supported by subclade dynamics having lower speciation rates than at the origin of *Papilio* (e.g. 0.122 for *Heraclides*, 0.178 for *Papilio*), except for *Menelaides* (0.386 events/Myr/lineage). The first radiation phase reaches an equilibrium before a short declining phase in half of the best scenarios. For the two other scenarios, the backbones have lower speciation rates at the origin but also contain more lineages (only three or two shifts). These results agree with the general tendency for diversification to slow down as evolution proceeds, considered as one of the most pervasive macroevolutionary principles (Rabosky, 2009; Morlon et al., 2010; Moen and Morlon, 2014; Condamine et al., 2019). Such diversification slowdowns have often been interpreted as the effect of competition for resources or niche availability (Rabosky, 2009) or the role of past environmental changes (Condamine et al., 2019). It is thus possible that the diversity of some *Papilio* lineages can be at equilibrium or limited by ecological resources, although these lineages are still expanding toward the present but at a slower pace than in the past, supporting the 'damped increase' hypothesis (Cornell, 2013).

Clade-specific rates of diversification can further explain differences in clades' species richness. Although the number of clade shifts varies from one scenario to another, this pattern of radiating subclades within the genus *Papilio* seems to be supported by a strong signal (Table S2). Clade-heterogeneous diversification has been unveiled across Papilionidae as a whole (Condamine et al., 2012; Allio et al., 2021) or for other swallowtail clades (Condamine et al., 2018b), which has been then linked to species' traits or environments. In the case of *Papilio*, the two most important shifts in diversification are recovered for subgenera *Achillides* and *Menelaides*. These two subclades are island-dwelling groups in the Indo-Malayan and Australasian Archipelagos (Condamine *et al.*, 2013b; Joshi and Kunte, 2022), a region where species diversity is the highest at the global scale for this genus (Fig. 2). Their diversification proceeded through repeated island colonizations since the middle Miocene leading to rampant allopatric speciation across the entire archipelago (Condamine *et al.*, 2013b; Joshi and Kunte 2022). In addition, the subgenus *Menelaides* shows an early-burst pattern with high speciation rate when it originated (0.386 events/Myr/lineage), followed by a decrease toward the present (0.145 events/Myr/lineage, Fig. 5). Besides their island distribution, *Menelaides* are well-known for their Batesian mimicry of aposematic and toxic troidine swallowtails (Kunte, 2009; Palmer and Kronforst, 2020; Kizhakke and Kunte, 2022), which may have spurred their high speciation rates. The subgenus *Menelaides* diversified faster than any other subgenus, and yet is the most recent clade with subgenus *Papilio* (Table S3). The latter is often found as a significant shift (Fig. 5), which may be linked to the host-plant shift on Apiaceae (Allio et al., 2021) and/or to the glaciation cycles initiated in the late Pliocene (Dupuis and Sperling, 2020).

We did not recover a model incorporating extinction rate in any subclade, which might be artifactual due to the difficulty of estimating extinction rates from phylogenies of extant species alone, or this may be biologically realistic if the nascent subclades did not have enough time to experience extinction to be detectable (e.g. Morlon et al., 2011). Assuming these estimates are not biased, this may suggest that subclades of *Papilio* evolved under a 'museum model of diversity' with a very low extinction rate. Altogether, while their geographic ranges contracted toward the equator (Fig. 4), the genus sustained high rates of species diversification thanks to multiple speciation shifts leading to a global increase of species accumulation toward the present (Fig. 5). This suggests that the tropics are not only the evolutionary source of the *Papilio* diversity but have also played an important role in mitigating their extinction.

3.6. Should more than one genus be recognized within Papilio?

The rank of genus is the most visible rung in the Linnaean hierarchy. It forms part of every species name, indicating shared relationships when there are multiple species in a genus, or unusual distinctness of a species when it constitutes a monotypic genus. Other than monophyly (evidenced by one or more synapomorphies), there are no objective criteria for determining genus boundaries, and even monophyly does not determine how high up on its tree a branch should be cut. Further, when subclades within a genus are elevated to the rank of genus, the information provided by the name about closer relationships is balanced by the lost information about the broader relationships of the group. Nonetheless, numerous attempts have been made to provide general guidelines for recognizing a genus (Talavera et al., 2012; Dorchin et al., 2018; Sigward et al., 2018; Zhang et al., 2019; Nakahara et al., 2020), with criteria including the compactness of a group, its distinctness from other such groups, the number of species it contains, its estimated age of divergence, comparability to related genera, consistency with established usage of a name, and degree of confidence in the assessment of its phylogenetic relationships (Ashlock and Mayr, 1991). Balancing and prioritizing potential conflicts among these criteria, while at the same time conveying as much information as possible about relationships between species to non-expert end-users of the names, remains one of the most subjective aspects of systematics.

It is clear that taxonomic stability is a primary aim, if not *the* primary aim, of the International Code of Zoological Nomenclature (ICZN, 1999, 2012). The genus *Papilio sensu lato* has been stable and unambiguously defined by morphology since Munroe (1961), and later supported by molecular evidence (e.g. Aubert et al., 1999; Caterino and Sperling, 1999; Zakharov et al., 2004a), although the clades within that genus

have continued to be volatile (Hancock, 1983; Miller, 1987; Zakharov et al., 2004a). Even *Chilasa* and '*Agehana*', which Hauser *et al.* (2005) treated as distinct, have continued to move back and forth with each new phylogeny (Zakharov et al., 2004a; Wu et al., 2015). Consequently, we have chosen here to retain the use of the genus *Papilio* in its broad sense, pending a more rigorous and objective integration of the multiple criteria that may be used to determine the boundary of a genus. For widely recognized and previously stable taxonomic groups, we consider it to be in the best interests of systematists to be conservative in proposing new name changes for anything other than well supported cases of non-monophyly. In our experience, continued name changes for high-profile taxa will consistently elicit frustration and disrespect from the larger community of biologists, conservationists, and the general public. At the same time, if names reflect current knowledge, then advances based on new evidence or even different weighting of evidence will inevitably result in name changes. Fortunately, a solution to this conflict is available through the use of subgeneric names, which allow systematists to refer clearly to their refined taxonomic concepts without putting the burden of increased confusion from name changes onto the users of these species and genus names in other fields.

4. Conclusion

Building a comprehensive species-level dated phylogeny for a globally distributed insect clade is challenging. We addressed this challenge with the genus *Papilio*, which currently includes 235 species worldwide. Our study assembles ~80% of the species diversity in a molecular supermatrix and provides a generally robust time-calibrated phylogenetic reference for *Papilio*. This new tree confirms previous results such as the early-diverging position of New World *Papilio* and the sister relationships between some subgenera and unveils novel relationships like the placement of *P. alexanor* and the non-monophyly of several subgenera. This phylogenetic framework provides the foundation for a systematic revision, but we urge specialists to refrain from changing the status of subgenera to the rank of genus. Deep nodes within *Papilio* remain poorly resolved and some subgenera are still poorly sampled, especially in Africa. Future studies with denser taxon sampling and full genomic data will undoubtedly provide a more accurate phylogeny for *Papilio* and justify a thorough systematic revision. Here we used this dated phylogeny to assess the evolutionary history of *Papilio,* with estimated divergence times that place its origin in the Oligocene (~30 Ma). Biogeographic analyses suggest a Beringian origin followed by southward dispersals into the tropics, with a pivotal role of the Paleotropics as a biogeographic crossroad for the worldwide colonization of *Papilio*.

5. Formal taxonomic changes

Our results confirm the status of '*Agehana*' (two species, *P. elwesi* and *P. maraho*) within subgenus *Pterourus* as proposed by Wu et al. (2015). *Papilio zalmoxis* Hewitson, 1864 is transferred from subgenus *Druryia* to subgenus *Nireopapilio*. This finding is also supported morphologically by Huxley (1976), who showed that the wing scale structure of *P. zalmoxis* is identical to *P. bromius* (valid name: *P. chrapkowskoides* Storace, 1952), both also containing blue-fluorescent pigment.

Two taxa in subgenus *Druryia* are recognized as separate species to *Papilio dardanus* Yeats, 1776, namely *Papilio meriones* C. Felder & R. Felder, 1864 (**stat. rev.**) from Madagascar and *Papilio humbloti* Oberthür, 1888 (**stat. rev.**) from Comoros Islands (Table 4).

In subgenus *Pterourus*, Tyler et al. (1994) placed *Papilio victorinus* Doubleday, 1844 from Mexico within *Pterourus menatius* (Hübner, [1819]), but we consider this Central American taxon as specifically distinct from the South American *Papilio menatius* (Owens et al. 2017, 2020), and we reinstate *Papilio victorinus* Doubleday, 1844 as a separate species (**stat. rev.**).

In subgenus *Papilio* two North American species are confirmed as

Table 4
Changes to subspecies nomenclature resulting from literature review and this study.

Species	New subspecies combinations
Papilio (*Achillides*) *orsippus* Godman & Salvin, 1888 (**stat. rev.**)	*Papilio orsippus orsippus* Godman & Salvin, 1888 *Papilio orsippus ambiguus* Rothschild, 1895 (**comb. nov.**) *Papilio orsippus gabrielis* Rothschild, 1898 (**comb. nov.**) *Papilio orsippus kallinikos* Fruhstorfer, 1903 (**comb. nov.**) *Papilio orsippus rothschildianus* Fruhstorfer, 1909 (**comb. nov.**)
Papilio (*Achillides*) *telegonus* C. Felder & R. Felder, 1860 (**stat. rev.**)	*Papilio telegonus telegonus* C. Felder & R. Felder, 1860 *Papilio telegonus dohertius* Rothschild, 1898 (**comb. nov.**) *Papilio telegonus morotaicus* Rothschild, 1908 (**comb. nov.**)
Papilio (*Euchenor*) *depilis* Rothschild, 1895 (**stat. rev.**)	*Papilio depilis depilis* Rothschild, 1895 *Papilio depilis neohannoveranus* Rothschild, 1898 (**comb. nov.**) *Papilio depilis novohibernicus* Rothschild, 1898 (**comb. nov.**)
Papilio (*Menelaides*) *aegeus* Donovan, 1805	*Papilio aegeus inopinatus* Butler, 1883 (**comb. nov.**) *Papilio aegeus komos* Fruhstorfer, 1904 (**comb. nov.**)
Papilio (*Menelaides*) *agenor* Linnaeus, 1758 (**stat. rev.**)	*Papilio agenor agenor* Linnaeus, 1758 *Papilio agenor polymnestor* Cramer, 1775 (**comb. nov.**) *Papilio agenor parinda* (Moore, 1881) (**comb. nov.**) *Papilio agenor nicobarensis* Hachitani, 1986 (**comb. nov.**) *Papilio agenor heronus* Fruhstorfer, 1902 (**comb. nov.**) *Papilio agenor iriomotensis* Fujioka, 2012 (**comb. nov.**) *Papilio agenor thunbergii* Siebold, 1824 (**comb. nov.**)
Papilio (*Menelaides*) *capaneus* Westwood, 1843 (**stat. rev.**)	*Papilio capaneus capaneus* Westwood, 1843 *Papilio capaneus beccarii* Oberthür, 1880 (**comb. nov.**) *Papilio capaneus gyrei* Tennent, 1999 (**comb. nov.**) *Papilio capaneus hasterti* Ribbe, 1907 (**comb. nov.**) *Papilio capaneus indicatus* Butler, 1876 (**comb. nov.**) *Papilio capaneus relmae* Tennent, 1999 (**comb. nov.**) *Papilio capaneus rotalita* (Swinhoe, 1893) (**comb. nov.**) *Papilio capaneus xenophilus* Mathew, 1886 (**comb. nov.**)
Papilio (*Menelaides*) *chaon* Westwood, 1844 (**stat. rev.**)	*Papilio chaon chaon* Westwood, 1844 *Papilio chaon annulus* Pendlebury, 1936 (**comb. nov.**) *Papilio chaon chaonulus* Fruhstorfer, 1902 (**comb. rev.**) *Papilio chaon rileyi* Fruhstorfer, 1913 (**comb. rev.**)
Papilio (*Menelaides*) *cilix* Godman & Salvin, 1879 (**stat. rev.**)	*Papilio cilix cilix* Godman & Salvin, 1879 *Papilio cilix lamponius* Fruhstorfer, 1904 (**comb. nov.**)
Papilio (*Menelaides*) *daksha* Moore, [1889] (**stat. rev.**)	*Papilio daksha daksha* Moore, [1889] *Papilio daksha mooreanus* Rothschild, 1895 (**comb. nov.**)
Papilio (*Menelaides*) *javanus* C. Felder, 1862 (**stat. nov.**)	*Papilio javanus javanus* C. Felder, 1862 *Papilio javanus theseus* Cramer, 1777 (**comb. nov.**) *Papilio javanus melanides* Haan, 1840 (**comb. nov.**)

(*continued on next page*)

Table 4 (*continued*)

Species	New subspecies combinations
	Papilio javanus vigellius Fruhstorfer, 1909 (**comb. nov.**)
	Papilio javanus messius Fruhstorfer, 1909 (**comb. nov.**)
	Papilio javanus sotira Jordan, 1909 (**comb. nov.**)
	Papilio javanus timorensis C. Felder & R. Felder, 1864 (**comb. nov.**)
	Papilio javanus alcindor Oberthür, 1879 (**comb. nov.**)
	Papilio javanus kurokawai Nakae, 2013 (**comb. nov.**)
	Papilio javanus tucanus Jordan, 1909 (**comb. nov.**)
Papilio (*Menelaides*) *mangarinus* Rothschild, 1908 (**stat. nov.**)	*Papilio mangarinus mangarinus* Rothschild, 1908
	Papilio mangarinus jindanus Rothschild, 1908 (**comb. nov.**)
	Papilio mangarinus tambora Rothschild, 1908 (**comb. nov.**)
	Papilio memnon memnon Linnaeus, 1758
Papilio (*Menelaides*) *memnon* Linnaeus, 1758	*Papilio memnon mayo* Atkinson, 1874 (**comb. nov.**)
Papilio (*Menelaides*) *oritas* Godman & Salvin, 1879 (**stat. rev.**)	*Papilio oritas oritas* Godman & Salvin, 1879
	Papilio oritas websteri Grose-Smith, 1894 (**comb. nov.**)
	Papilio oritas byronensis Talbot, 1932 (**comb. nov.**)
Papilio (*Menelaides*) *palawanicus* Staudinger, 1888 (**stat. nov.**)	*Papilio palawanicus palawanicus* Staudinger, 1888
	Papilio palawanicus enganius Doherty, 1891 (**comb. nov.**)
	Papilio palawanicus sinabangana Goode & Burk, 2013 (**comb. nov.**)
	Papilio palawanicus boloboca Page &Treadaway, 1996 (**comb. nov.**)
Papilio (*Papilio*) *bairdii* Edwards, 1866 (**stat. rev.**)	*Papilio bairdii bairdii* Edwards, 1866
	Papilio bairdii oregonia Edwards, 1876 (**comb. nov.**)
	Papilio bairdii dodi McDunnough, 1939 (**comb. nov.**)
Papilio (*Pterourus*) *victorinus* Doubleday, 1844 (**stat. rev.**)	*Papilio victorinus victorinus* Doubleday, 1844
	Papilio victorinus morelius Rothschild & Jordan, 1906 (**stat. rev.**)
	Papilio victorinus vulneratus Butler, 1872 (**stat. rev.**)

separate from *P. machaon* Linnaeus, 1758, namely *Papilio kahli* F. & R. Chermock, 1937 (**stat. nov.**) and *Papilio bairdii* Edwards, 1866 (**stat. rev.**). *Papilio bairdii* includes two subspecies, *oregonia* Edwards, 1876 and *dodi* McDunnough, 1939 (**comb. nov.**) as well as the nominate subspecies. *Papilio saharae* Oberthür, 1879 is treated as a subspecies of *P. machaon* as per Dupuis and Sperling (2020), subject to further investigation (Table 4).

We found two commonly recurrent patterns in several Asian groups, with speciation between taxa from mainland southern Asia and Sundaland, and between taxa in the Bismarck Archipelago from those in New Guinea, as detailed below.

Within subgenus *Achillides*, five taxa, three with more than one subspecies, are separated at species level based on molecular results from Condamine et al. (2013b) and Cabusas et al. (2020). *Papilio polyctor* Boisduval, 1836 (**stat. rev.**) from northern Pakistan and NW India is separated from *Papilio bianor* Cramer, 1777, and *Papilio hermeli* Nuyda, 1992 (**stat. rev.**) from Mindoro, Philippines is reinstated as a separate species to *P. chikae* Igarashi, 1965 from Luzon. *Papilio daedalus* C. Felder & R. Felder, 1861 (**stat. rev.**), with subspecies *angustatus* Staudinger, 1888, is confirmed as a separate species to *Papilio palinurus* Fabricius, 1787 (Table 4). Hiura and Alagar (1971) first separated *P. daedalus* from *P. palinurus* based on morphology, but Page and Treadaway (2003a) placed them as conspecific based on perceived similarity of genitalia. Molecular phylogenies clearly show they are separate species, and all subspecies of the two species can easily be separated based on the position of the postdiscal green hindwing band in relation to the anal eyespot. In *P. palinurus* the lowest point of the band is always above the eyespot, whereas in *P. daedalus* the band meets the eyespot. The *Papilio ulysses* group was found to consist of three separate species, *Papilio ulysses* Linnaeus, 1758 (South Moluccas, New Guinea and most associated islands, and Australia), *Papilio telegonus* C. Felder & R. Felder, 1860 (**stat. rev.**) from Northern Moluccas, and *Papilio orsippus* Godman & Salvin, 1888 (**stat. rev.**) from the Bismarck Archipelago and Solomon Islands (Table 4). Nakae (2021b) separated *Papilio arjuna* Horsfield, 1828, found on Sumatra and Java, from mainland Asian *P. paris* based on the findings of Condamine et al. (2013b), which is confirmed here due to paraphyly with *P. karna*. That study of *Achillides* also suggested that *P. maackii* and *P. syfanius* may be conspecific; but a recent genomic study clarified that mitochondrial gene exchange occurs between these two species in western China without significant exchange of nuclear genes (Xiong et al. 2022). As a result, we retain *P. maackii* and *P. syfanius* as separate species.

Subgenus *Euchenor* has previously been regarded as monobasic, but DNA analysis shows that the subgenus consists of two distinct species (Joshi and Kunte 2022): *Papilio euchenor* Guérin-Méneville, 1830 and *Papilio depilis* Rothschild, 1895 (**stat. rev.**). Jordan, in Rothschild (1895), stated that forewing scale structure in these two taxa are different but refrained from treating them as separate species (Table 4). Jordan (1896) then separated the two species, citing several morphological differences including genitalia. Jordan (1908-09) without explanation reunited *P. depilis* with *P. euchenor*, which has been followed by all authors until now. This specific separation of Bismarck Archipelago taxa from New Guinea taxa mirrors the separation of *P. orsippus* from *P. ulysses* in subgenus *Achillides*, and similar separation is found in subgenus *Menelaides* below.

In the *helenus* group of subgenus *Menelaides*, both Munroe (1961 - *nephelus*, *nubilus* and *chaon*) and Hancock (1985 - *noblei* and *antonio*) included species which have been found not to be monophyletic. *Papilio noblei* Nicéville, [1889] and *P. antonio* Hewitson, 1875 are shown here to belong to subgenus *Araminta*, which is also supported by morphology and wing pattern, particularly on the underside. *Papilio nubilus* has been shown to be a natural hybrid between *P. nephelus* and *P. polytes* rather than a valid species (Tsukada and Nishiyama 1980: 307). We confirmed that *Papilio hystaspes* C. Felder & R. Felder, 1862, separated from *helenus* by Hiura and Alagar (1971) based on male genitalia and treated as separate by Hancock (1983) but placed within *P. helenus* by Page and Treadaway (2003b), is a distinct species (**stat. rev.**) which branches off before the traditional *helenus* taxa and *sataspes* (Table 4). However, we also found that several taxa always considered to belong to *P. helenus* are separate species. The first species is *Papilio daksha* (**stat. rev.**) from Southern India. Joshi and Kunte (2022) analyzed specimens of ssp. *enganius* Doherty, 1891 from Sumatra, Java and Borneo and found that they are sister to *P. sataspes* from Sulawesi and associated islands, not *P. helenus* from mainland SE Asia and the Malay Peninsula (Table 4). The oldest name for the taxon in the Sundaic islands of Sumatra, Java, Borneo and Palawan, is *Papilio palawanicus* Staudinger, 1888 (**stat. nov.**).

In the Lesser Sunda Islands, two more taxa in the *helenus* group are also separate species, *Papilio mangarinus* Rothschild, 1908 (**stat. nov.**) and *Papilio biseriatus* Rothschild, 1895 (**stat. rev.**). *P. biseriatus* from Timor was previously treated as a separate species by Hancock (1983) and differs from the other species in the postdiscal white hindwing patch extending across four hindwing cells as in *P. hystaspes* from the Philippines, whereas in *P. helenus*, *palawanicus* and *mangarinus* the hindwing patch only covers three cells. As a result of this analysis the range of *P. helenus* is restricted to mainland Asia and the Malay Peninsula, Taiwan and Japan.

Within the *polytes* group, *Papilio protenor* Cramer, 1775 is sister to all

other species. This does not correspond to its expected position, as the early stages are very similar to *P. memnon*, not *polytes*; presumably this similarity is plesiomorphic. Five more species are recognised within the *polytes* group, two of which are newly separated in this work (Table 4). The two easternmost species, *Papilio ambrax* Boisduval, 1832 and *Papilio phestus* Guérin-Méneville, 1830, were treated as distinct species until Fujioka et al. (1997) combined them based on morphology and allopatry. However, molecular phylogenies (Joshi and Kunte, 2022) and population genetic analyses (Zhang et al., 2017; Deshmukh et al., 2022) show *P. ambrax* and *P. phestus* to be specifically distinct (**stat. rev.**). *Papilio alphenor* Cramer, 1776 was separated from *P. polytes* based on genitalia by Hiura and Alagar (1971), but subsequently sunk by Page and Treadaway (2003a) without explanation. Molecular phylogenies suggest that *P. alphenor* is sister to *phestus* + *ambrax*, not *polytes* (Joshi and Kunte, 2022), and thus must be treated as a separate species (**stat. rev.**). The remaining taxa placed in *P. polytes* also belong to two separate species: the mainland Asian populations are specifically distinct from the Sundaic island taxa, the oldest name for which is *Papilio javanus* C. Felder, 1862 (**stat. nov.**). *Papilio polytes*, *alphenor* and *javanus* also show prezygotic and postzygotic barriers to hybridization including assortative mating and low fitness of hybrid progeny, thus being reproductively isolated, with genome-wide and population genetic signatures of being highly diverged distinct species (Zhang et al., 2017; Deshmukh et al., 2022).

Specific separation between the mainland Asian taxa and those in the Sundaic islands was also found in the *memnon* group, but the status of taxa within this group is shown to be further complicated by several taxa traditionally considered as separate species being conspecific with related species (Joshi and Kunte, 2022). The larger *memnon* group then split into two clades, one containing the *bootes* and *alcmenor* species groups and the other *P. memnon* and close relatives. There are two branches within the first clade, one branch containing just two species, *Papilio bootes* Westwood, 1842 and *Papilio janaka* Moore, 1857 (**stat. rev.**, Table 4). Originally these were considered separate species, but they were treated as conspecific by Evans (1923), followed by Talbot (1939) and subsequent authors. Previous molecular phylogenies show that *P. bootes* and *P. janaka* are significantly genetically distinct (Joshi and Kunte, 2022), and are treated as different species although they are sympatric in NE Myanmar. The other branch of the first clade contains *Papilio acheron* Grose-Smith, 1887, *forbesi* Grose-Smith, 1883 and *lampsacus* Boisduval, 1836 (this last taxon not sequenced due to rarity) with *Papilio alcmenor* C. Felder & R. Felder, 1865 and *thaiwanus* Rothschild, 1898 sisters within this group. Previously *P. acheron*, *forbesi* and *lampsacus* were considered as very close to *P. memnon* due to their similar appearance, but this relationship was not confirmed by our analysis.

Molecular phylogenies confirmed that the Philippine taxon *rumanzovia* Eschscholtz, 1821 is genetically conspecific with Moluccan *P. deiphobus* (Joshi and Kunte, 2022), as stated on morphological grounds by Page and Treadaway (2003a,b). There is a wide genetic divergence between the mainland Asian and the Sundaland taxa previously united under *P. memnon*, which clearly represent two separate species, *Papilio agenor* Linnaeus, 1758 (**stat. rev.**) and *P. memnon* (Joshi and Kunte, 2022). Two traditionally well-regarded species, *Papilio polymnestor* Cramer, 1775 from peninsular India and Sri Lanka, and *Papilio mayo* Atkinson, [1874] from the Andaman Islands, do not merit species status. *Papilio polymnestor* was nested within *P. agenor* (Joshi and Kunte, 2022) and should therefore be treated as a subspecies, *Papilio agenor polymnestor* Cramer, 1775 (**stat. rev.**, Table 4). Likewise, *P. mayo* was nested within and therefore confirmed as conspecific with *P. memnon*, and should be treated as its subspecies, *Papilio memnon mayo* Atkinson, [1874] (**stat. rev.**).

The *nephelus* group comprises four species. *Papilio castor* Westwood, 1842 and *P. dravidarum* Wood-Mason, 1880 are sister species within this group, and the taxon *Papilio mahadeva* Moore, [1879] is confirmed to be conspecific with *P. castor* as stated on morphological grounds by Cotton and Racheli (2007). *Papilio chaon* Westwood, 1844 and *P. nephelus* Boisduval, 1836, from mainland SE Asia and Sundaland respectively, were originally described as separate species but were treated as conspecific by Igarashi (1979) based on early stages. Molecular phylogenies showed that genetically *nephelus* and *chaon* are highly divergent (Joshi and Kunte, 2022), and thus must be treated as separate species, *P. nephelus* and *P. chaon* (**stat. rev.**, Table 4).

Within the *aegeus* clade, *P. godeffroyi* was found to be the sister taxon to *Papilio oritas* Godman & Salvin, 1879 (**stat. rev.**) from New Britain, New Ireland and New Hannover, which was placed within *P. aegeus* by Jordan (1909). In molecular phylogenies, *P. oritas* diverged before the remaining species in the *aegeus* group (Joshi and Kunte, 2022), and thus is not conspecific with *P. aegeus*. This is another example of speciation within the Bismarck Islands. *Papilio inopinatus* Butler, 1883 from Timor was found to be nested within *P. aegeus* (**syn. nov.**), thus it becomes *Papilio aegeus inopinatus* (**comb. nov.**, Table 4).

The most complex clade within subgenus *Menelaides* is the *fuscus* group. The species *Papilio fuscus* Goeze, 1779 as recognised by Hancock (1992) in the most recent revision of this species is not a monophyletic entity. Hancock placed many taxa in his single species *Princeps fuscus* (Goeze, 1779) which we here recognise as five different species in two subclades of the *fuscus* group. Hancock placed *Papilio canopus* Westwood, 1842 and *Papilio hypsicles* Hewitson, 1868 as species level synonyms of *P. fuscus*, but we found that this arrangement is not monophyletic, and these two taxa represent separate species (**stat. rev.**, Table 4). He also placed four taxa traditionally treated within *P. fuscus* with two taxa previously known as *Papilio pitmani* Elwes & Nicéville, [1887] under the species name *Princeps prexaspes* (C. Felder & R. Felder, 1865). We confirmed that these six taxa form a single species which we treat as *Papilio prexaspes*, restricting the name *Princeps* Hübner, [1807] as the subgeneric name for the *Papilio demoleus* clade.

As stated above, Hancock (1992) treated *P. canopus* and *P. hypsicles* within *P. fuscus*, but molecular phylogenies showed that this would result in a paraphyletic species. Two groups of taxa traditionally treated within *Papilio fuscus* (e.g. Jordan, 1909) were found to be the sister to *P. hypsicles*, representing two distinct species not directly related to true *P. fuscus* (Joshi and Kunte, 2022). One of these consists of two taxa from New Britain (*lamponius* Fruhstorfer, 1904) and New Ireland (*cilix* Godman & Salvin, 1879) which represent a distinct species, *Papilio cilix* Godman & Salvin, 1879 (**stat. rev.**). This follows a pattern of speciation of New Britain and New Ireland taxa seen in other *Papilio* species as discussed above. The sister to *P. cilix* consists of the taxa traditionally placed in *P. fuscus* from eastern Australia, New Guinea and the Solomon Islands. The oldest available name for this species is *Papilio capaneus* Westwood, 1843 (**stat. rev.**, Table 4).

CRediT authorship contribution statement

Fabien L. Condamine: Conceptualization, Methodology, Validation, Investigation, Resources, Data curation, Writing – original draft, Project administration, Funding acquisition. **Rémi Allio:** Conceptualization, Methodology, Data curation, Writing – review & editing. **Eliette L. Reboud:** Conceptualization, Methodology, Writing – review & editing. **Julian R. Dupuis:** Resources, Validation, Writing – review & editing. **Emmanuel F.A. Toussaint:** Resources, Conceptualization, Methodology, Validation, Writing – review & editing. **Nathan Mazet:** Methodology, Software, Writing – review & editing. **Shao-Ji Hu:** Resources, Validation, Writing – review & editing. **Delano S. Lewis:** Resources, Validation, Writing – review & editing. **Krushnamegh Kunte:** Conceptualization, Resources, Validation, Writing – review & editing. **Adam M. Cotton:** Conceptualization, Resources, Validation, Writing – review & editing. **Felix A.H. Sperling:** Conceptualization, Validation, Investigation, Resources, Writing – review & editing, Funding acquisition.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (project GAIA, agreement no. 851188), and an 'Investissement d'Avenir' grant from the Agence Nationale de la Recherche (project CASMA, CEBA, ref. ANR-10-LABX-25-01) to F.L.C.; a Ramanujan Fellowship from the Dept. of Science and Technology, Govt. of India, to K.K., and support of the Dept. of Atomic Energy, Govt. of India, under project nos. 12-RandD-TFR-5.04-0800 and 12-RandD-TFR-5.04-0900 to TIFR/NCBS; a Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery Grant to F.A.H.S. Sequencing of most Indian *Papilio* was done at NCBS under research and voucher specimen collection permits issued to K.K. by the state forest departments in Kerala (permit no. WL 10-3781/2012 dated 18/12/2012, and GO (RT) No. 376/2012/FandWLD dated 26/07/2012) and Nagaland (permit no. CWL/GEN/240/522-39 dated 14 August 2012). Alberto Fracasso (Italy) and Laurie Wills (New Zealand) are thanked for providing samples of some hard to obtain taxa.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ympev.2023.107758.

References

Ae, S.A., 1979. The phylogeny of some *Papilio* based on interspecific hybridization data. Systematic Entomol. 4, 1–16.
Alfaro, M.E., Holder, M.T., 2006. The posterior and the prior in Bayesian phylogenetics. Annu. Rev. Ecol. Evol. Syst. 37, 19–42.
Allio, R., Scornavacca, C., Nabholz, B., Clamens, A.-L., Sperling, F.A.H., Condamine, F.L., 2020. Whole genome shotgun phylogenomics resolves the pattern and timing of swallowtail butterfly evolution. Syst. Biol. 69, 38–60.
Allio, R., Nabholz, B., Wanke, S., Chomicki, G., Pérez-Escobar, O.A., Cotton, A.M., Clamens, A.-L., Kergoat, G.J., Sperling, F.A.H., Condamine, F.L., 2021. Genome-wide macroevolutionary signatures of key innovations in butterflies colonizing new host plants. Nat. Commun. 12, 354.
Angelis, K., Álvarez-Carretero, S., Dos Reis, M., Yang, Z., 2018. An evaluation of different partitioning strategies for Bayesian estimation of species divergence times. Syst. Biol. 67, 61–77.
Aubert, J., Legal, L., Descimon, H., Michel, F., 1999. Molecular phylogeny of swallowtail butterflies of the tribe Papilionini (Papilionidae, Lepidoptera). Mol. Phylogenet. Evol. 12, 156–167.
Ashlock, P.D., Mayr, E., 1991. Principles of Systematic Zoology, second ed. MacGraw-Hill. Inc., New York.
Ayres, D.L., Darling, A., Zwickl, D.J., et al., 2012. BEAGLE: an application programming interface and high-performance computing library for statistical phylogenetics. Syst. Biol. 61, 170–173.
Baele, G., Li, W.L.S., Drummond, A.J., Suchard, M.A., Lemey, P., 2013. Accurate model selection of relaxed molecular clocks in Bayesian phylogenetics. Mol. Biol. Evol. 30, 239–243.
Blaimer, B.B., Ward, P.S., Schultz, T.R., Fisher, B.L., Brady, S.G., 2018. Paleotropical diversification dominates the evolution of the hyperdiverse ant tribe Crematogastrini (Hymenoptera: Formicidae). Insect Systematics and Diversity 2, 3.
Brake, I. and Bächli, G. (2013) Drosophilidae (Diptera). *World Catalogue of Insects, Volume 9*. Brill, Leyde.
Britton, T., Anderson, C.L., Jacquet, D., Lundqvist, S., Bremer, K., 2007. Estimating divergence times in large phylogenetic trees. Syst. Biol. 56, 741–752.
Cabusas, J.V.B., Barrion-Dupo, A.L.A., Latina, R.A., Caoili, B.L., 2020. Cytochrome c oxidase subunit I and II (COI and COII) sequences support recognition of *Papilio chikae* Igarashi and *Papilio hermeli* Nuyda as distinct species. Philippine Journal of Science 150, 111–119.
Cantino, P.D., De Queiroz, K. (Eds.), 2020. PhyloCode: A Phylogenetic Code of Biological Nomenclature. CRC Press.
Carpenter, F.M., 1992. Treatise on Invertebrate Paleontology, Part R, Arthropoda 3–4. Geological Society of America, Boulder, Colorado.
Caterino, M.S., Sperling, F.A.H., 1999. *Papilio* phylogeny based on mitochondrial cytochrome oxidase I and II genes. Mol. Phylogenet. Evol. 11, 122–137.
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.
Chernomor, O., von Haeseler, A., Minh, B.Q., 2016. Terrace aware data structure for phylogenomic inference from supermatrices. Syst. Biol. 65, 997–1008.
Collins, N.M., Morris, M.G., 1985. Threatened Swallowtail Butterflies of the World. The IUCN Red Data Book, Cambridge.
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., 2013a. Global biogeographic pattern in swallowtail diversification demonstrates alternate colonization routes in Northern and Southern Hemispheres. J. Biogeogr. 40, 9–23.
Condamine, F.L., Toussaint, E.F.A., Cotton, A.M., Genson, G.S., Sperling, F.A.H., Kergoat, G.J., 2013b. Fine-scale biogeographic and temporal diversification processes of peacock swallowtails (*Papilio* subgenus *Achillides*) in the Indo-Australian Archipelago. Cladistics 29, 88–111.
Condamine, F.L., Nagalingum, N., Marshall, C.R., Morlon, H., 2015. Origin and diversification of living cycads: A cautionary tale on the impact of the branching process prior on Bayesian molecular dating. BMC Evol. Biol. 15, 65.
Condamine, F.L., Nabholz, B., Clamens, A.-L., Dupuis, J.R., Sperling, F.A.H., 2018a. Mitochondrial phylogenomics, the origin of swallowtail butterflies, and the impact of the number of clocks in Bayesian molecular dating. Syst. Entomol. 43, 460–480.
Condamine, F.L., Rolland, J., Höhna, S., Sperling, F.A.H., Sanmartín, I., 2018b. 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., Rolland, J., Morlon, H., 2019. Assessing the causes of diversification slowdowns: temperature-dependent and diversity-dependent models receive equivalent support. Ecol. Lett. 22, 1900–1912.
Cornell, H.V., 2013. Is regional species diversity bounded or unbounded? Biol. Rev. 88, 140–165.
Cotton, A.M., Nakae, M., 2020. Replacement name for *Eques* Kirby, 1896 (Lepidoptera: Papilionidae). Butterflies 84, 47.
Cotton, A.M., Racheli, T., 2007. A preliminary annotated checklist of the Papilionidae of Laos with notes on taxonomy, phenology, distribution and variation (Lepidoptera, Papilionoidea). Fragmenta Entomologica 38, 279–378.
Deshmukh, R., Baral, S., Kuwalekar, M., Kizhakke, A.G., and Kunte K. (2022) Reproductive barriers and genomic hotspots of adaptation during allopatric species divergence. *bioRxiv*, https://doi.org/10.1101/2022.03.11.483945.
Dorchin, A., López-Uribe, M.M., Praz, C.J., Griswold, T., Danforth, B.N., 2018. Phylogeny, new generic-level classification, and historical biogeography of the *Eucera* complex (Hymenoptera: Apidae). Mol. Phylogenet. Evol. 119, 81–92.
Douady, C.J., Delsuc, F., Boucher, Y., Doolittle, W.F., Douzery, E.J.P., 2003. Comparison of Bayesian and maximum likelihood bootstrap measures of phylogenetic reliability. Mol. Biol. Evol. 20, 248–254.
Drummond, A.J., Bouckaert, R.R., 2015. Bayesian evolutionary analysis with BEAST. Cambridge University Press, Cambridge.
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., 2020. Phylogenomic test of mitochondrial clues to archaic ancestors in a group of hybridizing swallowtail butterflies. Mol. Phylogenet. Evol. 152, 106921.
Durden, C.J., Rose, H., 1978. Butterflies from the middle Eocene: the earliest occurrence of fossil Papilionidae. Pearce-Sellards Series Texas Memorial Museum 29, 1–25.
Economo, E.P., Narula, N., Friedman, N.R., Weiser, M.D., Guénard, B., 2018. Macroecology and macroevolution of the latitudinal diversity gradient in ants. Nat. Comm. 9, 1778.
Engel, M.S., 2015. Insect evolution. Curr. Biol. 25, R868–R872.
Evans, W.H., 1923. The identification of Indian butterflies. J. Bombay Nat. Hist. Soc. 29, 230–260.
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.
Fujioka, T., Tsukiyama, H. and Chiba, H. (1997) Japanese Butterflies and Their Relatives in the World. Vol. 1. Shuppan Geijutsu Sha, Tokyo. 200 + 304 pp., 162 pl.
Garnett, S.T., Christidis, L., Conix, S., Costello, M.J., Zachos, F.E., Bánki, O.S., Bao, Y., Barik, S.K., Buckeridge, J.S., Hovern, D., Lien, A., Montgomery, N., Nikolaeva, S., Pyle, R.L., Thomson, S.A., van Dijk, P.P., Whalen, A., Zhang, Z.-Q., Thiele, K.R., 2020. Principles for creating a single authoritative list of the world's species. PLoS Biol. 18, e3000736.
Gernhard, T., 2008. The conditioned reconstructed process. J. Theor. Biol. 253, 769–778.
Grimaldi, D., Engel, M.S., 2005. Evolution of the Insects. Cambridge University Press, Cambridge.
Guindon, S., Dufayard, J.F., Lefort, V., Anisimova, M., Hordijk, W., Gascuel, O., 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: Assessing the performance of PhyML 3.0. Syst. Biol. 59, 307–321.
Hancock, D.L., 1979. The systematic position of *Papilio anactus* Macleay (Lepidoptera: Papilionidae). Australian Entomological Magazine 6, 49–53.
Häuser, C.L., de Jong, R., Lamas, G., Robbins, R.K., Smith, C., Vane-Wright, R.I. 2005. Papilionidae revised GloBIS/GART species checklist (2nd draft). Available at: http://www.globis.insects-online.de/species&tree_h=.Papilionidae&sub=yes&tree_status=plus&tree_seq=0.
Hancock, D.L., 1983. Classification of the Papilionidae (Lepidoptera): a phylogenetic approach. Smithersia 2, 1–48.

Hancock, D.L., 1992. The *Princeps fuscus* complex (Lepidoptera: Papilionidae). Australian entomological Magazine 19, 1–8.
Hiura, I., Alagar, R.E., 1971. Studies on the Philippine butterflies chiefly collected by the co-operative survey by the Osaka Museum of Natural History and the National Museum of the Philippines, 1969. Part 1: Papilionidae. Bulletin of the Osaka Museum of Natural History 24, 29–44.
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.
Hölldobler, B., Wilson, E.O., 1990. The Ants. Harvard University Press, Harvard.
Huelsenbeck, J.P., Larget, B., Alfaro, M.E., 2004. Bayesian phylogenetic model selection using reversible jump Markov chain Monte Carlo. Mol. Biol. Evol. 21, 1123–1133.
Huxley, J. (1976) The coloration of *Papilio zalmoxis* and *P. antimachus*, and the discovery of Tyndall blue in butterflies. Proc. Royal Soc. Lond. Ser. B, Biol. Sci., 193, 441–453.
Iczn, 1999. International Code of Zoological Nomenclature, 4th edition. xxix, 306 pp. The International Trust for Zoological Nomenclature, London.
Iczn, 2012. Amendment of Articles 8, 9, 10, 21 and 78 of the International Code of Zoological Nomenclature to expand and refine methods of publication. ZooKeys 219, 1–10.
Igarashi, S. (1979) *Papilionidae and their early stages*. Vol. 1, 219 pp.; Vol. 2, [xvi] pp., 1-223 pls + 1-32 pls + 1-102 pls. Kodansha, Tokyo.
Igarashi, S., 1984. The classification of the Papilionidae mainly based on the morphology of their immature stages. Trans. Lepidopterist Soc. Japan 34, 41–96.
Jablonski, D., Roy, K., Valentine, J.W., 2006. Out of the tropics: evolutionary dynamics of the latitudinal diversity gradient. Science 314, 102–106.
de Jong, R., 2007. Estimating time and space in the evolution of the Lepidoptera. Tijdschrift voor Entomologie 150, 319–346.
de Jong, R., 2017. Fossil butterflies, calibration points and the molecular clock (Lepidoptera: Papilionoidea). Zootaxa 4270, 1–63.
Jordan, H.E.K., 1896. On mechanical selection and other problems. Novitates Zoologicae 3, 426–525.
Jordan, H.E.K. (1908-09) Family: Papilionidae, Swallowtails. In: Seitz, A. (Ed.), Die Gross-Schmetterlinge der Erde. Vol. 9, Fauna indoaustralica. Stuttgart, Alfred Kernen, pp. 11–109, 112.
Joshi, J. and Kunte, K. (2022) Polytypy and systematics: diversification of *Papilio* swallowtail butterflies in the biogeographically complex Indo-Australian Region. *bioRxiv*, doi.org/10.1101/2022.03.23.485569.
Kalyaanamoorthy, S., Minh, B.Q., Wong, T.K., Von Haeseler, A., Jermiin, L.S., 2017. ModelFinder: fast model selection for accurate phylogenetic estimates. Nat. Methods 14, 587–589.
Katoh, K., Standley, D.M., 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol. Biol. Evol. 30, 772–780.
Kizhakke, A.G. and Kunte, K. (2022) Mimicry promotes morphological diversification but not speciation in mimetic swallowtail butterflies. bioRxiv, https://doi.org/10.1101/2022.11.05.515277.
Kunte, K., 2009. The diversity and evolution of Batesian mimicry in *Papilio* swallowtails butterflies. Evolution 63, 2707–2716.
Kunte, K., Shea, C., Aardema, M.L., Scriber, J.M., Juenger, T.E., Gilbert, L.E., Kronforst, M.R., 2011. Sex chromosome mosaicism and hybrid speciation among tiger swallowtail butterflies. PLoS Genet. 7, e1002274.
Kunte, K., Zhang, W., Tenger-Trolander, A., et al., 2014. *doublesex* is a mimicry supergene. Nature 507, 229–232.
Lanfear, R., Frandsen, P.B., Wright, A.M., Senfeld, T., Calcott, B., 2017. PartitionFinder 2: new methods for selecting partitioned models of evolution for molecular and morphological phylogenetic analyses. Mol. Biol. Evol. 34, 772–773.
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., 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.
Lewis, D.S., Sperling, F.A.H., Nakahara, S., Cotton, A.M., Kawahara, A.Y., Condamine, F. L., 2015. Role of Caribbean Islands in the diversification and biogeography of Neotropical *Heraclides* swallowtails. Cladistics 31, 291–314.
Li, H.T., Yi, T.S., Gao, L.M., et al., 2019. Origin of angiosperms and the puzzle of the Jurassic gap. Nat. Plants 5, 461–470.
Maddison, W.P., Maddison, D.R., 2021. Mesquite: a modular system for evolutionary analysis. Version 3, 70. http://www.mesquiteproject.org.
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.
Mazet, N., Morlon, H., Fabre P.-H and Condamine F.L. (2023) Estimating clade-specific diversification rates and palaeodiversity dynamics from reconstructed phylogenies. *bioRxiv*, https://doi.org/10.1101/2022.05.10.490920.
Meseguer, A.S., Lobo, J.M., Cornuault, J., Beerling, D., Ruhfel, B.R., Davis, C.C., Jousselin, E., Sanmartín, I., 2018. Reconstructing deep-time palaeoclimate legacies in the clusioid Malpighiales unveils their role in the evolution and extinction of the boreotropical flora. Glob. Ecol. Biogeogr. 27, 616–628.
Miller, J.S., 1987. Phylogenetic studies in the Papilioninae (Lepidoptera: Papilionidae). Bull. Am. Mus. Nat. Hist. 186, 365–512.
Miller, M.A., Schwartz, T., Pickett, B.E., et al., 2015. A RESTful API for access to phylogenetic tools via the CIPRES science gateway. Evol. Bioinforma. 11, 43–48.
Minh, B.Q., Schmidt, H.A., Chernomor, O., Schrempf, D., Woodhams, M.D., Von Haeseler, A., Lanfear, R., 2020. IQ-TREE 2: New models and efficient methods for phylogenetic inference in the genomic era. Mol. Biol. Evol. 37, 1530–1534.
Moen, D., Morlon, H., 2014. Why does diversification slow down? Trends Ecol. Evol. 29, 190–197.
Morley, R.J., 2011. Cretaceous and Tertiary climate change and the past distribution of megathermal rainforests. In: Bush, M.B., Flenley, J.-.-R. (Eds.), Tropical Rainforest Responses to Climatic Change. Springer, Berlin, Heidelberg, pp. 1–31.
Morlon, H., Potts, M.D., Plotkin, J.B., 2010. Inferring the dynamics of diversification: a coalescent approach. PLoS Biol. 8, e1000493.
Morlon, H., Parsons, T.L., Plotkin, J., 2011. Reconciling molecular phylogenies with the fossil record. Proc. Natl. Acad. Sci. USA 108, 16327–16332.
Morlon, H., Lewitus, E., Condamine, F.L., Manceau, M., Clavel, J., Drury, J., 2016. RPANDA: an R package for macroevolutionary analyses on phylogenetic trees. Methods Ecol. Evol. 7, 589–597.
Munroe, E., 1961. The classification of the Papilionidae (Lepidoptera). Canadian Entomological: Supplement 17, 1–51.
Nakae, M., 2021a. Papilionidae of the World. Roppon-Ashi Entomological Books.
Nakae, M., 2021b. Taxonomic revision on five species of the family Papilionidae (Lepidoptera). Gekkan-Mushi 602, 35–39.
Nakahara, S., Lamas, G., Willmott, K., Espeland, M., 2020. Description of a new genus and species for a common and widespread Amazonian satyrine butterfly (Lepidoptera: Nymphalidae: Satyrinae: Satyrini). PeerJ 8, e10324.
O'Grady, P.M., DeSalle, R., 2018. Phylogeny of the genus *Drosophila*. Genetics 209, 1–25.
Owens, H., Lewis, D.S., Dupuis, J.R., et al., 2017. The latitudinal diversity gradient in New World swallowtail butterflies is caused by contrasting patterns of out-of- and into-the-tropics dispersal. Glob. Ecol. Biogeogr. 26, 1447–1458.
Owens, H.L., Lewis, D.S., Condamine, F.L., Kawahara, A.Y., Guralnick, R.P., 2020. Comparative phylogenetics of *Papilio* butterfly wing shape and size demonstrates independent hindwing and forewing evolution. Syst. Biol. 69, 813–819.
Page, M.G.P. and Treadaway, C.G. (2003b) Bauer & Frankenbach, Butterflies of the World. part 17. Papilionidae IX. Papilionidae of the Philippine Islands. Goecke & Evers, Keltern.
Page, M.G.P., Treadaway, C.G., 2003a. Papilionidae of the Philippine Islands I. Descriptions of New Subspecies and Changes in Classification. Bauer & Frankenbach, Butterflies of the World, Supplement 8, 1–6.
Palmer, D.H., Kronforst, M.R., 2020. A shared genetic basis of mimicry across swallowtail butterflies points to ancestral co-option of *doublesex*. Nat. Commun. 11, 6.
Pante, E., Puillandre, N., Viricel, A., Arnaud-Haond, S., Aurelle, D., Castelin, M., Chenuil, A., Destombe, C., Forcioli, D., Valero, M., Viard, F., Samadi, S., 2015. Species are hypotheses: Avoid connectivity assessments based on pillars of sand. Mol. Ecol. 24, 525–544.
Pound, M.J., Haywood, A.M., Salzmann, U., Riding, J.B., 2012. Global vegetation dynamics and latitudinal temperature gradients during the Mid to Late Miocene (15.97–5.33 Ma). Earth Sci. Rev. 112, 1–22.
Rabosky, D.L., 2009. Ecological limits and diversification rate: alternative paradigms to explain the variation in species richness among clades and regions. Ecol. Lett. 12, 735–743.
Rambaut, A., Drummond, A.J., Xie, D., Baele, G., Suchard, M.A., 2018. Posterior summarization in Bayesian phylogenetics using Tracer 1.7. Syst. Biol. 67, 901–904.
Ranwez, V., Harispe, S., Delsuc, F., Douzery, E.J.P., 2011. MACSE: Multiple Alignment of Coding SEquences accounting for frameshifts and stop codons. PLoS One 6, e22594.
Rebel, H. (1898) *Doritites bosniaskii*. Sitzungsberichte der akademie der wissenschaften. *Mathematischen-Naturwissenschaftliche Classe*. Abteilung 1: Mineralogie, Biologie, Erdkunde. Wien, **107**, 734–741, 745.
Ree, R.H., Sanmartín, I., 2018. Conceptual and statistical problems with the DEC+ J model of founder-event speciation and its comparison with DEC via model selection. J. Biogeogr. 45, 741–749.
Reed, R.D., Sperling, F.A., 1999. Interaction of process partitions in phylogenetic analysis: an example from the swallowtail butterfly genus *Papilio*. Mol. Biol. Evol. 16, 286–297.
Rolland, J., Condamine, F.L., Beeravolu, C.R., Jiguet, F., Morlon, H., 2015. Dispersal is a major driver of the latitudinal diversity gradient of Carnivora. Glob. Ecol. Biogeogr. 24, 1059–1071.
Rolland, J., Condamine, F.L., 2019. The contribution of temperature and continental fragmentation to amphibian diversification. J. Biogeogr. 46, 1857–1873.
Ronquist, F., Teslenko, M., van der Mark, P., et al., 2012. MrBayes 3.2: Efficient Bayesian phylogenetic inference and model choice across a large model space. Syst. Biol. 61, 539–542.
Rothschild, W., 1895. A Revision of the Papilios of the Eastern Hemisphere, Exclusive of Africa. Novitates Zoologicae 2, 167–463.
Sauquet, H., Ho, S.Y., Gandolfo, M.A., Jordan, G.J., Wilf, P., Cantrill, D.J., Bayly, M.J., Bromham, L., Brown, G.K., Carpenter, R.J., Lee, D.M., Murphy, D.J., Sniderman, J.M. K., Udovicic, F., 2012. Testing the impact of calibration on molecular divergence times using a fossil-rich group: the case of *Nothofagus* (Fagales). Syst. Biol. 61, 289–313.
Scriber, J.M., Tsubaki, Y., Lederhouse, R.C., 1995. Swallowtail Butterflies: Their Ecology and Evolutionary Biology. Scientific Publishers, Gainesville, Florida.
Scudder, S.H., 1875. Fossil butterflies. *Memoirs of the American Association for the*. Adv. Sci. 1, XI, + 99.
Shiraiwa, K., Cong, Q., Grishin, N.V., 2014. A new *Heraclides* swallowtail (Lepidoptera, Papilionidae) from North America is recognized by the pattern on its neck. Zookeys 468, 85–135.
Sigward, J.D., Sutton, M.D., Bennett, K.D., 2018. How big is a genus? Towards a nomothetic systematics. Zool. J. Linn. Soc. 183, 237–252.
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.
Sperling, F.A.H., 1987. Evolution of the *Papilio machaon* species group in western Canada. Quaestiones Entomologicae 23, 198–315.

Sperling, F.A.H., 1991. Mitochondrial DNA phylogeny, speciation, and host-plant coevolution of *Papilio* butterflies. Cornell Univ, Ithaca, New York. Ph.D. Dissertation,.

Sperling, F.A.H., 1993a. Mitochondrial DNA phylogeny of the *Papilio machaon* species group (Lepidoptera: Papilionidae). The Memoirs of the Entomological Society of Canada 165, 233–242.

Sperling, F.A.H., 1993b. Mitochondrial DNA variation and Haldane's rule in the *Papilio glaucus* and *Papilio troilus* species group. Heredity 71, 227–233.

Sperling, F.A.H., Harrison, R.G., 1994. Mitochondrial DNA variation within and between species of the *Papilio machaon* group of swallowtail butterflies. Evolution 48, 408–422.

Steinthorsdottir, M., Coxall, H.K., de Boer, A.M., Huber, M., Barbolini, N., Bradshaw, C. D., *et al.* (2021). The Miocene: The future of the past. Paleoceanogr. Paleoclimatol., 36, e2020PA004037.

Stork, N.E., 2018. How many species of insects and other terrestrial arthropods are there on Earth? Annu. Rev. Entomol. 63, 31–45.

Suchard, M.A., Lemey, P., Baele, G., Ayres, D.L., Drummond, A.J., Rambaut, A., 2018. Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. *Virus. Evolution* 4, vey016.

Talavera, G., Lukhtanov, V.A., Pierce, N.E., Vila, R., 2012. Establishing criteria for higher-level classification using molecular data: the systematics of *Polyommatus* blue butterflies (Lepidoptera, Lycaenidae). Cladistics 29, 166–192.

Talbot, G. (1939) *The Fauna of British India including Ceylon and Burma*. Butterflies Vol. I. Taylor & Francis, London. xxix + 600 pp.

Tennent, W.J., 1999. Notes on some Solomon Islands *Papilio* Linnaeus, with descriptions of four new subspecies (Lepidoptera: Papilionidae). Nachrichten des Entomologischen Vereins Apollo, N.F. 20, 207–230.

Tennent, W.J., Chandra, V., Müller, C.J., 2018. A remarkable new swallowtail butterfly from Fiji (Lepidoptera, Papilionidae). Nachrichten des Entomologischen Vereins Apollo 39, 53–61.

Teta, P., 2019. The usage of subgenera in mammalian taxonomy. Mammalia 83, 209–211.

Tsukada, E., Nishiyama, Y., 1980. Butterflies of the South East Asian Islands. Part I. Papilionidae. Plapac Co., Ltd, Tokyo, p. 457.

Tyler, H.A., Brown, K.S., Wilson, K., 1994. Swallowtail Butterflies of the Americas: A Study in Biological Dynamics, Ecological Diversity, Biosystematics, and Conservation. Scientific Publishers, Gainesville, Florida.

Vane-Wright, R.I., Raheem, D.C., Cieslak, A., Vogler, A.P., 1999. Evolution of the mimetic African swallowtail butterfly *Papilio dardanus*: Molecular data confirm relationships with *P. phorcas* and *P. constantinus*. Biol. J. Linn. Soc. 66, 215–229.

Wallace, A.R., 1865. On the phenomena of variation and geographical distribution as illustrated by the Papilionidae of the Malayan region. Transactions of the Linnean Society of London 25, 1–71.

Wiegmann, B.M., Trautwein, M.D., Winkler, I.S., Barr, N.B., Kim, J.-W., Lambkin, C., Bertone, M.A., Cassel, B.K., Bayless, K.M., Heimberg, A.M., Wheeler, B.M., Peterson, K.J., Pape, T., Sinclair, B.J., Skevington, J.H., Blagoderov, V., Caravas, J., Narayanan Kutty, S., Schmidt-Ott, U., Kampmeier, G.E., Thompson, F.C., Grimaldi, D.A., Beckenbach, A.T., Courtney, G.W., Friedrich, M., Meier, R. and Yeates, D.K. (2011) Episodic radiations in the fly tree of life. Proc. Natl. Acad. Sci. USA, 108, 5690–5695.

Winston, J.E., 1999. Describing Species: Practical Taxonomic Procedure for Biologists. Columbia University Press, Columbia, USA.

Wu, L.W., Yen, S.H., Lees, D.C., Lu, C.C., Yang, P.S., Hsu, Y.F., 2015. Phylogeny and historical biogeography of Asian *Pterourus* butterflies (Lepidoptera: Papilionidae): A case of intercontinental dispersal from North America to East Asia. PLoS One 10, e0140933.

Xie, W., Lewis, P.O., Fan, Y., Kuo, L., Chen, M.H., 2011. Improving marginal likelihood estimation for Bayesian phylogenetic model selection. Syst. Biol. 60, 150–160.

Xiong, T., Li, X., Yago, M., Mallet, J., 2022. Admixture of evolutionary rates across a butterfly hybrid zone. Elife 11, e78135.

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.

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

Zakharov, E.V., Smith, C.R., Lees, D.C., Cameron, A., Vane-Wright, R.I., Sperling, F.A.H., 2004b. Independent gene phylogenies and morphology demonstrate a Malagasy origin for a wide-ranging group of swallowtail butterflies. Evolution 58, 2763–2782.

Zhang, J., Cong, Q., Shen, J., Brockmann, E., Grishin, N.V., 2019. Genomes reveal drastic and recurrent phenotypic divergence in firetip skipper butterflies (Hesperiidae: Pyrrhopyginae). Proc. R. Soc. Lond.: Biol. Sci. 286, 20190609.

Zhang, W., Westerman, E., Nitzany, E., Palmer, S., Kronforst, M.R., 2017. Tracing the origin and evolution of supergene mimicry in butterflies. Nat. Commun. 8, 1269.

Ziegler, A., Eshel, G., Rees, P.M., Rothfus, T., Rowley, D., Sunderlin, D., 2003. Tracing the tropics across land and sea: Permian to present. Lethaia 36, 227–254.

Appendix III

This article has been submitted to *Insect Systematics and Diversity*

Museomics reveals patterns of within-species divergence and diversification in *Troides* birdwings (Lepidoptera: Papilionidae)

Corné F. H. van der Linden[1,3], Eliette L. Reboud[2], Emmanuelle Chevalier[2], Marcel Eurlings[3], Théo Léger[4], Stephen L. Sutton[5], Patrick Verbaarschot[1], Rob de Vos[3], Niklas Wahlberg[6], M. Eric Schranz[1], Fabien L. Condamine[2] & Sabrina Simon[1]

[1]*Biosystematics Group, Wageningen University & Research, 6708PB Wageningen, The Netherlands*
[2] *Institut des Sciences de l'Evolution de Montpellier (Université de Montpellier | CNRS | IRD | EPHE), Place Eugène Bataillon, 34095 Montpellier, France*
[3] *Naturalis Biodiversity Centre, Leiden, The Netherlands*
[4]*Museum für Naturkunde, Leibniz-Institut fuer Evolutions- und Biodiversitaetsforschung, Invalidenstr. 43, 10115 Berlin, Germany*
[5] *Institute for Tropical Biology & Conservation, University Malaysia Sabah, Kota Kinabalu 88400, Malaysia*
[6] *Department of Biology, Lund University, SWE-22362 Lund, Sweden*

Corresponding author

Corné Ferdinand Hendrikus van der Linden, Biosystematics Group, Wageningen University & Research, Droevendaalsesteeg 1, 6708 PB Wageningen, The Netherlands & Naturalis Biodiversity Centre, Leiden, The Netherlands, corne@lindenature.com, corne.vanderlinden@wur.nl

Abstract

Sundaland, a subcontinent situated in the Indo-Malay Archipelago, is characterised by a dynamic climatic, geological, and geographical history and a correspondingly diverse biota. The distribution of this biota follows complex biogeographical patterns, with high levels of local endemism and frequent instances of hidden diversity – greater differentiation than expected based on phenotype. Despite their high species richness and complex distributions, there is a paucity of biogeographical studies on Sundaic insects. By applying modern "museomics" techniques to a species group of *Troides* birdwing butterflies, we generated an inter- and intraspecific dataset comprising 103 genes and 10,594 amino acid positions from museum specimens collected between the 1890's and 1990's. Using this dataset, we were able to investigate the species status, historical biogeography, and diversification timing within the Sundaland endemic *amphrysus*-group. Diversity within the group initially arose in the late Miocene and early Pliocene and a subsequent diversification took place within the different species complexes during the Pleistocene. Borneo was identified as the most-likely ancestral range of the *amphrysus*-group and the source for multiple westward expansions within Sundaland. We found genetic evidence of hidden species-level diversity within the species complexes *Troides cuneifera* Oberthür 1879 and *Troides miranda* Butler 1869, for which we recommend future taxonomic re-evaluation. Together, these results provide a perspective on evolutionary dynamics and the role of Pleistocene glaciations in shaping patterns of insect diversity in Sundaland.

Keywords

Biogeography, Hidden diversity, Phylogenomics, Sundaland, *Troides*

Introduction

Southeast Asia is one of the world's megadiverse regions, hosting about a quarter of global biodiversity despite representing only 4% of the Earth's land surface (Myers et al. 2000, Mittermeier et al. 2005, Woodruff 2010, Corlett 2014). Situated within the region, the Indo-Malay Archipelago is particularly well-known for its exceptional biodiversity and hosts two global terrestrial biodiversity hotspots (Myers et al. 2000). The tropical position, complex geological history and insularity of the Archipelago have contributed to the evolution of its megadiverse species assemblages, shaping patterns of high regional and local endemism in various clades (Woodruff 2010, Hall 2013, De Bruyn et al. 2014, Von Rintelen et al. 2017).

The western section of the Archipelago, located on the Sunda Shelf, persists as a highly fragmented assemblage of islands; however this island setting is historically rare. For most of its geological history since its origin around 80 million years ago (Ma), the Sunda Shelf formed an exposed subcontinental landmass known as Sundaland (Hall and Morley 2004, Hall 2009, 2013, Sarr et al. 2019). Oscillations in global sea levels flooded (interglacial) and exposed (glacial) the Sunda Shelf during the late Pleistocene, culminating in the permanently flooded state from ~19,000 years ago to the present (Voris 2000, Cannon et al. 2009, Woodruff 2010, De Bruyn et al. 2014, Sarr et al. 2019). During glaciations, montane communities of flora and fauna were able to track the cooling climate downslope and correspondingly occupy larger ranges, with the opposite occurring during interglacial periods (Hewitt 2000, Qu et al. 2011).

Late Pleistocene dynamics are particularly thought to have had a major influence on diversification in Sundaland, which has been corroborated by studies of various clades including freshwater fishes, birds, and mammals (Mason et al. 2019, Cros et al. 2020, Husson et al. 2020, Sholihah et al. 2020). Despite high species richness and complex patterns of intraspecific diversity, the biogeography of Sundaic insects remains understudied. Recently diversified lineages with complex patterns of regional endemism are particularly suitable for investigating the role of Pleistocene changes on diversification in Sundaland.

Birdwing butterflies of the well-known genus Troides Hübner 1819 (Lepidoptera: Papilionidae) are an excellent example of such a recently diversified insect lineage. These birdwings have diversified into 21 species and 72 subspecies across their range, which straddles the Indo-Malay Archipelago from the Asian mainland to New Guinea (Tsukada and Nishiyama 1982, Matsuka 2001, Ohya 2001, Häuser et al. 2005, Condamine et al. 2015). Like other birdwing genera (*Trogonoptera* and *Ornithoptera*), *Troides* are host plant specialists, feeding on species in the genera *Aristolochia* and *Thottea,* both of which belong to the birthwort family

(Aristolochiaceae) (Nishida et al. 1993, Matsuka 2001). *Troides* are generally regarded as edge-tolerant, canopy species, and are typically strong flyers with representatives in most terrestrial habitats, from coastal mangroves to montane cloud forests (Haugum and Low 1979). The diversity within the genus shows significant geographical structuring and regional endemism, with many species being represented by allopatric subspecies.

Within *Troides*, four Sundaland endemic species form a clade known as the "*amphrysus*-group", these are *Troides amphrysus* Cramer 1782, *T. andromache* Staudinger 1892, *T. cuneifera* Oberthür 1879 and *T. miranda* Butler 1869 (Matsuka 2001, Condamine et al. 2015). These four species show significant intraspecific variation, with each species comprising several subspecific lineages, and we correspondingly refer to the four species as "species complexes" hereafter. The *amphrysus*-group has 14 "major" subspecies, which are the subspecies found on the four large Sundaland landmasses: Borneo, Sumatra, Java and the Malay Peninsula (Tsukada and Nishiyama 1982, Matsuka 2001). These subspecies have mostly been described during the late 19th century, often solely on the basis of wing color patterns in combination with allopatric distributions (Staudinger 1892, Fruhstorfer 1898). Despite their popularity among naturalists over the past ~150 years, taxonomic and systematic uncertainties persist within the *amphrysus*-group. Their scattered distribution and large morphological variation are an indication of the diversity that is found within each lineage, some of which may not be captured in current subspecies designations. Altogether, the *amphrysus*-group provides a suitable model for the study of within- and between-species diversification under the influence of late Pleistocene glaciations in Sundaland and allows the investigation of hidden diversity in an understudied biodiversity hotspot. The wealth of specimens present in natural history collections provides a supurb opportunity to address such research questions.

Natural history collections harbor the most complete catalogue of Earth's biodiversity and are therefore key to investigating the origins of biodiversity (Bakker et al. 2020, Card et al. 2021). The genomic potential of natural history collections is increasingly being unlocked through rapidly evolving high-throughput sequencing techniques combined with increasingly efficient bioinformatic pipelines for analysing low-coverage fragmented genomes (Bakker et al. 2020, Call 2020, Card et al. 2021, Raxworthy and Smith 2021). Museomics, the use of high-throughput sequencing to obtain genomic data from museum specimens (Call 2020, Twort et al. 2021), is able to provide unprecedented genomic insights into evolutionary histories and intraspecific diversity contained in natural history collections.

In this study, we make use of *Troides* museum specimens collected over the last ~130 years to carry out a fine-scale biogeographical study of the clade and to genetically investigate

numerous subspecies. With phylogenomic insights, we clarify morphological subspecies designations made by 19th century naturalists, test for hidden diversity and investigate the role of late Pleistocene glaciation dynamics in the diversification of Sundaland butterflies.

Materials and Methods

Taxon sampling

Specimens were obtained from Naturalis Biodiversity Centre (RMNH, Leiden, The Netherlands) and the Museum für Naturkunde (MfN, Berlin, Germany). Multiple specimens per subspecies were included where possible to account for difficulties in extracting degraded DNA from old specimens and to enable the detection of intraspecific variation. Accordingly, three to four specimens per subspecies were sampled within the *amphrysus*-group. In addition, one to two individuals were sampled for each of the three selected outgroup species. This resulted in a total sampling of 34 specimens, covering 12 out of 14 "major" *amphrysus*-group subspecies. Although we attempted to fully cover all *amphrysus*-group subspecies of the Greater Sunda islands and the Malay Peninsula, we unfortunately were unable to incorporate *T. andromache nishikawai* Kobayashi, 1992 and *T. miranda hayamii* Kobayashi, 1991, both rare subspecies occurring in southern Borneo. "Minor" subspecies, with restricted ranges on small islands along the coasts of the Greater Sunda Islands, were omitted as these subspecies are most likely derived from the subspecies present on the nearest larger landmass. This particularly applies to the numerous described subspecies of *T. amphrysus* on islands adjacent to Sumatra, e.g., *T. a. simeuluensis* (Ohya, 1985).

Specimen ages spanned across a hundred-year period: the oldest specimen was collected in 1892 (*T. amphrysus flavicollis*, specimen with DNA voucher MFNLEP329) – around the time that the first subspecies in the group were being described – and the youngest specimen was collected in 1990 (*T. cuneifera cuneifera*, specimen RMNH.INS.1047464). Different preservation methods were present within the selected taxa, although exact details are generally unclear for older specimens. Both papered specimens and pinned specimens were included in the sampled taxa. These are likely to have experienced different levels of degradation, since pinned specimens may have been relaxed prior to pinning, with this additional hydration potentially leading to additional DNA degradation. A complete overview of the specimens (collection locality, collection year and preservation method) is provided in Table 1.

Table 1. Museomics specimens accessed for this study. Indicated are sample number, specimen ID, collection locality, biogeographical sub-region, collection year or approximate timespan of collection and preservation method.

Group	Species	Subspecies	Sample ID	Specimen ID	Locality	Biogeographical Sub-region	Collection year	Preservation method
In	*Troides amphrysus* Cramer, 1779	*T. a. amphrysus* Fruhstofer, 1894	S4	RMNH.INS.1047702	Gunung Karang, Jonggol	Java	1977	Papered
		T. a. amphrysus Fruhstofer, 1894	S5	RMNH.INS.1047593	Gunung Karang, Jonggol	Java	1977	Papered
		T. a. ruficollis Butler, 1879	S16	MFNLEP328	Cameron Highlands	Malay Peninsula		Spread
		T. a. ruficollis Butler, 1879	S20	MFNLEP332	Malacca	Malay Peninsula		Spread
		T. a. euthydemus Fruhstorfer, 1913	S19	MFNLEP331	West Coast	Sumatra		Spread
		T. a. euthydemus Fruhstorfer, 1913	S28	RMNH.INS.326566	Tanjung Morawa	Sumatra		Spread
		T. a. euthydemus Fruhstorfer, 1913	S30	RMNH.INS.326548	Laut Tador	Sumatra	1950	Spread
		T. a. euthydemus Fruhstorfer, 1913	S40	RMNH.INS.326587	Tanjung Morawa	Sumatra		Spread
		T. a. euthydemus Fruhstorfer, 1914	S14	MFNLEP326		Sumatra		Spread
		T. a. flavicollis Duce, 1873	S17	MFNLEP329	Gunung Kinabalu	Borneo	1893	Spread
		T. a. actinodia Jordan, 1908	S18	MFNLEP330	Kutai	Borneo		Spread
		T. a. actinodia Jordan, 1909	S33	RMNH.INS.326653	South-East Borneo	Borneo		Spread
	Troides cuneifera Oberthür, 1879	*T. c. cuneifera* Oberthür, 1879	S12	RMNH.INS.1047464	Gunung Wayang,	Java	1990	Papered
		T. c. cuneifera Oberthür, 1879	S2	RMNH.INS.1048612	Oro-Oro Ombo	Java	1987	Papered
		T. c. cuneifera Oberthür, 1879	S3	RMNH.INS.1048603	Pujon Kidul	Java	1987	Papered
		T. c. cuneifera Oberthür, 1879	S38	RMNH.INS.326655	Tengger Mountains	Java		Spread
		T. c. paeninsulae Pendlebury, 1936	S24	MFNLEP336	Tapah	Malay Peninsula		Spread
		T. c. paeninsulae Pendlebury, 1936	S13	MFNLEP325	Tapah	Malay Peninsula		Spread
		T. c. sumatrana Hagen, 1894	S37	RMNH.INS.326654	Solok	Sumatra	1907	Spread
	Troides andromache Staudinger, 1892	*T. a. andromache* Staudinger, 1892	S6	MFNLEP256	Gunung Kinabalu	Borneo		Spread
		T. a. andromache Staudinger, 1892	S7	MFNLEP257	Gunung Kinabalu	Borneo		Spread
		T. a. andromache Staudinger, 1892	S9	RMNH.INS.1283435		Borneo		Spread
		T. a. marapokensis Fruhstorfer, 1898	S22	MFNLEP334	Gunung Merapok	Borneo		Spread
		T. a. marapokensis Fruhstorfer, 1898	S8	MFNLEP258	Gunung Merapok	Borneo		Spread
	Troides miranda Butler, 1869	*T. m. miranda* Butler, 1869	S36	RMNH.INS.326717	Gunung Kinabalu	Borneo	1907	Spread
		T. m. neomiranda Fruhstorfer, 1903	S15	MFNLEP327		Sumatra		Spread
		T. m. neomiranda Fruhstorfer, 1903	S21	MFNLEP333		Sumatra		Spread
		T. m. neomiranda Fruhstorfer, 1903	S31	RMNH.INS.326722	West Aceh, Seumanjam Meulaboh	Sumatra	1954	Spread
Out	*Troides magellanus* Felder, 1862	*Troides m. magellanus* Felder, 1862	S27	RMNH.INS.327160	Bilar	Philippines	1966	Spread
		Troides m. magellanus Felder, 1862	S29	RMNH.INS.327158	Bilar	Philippines	1966	Spread
	Troides hypolitus Cramer, 1775	*Troides h. hypolitus* Cramer, 1775	S11	RMNH.INS.1036109		Molucca		Papered
	Troides rhadamantus Staudinger, 1888	*Troides rhadamantus* Staudinger, 1888	S26	MFNLEP338	Mindoro	Philippines		Spread

In addition to the specimens from museum collections, reads from one recently collected specimen of each *amphrysus*-group species complex was incorporated in the dataset. These were downloaded from the publicly available sequence read archive (SRA BioProject PRJNA1013724) and belong to the following subspecies: *Troides amphrysus amphrysus* (SAMN37296348), *T. cuneifera cuneifera* (SAMN37296350), *T. andromache andromache (*SAMN37296349) and *T. miranda neomiranda* (SAMN37296351). Raw reads were filtered and paired with fastp using default options (Chen et al. 2018). Mitogenomes constructed by mapping reads to a refence mitogenome of *Ornithoptera richmondia* (Gray, 1853) (https://www.ncbi.nlm.nih.gov/nuccore/NC_037869.1) using Geneious R10 (https://www.geneious.com) and annotated with MitoFinder (Allio et al. 2020b). Draft genomes for these specimens were assembled separately as described in Allio et al. (2020a). Four published draft genomes of outgroup taxa were also incorporated into the dataset, these were: *Trogonoptera brookiana* (Wallace, 1855), *Ornithoptera priamus* (Linnaeus, 1758), *O. richmondia* (Gray, 1853) and *Troides plateni* (Staudinger, 1889) (Allio et al. 2020a). The sequencing and assembly of these genomes and mitogenomes is described in detail in Allio et al. (2020a).

DNA extraction

DNA was extracted from a single hindleg per specimen. For specimens with missing hindlegs, mesothoracic- or forelegs were sampled. Samples were crushed in Eppendorf tubes using a clean scalpel. DNA extraction was carried out using the QIAamp DNA Micro kit (Qiagen) using the Qiagen protocol "Isolation of Genomic DNA from Tissues" modified by Twort et al. (2021) for museomics. These modifications were: (i) samples were incubated in lysis buffer overnight at 56°C, (ii) samples were briefly centrifuged, and the supernatant was subsequently carried forward, and (iii) during DNA elution, samples were incubated on the filter column in elution buffer for 20 minutes prior to centrifugation (Twort et al. 2021). Extractions were carried out in an "ancient DNA" facility (Naturalis Biodiversity Centre, Leiden), to minimise potential contamination from other DNA sources. Other measures included working inside a flow cabinet, UV-sterilisation of working surfaces and the use of filter pipette tips. DNA concentration was measured for concentration and fragment size distribution using High Sensitivity D5000 ScreenTape Assays on the Agilent 4150 TapeStation System. All samples fulfilled the fragment length distribution criteria for library preparation (<250 bp), allowing us to directly proceed with library preparation.

Library preparation and sequencing

Whole-genome libraries were prepared for sequencing following the protocol by Meyer and Kircher (2010) with modifications according to Twort et al. (2021). Initially, we performed blunt-end repair to ensure homogeneous samples. During blunt-end repair, samples were incubated with a master mix consisting of: Tango buffer (1X), dNTP's (100 µM), ATP (1 mM), T4 PNK (0.5 U/µL), T4 DNA Polymerase (0.1 U/µL) and MilliQ. Incubation was carried out for 15 min at 25°C followed by 5 min at 12°C. Following blunt-end repair, a clean-up reaction was performed using filter columns (MinElute Reaction Cleanup Kit, Qiagen) and the samples were rinsed through three steps, first using PB buffer and then twice with PE buffer. Samples were removed from the column by incubation with elutriation buffer for 5 min at 37°C. Afterwards, adapter ligation (P7 & P5, main constituents of the Oligo Hybridisation Mix) was performed using T4 DNA Ligase Buffer (1X), PEG-400 (5%), T4 DNA Ligase (5 U/µL), MilliQ and a freshly diluted Oligo Hybridisation Mix (10 pmol) and incubated for 30 min at room temperature. Clean-up reactions were again performed as described above. Overhanging fragments were then filled-in by incubation with a mix of Isothermal Reaction Buffer (1X), dNTP's (250 µM), Bst Polymerase and MilliQ at 37°C for 20 min, this reaction was subsequently terminated by incubating for 20 min at 80°C. Each sample was subsequently amplified in six separate PCR reactions following a standardised protocol. After an initial denaturation of two min at 95°C, samples were amplified by 15 cycles of the following: 15 sec at 95°C, 30 sec at 60°C and one min at 68°C. The number of cycles was adjusted based on evaluation of the initial PCR reaction, where the number of cycles was increased or decreased to achieve approximately equimolar concentrations per sample. Finally, all six reactions per sample were pooled and 15 µL of the pooled sample was sent off for sequencing on the Illumina NovaSeq 6000 system (Novogene, UK) with paired-end reads of 150 bp, at a depth of 1.5 - 3 Gb. A detailed description of the protocol published by Twort et al. (2021) can be accessed online (https://doi.org/10.6084/m9.figshare.12927500).

Data cleaning and genome assembly

Random ligation may occur during the adapter ligation step during library preparation, and subsequently result in chimaera formation (Willerslev and Cooper 2005). Although not detrimental to the reliability of sequencing information, reads must be treated as single-end to obtain accurate results (Rowe et al. 2011). As in other museomic studies (Call 2020, Twort et al. 2021), our samples consisted of highly degraded genomes, and were therefore treated as

single-end reads in subsequent analyses. We only retained forward reads for subsequent downstream analysis steps as these suffered the least from adapter contamination.

Reads were cleaned with Trimmomatic 0.39 (Bolger et al. 2014) using parameters that have been used on museomic samples: LEADING: 3, TRAILING: 3, remove reads below 30bp, quality measured using a sliding window of 4 bp and quality higher than 25 (Call et al. 2021, Twort et al. 2021). Adapter sequences were removed, and overrepresented sequences were investigated for the presence of contaminants using online nucleotide BLAST (https://blast.ncbi.nlm.nih.gov/Blast.cgi) and removed accordingly, this consisted only of primer contamination.

Nuclear and mitochondrial genomes were subsequently assembled separately. Mitochondrial genomes were assembled for each specimen using MitoZ 2.3 (Meng et al. 2019) using default settings with following modifications: read length 150, insert size 150 in "quick mode". *De novo* nuclear genomes were assembled using SPAdes 3.15.2 (Nurk et al. 2013) with k-mer size set to 21, 33 and 55, following similar *de novo* assemblies by Call et al. (2021) and Twort et al. (2021).

To provide a preliminary assessment of genome completeness, we used a local tblastn approach in BLAST+ 2.11.0 (Camacho et al. 2009) to examine the percentage of recovered protein sequences from the Benchmarking Universal Single-Copy Orthologs (BUSCO) "Insecta" and "Lepidoptera" databases (https://busco-data.ezlab.org/v5/data/lineages/). The reference protein database (insecta_odb10.2020-09-10 or lepidoptera_odb10.2020-08-05) was treated as the query and our *de novo* genomes as the reference genome. This allowed comparison of the completeness across our genomes by scoring the amount of query protein sequences detected out of the known amount of ortholog sequences in the Insecta (n = 1,367) and Lepidoptera (n = 5,286) databases.

Ortholog detection and alignment

Orthologous genes were obtained from *de novo* genomes using a two-tier approach. Initially, gene models were identified in the genomes using MESPA 1.3, which constructed gene models for proteins with confident amino acid identity matches (>70%) to a closely-related reference proteome (Neethiraj et al. 2017). The proteome of *Danaus plexippus* (Lepidoptera: Nymphalidae) (Zhan et al. 2011) was selected as a reference proteome for the gene model prediction (divergence time ~98 Ma (Kawahara et al. 2019), BUSCO proteome completeness score 93%). Protein sequences of gene models predicted by MESPA were used as input for

Orthograph 0.7.1 (Petersen et al. 2017) to detect orthologous genes. The reference ortholog database consisted of 3,429 orthologous single-copy protein coding genes from *D. plexippus*, *Bombyx mori* (Lepidoptera: Bombycidae) and *Tribolium castaneum* (Coleoptera: Tenebrionidae) (Kawahara et al. 2019). Putative orthologs were summarised using Orthograph and aligned using the L-INS-I algorithm in MAFFT 7.471 (Katoh and Standley 2013). Aligned sequences were subsequently filtered for sections of random similarity using Aliscore 2.0 (Kück et al. 2010), and ambiguously aligned or randomly similar sections were subsequently removed using Alicut 2.3 (Kück 2016). The remaining sequences were concatenated into a supermatrix using FASconCAT-G 1.05 (Kück and Longo 2014).

From this supermatrix two different datasets were constructed by evaluating tree likeness and information content of each partition with MARE 0.1.2-rc (Meyer et al. 2011). The three datasets were constructed as follows: (1) in the initial MARE run we retained all partitions that received MARE Information Criterion above a threshold of 0 (IC>0), whilst specifying for all taxa to be kept in the analysis (dataset hereafter referred to as the "reduced" dataset); (2) the reduced dataset was used in a second MARE run, where all taxa were retained whilst selecting final partitions by maximizing tree-likeness and information content (dataset hereafter referred to as the "optimized" dataset); and (3) a third dataset was constructed using the 13 mitochondrial protein-coding genes (dataset hereafter referred to as the "mitogenome"). The quality of the different datasets was evaluated using Alistat 1.12 (Figure S1) (Wong et al. 2020).

Phylogenetic reconstructions

Supermatrix approach

Best scoring partition-specific substitution and rate models (Chernomor et al. 2016) were identified using ModelFinder in IQ-TREE 2.1.3 (Kalyaanamoorthy et al. 2017, Minh et al. 2020) for each of the three datasets (reduced, optimized and mitogenome). ModelFinder was restricted to the following nuclear and mitochondrial amino acid substitution models: DCMut (Kosiol and Goldman 2005), JTT (Jones et al. 1992), LG (Le and Gascuel 2008), Poisson, PMB (Veerassamy et al. 2003), VT (Müller and Vingron 2000), WAG (Whelan and Goldman 2001), mtART (Abascal et al. 2007), mtInv (Le et al. 2017). Furthermore, one additional protein mixture model, LG4X (Le et al. 2012), was selected to account for FreeRate heterogeneity. Substitution models were tested with the default rate heterogeneity models (+G, +GC, +R, +I and combinations thereof: +I+G & +I+R) (Yang 1994, Gu et al. 1995, Yang 1995, Soubrier et

al. 2012). Multiple FreeRate categories were also tested for, by setting -cmin 4 and -cmax 15. Model quality was evaluated based on maximum likelihood (ML) and the final model was selected for each partition using the corrected Akaike Information Criterion (AICc, Hurvich and Tsai (1989)). Subsequently, the partitioned alignment and the best scoring partition model was used to initiate 50 likelihood tree-search analyses; 25 with a random starting tree and 25 guided using the maximum-parsimony tree obtained from ModelFinder. From this set the final maximum-likelihood tree was selected and branch support was assessed using a non-parametric bootstrap analysis with 100 replicates mapped onto the ML tree. The number of alternative topologies recovered by the 50 ML tree searches was evaluated using UniqueTree 1.9. Taxa with unstable phylogenetic placement – "rogue taxa" – were identified in the three supermatrix phylogenies using RogueNaRok (Aberer et al. 2013), which was accessed online (https://rnr.h-its.org/submit) and run on each phylogeny (reduced, optimized & mitogenome) using 100 bootstrap trees and the ML tree obtained during the supermatrix tree search.

Supertree approach

Using individual alignments from the optimized dataset, 103 gene trees were generated with partition-specific substitution and rate models using ModelFinder and IQ-TREE 2.1.3 (Kalyaanamoorthy et al. 2017, Minh et al. 2020). Substitution and rate model selection and gene tree reconstructions were carried out simultaneously by specifying IQ-TREE option -MFP. The ML gene trees were concatenated into a combined tree set which was taken forward and used to generate the multi-species coalescent (MSC) subspecies tree using ASTRAL 5.7.8 (ASTRAL-III) (Zhang et al. 2018). Subspecies tree robustness was assessed using ASTRAL local posterior probability (localPP) support (Sayyari and Mirarab 2016).

Estimation of divergence times

Bayesian inferences were performed in BEAST 2.6.6 (Bouckaert et al. 2019) to estimate a temporal framework for the inter- and intraspecific relationships. Using the BEAUti interface, "StarBeastWithSTACEYops" was selected to prepare the input data file for the analysis. This template combines StarBEAST2 (Ogilvie et al. 2017) with a modification on tree parameters from the BEAST2 package STACEY (Jones 2017), allowing for a more efficient tree search analysis. Since there is no fossil record available for birdwings (*Trogonoptera*, *Ornithoptera* and *Troides*), node calibrations could not be directly specified using fossil information. Therefore, we relied on divergence time estimates from a recent comprehensive phylogeny of Papilionidae containing ~400 species calibrated by three papilionid fossils from Allio et al.

(2021) to set informative node age calibrations. Prior age constraints were specified by setting median, minimal and maximal ages to equal those from the 95% highest posterior density (HPD) obtained in Allio et al. (2021). Age calibrations were specified for three nodes, namely: (1) the crown age of birdwings between 17.5018 and 51.9644 Ma, (2), the divergence of *Troides* and *Ornithoptera* between 13.7414 and 41.0473 Ma, and (3) the crown age of *Troides* between 9.4524 and 29.7289 Ma (Allio et al. 2021). Higher relationships recovered using the MSC topology derived from the species tree in ASTRAL were kept fixed throughout the analysis, whilst allowing node heights and ultrametricity to be scaled. A subset of "clock genes" was selected from the total dataset to ensure realistic computation times and to avoid over-parameterisation, similar to the "gene-shopping" approach proposed by Smith et al. (2018). The three genes with the highest coverage across all taxa were selected to date the topology, these were: mitochondrially encoded ATP synthase membrane subunit 6 (ATP6), cytochrome oxidase subunit 2 (COX2), and NADH-ubiquinone oxidoreductase chain 3 (ND3). A shared substitution and rate model was assigned to the three genes using the BEAST2 package OBAMA to select the best-fit substitution model and rate parameters (Bouckaert 2020). The tree model was set to the birth-death tree model, deemed the most suitable choice for molecular dating of a mixed dataset including inter- and intraspecific divergences (Ritchie et al. 2017). Under the tab "Taxon Sets" each outgroup specimen was specified to species level and each *amphrysus*-group taxon was specified to subspecies level (distinct phenotypes with allopatric ranges – Haugum and Low (1979)). A log-normal relaxed clock prior was specified for each separate gene tree (Drummond et al. 2006). We specified runs of 300 million Markov chain Monte Carlo (MCMC) generations, sampling every 10,000 generations, ultimately yielding 90,004 posterior trees after removing 25% of posterior as burn-in. Four independent analyses were carried out, and we assessed the convergence of each run by evaluating the effective sample size (ESS, >200) and posterior distributions using Tracer 1.7.2 (Rambaut et al. 2018). Posterior trees from the four separate runs were combined using LogCombiner 2.7.5 (Bouckaert et al. 2019), and one maximum-clade credibility tree was generated using TreeAnnotator 2.7.5 using mean node heights (Bouckaert et al. 2019).

Mitochondrial haplotypes

Haplotype networks of mitochondrial genes cytochrome oxidase subunit 1 (COI) and NADH-ubiquinone oxidoreductase chain 4 (ND4) were constructed to examine the geographical

variation in mitochondrial divergence within each *amphrysus*-group species. Haplotypes of each gene were inferred for each species complex and a haplotype network was constructed from randomised minimum spanning trees of each haplotype with the functions "haplotype" and "rmst" in the R package pegas (Paradis 2010). Subspecies range maps were drawn based on the range extent provided in Matsuka (2001) using QGIS 3.22.0 (QGIS.org 2022). Pairwise p-distances of COI and ND4 sequences were calculated between subspecies using the function "dist.hamming" in the R package phangorn (Schliep 2011) and visualised using the R package ggplot2 (Wickham 2016).

Biogeographical analysis

We estimated ancestral ranges using the R package BioGeoBEARS 1.1.1 (Matzke 2018) to study the historical biogeography of the *amphrysus*-group. Outgroups from the genera *Trogonoptera* and *Ornithoptera* and the species *T. plateni* and *T. rhadamantus* were omitted from the dataset for this analysis since they are from species groups that are incompletely represented by our taxon selection and including incompletely sampled clades may bias the biogeographical analyses (Matzke 2018). For this reason, only the two most closely related outgroups, *T. hypolitus* and *T. magellanus* were retained. Taxon ranges coded at the level of biogeographical sub-region: Borneo, Sumatra, Java, Malay Peninsula, the Philippines and Wallacea. Accordingly, the model comprised of six different states and the maximum number of possible states per ancestral node was set to two. Using these geographical states and our time-calibrated subspecies tree as input data, two biogeographical models were evaluated. These were the Dispersal-Extinction-Cladogenesis (DEC; (Ree and Smith 2008)) with and without the parameter + *J*, which models founder-event speciation (thought to be appropriate in an island context; Condamine et al. 2015). The best-fit model was selected by comparing AICc scores between models. All R-based analyses were performed in R 4.1.2 (R Core Team 2021).

Species delimitation analyses

Aligned *amphrysus*-group COI sequences were used as input data for distance-based species delimitation methods. Automated Barcode Gap Discovery (ABGD) was applied to the data via https://bioinfo.mnhn.fr/abi/public/abgd/abgdweb.html (Puillandre et al. 2012). In ABGD, the

prior range for maximum intraspecific divergence was set at P = 0.01, which provides a threshold most suitable for comparison across partition analyses (Puillandre et al. 2012, Pozzi et al. 2020). Additionally, several distance metrics were evaluated to ensure consistency of results (uncorrected p-distance, JC69, K80 corrected distances). To evaluate consistency, the analysis was performed with a minimum slope increase of X = 1 and X = 1.5. Additionally, a second single-locus distance-based delimitation algorithm, Assemble Species by Automated Partitioning (ASAP), was applied to the data (Puillandre et al. 2021). ASAP does not require prior knowledge of maximum intraspecific divergence (Puillandre et al. 2021). Again, we evaluated the effect of the available distance metrics (uncorrected p-distance, JC69, K80 corrected distances). ASAP was accessed through its web portal: https://bioinfo.mnhn.fr/abi/public/asap/asapweb.html.

Additionally a tree-based delimitation method, the Bayesian implementation of Poisson Tree Processes (bPTP) (Zhang et al. 2013), was carried out on the supermatrix phylogeny obtained with the mitogenome dataset. The analysis was performed on the bPTP webserver (https://species.h-its.org/). The number of MCMC generations was set to 100,000 generations, sampling to every 1,000 generations. Outgroup taxa were excluded from the analysis. Four independent analyses were performed, and convergence of the MCMC chains was assessed after discarding the initial 10% of the posterior as burn-in by visually examining the bPTP trace log.

Results

Genomic coverage

As expected, DNA extracts from *Troides* birdwing butterfly museum specimens were highly fragmented. Extracts consisted mostly of fragments <250 bp, at low concentrations of on average 3.53 ng/uL (0.123 - 9.21 ng/uL). Despite the fragmented nature of DNA, we were able to construct 31 *de novo* genomes from 34 museum specimens sequenced, including in- and outgroup taxa, with collection dates ranging from 1893 to 1990. Complete mitogenomes were constructed for all specimens with *de novo* genomes, plus one additional specimen for which only the mitogenome could be recovered (total n = 32). Cleaned unpaired reads ranged from 3,128,525 reads from the worst (S9, RMNH.INS.1283435) to 23,452,134 reads for the best specimen (S26, MFNLEP338). The *de novo* assemblies were incomplete; the average

percentage of Lepidoptera BUSCO (*n* = 5,286) identified was 44.89% in samples sequenced at 1.5 Gb raw data and 80.52% for those sequenced for 3 Gb raw data (**Fig. 1**, Mann-Whitney U test $p = 0.004$, Table S1). This pattern remained consistent when considering only specimens with known collection years, which could only be reliably recovered for approximately half of the specimens examined (16/31) (**Fig. 1,** Mann-Whitney U test $p = 0.015$, Table S1). Although the number of specimens sampled was higher for the 3 Gb coverage subset (*n* = 10 vs 6), the distribution of specimen ages did not statistically differ for the two sequencing subsets (**Fig. 1**, Mann-Whitney U test $p = 0.703$, Table S1). Sequencing deeper, for 3 Gb instead of 1.5 Gb raw data, resulted in higher scores of genome completeness. Most markedly so for old specimens collected in the 1890's, where genome completeness for specimens sequenced at 1.5 Gb yielded 15-20 % genomes completeness and those sequenced at 3 Gb in 60-70% genomes completeness (**Fig. 1,** Table S1). A complete overview of sample age, number of reads recovered, proportion trimmed, number of contigs, contig size distribution, mitogenome quality and BUSCO completeness metrics for each sampled specimen can be found in Table S1.

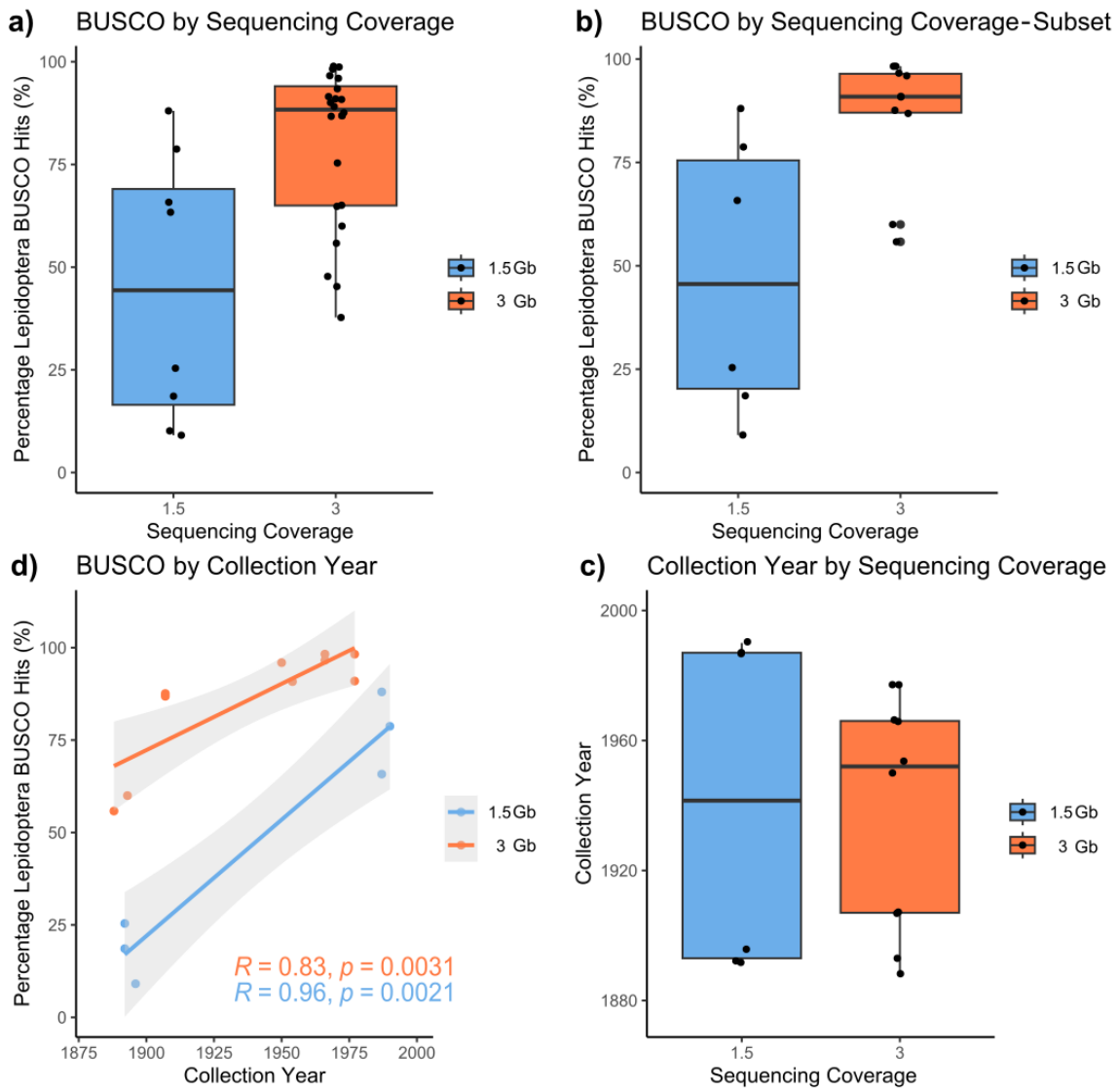


Figure 1. Genome coverage statistics of the *de novo* genomes generated from museum specimens. a) Percentage of Lepidoptera BUSCO's hits for specimens sequenced for 1.5 Gb and 3 Gb raw data. b) Percentage of Lepidoptera BUSCO's hits for specimens sequenced for 1.5 Gb and 3 Gb raw data for a subset of specimens for which collection years were available. c) Percentage of Lepidoptera BUSCO's hits for specimens in relation to collection year, depicted separately for specimens sequenced for 1.5 Gb and 3 Gb raw data. d) Distribution of collection years for specimens sampled for 1.5 Gb and 3 Gb raw data. As an indication of genome coverage, previously published high quality genomes of *Troides helena* (He et al. 2022) and *Troides oblongomaculatus* (Reboud et al. 2023) are approximately 330 Mb.

Ortholog datasets

Cleaning, ortholog identification and alignment quality checks of the *de novo* genomes yielded three different datasets. Mitogenome dataset consisted of the complete mitogenome comprising the 13 protein-coding genes for each of the 32 specimens. This dataset had a total alignment

length of 3,741 amino acid (AA) positions with 1.14% of missing data. The reduced dataset included 433 nuclear and 13 mitochondrial genes, the total alignment consisting of 36,069 AA positions from 446 orthologous sequences from 31 specimens, with 60.59% missing data. The optimized dataset consisted of 90 nuclear and 13 mitochondrial genes, together forming an alignment of 10,594 AA positions from 103 orthologous sequences from 32 specimens, with missing data accounting for 34.80% of sites. A complete overview of models and corresponding partition schemes used in subsequent analyses on the different datasets can be found in Table S2.

Dataset quality assessment

Unique tree topologies and rogue taxa

The number of unique tree topologies recovered during the 50 independent supermatrix tree searches showed variation across the three different datasets. The reduced dataset yielded six unique topologies, with 20 maximum agreement subtrees which contained 23 specimens out of a possible total of 40. The optimized dataset yielded three unique topologies and one maximum agreement subtree which contained 39 specimens out of a possible total of 40. The mitogenome dataset yielded 16 unique topologies, with five maximum agreement subtrees which contained 36 specimens out of a possible total of 40.

RogueNaRok identified rogue taxa in ML phylogenies obtained in all three datasets. Overall, rogue taxa were correctly assigned to the corresponding species complex level, whilst their intraspecific placement was resolved with low confidence (Fig. S2a, b, c). Specimens *T. amphrysus euthydemus* S28 and *T. amphrysus amphrysus* S4 were identified as rogue taxa in all three datasets. In the reduced dataset these two rogue taxa were accompanied by two additional rogue specimens: *T. amphrysus euthydemus* S14 and *T. amphrysus euthydemus* S19. The optimized dataset contained a total of three rogue taxa, consisting of the two shared rogue taxa together with *T. amphrysus flavicollis* S17. In the mitochondrial dataset the two shared rogue taxa were accompanied by *T. amphrysus amphrysus* S5 and *T. andromache andromache* S7, resulting in four rogue taxa. Rogue taxa identified by RogueNaRok were removed from each dataset, and tree search analyses were repeated following the described methodology. The number of unique trees subsequently resulted in only one topology being detected during the 50 independent supermatrix tree searches for the reduced and optimized datasets. In the mitogenome dataset the number of unique tree topologies decreased to four, with four samples

of *T. amphrysus euthydemus* S14, S19, S30 and S40 varying in position. These are all positioned within one intraspecific clade of the *T. amphrysus* species complex that shows very short branch lengths and low bootstrap values in the mitogenome dataset (Fig. S2c).

Conflicting nodes

Relationships within the *amphrysus*-group, particularly between the three species complexes *T. cuneifera*, *T. miranda* and *T. amphrysus*, showed topological disagreement across investigated datasets. Mitogenome and optimized datasets yielded ML trees with *T. miranda* and *T. amphrysus* as sister lineages, whilst the ML tree obtained using the reduced dataset assigned *T. cuneifera* and *T. amphrysus* as sister lineages. Therefore, the phylogenetic information content of all three datasets for this was evaluated for this incongruent node using Four-cluster Likelihood Mapping (FcLM) in IQ-TREE 2.1.3 (Strimmer and von Haeseler 1997, Minh et al. 2020). Furthermore, the reduced and optimized datasets were permuted using three different permutation schemes to uncover causal mechanism behind topological incongruence between datasets. The two datasets were permuted using three different permutation methods, these were: (1) homogenous permutation, in which all phylogenetic signal is removed and only missing and ambiguous positions remain in the alignment allowing the evaluation of heterogeneous and non-stationary patterns of missing and ambiguous sites; (2) standard permutation, where all non-ambiguous sites were replaced by randomly selected amino acids whilst keeping the distribution of missing and ambiguous sites in place, allowing evaluation of distribution of missing data; and (3) random permutation, where all sites were randomly permuted. FcLM across all three datasets unequivocally supported the sister lineage pair of *T. miranda* and *T. amphrysus* with *T. cuneifera* as the sister lineage of this species pair (Fig. S4). The origin of the contrasting topology of the reduced dataset with those of the mitogenome and optimized datasets could not be clearly revealed by permutation analyses (Fig. S4). Here, the signal could not be attributed to missing or ambiguous sites or to the distribution of such sites in the alignment, which is evident from the similar distribution of percentages to all three possible topologies in each scenario (Fig. S4). Ultimately, the optimized dataset was selected for the phylogenetic inferences and divergence time estimates since this dataset: (1) has the largest overall topological stability with the least variation in unique tree topologies, (2) yielded the largest maximum agreement subtree, which contained 39 out of 40 taxa examined (only excluding rogue taxon *T. amphrysus, euthydemus* S28), (3) consists of both nuclear and mitochondrial genes, and (4) recovered the most likely topology for the conflicting nodes.

Phylogenetic relationships

Supermatrix approach

We were able to infer interspecific relationships among the studied taxa with a high degree of confidence (**Fig. 2**). Relationships between and within outgroup taxa were well-resolved and were supported by high bootstrap percentage (BP) values (>95%). We recovered the *amphrysus*-group as a monophyletic lineage consisting of four divergent clades corresponding to the four species complexes *T. andromache*, *T. cuneifera*, *T. miranda* and *T. amphrysus* (BP=99%). *T. andromache* formed the earliest diverging group within the *amphrysus*-group and was confidently placed outside of the monophyletic clade consisting of *T. cuneifera*, *T. miranda* and *T. amphrysus* (BP=99%). *T. cuneifera* was recovered as the sister lineage to the clade containing *T. miranda* and *T. amphrysus*, albeit with weaker bootstrap support (BP=65%). *T. amphrysus* and *T. miranda* formed a sister lineage with high confidence (BP=100%).

Intraspecific relationships showed variable support across the species complexes (**Fig. 2**). Not all currently described subspecies were recovered as monophyletic using the supermatrix approach. Within *T. andromache*, subspecies *marapokensis* formed a weakly supported monophyletic group within the paraphyly of the nominate subspecies *andromache* (BP=60%). In *T. cuneifera*, specimens from subspecies *cuneifera* clustered together forming a well-supported monophyletic clade (BP=100%). However, subspecies *paeninsulae* was recovered as paraphyletic forming a well-supported monophyletic clade together with the *sumatrana* specimen (BP=100%). The two subspecies of *T. miranda* were reciprocally monophyletic, subspecies *neomiranda* was represented by multiple specimens in our dataset and formed a monophyletic sister clade to the single *miranda* specimen in the analysis (BP=100%). Intraspecific relationships within *T. amphrysus* showed the highest degree of incongruence with current subspecies taxonomy and none of the subspecies were recovered as monophyletic groups by the supermatrix analysis (**Fig. 2**). Overall, a large proportion of specimens could only be placed in the tree with very low to moderate confidence within *T. amphrysus* (BP ranging from 4 to 81%). A complete overview of bootstrap per node is provided in supplementary Fig. S2b.

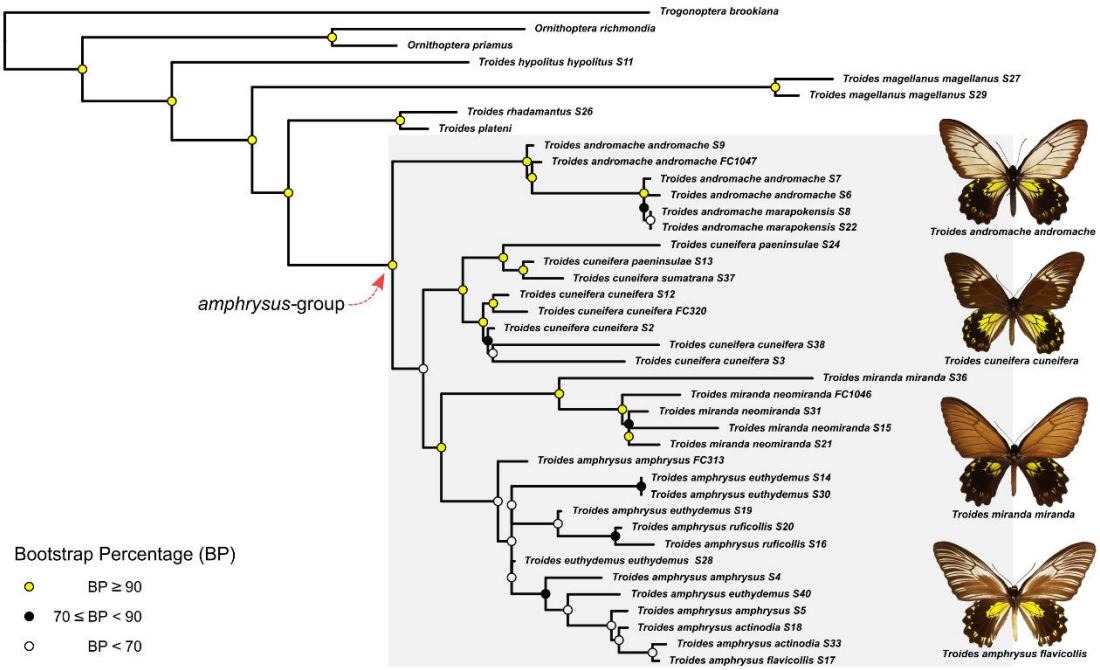


Figure 2. Maximum-likelihood phylogeny of the *amphrysus*-group and outgroup taxa, as inferred by IQTREE 2 from an amino acid alignment of 103 orthologous nuclear (n = 90) and mitochondrial (n = 13) sequences. Bootstrap support is indicated per node in percentage classes of high (>90%), moderate (70-90%) and low (<70%) support. Selected subspecies are shown, specimen figures adapted from the Natural History Museum (2014).

Supertree approach

The supertree topology reaffirmed the interspecific and intergeneric topology obtained using the supermatrix approach, although not all these relationships were recovered with high local posterior probability (localPP) (Fig. S3). Supertree and supermatrix approaches differed in topology of intraspecific relationships (**Fig. 2**, Fig. S3). As in the supermatrix approach, the supertree indicated low phylogenetic confidence for the exact placement of specimens at the subspecies level. Within *T. andromache,* subspecies were recovered as paraphyletic, however only with moderately high to low support (localPP = 0.5 – 0.91). Within *T. cuneifera*, all subspecies were recovered as monophyletic groups with relatively high support (localPP = 0.94 – 0.99). *T. miranda* subspecies were recovered as monophyletic, the well-supported clade (localPP = 1) consisting of *T. miranda neomiranda* was sister to that of *T. miranda miranda* with moderate support (localPP = 0.82). Intraspecific relationships within *T. amphrysus* were recovered with varying degrees of phylogenetic confidence from low to moderately high support (localPP = 0.38 – 0.9). Within this species complex, none of the subspecies were recovered as monophyletic clades. However, geographically proximate subspecies clustered together in the phylogeny. Both western Sundaland subspecies *T. amphrysus euthydemus* and *T. amphrysus ruficollis*, formed a monophyletic clade (localPP = 0.71). Both Bornean

subspecies, *T. amphrysus flavicollis* and *T. amphrysus actinodia* also formed a monophyletic clade (localPP = 0.67). Specimens belonging to the nominate subspecies, *T. amphrysus amphrysus*, were placed either as outgroup to the Bornean clade (localPP = 0.59) or to the western Sundaland clade (localPP = 0.56). The exception being specimen *T. amphrysus amphrysus* S4, placed as outgroup to the entire *T. amphrysus* clade with moderately high confidence (localPP = 0.9).

Misidentified specimens

Instances of misidentifications were revealed among the sampled specimens. Some of these were identified as suspicious cases *a priori*, as their locality information did not match with their subspecies assignment. Two specimens were confidently placed in the tree in positions in conflict with their collection labels. Label information of specimen S26 indicated it to be a *T. cuneifera paeninsulae,* known only from the Malay Peninsula, but its locality label read "Mindoro" which is situated in the Philippines. The phylogenetic analyses placed sample S26 as a sister lineage to the *Troides plateni* specimen with high support, revealing its phylogenetic identity as *T. rhadamantus* (**Fig. 2**, BP=100%). In other instances, misidentifications were revealed for specimens that occur in sympatry and are highly similar phenotypically. Specimen S14, was labelled as a *T. cuneifera sumatrana*, however it was placed confidently within *T. amphrysus* in the ML tree analyses (Fig. 2, BP=100%), and was subsequently re-identified as *T. amphrysus euthydemus*. Lastly, specimen S19 was labelled as *T. amphrysus ruficollis* whilst originating from the West coast of Sumatra, which is inside the known range of the phenotypically similar conspecific *T. amphrysus euthydemus*. Since this subspecies pair is virtually indistinguishable genetically, only phenotypic characters could be relied upon for re-identification. Subsequent re-evaluation of the phenotypic characters of the museum specimens confirmed the suspected re-identifications in all three cases.

Divergence time analyses

Bayesian divergence time analyses recovered age estimates for the higher relationships and intraspecific relationships of the previously obtained phylogenetic topology (**Fig. 3**). The four independent analyses converged after 300 million generations with high ESS values (ESS = 254.2 – 263.1). After discarding 25% of the posterior as burn-in and combining the separate runs, divergence time estimates were obtained from the combined posterior trees. The crown age of birdwings was estimated around 24.47 Ma (95% HPD = 17.07–33.17 Ma). The

divergence of the genera *Ornithoptera* from *Troides*, was estimated to have occurred around 16.80 Ma (95% HPD = 12.50–21.84 Ma). The most recent common ancestor of *Troides* was placed at 12.69 Ma (95% HPD = 9.20–16.80 Ma) and, subsequently, the *amphrysus*-group diverged from other *Troides* around 10.37 Ma (95% HPD = 7.03–14.21 Ma). Within the *amphrysus*-group, species complexes diverged approximately between 7 and 5 Ma. *T. andromache* was the first lineage to diverge from the other three species complexes around 7.00 Ma (95% HPD = 4.40–9.98 Ma). The two *T. andromache* subspecies lineages investigated, the northern *andromache* and the southern *marapokensis*, diverged around 0.26 Ma (95% HPD = 1.35e-06–0.78 Ma). *T. cuneifera* diverged from the monophyletic clade consisting of *T. miranda* and *T. amphrysus* around 5.57 Ma (95% HPD = 3.39–8.05 Ma). Javan subspecies *cuneifera* diverged from Sumatran and Peninsular counterparts 1.86 Ma (95% HPD = 0.49–3.53 Ma), which subsequently diverged 0.59 Ma (95% HPD = 8.06e-05–1.41 Ma) into the Sumatran *sumatrana* and the Peninsular *paeninsulae*. *T. miranda* split from *T. amphrysus* around 5.05 Ma (95% HPD = 2.92–7.48 Ma). The Bornean subspecies *miranda* diverged from the Sumatran *neomiranda* approximately 1.05 Ma (95% HPD = 0.19–1.99 Ma). The most recently diversified species complex within the group, *T. amphrysus*, split into Bornean and western Greater Sunda clades approximately 1.00 Ma (95% HPD = 0.22–1.88 Ma). Subsequently, the Bornean subspecies *actinodia* and *flavicollis* diverged approximately 0.44 Ma (95% HPD = 6.59e-06–1.06 Ma). Respectively, the Javan subspecies *amphrysus* diverged around 0.32 Ma (95% HPD = 0.06–0.65 Ma) from the Sumatran *euthydemus* and Peninsular *ruficollis* which diverged 0.17 Ma (95% HPD = 8.81e-04–0.40 Ma).

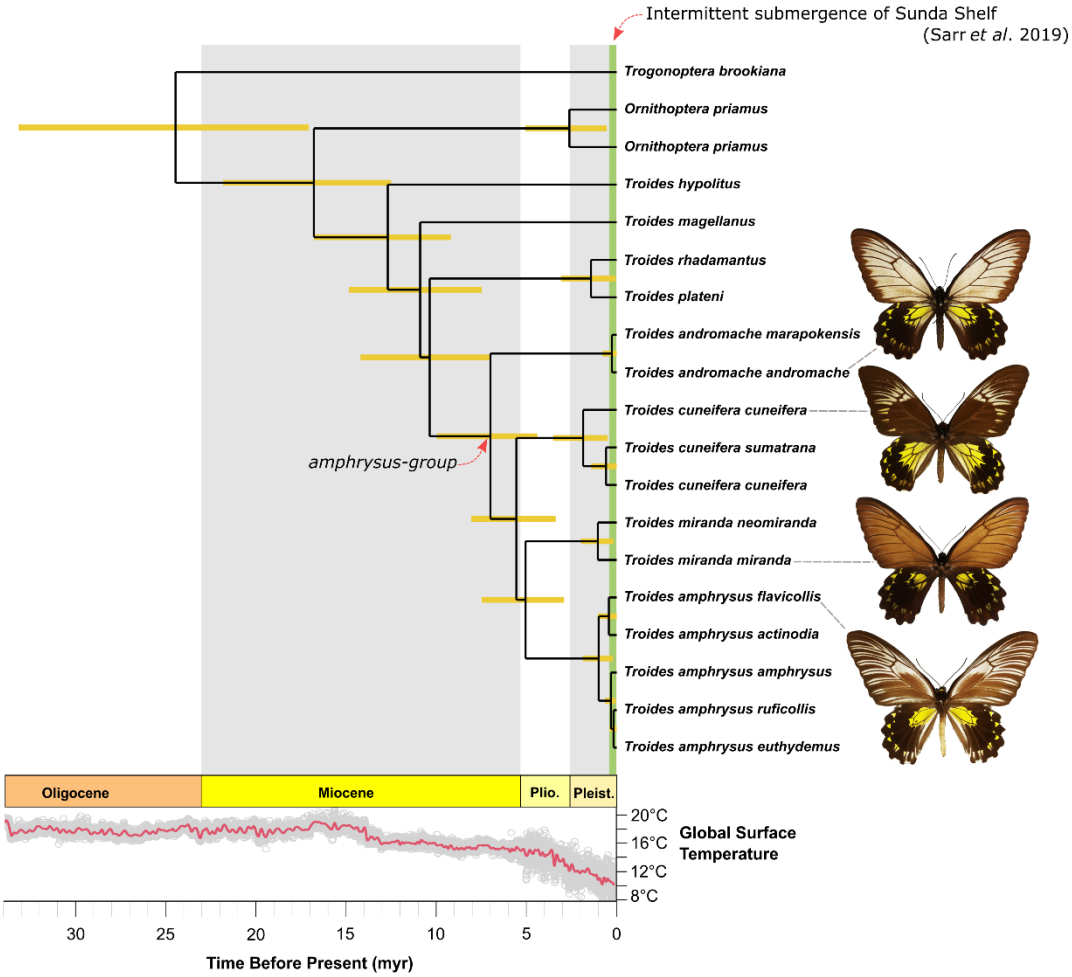


Figure 3. Time calibrated phylogeny of the *amphrysus*-group and selected outgroup taxa. Interspecific and intergeneric topology as inferred by ASTRAL from 103 orthologous nuclear (n = 90) and mitochondrial (n = 13) gene trees. Intraspecific relationships & time calibrations inferred with Bayesian Inference in StarBEAST 2. Bars depict 95% HPD intervals of divergence time posterior estimates. Scale indicated both as millions of years before the present and as a geologic time scale. Selected subspecies are shown, specimen figures adapted from the Natural History Museum (2014).

Mitochondrial haplotypes

Mitochondrial haplotype networks (**Fig. 4**) illustrate genetic differences and similarities in subspecies identity and geographical origin. Results from both genes examined (COI and ND4) depict similar patterns of intraspecific variation between subspecies pairs (**Fig. 4**). In general, haplotypes clustered closely together based on geographical proximity and subspecies affiliation. Mitochondrial divergence between subspecies pairs showed a high degree of variation across species complexes. Additionally, haplotype networks revealed a high degree of variation within subspecies, with subspecies represented by more than one sample either sorted into multiple (e.g., *T. miranda neomiranda*, **Fig. 4c**) or one single haplotype (e.g., *T.*

andromache marapokensis, **Fig. 4d**). Uncorrected p-divergences are discussed only for COI comparisons as ND4 divergences followed the same patterns (see Table S3).

Intraspecific COI variation between *T. amphrysus* subspecies reflected haplotype network patterns for this species complex, with pairwise uncorrected p-divergences generally indicating greatest divergence between subspecies from distant parts of the range (**Fig. 4a**, **Fig. 5**, Table S3). Western Greater Sunda subspecies (*T. amphrysus ruficollis*, *T. amphrysus euthydemus* and *T. amphrysus amphrysus*) showed 0.33 – 0.55% COI divergence with their Bornean counterparts (*T. amphrysus flavicollis* and *T. amphrysus actinodia*). Between subspecies occurring on the western Greater Sundas and the Malay Peninsula, comprising subspecies *T. amphrysus ruficollis, T. amphrysus euthydemus* and *T. amphrysus amphrysus*, we found COI divergence ranging 0.07 – 0.26%. Subspecies comparison from the two subspecies from Borneo, *T. amphrysus actinodia* and *T. amphrysus flavicollis*, revealed 0.33% COI divergence. Samples from subspecies *T. amphrysus ruficollis* (Malay Peninsula) and *T. amphrysus euthydemus* (Sumatra) were found to be closely affiliated and show complete overlap in ND4 and near complete overlap in COI haplotypes.

Mitochondrial haplotype networks of the species complex *T. cuneifera* show a separation of its subspecies into two distinct groups (**Fig. 4b**). One cluster, formed by subspecies *T. cuneifera paeninsulae* (Malay Peninsula) and *T. cuneifera sumatrana* (Sumatra), and another consisting of the nominate subspecies *T. cuneifera cuneifera* (Java). *T. cuneifera paeninsulae* and *T. cuneifera sumatrana* showed pairwise COI divergence of 0.45-0.52%. Comparison of the *T. cuneifera cuneifera* with both these two subspecies showed approximately four times higher COI divergence (2.4-2.6 %, **Fig. 5**, Table S3).

Mitochondrial haplotypes of the representatives of the *T. miranda* species complex cluster into two distinct groups consisting of the *T. miranda neomiranda* (Sumatra) and *T. miranda miranda* (northern Borneo) subspecies (**Fig. 4c**). Between this subspecies pair we recovered COI divergence of 2.02% (**Fig. 5**, Table S3).

The *T. andromache* species complex, here represented by *T. andromache marapokensis* and *T. andromache andromache*, showed different mitochondrial haplotypes occurring in relatively close geographical proximity (**Fig. 4d**). *T. andromache andromache* was found to be 0.19 – 0.33% divergent from *T. andromache marapokensis,* with samples of the latter represented by a single haplotype for both genes investigated (**Fig. 4d**, **Fig. 5**, Table S3). High haplotype diversity was recovered within *T. andromache andromache*, with levels of intra-subspecies divergence comparable to its divergence with the conspecific *T. andromache marapokensis* (**Fig. 4d**).

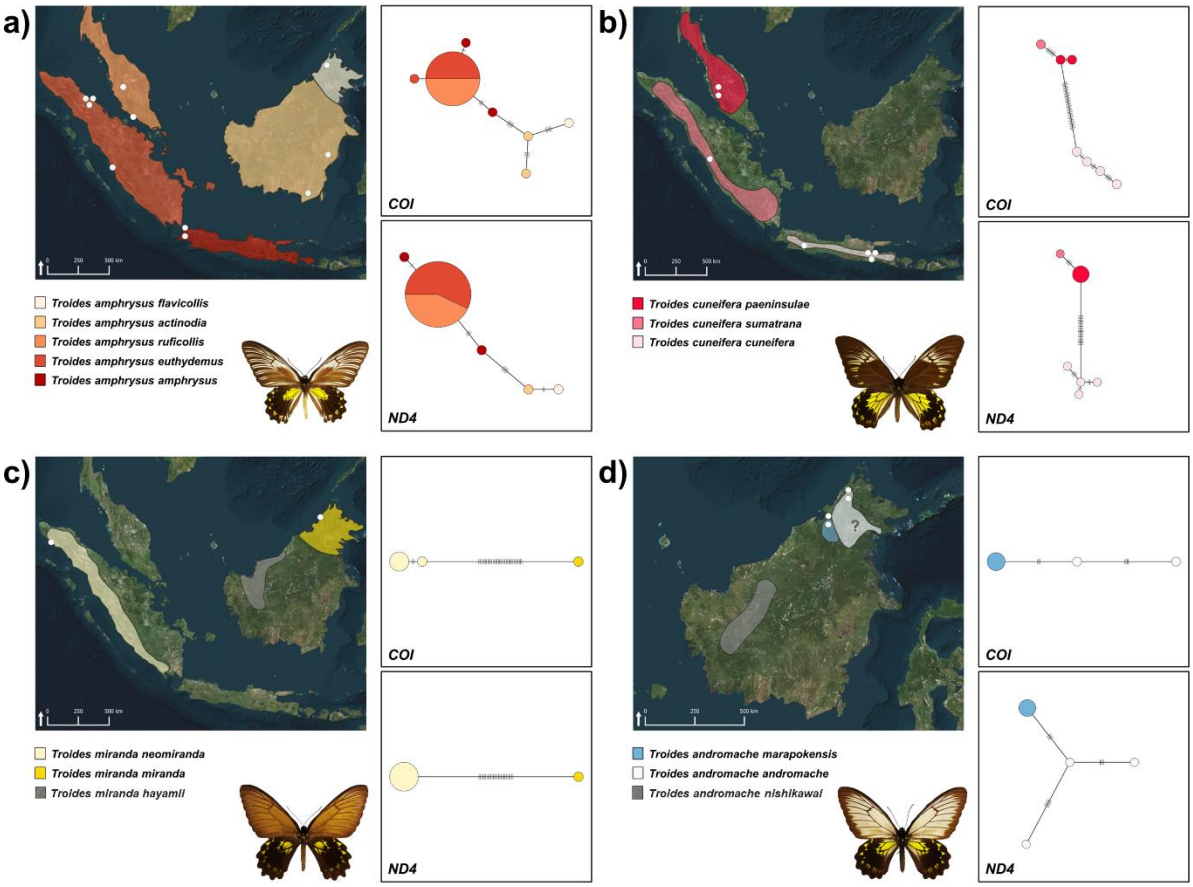


Figure 4. Minimum spanning networks of mitochondrial haplotypes found within each *amphrysus*-group species complex. Haplotypes are based on coding sequences of cytochrome oxidase subunit 1 (COI) and NADH-ubiquinone oxidoreductase chain 4 (ND4) sequences. Substitutions are depicted as vertical bars between different haplotypes. Circle colors confer subspecies identity and circle size is relative to the number of individuals belonging to that haplotype (ranging between 1 - 6). Maps depict approximate subspecies ranges based on Matsuka (2001) and sampling localities used in this study are indicated in white (based on collection sites listed in Table **1**). For reference, ranges of omitted subspecies *Troides andromache nishikawai* and *Troides miranda hayamii* are shown in grey). Selected subspecies are shown, specimen figures adapted from Natural History Museum (2014), maps produced in QGIS 3.22.0.

Ancestral range estimation

Model comparison based on AICc scores for the two tested models yielded DEC+*J* as the most likely model for the biogeographical history of the *amphrysus*-group (Table 2). Under the DEC+*J* model, the ancestors of the *amphrysus*-group most likely dispersed from an ancestral Philippine – Wallacean range into Sundaland (**Fig. 6a**). Borneo was recovered as the most likely ancestral range of the *amphrysus*-group and all interspecific nodes within the *amphrysus*-group (**Fig. 6a**).

The ancestral range of the earliest diverging *amphrysus*-group species complex, *T. andromache*, was estimated to lie in Borneo. The ancestral range of the *T. cuneifera* species complex was inferred on Java. The ancestral range of the clade consisting of subspecies *T.*

cuneifera paeninsulae and *T. cuneifera sumatrana* was recovered to lie on Sumatra, although the Malay Peninsula received a near-equal probability as the ancestral range for this node. Borneo was also recovered as the most-likely ancestral range of the *T. miranda* species complex. Likewise, Borneo was estimated to be the ancestral range of *T. amphrysus* species complex as well as that of its two Bornean subspecies *T. amphrysus flavicollis* and *T. amphrysus actinodia*. The ancestral range of the *T. amphrysus* clade occurring in the western Greater Sunda Islands was inferred lie on Java. The ancestral range of subspecies *T. amphrysus ruficollis* and *T. amphrysus euthydemus* was marginally attributed to Sumatra.

Table 2. Model comparison overview of the two ancestral range estimation models (DEC & DEC + *J*) evaluated using BioGeoBEARS for the *amphrysus*-group and selected outgroup taxa.

	LnL	Number of parameters	d	e	j	AICc
DEC	-35,17	2	0,015	2,00E-08	0	75,43
DEC+J	-22,09	3	1,00E-12	1,00E-12	0,26	52,58

Species delimitation

Both distance-based species delimitation methods detected six species-level lineages within the *amphrysus*-group from the COI alignment (**Fig. 6b**). ASAP reported six species-level partitions within the *amphrysus*-group, which were stable regarding distance metric choice (simple distance p = 3.01e-02, JC69 p = 2.80e-02, K80 p = 3.26e-02). At the recommended maximum intraspecific divergence prior P = 0.01, ABGD recovered six species-level partitions. Partitions recovered by ABGD were stable regarding distance metric choice (JC69 or K80) and the choice of minimum slope increase (X = 1 or 1.5). ABGD with simple distances did not reach the recommended maximum intraspecific divergence prior P = 0.01 and were therefore disregarded. The tree-based bPTP analysis yielded the same species-level clades, this result was consistent for both single-locus (COI) or whole mitogenome analyses (**Fig. 6b**). Overall, the three species delimitation methods delimited the same six monophyletic species-level lineages corresponding to: (1) *T. andromache*, (2) *T. cuneifera cuneifera*, (3) *T. cuneifera paeninsulae* + *T. cuneifera sumatrana*, (4) *T. miranda neomiranda*, (5) *T. miranda miranda*, and (6) *T. amphrysus* (**Fig. 6 b**).

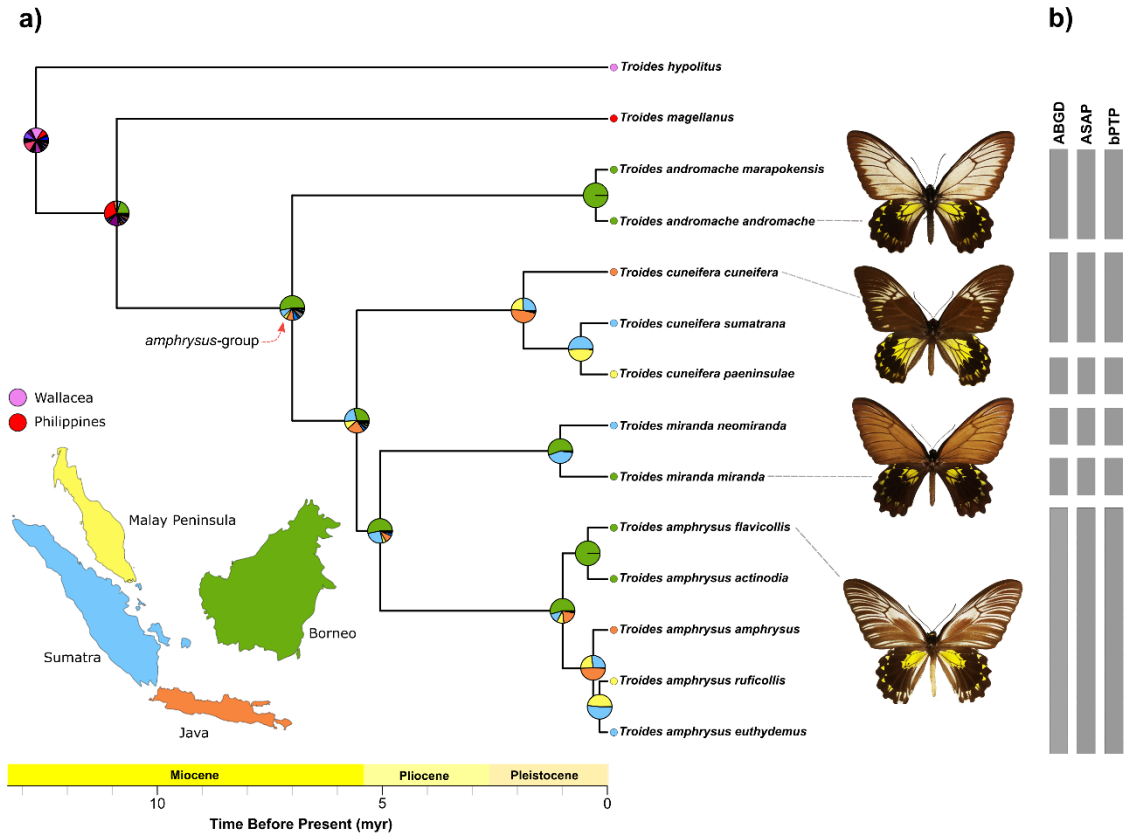


Figure 6. a) Time-calibrated phylogeny with ancestral range estimates of the *amphrysus*-group and selected outgroup taxa as proposed under the DEC + *J* model in BioGeoBEARS. Pie charts indicate the probability of ancestral ranges of a certain node. Range colors indicated on the phylogeny and inset map refer to: pink (Wallacea), red (Philippines), green (Borneo), orange (Java), blue (Sumatra) and yellow (Malay Peninsula). Ranges of outgroups in Wallacea and the Philippines are not depicted on the inset map. Additional colors in piecharts indicate the probability of mixed ancestral ranges. b) Species delimitations based on three species delimitation analyses: Automated Barcode Gap Discovery (ABGD), Assemble Species by Automated Partitioning (ASAP) and Bayesian implementation of Poisson Tree Processes (bPTP). Horizontal bars indicate species partitions proposed per analysis. Selected subspecies are shown, specimen figures adapted from Natural History Museum (2014).

Discussion

We studied the diversification of a well-known clade of *Troides* birdwing butterflies by drawing on the specimen record amassed in natural history museums over a 100-year period of biological exploration. Through the application of museomics – high-throughput sequencing to obtain genomic data from museum specimens – to butterfly specimens collected between the 1890's and 1990's, we could gather nuclear and mitochondrial data spanning the geographical range of the *amphrysus*-group. In doing so, we were able to reveal the phylogenetic relationships, timing of diversification, historical biogeography, and intraspecific diversity and in turn provide an insect's perspective on diversification in the megadiverse region Sundaland.

Phylogenetic relationships of the *amphrysus*-group

Interspecific and intergeneric relationships recovered in our study are in high agreement with earlier studies on the phylogenetic relationships of birdwings and Papilionidae (**Figs. 2, 3**) (Condamine et al. 2015, Allio et al. 2020a, Allio et al. 2021). *Trogonoptera* and *Ornithoptera* are well-supported as separate genera and form the outgroup to *Troides*. The monophyly of the *amphrysus*-group has previously been recognised on the basis of morphological (Haugum and Low 1979) and molecular studies (Condamine et al. 2015), and has been confirmed here using whole-genome data (**Figs. 2, 3**).

Within the *amphrysus*-group we initially recovered two alternative topologies of the relationships between the three species complexes *T. cuneifera*, *T. miranda* and *T. amphrysus*, for the reduced dataset versus the mitogenome and optimized datasets (**Fig. 2**, Fig. S2b, c). The alternative topology indicated *T. cuneifera* to be the sister taxon of *T. amphrysus* instead of *T. miranda* (Fig. S2b). All likelihood mapping analyses on the original datasets and on the datasets with the rogue taxa removed unequivocally indicated the topology presented in **Figure 2** as the most stable topology (Fig. S4). Missing data accounted for 60.59% of the reduced dataset, and this most likely played a role in forming topologically incongruent relationships with the other datasets which suffered less from missing data (mitogenome 1.14%, optimized 34.8%). Permutation analyses did not reveal the exact cause of the conflict, indicating it could be a combination of factors (Fig. S4). However, a large amount of missing data is known to be a source of phylogenetic inconsistencies in phylogenomic studies (Roure et al. 2012), and this is especially the case in museomics studies where the degraded nature of the sequenced tissues increases the likelihood of alignments with large amounts of missing data (Smith et al. 2020).

Multiple unique topologies recovered during the tree searches for each of the three datasets could be attributed to the shifting positions of rogue taxa (Fig. S2a, b, c). Their removal resulted in only one stable backbone topology being recovered for all three datasets. These rogue taxa had especially low coverage (Table S1), and only locally shifted position but were confidently assigned within their respective species complex. Relationships between species complexes within the *amphrysus*-group recovered by this study corroborates the previous work by Condamine et al. (2015).

Timing of diversification within the *amphrysus*-group

Placing diversification in a temporal framework

Intraspecific relationships, especially shallow intraspecific relationships, are generally difficult to resolve using phylogenetic methods due to the interaction of phylogenetic and population genetic processes at play during speciation (Mirarab et al. 2021). Some, but not all, of these processes are accommodated for by careful model selection (Mirarab et al. 2021). Although divergence dated "specimen trees" are commonplace in the literature, such trees commonly violate model assumptions for speciation and coalescent-based tree priors and may correspondingly provide inaccurate divergence time estimates (Ritchie et al. 2017). The MSC provides the most stable model to discern relationships within recently diversified clades, as it accommodates different genealogical histories within one species tree (Rannala and Yang 2003). The variation of genetic differentiation between different subspecies lineages (**Fig. 6b**) highlights that some lineages within the *amphrysus*-group may be further along the speciation continuum than others. This is reflected in the non-monophyly of subspecies lineages under the supermatrix approach (**Fig. 2**). The supertree (Fig. S3) created under the MSC model based on 103 separate gene trees, shows higher agreement with the subspecies classifications than the species tree inferred from the concatenated alignment of the same dataset (**Fig. 2**). In the supertree, branches in conflict with current subspecies designations were characterised by low to moderate support, indicating near-random associations (Fig. S3). Whilst acknowledging that population genetic processes may interfere with phylogenetic processes among the subspecies lineages investigated, the topology inferred using the supertree approach was selected for divergence time analyses.

Pleistocene diversification within the amphrysus-group

The species complexes that make up the *amphrysus*-group are inferred to have diverged 5-7 Ma (**Fig. 3**). This period coincides with marked changes in paleoclimate and paleogeography in the Southeast Asian region (Hall 2009, 2013, Holbourn et al. 2018). During the mid- to late Miocene, Borneo became a higher and larger landmass, and mountain ranges rapidly emerged in Sumatra and West Java (Hall 2013). Additionally, intense climatic changes took place during this timeframe as a result of late Miocene cooling and the intensification of the Southeast Asian winter monsoon 5.5–7 Ma (Holbourn et al. 2018). Following this initial diversification, no new extant lineages arose for approximately 3 million years until a second burst of diversification which is inferred to have taken place during the Pleistocene (2.6–0.012 Ma). This last burst of diversification giving rise to the currently recognized subspecies (**Fig. 3**).

Pleistocene glaciations have globally impacted evolutionary processes, not only in high-latitude regions but also in the tropics (Hewitt 2000, Hewitt 2004, Cros et al. 2020, Garg et al. 2020, Garg et al. 2022). Glaciations allowed montane communities to periodically track the cooling climate downslope and correspondingly occupy larger ranges, with the opposite occurring during interglacial periods (Hewitt 2000, Qu et al. 2011). Glaciations likely facilitated population expansions of montane taxa (Garg et al. 2020), and thereby allowed gene flow between previously disjunct populations (Hewitt 2000, Qu et al. 2011, Capurucho et al. 2018). In addition, glaciation-driven shifts in vegetation zones are expected to have a strong impact on distributions of phytophagous insects, through their own microclimatic niche requirements and tight ecological links with their host plants.

The burst of diversification during the Pleistocene highlights the role of paleoclimate dynamics on recent diversification processes in Sundaland insects (**Fig. 3**). Concordantly with other groups in Sundaland, *amphrysus*-group divergence times do not follow a "one-size-fits-all" pattern. Similar to bulbuls and babblers (Cros et al. 2020), and mammals such as colugos, mouse deer (Mason et al. 2019) and tree squirrels (Den Tex et al. 2010, Hinckley et al. 2020), some lineages show relatively deep interspecific divergence across the Greater Sunda Islands (e.g., *T. cuneifera*), whilst others are only weakly diverged (e.g., *T. amphrysus*). Ecological characteristics may drive such patterns; such as shown for edge-tolerant or forest dwelling birds (Cros et al. 2020) and mountain or lowland tree squirrels (Hinckley et al. 2020). Several forms of habitat specialization are thought to occur among these *Troides*, and a clear distinction between lowland and montane diversification patterns could not be recovered within the *amphrysus*-group. In the case of the *T. miranda* species complex, different lineages have colonized both montane (e.g., *Troides miranda neomiranda)* and lowland (e.g., *Troides*

miranda miranda) habitats. However, in general, records from most *amphrysus*-group representatives span a broad elevation gradient, such as reported by Matsuka (2001) and for most detailed descriptions are lacking. Investigation of montane diversification may be aided by sampling more densely, and sampling across elevational gradients of known montane specialists, e.g., *T. andromache*. The most recently diverged lineage, *T. amphrysus* is the only species complex of which all subspecies are known to occur from sea level to the lower montane zone (Haugum and Low 1979, Tsukada and Nishiyama 1982, Matsuka 2001). In contrast to the other predominantly montane *amphrysus*-group representatives, glaciation-induced changes to lowland ecosystems may have had a more significant impact on patterns in this species complex.

Sea-level fluctuations and the amphrysus-group

During glaciations, global sea levels decreased as water became locked in expanding polar ice sheets, with local sea levels generally 40-60 m below present in the Sundaland region, frequently exposing the Sunda Shelf (Voris 2000, Hall and Morley 2004, De Bruyn et al. 2014). Geological and biological evidence indicates that the Sunda Shelf acted as a land bridge within Sundaland, connecting Greater Sunda Islands (Borneo, Sumatra and Java) to the Malay Peninsula during times of lower sea level (Voris 2000, Hall and Morley 2004, Cannon et al. 2009, Woodruff 2010, De Bruyn et al. 2014). The most recent scenario, based on re-evaluated geological and biogeographical evidence by Sarr et al. (2019), indicates that the Sunda Shelf may have been continuously exposed until ~400,000 years ago, and intermittently during glaciations from then onwards (**Fig. 3**). When Sundaland was exposed, evergreen rainforests are thought to have covered most of its plains, eventually forming an expanse of rainforest habitat (Cannon et al. 2009, Woodruff 2010). Such extensive rainforests, of both lowland and upland vegetation zones, currently exist in a refugial state (Cannon et al. 2009, Woodruff 2010, De Bruyn et al. 2014).

Particularly for lowland taxa, denizens of evergreen lowland rainforests, eustatic fluctuations are expected to have resulted in losses of habitat and connectivity. Recently, a meta-analysis of intraspecific divergence times in different (largely vertebrate) clades found broad biogeographical support for divergences between 400,000–18,000 years ago, coinciding with periodical flooding of the Sunda Shelf (Husson et al. 2020). Phylogenetic evidence indicates recent diversification and relatively low genetic divergence between some *T. amphrysus* subspecies (**Figs. 3, 4, 5**). Although Bornean and western Sundaic *T. amphrysus*

populations diverged approximately 1.00 Ma, subspecies in the western Sundaland clade diverged from one another between approximately 320,000-170,000 years ago (**Fig. 3**). Divergence times coincide with the onset of major marine incursions reported by Sarr et al. (2019), which may have separated *T. amphrysus* populations into the western Greater Sunda and the Malay Peninsula subspecies (*T. a. amphrysus* vs. *T. amphrysus euthydemus* and *T. amphrysus ruficollis*; **Fig. 3**).

Increasing the geographical coverage of this wide-ranging, subspecies-rich, and recently diversified "supertramp" species (*sensu* Diamond (1974)) may provide additional insights on the potential impact of eustatic fluctuations in Sundaland on its diversification. Here, Sunda Shelf islands may provide additional biogeographical clues, by harboring intermediate forms between eastern and western Sundaland subspecies groups. Recently it has been demonstrated that populations of birds and mammals on such islands show genetic affinity with the landmasses they were most recently connected with during Pleistocene glaciations, which does not necessarily correspond to the nearest landmass, in contrast to classic island biogeographical theory (Mason et al. 2019, Garg et al. 2022). Particularly insightful would be the incorporation of island forms occurring on islands between Borneo and the western Sundaic landmasses such as the Natuna Islands.

Borneo as an evolutionary source

Borneo was likely intermittently connected with the Philippines along an arc of volcanic islands in the Sulu Sea during the late Miocene (~10 Ma) (Hall 2013), which may have functioned as the entry point to Sundaland for ancestral representatives of the *amphrysus*-group from their former range in the Philippines and Wallacea (Condamine et al. 2015). Subsequently, the four different species complexes within the group most likely differentiated on Borneo (**Fig. 6a**). In a meta-analysis of divergence times De Bruyn et al. (2014) identified Borneo as a major evolutionary hotspot within South-East Asia, and within Sundaland. With a diverse range of organisms having significantly higher levels of *in situ* diversification in Borneo, and correspondingly higher levels of emigration out of Borneo than on any other Sundaic landmass (De Bruyn et al. 2014). The importance of Borneo as a source and centre of diversity for many Sundaic taxa has been coined the "Out-of-Borneo" hypothesis and is applicable to a wide range of taxa (De Bruyn et al. 2014, Grismer et al. 2017, Williams et al. 2017, Flury et al. 2021). Results from the biogeographical analysis corroborate the "Out-of-Borneo" hypothesis for the

amphrysus-group, as Borneo harbors the greatest diversity of *amphrysus*-group representatives within Sundaland and has been the source of repeated westward expansions across the Sunda Shelf to the western Greater Sunda Islands (**Fig. 6a**).

Following initial dispersal from Borneo, *T. cuneifera* and *T. amphrysus* show remarkable overlap in the pattern of colonization of the western Greater Sunda Islands (**Fig. 6a**). The earliest diverging subspecies of both species complexes, *T. cuneifera cuneifera* and *T. amphrysus amphrysus* occur on Java, suggesting that dispersal from Borneo to Java most likely functioned as the founding event of the western Sundaic cohort of both species complexes. Important to note is that the divergence event between *T. cuneifera cuneifera* and the common ancestor of *T. cuneifera paeninsulae* and *T. cuneifera sumatrana* took place approximately 1.5 million years earlier than *T. amphrysus amphrysus* from the common ancestor of *T. amphrysus euthydemus* and *T. amphrysus ruficollis* (**Fig. 3**). This suggests that the dispersal or range expansion of *amphrysus*-group from Borneo to Java may have occurred at least twice over a two-million-year timespan. Sumatran and Peninsular representatives of both species complexes are most closely related, as is expected by the short distance between their respective ranges and the frequent connection of Sumatra and the Malay Peninsula (Voris 2000).

The *T. miranda* species complex follows a Bornean-Sumatran distribution with potentially hidden species-level differentiation between the two islands (Tsukada and Nishiyama 1982). This biogeographical connection between Borneo and Sumatra is well-known within Sundaland, with a diverse array of species sharing recent common ancestry between the two islands, references in Husson et al. (2020). Bathymetry indicates the land bridge between Borneo and Sumatra was likely the most frequently occurring corridor between Borneo and the western Greater Sunda Islands (Voris 2000). Ancestral representatives of this species complex likely dispersed from Borneo to Sumatra over such land bridges, which may even have been continuously exposed during the timeframe of divergence between the two lineages (Sarr et al. 2019). The inclusion of *T. miranda hayamii*, the southernmost representative of the *T. miranda* complex in Borneo, would shed additional light on diversification and biogeographical patterns of this species complex.

Over-water dispersal may also play a role in the distribution of the *amphrysus*-group, however given the divergence time estimates (**Fig. 3**) and the genetic distance (**Fig. 5**), it is unlikely that overwater dispersal continuously takes place between the larger landmasses of Sundaland. One exception may be between subspecies *T. amphrysus euthydemus* and *T. amphrysus ruficollis*, which are highly similar genetically (**Figs. 4, 5**) and both occupy

relatively low-lying areas that are only separated by a narrow strait that was frequently exposed at slightly lower sea-levels (Voris 2000).

Hidden diversity revealed?

Currently recognised subspecies of the *amphrysus*-group have generally been described on the basis of coloration and allopatric distribution ranges (e.g., Staudinger (1892). However, color patterns within the *amphrysus*-group may be relatively plastic and vary widely between closely related subspecies, whilst showing overlap between co-occurring but more distantly related taxa, possibly governed by Müllerian mimicry associations (Haugum and Low 1979). Nevertheless, results from phylogenomic analyses closely match with the 19th and early 20th century classifications within the *amphrysus*-group. Cryptic diversity – phenotypically identical but widely genetically differentiated lineages – is generally not present within the species group. However, hidden diversity – greater genetic differentiation than expected based on phenotypic differences – was encountered among the investigated taxa, and analyses indicate that the *amphrysus*-group may putatively be represented by six species-level lineages rather than the originally described four (**Fig. 6b**).

The putative species delimitation within the *T. miranda* species complex corresponds with phenotypic and ecological clues of possible species-level differentiation between the two lineages. Indeed, *T. miranda miranda* and *T. miranda neomiranda* differ markedly in phenotype, particularly in the coloration of female forewings, *T. miranda neomiranda* females having more strikingly patterned forewings in comparison to the more uniformly colored *T. miranda miranda* females (e.g., inset Fig. 2). Haugum and Low (1979) describe significant differences in male genitalia between *T. miranda miranda* and *T. miranda neomiranda*, especially noting a different degree of serration on the ventral hook of the harpe. Not only do the two subspecies occur on distant areas (Borneo and Sumatra) (**Fig. 4c**), their habitat and altitudinal ranges are also markedly different. *T. miranda miranda* is known predominantly as a lowland species, even frequently recorded in sea-level habitats such as mangroves and lowland rainforest and is rarely recorded at higher elevations and unknown from elevations exceeding 500 m above sea level (Haugum and Low 1979). Whilst *T. miranda neomiranda* is a high-altitude species, generally thought to be restricted to mountain ranges around 1000 m above sea level (Haugum and Low 1979, Tsukada and Nishiyama 1982). Both authors hint at possible species-level differentiation (Haugum and Low 1979, Tsukada and Nishiyama 1982).

Along with the multiple ecological and morphological distinctions, the new genetic evidence presented in this study provides the first line of genetic evidence for a species status for *T. miranda miranda* and *T. miranda neomiranda*. However, we remain cautious because a single specimen for *T. miranda miranda* has been included and the subspecies *T. miranda hayamii* is missing from our taxon set.

The second case of putative species-level differentiation was uncovered in the *T. cuneifera* species complex, with consensus delimitation methods indicating a species-level split between *T. cuneifera cuneifera* and *T. cuneifera paeninsulae* + *T. cuneifera sumatrana* (**Fig. 6b**). We found a high degree of genomic differentiation between these two putative species (**Fig. 5**). However, the potential species-level differentiation within *T. cuneifera* is less obvious with respect to phenotypic differentiation, as phenotypes from the three subspecies are similar, which may be driven by Müllerian mimicry associations with co-occurring *Troides* species. With respect to male genitalia, the clasper and harpe of *T. cuneifera cuneifera* slightly differ to those of *T. cuneifera paeninsulae* and *T. cuneifera sumatrana* (Haugum and Low 1979). Observations on the ecology of the subspecies, such as habitat requirements and host-plant associations, are scant and cannot at present provide additional evidence for taxonomical considerations within this species complex. On the contrary, *T. cuneifera paeninsulae* and *T. cuneifera sumatrana* are highly similar and multiple authors have questioned the validity of their respective subspecies status e.g., Haugum and Low (1979). Genetic distance and phylogenomic evidence, however, indicate some degree of genomic differentiation between these subspecies and their status as distinct subspecies remains valid from this perspective (**Figs. 2 - 5**, Fig. S3).

Aside from the two putative species-level splits, subspecies designations seemed to accurately reflect the genetic differentiation between the different taxa examined. One other exception being the subspecies *T. amphrysus euthydemus* and *T. amphrysus ruficollis*, which showed significant overlap in mitochondrial haplotypes (**Fig. 4**) and virtually no differentiation for either of the examined genes (**Fig. 5**), indicating that their respective subspecies status may not accurately reflect the level of divergence between the two forms. This has long been recognised, with several authors noting no constant morphological characters separating the Peninsular and Sumatran populations (e.g., d'Abrera (1975)), however their distinct status has persisted since their description. The lack of distinction between Peninsular and Sumatran representatives is not completely unsurprising as *T. amphrysus* is known to have a relatively broad elevational and geographical range, along with the frequent and recent land connection between Sumatra and the Malay Peninsula during periods of lower sea-level (Voris 2000). The

paucity of morphological distinctions in conjunction with the lack of genetic differentiation recovered by this study together suggest that *T. a. ruficollis* and *T. a. euthydemus* may need to be merged into a single taxon.

The ability of species delimitation methods to differentiate genetic clusters into species-level lineages varies between approaches and the speciation model (Carstens et al. 2013). Comparison of several species delimitation methods to form a consensus delimitation scheme is generally advocated (Carstens et al. 2013, Suárez-Villota et al. 2018, Jin et al. 2020, Pozzi et al. 2020). Together, the different delimitation methods reveal a plausible consensus species partition scheme among the studied taxa (**Fig. 6b**). Molecular species delimitation requires critical evaluation and should not be seen as a replacement to traditional taxonomic treatment, as stressed by many developers of species delimitation algorithms (e.g. Puillandre et al. (2012)). Classical taxonomical revision of the *amphrysus*-clade should be carried out to validate suggested species-level splits within the *T. cuneifera* and *T. miranda* species complexes. Such work should encompass all subspecies lineages and be carried out in an integrative taxonomic framework that considers morphological, ecological, and genomic lines of evidence to investigate species boundaries.

Conclusion

The complex distribution patterns of island biota have been a source of fascination to evolutionary biologists since Alfred Russel Wallace founded the field of biogeography (Wallace 1855). In Sundaland, complex geography, geological and paleoclimatic processes have shaped evolution in a way that exemplifies this fascination to the extreme. In this study we applied museomics to a well-known and diverse lineage of butterflies from one of the world's most famous archipelagos. Using modern-day phylogenomic methods, even specimens collected ~150 years ago by 19th century naturalists could be successfully incorporated into a phylogenetic framework, and thereby contribute to our understanding of the tree of life today. Here, we provide evidence for intraspecific biogeographic processes in Sundaic insects, and thereby address a knowledge gap in our understanding of biogeographical patterns in this biodiversity hotspot. After initially originating in the late Miocene, the formation of amphrysus-group species complexes dates to the Pliocene and within-species diversification to the Pleistocene, reaffirming the role of this geological period in shaping the Sundaic biota. Borneo was identified as the centre of diversity, and the source of numerous westward

colonization events by the amphrysus-group, so joining the diverse array of taxa with an “Out of Borneo” evolutionary history. Furthermore, through species delimitation analyses we provide the first holdfasts for describing hidden species within the group and suggest detailed taxonomical examination of the T. cuneifera and T. miranda complexes.

Acknowledgements

We would like to thank Mónica Guimarães Cruz (Naturalis Biodiversity Centre, Leiden) and Viola Richter (Museum für Naturkunde, Berlin) for providing most of the specimens used in this study. Furthermore, we are grateful for the help of Elsa Call, Freek Bakker, Harald Letsch, Siavash Mirarab and Chris Wheat who provided a wealth of advice and guidance regarding their respective areas of expertise. We would also like to thank Martin Partridge of the Swallowtail & Birdwing Butterfly Trust for sharing relevant literature and Thomas Wong for providing the program UniqueTree 1.9. We would also like to acknowledge Djunijanti Peggie (Bogor Zoology Museum, Bogor), Ulmar Grafe (Universiti Brunei Darussalam, Brunei Darussalam) and Wei Song Hwang (Lee Kong Chian Natural History Museum, Singapore) for searching for *T. andromache* specimens in the collections under their care. We are grateful for comments by Tobias Bader, Esther Blommert, Mikay Breet, Klaas Bouwmeester, Nora Damaris Pasquali Medici de Biron, Liana Greenburg, Iris Haverkorn, Sybren Kooistra, Thijmen van der Loop and Isabel Lugtenburg on early versions of the manuscript. Additionally, we would like to thank Bart Burger for his suggestions on components of the analysis and Isabela Pombo Geertsma for her suggestions to improve the final manuscript. Finally, we thank Robin van Velzen for his continued support during the final stages and the submission of this work. This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (project GAIA, agreement no. 851188 to F.L.C.).

Conflict of Interest

The authors declare no conflicts of interest.

Data Availability Statement

Raw sequences, *de novo* genome and mitogenome assemblies are available on the European Nucleotide Archive under the project accession PRJEB64430.

Author contributions

CFHvdL: Conceptualization; Data curation; Formal analysis; Investigation; Writing – original draft; Visualization; Writing – review & editing. **ELR:** Resources; Data curation; Writing – review & editing. **EC:** Resources; Writing – review & editing. **ME:** Methodology; Resources; Writing – review & editing. **TL:** Resources; Writing – review & editing. **SLS:** Conceptualization; Writing – review & editing. **PV:** Supervision; Methodology; Resources; Writing – review & editing. **RdV:** Resources; Writing – review & editing. **NW:** Resources; Writing – review & editing. **MES:** Supervision; Project administration; Writing – review & editing. **FLC:** Supervision; Resources; Methodology; Funding acquisition; Writing – review & editing. **SS:** Supervision; Methodology; Funding acquisition; Conceptualization; Project administration; Writing – review & editing.

References

Abascal, F., D. Posada, and R. Zardoya. 2007. MtArt: a new model of amino acid replacement for Arthropoda. Molecular Biology and Evolution 24: 1-5.

Aberer, A. J., D. Krompass, and A. Stamatakis. 2013. Pruning rogue taxa improves phylogenetic accuracy: an efficient algorithm and webservice. Systematic biology 62: 162-166.

Allio, R., C. Scornavacca, B. Nabholz, A.-L. Clamens, F. A. H. Sperling, and F. L. Condamine. 2020a. Whole Genome Shotgun Phylogenomics Resolves the Pattern and Timing of Swallowtail Butterfly Evolution. Systematic biology 69: 38-60.

Allio, R., A. Schomaker-Bastos, J. Romiguier, F. Prosdocimi, B. Nabholz, and F. Delsuc. 2020b. MitoFinder: Efficient automated large-scale extraction of mitogenomic data in target enrichment phylogenomics. Molecular Ecology Resources 20: 892-905.

Allio, R., B. Nabholz, S. Wanke, G. Chomicki, O. A. Pérez-Escobar, A. M. Cotton, A.-L. Clamens, G. J. Kergoat, F. A. H. Sperling, and F. L. Condamine. 2021. Genome-wide macroevolutionary signatures of key innovations in butterflies colonizing new host plants. Nature Communications 12: 354.

Bakker, F. T., A. Antonelli, J. A. Clarke, J. A. Cook, S. V. Edwards, P. G. Ericson, S. Faurby, N. Ferrand, M. Gelang, and R. G. Gillespie. 2020. The Global Museum: natural history collections and the future of evolutionary science and public education. PeerJ 8: e8225.

Bolger, A. M., M. Lohse, and B. Usadel. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics (Oxford, England) 30: 2114-2120.
Bouckaert, R., T. G. Vaughan, J. Barido-Sottani, S. Duchêne, M. Fourment, A. Gavryushkina, J. Heled, G. Jones, D. Kühnert, N. De Maio, M. Matschiner, F. K. Mendes, N. F. Müller, H. A. Ogilvie, L. du Plessis, A. Popinga, A. Rambaut, D. Rasmussen, I. Siveroni, M. A. Suchard, C.-H. Wu, D. Xie, C. Zhang, T. Stadler, and A. J. Drummond. 2019. BEAST 2.5: An advanced software platform for Bayesian evolutionary analysis. PLOS Computational Biology 15: e1006650.
Bouckaert, R. R. 2020. OBAMA: OBAMA for Bayesian amino-acid model averaging. PeerJ 8: e9460-e9460.
Call, E. 2020. The age of museomics: How to get genomic information from museum specimens of Lepidoptera. PhD Doctoral Thesis (compilation), Lund University Lund.
Call, E., C. Mayer, V. Twort, L. Dietz, N. Wahlberg, and M. Espeland. 2021. Museomics: Phylogenomics of the Moth Family Epicopeiidae (Lepidoptera) Using Target Enrichment. Insect Systematics and Diversity 5.
Camacho, C., G. Coulouris, V. Avagyan, N. Ma, J. Papadopoulos, K. Bealer, and T. L. Madden. 2009. BLAST+: architecture and applications. BMC Bioinformatics 10: 421.
Cannon, C. H., R. J. Morley, and A. B. G. Bush. 2009. The current refugial rainforests of Sundaland are unrepresentative of their biogeographic past and highly vulnerable to disturbance. Proceedings of the National Academy of Sciences 106: 11188.
Capurucho, J. M. G., M. V. Ashley, C. C. Ribas, and J. M. Bates. 2018. Connecting Amazonian, Cerrado, and Atlantic forest histories: Paraphyly, old divergences, and modern population dynamics in tyrant-manakins (Neopelma/Tyranneutes, Aves: Pipridae). Molecular Phylogenetics and Evolution 127: 696-705.
Card, D. C., B. Shapiro, G. Giribet, C. Moritz, and S. V. Edwards. 2021. Museum Genomics. Annual Review of Genetics 55: 633-659.
Carstens, B. C., T. A. Pelletier, N. M. Reid, and J. D. Satler. 2013. How to fail at species delimitation. Molecular Ecology 22: 4369-4383.
Chen, S., Y. Zhou, Y. Chen, and J. Gu. 2018. fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics (Oxford, England) 34: i884-i890.
Chernomor, O., A. von Haeseler, and B. Q. Minh. 2016. Terrace Aware Data Structure for Phylogenomic Inference from Supermatrices. Systematic biology 65: 997-1008.
Condamine, F. L., E. F. Toussaint, A.-L. Clamens, G. Genson, F. A. Sperling, and G. J. Kergoat. 2015. Deciphering the evolution of birdwing butterflies 150 years after Alfred Russel Wallace. Scientific Reports 5: 11860.
Corlett, R. 2014. The ecology of tropical East Asia, Oxford University Press, USA.
Cros, E., B. Chattopadhyay, K. M. Garg, N. S. R. Ng, S. Tomassi, S. Benedick, D. P. Edwards, and F. E. Rheindt. 2020. Quaternary land bridges have not been universal conduits of gene flow. Molecular Ecology 29: 2692-2706.
d'Abrera, B. 1975. Birdwing butterflies of the world, Lansdowne Press, Melbourne.
De Bruyn, M., B. Stelbrink, R. Morley, R. Hall, G. Carvalho, C. Cannon, G. Van den Bergh, E. Meijaard, I. Metcalfe, L. Boitani, L. Maiorano, R. Shoup, and T. Rintelen. 2014. Borneo and Indochina are Major Evolutionary Hotspots for Southeast Asian Biodiversity. Systematic biology 63: 879-901.
Den Tex, R.-J., R. Thorington, J. E. Maldonado, and J. A. Leonard. 2010. Speciation dynamics in the SE Asian tropics: Putting a time perspective on the phylogeny and biogeography of Sundaland tree squirrels, Sundasciurus. Molecular Phylogenetics and Evolution 55: 711-720.
Diamond, J. M. 1974. Colonization of Exploded Volcanic Islands by Birds: The Supertramp Strategy. Science 184: 803-806.

Drummond, A. J., S. Y. W. Ho, M. J. Phillips, and A. Rambaut. 2006. Relaxed Phylogenetics and Dating with Confidence. PLOS Biology 4: e88.

Flury, J. M., A. Haas, R. M. Brown, I. Das, Y. M. Pui, K. Boon-Hee, U. Scheidt, D. T. Iskandar, A. Jankowski, and S. T. Hertwig. 2021. Unexpectedly high levels of lineage diversity in Sundaland puddle frogs (Dicroglossidae: Occidozyga Kuhl and van Hasselt, 1822). Molecular Phylogenetics and Evolution 163: 107210.

Fruhstorfer, H. 1898. Neue Papilio-Formen aus dem Malayischen Archipel. Berliner Entomologischer Zeitschrift 43: 419.

Garg, K. M., B. Chattopadhyay, B. Koane, K. Sam, and F. E. Rheindt. 2020. Last Glacial Maximum led to community-wide population expansion in a montane songbird radiation in highland Papua New Guinea. BMC Evolutionary Biology 20: 82.

Garg, K. M., B. Chattopadhyay, E. Cros, S. Tomassi, S. Benedick, D. P. Edwards, and F. E. Rheindt. 2022. Island Biogeography Revisited: Museomics Reveals Affinities of Shelf Island Birds Determined by Bathymetry and Paleo-Rivers, Not by Distance to Mainland. Molecular biology and evolution 39: msab340.

Grismer, L. L., P. L. Wood, Jr., A. Aowphol, M. Cota, M. S. Grismer, M. L. Murdoch, C. Aguilar, and J. L. Grismer. 2017. Out of Borneo, again and again: biogeography of the Stream Toad genus Ansonia Stoliczka (Anura: Bufonidae) and the discovery of the first limestone cave-dwelling species. Biological Journal of the Linnean Society 120: 371-395.

Gu, X., Y. X. Fu, and W. H. Li. 1995. Maximum likelihood estimation of the heterogeneity of substitution rate among nucleotide sites. Molecular Biology and Evolution 12: 546-557.

Hall, R. 2009. Southeast Asia's changing palaeogeography. Blumea-Biodiversity, Evolution and Biogeography of Plants 54: 148-161.

Hall, R. 2013. The palaeogeography of Sundaland and Wallacea since the Late Jurassic. Journal of Limnology 72: e1.

Hall, R., and C. K. Morley. 2004. Sundaland basins, pp. 55-85. In P. Clift, W. Kuhnt, P. Wang and D. Hayes (eds.), Continent-Ocean Interactions Within East Asian Marginal Seas, vol. 149. American Geophysical Union.

Haugum, J., and A. M. Low. 1979. A Monograph of the Birdwing Butterflies, Troides; amphrysus & haliphron groups, vol. 2, part 2, Scandinavian Science Press LTD., Klampenborg, Denmark.

Häuser, C. L., J. Holstein, and A. Steiner. 2005. The Global Butterfly Information System.

He, J. W., R. Zhang, J. Yang, Z. Chang, L. X. Zhu, S. H. Lu, F. A. Xie, J. L. Mao, Z. W. Dong, G. C. Liu, P. Hu, Y. Dong, W. T. Wan, R. P. Zhao, T. Z. Xiong, J. L. León-Cortés, C. Y. Mao, W. Zhang, S. Zhan, J. Li, L. Chen, W. Wang, and X. Y. Li. 2022. High-quality reference genomes of swallowtail butterflies provide insights into their coloration evolution. Zoological research 43: 367-379.

Hewitt, G. 2000. The genetic legacy of the Quaternary ice ages. Nature 405: 907-913.

Hewitt, G. M. 2004. Genetic consequences of climatic oscillations in the Quaternary. Philos Trans R Soc Lond B Biol Sci 359: 183-195.

Hinckley, A., M. T. R. Hawkins, A. S. Achmadi, J. E. Maldonado, and J. A. Leonard. 2020. Ancient Divergence Driven by Geographic Isolation and Ecological Adaptation in Forest Dependent Sundaland Tree Squirrels. Frontiers in Ecology and Evolution 8.

Holbourn, A. E., W. Kuhnt, S. C. Clemens, K. G. D. Kochhann, J. Jöhnck, J. Lübbers, and N. Andersen. 2018. Late Miocene climate cooling and intensification of southeast Asian winter monsoon. Nature Communications 9: 1584.

Hurvich, C. M., and C.-L. Tsai. 1989. Regression and time series model selection in small samples. Biometrika 76: 297-307.

Husson, L., F. C. Boucher, A.-C. Sarr, P. Sepulchre, and S. Y. Cahyarini. 2020. Evidence of Sundaland's subsidence requires revisiting its biogeography. Journal of Biogeography 47: 843-853.

Jin, M., A. Zwick, A. Ślipiński, R. de Keyzer, and H. Pang. 2020. Museomics reveals extensive cryptic diversity of Australian prionine longhorn beetles with implications for their classification and conservation. Systematic Entomology 45: 745-770.

Jones, D. T., W. R. Taylor, and J. M. Thornton. 1992. The rapid generation of mutation data matrices from protein sequences. Bioinformatics (Oxford, England) 8: 275-282.

Jones, G. 2017. Algorithmic improvements to species delimitation and phylogeny estimation under the multispecies coalescent. Journal of Mathematical Biology 74: 447-467.

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

Katoh, K., and D. M. Standley. 2013. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. Molecular Biology and Evolution 30: 772-780.

Kawahara, A. Y., D. Plotkin, M. Espeland, K. Meusemann, E. F. A. Toussaint, A. Donath, F. Gimnich, P. B. Frandsen, A. Zwick, M. dos Reis, J. R. Barber, R. S. Peters, S. Liu, X. Zhou, C. Mayer, L. Podsiadlowski, C. Storer, J. E. Yack, B. Misof, and J. W. Breinholt. 2019. Phylogenomics reveals the evolutionary timing and pattern of butterflies and moths. Proceedings of the National Academy of Sciences 116: 22657.

Kosiol, C., and N. Goldman. 2005. Different versions of the Dayhoff rate matrix. Molecular Biology and Evolution 22: 193-199.

Kück, P. 2016. ALICUT: A Perlscript which cuts ALISCORE identified RSS computer program, version 2.3. By Kück, P., Bonn, Germany.

Kück, P., and G. C. Longo. 2014. FASconCAT-G: extensive functions for multiple sequence alignment preparations concerning phylogenetic studies. Frontiers in Zoology 11: 1-8.

Kück, P., K. Meusemann, J. Dambach, B. Thormann, B. M. von Reumont, J. W. Wägele, and B. Misof. 2010. Parametric and non-parametric masking of randomness in sequence alignments can be improved and leads to better resolved trees. Frontiers in Zoology 7: 1-12.

Le, S. Q., and O. Gascuel. 2008. An improved general amino acid replacement matrix. Molecular Biology and Evolution 25: 1307-1320.

Le, S. Q., C. C. Dang, and O. Gascuel. 2012. Modeling Protein Evolution with Several Amino Acid Replacement Matrices Depending on Site Rates. Molecular Biology and Evolution 29: 2921-2936.

Le, V. S., C. C. Dang, and Q. S. Le. 2017. Improved mitochondrial amino acid substitution models for metazoan evolutionary studies. BMC Evolutionary Biology 17: 136.

Mason, V. C., K. M. Helgen, and W. J. Murphy. 2019. Comparative phylogeography of forest-dependent mammals reveals Paleo-forest corridors throughout Sundaland. Journal of Heredity 110: 158-172.

Matsuka, H. 2001. トリバネチョウ生態図鑑: Natural history of birdwing butterflies, 松香出版.

Matzke, N. J. 2018. BioGeoBEARS: BioGeography with Bayesian (and likelihood) Evolutionary Analysis with R Scripts computer program, version 1.1.1. By Matzke, N. J., GitHub published on GitHub on November 6, 2018.

Meng, G., Y. Li, C. Yang, and S. Liu. 2019. MitoZ: a toolkit for animal mitochondrial genome assembly, annotation and visualization. Nucleic Acids Research 47: e63-e63.

Meyer, B., K. Meusemann, and B. Misof 2011. MARE: MAtrix REduction - A tool to select optimized data subsets from supermatrices for phylogenetic inference computer

program, version 0.1.2-rc. By Meyer, B., K. Meusemann, and B. Misof, Adenauerallee 160, 53113 Bonn, Germany.
Meyer, M., and M. Kircher. 2010. Illumina sequencing library preparation for highly multiplexed target capture and sequencing. Cold Spring Harbor protocols 2010: pdb.prot5448.
Minh, B. Q., H. A. Schmidt, O. Chernomor, D. Schrempf, M. D. Woodhams, A. von Haeseler, and R. Lanfear. 2020. IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. Molecular Biology and Evolution 37: 1530-1534.
Mirarab, S., L. Nakhleh, and T. Warnow. 2021. Multispecies Coalescent: Theory and Applications in Phylogenetics. Annual Review of Ecology, Evolution, and Systematics 52: 247-268.
Mittermeier, R., P. Robles-Gil, M. Hoffmann, J. Pilgrim, T. Brooks, C. Mittermeier, J. Lamoreux, and G. da Fonseca. 2005. Hotspots Revisited: Earth's Biologically Richest and Most Endangered Ecoregions. Conservation International, Washington.
Müller, T., and M. Vingron. 2000. Modeling amino acid replacement. Journal of Computational Biology 7: 761-776.
Myers, N., R. A. Mittermeier, C. G. Mittermeier, G. A. B. da Fonseca, and J. Kent. 2000. Biodiversity hotspots for conservation priorities. Nature 403: 853-858.
Natural History Museum. 2014. Dataset: Collection specimens. Resource: Specimens. Natural History Museum Data Portal (data.nhm.ac.uk).
Neethiraj, R., E. A. Hornett, J. A. Hill, and C. W. Wheat. 2017. Investigating the genomic basis of discrete phenotypes using a Pool-Seq-only approach: New insights into the genetics underlying colour variation in diverse taxa. Molecular Ecology 26: 4990-5002.
Nishida, R., J. D. Weintraub, P. Feeny, and H. Fukami. 1993. Aristolochic acids from Thottea spp. (Aristolochiaceae) and the osmeterial secretions of Thottea-feeding Troidine swallowtail larvae (Papilionidae). Journal of Chemical Ecology 19: 1587-1594.
Nurk, S., A. Bankevich, D. Antipov, A. Gurevich, A. Korobeynikov, A. Lapidus, A. Prjibelsky, A. Pyshkin, A. Sirotkin, Y. Sirotkin, R. Stepanauskas, J. McLean, R. Lasken, S. R. Clingenpeel, T. Woyke, G. Tesler, M. A. Alekseyev, and P. A. Pevzner. Year. Published. Assembling Genomes and Mini-metagenomes from Highly Chimeric Reads, pp. 158-170. *In* M. Deng, R. Jiang, F. Sun and X. Zhang (eds.), Research in Computational Molecular Biology, 2013// 2013, Berlin, Heidelberg. Springer Berlin Heidelberg.
Ogilvie, H. A., R. R. Bouckaert, and A. J. Drummond. 2017. StarBEAST2 Brings Faster Species Tree Inference and Accurate Estimates of Substitution Rates. Molecular Biology and Evolution 34: 2101-2114.
Ohya, T. 2001. Checklist of Birdwing Butterfly, pp. 346-348. In H. Matsuka (ed.), Natural History of Birdwing Butterflies. Matsuka Shuppan, Tokyo, Japan.
Paradis, E. 2010. pegas: an R package for population genetics with an integrated–modular approach. Bioinformatics (Oxford, England) 26: 419-420.
Petersen, M., K. Meusemann, A. Donath, D. Dowling, S. Liu, R. S. Peters, L. Podsiadlowski, A. Vasilikopoulos, X. Zhou, B. Misof, and O. Niehuis. 2017. Orthograph: a versatile tool for mapping coding nucleotide sequences to clusters of orthologous genes. BMC Bioinformatics 18: 111.
Pozzi, L., A. Penna, S. K. Bearder, J. Karlsson, A. Perkin, and T. R. Disotell. 2020. Cryptic diversity and species boundaries within the Paragalago zanzibaricus species complex. Molecular Phylogenetics and Evolution 150: 106887.
Puillandre, N., S. Brouillet, and G. Achaz. 2021. ASAP: assemble species by automatic partitioning. Molecular Ecology Resources 21: 609-620.

Puillandre, N., A. Lambert, S. Brouillet, and G. Achaz. 2012. ABGD, Automatic Barcode Gap Discovery for primary species delimitation. Molecular Ecology 21: 1864-1877.
QGIS.org 2022. QGIS 3.22.0 computer program, version By QGIS.org.
Qu, Y., X. Luo, R. Zhang, G. Song, F. Zou, and F. Lei. 2011. Lineage diversification and historical demography of a montane bird Garrulax elliotii - implications for the Pleistocene evolutionary history of the eastern Himalayas. BMC Evolutionary Biology 11: 174.
R Core Team 2021. R: A language and environment for statistical computing. computer program, version By R Core Team, Vienna, Austria.
Rambaut, A., A. J. Drummond, D. Xie, G. Baele, and M. A. Suchard. 2018. Posterior Summarization in Bayesian Phylogenetics Using Tracer 1.7. Systematic biology 67: 901-904.
Rannala, B., and Z. Yang. 2003. Bayes estimation of species divergence times and ancestral population sizes using DNA sequences from multiple loci. Genetics 164: 1645-1656.
Raxworthy, C. J., and B. T. Smith. 2021. Mining museums for historical DNA: advances and challenges in museomics. Trends in Ecology & Evolution.
Reboud, E. L., B. Nabholz, E. Chevalier, M.-k. Tilak, D. Bito, and F. L. Condamine. 2023. Genomics, Population Divergence, and Historical Demography of the World's Largest and Endangered Butterfly, The Queen Alexandra's Birdwing. Genome biology and evolution 15.
Ree, R. H., and S. A. Smith. 2008. Maximum likelihood inference of geographic range evolution by dispersal, local extinction, and cladogenesis. Systematic biology 57: 4-14.
Ritchie, A. M., N. Lo, and S. Y. W. Ho. 2017. The Impact of the Tree Prior on Molecular Dating of Data Sets Containing a Mixture of Inter- and Intraspecies Sampling. Systematic biology 66: 413-425.
Roure, B., D. Baurain, and H. Philippe. 2012. Impact of Missing Data on Phylogenies Inferred from Empirical Phylogenomic Data Sets. Molecular Biology and Evolution 30.
Rowe, K. C., S. Singhal, M. D. Macmanes, J. F. Ayroles, T. L. Morelli, E. M. Rubidge, K. E. Bi, and C. C. Moritz. 2011. Museum genomics: low-cost and high-accuracy genetic data from historical specimens. Molecular Ecology Resources 11: 1082-1092.
Sarr, A. C., L. Husson, P. Sepulchre, A. M. Pastier, K. Pedoja, M. Elliot, C. Arias-Ruiz, T. Solihuddin, S. Aribowo, and Susilohadi. 2019. Subsiding Sundaland. Geology 47: 119-122.
Sayyari, E., and S. Mirarab. 2016. Fast Coalescent-Based Computation of Local Branch Support from Quartet Frequencies. Molecular Biology and Evolution 33: 1654-1668.
Schliep, K. P. 2011. phangorn: phylogenetic analysis in R. Bioinformatics (Oxford, England) 27: 592-593.
Sholihah, A., E. Delrieu-Trottin, T. Sukmono, H. Dahruddin, R. Risdawati, R. Elvyra, A. Wibowo, K. Kustiati, F. Busson, S. Sauri, U. Nurhaman, E. Dounias, M. S. A. Zein, Y. Fitriana, I. V. Utama, Z. A. Muchlisin, J.-F. Agnèse, R. Hanner, D. Wowor, D. Steinke, P. Keith, L. Rüber, and N. Hubert. 2020. Disentangling the taxonomy of the subfamily Rasborinae (Cypriniformes, Danionidae) in Sundaland using DNA barcodes. Scientific Reports 10: 2818.
Smith, B. T., W. M. Mauck, III, B. W. Benz, and M. J. Andersen. 2020. Uneven Missing Data Skew Phylogenomic Relationships within the Lories and Lorikeets. Genome biology and evolution 12: 1131-1147.
Smith, S. A., J. W. Brown, and J. F. Walker. 2018. So many genes, so little time: A practical approach to divergence-time estimation in the genomic era. PLOS ONE 13: e0197433.

Soubrier, J., M. Steel, M. S. Y. Lee, C. Der Sarkissian, S. Guindon, S. Y. W. Ho, and A. Cooper. 2012. The Influence of Rate Heterogeneity among Sites on the Time Dependence of Molecular Rates. Molecular Biology and Evolution 29: 3345-3358.

Staudinger, O. 1892. Ornithoptera Andromache n. sp. . Deutsche Entomologische Zeitschrift 5: 393-394.

Strimmer, K., and A. von Haeseler. 1997. Likelihood-mapping: A simple method to visualize phylogenetic content of a sequence alignment. Proceedings of the National Academy of Sciences 94: 6815-6819.

Suárez-Villota, E. Y., C. A. Quercia, L. M. Díaz, V. Vera-Sovier, and J. J. Nuñez. 2018. Speciation in a biodiversity hotspot: Phylogenetic relationships, species delimitation, and divergence times of Patagonian ground frogs from the Eupsophus roseus group (Alsodidae). PLOS ONE 13: e0204968.

Tilak, M.-K., F. Justy, M. Debiais-Thibaud, F. Botero-Castro, F. Delsuc, and E. J. P. Douzery. 2015. A cost-effective straightforward protocol for shotgun Illumina libraries designed to assemble complete mitogenomes from non-model species. Conservation Genetics Resources 7: 37-40.

Tsukada, E., and Y. Nishiyama. 1982. Butterflies of the South East Asian islands, Vol. 1, Papilionidae. Plapac Co., Ltd, Tokyo, English edition, 1982, 457 pp.

Twort, V. G., J. Minet, C. W. Wheat, and N. Wahlberg. 2021. Museomics of a rare taxon: placing Whalleyanidae in the Lepidoptera Tree of Life. Systematic Entomology n/a.

Veerassamy, S., A. Smith, and E. R. Tillier. 2003. A transition probability model for amino acid substitutions from blocks. Journal of Computational Biology 10: 997-1010.

Von Rintelen, K., E. Arida, and C. Häuser. 2017. A review of biodiversity-related issues and challenges in megadiverse Indonesia and other Southeast Asian countries. Research Ideas and Outcomes 3: e20860.

Voris, H. K. 2000. Special Paper 2: Maps of Pleistocene Sea Levels in Southeast Asia: Shorelines, River Systems and Time Durations. Journal of Biogeography 27: 1153-1167.

Wallace, A. R. 1855. XVIII.—On the law which has regulated the introduction of new species. Annals and Magazine of Natural History 16: 184-196.

Whelan, S., and N. Goldman. 2001. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. Molecular biology and evolution 18: 691-699.

Wickham, H. 2016. ggplot2: Elegant Graphics for Data Analysis, Springer-Verlag, New York.

Willerslev, E., and A. Cooper. 2005. Review Paper. Ancient DNA. Proceedings of the Royal Society B: Biological Sciences 272: 3-16.

Williams, E. W., E. M. Gardner, R. Harris, III, A. Chaveerach, J. T. Pereira, and N. J. C. Zerega. 2017. Out of Borneo: biogeography, phylogeny and divergence date estimates of Artocarpus (Moraceae). Annals of Botany 119: 611-627.

Wong, T. K. F., S. Kalyaanamoorthy, K. Meusemann, D. K. Yeates, B. Misof, and L. S. Jermiin. 2020. A minimum reporting standard for multiple sequence alignments. NAR Genomics and Bioinformatics 2: lqaa024.

Woodruff, D. S. 2010. Biogeography and conservation in Southeast Asia: how 2.7 million years of repeated environmental fluctuations affect today's patterns and the future of the remaining refugial-phase biodiversity. Biodiversity and Conservation 19: 919-941.

Yang, Z. 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: Approximate methods. Journal of Molecular Evolution 39: 306-314.

Yang, Z. 1995. A space-time process model for the evolution of DNA sequences. Genetics 139: 993.

Zhan, S., C. Merlin, J. L. Boore, and S. M. Reppert. 2011. The monarch butterfly genome yields insights into long-distance migration. Cell 147: 1171-1185.
Zhang, C., M. Rabiee, E. Sayyari, and S. Mirarab. 2018. ASTRAL-III: polynomial time species tree reconstruction from partially resolved gene trees. BMC Bioinformatics 19: 153.
Zhang, J., P. Kapli, P. Pavlidis, and A. Stamatakis. 2013. A general species delimitation method with applications to phylogenetic placements. Bioinformatics (Oxford, England) 29: 2869-2876.

Figure S1. Multiple sequence alignment completeness as reported by Alistat. Information content of the amino acid alignment for each pairwise comparison of specimens depicted separately for the a) reduced dataset, b) optimized dataset and c) mitogenome dataset.

Figure S2. Maximum-likelihood phylogeny of the *amphrysus*-group and outgroup taxa, as inferred by IQTREE 2 on the a) reduced dataset b) optimized dataset and c) mitogenome dataset. Rogue taxa within each dataset are indicated by red arrows.

Figure S3. Maximum-likelihood phylogeny of the *amphrysus*-group and outgroup taxa, as inferred by ASTRAL on 103 gene trees generated with partition-specific substitution and rate models using ModelFinder in IQTREE 2. Branch support is indicated as ASTRAL local posterior probability (localPP) support values.

Figure S4. Results of Four-cluster Likelihood Mapping (FcLM) performed in IQTREE 2 on a) the reduced dataset and b) the optimized dataset. Likelihood mapping was performed fivefold per dataset, (i) on the original alignment, (ii) the alignment with the rogue taxa removed and on the alignment after one of the following three different permutation schemes: (iii) homogenous permutation, (iv) standard permutation and (v) random permutation.

Table S1. Overview of sample age, number of reads recovered, proportion trimmed, number of contigs, contig size distribution, mitogenome quality and BUSCO completeness metrics for each museomics specimen accessed for this study.

Table S2. Maximum-likelihood substitution models as proposed by ModelFinder for the reduced, optimized and mitogenome datasets.

Table S3. Pairwise uncorrected p-divergence of cytochrome oxidase subunit 1 (COI) and NADH-ubiquinone oxidoreductase chain 4 (ND4) sequences between selected subspecies comparisons within each *amphrysus*-group species complex. The number of samples compared is also given.

Appendix IV

This article is accepted with minor revision in *Genes, Genomes, Genetics*

Genomics of the relict species *Baronia brevicornis* sheds light on its demographic history and genome size evolution across swallowtail butterflies

Authors

Alba Marino [1,†], Eliette L. Reboud [1†], Emmanuelle Chevalier [1], Marie-Ka Tilak [1], Jorge Contreras-Garduño [2], Benoit Nabholz [1,3†] and Fabien L. Condamine [1,*†]

Affiliations

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

[2] *Universidad Nacional Autónoma de México, Escuela Nacional de Estudios Superiores, campus Morelia, Antigua Carretera a Pátzcuaro #8701, Col. Ex-Hacienda San José de la Huerta, 58190 Morelia, Michoacán, Mexico*

[3] *Institut Universitaire de France (IUF), Paris, France*

† These authors contributed equally.

Running head: Genomics of the relict *Baronia* butterfly

Corresponding authors:

E-mail: alba.marino@umontpellier.fr; fabien.condamine@gmail.com

Abstract

Relict species, like coelacanth, gingko, tuatara, are the remnants of formerly more ecologically and taxonomically diverse lineages. It raises the questions of why they are currently species-poor, have restrained ecology, and are often vulnerable to extinction. Estimating heterozygosity level and demographic history can guide our understanding of the evolutionary history and conservation status of relict species. However, few studies have focused on relict invertebrates compared to vertebrates. We sequenced the genome of *Baronia brevicornis* (Lepidoptera: Papilionidae), which is an endangered species, the sister species of all swallowtail butterflies, and is the oldest lineage of all extant butterflies. From a dried specimen, we were able to generate both long-read and short-read data and assembled a genome of 406 Mb for *Baronia*. We found a fairly high level of heterozygosity (0.58%) compared to other swallowtail butterflies, which contrasts with its endangered and relict status. Taking into account the high ratio of recombination over mutation, demographic analyses indicated a sharp decline of the effective population size initiated in the last million years. Moreover, the *Baronia* genome was used to study genome size variation in Papilionidae. Genome sizes are mostly explained by transposable elements activities, suggesting that large genomes appear to be a derived feature in swallowtail butterflies as transposable elements activity is recent and involves different transposable elements classes among species. This first *Baronia* genome provides a resource for assisting conservation in a flagship and relict insect species as well as for understanding swallowtail genome evolution.

Key words: conservation genomics, genome size, Papilionidae, relict species, transposable elements.

Article summary

Sequencing the genome of *Baronia brevicornis*, an endangered species and the oldest butterfly lineage, revealed a high level of heterozygosity but a sharp population decline in the past million years. As the sister species of all swallowtail butterflies, this study also highlights genome size variation in Papilionidae and the role of transposable element activity in their evolution. The *Baronia* genome provides valuable resources for understanding swallowtail genome evolution and assisting in conservation efforts.

Introduction

Since Darwin (1859), relict species have captivated evolutionary biologists, who often view them as curious 'living fossils' or remnants of old times. Some plants such as *Ginkgo* (*ca.* 80 million years ago, Ma, Royer et al. 2003) and vertebrates such as the *Solenodon* (*ca.* 60 Ma, Brace et al. 2016) and the coelacanth (*ca.* 75 Ma, Cavin and Guinot 2014) are perhaps the most famous relict organisms. Because of their ancestry, they are thought to provide interesting and important information on a vanished past and can be used to understand the evolution of clades and biotas (Grandcolas et al. 2014). Relict species raise the questions of why they are currently depauperate, have restrained (highly specialized) ecology, and are vulnerable to extinction (often ranked as endangered).

Among invertebrates, the swallowtail butterfly *Baronia brevicornis* Salvin, 1893 (Lepidoptera: Papilionidae) is considered a relict species and is regarded as one of the most mysterious butterfly species in the world (**Figure 1A**). Molecular phylogenetic and phylogenomic studies have shown that *B. brevicornis* is the sister species of all swallowtail butterflies (Condamine et al. 2012; Espeland et al. 2018; Allio et al. 2020a). Depending on divergence time estimates from phylogenies, *B. brevicornis* diverged about 55-75 Ma (Condamine et al. 2012; Allio et al. 2020a), and is the oldest lineage of all extant butterflies, which originated about 100 Ma (Espeland et al. 2018; Chazot et al. 2019; Kawahara et al. 2019). It has been argued that the lineage leading to *B. brevicornis* survived the Cretaceous/Paleogene mass extinction and has maintained a relatively unchanged morphology at least for about 80-90 Ma (Heikkilä et al. 2012; Machkour-M'Rabet et al. 2014; Legal et al. 2015; Galicia-Mendoza et al. 2021). One of the most interesting features of *B. brevicornis*, contrary to all other Papilionidae, is that their caterpillars feed strictly on *Vachellia cochliacantha* (formerly *Acacia cochliacantha*: Fabaceae; **Figure 1B**). According to some results, Fabaceae would be the ancestral host-plant preference of all butterflies (Kawahara et al. 2023) and potentially swallowtail butterflies (Simonsen et al. 2011 but see Condamine et al. 2012; Allio et al. 2021), which highlights the importance of studying *Baronia* to understand early evolution of butterflies.

Although the geographic distribution of its host plant is large, *B. brevicornis* has a restricted natural range (León-Cortés et al. 2004). *B. brevicornis* is strictly endemic to Mexico within tropical dry forests (Machkour-M'Rabet et al. 2014; Legal et al. 2015; Galicia-Mendoza et al. 2021), located in central southern Mexico, including the Rio Balsas (Balsas River) drainage basin that extends from the western state of Jalisco to the southern state of Oaxaca

(**Figure 1C**). There is also a small population in the state of Chiapas, specifically in the Central Valley, close to the city of Tuxtla Gutiérrez (**Figure 1C**). Based on this disjunct distribution (more than 450 km) between the central Mexico and Chiapas populations, previous authors have proposed the presence of two subspecies: the nominal *B. b. brevicornis* Salvin, 1893 and *B. brevicornis rufodiscalis* De la Maza and White, 1987 for the isolated populations of Chiapas. Due to its particular phenotype, mysterious origin and localized distribution, *B. brevicornis* has been listed as a Near Threatened (NT) species since 1986 by the IUCN Red List of Threatened Species but has been recently ranked as Endangered (EN) by the IUCN (Puttick et al. 2018, **Figure 1**).

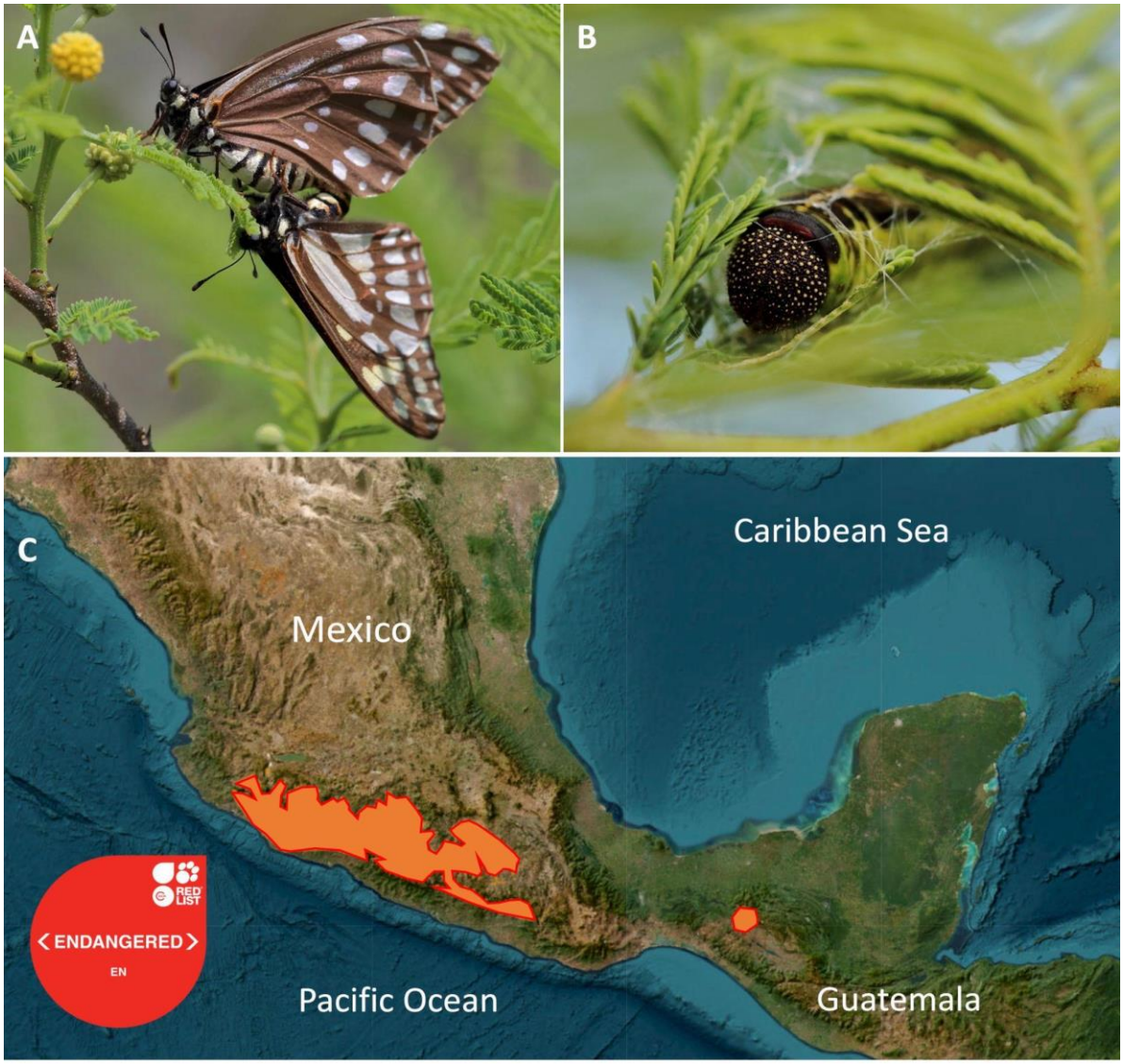


Figure 1. (A) Two adults of *Baronia brevicornis* mating. (B) A caterpillar of *Baronia brevicornis* hiding in a shelter built with leaves of its host plant *Vachellia cochliacanha* (formerly *Acacia cochliacanha*). (C) The distribution range of *Baronia brevicornis* is composed of two populations, shown in orange. Photos: Jorge Contreras-Garduño. The map has been designed on MapMaker: https://mapmakerclassic.nationalgeographic.org based on data from the IUCN Red List https://www.iucnredlist.org/ (Puttick et al. 2018).

Despite being a remarkable and taxonomically important species, knowledge about the ecology and host-plant use has remained very scarce until recently (Machkour-M'Rabet et al. 2014; Legal et al. 2015; Galicia-Mendoza et al. 2021). Yet, natural history data on this species are needed to help establish conservation programs. Some studies have shown that *B. brevicornis* populations cannot occur if the host plant does not cover at least two-thirds of the locality, and that even in the most favorable zones, the occupancy of the butterfly does not exceed 2.5% of the available habitat even when its host plant covers 50% of the area (Legal et al. 2015). Moreover, it has been suggested that the locality where *B. brevicornis* resides should provide not only host plants for the caterpillars but also food sources for the adult butterflies (Galicia-Mendoza et al. 2021). A few molecular studies have assessed the genetic diversity of *B. brevicornis*. Machkour-M'Rabet et al. (2014) have found high genetic diversity within the species. These high values can be explained by the population dynamic of this species characterized by a very high density of individuals over very limited areas (Legal et al. 2015), while inter-population variation is thought to reflect both the age of colonization and locality perturbation level. Genetic structure analysis further revealed three populations, suggesting that both habitat and host-plant specificity probably limit the exchange of individuals between populations.

With the advent of genomic data and analyses, several studies have addressed a range of evolutionary questions using relict species as biological models. For example, the coelacanth genome has been instrumental to reveal tetrapod evolution (Amemiya et al. 2013) or to elucidate how ancestral aquatic vertebrates evolved into terrestrial animals (Nikaido et al. 2013). In addition, genomes of relict and endangered species can also be used to assist conservation actions. For instance, assessing genome-wide diversity provides a good proxy for the loss of adaptive potential that an endangered species has experienced. Similarly, inferring demographic history of an endangered species has emerged as a powerful tool for unveiling possible past bottlenecks in the population dynamics (Nadachowska-Brzyska et al. 2015). Estimation of genome-wide diversity and demography can be estimated by computing heterozygosity and inbreeding depression obtained from whole-genome analyses (Robinson et al. 2022). However, these analyses require generating a relatively high-quality whole genome to obtain less biased estimates of heterozygosity.

Within swallowtail butterflies, whole-genome sequencing has allowed addressing several questions in evolutionary biology and ecology such as identifying molecular adaptations during host-plant shifts (Allio et al. 2021), studying the origin and evolution of color diversity (He et al. 2022), or investigating population genomic and demographic history

(Reboud et al. 2023). The availability of swallowtail genomes also brings the question of genome size evolution across the family (Liu et al. 2020) as we have gleaned knowledge from genome assembly (e.g. Allio et al. 2020a; Podsiadlowski et al. 2021; He et al. 2022). It appears that the subfamily Parnassiinae shows large genome sizes (between 500 Mb to 1.4 Gb), while the subfamily Papilioninae has smaller genome sizes in general (around 300 Mb but can go up to 930 Mb in *Graphium*). Therefore, *Baronia brevicornis*, as the sister subfamily to all other swallowtails, stands out as a key lineage to study genome size evolution across the family Papilionidae. The draft genome of *B. brevicornis* published by Allio et al. (2020a) is around 480 Mb but is only assembled with short-read data at low depth of coverage, which impedes an accurate estimate of its genome size and restrains demographic inferences as well as computation of heterozygosity level.

In this study, we present a first genome assembly of *B. brevicornis* generated with a long-read sequencing approach, complemented with genomic polishing using short-read and homology-based annotation. Using this newly assembled genome, we (1) assess the level of nuclear heterozygosity to be compared with other swallowtail butterflies, (2) estimate the demographic history of the species to correlate with past events (in particular climate fluctuations) by taking into account the ratio of recombination rate over mutation rate, and (3) provide insights into the genome evolution across swallowtail butterflies by annotating transposable elements for 18 available high-quality whole genomes of Papilionidae. This reference genome will enable future population genomic studies with *Baronia brevicornis*.

Materials and Methods

Sample, library preparations and sequencing

We collected in May 2019 a single specimen of the nominal species *Baronia brevicornis brevicornis*, an adult female (voucher FC897), coming from the Northern population of Mexico (Morelos state, south of Mexico City). The specimen was dried and stored in a freezer at -20°C without any additional preservation product. Tissues from the thorax were used to extract high-molecular weight DNA. Following Reboud et al. (2023) who tested two different extraction methods, we used the Qiagen genomic DNA kit to obtain a better 260/230 ratio as estimated with Nanodrop assays guaranteeing DNA purity for long-read sequencing with Oxford Nanopore Technology (ONT). Final DNA purity and concentrations were measured using both Nanodrop (Thermo Fisher, USA) and Qubit (Thermo Fisher, USA).

Whole-genome libraries were constructed using the resulting high-molecular-weight DNA as input for the Nanopore LSK-109 ligation kit (Oxford Nanopore Technologies, UK) following the manufacturer's protocol. Long-read sequencing was performed on a GridION device with three R9.4.1 flow cells. Remaining DNA extractions were sent to Novogene Europe (Cambridge, UK) for two library preparations. Libraries were generated using NEBNext DNA Library Prep Kit following manufacturer's recommendations and indices were added to each library. Genomic DNA was randomly fragmented to a size of 350 bp by shearing, then DNA fragments were end-polished, A-tailed, and ligated with the NEBNext adapter for Illumina sequencing, and further PCR enriched by P5 and indexed P7 oligos. The PCR products were purified (AMPure XP system) and the resulting libraries were analyzed for size distribution by Agilent 2100 BioAnalyzer and quantified using real-time PCR. Since the genome sizes for *Baronia* was estimated to be about 480 Mb (Allio et al. 2020a), Illumina 150 bp paired-end sequencing was run on a NovaSeq 6000 instrument to obtain about 30 Gb per library corresponding to a genome depth-of-coverage of >100x after combining the two libraries.

Genome assembly

For ONT sequencing, all raw long-read sequence data (fast5 files) were basecalled using Guppy 5.0.15 with the super-high accuracy mode and a quality control of 10 (min_score 10). Sequencing adapters were trimmed using Porechop 0.2.3 (https://github.com/rrwick/Porechop). Draft genome assemblies were performed with the long-read assembler Flye 2.8.3 (Kolmogorov et al. 2019, https://github.com/fenderglass/Flye) with default options. The Illumina raw reads were cleaned, filtered, and paired using fastp 20.0 (Chen et al. 2018) with default options. To improve base accuracy and reduce assembly errors, the long-read Flye draft assemblies were polished with POLCA (Ziming and Salzberg 2020) implemented in MaSuRCA 4.0.1 (Ziming et al. 2013). Assembly statistics were then assessed using the gVolante2 platform (Nishimura et al. 2017) to retrieve the number and size of contigs, the presence, completeness, and duplication of BUSCO genes of the Lepidoptera *odb10* database (Manni et al. 2021). We also checked for haplotype duplication using the spectra-asm plot with Merqury (Rhie et al. 2020). Merqury identifies the number of times each k-mer identified in the reads is present in the assembly allowing an assessment of whether some k-mers are present twice or more in the assembly, indicating the presence of haplotype duplication. Before submitting genomes assemblies to GenBank, we checked for possible contaminations using BlobTools 1.1.1 (Laetsch and Blaxter 2017) set to the ncbi and diamond

databases. We found no evidence of artificial contamination coming from laboratory manipulation, but some contigs were clearly identified as belonging to exogenous organisms such as host plants and symbionts (Figure S1). We removed all contigs that were belonging to Bacillota (formerly Firmicutes) (all belonging to Order Lacterobacillales) or Pseudomonadota (all belonging to Order Enterobactererales) phylum (Table S1).

We used GetOrganelle 1.7.7 (Jin et al. 2020), with the cleaned and paired short-reads and a *Baronia brevicornis* reference mitogenome from a previous study (GenBank accession number: LT999970, Condamine et al. 2018) to assemble a new *B. brevicornis* mitogenome. The resulting assembly was given to MitoFinder 1.4 (Allio et al. 2020b) to annotate and extract protein-coding genes, tRNA and rRNA genes. The new mitogenome of *Baronia brevicornis* is OR063968. This mitogenome was aligned with the mitogenome of *B. brevicornis* LT999970 (Condamine et al. 2018) and variable sites were counted with the "Statistics" option in Seaview (Gouy et al. 2010).

Genome annotations and whole-genome alignment

We performed the MAKER2 pipeline of gene annotations. First, the repeat sequences were reconstructed using RepeatModeler 2.0.1 (Flynn et al. 2020). These newly identified repeats were used to annotate the repeat sequences using RepeatMasker (Smit et al. 2013-2015) in association with the Dfam libraries (Storer et al. 2021) setting the parameter "-species Arthropoda". Second, we ran MAKER 2.31.11 (Holt and Yandell 2011) using the repeat annotated by RepeatMasker and homology information obtained with the protein sequences of *Heliconius melpomene*, *Melitaea cinxia*, *Papilio machaon*, *Papilio xuthus*, *Papilio glaucus*. Third, SNAP (Korf et al. 2004) and AUGUSTUS (Stanke et al. 2006) were used to produce gene prediction models from the first round of MAKER. BUSCO 5 (Simão et al. 2015) with options "--long" and "--augustus" and the Endopterygota database was used to produce the gene prediction model of AUGUSTUS. Finally, we ran again MAKER using the annotation from the first round and the gene models of SNAP and AUGUSTUS to produce a final round of annotations.

We also used the BRAKER2 pipeline, specifically designed to perform annotation for assembly without RNA-seq data (Brůna et al. 2021). This pipeline uses the assembly soft-masked for the repeat sequences and a large set of coding sequences from distantly related species to perform the annotation. Here, we used the Arthropoda set of OrthoDB 11 (Kuznetsov et al. 2023) available at https://bioinf.uni-greifswald.de/bioinf/partitioned_odb11/ in which we added the sequences of *Papilio xuthus* and *Ornithoptera alexandrae.*

As quality control, we evaluated the mismatches between the hemizygous Z contigs and Illumina read consensus. We performed a genome assembly with the Illumina reads only using MEGAHIT (Li et al. 2015). We then aligned the MEGAHIT contigs onto our long-read-based assembly using Minimap2 (Li 2018). Alignments were filtered based on maximal quality (60) and must represent more than 80% of the MEGAHIT contig size. We then estimated the divergence between the Z contigs and the Illumina reads.

Finally, we performed whole-genome alignments using Progressive Cactus (https://github.com/ComparativeGenomicsToolkit/cactus) (Armstrong *et al.* 2020) with the genomes of *Papilio bianor* (Lu et al. 2019), *P. machaon* (Lohse et al. 2022), and *Ornithoptera alexandrae* (Reboud et al. 2023). The contigs will less that 10% of their size aligned to Z or W chromosomes were excluded. This alignment will allow identifying sequences of *Baronia* genome assembly that are homologs to chromosome Z and W of *Papilio* assemblies and Z of *Ornithoptera* assemblies.

Nuclear heterozygosity

Genotype calling was performed for both Illumina and Nanopore data. For Illumina data, read mapping was performed using the SpeedSeq pipeline (Chiang et al. 2015) that relies on BWA (Li 2013) excluding duplicated reads using SAMtools (Li et al. 2009). Genotype calling was performed using Freebayes 1.3.2 (Garrison and Marth 2012). We excluded positions with a coverage below 15x and higher 200x. SNP with a quality below 200 were excluded. For ONT data, read mapping was performed using Minimap2 (Li 2018) and genotype calling with LongShot 0.4.1 (Edge and Bansal 2019), using a threshold of 10x minimum and 150x maximum for the depth of coverage (minimal quality of 50) and applying a transition/transversion rate for genotype prior estimation (*ts_tv_ratio*) of 2.0 (Edge and Bansal 2019). SNP with a quality below 300 as provided by Freebayes in the VCF were excluded. For ONT, a SNP is considered only if the alternate allele is supported by at more than 20% of the reads and less than 80% of the reads. We used python programs VCF2fasta_no_mono.py and VCF2FastaLongshot.py (https://github.com/benoitnabholz/VCF2Fasta) to convert the VCF file into fasta files for Illumina and ONT data, respectively. Next, sites annotated as repeat by RepeatModeler/RepeatMasker were excluded. Finally, heterozygosity was computed on the whole genome using the program heterozygosity.py (https://github.com/benoitnabholz/popgen_python) or focusing on BUSCO single-copy orthologs to compute synonymous and 4-fold degenerated heterozygosity using the programs

seq_stat_coding (https://github.com/benoitnabholz/seq_stat) and *selectClassSite* (https://github.com/benoitnabholz/selectClassSite).

Estimation of the demographic history and effective population size

We relied on a Sequential Markovian Coalescent (SMC) model (Schiffels and Wang 2020) to estimate the ancestral effective population size (*Ne*) of *Baronia*. The SMC model needs to be calibrated, in particular with a value of mutation rate. However, it has been shown that SMC models do not perform well when the ratio of recombination rate (r) over mutation rate (μ) becomes greater than one (Sellinger et al. 2021). This phenomenon seems to be prevalent in invertebrate genomes. Assuming a single crossover per tetrad per male meiosis and 30 chromosomes, Reboud et al. (2023) have estimated that the average recombination rate for *Ornithoptera alexandrae* (a swallowtail butterfly) is 2.7e-8, which is more than ten times higher than its average mutation rate estimated at 1.316e-9 mutations per site per generation. To investigate the sensitivity of SMC analyses to parameters, Reboud et al. (2023) have simulated data with the range of recombination and mutation rate parameters with r being ten times higher than μ such as those observed in invertebrates. Fitting a SMC model with default options (*-rhoOverMu = 0.25*) versus an adjusted ratio of r over μ (*-rhoOverMu = 10*) recovered different demographic histories, with the model using the adjusted ratio showing a good fit to the simulated data (Reboud et al. 2023). Accordingly, the ratio of r over μ must be well adjusted to recover trustworthy demographic inferences in invertebrates.

It is likely that *Baronia* has a mutation rate close to *Ornithoptera* as it is also in the range of the mutation rates estimated for *Heliconius* (Nymphalidae) between 1.3e-9 and 5.5e-9 (Keightley et al. 2015). In addition, *Baronia* is the only species in the subfamily Baroniinae, which is sister to all remaining Papilionidae (Allio et al. 2020a, 2021), and is thus too distantly related to any extant species to be able to accurately estimate the mutation rate using neutral divergence. Therefore we used the same mutation rate of *Ornithoptera alexandrae* (μ = 1.3e-9). Assuming *B. brevicornis* genome is 406 Mb long, distributed in 30 chromosomes and that there is a single crossover per tetrad per male meiosis, the recombination rate would be r = 3.38e-8.

We relied on the Multiple Sequentially Markovian Coalescent (MSMC) model as implemented in MSMC2 (Schiffels and Wang 2020; https://github.com/stschiff/msmc2). We used the VCF files generated using Longshot (as described in the *Nuclear heterozygosity* section) and created the so-called "mask file" for each individual based on the depth of coverage thresholds of >20x and <150x using a custom python script (available in FigShare).

These files were then combined using the "generate_multihetsep.py" of MSMC2 to generate "multihetsep.txt" input files (https://github.com/stschiff/msmc-tools/blob/master/msmc-tutorial/guide.md). We followed the recommendations of Reboud et al. (2023) using MSMC2, which was thus set with the option *-rhoOverMu = 10* to better account for the ratio r over μ that is higher than 10. Contigs shorter than 500 kb and of the Z chromosomes were excluded representing a total of 171 Mb (42% of the genome size). We generated 10 bootstraps using the multihetsep_bootstrap.py (https://github.com/stschiff/msmc-tools) and generated all graphs with the R package ggplot2 (Wickham 2016) by considering a generation time of 1 for *B. brevicornis* (Legal et al. 2015).

Transposable elements and repeat sequence dynamics

Transposable elements (TEs) were annotated for the genome assemblies of *Baronia* and 17 other high-quality genomes of swallowtail butterfly species (Li et al. 2015; Nishikawa et al. 2015; Liu et al. 2019; Podsiadlowski et al. 2021; He et al. 2022; Mackintosh et al. 2022; Reboud et al. 2023; see Table S2 for details). We used Earl Grey 1.2 (https://github.com/TobyBaril/EarlGrey), which is a pipeline combining several tools for TE detection and performing automated consensus curation. Firstly, repeats in the assembly are masked with RepeatMasker 4.1.2 by homology to known metazoan sequences from Dfam 3.5 and RepBaseRepeatMaskerEdition-20181026 (-r metazoa). RepeatModeler 2.0.3 is then employed on the hard-masked genome to search novel TE families. The new detected sequences undergo an iterative process of consensus refinement and elongation (-i 5 -f 1000, default options): this aims to reduce the redundancy and improve the quality of the *de novo* library. Finally, the unmasked genome is mined with LTR_finder 1.0.7 and again with RepeatMasker using the combined curated *de novo* and public metazoan libraries. The resulting annotations are merged and defragmented and overlappings removed. For a detailed description of the pipeline see Baril et al. (2022). Earl Grey computes the genetic distance between the annotated TE copy and their respective consensus sequences. Consensus sequences correspond to a reconstruction of the ancestral sequences of the TEs and, therefore, the genetic distance could be used as an estimation of the age of the TE insertion in the genome. To compare the genetic distance between the TEs and their consensus sequences and the divergence among species, we compute the genetic distance among the species using the same method. The alignments of TEs to their consensus produced by the second RepeatMasker run were extracted with a custom script (available in FigShare) and used to compute Kimura 2P distance (i.e. Kimura K80) copy by copy with the *dist.dna* function of the R package ape 5.7-

1 (Paradis and Schliep 2019). To estimate the divergence among species, we used the orthologous sequences provided by BUSCO (Manni *et al.* 2021) selecting the 524 genes present in all species. BUSCO genes were aligned using the OMM_MACSE pipeline that is a codon aware alignment method (Ranwez et al. 2018; Scornavacca et al. 2019). Third-codon positions were extracted from the alignment and the same pairwise genetic Kimura 2P distance between species was computed. A phylogeny of the 18 species was obtained from Allio et al. (2020a, 2021) and associated with the obtained median genetic distances. The interspecies distances were then compared to the genetic distances between TEs and their consensus to determine if TE insertions were anterior or posterior of species divergences.

Results and Discussion

Genome sequence statistics

For one adult female specimen collected in 2019, stored in dry conditions, we sequenced the DNA combining a mean of 54x of long reads (Oxford Nanopore, 32 Gb: N50 = 3,027 bp, the mean = 2,160 bp, the median = 1,585 bp, and 1.16 Gb of reads have a length >10 kb) for draft assembly and 141x of short reads (Illumina, 56.9 Gb) for polishing (see *Materials and Methods*). Using Flye assembler (Kolmogorov et al. 2019) and POLCA polisher (Ziming and Salzberg 2020), we assembled the genome of *B. brevicornis* that is 406 Mb, which has 4,834 contigs (after removing 152 contigs that were detected as exogenous) and a N50 of 0.4 Mb (**Table 1**). Over a total of 5,286 core genes of the Lepidoptera database (*odb10,* Manni et al. 2021), BUSCO recovered 93.7% single complete genes, 1.1% duplicated genes, 1.3% fragmented genes and 3.9% missing genes (**Table 1**). The very low proportion k-mers present in the Illumina reads were present twice or more in the assembly confirming a very low proportion of haplotype duplication in the assembly (Figure S2). The genome size and gene completeness of our *B. brevicornis* assembly has a lower quality though fairly comparable to previously published genomes of swallowtail species: *Papilio demoleus* (Papilionini: 240 Mb, BUSCO recovered 98.1% single complete genes, 0.2% duplicated genes, 0.7% fragmented genes and 1.0% missing genes), *Troides helena* (Troidini: 330 Mb, BUSCO recovered 95.9% single complete genes, 0.3% duplicated genes, 0.5% fragmented genes and 3.3% missing genes), *Lamproptera curius* (Leptocircini: 550 Mb, BUSCO recovered 89.3% single complete genes, 0.2% duplicated genes, 1.7% fragmented genes and 8.8% missing genes), or *Parnassius orleans* (Parnassiini: 1.18 Gb, BUSCO recovered 92.2% single complete genes, 1.1% duplicated genes, 1.2% fragmented genes and 5.5% missing genes), which were assembled

with similar data and methods (He et al. 2022). Furthermore, the genome size of *Baronia* stands between the generally small genomes in the subfamily Papilioninae and the large genomes mostly found in the subfamily Parnassiinae (Liu et al. 2020; He et al. 2022), which brings questions on the evolutionary dynamic of genome size across the family Papilionidae. Gene annotation with MAKER and BRAKER2 leads to 14,362 and 17,747 annotated protein-coding sequences of mean length 7,150 bp and 6,703 bp, respectively (available at https://figshare.com/s/f26e4b4a9d0e47e2832d). The BUSCO scores for the protein sets were 93.7% complete including 91.1% single and 2.6% duplicated for BRAKER2 and 92.5% complete including 90.0% single and 2.5% duplicated for MAKER.

Table 1. Assembly statistics for the genomes of *Baronia brevicornis*. LR = long reads, SR = short reads, bp = base pairs. For BUSCO scores, S = single-copy genes, D = duplicated genes, F = fragmented genes, and M = missing genes out of 5,286 genes in *odb10* lepidopteran database. Amount of raw data and final mean coverage of Voucher 167 were approximated from Allio et al. (2020a).

	Raw data sequenced (Gb) (LR + SR)	Final mean coverage (LR + SR)	Assembly size (bp)	Number of contigs	N50 (bp)	Max length (bp)	Number of gaps (≥5 N's)	BUSCO score (%)
Voucher 167 (Allio et al. 2020)	0 + ~11	0 + ~23x	488,028,434	973,148	1,169	33,471	35,404	S:50.5; D:0.5; F:19.2; M:29.8
Voucher FC897 (this sudy)	31.96 + 56.9	54x + 141x	405,627,949	4,834	412,514	2,621,275	39	S:93.7; D:1.1; F:1.3; M:3.9

Identification of sex chromosomes

We sequenced a female, allowing us to identify the Z/W chromosomes based on coverage and heterozygosity information. Median coverage after read cleaning and excluding mapping position quality lower than 50.0 and base quality lower than 30.0 leads to 54x for ONT data and 75x for Illumina data (after removing PCR duplicates). The classification was made difficult due to the strong correlation between GC content and depth of coverage observed with both ONT and Illumina data (Figure S2). However, a visual inspection of the relationship allows identifying two categories of contigs with one category having a lower depth of coverage considering their GC content (contigs above the line in Figure S2). These contigs are probably from the sex chromosome. This is confirmed by the lower heterozygosity (=3e-4) and 76.3% of the contigs (148 out of the 194) have no SNP. Finally, contigs that have been aligned to the Z chromosome of *Papilio bianor* (contig no. 30) and of *P. machaon* all fall in this category. In contrast, none of the five contigs (392 kb in total) aligned to the W chromosome of *P. machaon* show a pattern of coverage or genetic diversity compatible with a haploid status, thus corresponding to autosomal regions (Figure S3). One *Baronia* specimen was previously

sequenced, but at low (25x) coverage, and was an adult male (NCBI accession SRR8954515; Allio et al. 2020a). We compared this male with the present female specimen to identify and exclude a few more contigs that belong to the Z chromosome (Figure S4; in total 193 contigs representing 18 Mb).

Finally, the haploid Z chromosome sequence provides an opportunity to evaluate our genome quality by quantifying the mismatch between the Illumina reads and the Z contigs. We found a median divergence of 0.02%, corresponding to 2 differences per 10 kb, between the contigs obtained using the Illumina reads and the Z contigs of long-reads assembly. Assuming that this divergence provides an estimate of the assembly error rate, it is 25 times lower than the estimated heterozygosity level.

Baronia has unexpectedly elevated genomic diversity levels

Using Illumina data and annotation with MitoFinder, the *Baronia* mitogenome was reconstructed with a base coverage of 468x and we retrieved all genes including the 13 protein-coding genes and 2 rRNA genes. Comparing this new mitogenome with the previous mitogenome of the same population (LT999970), the mitogenomic diversity (π-diversity) including coding and non-coding regions was calculated at ~1.6% (with 236 variable sites over 14,728 aligned positions). For comparison, the mitochondrial diversity of the endangered butterfly *Ornithoptera alexandrae* is 22 times lower (~0.0704%; Reboud et al. 2023) than that of *Baronia brevicornis*. Recovering an elevated mitogenomic diversity for an endangered species is unexpected, in particular when comparing individuals of the same population. However, it is difficult to conclude with mitogenomic evidence only because we lack a large-scale estimation of mitochondrial diversity based on whole mitogenomes for butterflies since studies estimating mitochondrial diversity usually relied on the cytochrome *c* oxidase subunit I DNA barcode marker (e.g. Dincă et al. 2021 for European butterflies). In addition, high mitochondrial diversity does not necessarily equate to high autosomal diversity (Mackintosh et al. 2019).

Excluding the Z chromosome and contigs shorter than 30 kb, Illumina and ONT data lead to very similar estimates of autosomal heterozygosity around 0.59% and 0.56%, respectively. Heterozygosity is highly variable among contigs. Even selecting contigs larger than 200 kb, heterozygosity varies from less than 0.25% to 1.0% among contigs. Focusing on the BUSCO single-copy orthologs, synonymous nucleotide diversity is 0.78%, nucleotide diversity of 4-fold degenerated position is 0.75% (n=4,533 genes), and the ratio of non-

synonymous nucleotide diversity over synonymous nucleotide diversity (pN/pS) is 0.136. Compared to other butterflies, the genetic diversity of *Baronia* is relatively low (Mackintosh et al. 2019) but is much higher than other endangered swallowtail butterflies (Reboud et al. 2023).

Our understanding of the determinants of heterozygosity remains debated (e.g. Galtier and Ellegren 2016; Mackintosh et al. 2019). Neutral genetic diversity is proportional to effective population size and mutation rate. Therefore, range size, endangered status or life-history traits may influence *Ne* and be associated with the level of genetic diversity (e.g. Romiguier et al. 2014; Mackintosh et al. 2019; Buffalo 2021). Given the specific features of *B. brevicornis* like its relict status, ancestral habits, and endangered status, one may be tempted to explain the inferred heterozygosity level. Indeed, the endangered swallowtail *Luehdorfia taibai* has a restricted distribution range in the Qinling Mountains in China (Fang et al. 2019) and shows a very low heterozygosity level of 0.057%, similar to that of the giant panda (Guan et al. 2022). However, it seems that species' specific traits do not correlate well with heterozygosity levels in swallowtail butterflies (Mackintosh et al. 2019). For instance, in the threatened troidine swallowtails (*Ornithoptera alexandrae*, *O. priamus*, and *Troides oblongomaculatus*), there are heterogeneous levels of heterozygosity ranging from a very low level (0.0737% autosomal, 0.0704% neutral for *T. oblongomaculatus*) to a medium level (autosomal 0.433%, neutral diversity 0.708% for *O. priamus*). Within the species-rich genus *Papilio*, these heterozygosity levels are much higher with estimates ranging from 1.0 to 2.3% even if estimated with a different method (Lu et al. 2019). Therefore, our estimate of the heterozygosity level for *Baronia* cannot be explained by the species' range or body size that could have an indirect effect on *Ne* or mutation rate, as suggested in Mackintosh et al. (2019).

In a genetic study, Machkour-M'Rabet et al. (2014) also found a high level of heterozygosity and suggested that the peculiar population dynamic of *Baronia* in high population densities over very restricted areas can explain such a high genetic diversity at the species level. This important density of individuals would favor significant intra-population genetic diversity over a long period of time. Finally, *Baronia* is well known for its color polymorphism with three male and three female color wing morphs observed in the Northern population (Galicia-Mendoza et al. 2021). This raises the question of: How are multiple morphs maintained in local populations against genetic drift that leads to loss of genetic variation and recombination that breaks up non-random trait associations? Balancing selection such as negative frequency-dependent selection (rare-morph advantage) or over dominant selection (heterozygote advantage) could be at play. One possible research avenue would involve a

population genomic study to link this color polymorphism with fundamental evolutionary processes like frequency-dependent selection, gene flow, recombination, and to better understand the maintenance of heterozygosity level. With more genomic data, it would also be possible to compare with other butterfly cases, such as in the genus *Papilio* (Kunte et al. 2014; Timmermans et al. 2017), to better understand the origin and maintenance of sex-limited morphs.

Sharp decline in the demographic history of Baronia in the last million years

We estimated how the genetic diversity translates into the demography dynamic for the species. Given the extinction risks and the relict status of *Baronia*, we expected a prolonged decline of effective population size such as those inferred for the endangered Queen Alexandra's birdwing butterfly (*O. alexandrae*, Reboud et al. 2023) or the Apollo butterfly (*Parnassius apollo*, Kebaïli et al. 2022); the latter has population declines throughout different mountain massifs even with high heterozygosity levels. However, recent studies have also unveiled long-term low but stable effective population size for endangered species such as in the vaquita porpoise (Morin et al. 2021; Robinson et al. 2022). Except the cases on *O. alexandrae* (Reboud et al. 2023) and *Parnassius apollo* (Kebaïli et al. 2022), there are still few examples of demographic history in insects, other than pest insects (but see Walton et al. 2021; Manthey et al. 2022; García-Berro et al. 2023).

Demographic analyses with MSMC2 traced relatively similar demographic histories for *Baronia* with Illumina and ONT data analyzed independently (**Figure 2**, Figure S5). The effective population size *Ne* of *Baronia* seems to have been at a low but continuously increasing number from ~1,500,000 to 2,500,000 in effective population size from its origin until the last million years, when the *Ne* reached its maximum. However, we inferred a drastic demographic decline in the last million years or so (from ~900,000 years with Illumina data to ~2 million years with ONT data) that stabilized to low *Ne* (less than 250,000 individuals) around 120,000 years ago until the present (**Figure 2**). This demography dynamic resembles that of endangered vertebrate species such as the brown hyena (Westburry et al. 2018) or the Californian condor (Robinson et al. 2021), whose effective population sizes were more elevated but declined in the last million years. Altogether, these results suggest that the ancestral effective population of *Baronia* has been larger than today, thus suggesting that its distribution range was probably larger.

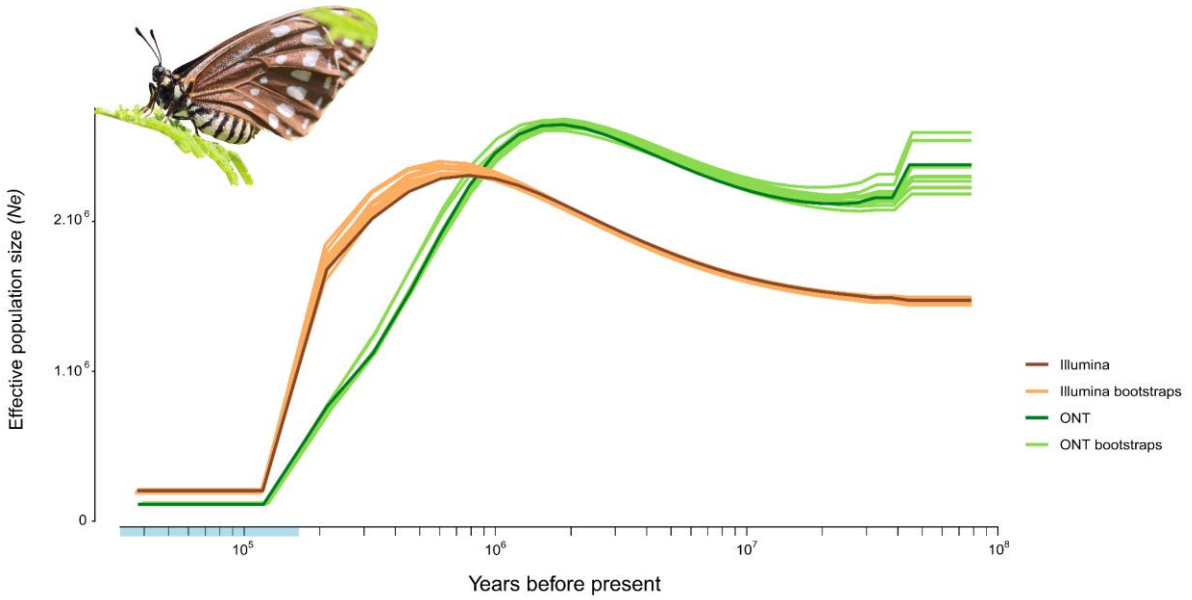


Figure 2. Estimated historical demography of *Baronia brevicornis*. MSMC2 estimates of the effective population size (*Ne*) with both Illumina and Nanopore data. Bootstraps are represented in clear lines. MSMC2 analyses were performed with *-rhoOverMu = 10* to consider a recombination rate ten times higher than the mutation rate. The discrepancy between the Illumina and ONT datasets is likely due to the repeated sequences (Figure S5). The pale blue rectangle along the time bar indicates the limits of the last glaciation period in the late Pleistocene. Photo: Jorge Contreras-Garduño.

Temporal variations of effective population sizes are usually compared with past climatic fluctuations such as temperature and/or sea level, in line with Quaternary glaciations (e.g. Nadachowska-Brzyska et al. 2015; Westbury et al. 2018; Morin et al. 2021). We can tentatively associate the inferred *Ne* variations of *Baronia* with the Pleistocene glaciation cycles, although it remains difficult to extract a correlation because of climatic heterogeneity and uncertainties on demographic parameters estimates. The last million years was mostly a glacial period that has also been documented in Mexico (e.g. Vazquez-Selem and Heine 2004). A drop in temperature of 5–9°C supports marked cooling over tropical land and oceans during the Pleistocene (Vazquez-Selem and Heine 2004). The cooling period coincides with the decrease of *Ne* for *B. brevicornis*.

Effective population sizes have also been compared with conservation status from IUCN data (Nadachowska-Brzyska et al. 2015; Wilder et al. 2023). It has been suggested that historical demography can inform contemporary resilience. For instance, mammal species with small historical effective population sizes show a large burden of deleterious alleles due to long-term accumulation and fixation of genetic load and have a higher risk of extinction

(Wilder et al. 2023). The strong demographic contraction inferred with both types of data for *Baronia* is in line with the ideas that historical population size can be relevant to contemporary extinction risk and that genomic information can help predict extinction risk.

Genome size evolution in Papilionidae

Genome size is extensively variable among species of Papilionidae (Liu *et al.* 2020; Podsiadlowski *et al.* 2021). The largest assembly (*Parnassius apollo*, 1,392 Mb) is 5.7 times larger than the smallest genome (*Papilio xuthus*, 244 Mb). In accordance with other genomic studies (Petersen et al. 2019; Wu and Lu 2019; Cicconardi et al. 2023), this variation is largely explained by transposable elements (Pearson's correlation coefficient = 0.98; correlation coefficient under phylogenetic generalized least squares = 0.47, $P < 0.0001$), with the genomes of *P. apollo* and *P. xuthus* composed of 67% and 15% of transposable elements, respectively. However, the non-repeat fraction of the genome is also variable among species and is not closely following overall genome size. For example, *Lamproptera curius* has a genome of 623 Mb with only 32% of annotated TEs whereas the similar size *Sericinus montela* (594 Mb) has 50% of TEs (**Figure 3**, Table S2). This can be due to ancient TEs persisting in the genome: indeed, our analysis is limited to relatively young TEs that could be recognized as such, while it is not able to recover old TEs which have substantially diverged from their ancestral sequence. However, the fact that genome size variation is mostly explained by TE content suggests that unannotated TE relics probably just moderately contribute to such variation. Another limit lies in the quality of the genome: as repeated sequences are the main responsible for gaps, the fragmented assembly of Baronia could hinder the recovery of its complete TE complement. However, the genome size estimated using GenomeScope is close to our assembly size (380 Mb for GenomeScope and 405 Mb for the assembly), so it is perhaps unlikely that a large proportion of TE sequences is missing.

Among the known TE subclasses, we find Long Interspersed Nuclear Elements (LINEs) to be the most abundant across all species (11% of the genome assembly on average), followed by DNA and Short Interspersed Nuclear Elements (SINEs) (3%). At the same time, a high variability of TE landscapes is generally reported across and within insect orders (Petersen et al. 2019; Gilbert et al. 2021): this is also found at the level of swallowtail butterflies as some species show lineage-specific element expansions as compared to their closer relatives. For instance, *Papilio protenor* displays an increase in Rolling Circle (RC) elements (9%), *Luehdorfia chinensis* has 11% of its genome covered by SINEs, and *P. apollo* genome underwent an expansion of Long Terminal Repeat elements (LTRs) (7%). It should be noted

that in many cases unknown elements take up a relevant portion of the genome, partly limiting our understanding of TE variability. Indeed, unknown elements are putative TE sequences that were identified as repeated but could not be classified by the pipeline, and correspond in broad terms to new, lineage-specific TE families. Because manual curation of *de novo* TE libraries was not performed, in some cases unknown sequences might be however spurious hits resulting for example from gene duplications (Goubert et al. 2022).

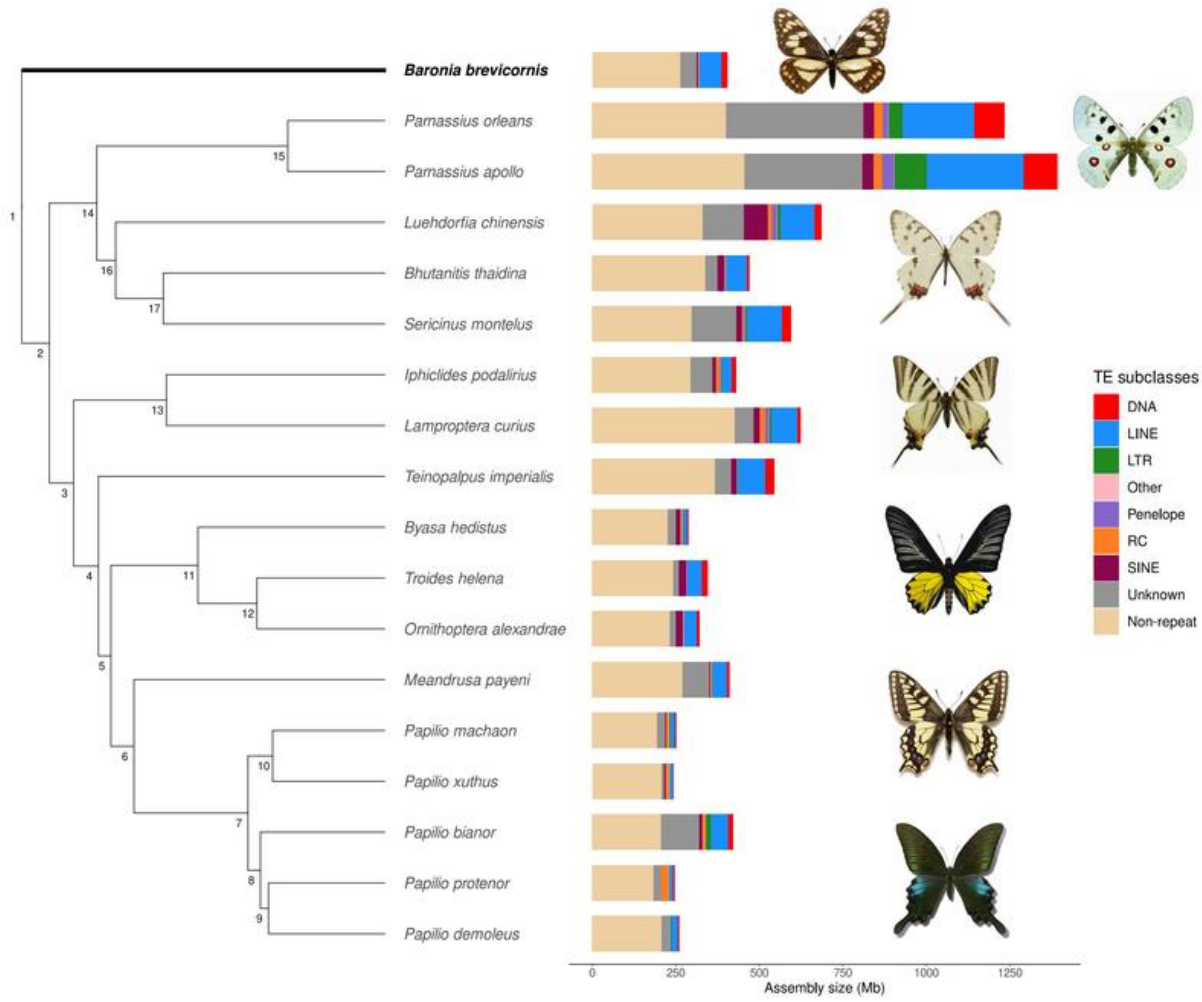


Figure 3. Variation in genome sizes (Mb) and transposable elements content across Papilioninae, Parnassiinae, and the monotypic Baroniinae represented by *Baronia brevicornis*. Node numbers correspond to the divergence points between species calculated as median pairwise Kimura 2P distance and reported in Figure 4 and Figures S3-S6. The phylogeny was extracted from Allio et al. (2020a, 2021). 'Other' includes simple repeats, microsatellites, and RNA sequences. LINE, Long Interspersed Nuclear Elements; LTR, Long Terminal Repeats; RC, Rolling Circles; SINE, Short Interspersed Nuclear Elements. Photos: Fabien L. Condamine.

Given that genome size has a patchy distribution across the phylogeny of Papilionidae and TEs appear to be major determinants of genome size variation, the condition of the common ancestor of swallowtail butterflies is uncertain: current genome size differences in this group could have arisen either through a differential erosion of a genome originally replete with TEs, or through derived TE expansions inflating genomes in certain lineages.

To understand whether the contribution of TEs to genome size variation is more recent or ancient, we investigated the distribution of TE copies' ages according to their distance from the "parental" sequence that originated upon transposition in each species. This analysis was coupled with the assessment of the approximate divergence between species to pinpoint shared and lineage-specific patterns of TE expansion. Because elements diverging by more than 50% from the consensus generally cannot be detected, TE landscapes could be compared only between the most closely related species (**Figure 4**, Figures S6-S9). Despite their modest sizes, most of the TEs in the genomes of Papilionini seem to have originated independently after lineage splits in all species: RC mostly characterizes the recent activity in *P. xuthus* and *P. protenor* genomes, while LINEs are mainly active in the other species (**Figure 4**, Figure S6). Within Troidini, mainly LINEs and SINEs were active before the split of the three species (node 11, 28% divergence), after which *B. hedistus* lineage accumulated TEs independently from *O. alexandrae* and *T. helena* which share a clear TE burst before their separation (node 12, 11% divergence) (**Figure 4**, Figure S7). As for Zerynthiini and Luehdorfiini, most of the TEs seem to have accumulated independently in each lineage (node 17, 36% divergence) (**Figure 4**, Figure S8). Finally, in the two *Parnassius* genomes both common and novel TE expansions can be observed. Overall, the TE profiles of all species point at a mostly recent and lineage-specific TE activity as responsible for the present genome size variation across swallowtail butterflies and strongly suggest that big genomes such as those of Parnassiini are derived. That being said, it should be noted that this analysis provides landscapes of relatively recent TEs, which are able to explain current genome size variation in swallowtail butterflies. Indeed, we found that the TE activity is recent or ongoing even in Troidini, in the genera *Papilio* and in *Baronia*, which all have typically small-sized genomes. As it is clear that TE dynamics actively fuel current genome plasticity, we cannot rule out similar TE outbreaks that we are not able to track down with current methods to have shaped the early genome evolution of Papilionidae. Moreover, it should be emphasized that changes in genome size are the result of both repeat sequences expansion and deletion processes acting jointly (Kapusta et al. 2017): the high turnover rate of elements reported in Nymphalidae (Lavoie et al. 2013; Baril and

Hayward 2022), and the reconstruction of both genomic expansion and shrinkage events along the evolution of the same family (Cicconardi et al. 2023) support this view.

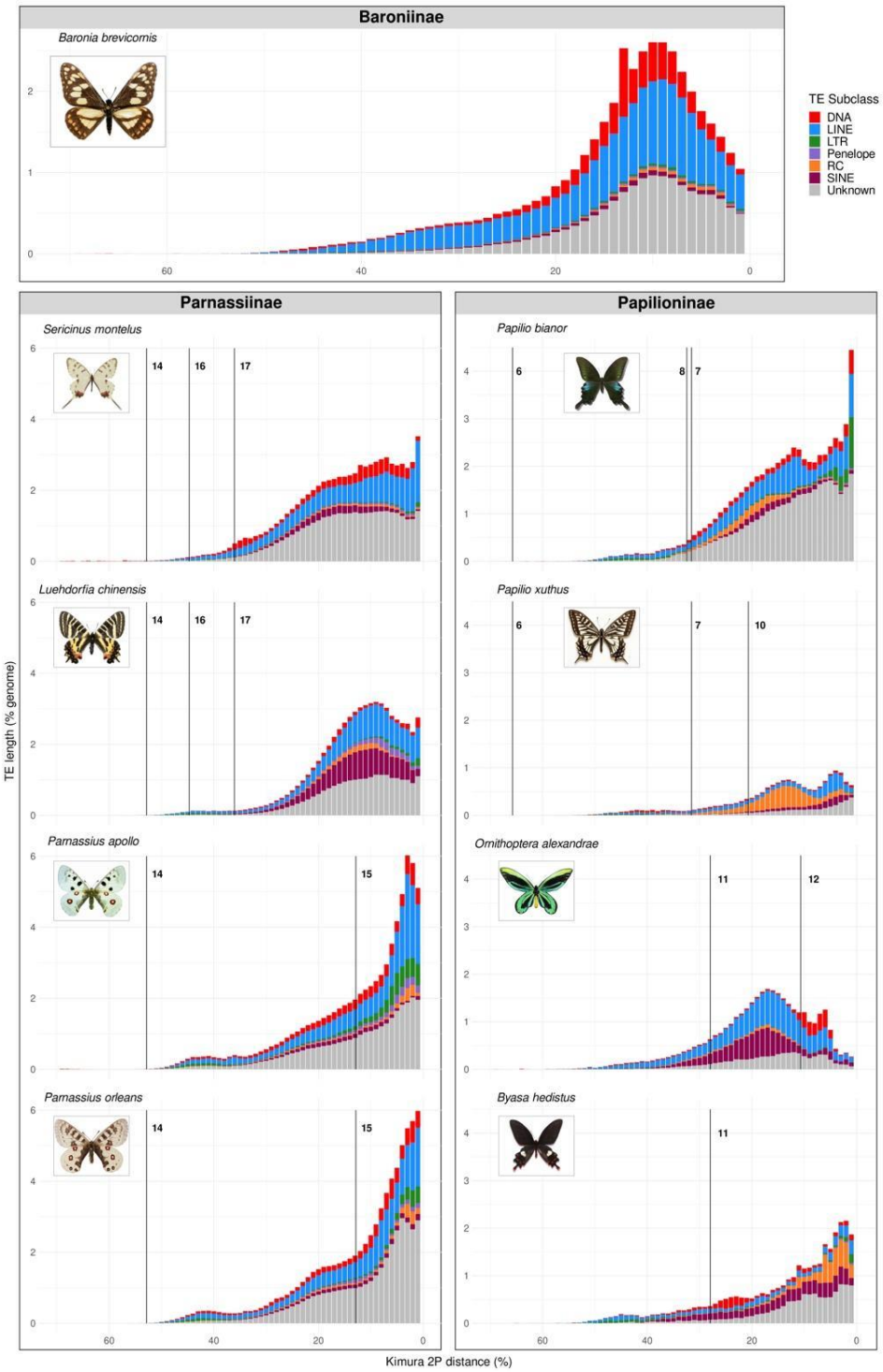


Figure 4. Distribution of transposable element copies in genome percentage according to their divergence from consensus. *Baronia*'s landscape is shown along with some of the main lineages from the two other subfamilies whose landscapes can be compared by the divergence nodes shown in Figure 3. Low distance copies (right side of the plots) correspond to recently active ones, while highly divergent copies (left side of the plots) are remnants of old TE activity. Photos: Fabien L. Condamine.

Why the genomes of some groups tend to be refractory to repeat accumulation, while others do not, remains an open question. Lynch and Conery (2003) proposed that TEs might be able to accumulate in lineages with low effective population size as an effect of the reduced efficacy of selection against these mildly deleterious mutations. The reduced polymorphism associated with increased genome size in some populations of the genus *Leptidea* agrees with this view (Talla et al. 2017). However, no relationship between genetic diversity and genome size was observed when considering a wide set of lepidopteran species (Mackintosh et al. 2019). Small-RNA-based epigenetic regulation is known to be an important and widespread mechanism put in place by the host to tame selfish element proliferation (Blumenstiel 2011). As horizontal transfer of TEs seems to occur very frequently in Lepidoptera (Reiss et al. 2019), this might be an important factor repeatedly facilitating genome size variation in this group. Indeed, it has been suggested that foreign elements coming from distantly related taxa could be able to elude the host defense machinery and more promptly invade the new genome (Schaak et al. 2010; Venner et al. 2017). Given the dynamicity of their genomes, swallowtail butterflies are therefore an ideal group to investigate the evolutionary mechanisms underlying genome size variation.

Limitations

Because of the preservation in dry conditions of the specimen, we could not sequence a greater amount of long reads data and transcriptomic data that would have helped to assemble a more contiguous genome with a proper RNAseq-based annotation. As a result, our genome has a fairly low N50 and is more fragmented, compared to previous genome assemblies in Papilionidae based on similar types of data (e.g. He et al. 2022; Reboud et al. 2023). However, this genome still represents an interesting resource for this rare and endangered species that fills a knowledge gap in this insect family. Although we recovered a similar demographic trend for *Baronia* between Illumina and ONT data, our MSMC inferences show some level of discrepancy, which should then be interpreted with caution and better understood. By replicating the MSMC2 analyses with the repeat masked, we found that the Illumina and ONT datasets produce similar demographic trends (Figure S5), suggesting that the discrepancy between the Illumina and ONT datasets is due to the repeat sequences. The ONT are likely to be mapped more accurately on the repeat portion of the genome and excluding these regions leads to a homogenization of all results. Despite lineage-specific features unveiled, we also remain careful on our results and interpretations of genome size evolution because a more accurate genus-level sampling within Parnassiinae and Papilioninae can bring new information.

Conclusion

In this study, using a combination of long and short reads, we presented the genome of *Baronia brevicornis*, an endangered butterfly species. The genome was found to be of comparable quality to other published genomes of swallowtail species. The heterozygosity level was found to be unexpectedly elevated, given its endangered status. However, the genetic diversity of this butterfly was found to be relatively low compared to other butterflies, but much higher than other endangered swallowtail butterflies. The historical demography is characterized by a strong decline of the effective population size initiated in the last million years that stabilized to a low effective population size in the last 100,000 years. As the sister species of swallowtail butterflies and the oldest lineage of all butterflies, the *Baronia* genome was pivotal to study genome size variation in Papilionidae. The activity of transposable elements is the primary driver of genome size evolution in swallowtail butterflies, indicating that the emergence of large genomes is a recent characteristic due to various transposable element classes. Overall, this study provides important insights into the genome and genetic diversity of this endangered butterfly species.

Data Availability Statement

The genome, mitogenome and sequencing data of the present study, including Nanopore, Illumina, and RNA data are available from the Genome database and Sequence Read Archive under the Bioproject accession number PRJNA971161, with the corresponding BioSamples accession number SAMN35020972, genome accession number JASFAT000000000. The genome annotations made with the BRAKER2 and MAKER pipelines are both available in the FigShare online repository: https://figshare.com/s/f26e4b4a9d0e47e2832d.

Acknowledgments

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (project GAIA, agreement no. 851188). This project benefited from the Montpellier Bioinformatics Biodiversity platform supported by an "Investissements d'avenir" program managed by the ANR (CeMEB, ref. ANR-10-LABX-04-01), from an "Investissements d'Avenir" program

managed by the ANR (CEBA, ref. ANR-10-LABX-25-01), and from an ANR project (NeGA, ref. ANR-20-CE02-0008).

Author Contributions

A.M., E.L.R., J.C.G., B.N., and F.L.C. conceived the study. E.C. and M.-K.T. performed DNA extraction, library preparations, ONT sequencing and Illumina sequencing. E.L.R. performed the genome assembly and genome-wide statistics. A.M. performed the analysis of transposable elements. B.N. conducted homology-based annotation, molecular diversity analyses, and historical demography. F.L.C., B.N. and A.M. wrote the manuscript. All authors edited and approved the manuscript.

References

Allio R, Scornavacca C, Nabholz B, Clamens A-L, Sperling FAH, Condamine FL 2020a. Whole genome shotgun phylogenomics resolves the pattern and timing of swallowtail butterfly evolution. Systematic Biology 69: 38-60.

Allio R, Schomaker-Bastos A, Romiguier J, Prosdocimi F, Nabholz B, Delsuc F. 2020b. MitoFinder: efficient automated large-scale extraction of mitogenomic data in target enrichment phylogenomics. Molecular Ecology Resources 20: 892–905.

Allio R, Nabholz B, Wanke S, Chomicki G, Pérez-Escobar OA, Cotton, AM, et al. 2021. Genome-wide macroevolutionary signatures of key innovations in butterflies colonizing new host plants. Nature Communications 12: 354.

Amemiya CT, Alföldi J, Lee AP, Fan S, Philippe H, MacCallum I, et al. 2013. The African coelacanth genome provides insights into tetrapod evolution. Nature 496: 311-316.

Armstrong J, Hickey G, Diekhans M, Fiddes IT, Novak AM, Deran A, Fang Q, Xie D, Feng S, Stiller J, et al. 2020. Progressive Cactus is a multiple-genome aligner for the thousand-genome era. Nature 587: 246-251.

Baril T, Hayward A. 2022. Migrators within migrators: Exploring transposable element dynamics in the monarch butterfly, *Danaus plexippus*. Mobile DNA 13: 5.

Baril T, Imrie RM, Hayward A. 2022. Earl Grey: a fully automated user-friendly transposable element annotation and analysis pipeline. https://doi.org/10.1101/2022.06.30.498289 [accessed 2023 April 30]

Blumenstiel, J. P. (2011). Evolutionary dynamics of transposable elements in a small RNA world. Trends in Genetics 27: 23–31.

Brace S, Thomas JA, Dalén L, Burger J, MacPhee, RD, Barnes I, Turvey ST. 2016. Evolutionary history of the Nesophontidae, the last unplaced recent mammal family. Molecular Biology and Evolution 33: 3095-3103.

Brůna T, Hoff KJ, Lomsadze A, Stanke M, Borodovsky M. 2021. BRAKER2: Automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein database. NAR Genomics and Bioinformatics 3: lqaa108.

Buffalo V. 2021 Quantifying the relationship between genetic diversity and population size suggests natural selection cannot explain Lewontin's paradox. eLife 10: e67509.

Cavin L, Guinot G. 2014. Coelacanths as "almost living fossils". Frontiers in Ecology and Evolution 2: 49.

Chazot N, Wahlberg N, Freitas AVL, Mitter C, Labandeira C, Sohn JC, et al. 2019. Priors and posteriors in Bayesian timing of divergence analyses: the age of butterflies revisited. Systematic Biology 68: 797-813.

Chen S, Zhou Y, Chen Y, Gu J. 2018. fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics 34: i884-i890.

Chiang C, Layer RM, Faust GG, Lindberg MR, Rose DB, Garrison EP, Marth GT, Quinlan AR, Hall IM. 2015. SpeedSeq: ultra-fast personal genome analysis and interpretation. Nature Methods 12: 966–968.

Cicconardi F, Milanetti E, Pinheiro de Castro EC, et al. 2023. Evolutionary dynamics of genome size and content during the adaptive radiation of Heliconiini butterflies. Nature Communications 14: 5620.

Condamine FL, Sperling FAH, Wahlberg N, Rasplus J-Y, Kergoat GJ. 2012. What causes latitudinal gradients in species diversity? Evolutionary processes and ecological constraints on swallowtail biodiversity. Ecology Letters 15: 267-277.

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

Darwin C. 1859. *On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life*. John Murray, London.

Díez-del-Molino D, Sánchez-Barreiro F, Barnes I, Gilbert MTP, Dalén L. 2018. Quantifying temporal genomic erosion in endangered species. Trends in Ecology and Evolution 33: 176-185.

Dincă V, Dapporto L, Somervuo P, Vodă R, Cuvelier S, Gascoigne-Pees M, Huemer P, Mutanen M, Hebert PDN, Vila, R. 2021. High resolution DNA barcode library for European butterflies reveals continental patterns of mitochondrial genetic diversity. Communications Biology 4: 315.

Espeland M, Breinholt J, Willmott KR, Warren AD, Vila R, Toussaint EFA, et al. 2018. A comprehensive and dated phylogenomic analysis of butterflies. Current Biology 28: 770-778.

Galicia-Mendoza I, Pineda-García F, Oyama K, Cordero-Rivera A, Osorio-Beristain M, Contreras-Garduño J. 2021. Biology of the oldest butterfly species in the world, *Baronia brevicornis*: Food, abundance, polymorphism, and survival. Revista Mexicana de Biodiversidad 92: 923503.

García-Berro A, Talla V, Vila R, Wai HK, Shipilina D, Chan KG, Pierce NE, Backström N, Talavera G. 2023. Migratory behavior is positively associated with genetic diversity in butterflies. Molecular Ecology 32: 560-574.

Goubert C, Craig RJ, Bilat AF, Peona V, Vogan AA, Protasio AV. 2022. A beginner's guide to manual curation of transposable elements. Mobile DNA 13: 7.

Edge P, Bansal V. 2019. Longshot enables accurate variant calling in diploid genomes from single-molecule long read sequencing. Nature Communications 10: 4660.

Ellegren H, Galtier N. 2016. Determinants of genetic diversity. Nature Review Genetics 17: 422-433.

Fang LJ, Zhang YL, Gao K, Ding CP, Zhang YJ. 2019. Butterfly communities along the Heihe River Basin in Shaanxi Province, a biodiversity conservation priority area in China. Journal of Insect Conservation 23: 873–883.

Flynn JM, Hubley R, Goubert C, Rosen J, Clark AG, Feschotte C, Smit AF. 2020. RepeatModeler2 for automated genomic discovery of transposable element families. Proceedings of the National Academy of Sciences of the USA 117: 9451-9457.

Gilbert C, Peccoud J, Cordaux R. 2021. Transposable elements and the evolution of insects. Annual Review of Entomology 66: 355-372.

Gouy M, Guindon S, Gascuel O. 2010. SeaView version 4: a multiplatform graphical user interface for sequence alignment and phylogenetic tree building. Molecular Biology and Evolution 27: 221-224.

Grandcolas P, Nattier R, Trewick S. 2014. Relict species: a relict concept? Trends in Ecology and Evolution 29: 655-663.

Guan D-L, Zhao L, Li Y, Xing L-X, Huang H, Xu S-Q. 2022. Genome assembly of *Luehdorfia taibai*, an endangered butterfly endemic to Qinling Mountains in China with extremely small populations. Frontiers in Ecology and Evolution 10: 955246.

Heikkilä M, Kaila L, Mutanen M, Pena C, Wahlberg N. 2012. Cretaceous origin and repeated tertiary diversification of the redefined butterflies. Proceedings of the Royal Society B: Biological Sciences 279: 1093-1099.

Holt C, Yandell M. 2011. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. BMC Bioinformatics 12: 1-14.

Jin JJ, Yu WB, Yang JB, Song Y, DePamphilis CW, Yi TS, Li DZ. 2020. GetOrganelle: a fast and versatile toolkit for accurate de novo assembly of organelle genomes. Genome Biology 21: 241.

Kapusta A, Suh A, Feschotte C. 2017. Dynamics of genome size evolution in birds and mammals. Proceedings of the National Academy of Sciences of the USA 114: E1460-E1469.

Kawahara AY, Plotkin D, Espeland M, Meusemann K, Toussaint EFA, Donath A, et al. 2019. Phylogenomics reveals the evolutionary timing and pattern of butterflies and moths. Proceedings of the National Academy of Sciences of the USA 116: 22657-22663.

Kawahara AY, Storer C, Carvalho APS, Plotkin DM, Condamine FL, Braga MP, Ellis EA, St Laurent RA, Li X, Barve V, et al. 2023 A global phylogeny of butterflies reveals their evolutionary history, ancestral hosts and biogeographic origins. Nature Ecology and Evolution 7: 903-913.

Kebaïli C, Sherpa S, Rioux D, Després L. 2022. Demographic inferences and climatic niche modelling shed light on the evolutionary history of the emblematic cold-adapted Apollo butterfly at regional scale. Molecular Ecology 31: 448-466.

Keightley PD, Pinharanda A, Ness RW, Simpson F, Dasmahapatra KK, Mallet J, Davey JW, Jiggins CD. 2015. Estimation of the spontaneous mutation rate in *Heliconius melpomene*. Molecular Biology Evolution 32: 239-243.

Kolmogorov M, Yuan J, Lin Y, Pevzner PA. 2019. Assembly of long, error-prone reads using repeat graphs. Nature Biotechnology 37: 540-546.

Korf I. 2004. Gene finding in novel genomes. BMC Bioinformatics 5: 1-9.

Kunte K, Zhang W, Tenger-Trolander A, Palmer DH, Martin A, Reed RD, et al. 2014. *Doublesex* is a mimicry supergene. Nature 507: 229-232.

Kuznetsov D, Tegenfeldt F, Manni M, Seppey M, Berkeley M, Kriventseva EV, Zdobnov EM. 2023. OrthoDB v11: annotation of orthologs in the widest sampling of organismal diversity. Nucleic Acids Research 51: D445-D451.

Lavoie CA, Platt RN, Novick PA, Counterman BA, Ray DA. 2013. Transposable element evolution in *Heliconius* suggests genome diversity within Lepidoptera. Mobile DNA 4: 21.

Laetsch DR, Blaxter ML. 2017. BlobTools: Interrogation of genome assemblies. F1000Research 6: 1287.

Legal L, Dorado O, Machkour-M'Rabet S, Leberger R, Albre J, Mariano NA, Gers C. 2015. Ecological constraints and distribution of the primitive and enigmatic endemic Mexican butterfly *Baronia brevicornis* (Lepidoptera: Papilionidae). Canadian Entomologist 147: 71-88.

León-Cortés JL, Pérez-Espinoza F, Marín L, Molina-Martínez A. 2004. Complex habitat requirements and conservation needs of the only extant Baroniinae swallowtail butterfly. Animal Conservation 7: 241–250.

Li H. 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. ArXiv Prepr. ArXiv13033997.

Li H. 2018. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics 34: 3094–3100.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25: 2078–2079.

Li X, Fan D, Zhang W, Liu G, Zhang L, Zhao L, Fang X, Chen L, Dong Y, Chen Y, et al. 2015. Outbred genome sequencing and CRISPR/Cas9 gene editing in butterflies. Nature Communications 6: 8212.

Liu G, Chang Z, Chen L, He J, Dong Z, Yang J, Lu S, Zhao R, Wan W, Ma G, Li J, Zhang R, Wang W, Li X. 2020. Genome size variation in butterflies (Insecta, Lepidoptera, Papilionoidea): a thorough phylogenetic comparison. Systematic Entomology 45: 571-582.

Lu S, Yang J, Dai X, Xie F, He J, Dong Z, Mao J, Liu G, Chang Z, Zhao R, et al. 2019. Chromosomal-level reference genome of Chinese peacock butterfly (*Papilio bianor*) based on third-generation DNA sequencing and Hi-C analysis. GigaScience 8: giz128.

Lynch M, Conery JS. 2003. The origins of genome complexity. Science 302: 1401–1404.

Mackintosh A, Laetsch DR, Hayward A, Charlesworth B, Waterfall M, Vila R, Lohse K. 2019. The determinants of genetic diversity in butterflies. Nature Communications 10: 3466.

Mackintosh A, Laetsch DR, Baril T, Ebdon S, Jay P, Vila R, Hayward A, Lohse K. 2022. The genome sequence of the scarce swallowtail, *Iphiclides podalirius*. Genes Genomes Genetics 12: jkac193.

Machkour-M'Rabet S, Leberger R, León-Cortés JL, Gers C, Legal L. 2014. Population structure and genetic diversity of the only extant Baroninae swallowtail butterfly, *Baronia brevicornis*, revealed by ISSR markers. Journal of Insect Conservation 18: 385-396.

Manni M, Berkeley MR, Seppey M, Simão FA, Zdobnov EM. 2021. BUSCO update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. Molecular Biology and Evolution 38: 4647-4654.

Manthey JD, Girón JC, Hruska JP. 2022. Impact of host demography and evolutionary history on endosymbiont molecular evolution: A test in carpenter ants (genus *Camponotus*) and their *Blochmannia* endosymbionts. Ecology and Evolution 12: e9026.

Morin PA, Archer FI, Avila CD, Balacco JR, Bukhman YV, Chow W, Fedrigo O, Formenti G, Fronczek JA, Fungtammasan A, et al. 2021. Reference genome and demographic history of the most endangered marine mammal, the vaquita. Molecular Ecology Resources 21: 1008-1020.

Nikaido M, Noguchi H, Nishihara H, Toyoda A, Suzuki Y, Kajitani R, et al. 2013. Coelacanth genomes reveal signatures for evolutionary transition from water to land. Genome Research 23: 1740-1748.

Nishikawa H, Iijima T, Kajitani R, Yamaguchi J, Ando T, Suzuki Y, Sugano S, Fujiyama A, Kosugi S, Hirakawa H, et al. 2015. A genetic mechanism for female-limited Batesian mimicry in *Papilio* butterfly. Nature Genetics 47: 405-409.

Nishimura O, Hara Y, Kuraku S. 2017. gVolante for standardizing completeness assessment of genome and transcriptome assemblies. Bioinformatics 33: 3635-3637.

Paradis E, Schliep K. 2019. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. Bioinformatics 35: 526-528.

Petersen M, Armisén D, Gibbs RA, Hering L, Khila A, Mayer G, Richards S, Niehuis O, Misof B. 2019. Diversity and evolution of the transposable element repertoire in arthropods with particular reference to insects. BMC Ecology and Evolution 19: 11.

Podsiadlowski L, Tunström K, Espeland M, Wheat CW. 2021. The genome assembly and annotation of the Apollo butterfly *Parnassius apollo*, a flagship species for conservation biology. Genome Biology and Evolution 13: evab122.

Puttick A, Leon-Cortes J, Legal L. 2018. *Baronia brevicornis*. The IUCN Red List of Threatened Species 2018: e.T2594A119581233.

Reboud EL, Nabholz B, Chevalier E, Tilak M-k, Bito D, Condamine FL. 2023 Genomics, population divergence and historical demography of the world's largest and endangered butterfly, the Queen Alexandra's birdwing. Genome Biology and Evolution 15: evad040.

Reiss D, Mialdea G, Miele V, de Vienne DM, Peccoud J, Gilbert C, Duret L, Charlat S. 2019. Global survey of mobile DNA horizontal transfer in arthropods reveals Lepidoptera as a prime hotspot. PLoS Genetics 15: e1007965.

Rhie A, Walenz BP, Koren S, Phillippy AM. 2020. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. Genome Biology 21: 245.

Robinson JA, Kyriazis CC, Nigenda-Morales SF, Beichman AC, Rojas-Bracho L, Robertson KM, Fontaine MC, Wayne RK, Lohmueller KE, Taylor BL, et al. 2022. The critically

endangered vaquita is not doomed to extinction by inbreeding depression. Science 376: 635-639.

Romiguier J, Gayral P, Ballenghien M, Bernard A, Cahais V, et al. 2014 Comparative population genomics in animals uncovers the determinants of genetic diversity. Nature 515: 261–263.

Royer DL, Hickey LJ, Wing SL. 2003. Ecological conservatism in the "living fossil" Ginkgo. Paleobiology 29: 84-104.

Schaack S, Gilbert C, Feschotte C. 2010. Promiscuous DNA: Horizontal transfer of transposable elements and why it matters for eukaryotic evolution. Trends in Ecology and Evolution 25: 537–546.

Schiffels S, Wang K. 2020. MSMC and MSMC2: The Multiple Sequentially Markovian Coalescent. In: Dutheil JY, editor. Statistical Population Genomics. New York (NY): Springer New York. p. 147–166.

Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. 2015. BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics 31: 3210-3212.

Simonsen TJ, Zakharov EV, Djernaes M, Cotton AM, Vane-Wright RI, Sperling FAH 2011. Phylogenetics and divergence times of Papilioninae (Lepidoptera) with special reference to the enigmatic genera *Teinopalpus* and *Meandrusa*. Cladistics 27: 113-137.

Smit AF, Hubley R, Green P. 2015. RepeatMasker Open-4.0. 2013–2015.

Stanke M, Tzvetkova A, Morgenstern B. 2006. AUGUSTUS at EGASP: using EST, protein and genomic alignments for improved gene prediction in the human genome. Genome Biology 7: 1-8.

Storer J, Hubley R, Rosen J, Wheeler TJ, Smit AF. 2021. The Dfam community resource of transposable element families, sequence models, and genome annotations. Mobile DNA 12: 1-14.

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 Biology and Evolution 9: 2491–2505.

Timmermans MJ, Thompson MJ, Collins S, Vogler AP. 2017. Independent evolution of sexual dimorphism and female-limited mimicry in swallowtail butterflies (*Papilio dardanus* and *Papilio phorcas*). Molecular Ecology 26: 1273-1284.

Vazquez-Selem L, Heine K. 2004. Late quaternary glaciation of Mexico. In Developments in Quaternary Sciences (Vol. 2, pp. 233-242). Elsevier.

Venner S, Miele V, Terzian C, Biémont C, Daubin V, Feschotte C, Pontier D. 2017. Ecological networks to unravel the routes to horizontal transposon transfers. PLoS Biology 15: e2001536.

Walton W, Stone GN, Lohse K. 2021. Discordant Pleistocene population size histories in a guild of hymenopteran parasitoids. Molecular Ecology 30: 4538-4550.

Wilder AP, Supple MA, Subramanian A, Mudide A, Swofford R, Serres-Armeno A, Steiner C, Koepfli K-P, Genereux DP, Karlsson EK, et al. 2023. The contribution of historical processes to contemporary extinction risk in placental mammals. Science 380: eabn5856.

Wu C, Lu J. 2019. Diversification of transposable elements in arthropods and its impact on genome evolution. Genes 10: 338.

Zimin AV, Marçais G, Puiu D, Roberts M, Salzberg SL, Yorke JA. 2013. The MaSuRCA genome assembler. Bioinformatics 29: 2669-2677.
Zimin AV, Salzberg SL. 2020. The genome polishing tool POLCA makes fast and accurate corrections in genome assemblies. PLoS Computational Biology 16: e1007981.

Supplemental Material

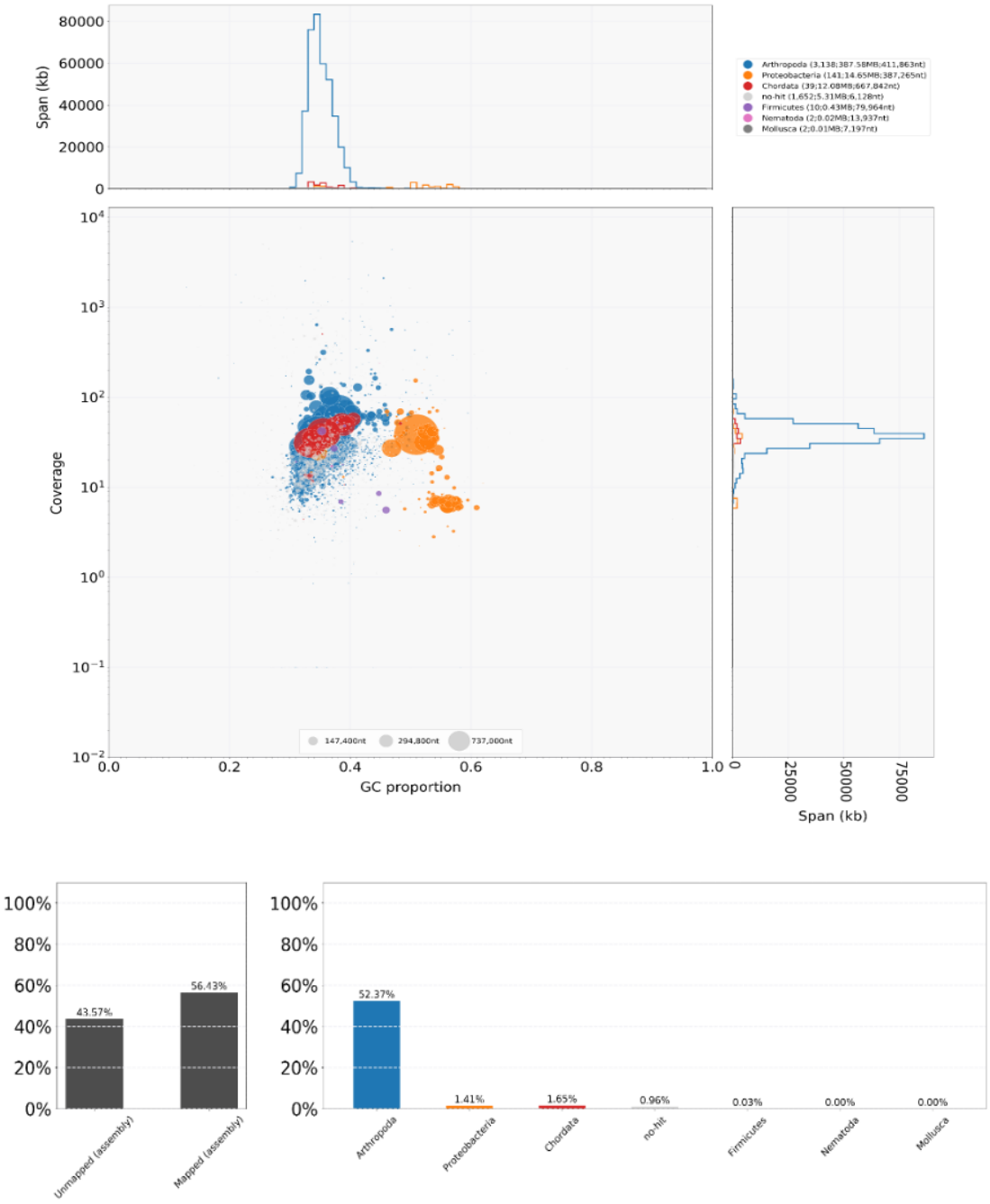


Figure S1. BlobTools plots for the genome assembly of *Baronia brevicornis*.

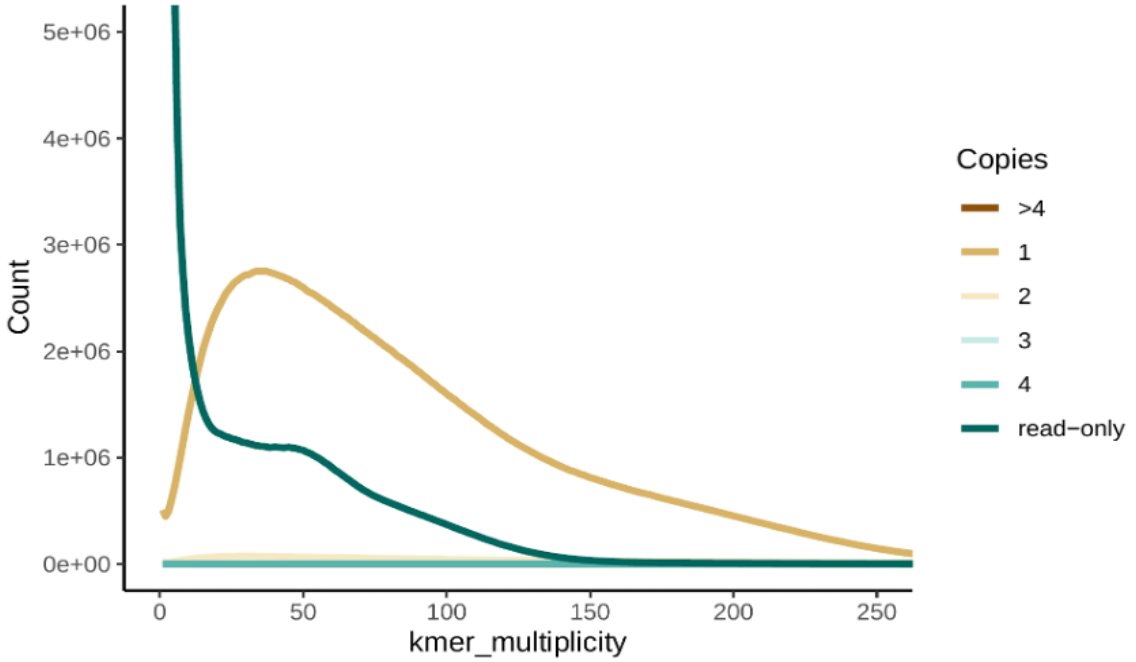


Figure S2. Spectra asm plot in Merqury.

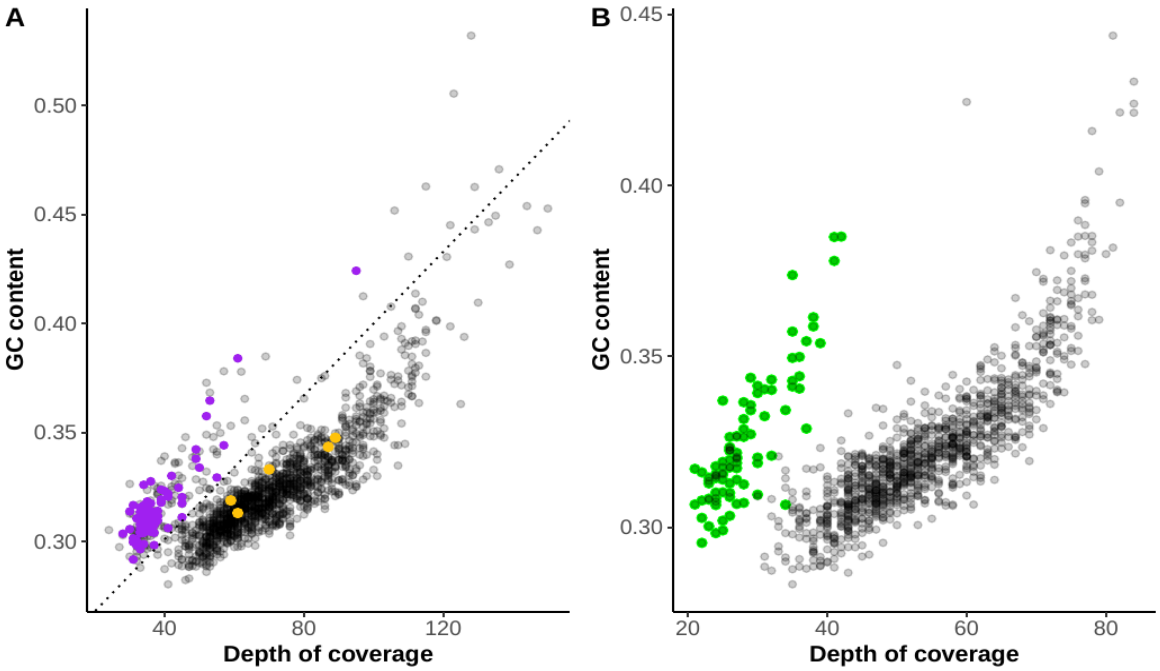


Figure S3. Plots of depth of coverage as a function of GC content per contigs. Only contigs longer than 30 kb are represented. A) Illumina data. The dotted line separates the two categories of contigs. Contigs above the line and with a coverage < 82x are probably from the Z chromosome. The purple points are the contigs that have been aligned to the Z chromosome (contig 30) of *Papilio bianor* and *P. machaon* using Cactus. The yellow points are the contigs that have been aligned to the W chromosome of *P. machaon*. B) Oxford Nanopore Technology data. The green points are the contigs that have been identified as the Z chromosome using the Illumina data.

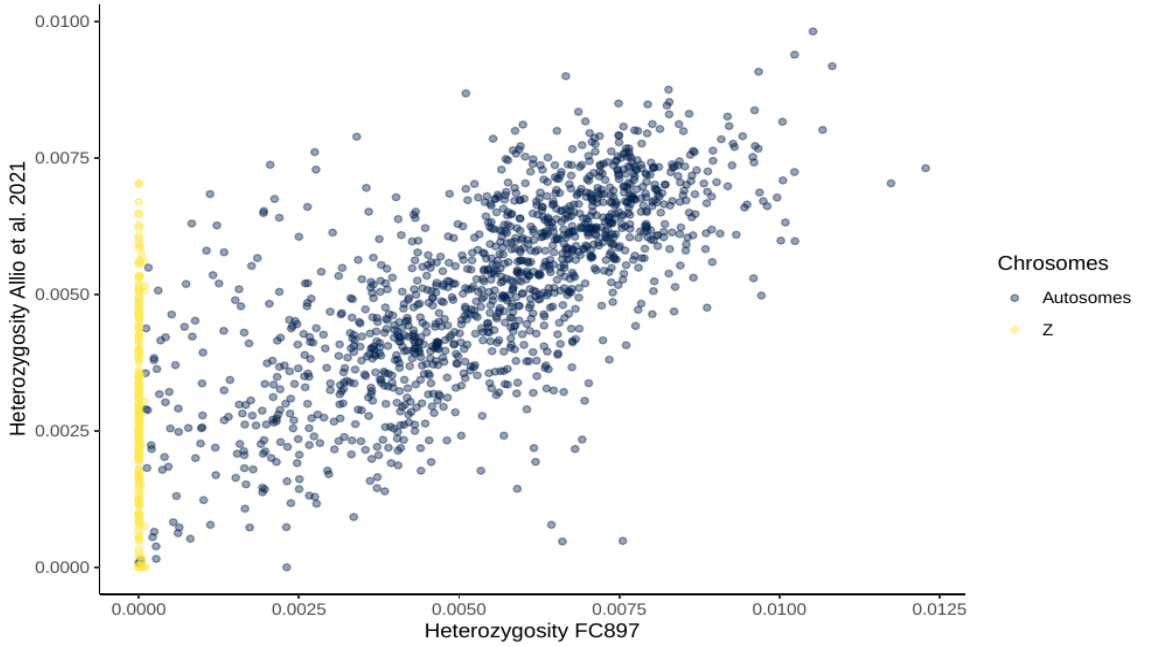


Figure S4. Heterozygosity of the contigs larger than 30 kb computed for the current genome assembly of a female and compared to the male sequenced in Allio *et al.* (2020a). Yellow points indicate the contigs belonging to the Z chromosome, while gray points indicate the contigs belonging to the autosomes.

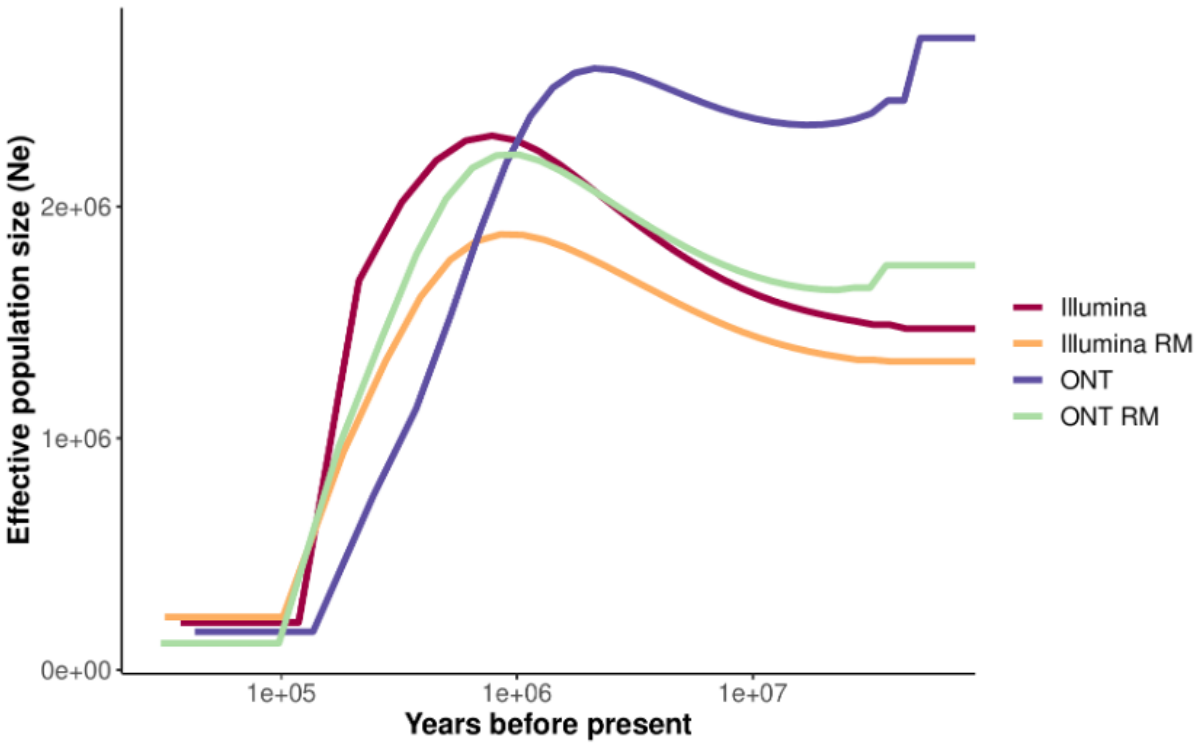


Figure S5. MSMC2 estimates of the effective population size (Ne) with both Illumina and Nanopore data original data or with repeat masked (labeled RM).

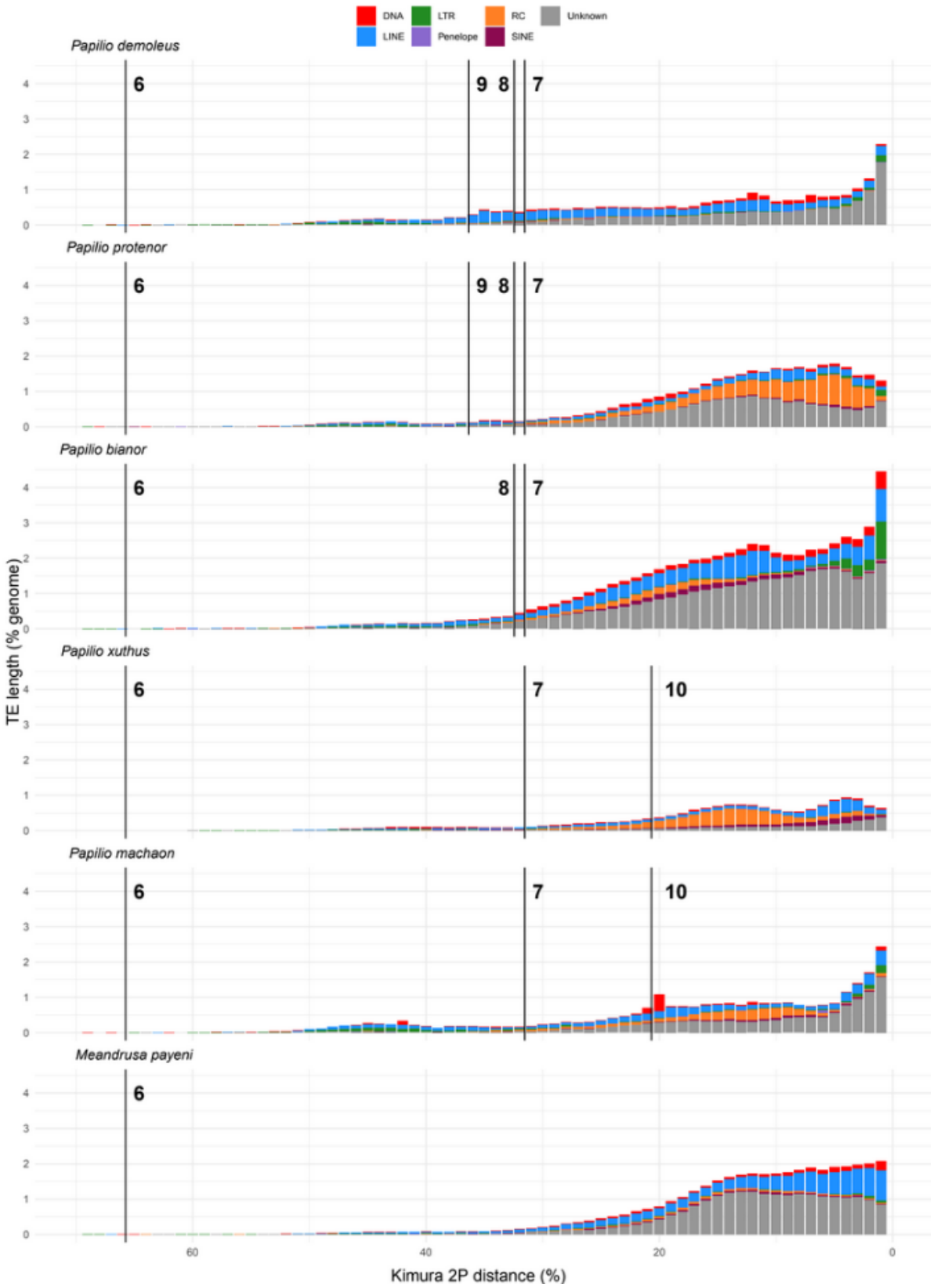


Figure S6. TE copies distribution according to their divergence from consensus across all Papilionini. Past lineage splits are reported for each species and refer to node numbers in Figure 3.

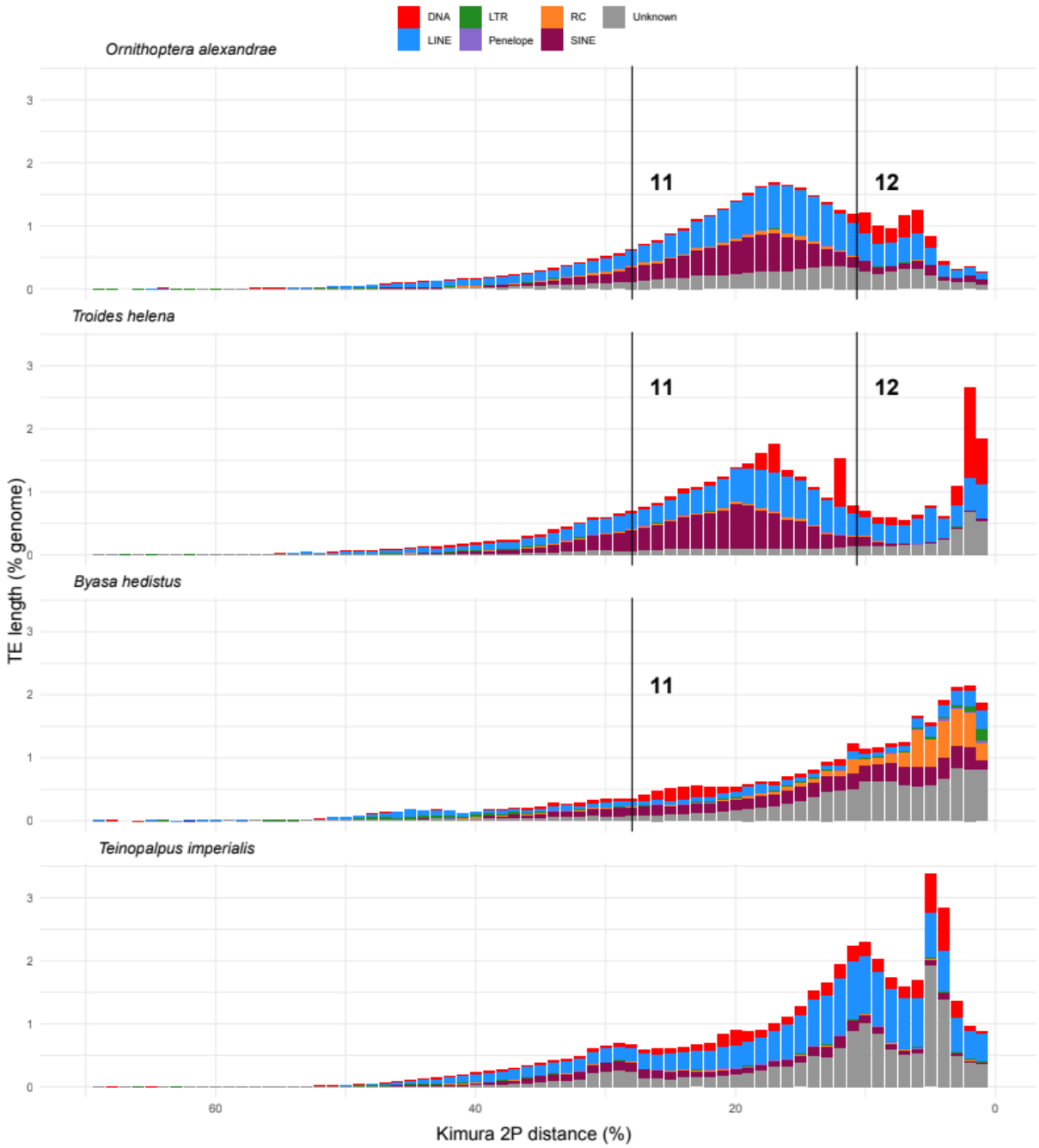


Figure S7. TE copies distribution according to their divergence from consensus across all Troidini and Teinopalpini. Past lineage splits are reported for each species and refer to node numbers in Figure 3.

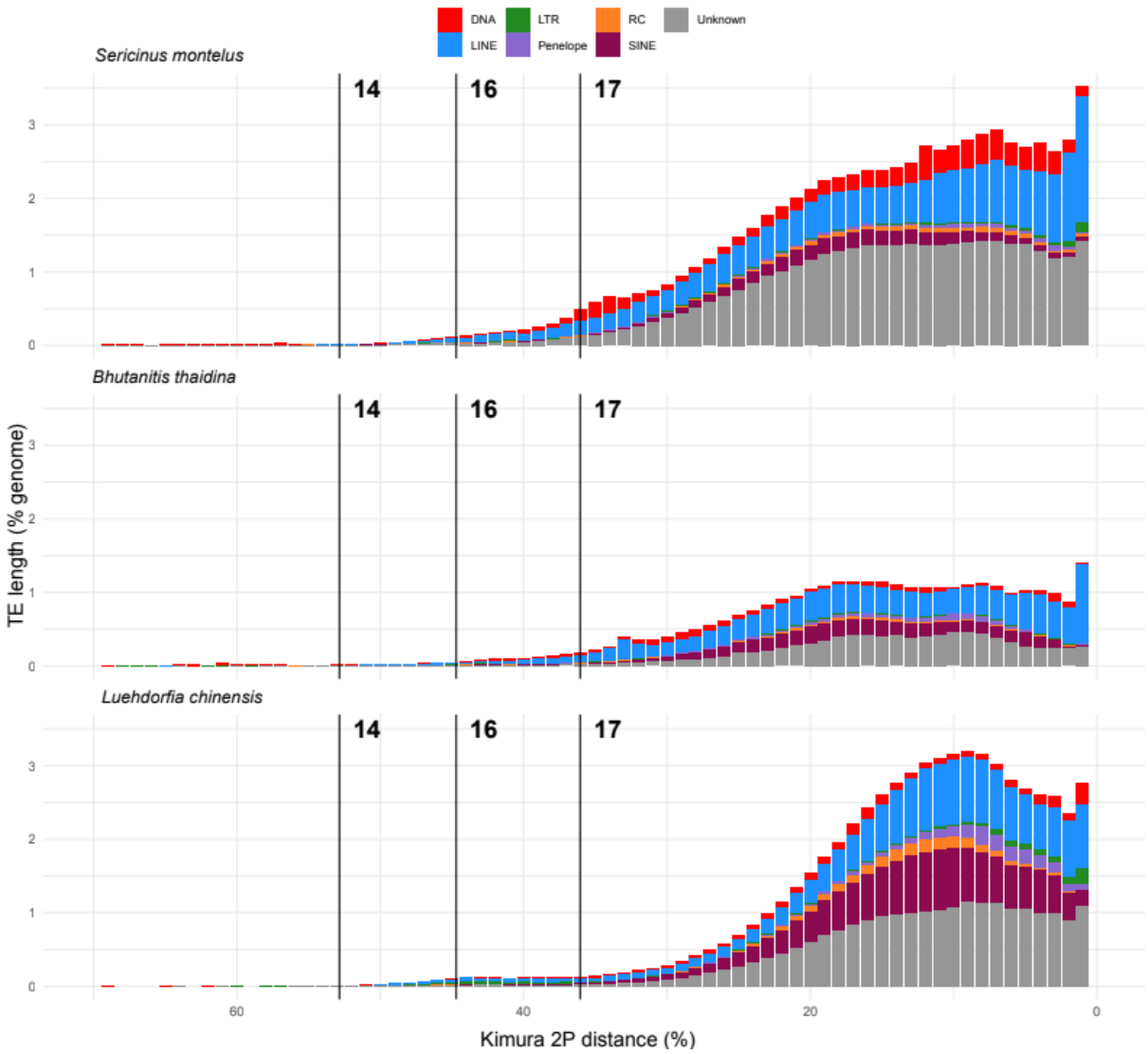


Figure S8. TE copies distribution according to their divergence from consensus across Zerynthiini and Luehdorfiini. Past lineage splits are reported for each species and refer to node numbers in Figure 3.

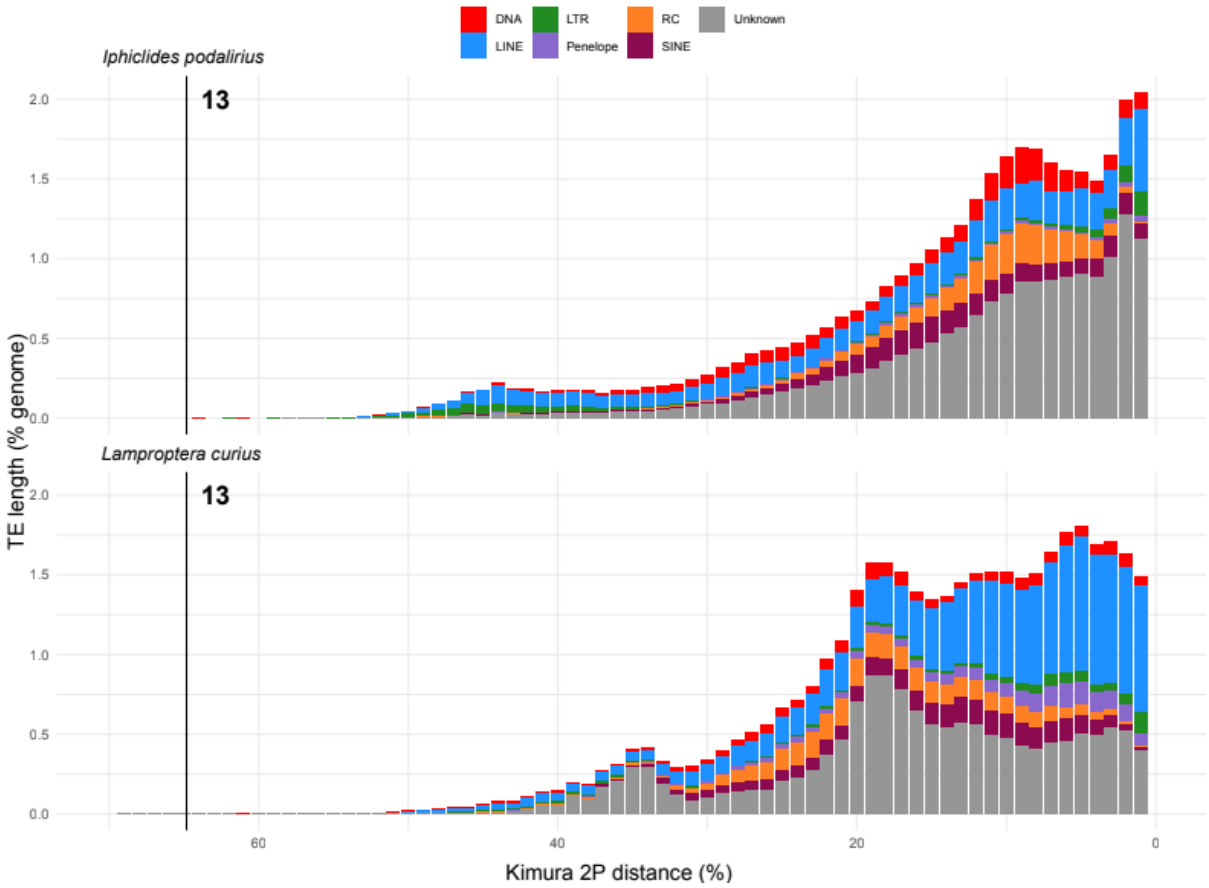


Figure S9. TE copies distribution according to their divergence from consensus across Leptocircini. Past lineage splits are reported for each species and refer to node numbers in Figure 3.

Phylum	Genus	FC897
Bacillota	*Vagococcus*	2
Pseudomonadota	*Citrobacter*	48
	Enterobacter	1
	Enterococcus	8
	Escherichia	9
	Klebsiella	75
	Proteus	1
	Raoultella	1
	Salmonella	7
	Total Size (bp)	**15,113,703**

Table S1. Number of contigs removed and their phylum rank (Bacillota, Pseudomonadota) after BlobTools analysis. The total number of contigs removed is 152.

Table S2. Taxonomy, references, assembly sizes and TE annotation results for *Baronia* and the 17 species used in the analysis of GS variation.

Not provided in this document because of the large size.

Appendix V

M1 Report, Université de Montpellier, MEME

Host-plant shifts in swallowtail butterflies did not happen in isolation

Eleanor Sheridan

Supervised by Dr. Fabien Condamine and Eliette Reboud

Institut des Sciences de l'Evolution-Montpellier (ISEM), Montpellier, France

Abstract

Insects are the most species-rich eukaryotic group on the planet. This huge diversity is often credited to their coevolutionary interactions with plants, with these high diversification rates then further increased by repeated shifts between host-plants. However, possible associations between shifts in host-plant and shifts in other biogeographic traits that might also fuel increased diversification rates remain untested. Moreover, the precise effects of these host-plant shifts on phenotypic diversification are unknown. To investigate whether changes in host-plant preference and changes in other traits occurred in parallel in swallowtail butterflies, we estimated the ancestral state of host-plant, region, biome, and wingspan, and calculated the disparity in colour between different species. We found that shifts in regions but not biomes occurred concurrently with colonisation of new host-plants. This may have been due to expansion into new regions exposing species to certain plant families for the first time. Moreover, we found that species differed more in colour if they were separated by at least one host-plant shift and that the rate of wingspan evolution slowed following host-plant shifts. This is consistent with host-plant shifts influencing morphological evolution, appearing to promote faster evolution of colour but slower evolution of wingspan. These results highlight the complexity of host-plant shifts and the need to consider multiple traits when seeking to understand a system.

Word count: 3988

Introduction

Of all the eukaryotic groups on the planet, insects have the most described species (Stork 2018) but such an evolutionary success remains puzzling. It is now almost 60 years since a seminal paper by Ehrlich and Raven (1964) proposed an answer to this very question, suggesting that coevolutionary interactions between insects and their host-plants might be responsible for driving diversification. In particular, they theorised that arms races would result from attacks by herbivorous insects and the evolution of novel defences by food plants. These would repeatedly lead to the creation of new 'adaptive zones' in which a lineage could diversify. Since then, many studies have provided support for this model, demonstrating that shifts in host-plant, and thus expansion into new adaptive zones, do indeed often coincide with increases in diversification rates (Mitter et al. 1988; Farrell 1998; Winkler et al. 2009; Fordyce 2010; Edger et al. 2015; Sahoo et al. 2017; Condamine et al. 2018; Yamamoto et al. 2020; Allio et al. 2021). In addition, widespread signals of selection resulting from host-plant shifts are visible in genomes (Edger et al. 2015; Allio et al. 2021), indicating that selection and adaptation can be induced from such shifts. However, consequences of host-plant shifts may be more complex than simply an increase in diversification rates resulting from the opportunity for exploitation of new adaptive zones.

Swallowtail butterflies (Lepidoptera: Papilionidae) are a charismatic and diverse insect family (Ehrlich and Raven 1964; Collins and Morris 1985) well-suited to asking questions about the complex nature of host-plant shifts. Across the swallowtail family, species primarily feed on thirteen different plant families (Hancock 1983; Figure 1A), with a strong signal of phylogenetic conservatism in this preference (Condamine et al. 2012; Allio et al. 2021). Rare shifts between these host-plants have accelerated diversification rates (Condamine et al. 2012; Allio et al. 2021). Swallowtails are found in every global region and biome (Figures 1C and D), meaning their range encapsulates the entirety of global ecological diversity. Consequently, there is a wealth of other data to combine with data on host-plant shifts.

A change of host-plant constitutes a major ecological shift because it opens up novel niches for exploitation; but it is far from being the only ecological shift a species can undergo. Diversification rates also differ by biogeographic location, varying with latitude (Condamine et al. 2012; Pyron and Wiens 2013; Rolland et al. 2014; Igea and Tanentzap 2020) and continent (Igea and Tanentzap 2019). Moreover, even just within a species, location affects many phenotypic traits, from colouration (Bishop et al. 2016; Xing et al. 2016; Pinkert et al. 2017; Stelbrink et al. 2019), to predator responses (Laurila et al. 2008), to dispersal ability (Sheard et al. 2020). Changes in location can therefore induce novel selection pressures on species and accelerate diversification, just as changes in host-plant preference can. Perhaps these could shift in tandem. This possibility, that host-plant preference could go hand in hand with biome or regional presence, is exemplified by the *Parnassius* genus: the only swallowtail genus to feed on Papaveraceae (Hancock 1983) or to live at high altitudes (Weiss 1991). Understanding

the potentially confounding effects of biogeographic context with host-plant preference is important for accurately assessing the true contribution of host-plant shifts to species diversification.

In addition to understanding how external pressures on species change, it is also important to understand the evolutionary consequences of these pressures. Host-plant shifts left genome-wide signals of selection in swallowtails (Allio et al. 2021), but what about changes to morphological traits? Swallowtails are markedly diverse in colouration (Ehrlich and Raven 1964; Stelbrink et al. 2019) and size (Collins and Morris 1985; Figure 1B). For instance, *Parnassius* species can have wingspans as little as 3 cm (Weiss 1991) while *Ornithoptera alexandrae* can reach widths of up to 25 cm (Collins and Morris 1985). Perhaps some of this morphological diversity could be attributed to adaptation resulting from host-plant shifts throughout the family's evolutionary history – as observed in the nymphalid genus *Morpho* (Chazot et al. 2021). By analysing the effect of host-plant shifts on wingspan evolution and colour disparity in the older and more species-rich swallowtail clade, one can ask whether such an evolutionary effect of host-plant shifts on morphology exists at this larger phylogenetic level.

In this study, we assembled a comprehensive database of host-plant preference, wingspan, colouration, and distribution across regions and biomes to define the traits of modern day Papilionidae (Figure 1). Using a time-calibrated phylogeny of 421 Papilionidae species, we were able to estimate the ancestral states of traits and determine the timing of changes in these traits. This allowed us to investigate potential cooccurrences between shifts in host-plant preference and shifts in other traits to explore two complimentary questions. First, we tested the hypothesis that shifts in biome and region happened in parallel with shifts in host-plant. Then, we examined whether colonisation of new host-plants increased the rate of change of colouration or wingspan.

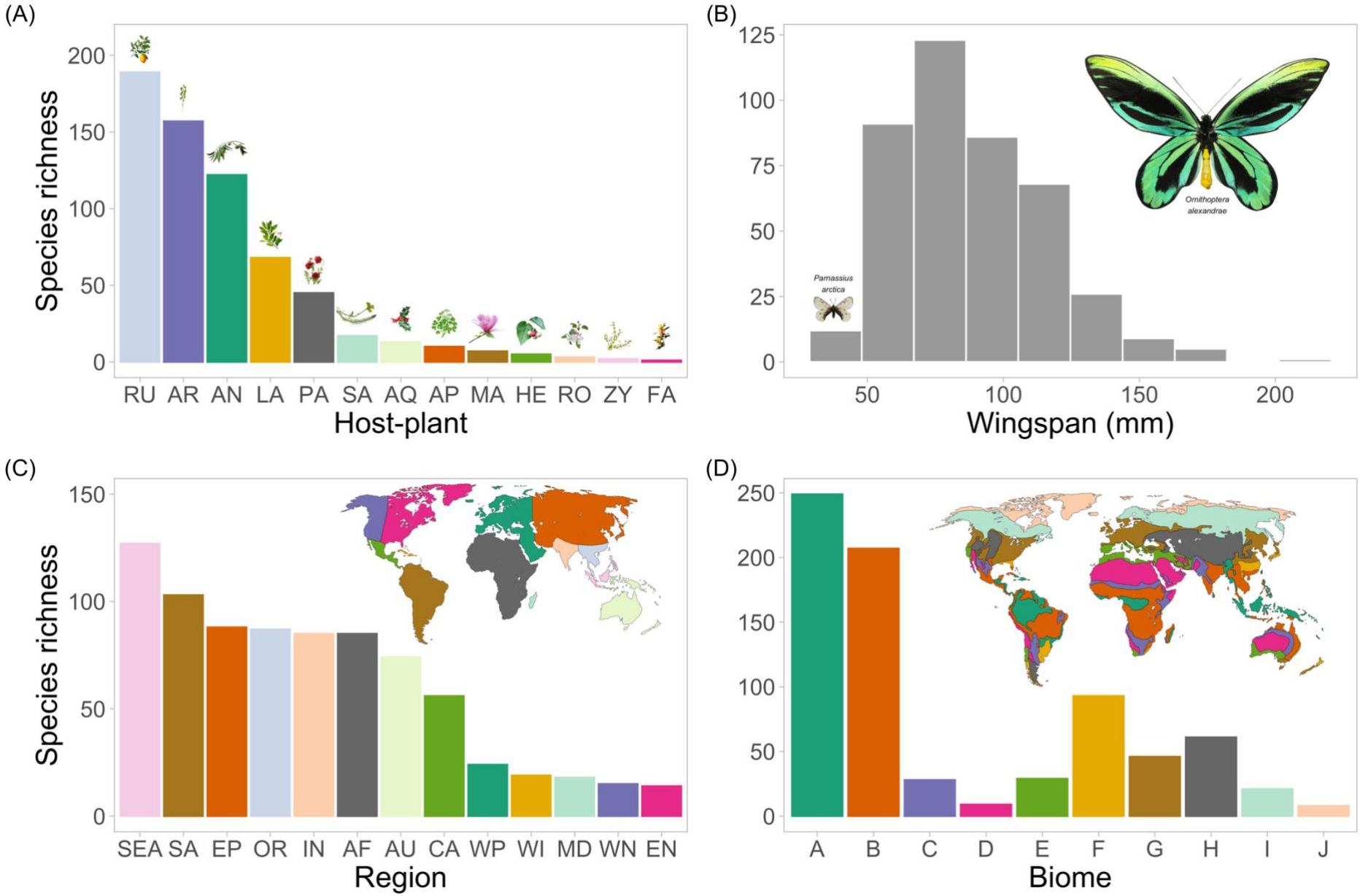

(A)
Species richness
200
150
100
50
0
RU AR AN LA PA SA AQ AP MA HE RO ZY FA
Host-plant
(B)
125
100
75
50
25
0
Parnassius arctica
Ornithoptera alexandrae
50 100 150 200
Wingspan (mm)
(C)
Species richness
150
100
50
0
SEA SA EP OR IN AF AU CA WP WI MD WN EN
Region
(D)
250
200
150
100
50
0
A B C D E F G H I J
Biome

Figure 1: Diversity in the extant Papilionidae species. (A) The frequency of species feeding on each of the principal host-plant families: Annonaceae (AN), Apiaceae (AP), Aquifoliaceae (AQ), Aristolochiaceae (AR), Crassulaceae (SA), Fabaceae (FA), Hernandiaceae (HE), Lauraceae (LA), Magnoliaceae (MA), Papaveraceae (PA), Rosaceae (RO), Rutaceae (RU), and Zygophyllaceae (ZY). Species that feed on multiple plant families are represented in the bars for all these families. **(B)** A histogram of wingspans of all extant species. Images of butterflies are sized relative to each other: *Parnassius arctica* is the smallest Papilionidae species and *Ornithoptera alexandrae* the largest. **(C)** The frequency of species in each geographic region: Western Palearctic (WP), Eastern Palearctic (EP), Western Nearctic (WN), Eastern Nearctic (EN), Central America (CA), Caribbean Islands (WI), South America (SA), Africa (AF), Madagascar and surrounding islands (MA), India (IN), continental South East Asia (OR), insular Southeast Asia (SEA), and Australia and surrounding islands (AU). Species found in multiple regions are represented in multiple bars. The inset map shows how we defined regions, with colours on the map corresponding to the colours of the bars. **(D)** The frequency of species in each biome: evergreen equatorial rainforest (A), tropical deciduous woodland (B), savanna (C), subtropical desert (D), sclerophyllous woodland and shrubland (E), temperate evergreen forest (F), broadleaf deciduous forest (G), steppe (H), boreal coniferous forest (I), and tundra (J). Species found in multiple biomes are represented in multiple bars. The inset map shows how the location of each biome was defined (Olson et al. 2001). Colours on the map correspond to the colours of the bars.

Materials and Methods

Data compilation

An existing maximum-clade credibility dated phylogenetic tree based on five mitochondrial (COI, COII, ND1, ND5, and rRNA 16S) and two nuclear (EF-1a and Wg) genes was used in this study. This tree was based on that generated by Allio et al. (2021), with the addition of thirteen species as well as taxonomic updates for species that had since been renamed. This tree represented 421 of the described 622 Papilionidae species – 68% of the total species diversity – and all genera. However, taxonomic coverage was not equal across the family. Considering only the most species rich tribes, the tree contained: 77% of Papilionini species, 64% of Troidini species, 94% of Parnassiini species, 93% of Zerynthiini species, 100% of Luehdorfiini species, and 42% of Leptocircini species.

Host-plant data was based on Allio et al. (2021) and Hancock (1983). Species were scored for larval presence or absence on their 13 primary host-plant families: Annonaceae (AN), Apiaceae (AP), Aquifoliaceae (AQ), Aristolochiaceae (AR), Crassulaceae (SA), Fabaceae (FA), Hernandiaceae (HE), Lauraceae (LA), Magnoliaceae (MA), Papaveraceae (PA), Rosaceae (RO), Rutaceae (RU), and Zygophyllaceae (ZY).

The world was separated into 13 geographic regions (Figure 1C). This separation was based on present-day biome and geography, while also taking into account the paleogeography of the regions over the last 70 million years (Blakey 2008; Seton et al. 2012). These regions were: Western Palearctic (WP), Eastern Palearctic (EP), Western Nearctic (WN), Eastern Nearctic (EN), Central America (CA), Caribbean Islands (WI), South America (SA), Africa (AF), Madagascar and surrounding islands (MA), India (IN), continental South East Asia (OR), insular Southeast Asia (SEA), and Australia and surrounding islands (AU). Species were scored for presence and absence in each region.

Biomes were defined based on climate and location (Olson et al. 2001; Figure 1D). These 10 biomes were: evergreen equatorial rainforest (A), tropical deciduous woodland (B), savanna (C), subtropical desert (D), sclerophyllous woodland and shrubland (E), temperate evergreen forest (F), broadleaf deciduous forest (G), steppe (H), boreal coniferous forest (I), and tundra (J). Species were scored for presence and absence in each biome.

Morphological data consisted of wingspan and colouration. Wingspan - the distance between the tips of the two forewings - was obtained through a literature search (Appendix 1). This resulted in a dataset of wingspans ranging from 35 to 208 mm. For 30 species, it was only possible to obtain data on wing length: a measure of the distance from the tip of a single forewing to the body. A linear regression was therefore computed between wingspan and wing length for the 125 species with data for

both measures, and then used to convert wing length to wingspan for the 30 species only possessing wing length data. After this, 58 of the 421 species in the phylogeny (14%) still lacked wingspan data. These species were assigned the value of the nearest sister species, or the median value of all data-possessing species if multiple equally close species all had data. Although this may seem circular, closely related species often have very similar morphology (Murillo-Hiller, 2007; Hu et al., 2018). Furthermore, such assignment occurred across the phylogeny, so any errors should affect ancestral state estimation equally.

Colouration data was generated in the form of pairwise comparisons between species (Pinkert, unpublished data) using the R package *colordistance* (Weller 2021) following Weller and Westneat (2019). This recently developed method allows for an objective way to easily classify colour differences between species. In brief, we obtained 426 images from 311 species (thus some species were represented multiple times in the dataset) and all genera of Papilionidae. All of the pixels of an image were treated as coordinates and plotted in a 3D space with the R, G, and B colour channels corresponding to the x, y, and z axis respectively. Each channel/axis was then separated in half, giving a 2 by 2 by 2 grid. As a result, there were eight (2^3) potential bins a pixel could be found within. For each bin, the relative frequency of pixels and the mean RGB value of these pixels was calculated. These two measures were compared between images using the Earth mover's distance metric (Rubner et al. 2000) to give a value for colour difference. This resulted in a dataset containing pairwise comparisons between every combination of images.

Ancestral state estimation

Ancestral states were estimated for host-plant, region, and biome using the Dispersion, Extinction, Cladogenesis "eXtended" (DECX) model (Beeravolu and Condamine 2016). This is an extension of the Dispersion, Extinction, Cladogenesis (DEC) model (Ree and Smith 2008) which allows for estimation of the ancestral state of species ranges by simulating local dispersal and extinction. The DECX model improves upon this by allowing for changes in connectivity over time as a result of paleogeographic changes. Although initially developed for geographic traits, the DECX model can be used for any traits with multiple states encoded in a binary presence/absence manner. For both region and biome presence, adjacency matrices were used to only permit transitions between connected regions/biomes. Tectonic and climatic processes mean that these biomes and regions have not been constant through time (Beeravolu and Condamine 2016), and thus we provided different matrices for different timepoints. For regions, the matrix was set to vary across six time slices, reflecting historic changes in geography (Blakey 2008; Seton et al. 2012). The time slices used were: 0 to 5.33 MYA (Pleistocene and Pliocene), 5.33 to 23.03 MYA (Miocene), 23.03 to 33.9 MYA (Oligocene), 33.9 to 56 MYA (Eocene), 56 to 66

MYA (Paleocene), and 66 MYA to the root age of Papilionidae (late Cretaceous). For biomes, the matrix was set to vary across five time slices: 0 to 5.3 MYA, 5.3 to 11.6 MYA, 11.6 to 15.9 MYA, 15.9 to 33.8 MYA, and 33.8 MYA to the root age of Papilionidae, reflecting the emergence and movement of biomes over time (Ziegler et al. 2003; Pound et al. 2012; Pound and Salzmann 2017). Models were run allowing two or three states per node for host-plant and four or five states for region and biome. This decision was based on the maximum number of states extant species were present in. The model with the best log-likelihood was selected for each trait.

Wingspan data were continuous so a different model was needed to estimate the ancestral states. The ancestral state of wingspan was thus estimated using the fastAnc function from the phytools package (Revell 2012) in R (R Core Team 2021).

Correlated shifts with host-plant shifts

DECX models produce the most likely states for each node, with any one state being made up of one or more of the possible traits - e.g. for host-plant preference, a state could be one plant (giving a specialist species) or multiple plants (giving a generalist species). For each node, we took the most likely state to carry forward for our analysis. To be consistent with how we defined a shift, we looked specifically at gains of a trait to reflect the novel selective pressures brought about by colonisation of a new niche (Schluter 2000; Terekhanova et al. 2014; Marques et al. 2018). Nodes were thus scored for whether they represented a gain of a new host-plant, biome, and geographic region (i.e. a new trait had been acquired since the ancestral node; Figure 2). We then used chi-squared tests to evaluate whether gains of region or biome coincided with gains of host-plant. For this, each node was scored as either gain/gain, gain/no gain, no gain/gain, or no gain/no gain for host-plant and region/biome respectively.

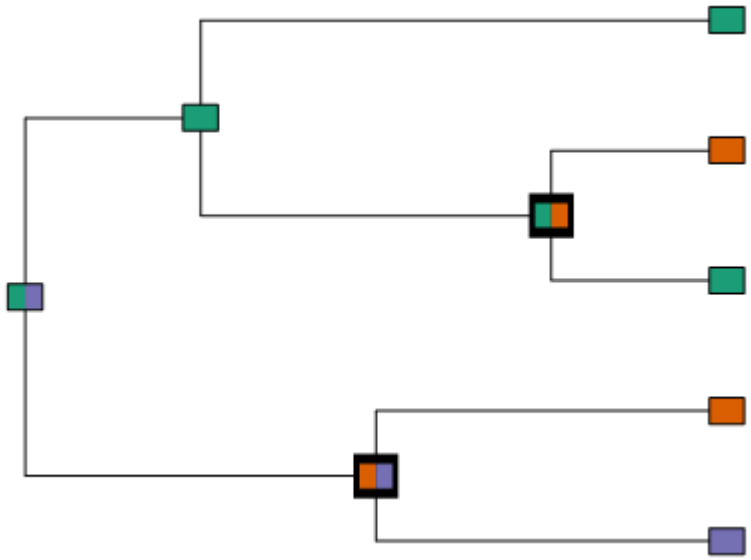


Figure 2: Illustration of the scoring of nodes for gains of a trait on a hypothetical phylogeny. Each colour represents a different state of the trait, with multiple colours at one node showing that the species was present in all of those states. States are shown for tips and internal nodes. Bold outlines denote nodes that would be scored as a gain in our analysis.

Wingspan was a continuous rather than a categorical trait, so nodes could not be scored for gains in the same manner as for biogeographic traits. Instead, the rate of change of wingspan was calculated for each branch. To do so, we computed the difference in wingspan (mm) between every node (including tips) and its most recent ancestor. This value was then divided by the wingspan (mm) of the most recent ancestor (i.e. the start of the branch) so as to be proportional to the size of the species. To convert this value to a rate, it was divided by the branch length in millions of years, giving a final value for each branch in units of MY^{-1}. Each branch was then scored for whether or not a host-plant gain occurred at the ancestral node. We used a Mann-Whitney test to evaluate whether host-plant shifts at the start of a branch affected the rate of change in wingspan. This non-parametric test was selected as it does not assume equal sample sizes or normally distributed data.

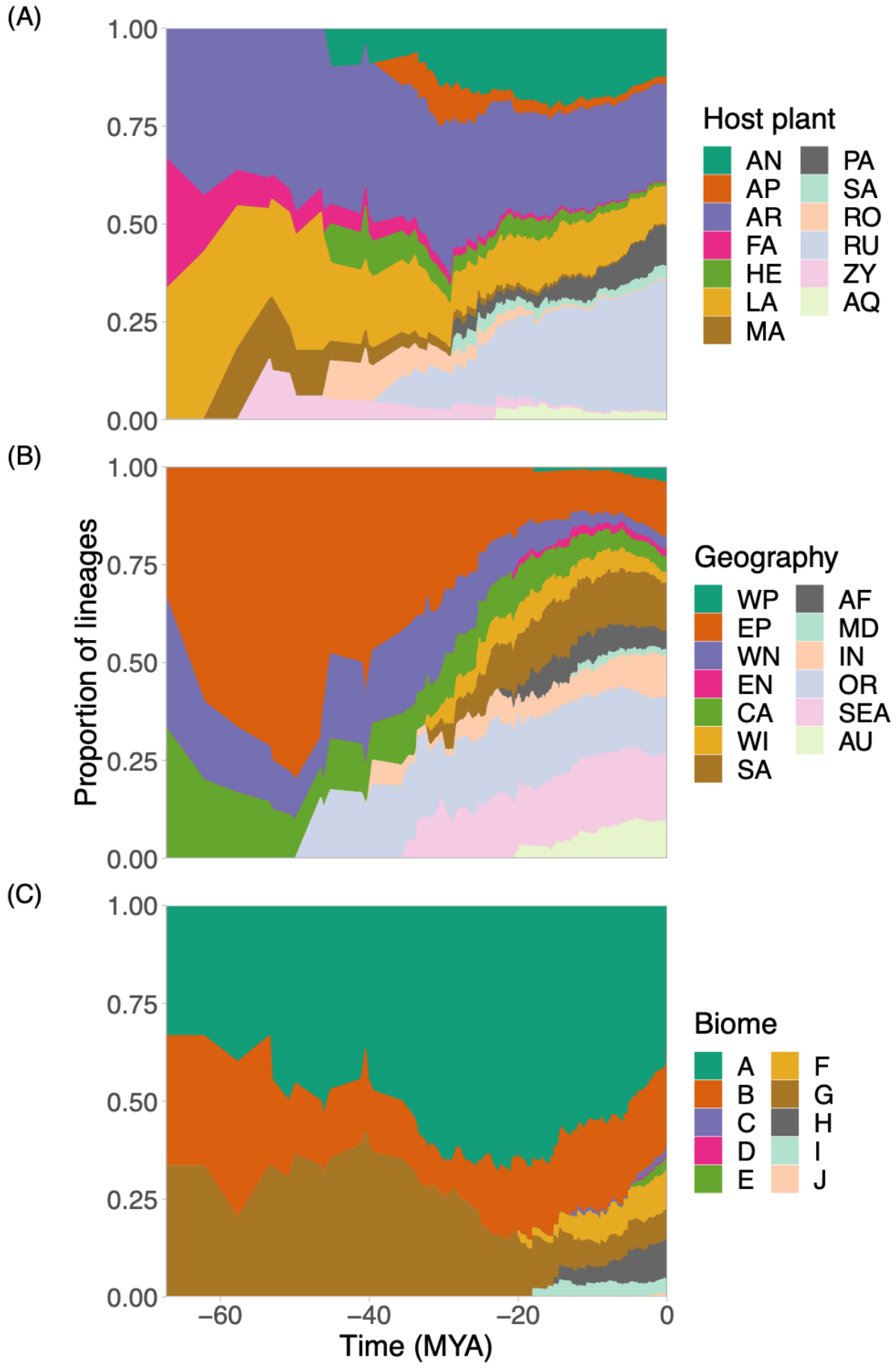
Colour data was purely comprised of comparisons between species pairs so no ancestral estimation was possible. Therefore, we used the ancestral state estimation for host-plant preference to categorise each colour comparison as being between two species with (1) the same history of host-plant preference or (2) different histories of host-plant preference. A different evolutionary history of host-plant preference was used as a proxy for a host-plant shift as it meant that there had been at least one host-plant shift since the pair of species diverged. For each species pair, the time since divergence (i.e. branch length) was calculated using the *cophenetic.phylo* function from the R package *ape* (Paradis and Schliep 2019). Below a divergence time of 14.4 MY, all species pairs shared a history of host-plant preference; and above a divergence time of 49.9 MY, all species pairs had different histories of host-plant preference. So that the two host-plant categories were comparable when accounting for time since

species divergence, we retained only those species pairs that fell between 14.4 and 49.9 MY since divergence. This left us with 15910 data points, with each point representing a comparison between two species. Colour differences were log transformed to normalise the distribution. We were then able to use linear models to analyse this data, with the response variable being the log transformed colour difference between pairs of species and the predictive variables being (1) whether or not there was a shared history of host-plant preference, (2) time since species divergence, and (3) the interaction between these two variables. Non-significant variables were dropped from the model.

Results

Ancestral Papilionidae

The ancestral state was estimated for host-plant, biome, region, and wingspan. For host-plant, all of the most likely states included Aristolochiaceae and Fabaceae plus a third additional plant family – with the most likely of these being Lauraceae. All four most likely estimations for the regions in which the ancestral Papilionidae was found included the eastern Palearctic and western Nearctic, with the most likely also including Central America. We estimated the ancestral biome to be predominately tropical, with the most likely state spanning three biomes: evergreen equatorial rainforest, tropical deciduous woodland, and broadleaf deciduous forest. This ancient Papilionidae was estimated to have a wingspan of 76.5 mm. This is slightly lower than the average wingspan of all extant species: 88.5 ± 27.63 mm (mean ± standard deviation). Over the course of the last 66 million years, Papilionidae have spread to all geographic regions and biomes, colonised 13 different families of host-plant (Figure 3), and diversified in size so that they now range in wingspans from 35 to 210 mm.


(A)
(B)
(C)
1.00
0.75
0.50
0.25
0.00
Proportion of lineages
Host plant
AN
AP
AR
FA
HE
LA
MA
PA
SA
RO
RU
ZY
AQ
Geography
WP
EP
WN
EN
CA
WI
SA
AF
MD
IN
OR
SEA
AU
Biome
A
B
C
D
E
F
G
H
I
J
−60
−40
−20
0
Time (MYA)

234

Figure 3: Lineages through time for host-plant, geography, and biome. The proportion of lineages in each state was plotted against time for **(A)** host-plant, **(B)** region and **(C)** biome. Lineages present in multiple states are included in shaded areas for all of these states. Abbreviations are as follows. Host-plants: Annonaceae (AN), Apiaceae (AP), Aquifoliaceae (AQ), Aristolochiaceae (AR), Crassulaceae (SA), Fabaceae (FA), Hernandiaceae (HE), Lauraceae (LA), Magnoliaceae (MA), Papaveraceae (PA), Rosaceae (RO), Rutaceae (RU), and Zygophyllaceae (ZY). Regions: Western Palearctic (WP), Eastern Palearctic (EP), Western Nearctic (WN), Eastern Nearctic (EN), Central America (CA), Caribbean Islands (WI), South America (SA), Africa (AF), Madagascar and surrounding islands (MA), India (IN), continental South East Asia (OR), insular Southeast Asia (SEA), and Australia and surrounding islands (AU). Biomes: evergreen equatorial rainforest (A), tropical deciduous woodland (B), savanna (C), subtropical desert (D), sclerophyllous woodland and shrubland (E), temperate evergreen forest (F), broadleaf deciduous forest (G), steppe (H), boreal coniferous forest (I), and tundra (J).

Biogeographic shifts

Host-plant gains coincided with expansion only into new regions (Figure 4). Specifically, gaining a new host-plant at a node was significantly associated with gaining a new region (χ^2 = 12.7, n = 841, p < 0.001; Figure 4A), with expansion into a new region occurring at 69% (9/13) of the host-plant-gain nodes and 23% (190/828) of other nodes. Colonisation of a new biome was not associated with gaining a new host-plant (χ^2 = 1.28, n = 841, p = 0.259; Figure 4B). A new biome was colonised at 46% of the host-plant-gain nodes (6/13) and 28% (232/828) of other nodes.

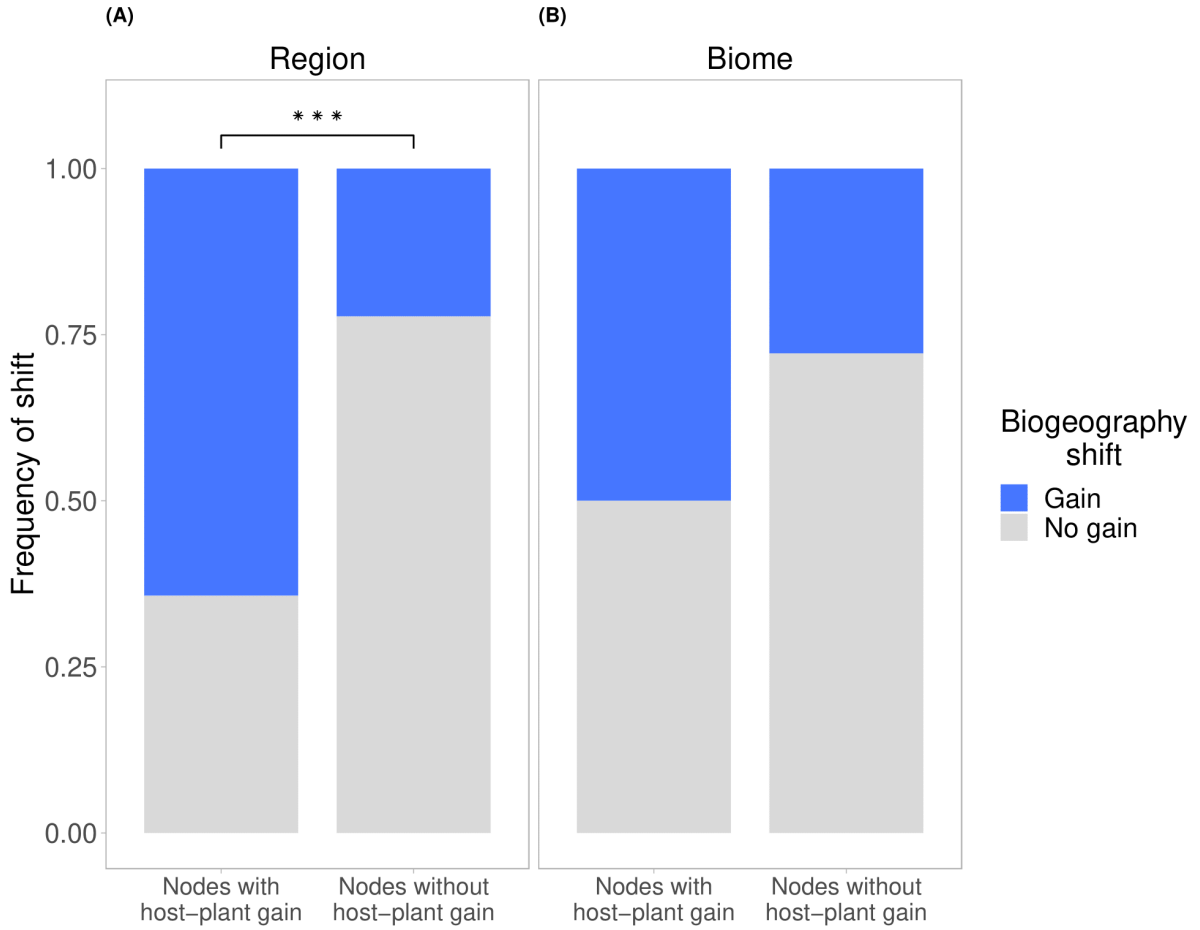


Figure 4: Expansion into new regions and biomes. The frequency of expansion into new regions (A) and biomes (B) for host-plant-gain nodes (n = 13) was compared to that for all other nodes (n = 828) using a chi-squared test.

Morphological shifts

Host-plant shifts affected the evolution of both colouration and wingspan. Colouration was more similar between species with a shared evolutionary history of host-plant preference than between those that were separated by host-plant shifts, as well as between species that were more closely related ($F = 572$, $p < 0.001$ and $F = 18.6$, $p < 0.001$ respectively; Figure 5). We did not detect a significant interaction between time since species divergence and the evolutionary history of host-plant preference ($F = 1.83$, $p = 0.176$) so this term was dropped from the model.

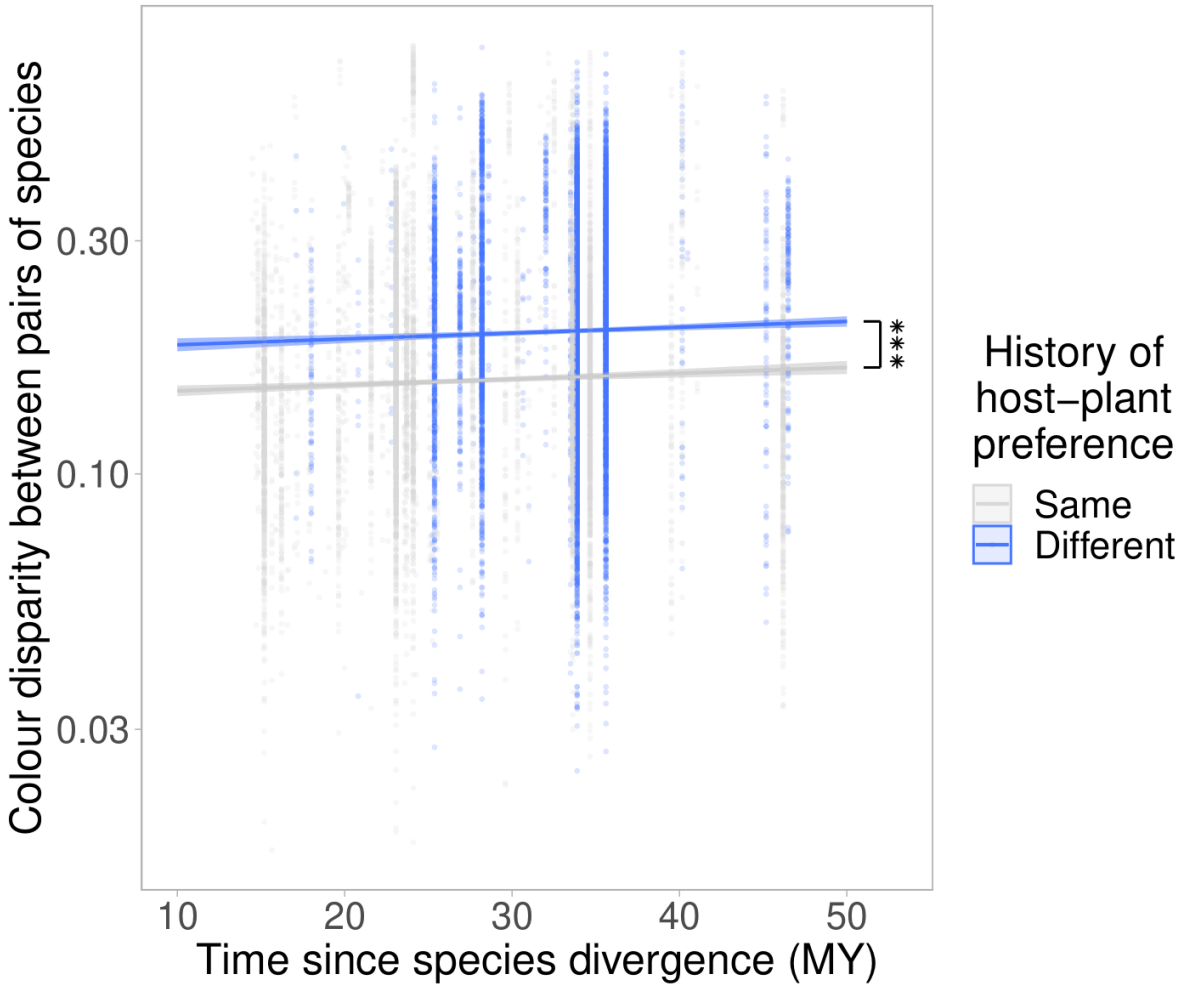


Figure 5: The effect of historical host-plant preference on modern day colour disparity. Species pairs separated by at least one host-plant shift (blue, n = 8384) were compared to those pairs where no host-plant shifts have occurred for either since divergence (grey, n = 7526), accounting for the time since the species diverged.

Wingspan evolution slowed down on branches immediately following the gain of a new host-plant. Branches with host-plant gains at the previous node showed significantly lower log rates of change in wingspan per unit time compared to those branches not directly after a host-plant gain ($t = -3.40$, $df = 21$, $p = 0.003$; Figure 6). Specifically, the wingspan on branches directly following host-plant gains changed at a rate of $8.16 \times 10^{-4} \pm 8.31 \times 10^{-4}$ MY^{-1} (mean ± standard deviation) whilst the wingspan on all other branches changed at a rate of $2.954 \times 10^{-3} \pm 5.134 \times 10^{-3}$ MY^{-1}.

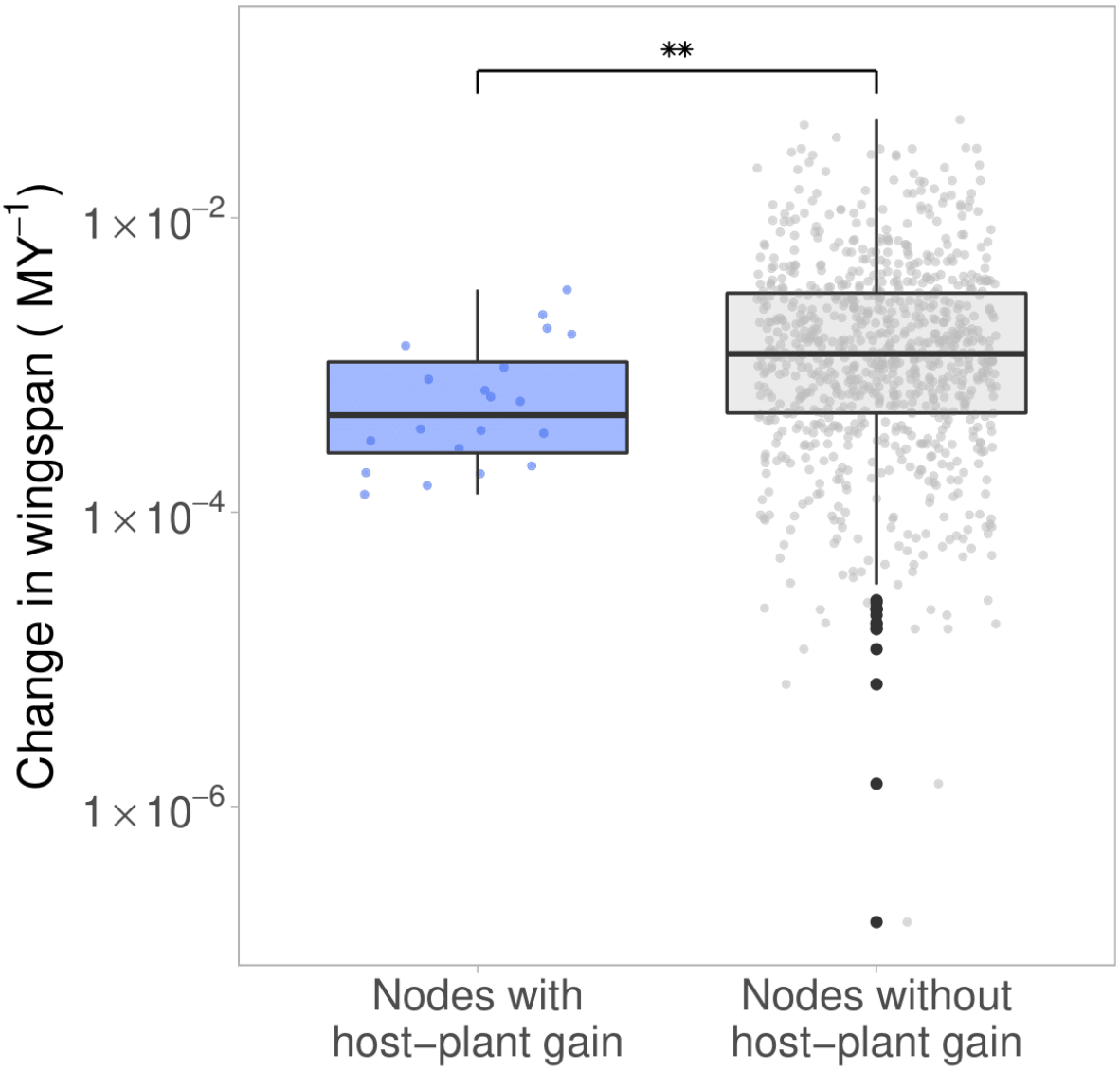


Figure 6: Nodes with host-plant gains were followed by slower evolution of wingspan. The change in wingspan along a branch was significantly lower ($p = 0.003$) for branches directly following host-plant gains ($n = 20$) than for all other branches ($n = 820$).

Discussion

In this study we estimated the ancestral state of biogeographic and morphological traits of Papilionidae and the cooccurrence of shifts in these traits with shifts in host-plants. Our estimation of the ancestral host-plant matched the most recent estimation (Allio et al. 2021). Similarly, our estimation of the regional distribution was in accord with previous work (Condamine et al. 2012, 2013) despite using a more up-to-date phylogeny with twice as many species (421 compared to 203). Previous work had not explicitly estimated the ancestral biome of Papilionidae. However, our estimation of the ancestral biome as predominantly tropical is in accordance with boreal regions - which include the Western Nearctic and Eastern Palearctic - being tropical at the time when Papilionidae originated (Ziegler et al. 2003; Sluijs et al. 2006; Suan et al. 2017). To the best of our knowledge, the ancestral Papilionidae wingspan has not been estimated before and so we present the first estimation of this.

Fossil evidence in the literature provides an opportunity to assess the accuracy of our models. Our estimation of the Western Nearctic being included in the ancestral region matches the location of the oldest known Papilionidae fossil: *Praepapilio* Durden and Rose, 1978. However, the other two known fossils - *Thaites ruminiana* Scudder, 1875 and *Doritites bosniackii* Rebel, 1989 - were both found in the Western Palearctic (Scudder 1875; Rebel 1898) but our model recovered their closest ancestors as being found in the Eastern Palearctic, where Parnassiinae species richness is highest today. Similarly, although our models for biome preference recovered the root as matching the *Praepapilio* biome (Golovneva 2000; Spicer and Herman 2010), there was a mismatch for *Thaites ruminiana* and *Doritites bosniackii*. These two fossils were from regions thought to be temperate evergreen forests (Pound et al. 2012; Pound and Salzmann 2017), but we estimated their closest ancestor to have lived in broad-leaf deciduous forests. Both of these fossils correspond to extinct lineages, rather than being directly representative of any node (Nazari et al. 2007), so perhaps the mismatch is a result of lineages that dispersed but ultimately failed to persist. Finally, although measuring the exact wingspan from a fossil will not be as precise as inferring location, it can be used as a rough gauge. *Praepapilio* is the only fossil of a high enough quality to estimate the wingspan, and does appear to have had a wingspan similar to our estimation of 76.5 mm (Durden and Rose 1978), lending support to our estimation of wingspan.

We found that colonisation of new host-plants occurred in tandem with expansion into new regions. This suggests that some of the signals of selection present around host-plant shifts (Allio et al. 2021) may also have been a result of adaptation to novel selection pressures imposed by a new region. Whether this association was causative or purely correlational is not something that can be easily evaluated. However, it is eminently plausible that there was a causative link between at least some of the cooccurrences of host-plant gains and regional expansions. Perhaps expansion into a new region enabled colonisation of a host-plant that was not present in the ancestral region. For instance, the

colonisation of Hernandiaceae and Annonaceae at the crown of the Leptocircini coincided with expansion into the Oriental region from the Eastern Palearctic. With neither plant family present in this ancestral region (Michalak et al. 2010; Couvreur et al. 2011), the expansion into the Oriental region likely facilitated colonisation of these novel host-plants. In the early stages of Papilionidae diversification, the family was co-distributed with Aristolochiaceae (Allio et al. 2021) - an ancestral host-plant - further indicating that plant distribution can indeed play an important role in host-plant preference. Future work should therefore try to marry host-plant biogeography with Papilionidae biogeography to more systematically evaluate such patterns.

Species were found to be more similar in colouration if they had a shared history of host-plant preference. It has been suggested that toxic species evolve resemblance to each other (Mullerian mimicry: Müller 1878; Sherratt 2008) to send consistent signals to predators. Perhaps, at least for species feeding on toxic host-plants, this has contributed to those species sharing a host-plant more closely resembling each other. Additionally, this resemblance might be indicative of rapid colour evolution following a host-plant shift, resulting in lineages that diverged in host-plant preference also diverging in colour. This could come about if different host-plants provided different selective pressures. For instance, shifting to a more toxic plant could have led to selection for aposematic colouration (Mallet and Joron 1999), and thus a divergence in colour from a sister lineage that fed on non-toxic plants. However, this does seem to be at odds with the high levels of Batesian mimicry within the family, with non-toxic species mimicking toxic ones (Kunte 2009). Interestingly, we found the exact opposite for wingspan, with slower rates of wingspan evolution directly following colonisation of a new host-plant. The reasons for this are not clear. Moreover, it is the opposite pattern to that found in *Morpho* butterflies, where the most pronounced changes in wingspan followed host-plant shifts (Chazot et al. 2021). Perhaps host-plant shifts acted as bottlenecks (Munday et al. 2004), with only very few species switching their preference. The resulting lack of genetic diversity might have slowed the rate of evolution (Barrett and Schluter 2008). If this was the case, it would imply that there must have been strong selective pressures on colouration to enable it to overcome this constraint and rapidly evolve. Regardless of the underlying reasons, these findings suggest that some of the genetic signals of selection seen in swallowtails around host-plant shifts (Allio et al. 2021) may be in genes involved in controlling colouration but not in genes underpinning size.

By improving our understanding of concurrently occurring events, we highlight why host-plant shifts should not be studied in isolation but rather in concert with other traits that may covary with plant preferences. We also provide suggestions for some of the morphological traits that might have been evolving rapidly around such host shifts. Taking into account the possibility of confounding effects resulting from parallelly shifting traits will allow more accurate assessments of the impact of host-plant shifts in the future.

References

Allio, R., B. Nabholz, S. Wanke, G. Chomicki, O. A. Pérez-Escobar, A. M. Cotton, A.-L. Clamens, G. J. Kergoat, F. A. H. Sperling, and F. L. Condamine. 2021. Genome-wide macroevolutionary signatures of key innovations in butterflies colonizing new host plants. Nature Communications 12:354.

Barrett, R. D. H., and D. Schluter. 2008. Adaptation from standing genetic variation. Trends in Ecology & Evolution 23:38–44.

Beeravolu, C. R., and F. L. Condamine. 2016. An Extended Maximum Likelihood Inference of Geographic Range Evolution by Dispersal, Local Extinction and Cladogenesis. bioRxiv 038695.

Bishop, T. R., M. P. Robertson, H. Gibb, B. J. Van Rensburg, B. Braschler, S. L. Chown, S. H. Foord, T. C. Munyai, I. Okey, and P. G. Tshivhandekano. 2016. Ant assemblages have darker and larger members in cold environments. Global Ecology and Biogeography 25:1489–1499.

Blakey, R. C. 2008. Gondwana paleogeography from assembly to breakup—A 500 my odyssey. Geological Society of America Special Papers 441:1–28.

Chazot, N., P. Blandin, V. Debat, M. Elias, and F. L. Condamine. 2021. Punctual ecological changes rather than global factors drive species diversification and the evolution of wing phenotypes in Morpho butterflies. bioRxiv 2021.04.18.440344.

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

Condamine, F. L., J. Rolland, S. Höhna, F. A. H. Sperling, and I. Sanmartín. 2018. Testing the Role of the Red Queen and Court Jester as Drivers of the Macroevolution of Apollo Butterflies. Systematic Biology 67:940–964.

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

Condamine, F. L., F. A. H. Sperling, N. Wahlberg, J.-Y. Rasplus, and G. J. Kergoat. 2012. What causes latitudinal gradients in species diversity? Evolutionary processes and ecological constraints on swallowtail biodiversity. Ecology Letters 15:267–277.

Couvreur, T. L. P., M. D. Pirie, L. W. Chatrou, R. M. K. Saunders, Y. C. F. Su, J. E. Richardson, and R. H. J. Erkens. 2011. Early evolutionary history of the flowering plant family Annonaceae: steady diversification and boreotropical geodispersal. Journal of Biogeography 38:664–680.

Durden, C. J., and H. Rose. 1978. Butterflies from the Middle Eocene: The Earliest Occurrence of Fossil Papilionoidea (Lepidoptera). Texas Memorial Museum, The University of Texas at Austin.

Edger, P. P., H. M. Heidel-Fischer, M. Bekaert, J. Rota, G. Glöckner, A. E. Platts, D. G. Heckel, J. P. Der, E. K. Wafula, M. Tang, J. A. Hofberger, A. Smithson, J. C. Hall, M. Blanchette, T. E.

Bureau, S. I. Wright, C. W. dePamphilis, M. E. Schranz, M. S. Barker, G. C. Conant, N. Wahlberg, H. Vogel, J. C. Pires, and C. W. Wheat. 2015. The butterfly plant arms-race escalated by gene and genome duplications. Proceedings of the National Academy of Sciences of the USA 112:8362–8366.

Ehrlich, P. R., and P. H. Raven. 1964. Butterflies and Plants: A Study in Coevolution. Evolution 18:586–608.

Farrell, B. D. 1998. "Inordinate Fondness" Explained: Why Are There So Many Beetles? Science 281:555–559.

Fordyce, J. A. 2010. Host shifts and evolutionary radiations of butterflies. Proceedings of the Royal Society B: Biological Sciences 277:3735–3743.

Golovneva, L. B. 2000. The Maastrichtian (Late Cretaceous) climate in the Northern Hemisphere. Geological Society, London, Special Publications 181:43–54.

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

Igea, J., and A. J. Tanentzap. 2020. Angiosperm speciation cools down in the tropics. Ecology Letters 23:692–700.

Igea, J., and A. J. Tanentzap. 2019. Multiple macroevolutionary routes to becoming a biodiversity hotspot. Science Advances 5:eaau8067.

Kunte, K. 2009. The Diversity and Evolution of Batesian Mimicry in Papilio Swallowtail Butterflies. Evolution 63:2707–2716.

Laurila, A., B. Lindgren, and A. T. Laugen. 2008. Antipredator Defenses Along a Latitudinal Gradient in Rana Temporaria. Ecology 89:1399–1413.

Mallet, J., and M. Joron. 1999. Evolution of diversity in warning color and mimicry: polymorphisms, shifting balance, and speciation. Annual review of ecology and systematics 30:201–233.

Marques, D. A., F. C. Jones, F. Di Palma, D. M. Kingsley, and T. E. Reimchen. 2018. Experimental evidence for rapid genomic adaptation to a new niche in an adaptive radiation. Nat Ecol Evol 2:1128–1138.

Michalak, I., L.-B. Zhang, and S. S. Renner. 2010. Trans-Atlantic, trans-Pacific and trans-Indian Ocean dispersal in the small Gondwanan Laurales family Hernandiaceae. Journal of Biogeography 37:1214–1226.

Mitter, C., B. Farrell, and B. Wiegmann. 1988. The Phylogenetic Study of Adaptive Zones: Has Phytophagy Promoted Insect Diversification? The American Naturalist 132:107–128.

Müller, F. 1878. Über die vortheile der mimicry bei schmetterlingen. Zoologischer Anzeiger 1:54–55.

Munday, P. L., L. van Herwerden, and C. L. Dudgeon. 2004. Evidence for Sympatric Speciation by Host Shift in the Sea. Current Biology 14:1498–1504.

Nazari, V., E. V. Zakharov, and F. A. H. Sperling. 2007. Phylogeny, historical biogeography, and taxonomic ranking of Parnassiinae (Lepidoptera, Papilionidae) based on morphology and seven

genes. Molecular Phylogenetics and Evolution 42:131–156.
Olson, D. M., E. Dinerstein, E. D. Wikramanayake, N. D. Burgess, G. V. Powell, E. C. Underwood, J. A. D'amico, I. Itoua, H. E. Strand, and J. C. Morrison. 2001. Terrestrial Ecoregions of the World: A New Map of Life on Earth. BioScience 51:933–938.
Paradis, E., and K. Schliep. 2019. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. Bioinformatics 35:526–528.
Pinkert, S. unpublished data.
Pinkert, S., R. Brandl, and D. Zeuss. 2017. Colour lightness of dragonfly assemblages across North America and Europe. Ecography 40:1110–1117.
Pound, M. J., A. M. Haywood, U. Salzmann, and J. B. Riding. 2012. Global vegetation dynamics and latitudinal temperature gradients during the Mid to Late Miocene (15.97–5.33Ma). Earth-Science Reviews 112:1–22.
Pound, M. J., and U. Salzmann. 2017. Heterogeneity in global vegetation and terrestrial climate change during the late Eocene to early Oligocene transition. Scientific Reports 7:43386.
Pyron, R. A., and J. J. Wiens. 2013. Large-scale phylogenetic analyses reveal the causes of high tropical amphibian diversity. Proceedings of the Royal Society B: Biological Sciences 280:20131622.
R Core Team. 2021. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.
Rebel, H. 1898. Doritites bosniaskii. Sitzungsberichte der akademie der wissenschaften. Mathematischen-naturwissenschaftliche classe. Abteilung 1: Mineralogie, biologie, erdkunde.Wien. 107:734–741, 745.
Ree, R. H., and S. A. Smith. 2008. Maximum likelihood inference of geographic range evolution by dispersal, local extinction, and cladogenesis. Systematic Biology 57:4–14.
Revell, L. J. 2012. phytools: an R package for phylogenetic comparative biology (and other things). Methods in Ecology and Evolution 3:217–223.
Rolland, J., F. L. Condamine, F. Jiguet, and H. Morlon. 2014. Faster Speciation and Reduced Extinction in the Tropics Contribute to the Mammalian Latitudinal Diversity Gradient. PLOS Biology 12:e1001775.
Rubner, Y., C. Tomasi, and L. J. Guibas. 2000. The Earth Mover's Distance as a Metric for Image Retrieval. International Journal of Computer Vision 40:99–121.
Sahoo, R. K., A. D. Warren, S. C. Collins, and U. Kodandaramaiah. 2017. Hostplant change and paleoclimatic events explain diversification shifts in skipper butterflies (Family: Hesperiidae). BMC Evolutionary Biology 17:174.
Schluter, D. 2000. The ecology of adaptive radiation. OUP Oxford.
Scudder, S. H. 1875. Fossil butterflies. Memoirs of the American Association for the Advancement of Science I: I-XI:1–99.
Seton, M., R. D. Müller, S. Zahirovic, C. Gaina, T. Torsvik, G. Shephard, A. Talsma, M. Gurnis, M.

458 Turner, and S. Maus. 2012. Global continental and ocean basin reconstructions since 200 Ma.
459 Earth-Science Reviews 113:212–270.
460 Sheard, C., M. H. C. Neate-Clegg, N. Alioravainen, S. E. I. Jones, C. Vincent, H. E. A. MacGregor, T.
461 P. Bregman, S. Claramunt, and J. A. Tobias. 2020. Ecological drivers of global gradients in
462 avian dispersal inferred from wing morphology. Nature Communications 11:2463.
463 Sherratt, T. N. 2008. The evolution of Müllerian mimicry. Naturwissenschaften 95:681.
464 Sluijs, A., S. Schouten, M. Pagani, M. Woltering, H. Brinkhuis, J. S. S. Damsté, G. R. Dickens, M.
465 Huber, G.-J. Reichart, and R. Stein. 2006. Subtropical Arctic Ocean temperatures during the
466 Palaeocene/Eocene thermal maximum. Nature 441:610–613.
467 Spicer, R. A., and A. B. Herman. 2010. The Late Cretaceous environment of the Arctic: A quantitative
468 reassessment based on plant fossils. Palaeogeography, Palaeoclimatology, Palaeoecology
469 295:423–442.
470 Stelbrink, P., S. Pinkert, S. Brunzel, J. Kerr, C. W. Wheat, R. Brandl, and D. Zeuss. 2019. Colour
471 lightness of butterfly assemblages across North America and Europe. Scientific Reports 9:1760.
472 Stork, N. E. 2018. How many species of insects and other terrestrial arthropods are there on Earth?
473 Annual Review of Entomology 63:31–45.
474 Suan, G., S.-M. Popescu, J.-P. Suc, J. Schnyder, S. Fauquette, F. Baudin, D. Yoon, K. Piepjohn, N. N.
475 Sobolev, and L. Labrousse. 2017. Subtropical climate conditions and mangrove growth in
476 Arctic Siberia during the early Eocene. Geology 45:539–542.
477 Terekhanova, N. V., M. D. Logacheva, A. A. Penin, T. V. Neretina, A. E. Barmintseva, G. A. Bazykin,
478 A. S. Kondrashov, and N. S. Mugue. 2014. Fast Evolution from Precast Bricks: Genomics of
479 Young Freshwater Populations of Threespine Stickleback Gasterosteus aculeatus. PLOS
480 Genetics 10:e1004696.
481 Weiss, J.-C. 1991. The Parnassiinae of the world. Sciences Nat; Hillside Books.
482 Weller, H. 2021. colordistance: Distance Metrics for Image Color Similarity.
483 Weller, H. I., and M. W. Westneat. 2019. Quantitative color profiling of digital images with earth
484 mover's distance using the R package colordistance. PeerJ 7:e6398.
485 Winkler, I. S., C. Mitter, and S. J. Scheffer. 2009. Repeated climate-linked host shifts have promoted
486 diversification in a temperate clade of leaf-mining flies. Proceedings of the National Academy
487 of Sciences of the USA 106:18103–18108.
488 Xing, S., T. C. Bonebrake, C. C. Tang, E. J. Pickett, W. Cheng, S. E. Greenspan, S. E. Williams, and
489 B. R. Scheffers. 2016. Cool habitats support darker and bigger butterflies in Australian tropical
490 forests. Ecology and evolution 6:8062–8074.
491 Yamamoto, T., M. Hattori, Y. Matsumoto, S. Ueda, and T. Itino. 2020. Evolutionary diversification of
492 Japanese Stomaphis aphids (Aphididae, Lachninae) in relation to their host plant use and ant
493 association. The Science of Nature 107:14.
494 Ziegler, A., G. Eshel, P. M. Rees, T. Rothfus, D. Rowley, and D. Sunderlin. 2003. Tracing the tropics

across land and sea: Permian to present. Lethaia 36:227–254.

Abstract

Understanding why biodiversity is not evenly distributed on Earth across space, time, and lineages is a key question in macroevolutionary studies. In particular, the great diversity of insects raises the question of the role of biotic factors and key innovations, such as the ability to circumvent plant defenses, in the diversification of organisms. Although this is one of the main theories for the diversification of phytophagous insects, numerous climatic, morphological or contingent factors could have also played an important role in their evolutionary history. To address these questions, it is necessary to build a robust and well-sampled phylogenetic framework at the species level, which can be challenging in the case of non-model organisms. In this thesis, I address this issue by reconstructing nearly complete, robust, and dated phylogenies of two tribes of swallowtail butterflies, a charismatic family of Lepidoptera with a worldwide distribution. I revise the taxonomy of the two groups using genomic approaches for species delimitation at the population level, production of reference genomes for several key genera and over 280 whole-genomes, and in-depth phylogenetic studies. Based on this work, I infer the evolutionary history of the two tribes in a biotic (host plant) and biogeographic framework. This comparative study represents a major advance in our understanding of the evolutionary history of Papilionidae, and emphasizes the importance of well-sampled studies in accurately unraveling the factors behind their diversification.

Key words: Biogeography, Diversification, Host-plant shifts, Leptocircini, Molecular dating, Phylogenomics, Reference genomes, Troidini

Résumé

La diversité des espèces vivantes sur Terre n'est pas répartie équitablement dans l'espace, le temps, et entre lignées. L'origine de cette disparité est un questionnement phare des études macroévolutives. La grande diversité des insectes, en particulier, pose la question du rôle des facteurs biotiques et des innovations clefs tel que le contournement de défense des plantes dans la diversification des organismes. Bien que cela constitue l'une des théories principales de la diversification des insectes phytophages, de nombreux facteurs climatiques, morphologiques, ou contingents pourraient également avoir été des acteurs majeurs dans leur histoire évolutive. Pour explorer ces questions, il est nécessaire de se placer dans un cadre phylogénétique robuste et bien échantillonné, ce qui peut représenter un défi important pour des organismes non-modèles. Dans cette thèse, je traite cette question en reconstruisant les phylogénies quasi exhaustives, robustes et datées de deux tribus de la famille des Papilionidae, des papillons charismatiques à la répartition mondiale. A l'aide d'approches génomiques de délimitation d'espèces à l'échelle populationnelle, de productions de génomes complets et de génomes de référence, ainsi que des études phylogénomiques poussées, je révise la taxonomie des deux groupes, et infère leur histoire évolutive d'un point de vue biotique (plante hôte) et biogéographique. Cette étude comparative représente une avancée majeure pour la compréhension de l'histoire évolutive de la famille des Papilionidae, et souligne l'importance d'un échantillonnage exhaustif pour appréhender précisément les facteurs de leur diversification.

Mots-clefs : Biogéographie, Datation moléculaire, Diversification, Génomes de référence, Interactions plantes-insectes, Leptocircini, Phylogénomiques, Troidini