Phylogenomics and comparative genomics in ant-eating mammals

Rémi Allio

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

En biologie des populations et écologie

École doctorale GAIA

Unité de recherche UMR 5554 ISEM

Phylogenomics and comparative genomics in myrmecophagous mammals

Présentée par Rémi ALLIO
Le 11 Février 2021

Sous la direction de Frédéric DELSUC

Devant le jury composé de

Marie SEMON, Maitre de conférence, ENS Lyon, France	Rapporteur
Robert WATERHOUSE, Professeur assistant, Université de Lausane, Suisse	Rapporteur
Emmanuelle JOUSSELIN, Professeure, INRAe, Montpellier, France	Examinateur
Carole SMADJA, Directrice de recherche, CNRS, Montpellier, France	Président
Ludovic ORLANDO, Directeur de recherche, CNRS, Toulouse, France	Examinateur
Michael FONTAINE, Chargé de recherche, CNRS, Montpellier, France	Invité
Benoit NABHOLZ, Maitre de conférence, Université de Montepellier, France	Invité
Frédéric DELSUC, Directeur de recherche, CNRS, Montpelllier, France	Directeur de thèse

UNIVERSITÉ DE MONTPELLIER

« Nothing in biology makes sense except in the light of evolution »

Theodosius Dobzhansky

– SUMMARY –

SCIENTIFIC CONTRIBUTIONS

Publications

Allio, R. *, Teullet, S.*, Lutgen, D.*, Magdeleine, A., Koual, R., Tilak, M.-K., Emerling, C. A., Lefebure, T., & Delsuc, F. (2021). Comparative transcriptomics reveals divergent paths of chitinase evolution underlying dietary convergence in ant-eating mammals. *In prep*.

Condamine, F. L., **Allio, R.**, Cotton, A. M., Hu, S.-J., Kunte, K., & Sperling, F. A. H. (2021). A comprehensive time-calibrated phylogeny illuminates the evolutionary origins of the diverse and globally distributed butterfly genus *Papilio*. *In prep.*

Allio, R., Tilak, M. K., Scornavacca, C., Avenant, N. L., Corre, E., Nabholz, B., & Delsuc, F. (2020). High-quality carnivore genomes from roadkill samples enable species delimitation in aardwolf and bat-eared fox. *eLife. Review round 2*. (link)

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. (2020). Genome-wide macroevolutionary signatures of key innovations in butterflies colonizing new host plants. *Nature communications*. *Accepted*. (link)

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. *Molecular Ecology Resources. 2020; 20: 892– 905.* (link)

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. *Systematic Biology*, 69(1), 38-60. (link).

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. *Molecular Biology and Evolution*, 34(11), 2762-2772. (link)

** equal contribution*

Awards

"**Best Graduate Student Paper Award**" for Mol. Biol. Evol. (2017)
Society for Molecular Biology & Evolution (Yokohama, Japan 2018)

Scientific communications

Allio R., Tilak M.-K., Scornavacca, C., Avenant N.L., Corre E., Nabholz B. & Delsuc F. (5-9 October). Genomics from roadkill enable species delimitation in aardwolf and bat-eared fox. **Biodiversity Genomics 2020** (Sanger Institute, UK). [Virtual talk]

Allio R., Tilak M.-K., Avenant N.L., Corre E., Nabholz B. & Delsuc F. (4-5 February 2020). Roadkill genomics: high quality mammalian genomes from hybrid assembly of short Illumina reads and MinION long reads. **Rencontres ALPHY : Génomique Evolutive, Bioinformatique, Alignement et Phylogénie** (Lyon, France). [Oral communication]

Allio, R., Tilak, M.-K., Magdeleine, A., Nabholz, B., & Delsuc, F. **(3** February 2020). How roadkill can become a valuable resource for genome-wide analyses. **LEHNA** (Lyon, France). [Seminar communication]

Allio R., Romiguier J., Nabholz B. & Delsuc F. (21-25 July 2019). Extracting complementary mitogenomic data from target enrichment experiments: a case study with 501 ant UCE libraries. **Annual Meeting of the Society for Molecular Biology and Evolution** (Manchester, UK). [Poster]

Allio R., Romiguier J., Nabholz B. & Delsuc F. (7-8 February 2019). In search of mitochondrial DNA from Ultra Conserved Elements sequencing data. **ALPHY: Bioinformatics and Evolutionary Genomics** (Paris, France). [Oral Communication]

Allio R., Koual R., Tilak M.-K., Avenant N.L., Nabholz B. & Delsuc F. (16-17 October 2018). Testing the hypothesis of allopatric speciation through biogeographical disjunction in three species of African carnivores. **9th Annual Oppenheimer De Beers Research Conference** (Johannesburg, South Africa). [Poster]

Allio R., Koual R., Tilak M.-K., Avenant N.L., Nabholz B. & Delsuc F. (19-22 August 2018). Testing the hypothesis of allopatric speciation through biogeographical disjunction in three species of African carnivores (aardwolf, bat-eared fox, and black-backed jackal)**. 2nd Joint Congress in Evolutionary Biology** (Montpellier, France). [Poster]

Remerciements

"Mais, vous savez, moi je ne crois pas qu'il y ait de bonne ou de mauvaise situation. Moi, si je devais résumer ma vie aujourd'hui avec vous, je dirais que c'est d'abord des rencontres, des gens qui m'ont tendu la main, peut-être à un moment où je ne pouvais pas, où j'étais seul chez moi. Et c'est assez curieux de se dire que les hasards, les rencontres forgent une destinée... Parce que quand on a le goût de la chose, quand on a le goût de la chose bien faite, le beau geste, parfois on ne trouve pas l'interlocuteur en face, je dirais, le miroir qui vous aide à avancer. Alors ce n'est pas mon cas, comme je le disais là, puisque moi au contraire, j'ai pu ; et je dis merci à la vie, je lui dis merci, je chante la vie, je danse la vie... Je ne suis qu'amour ! Et finalement, quand beaucoup de gens aujourd'hui me disent "Mais comment fais-tu pour avoir cette humanité ?", eh ben je leur réponds très simplement, je leur dis que c'est ce goût de l'amour, ce goût donc qui m'a poussé aujourd'hui à entreprendre une construction mécanique, mais demain, qui sait, peut-être seulement à me mettre au service de la communauté, à faire le don, le don de soi..." Edouard Baer (Otis)

Pour commencer, je voudrais remercier mon encadrement au sens large : Marie-Ka, Fabien, Benoit et Fred !

Marie-ka, merci pour ces trois ans !
Sans toi ma thèse ne serait pas celle quelle est aujourd'hui. Tu as été très largement impliquée depuis le début. J'ai toujours senti que tu voulais vraiment que tout marche pour le bon déroulement de ma thèse. De manière générale c'est ton souhait pour toutes les thèses au labo et pour tous les projets qui passent entre tes mains. J'ai énormément apprécié la place que tu m'as donnée dans le développement du protocole de séquençage MinION que nous avons pu développer ensemble. Ta motivation et ton professionnalisme font de toi, de mon point de vue, le cœur de l'équipe PhylEvolMol. Je tiens aussi à te remercier pour ton soutien cette fois ci plus personnel tout le long de ma thèse. Ça a toujours été facile de me confier et tu as toujours répondu présente lorsque j'avais besoin de partager un sentiment quel qu'il soit. Un grand merci encore.

Fred, un immense merci pour m'avoir permis de réaliser cette thèse. Pendant ces trois années, j'ai vraiment apprécié la façon dont nous avons pu, ensemble, faire avancer ce projet. Tellement de bons moments ont marqué cette thèse avec notamment : l'Afrique du Sud et la Guyane, des barbecues, de la Belle Cabresse, un décollage de fusée, des papiers plus ou moins acceptés, des congrès, des contrôles de douane (:D), etc… La liste est longue. J'ai vraiment apprécié notre intéraction, je me suis senti à l'aise dès le début et soutenu jusqu'à la dernière minute ! Tu m'as permis d'évoluer dans le monde de la recherche sur plein d'aspects et pour tout ça je te dis un grand merci. Il nous reste à faire pour ce projet et je suis content que l'aventure ne s'arrête pas avec la thèse !

Benoit, merci pour m'avoir donné ma chance en recherche en me proposant, sans même me demander un CV ou une lettre de motivation, mon stage de Master 1. Dès que je t'ai rencontré j'ai su que tu étais quelqu'un de fondamentalement gentil avec qui je pourrais toujours apprendre quelque chose de nouveau quelle qu'en soit la difficulté. Depuis mon M1 jusqu'à maintenant, j'ai toujours apprécié notre façon d'interagir et ta façon de m'encadrer. À la fois très distant, ce qui m'a permis de faire mon propre chemin, et hyper présent lorsque j'en avais besoin. Merci pour tout.

Fabien, je voudrais te remercier pour tous les moments passés ces dernières années. Avec Benoit vous étiez mes co-bureaux et dès le premier jour ça a marché. Depuis mon M1 je me suis régalé de faire de la recherche à tes côtés. Tu m'as proposé un stage de M2 que j'ai sans jamais te l'avoir dit, hésité à accepter, mais heureusement que j'ai fait ce choix là car je serai passé à côté d'une énorme opportunité tant scientifique qu'humaine. Je te remercie pour tout et j'espère qu'on aura l'occasion de retravailler ensemble un jour !

Je tiens à remercier l'ERC pour avoir financé ce projet, le CEBA pour nous avoir permis d'aller en Guyane, participer aux journées CEBA, auxquelles nous avons combinées nos sessions terrain. Merci également au MBB pour toutes les ressources informatiques dont j'ai pu disposer pendant ma thèse. Je tiens particulièrement à remercier Remy Dernat qui à fait tout son possible pour que je puisse profiter de ces ressources malgré un certains nombres de difficultés matérielles rencontrées (disque dur H.S. pendant les vacances de Noël, machine de calcul H.S. pour raison inconnue etc.).

Enfin, je voudrais remercier tous les gens qui m'ont entouré et m'entourent encore chaque jour au labo et qui contribuent à la bonne humeur générale qui définie notre vie à l'ISEM. Les non permanents (Alex, Clémentine, Eliette, Jacob, Manon B., Manon H., Marjo, Marianne, Mathilde, Nathan, Nico, Nicola, Paul, Quentin, Romain, Sergio, Sophie, Yoann le grand et Yoann le petit) mais aussi les permanents. D'ailleurs, un petit message aux permanents de l'équipe PhylEvolMol : surtout ne changez rien. Dès le premier jour où je suis arrivé dans l'équipe, je me suis senti bien, intégré et considéré. Je sais que tous les doctorants n'ont pas toujours la chance d'avoir une telle relation avec les permanents d'autres équipes, qui ne fonctionnent pas mal, mais différemment. Pour ma part, cet environnement m'a permis d'évoluer et de trouver ma place dans un milieu scientifique que j'ai vraiment apprécié et pour ça, je vous dis un grand merci.

Je voudrais tout particulièrement remercier Quentin et Sergio. Nous avons été pendant ces trois ans, le trio des glandouilles au labo. Vous savez mieux que personne comment se sont déroulées ces trois années de thèse au labo pour moi et sachez que sans vous ça n'aurait pas été pareil. Vous êtes tout les deux des personnes exceptionnelles qui m'ont aidé à avoir confiance en ce que je faisais. Vous êtes tous les deux de très bons amis avec qui j'ai passé de très bons moments, sur le terrain, au labo, en réunion, en ville etc… et j'espère qu'on se retrouvera dans le futur pour continuer à faire ce qu'on fait le mieux : rire, râler et refaire le monde.

Je voudrais remercier mon jury de thèse pour avoir accepté de lire et d'évaluer mon travail malgré le travail qui s'accumule en cette année 2020 dont je suis sûr qu'on se souviendra longtemps.

Puisque j'en ai l'occasion ici, je voudrais remercier les personnes qui m'ont donné ma chance et qui ont donc largement participé au fait que j'écrive cette thèse. Je remercie donc Hélène Terzian et Zohra Benfodda pour m'avoir encouragé à continuer dans cette voix. Je remercie aussi Sylvia Campagna pour m'avoir proposé le stage que j'ai réalisé sur la communication chimique chez les mammifères marins. Stage qui m'a certainement permis d'intégrer le Master Darwin de Montpellier. Je remercie Philippe Berta pour m'avoir largement soutenu dans mon choix de faire de la recherche fondamentale malgré des discussions toujours très animées, mais constructives. Je remercie Pierrick Labbé pour m'avoir, sans me connaître, conforté dans mon choix de candidater au Master Darwin. Enfin, je remercie les équipes pédagogiques des masters 1 et 2 de Darwin pour m'avoir permis de suivre cette formation.

Je voudrais aussi remercier les gens du master Darwin et tout particulièrement notre bande d'inséparables avec Aude, Cécile, Quentin et Gopal.
Gopal, je tiens à te remercier pour tous les bons moments qu'on a passé ensemble et pour m'avoir transmis ta passion pour les reptiles, m'avoir appris à chercher et marquer les serpents. Je te remercie, entre autre, pour ce séjour à Chizé où nous avons réussi à attraper et marquer autours de 200 serpents en deux semaines ! C'était vraiment une expérience exceptionnelle pour moi. J'espère qu'on se retrouvera très vite.

Je voudrais remercier deux sports (en particulier), la natation et le volley (en salle comme sur le sable) qui m'ont permis de m'exprimer, de me défouler et de me changer les idées quand il le fallait. Je remercie les différents joueurs avec qui j'ai partagé cette passion et notamment Loïs, Mat' (& Betty), Nico A., Hugo et Noubet. J'espère qu'on se retrouvera vite sur le terrain !
J'en profite pour remercier notre labo pour avoir soutenu le volley à l'ISEM ! Dans ce contexte nous avons partager quelques parties d'enfer avec par exemple : Fabien, Fred, Manu, Manon, Marjo, Max, Sergio, Yoann.

Je voudrais remercier les copains du CEFE : Manon (& Matthieu), Benjamin, Pauline, Sam, Suzanne & Max, Tangi et Bertrand !

Je voudrais remercier les étudiants que j'ai encadré qui ont contribué à l'avancée de ma thèse : Dave Lutgen, Mathilde Barthe, et Sophie Teullet.

Je voudrais remercier les copains de Nîmes : Manon, Thomas, Astrid, Claire, Océane, Ana, Marion, Camille, Tom, Youssef, Kévin, Raphaël, Clément, Julia, Méline, Mathilde, Nathan, Pauline etc et tout particulièrement Paul que j'ai retrouvé en licence et avec qui, entre autres, j'ai révisé mes partiels en rigolant au téléphone souvent la veille au soir et en revoyant les concepts importants.

Un grand merci à Dorian et Julien (et à vos familles respectives). C'est de plus en plus dur de se voir vu qu'on évolue dans des pays différents mais sachez que vous comptez beaucoup pour moi et je souhaite vraiment que notre amitié continue à résister à la distance qui nous sépare. Toujours un immense plaisir de vous revoir et de reprendre la vie comme si nous n'avions jamais été séparés. À très bientôt j'espère.
Un grand merci aussi à Élara. T'as beau être une pièce rapportée, on a quand même passé de très bons moments tous ensemble comme ces parties enflammées de Catane, ces soirées séries ou Koh Lanta ou encore ces heures au téléphone à discuter potins en haut parleur. Merci à toi !

Aude, ma p'tite tête, un grand merci à toi pour être là depuis maintenant plus de 4 ans. Je te remercie du fond du cœur pour l'équilibre que tu apportes dans ma vie. Merci pour toutes ces discussions scientifiques qui parfois vont jusqu'à te saouler au plus au point (et oui, je sais que j'ai beaucoup trop utilisé le mot MitoFinder). Merci pour ta complicité, ta bonne humeur et pour tout ce que tu m'apportes au quotidien sans même t'en rendre compte. Enfin et surtout, merci pour ta joie de vivre qui me réchauffe le cœur chaque jour lorsqu'on se retrouve après nos journées de travail respectives. Merci pour ton dessin (cf part 1 de la thèse) et merci pour tes conseils tout le long de la thèse. Enfin, merci également à tes grands parents et à ta famille pour m'avoir accueilli les bras ouverts dans le calme et la bonne humeur ! ;-)

Maman, Papa, Marie, je tiens à vous remercier pour tout ce que vous m'avez apporté. Ça fait maintenant un peu plus de trois ans que j'ai commencé ma thèse, mais vous ça fait bien plus longtemps que vous me soutenez. C'est difficile à expliquer en quelques mots, mais d'avoir pu évoluer dans une atmosphère si équilibrée, juste, joyeuse, pleine d'amour et d'humour, fait de moi la personne que je suis aujourd'hui. Je vous remercie pour tout ce que vous avez fait pour moi, pendant les trois années de cette thèse, mais surtout pendant les années qui l'ont précédée. Merci aussi à mes grands parents, qui sont à eux seuls la définition de l'amour et la gentillesse, ainsi qu'à toute ma famille pour tout les moments qu'on a pu passer ensemble à la montagne, à la mer, mais aussi à table (et oui, les fameux repas de famille ☺). J'ai de la chance de tous vous avoir et vraiment merci d'être comme vous êtes.

Pour finir, si vous êtes en train de lire ces remerciements et que vous n'êtes pas remercié c'est sans aucun doute une erreur de ma part. Merci à vous !

– INTRODUCTION –

1- The birth of phylogenetics

Phylogenetics - the reconstruction of organisms' evolutionary history and relationships between groups - owes its existence to the revolutionizing contributions of Jean Baptiste Lamarck (1809, **Fig. 1a**), Alfred Russel Wallace (1858) and Charles Darwin (1859, **Fig. 1b**) in the understanding of the evolution of living organisms on Earth. First, Lamarck suggested that species constantly appear by spontaneous generation and gradually evolved from simple to more complex forms by adapting to their environment ("transmutation" theory, 1809). This is the first evolutionary scheme proposed by biologists to explain Earth's biodiversity. Then, the main contribution of Wallace and Darwin to the theory of evolution was to add the notion of variance and random characters apparition in populations, allowing organisms to adapt to various environmental conditions (Wallace & Darwin 1858). The filter of this variance was called "natural selection" and characterized as one of the main drivers of organisms' evolution, being the process responsible for divergence between populations and ultimately the apparition of distinct species (i.e. speciation). With these two different theories, these three world-renowned biologists were the first to suggest that organisms evolved from an ancestor, which evolved itself from another ancestor, and so on. Following this logic, Charles Darwin drew the first phylogenetic tree in the history of biology in his famous 1859 book, "*(On) The origin of species*" (**Fig. 1b**). Even if the notion of heredity was implicitly formulated in the Darwinian theory of evolution, the rules of heredity were understood only years later by Gregor Mendel via his famous experiments on peas (Mendel 1866).

Although taxonomy - the science of naming, defining and classifying species - was already well developed in the XVIIIth century, notably thanks to the huge work of Carl Linnaeus (or Carl von Linné; 1758), both the notion of natural selection on organisms and the notion of inheritance of traits through generations, marked the beginning of a new way of classifying living organisms. Indeed, since the XIXth century, the evolutionary history of species (i.e. their ancestors and their relationships) is taken into account when classifying the diversity of organisms on Earth. This is the birth of the phylogeny, which was formalized for the first time in Ernst Haeckel's recapitulation theory (1866, **Fig. 1c**). The phylogeny can be viewed as the summary of the history of organisms and the representation of their evolutionary relationships. It materializes the historical evolution of organisms, which is essential to understand the evolution of biodiversity on Earth. Indeed, for example, historical context is crucial to explain most evolutionary processes such as convergence (i.e. the apparition of similar characters within two independent phylogenetic lineages).

Phylogenetics quickly played an important role in systematics. To infer the relationships of organisms, several types of characters have been and are still used by biologists. Importantly, to be comparable and usable for phylogenetic inference, a character has to be inherited from a common ancestor of all the species included in the inference and transmitted through speciation events. This is

the notion of homology. Based on this idea, the burgeoning of phylogenetic inferences in the early XIXth century was allowed by the definition of homologous morphological characters. These characters were meticulously defined by specialists in morphology and anatomy and the first inferences based on morphological/phenotypic characteristics allowed, for example, the confirmation or the identification of some of today's major clades. At the end of the XIXth, Willi Hennig's book, *Phylogenetic systematics* (1966), exposed the importance of phylogeny in systematics and the support that phylogenetic inferences provide for studying biodiversity evolution.

During the next decades, phylogenetic inferences were essentially conducted using phenotypic characters. However, progress in molecular biology made in the second half of the XXth century led to the development of the knowledge of the cell machinery including proteins, RNA, and DNA sequences, which were recognized as a new promising way to study organismal evolution (Zuckerkandl & Pauling 1965). Being the support of the genetic information, inherited from generations to generations, scientists rapidly took advantage of these molecules to infer phylogenetic relationships. Firstly, by analysing hemoglobin structures by a combination of electrophoresis and chromatography on paper, Zuckerkandl et al. (1960) found that protein structures were more similar between close-related species. However, this method was not quantitative and thus not applicable for phylogeny reconstruction. The first quantitative methods consisted in comparing the affinity between molecular sequences of different species. Based on their knowledge in immunology, Sarich and Wilson (1967) proposed a quantitative measure of distances based on the intensity of the immune antigen/antibody reaction between pairs of closely related species. Later, a similar strategy was developed to compare DNA sequences based on hybridising DNA strands of two different species and measuring the strength of their interaction ($T_{50}H$ values, Kohne 1970; Sibley & Ahlquist 1980). Following from these different advances, the fundamentals of molecular evolution and modern phylogenetics were already in place at the end of the 1970s (Wilson et al. 1977).

Soon after, the rise of new technologies in DNA amplification and sequencing revolutionized our way to infer phylogenetic relationships. Indeed, the development of Sanger sequencing (Sanger et al. 1977) coupled with the invention of the Polymerase Chain Reaction (Mullis et al. 1987) allowed sequencing homologous portions of DNA for different organisms. As for morphological data, the comparison of these sequences allowed inferring phylogenetic relationships using character-based methods such as maximum parsimony (Farris 1970; Fitch 1971) and maximum likelihood (Felsenstein 1973) considering each homologous site of DNA sequences as an independent character. Hence, the phylogenetic matrices evolved from dozens to hundreds of characters. Additionally, the universality of the characters encountered in DNA sequences (i.e. the four nucleotides Adenosine, Cytosine, Guanine, and Thymine) made possible the comparison of distant organisms much more easily than with morphological data with which it may be difficult to determine homologous characters such as for prokaryotes (Woese 1987). Thanks to the generalization of Sanger sequencing, both the number of genetic markers and the number of samples used in phylogenetic inferences have

increased, which has led to the improvement in our understanding of the evolutionary history of several groups across the Tree of Life. However, for several anciently diverged groups or for groups with high diversity and/or rapid radiation, phylogenetic studies based on a handful of molecular markers resulted in poorly supported phylogenetic relationships due to a limited number of phylogenetically informative characters (*e.g.* Lara et al. 1996). But again with new technological developments, scientists were able to overcome this problem by having access to a large amount of genome scale data. Indeed, with the development of, first, the so-called Next Generation Sequencing (NGS) methods and, more recently, the third generation sequencing technologies, it became possible to sequence the entire genome of an ever increasing number of organisms at an ever decreasing cost. The field of phylogenetic reconstruction has entered into the genomic era (Delsuc et al., 2005; Scornavacca et al. 2020).

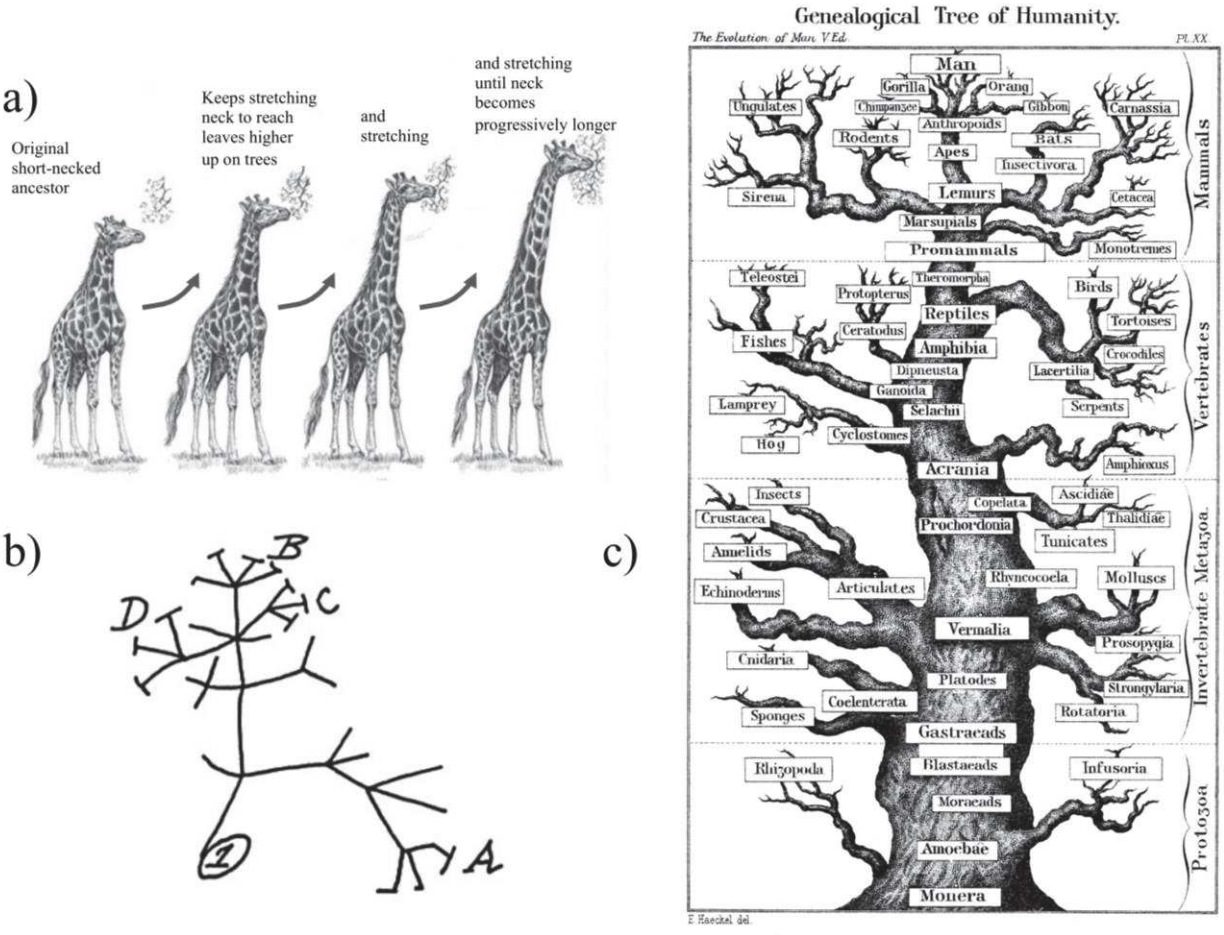


Figure 1 | The theory of the evolution as seen by a) Lamarck, b) Darwin and c) Haeckel.

2- The genomic era

With the impressive improvements of sequencing technologies in the last decades, while the sequencing of the first human genome took 13 years and cost approximately $3 billion (The Human Genome Project, Collins et al. 2003), it is now possible to sequence the genome of non-model organisms in a few days at affordable cost for most ecology and evolution laboratories. In addition, numerous large-scale genome-sequencing efforts such as the Earth BioGenome Project (Lewin et al. 2018), Genome 10K (Koepfli et al. 2015), the Vertebrate Genomes Project (https://vertebrategenomesproject.org/), Bat 1K (Teeling et al. 2018), Bird 10K (Feng et al. 2020), the Arthropods i5K project (i5k consortium 2013) and the DNA Zoo (https://www.dnazoo.org/), are now underway. These international projects/consortiums aim to sequence thousands of genomes from different clades of organisms around the world. The birth of such projects was indeed facilitated by the development of NGS technologies which led to the large decrease in sequencing cost from $10,000 per megabase in the 90's (Collins et al. 2003) to about $0.08 per megabase today (National Human Genome Research Institute, https://www.genome.gov/about-genomics/fact-sheets/DNA-Sequencing-Costs-Data), associated with sequencing yield increase in similar proportions. Additionally, even if more and more genomes are being sequenced around the world, another important aspect of the evolution of the genomic field in the past decades is the evolution of the quality of genome assemblies. This improvement can be associated with the development of the third generation sequencing technologies such as the "Single-Molecule Real-Time" method developed by PacBio (Rhoads & Au 2015) or the sequencing of long DNA molecules through nanopores developed by Oxford Nanopore technologies (Jain et al., 2016). These new long read technologies increased the size of sequence reads from typically 150-250 base pairs in NGS sequencing to 1,000-1,000,000 base pairs, which greatly helped reducing the fragmentation of genome assemblies by permitting the assembly of complex genomic portions exclusively composed of repeat elements for example.

Overall, the improvement of sequencing technologies led to the drastic augmentation of the number of molecular markers available for phylogenetic inferences and marked the beginning of the phylogenomic era. Indeed, the discipline of phylogenomics is a recent field of biology (Eisen 1998), at the intersection between phylogenetic inferences and genomic comparisons. Phylogenomic studies rely on the use of a large amount of information extracted from DNA/protein sequences (*i. e.* genomic data) to better understand organismal evolution (Philippe & Blanchette 2007; Pennisi 2008). Of course, as is usually the case with new technologies, the facilitation of DNA sequencing and extraction of genetic markers bring new objectives and challenges for phylogenetic studies (Scornavacca et al 2020). Given that my PhD project grew in this context, this section presents (i) the main practices in phylogenetic studies to obtain informative genomic markers and their pros and cons, (ii) the challenges associated with the use of such large datasets, and (iii) the importance of the phylogenetic backbone for understanding evolutionary history.

The main approaches currently applied in phylogenomic studies to obtain informative phylogenetic markers consist either in specifically sequencing subparts of whole genomes - using DNA capture methods (Ultra Conserved Elements, Anchored Hybrid Enrichments and RAD-Seq) or RNA-seq sequencing - or in extracting markers of interest from whole genome sequencing data (Zhang et al. 2020). The first strategy, consisting of diminishing the sequencing efforts, has the two advantages of reducing the sequencing cost given its specificity and facilitating subsequent bioinformatic analyses to obtain phylogenetic markers. On the one hand, transcriptomic sequencing allows the sequencing of the portions of the genome that are expressed, including the protein-coding genes. However, this type of sequencing is only possible with well-preserved tissue samples, conserved in RNAlater or flash frozen. This makes RNA-Seq difficult to conduct in practice for large phylogenomic studies. On the other hand, capture methods such as Ultra Conserved Elements (UCEs) sequencing or anchored hybrid enrichment (AHE) use DNA/RNA probes to hybridize and selectively capture target sequences from DNA extractions. This strategy, initially developed for ancient DNA, has proven its efficiency with ethanol-preserved tissues, old DNA extractions, and museum specimens (e.g., Guschanski et al. 2013; Blaimer et al. 2016). However, although the capture methods offer a perfect compromise between sequencing and bioinformatic efforts and the quantity of genetic informative markers, the specificity of the data obtained limits their application to phylogenetic studies. In that context, with the decrease in sequencing costs, the strategy consisting of sequencing the whole genome became a viable alternative. Applicable to most tissues, whole genome shotgun (WGS) sequencing permits to obtain a large amount of informative phylogenetic markers (e.g. Allen et al. 2017, Zhang 2019, Allio et al. 2020a: Annexe 1) but also to access other parts of the genome to further study organismal evolution (e.g. transposable elements, Platt et al. 2016; untranslated regions, Sackton et al 2019). Furthermore, the burgeoning of such genomic datasets led to the recent development of efficient bioinformatic methods to either specifically assemble phylogenomic markers from raw sequencing data (e.g. Schwartz et al. 2015; Allen et al. 2015, 2017; Pouchon et al. 2018; Hughes & Teeling 2019) or to extract orthologous genes after assembling the data (e.g. Zhang et al. 2019, Allio et al. 2020a: Annexe 1). In the light of these methods, whole genome sequencing methods provide a promising approach for phylogenomic studies with evolutionary interests.

In addition to the challenge of generating genome-scale datasets due to the sequencing and bioinformatic efforts required, analyzing such datasets brings new challenges for phylogenetic studies. Indeed, inferring phylogenetic relationships from large amounts of sequencing data is difficult from the computational, statistical and modeling points of view. Here, I focus on the challenge of modelling the evolutionary history of organisms through genomic markers. On the one hand, to be completely accurate, one could expect that phylogenetic inferences with such variable phylogenetic markers should take into account the different evolutionary history of each independent marker (Bryant & Hahn 2020). Indeed, although phylogenetic markers are often used conjointly to infer phylogenetic relationships using concatenation (supermatrix approach), it is now commonly accepted

that the different genes of an organism may have a different evolutionary history. Biological phenomenons such as genetic introgression (Mallet et al. 2016) or incomplete lineage sorting indeed lead to complex gene evolution and incongruent gene trees (Degnan & Rosenberg 2006; Bryant & Hahn 2020; Rannala et al. 2020). Although incongruences in gene trees were initially associated with weak phylogenetic information for each loci, many recent examples confirm the biological causes of this topological variation in recently diverged species (e.g. Fontaine et al 2015, Pease et al. 2016, Rogers et al. 2019). To address this issue, specific methodological approaches taking into account the evolutionary history of each marker independently have been developed (Rannala & Yang 2003; Mirarab et al. 2006; Rannala et al. 2020). The objective of these approaches, known as "coalescent" approaches (Maddison 1997, Rosenberg & Nordborg 2002), is to take advantage of gene tree information instead of suffering from gene tree/species tree discrepancies as in concatenation approaches (Rannala et al. 2020). However, gene tree versus species tree discordance is far from restricted to ILS and the other sources of errors including introgression, gene duplication and homoplasy can exaggerate the importance of ILS by inflating gene tree versus species tree discordance (Gatesy & Springer 2014; Springer & Gatesy 2016). Nevertheless, using both supermatrix and gene tree/species tree approaches often leads to point out interesting nodes where evolutionary history of species is trickier than expected (e.g. in Primates Vanderpool et al. 2020; in Papilionidae Allio et al. 2020a: Annexe 1 ; and in Carnivora Allio et al. 2020b: Part 2.2).

On the other hand, even if using more and more genetic markers led to the diminishing of phylogenetic incongruences thanks to the large amount of information used to infer phylogenetic relationships, in some cases, only adding more sequences was not enough to resolve the inconsistencies between studies (e.g. Dunn et al. 2008, Philippe et al 2009, Schierwater et al 2009, Philippe et al. 2011). However, using appropriate evolutionary models for multi-marker datasets permits to avoid inference artifacts (e.g. Simion et al. 2017). As discussed above, phylogenomic datasets can present heterogeneity in the evolutionary history of their constitutive markers. In fact, such heterogeneity is also observed across sites and lineages due to the variation of nucleotide substitution rates along and between genomes and over time, making it complex to model sequence evolution in phylogenomic inferences (Simion et al. 2020). As pointed out by Simion et al. (2020), current phylogenomic methods are still a long way from implementing a realistic model of genome evolution. Given the number of processes to take into account, even the current best practices in phylogenomics rely on several simplifying hypotheses. However, some interesting advances have been made during the last decades. For example, the first models of the nucleotide frequencies and evolutionary rates considered limited complexity in sequence evolution with either similar exchange rates between the four nucleotides (Jukes & Cantor 1969) or allowing different substitution rates for transition and transversion events (Kimura 1980). Now, the most recently developed and most widely used model implements distinct and independent exchange rates for all substitution types (Generalised time-reversible model, Rodriguez et al. 1990). Often combined with a Gamma distribution (Yang

1994), this model takes into account the variation in substitution rates across sites of multiple sequence alignments. Additionally, to efficiently take into account the variation in evolutionary rates across both sites and loci, an interesting approach resides in partitioned models. In phylogenomic datasets, the most basic partitioning scheme consists in partitioning supermatrix alignments by genes/loci. However, given the fact that the variation in evolutionary rate is also observed across sites, especially in protein-coding sequences, methods have been implemented to define partitioning schemes directly from the sequence properties. Using these methods, sites presenting similar substitution patterns are grouped in the phylogenetic inference procedure (Lanfear 2012, 2014, 2017; Frandsen 2015). Finally, mixture models, such as the CAT model (Lartillot & Philippe 2004) initially developed to deal with evolutionary substitution rates of amino acid sequences, presents an interesting strategy to partitioning sites by creating site categories for each distinct evolutionary scheme observed in the data. This implementation has proven its efficiency to overcome some known biases in phylogenetic reconstruction such as the long-branch attraction artifact, which is caused by multiple substitutions observed at sites presenting rapid evolutionary rates in independent lineages (e.g. Lartillot et al. 2007; Rodríguez-Ezpeleta et al. 2007; Simion et al. 2017).

Overall, the incorporation of the solutions found to the different challenges imposed by phylogenomic datasets led to the improvement of phylogenomic inferences. This achievement was crucial to be able to appropriately compare genomes among species and to understand their evolution. Indeed, the development of the sequencing technologies has also impacted our way to study biodiversity. Genomic scans for adaptive selection signatures, associated with robust phylogenetic background, could allow to associate evolutionary events with molecular signatures, representing a promising strategy to link molecular evolution to life historical traits. For instance, a recent study has pointed out that changes in thermal niches drove penguin diversification and were accompanied by adaptive signatures in genes that govern thermoregulation and oxygen metabolism (Vianna et al. 2020). Similarly, following my master project, we were able to infer a robust phylogeny for swallowtail butterflies using genome-scale data for 61 species (45 Papilionidae and 16 outgroup species), representing all described genera (Allio et al. 2020a: Annexe 1). Then, using an additional dataset composed of Sanger sequencing data for 408 species of swallowtail butterflies (~71% of Papilionidae diversity) and 247 species of birthworts (~49% of the Aristolochiaceae diversity), we showed that the antagonist interaction between Papilionidae and birthworts - their highly toxic host plants - began 55 millions years ago in Beringia. Despite their relatively high level of host plant conservation, likely due to the specificity of the mechanisms used by butterflies to colonise their host, the evolutionary history of Papilionidae was punctuated by several host plant shifts. By conducting both diversification and genome scan analyses for positive selection, we showed that changes in host plants were associated with boosts in swallowtail butterfly diversification and with more adaptive genomic signatures than non host plant shift lineages (Allio et al. 2020c: Annexe 2).

To conclude, our capacity to generate highly accurate and robust phylogenetic trees is crucial to understand the role of different natural processes in shaping biodiversity. After having briefly presented some examples for which the phylogenetic backbones were necessary to understand organismal evolution, in the next section I focus on a fascinating phenomenon understandable only in the light of a robust phylogenetic framework: evolutionary convergence.

3- Evolutionary convergence: The case of ant-eating mammals

The phenomenon of evolutionary convergence is a fascinating process in which distantly related species independently acquire similar characteristics in response to similar selection pressures. As introduced by Wallace and Darwin (1958), natural selection drives the evolution of organisms by playing the role of the filter to the variance observed *in natura*. For a given environment, the individuals passing through this filter are the most adapted ones. Given the numerous possibilities to efficiently adapt to an environment, and considering the large diversity of forms on Earth, one could assume that only few cases of convergence are expected. By adding the notion of contingency, the fact that organismal evolution depends on the evolutionary history of their ancestors, Stephen Jay Gould concluded his famous book *Wonderful life* by writing: "Replay the tape a million times…and I doubt that anything like *Homo sapiens* would ever evolve again.". Gould's arguments were inspired by his observations on the animal fossil record of the Burgess Shale (Cambrian). He argued that the diversity observed in fossils was much higher than the diversity observed in living organisms and that the similarity observed between fossil and extant species was explained by their close relationships. Hence, if different species had survived in the past, today's biodiversity could be extremely different having adapted to the environment by a different path due to its different ancestral characteristics. This is the notion of historical contingency (Gould 2002). However, bothered by the idea that the evolution of the current diversity, including *Homo sapiens*, was the fruit of chance, one of Gould's students, Simon Conway Morris, drew diametrically opposed conclusions. Based on the same fossil record, by observing similar characteristics between extant and extinct species, Conway Morris has become the leading proponent of the view that convergent evolution is the dominant story behind life's diversity as exposed in his book "*The Crucible of Creation*" (Conway Morris 1998). Of course, Conway Morris' conclusion was: "Rerun the tape of life as often as you like, and the end result will be much the same". Years after the debate between Gould and Conway Morris, we know that both historical contingency and evolutionary convergence have impacted the evolution of the current biodiversity (Blount et al. 2018) and the major question relies on evaluating the relative impact of these two evolutionary processes.

The burgeoning of evolutionary convergence examples can be associated with the utilization of DNA markers in phylogenetic inferences (Madsen et al. 2001, Murphy et al. 2001, Delsuc et al. 2002). Indeed, phenotypical-based phylogenetic matrices are by definition sensible to convergent

evolution. Hence, adding complementary DNA information led to the discovery of numerous convergence cases. Species found as sister-species in phenotypical-based phylogenetic inferences were instead revealed as very distant species presenting similar characteristics due to similar selection pressures (see McGhee 2011 or Losos 2017 for many examples). One of the most famous recent examples of convergent evolution concerns the convergence towards echolocation in mammals. Research on the candidate gene coding for the prestin protein, which is involved in hearing, pointed out a number of convergent adaptive substitutions shared by echolocating bats and toothed whales (Li et al. 2010a; Liu et al. 2010a,b). Given the likely implication of hearing genes in echolocation, Davies et al. (2012) compared the evolution of additional candidate genes between echolocating and non-echolocating bats and found additional convergent substitutions, suggesting that adaptive convergent molecular evolution might be more widespread than generally accepted. In that context, the first study looking at convergent adaptations in echolocating mammals at the genomic scale led to the discovery of 200 potentially convergent genes between echolocating bats and dolphins (Parker et al. 2013). However, this study suffered from the lack of an appropriate null model for detecting adaptive convergent substitutions (Thomas & Hahn 2015; Zou & Zhang 2015). Indeed, although natural selection may lead to evolutionary convergence due to similar selection pressures, other biological mechanisms may also lead to molecular convergences without the need to involve selection pressure (Losos 2011). For instance, developmental constraints may limit evolutionary possibilities and lead by coincidence to similar phenotypes (Maynard Smith et al. 1985). Similarly, since only four nucleotides and twenty one amino acids can be found at a specific site, it is likely that similar substitutions can be observed by chance between distant lineages. Additionally, mutational biases, changes in recombination rates or biased gene conversion (bGC) may inflate or diminish the chance to detect convergent substitutions (Lartillot 2013; Rey et a. 2019). These two types of molecular convergence may be distinguished as foreground convergent substitutions - associated with the convergent phenotype - or background convergent substitutions - independent of the convergent phenotype. Hence, the search for molecular convergence ideally consists in searching for molecular convergence between lineages by controlling for background convergent substitutions. In that context, by re-analysing the dataset of Parker et al. (2013), two independent studies showed that most convergent substitutions are common in echolocating and non-echolocating mammal genomes, and thus not necessarily linked to convergent phenotypes (Thomas & Hahn 2015; Zou & Zhang 2015). These two studies argued that using an appropriate null model to detect adaptative convergent substitutions is mandatory. In this light, recent practices consist in using the evolutionary history of close-related non-convergent species and controlling for random substitution probability to point out probable foreground convergent evolution (e.g. Hu et al. 2017). Alternatively, the detection of convergent evolutionary rates can permit to point out possible convergent adaptation to similar pressure (oftenly detects pseudogenization; e.g. Partha et al. 2017; Kowalczyk et al. 2019).

The ant- and termite-eating mammals are among the most famous examples of morphological convergence (Redford 1987, McGhee 2011). This particular lifestyle evolved in five distinct lineages of placental mammals: the aardvark, the aardwolf, the anteaters, the giant armadillo, and the pangolins. The high specialization towards the consumption of ants and termites led to incredible morphological convergences (**Fig 2**). Although ant and termite nests are composed of thousands of individuals, their social organisation makes them difficult to exploit as a resource. Myrmecophagous mammals have evolved long claws and strong forelimbs allowing them to tear nests apart and some of them present very long, sticky tongues facilitating the capture of ants and termites in very large numbers. Additionally, both ants and termites present chitinous skeletons, making them difficult to digest. In response to this constraint, anteaters, pangolins and aardvark evolved hypertrophied salivary glands and strong muscular pyloric regions in their stomachs to assist in digesting ants (Lecointre and Le Guyader 2006). As a consequence of this behaviour, the anteaters and pangolins have totally lost their teeth, and the aardvark presents enamel-less teeth (Ferreira-Cardoso et al. 2019). Hence, based on their morphological similarities, all but one (the aardwolf) myrmecophagous mammals were initially classified together as Edentata (Vicq-d'Azyr 1742, Cuvier 1798). Then, subsequent morphological observations and the attribution of some characters to convergent evolution toward myrmecophagy led to several rearrangements in ant-eating mammal classification (Huxley 1872, Weber 1904, McKenna 1975). However, it was only in light of molecular markers that the polyphyly of the Edentata was discovered and fully exposed (Delsuc et al. 2001, Madsen et al. 2001, Murphy et al. 2001, Delsuc et al. 2002). Indeed, morphological adaptations associated with the shift to the myrmecophagous diet are so preponderant that even using 4541 morphological characters, a recent cladistic study misleadingly inferred the monophyly of Edentata (**Fig. 3**; O'Leary et al. 2013; Springer et al. 2013). Their extreme level of convergence made the myrmecophagous mammals an excellent case to study the processes underlying convergent evolution in a genomic context. In fact, the molecular bases of such an extraordinary level of convergence observed in ant-eating mammals remains unknown.

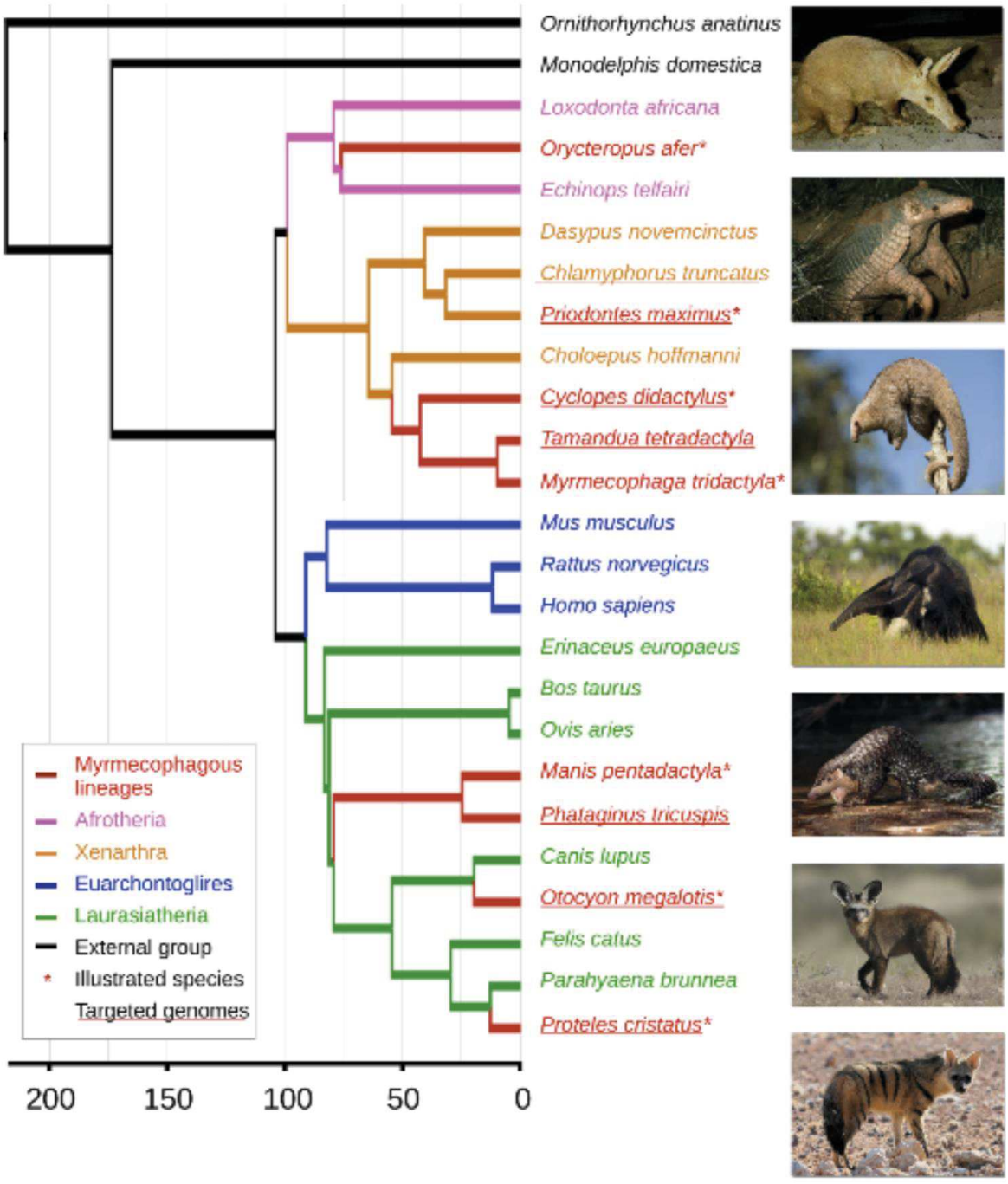


Figure 2 | Simplified phylogeny representing five independent origins of myrmecophagy in mammals (red branches).

4- Dissertation's main goals

To further study the evolutionary processes associated with dietary specialization towards ant and termite consumption, my PhD thesis forms part of an ERC-funded project named "ConvergeAnt". This project focuses on the convergence in myrmecophagous mammals through an integrative approach investigating the question from the (i) morphological (ii) genomic and (iii) microbiome points of view. Involving aspects of phylogenomics and comparative genomics, my PhD project naturally forms part of the second work package of this ERC project.

In the first chapter, I present analyses conducted to characterize the precise diet of myrmecophagous mammals. To do so, both fecal samples collected on native home range of myrmecophagous mammals, and ants and termites encountered in these areas were sequenced using shallow coverage shotgun Illumina sequencing. The latter allowed us to construct a mitogenomic database to be able to identify myrmecophagous preys based on mitochondrial sequences extracted from their feces. Considering the huge number of samples a user-friendly pipeline, called MitoFinder, was developed to automatically assemble, extract, and annotate mitochondrial sequences from high throughput sequencing data. Both the strategy used to identify myrmecophagous mammals' preys and the MitoFinder pipeline are presented in sections 1 and 2, respectively.

At the beginning of the thesis, it was planned to generate high quality genomes with a hybrid assembly strategy (combining short reads and long reads). As we only had particular tissues (mostly issued from roadkill samples), we had to develop a specific extraction protocol for MinION long read sequencing. In the second chapter, I explain how we developed this protocol and the associated bioinformatic workflow to assemble and annotate nine mammalian genomes comprising mainly strictly myrmecophagous species. Then, in the second section of this chapter, I present an example of application of genomic data from roadkill samples for species delineation. In this example, we focused on carnivores, that include two myrmecophagous species. We carried out species delineation analyses at the genomic scale for the two species of myrmecophagous carnivores (*Proteles cristatus* and *Otocyon megalotis*) as each of them has two isolated populations described as subspecies in Eastern and Southern Africa.

Finally, in the third chapter, I present gene expression analyses based on transcriptomes obtained from salivary glands of 24 mammals, including ant-eating mammals. Myrmecophagous mammals often present hypertrophied salivary glands, suggesting a possible role in the adaptation to myrmecophagy. Gene expression analyses were conducted with a particular focus on the chitinase gene family involved in the degradation of insect chitin since convergently evolved anteaters and pangolins were previously shown to present distinct chitinase gene repertories. After reconstructing chitinase gene family evolution, we also compared the expression of the different paralogs in several digestive and non-digestive organs between the lesser anteater and the Malayan pangolin to gain further insight into the molecular mechaisms underlying their convergent dietary adaptation.

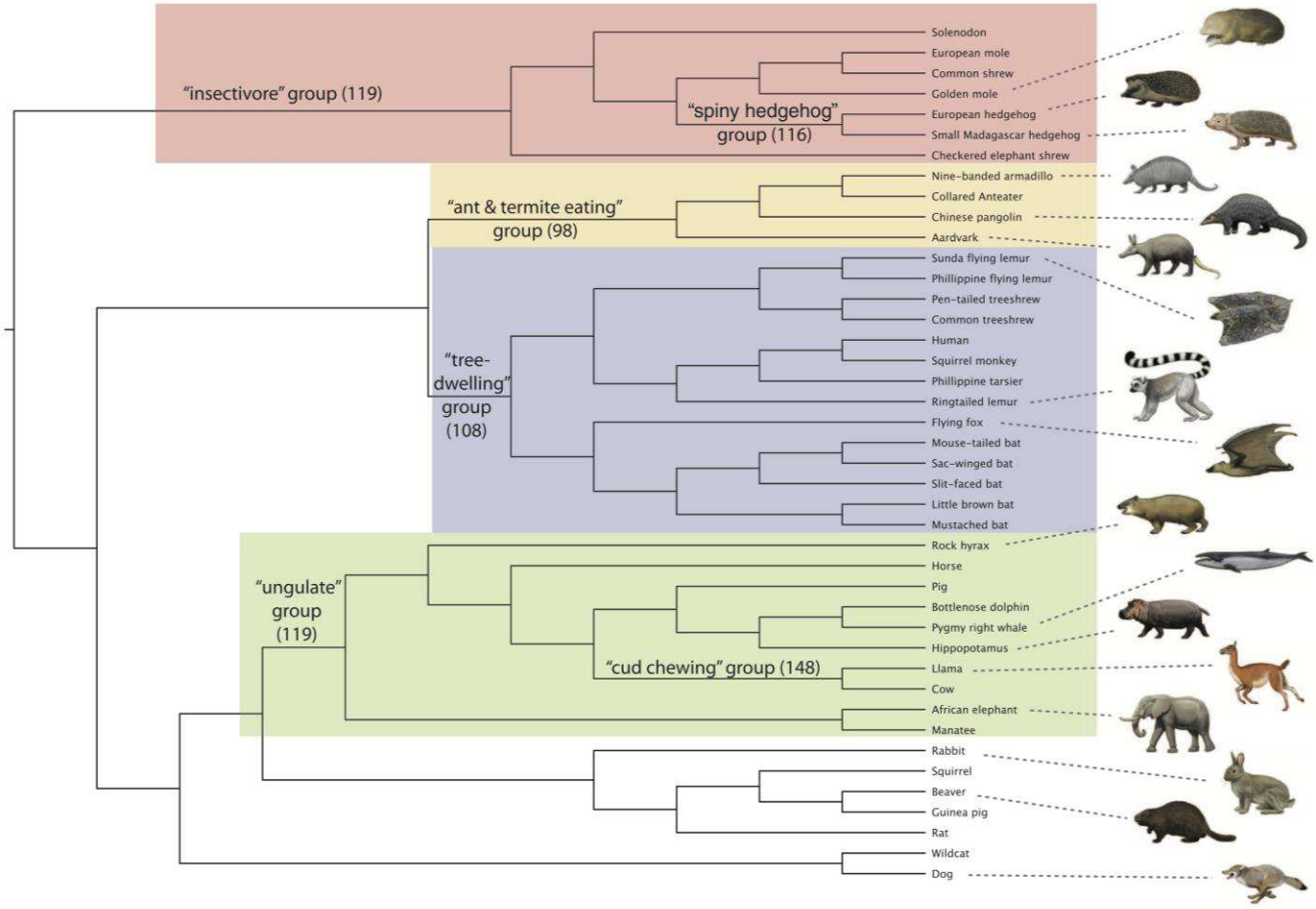


Figure 3 | Figure and caption from Springer et al. 2013.
Polyphyletic ecomorphology groups. Polyphyletic ecomorphology groups from figure S2 (phenomic tree) of O'Leary et al. (2013). Fossils and nonplacentals were pruned to highlight relationships among living placentals. Numbers next to groups indicate the number of "apomorphies" with Deltran optimization (Swofford 2001). [Paintings by Carl Buell]

References

⇧Back to summary⇧

Allen JM, Boyd B, Nguyen N, Vachaspati P, Warnow T, Huang DI, Grady PGS, Bell KC, Cronk QCB, Mugisha L, Pittendrigh BR, Soledad Leonardi M, Reed DL, Johnson KP. 2017. Phylogenomics from whole genome sequences using aTRAM. *Syst Biol* **66**:syw105. doi:10.1093/sysbio/syw105

Allen JM, Huang DI, Cronk QC, Johnson KP. 2015. aTRAM - automated target restricted assembly method: a fast method for assembling loci across divergent taxa from next-generation sequencing data. *BMC Bioinformatics* **16**:98. doi:10.1186/s12859-015-0515-2

Allio R, Nabholz B, Wanke S, Chomicki G, Pérez-Escobar OA, Cotton AM, Clamens A-L, Kergoat GJ, Sperling FAH, Condamine FL. 2020a. Genome-wide macroevolutionary signatures of key innovations in butterflies colonizing new host plants. *bioRxiv* 2020.07.08.193086. doi:10.1101/2020.07.08.193086

Allio R, Scornavacca C, Nabholz B, Clamens A-L, Sperling FA, Condamine FL. 2020b. Whole genome shotgun phylogenomics resolves the pattern and timing of swallowtail butterfly evolution. *Syst Biol* **69**:38–60. doi:10.1093/sysbio/syz030

Allio R, Tilak M-K, Scornavacca C, Avenant NL, Corre E, Nabholz B, Delsuc F. 2020c. High-quality carnivore genomes from roadkill samples enable species delimitation in aardwolf and bat-eared fox. *bioRxiv* 2020.09.15.297622. doi:10.1101/2020.09.15.297622

Blaimer BB, LaPolla JS, Branstetter MG, Lloyd MW, Brady SG. 2016. Phylogenomics, biogeography and diversification of obligate mealybug-tending ants in the genus Acropyga. *Mol Phylogenet Evol* **102**:20–29. doi:10.1016/J.YMPEV.2016.05.030

Blount ZD, Lenski RE, Losos JB. 2018. Contingency and determinism in evolution: Replaying life's tape. *Science (80-)* **362**. doi:10.1126/SCIENCE.AAM5979

Bryant D, Hahn MW. 2020. The concatenation question. *Phylogenetics Genomic Era* 3–4.

Collins FS, Morgan M, Patrinos A. 2003. The human genome project: Lessons from large-scale biology. *Science (80-)* **300**:286–290. doi:10.1126/SCIENCE.1084564

Cuvier JLNF. 1798. Tableau élémentaire de l'histoire naturelle des animaux. Baudouin.

Darwin C. n.d. No TitleThe Origin of Species.

Darwin C, Wallace AR. 1858. On the tendency of species to form varieties; and on the perpetuation of varieties and species by natural means of selection. *J Proc Linn Soc London Zool* **3**:45–62.

Davies KTJ, Cotton JA, Kirwan JD, Teeling EC, Rossiter SJ. 2012. Parallel signatures of sequence evolution among hearing genes in echolocating mammals: an emerging model of genetic convergence. *Heredity (Edinb)* **108**:480–489. doi:10.1038/hdy.2011.119

Degnan JH, Rosenberg NA. 2006. Discordance of species trees with their most likely gene trees. *PLoS Genet* **2**:e68. doi:10.1371/journal.pgen.0020068

Delsuc F, Brinkmann H, Philippe H. 2005. Phylogenomics and the reconstruction of the tree of life. *Nat Rev Genet* **6**:361–375. doi:10.1038/nrg1603

Delsuc F, Ctzeflis FM, Stanhope MJ, Douzery EJP. 2001. The evolution of armadillos, anteaters and sloths depicted by nuclear and mitochondrial phylogenies: implications for the status of the enigmatic fossil *Eurotamandua*. *Proc R Soc London Ser B Biol Sci* **268**:1605–1615. doi:10.1098/rspb.2001.1702

Delsuc F, Scally M, Madsen O, Stanhope MJ, de Jong WW, Catzeflis FM, Springer MS, Douzery EJP. 2002. Molecular phylogeny of living xenarthrans and the impact of character and taxon sampling on the placental tree rooting. *Mol Biol Evol* **19**:1656–1671. doi:10.1093/oxfordjournals.molbev.a003989

Dunn CW, Hejnol A, Matus DQ, Pang K, Browne WE, Smith SA, Seaver E, Rouse GW, Obst M,

Edgecombe GD, Sørensen M V., Haddock SHD, Schmidt-Rhaesa A, Okusu A, Kristensen RM, Wheeler WC, Martindale MQ, Giribet G. 2008. Broad phylogenomic sampling improves resolution of the animal tree of life. *Nature* **452**:745–749. doi:10.1038/nature06614

Eisen JA. 1998. Phylogenomics: Improving functional predictions for uncharacterized genes by evolutionary analysis. *Genome Res* **8**:163–167. doi:10.1101/GR.8.3.163

Farris JS. 1970. Methods for computing Wagner trees. *Syst Biol* **19**:83–92. doi:10.1093/sysbio/19.1.83

Felsenstein J. 1973. Maximum likelihood and minimum-steps methods for estimating evolutionary trees from data on discrete characters. *Syst Biol* **22**:240–249. doi:10.1093/sysbio/22.3.240

Feng S, Stiller J, Deng Y, Armstrong J, Fang Q, Reeve AH, Xie D, Chen G, Guo C, Faircloth BC, Petersen B, Wang Z, Zhou Q, Diekhans M, Chen W, Andreu-Sánchez S, Margaryan A, Howard JT, Parent C, Pacheco G, Sinding M-HS, Puetz L, Cavill E, Ribeiro ÂM, Eckhart L, Fjeldså J, Hosner PA, Brumfield RT, Christidis L, Bertelsen MF, Sicheritz-Ponten T, Tietze DT, Robertson BC, Song G, Borgia G, Claramunt S, Lovette IJ, Cowen SJ, Njoroge P, Dumbacher JP, Ryder OA, Fuchs J, Bunce M, Burt DW, Cracraft J, Meng G, Hackett SJ, Ryan PG, Jønsson KA, Jamieson IG, da Fonseca RR, Braun EL, Houde P, Mirarab S, Suh A, Hansson B, Ponnikas S, Sigeman H, Stervander M, Frandsen PB, van der Zwan H, van der Sluis R, Visser C, Balakrishnan CN, Clark AG, Fitzpatrick JW, Bowman R, Chen N, Cloutier A, Sackton TB, Edwards S V., Foote DJ, Shakya SB, Sheldon FH, Vignal A, Soares AER, Shapiro B, González-Solís J, Ferrer-Obiol J, Rozas J, Riutort M, Tigano A, Friesen V, Dalén L, Urrutia AO, Székely T, Liu Y, Campana MG, Corvelo A, Fleischer RC, Rutherford KM, Gemmell NJ, Dussex N, Mouritsen H, Thiele N, Delmore K, Liedvogel M, Franke A, Hoeppner MP, Krone O, Fudickar AM, Milá B, Ketterson ED, Fidler AE, Friis G, Parody-Merino ÁM, Battley PF, Cox MP, Lima NCB, Prosdocimi F, Parchman TL, Schlinger BA, Loiselle BA, Blake JG, Lim HC, Day LB, Fuxjager MJ, Baldwin MW, Braun MJ, Wirthlin M, Dikow RB, Ryder TB, Camenisch G, Keller LF, DaCosta JM, Hauber ME, Louder MIM, Witt CC, McGuire JA, Mudge J, Megna LC, Carling MD, Wang B, Taylor SA, Del-Rio G, Aleixo A, Vasconcelos ATR, Mello C V., Weir JT, Haussler D, Li Q, Yang H, Wang J, Lei F, Rahbek C, Gilbert MTP, Graves GR, Jarvis ED, Paten B, Zhang G. 2020. Dense sampling of bird diversity increases power of comparative genomics. *Nature* **587**:252–257. doi:10.1038/s41586-020-2873-9

Ferreira-Cardoso S, Delsuc F, Hautier L. 2019. Evolutionary tinkering of the mandibular canal linked to convergent regression of teeth in placental mammals. *Curr Biol* **29**:468–475.e3. doi:10.1016/J.CUB.2018.12.023

Fitch WM. 1970. Distinguishing homologous from analogous proteins. *Syst Zool* **19**:99. doi:10.2307/2412448

Fontaine MC, Pease JB, Steele A, Waterhouse RM, Neafsey DE, Sharakhov I V., Jiang X, Hall AB, Catteruccia F, Kakani E, Mitchell SN, Wu Y-C, 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 (80-)* **347**. doi:10.1126/SCIENCE.1258524

Frandsen PB, Calcott B, Mayer C, Lanfear R. 2015. Automatic selection of partitioning schemes for phylogenetic analyses using iterative k-means clustering of site rates. *BMC Evol Biol* **15**:13. doi:10.1186/s12862-015-0283-7

Gatesy J, Springer MS. 2014. Phylogenetic analysis at deep timescales: Unreliable gene trees, bypassed hidden support, and the coalescence/concatalescence conundrum. *Mol Phylogenet Evol* **80**:231–266. doi:10.1016/J.YMPEV.2014.08.013

Gould SJ. 2002. The structure of evolutionary theory. Cambridge, MA: Belknap Press of Harvard Univ. Press.

Gould SJ. 1990. Wonderfull life: the Burgess Shale and the nature of history. WW Norton & Company.

Guschanski K, Krause J, Sawyer S, Valente LM, Bailey S, Finstermeier K, Sabin R, Gilissen E, Sonet G, Nagy ZT, Lenglet G, Mayer F, Savolainen V. 2013. Next-generation museomics disentangles one of the largest primate radiations. *Syst Biol* **62**:539–554. doi:10.1093/sysbio/syt018

Haeckel E. 1866. Generelle Morphologie der Organismen. Allgemeine Grundzüge der organischen Formen-Wissenschaft, mechanisch begründet durch die von C. Darwin reformirte Descendenz-Theorie, etc.

Hennig W. 1966. Phylogenetic systematics. University of Illinois Press.

Hu Y, Wu Q, Ma S, Ma T, Shan L, Wang X, Nie Y, Ning Z, Yan L, Xiu Y, Wei F. 2017. Comparative genomics reveals convergent evolution between the bamboo-eating giant and red pandas. *Proc Natl Acad Sci U S A* **114**:1081–1086. doi:10.1073/pnas.1613870114

Hughes GM, Teeling EC. 2019. AGILE: an assembled genome mining pipeline. *Bioinformatics* **35**:1252–1254. doi:10.1093/bioinformatics/bty781

Huxley TH. 1880. A manual of the anatomy of vertebrated animals. D. Appleton.

i5K Consortium. 2013. The i5K initiative: Advancing arthropod genomics for knowledge, human health, agriculture, and the environment. *J Hered* **104**:595–600. doi:10.1093/jhered/est050

Jain M, Olsen HE, Paten B, Akeson M. 2016. The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. *Genome Biol* **17**:239. doi:10.1186/s13059-016-1103-0

Jukes TH, Cantor CR. 1969. Evolution of protein molecules. *Mamm protein Metab* **3**:21–132.

Kimura M. 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J Mol Evol* **16**:111–120.

Koepfli K-P, Paten B, O'Brien SJ, O'Brien SJ. 2015. The Genome 10K Project: A way forward. *Annu Rev Anim Biosci* **3**:57–111. doi:10.1146/annurev-animal-090414-014900

Kohne DE. 1970. Evolution of higher-organism DNA. *Q Rev Biophys* **3**:327–375. doi:10.1017/S0033583500004765

Kowalczyk A, Partha R, Clark NL, Chikina M. 2020. Pan-mammalian analysis of molecular constraints underlying extended lifespan. *Elife* **9**. doi:10.7554/eLife.51089

Lamarck JBD. 1809. No TitlePhilosophie Zoologique. Paris, France.

Lanfear R, Calcott B, Ho SYW, Guindon S. 2012. PartitionFinder: Combined selection of partitioning schemes and substitution models for phylogenetic analyses. *Mol Biol Evol* **29**:1695–1701. doi:10.1093/molbev/mss020

Lanfear R, Calcott B, Kainer D, Mayer C, Stamatakis A. 2014. Selecting optimal partitioning schemes for phylogenomic datasets. *BMC Evol Biol* **14**:82. doi:10.1186/1471-2148-14-82

Lanfear R, Frandsen PB, Wright AM, Senfeld T, Calcott B. 2016. PartitionFinder 2: New methods for selecting partitioned models of evolution for molecular and morphological phylogenetic analyses. *Mol Biol Evol* **34**:772–773. doi:10.1093/molbev/msw260

Lara MC, Patton JL, da Silva MNF. 1996. The simultaneous diversification of south american echimyid rodents (Hystricognathi) based on complete cytochrome b sequences. *Mol Phylogenet Evol* **5**:403–413. doi:10.1006/MPEV.1996.0035

Lartillot N, Brinkmann H, Philippe H. 2007. Suppression of long-branch attraction artefacts in the animal phylogeny using a site-heterogeneous model. *BMC Evol Biol* **7**:S4. doi:10.1186/1471-2148-7-S1-S4

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. doi:10.1093/molbev/msh112

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. doi:10.1093/sysbio/syt022

Lecointre G, Le Guyader H. 2006. The tree of life: a phylogenetic classification. Harvard University Press.

Lewin HA, Robinson GE, Kress WJ, Baker WJ, Coddington J, Crandall KA, Durbin R, Edwards S V, Forest F, Gilbert MTP, Goldstein MM, Grigoriev I V, Hackett KJ, Haussler D, Jarvis ED, Johnson WE, Patrinos A, Richards S, Castilla-Rubio JC, van Sluys M-A, Soltis PS, Xu X, Yang H, Zhang G. 2018. Earth BioGenome Project: Sequencing life for the future of life. *Proc Natl Acad Sci U S A* **115**:4325–4333. doi:10.1073/pnas.1720115115

Li Y, Liu Z, Shi P, Zhang J. 2010. The hearing gene Prestin unites echolocating bats and whales. *Curr Biol* **20**:R55–R56. doi:10.1016/J.CUB.2009.11.042

Linné C. 1758. Systema naturaeSystema Naturae.

Liu Y, Cotton JA, Shen B, Han X, Rossiter SJ, Zhang S. 2010a. Convergent sequence evolution between echolocating bats and dolphins. *Curr Biol* **20**:R53–R54. doi:10.1016/J.CUB.2009.11.058

Liu Y, Rossiter SJ, Han X, Cotton JA, Zhang S. 2010b. Cetaceans on a molecular fast track to ultrasonic hearing. *Curr Biol* **20**:1834–1839. doi:10.1016/J.CUB.2010.09.008

Losos JB. 2017. Improbable destinies: Fate, chance, and the future of evolution. Penguin.

Losos JB. 2011. Convergence, adaptation, and constraint. *Evolution (N Y)* **65**:1827–1840. doi:10.1111/j.1558-5646.2011.01289.x

Maddison WP. 1997. Gene trees in species trees. *Syst Biol* **46**:523–536. doi:10.1093/sysbio/46.3.523

Madsen O, Scally M, Douady CJ, Kao DJ, DeBry RW, Adkins R, Amrine HM, Stanhope MJ, de Jong WW, Springer MS. 2001. Parallel adaptive radiations in two major clades of placental mammals. *Nature* **409**:610–614. doi:10.1038/35054544

Mallet J, Besansky N, Hahn MW. 2016. How reticulated are species? *BioEssays* **38**:140–149. doi:10.1002/bies.201500149

Maynard Smith J, Burian R, Kauffman S, Alberch P, Campbell J, Goodwin B, Lande R, Raup D, Wolpert L. 1985. Developmental constraints and evolution: A perspective from the mountain lake conference on development and evolution. *Q Rev Biol* **60**:265–287. doi:10.1086/414425

McGhee GR. 2011. Convergent evolution: limited forms most beautiful. MIT Press.

McKenna MC. 1975. Toward a phylogenetic classification of the MammaliaPhylogeny of the Primates. Boston, MA: Springer US. pp. 21–46. doi:10.1007/978-1-4684-2166-8_2

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

Morris SC. 1998. The crucible of creation: the Burgess Shale and the rise of animals. Peterson's.

Mullis K, Faloona F, Scharf S, Saiki R, Horn G, Erlich H. 1986. Specific enzymatic amplification of DNA in vitro: The polymerase chain reaction. *Cold Spring Harb Symp Quant Biol* **51**:263–273. doi:10.1101/SQB.1986.051.01.032

Murphy WJ, Eizirik E, O'Brien SJ, Madsen O, Scally M, Douady CJ, Teeling E, Ryder OA, Stanhope MJ, Jong WW de, Springer MS. 2001. Resolution of the early placental mammal radiation using bayesian phylogenetics. *Science (80-)* **294**:2348–2351. doi:10.1126/SCIENCE.1067179

O'Leary MA, Bloch JI, Flynn JJ, Gaudin TJ, Giallombardo A, Giannini NP, Goldberg SL, Kraatz BP, Luo Z-X, Meng J, Ni X, Novacek MJ, Perini FA, Randall ZS, Rougier GW, Sargis EJ, Silcox MT, Simmons NB, Spaulding M, Velazco PM, Weksler M, Wible JR, Cirranello AL. 2013. The placental mammal ancestor and the post–K-Pg radiation of placentals. *Science (80-)* **339**:662–667. doi:10.1126/SCIENCE.1229237

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. doi:10.1038/nature12511

Partha R, Chauhan BK, Ferreira Z, Robinson JD, Lathrop K, Nischal KK, Chikina M, Clark NL. 2017. Subterranean mammals show convergent regression in ocular genes and enhancers, along with adaptation to tunneling. *Elife* **6**:e25884. doi:10.7554/eLife.25884

Pease JB, Haak DC, Hahn MW, Moyle LC. 2016. Phylogenomics reveals three sources of adaptive

variation during a rapid radiation. *PLOS Biol* **14**:e1002379. doi:10.1371/journal.pbio.1002379

Pennisi E. 2008. Building the tree of life, genome by genome.

Philippe H, Blanchette M. 2007. Overview of the first phylogenomics conference. *BMC Evol Biol* **7**:S1. doi:10.1186/1471-2148-7-S1-S1

Philippe H, Brinkmann H, Lavrov D V., Littlewood DTJ, Manuel M, Wörheide G, Baurain D. 2011. Resolving difficult phylogenetic questions: Why more sequences are not enough. *PLoS Biol* **9**:e1000602. doi:10.1371/journal.pbio.1000602

Philippe H, Derelle R, Lopez P, Pick K, Borchiellini C, Boury-Esnault N, Vacelet J, Renard E, Houliston E, Quéinnec E, Da Silva C, Wincker P, Le Guyader H, Leys S, Jackson DJ, Schreiber F, Erpenbeck D, Morgenstern B, Wörheide G, Manuel M. 2009. Phylogenomics revives traditional views on deep animal relationships. *Curr Biol* **19**:706–712. doi:10.1016/J.CUB.2009.02.052

Platt RN, Blanco-Berdugo L, Ray DA. 2016. Accurate transposable element annotation is vital when analyzing new genome assemblies. *Genome Biol Evol* **8**:403–410. doi:10.1093/gbe/evw009

Pouchon C, Fernández A, Nassar JM, Boyer F, Aubert S, Lavergne S, Mavárez J. 2018. Phylogenomic analysis of the explosive adaptive radiation of the *Espeletia* complex (Asteraceae) in the tropical andes. *Syst Biol* **67**:1041–1060. doi:10.1093/sysbio/syy022

Rannala B, Edwards S, Leaché A. 2020. The multi-species coalescent model and species tree inference. *Phylogenetics Genomic Era* 3–3.

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

Redford KH. 1987. Ants and termites as foodCurrent Mammalogy. Boston, MA: Springer US. pp. 349–399. doi:10.1007/978-1-4757-9909-5_9

Rey C, Lanore V, Veber P, Guéguen L, Lartillot N, Sémon M, Boussau B. 2019. Detecting adaptive convergent amino acid evolution. *Philos Trans R Soc B Biol Sci* **374**:20180234. doi:10.1098/rstb.2018.0234

Rhoads A, Au KF. 2015. PacBio sequencing and its applications. *Genomics Proteomics Bioinformatics* **13**:278–289. doi:10.1016/J.GPB.2015.08.002

Rodríguez-Ezpeleta N, Brinkmann H, Roure B, Lartillot N, Lang BF, Philippe H. 2007. Detecting and overcoming systematic errors in genome-scale phylogenies. *Syst Biol* **56**:389–399. doi:10.1080/10635150701397643

Rodríguez F, Oliver JL, Marín A, Medina JR. 1990. The general stochastic model of nucleotide substitution. *J Theor Biol* **142**:485–501. doi:10.1016/S0022-5193(05)80104-3

Rogers J, Raveendran M, Harris RA, Mailund T, Leppälä K, Athanasiadis G, Schierup MH, Cheng J, Munch K, Walker JA, Konkel MK, Jordan V, Steely CJ, Beckstrom TO, Bergey C, Burrell A, Schrempf D, Noll A, Kothe M, Kopp GH, Liu Y, Murali S, Billis K, Martin FJ, Muffato M, Cox L, Else J, Disotell T, Muzny DM, Phillips-Conroy J, Aken B, Eichler EE, Marques-Bonet T, Kosiol C, Batzer MA, Hahn MW, Tung J, Zinner D, Roos C, Jolly CJ, Gibbs RA, Worley KC, Consortium BGA. 2019. The comparative genomics and complex population history of *Papio* baboons. *Sci Adv* **5**:eaau6947. doi:10.1126/sciadv.aau6947

Rosenberg NA, Nordborg M. 2002. Genealogical trees, coalescent theory and the analysis of genetic polymorphisms. *Nat Rev Genet* **3**:380–390. doi:10.1038/nrg795

Sackton TB, Clark N. 2019. Convergent evolution in the genomics era: new insights and directions. *Philos Trans R Soc B Biol Sci* **374**:20190102. doi:10.1098/rstb.2019.0102

Sanger F, Nicklen S, Coulson AR. 1977. DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci* **74**:5463–5467. doi:10.1073/PNAS.74.12.5463

Sarich VM, Wilson AC. 1967. Immunological time scale for hominid evolution. *Science (80-)* **158**:1200–1203. doi:10.1126/SCIENCE.158.3805.1200

Schierwater B, Eitel M, Jakob W, Osigus H-J, Hadrys H, Dellaporta SL, Kolokotronis S-O, DeSalle

R. 2009. Concatenated analysis sheds light on early metazoan evolution and fuels a modern "Urmetazoon" hypothesis. *PLoS Biol* **7**:e1000020. doi:10.1371/journal.pbio.1000020

Schwartz RS, Harkins KM, Stone AC, Cartwright RA. 2015. A composite genome approach to identify phylogenetically informative data from next-generation sequencing. *BMC Bioinformatics* **16**:193. doi:10.1186/s12859-015-0632-y

Scornavacca C, Delsuc F, Galtier N. 2020. Phylogenetics in the genomic era p.p. 1-568.

Sibley CG, Ahlquist JE. 1984. The phylogeny of the hominoid primates, as indicated by DNA-DNA hybridization. *J Mol Evol* **20**:2–15. doi:10.1007/BF02101980

Simion P, Delsuc F, Philippe H. 2020. To what extent current limits of phylogenomics can be overcome? *Phylogenetics Genomic Era* 2.1:1–2.1:34.

Simion P, Philippe H, Baurain D, Jager M, Richter DJ, Di Franco A, Roure B, Satoh N, Quéinnec É, Ereskovsky A, Lapébie P, Corre E, Delsuc F, King N, Wörheide G, Manuel M. 2017. A large and consistent phylogenomic dataset supports sponges as the sister group to all other animals. *Curr Biol* **27**:958–967. doi:10.1016/j.cub.2017.02.031

Springer MS, Gatesy J. 2016. The gene tree delusion. *Mol Phylogenet Evol* **94**:1–33. doi:10.1016/J.YMPEV.2015.07.018

Springer MS, Meredith RW, Teeling EC, Murphy WJ. 2013. Technical comment on "the placental mammal ancestor and the Post–K-Pg radiation of placentals." *Science (80-)* **341**:613–613. doi:10.1126/SCIENCE.1238025

Swofford DL. 2001. PAUP*: Phylogenetic analysis using parsimony (and other methods) version 4.0 beta.

Teeling EC, Vernes SC, Dávalos LM, Ray DA, Gilbert MTP, Myers E, Consortium B. 2018. Bat biology, genomes, and the Bat1K project: To generate chromosome-level genomes for all living bat species. *Annu Rev Anim Biosci* **6**:23–46. doi:10.1146/annurev-animal-022516-022811

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. doi:10.1093/molbev/msv013

Vanderpool D, Minh BQ, Lanfear R, Hughes D, Murali S, Harris RA, Raveendran M, Muzny DM, Hibbins MS, Williamson RJ, Gibbs RA, Worley KC, Rogers J, Hahn MW. 2020. Primate phylogenomics uncovers multiple rapid radiations and ancient interspecific introgression. *PLOS Biol* **18**:e3000954. doi:10.1371/journal.pbio.3000954

Vianna JA, Fernandes FAN, Frugone MJ, Figueiró H V., Pertierra LR, Noll D, Bi K, Wang-Claypool CY, Lowther A, Parker P, Bohec C Le, Bonadonna F, Wienecke B, Pistorius P, Steinfurth A, Burridge CP, Dantas GPM, Poulin E, Simison WB, Henderson J, Eizirik E, Nery MF, Bowie RCK. 2020. Genome-wide analyses reveal drivers of penguin diversification. *Proc Natl Acad Sci* **117**:22303–22310. doi:10.1073/PNAS.2006659117

Vicq-D′Azyr F. 1792. Systême anatomique..: Quadrupèdes. Panckoucke.

Weber MWC. 1904. Die Säugetiere; Einführung in die Anatomie und Systematik der recenten und fossilen Mammalia. Fischer.

Wilson AC, Carlson SS, White TJ. 1977. Biochemical evolution. *Annu Rev Biochem* **46**:573–639. doi:10.1146/annurev.bi.46.070177.003041

Woese CR. 1987. Bacterial evolution. *Microbiol Rev* **51**:221.

Yang Z. 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *J Mol Evol* **39**:306–314.

Zhang C, Rabiee M, Sayyari E, Mirarab S. 2018. ASTRAL-III: polynomial time species tree reconstruction from partially resolved gene trees. *BMC Bioinforma 2018 196* **19**:15–30. doi:10.1186/s12859-018-2129-y

Zhang F, Ding Y, Zhu C, Zhou X, Orr MC, Scheu S, Luan Y. 2019. Phylogenomics from low-

coverage whole-genome sequencing. *Methods Ecol Evol* **10**:507–517. doi:10.1111/2041-210X.13145

Zhang J, Lai J. 2020. Phylogenomic approaches in systematic studies. *Zool Syst* **43**:151–162.

Zou Z, Zhang J. 2015. No genome-wide protein sequence convergence for echolocation. *Mol Biol Evol* **32**:1237–1241. doi:10.1093/molbev/msv014

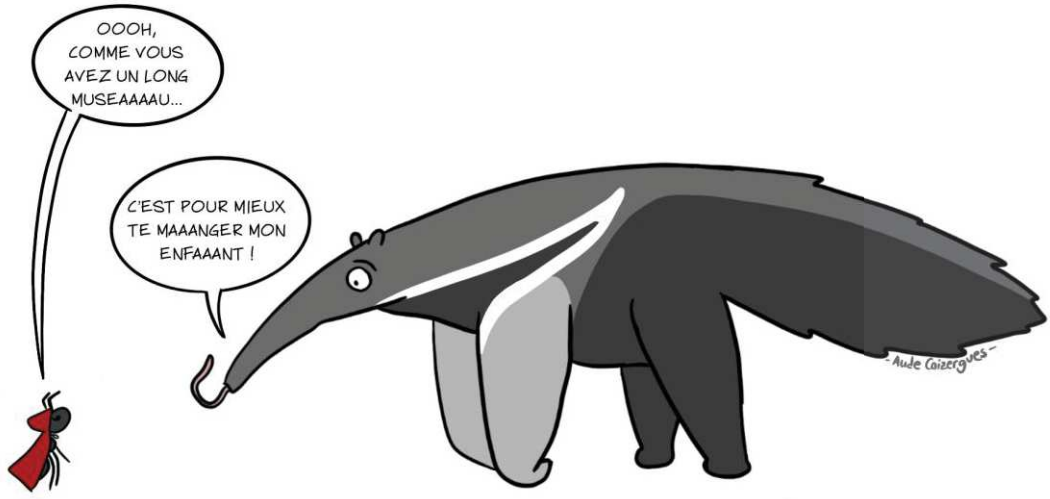
Zuckerkandl E, Jones RT, Pauling L. 1960. A comparison of animal hemoglobins by tryptic peptide pattern analysis. *Proc Natl Acad Sci U S A* **46**:1349–60. doi:10.1073/pnas.46.10.1349

Zuckerkandl E, Pauling L. 1965. Molecules as documents of evolutionary history. *J Theor Biol* **8**:357–366. doi:10.1016/0022-5193(65)90083-4

⇧Back to summary⇧

– PART I –

Ant and termite mitogenomic databases for diet characterisation from fecal samples



1- Diet characterisation using fecal samples

Although myrmecophagous mammals such as the lesser anteater (*Tamandua tetradactyla*), the Cape pangolin (*Smutsia temminckii*) or the aardvark (*Orycteropus afer*) are known to have a specialized diet based on ants and/or termites, the precise characterisation of which species are consumed and the seasonal variation in their diet remain unclear. Indeed, due to the elusivity of myrmecophagous mammal species and their particular lifestyles (essentially nocturnal for the Cape pangolin and aardvark and semi-arboreal for the lesser anteater, for example) simple field observations can be difficult to conduct to precisely determine their diet. In practice, studying the diet of myrmecophagous mammals usually involves investigations of chitinous parts of termites and ants obtained from the animal stomach, gut, or feces (e.g. Wu et al. 2005; Miranda et al. 2009, Sun et al. 2020). However, because the feces of myrmecophagous mammals also often contain soil, diet analyses from macroscopic prey remains are time- and labour- intensive (Sun et al. 2020). In this context, molecular screening of fecal samples, giving access to remnant DNA signatures of consumed preys, could be a promising strategy to efficiently characterize the regime of these rare species. Indeed, feces molecular screening had already been used successfully to characterize the diet of species for which field observations are complicated (see Pompanon et al. 2012 for a review). Briefly, once again, the development of next generation sequencing methods led to the apparition of interesting methods to efficiently sequenced the remnant DNA signatures of consumed preys in field-collected fecal samples. Overall, two approaches are commonly used and rely either on PCR amplification or target capture enrichment of the sequences associated with the potential preys and their comparison with reference barcoding databases (e.g. Pompanon et al. 2012; Shehzad et al. 2012; Alberti et al. 2018; Galan et al. 2018, Gauthier et al. 2020).

As presented in my introduction, the third part of the ERC ConvergeAnt project is dedicated to the investigation of the potential role of the microbiome in dietary convergence among ant-eating mammals. In that context, fecal samples of myrmecophagous mammals, among others, have been collected in the five past years and were available for sequencing at the beginning of my PhD. In addition to microbiome characterization, metagenomics could also potentially allow (i) the confirmation of the host species associated with the sample, and (ii) the precise characterization of the diet of myrmecophagous species. In the meantime, since the number of ant reference mitogenomes currently available is relatively low, a maximum of ant and termite species encountered during the fieldwork were sampled. To build a larger and more exhaustive database usable to precisely determine the diet of three species of ant-eating mammals, these species were collected in the same areas as the fecal samples (see sampling and approach section). This section presents a first attempt at a fully metagenomic diet analysis by describing (i) how fecal samples and ants and termites were collected in the field, (ii) how the ant and termite mitochondrial reference databases were constructed using a

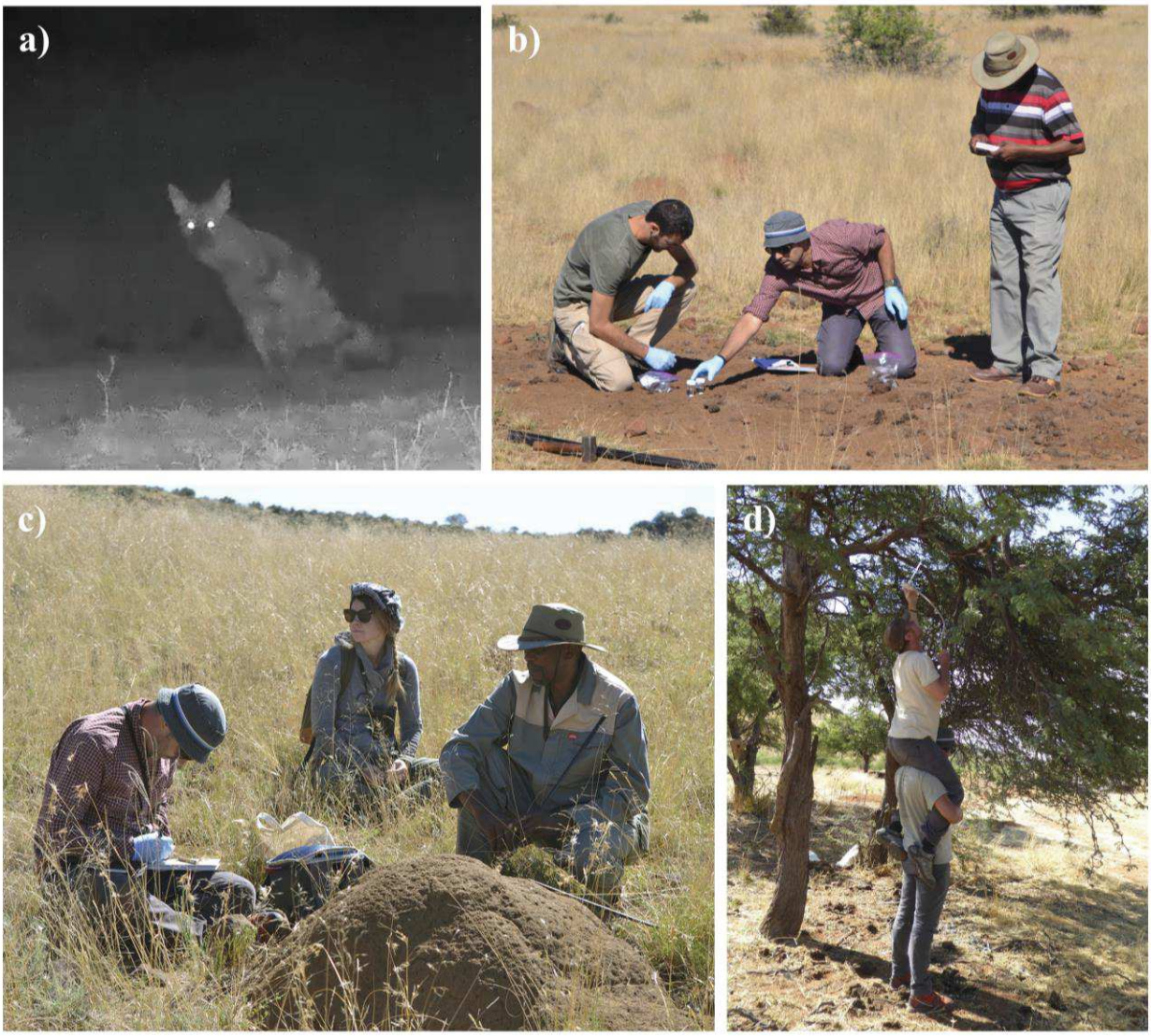


Figure 1 | a) Camera trap record of an aardwolf defecating in a lattrine. b) Sampling of fecal samples in an aardwolfs' lattrine. Tussen-die-Riviere, South Africa.. Sampling of c) termites and d) ants in myrmecophagous mammals native home range. Tussen-die-Riviere, South Africa.

newly developed bioinformatic tool, and (iii) a preliminary assessment of ant and or termite consumption in myrmecophagous mammals using metagenomic data from field-collected fecal samples.

1.1 - Sampling and approach

Sampling of fecal samples

All field sampling campaigns took place within each species native range in two natural reserves of South Africa between 2016 and 2018: for the aardwolf (*Proteles cristatus*) fecal samples were collected at Tussen-die-Riviere nature reserve in the Free State province (S 30° 28" 3.426', E 26° 7" 5.828'), and for the Cape pangolin (*Smutsia temminckii*) and the aardvark (*Orycteropus afer*), samples were collected at Tswalu Kalahari reserve in the Northern Cape province (S 27° 17' 49.626", E 22° 23' 43.569"). To reduce potential contamination by external DNA (such as feces-eating larvae), reduce DNA degradation, and for optimal use in microbiome characterization, fecal samples were collected as fresh as possible, but a number of dry fecal samples were also collected for diet characterization. The strategy used for collecting fresh fecal samples varied depending on species and habitat involved. For example, aardwolves (*Proteles cristatus*) urinate and defecate in collective places commonly called "latrines" (**Fig 1**). This peculiar behaviour facilitates the collecting of fresh fecal samples by regularly inspecting previously localised latrines in the early morning with the help of Nico L. Avenant and Tshediso Putsane (National Museum Bloemfontein) who previously conducted ecological studies at Tussen-die-Riviere. Regarding the aardvark and Cape pangolin, we benefited from the respective experience of Nora Weyer and Wendy Panaino (University of the Witwatersrand, Johannesburg), who both monitored and radio-tracked populations of these two elusive species during their respective PhDs conducted at Tswalu Kalahari (Weyer 2018; Panaino 2020). We mainly collected aardvark samples by surveying burrows surroundings in the early morning following the Guidance of Nora Weyer and Wendy Panaino allowed us to follow her at night to directly collect feces samples from her radio-tracked individuals and also provided access to frozen samples she has collected all year round. For other ant-eating species found in the reserves (i.e. the bat-eared fox and the black-backed jackal), the sampling was done opportunistically by making transects in different portions of both reserves. A total of 128 fecal samples preserved in 95% ethanol from myrmecophagous mammal species have been collected during the different fieldwork sessions.

Sampling of ants and termites

The objective of this sampling was to create a reference mitogenomic database of ants and termites to identify the species detected in fecal samples. In fact, even if a relatively large number (~500) of termite mitogenomes were publicly available three years ago, about 2600 species of termites are actually described. Regarding the ants, the gap between the number of species and the number of available mitogenomes was even more impressive. Indeed, among the ~13,000 species of ants described (AntCat 2014; Ward 2014), only 29 mitogenomes had been assembled and deposited in GenBank (as of March 29, 2018). To fill these gaps, we first tried to collect as many species of ants and termites as possible from the two South African reserves where we also collected fecal samples. In order to increase the number of reference mitogenomes for future diet metabarcoding studies, ants and termites were also collected opportunistically during field work sessions conducted in French Guiana and in Montpellier. In addition, regarding ants, for which the gap was the most important, we used another potential source of mitochondrial data. Indeed, a large amount of Ultra Conserved Elements (UCEs) have been sequenced in the past five years on numerous ant species to reconstruct different parts of the ant phylogeny (Blaimer et al. 2015; Faircloth et al. 2015; Blaimer et al. 2016; Branstetter et al. 2017a,b,c; Jesovnik et al. 2017; Pierce et al. 2017; Prebus et al. 2017; Ward & Branstetter 2017). Even if DNA sequence capture methods are used to efficiently enrich targeted DNA regions such as UCEs in library preparation prior to sequencing, up to 40% of the resulting reads can belong to non-targeted regions (Chilamakuri et al. 2014). Interestingly, mitochondrial DNA can often be assembled from these "off-target reads" (e.g. Smith et al. 2014; do Amaro et al. 2015). So to obtain as many ant mitogenomes as possible, we decided to use the 501 UCE libraries available as of March 29th, 2018 in the Short Read Archive (SRA) of the National Center for Biotechnology Information (NCBI).

Illumina Sequencing

DNA from 119 fresh and dry fecal samples out of the 128 preserved in 95% ethanol directly in the field was extracted using a protocol initially developed for extracting extracellular DNA from large amounts of soil material (Taberlet et al. 2012). For ants and termites, DNA was extracted from the abdomen of 219 whole specimens preserved in 95% ethanol using the Qiagen DNeasy blood and tissue extraction kit following manufacturer's instructions. Illumina libraries were then prepared from both types of DNA extracts following the cost-effective protocol developed by Tilak et al. (2015). Low-coverage shotgun sequencing using single 100bp reads was then performed on an Illumina HiSeq 2500 instrument at the Montpellier GenomiX Platform (MGX) at a sequencing depth of coverage of a few million reads per sample.

Mitogenome assemblies

Adapters and bad quality reads were removed from raw sequencing data using fastp (Chen et al. 2018). Given the large quantity of samples to analyse for both ants and termites (n = 318 + 501 UCE libraries) and feces (n = 119), we decided to develop a user-friendly bioinformatic pipeline. This pipeline, called MitoFinder, was designed to automatically assemble, extract, and annotate mitochondrial genomes from high throughput sequencing data (Allio et al. 2020, see Part 1.2). The first step of the pipeline consists in *de novo* assembling short reads into contigs using one of the three implemented metagenomic assemblers: MEGAHIT (Li et al. 2016), MetaSPAdes (Nurk et al. 2017) and IDBA-UD (Peng et al. 2012). Based on similarity searches (BLAST) against reference mitochondrial genomes, the second step consists of identifying mitochondrial contig(s) among the assemblies. Finally, mitochondrial genes are annotated using reference mitochondrial genes (at the amino acid level for protein-coding genes and at the nucleotide level for ribosomal RNAs), and tRNA genes are annotated using either ARWEN (Laslett & Canbäck 2007), tRNAscan (Chan & Lowe 2019) or MiTFi (Juhling et al. 2012).

Given that MitoFinder extraction and annotation steps are based on user-provided reference mitochondrial genomes, two specific databases were created. On the one hand, to extract mitogenomic data from ants and termites samples, a database of 1803 insect reference mitogenomes including all species of termites and ants available in GenBank as of July 20th, 2018 was compiled. On the other hand, a second database including reference mitogenomes of 10 ant-eating mammal species found in the two reserves was compiled to allow mitochondrial genome extraction of the host from fecal samples. We first ran MitoFinder using default parameters independently for ant and termite samples and for fecal samples with the first reference insect database. Fecal samples were then analysed a second time using the second mammal reference database to confirm host identity via mitochondrial DNA barcoding. Finally, shotgun reads obtained from all fecal samples were mapped using the Geneious R10 read mapper (Kearse et al. 2012) with stringent “Low Sensitivity” default parameters against both the potential host mammal species mitogenome reference database to identify host reads, and against the mitochondrial contigs of ants and termites previously annotated by MitoFinder to identify prey reads.

Barcoding and phylogenetic inferences

Ant and termite species were first tentatively identified in the field based on morphological observations against a field guide. Then, cytochrome c oxidase subunit 1 (COX1) sequences extracted by MitoFinder were compared with Species Level Barcode Records (including more than four millions COX1 sequences) through the identification server of the Barcode Of Life Data System v4 (Ratnasingham & Hebert, 2007). Finally, phylogenetic inferences were conducted to identify species for which no COX1 sequences were found. To do so, each mitochondrial loci was first aligned independently using MAFFT with mitochondrial loci extracted from 170 and 696 published

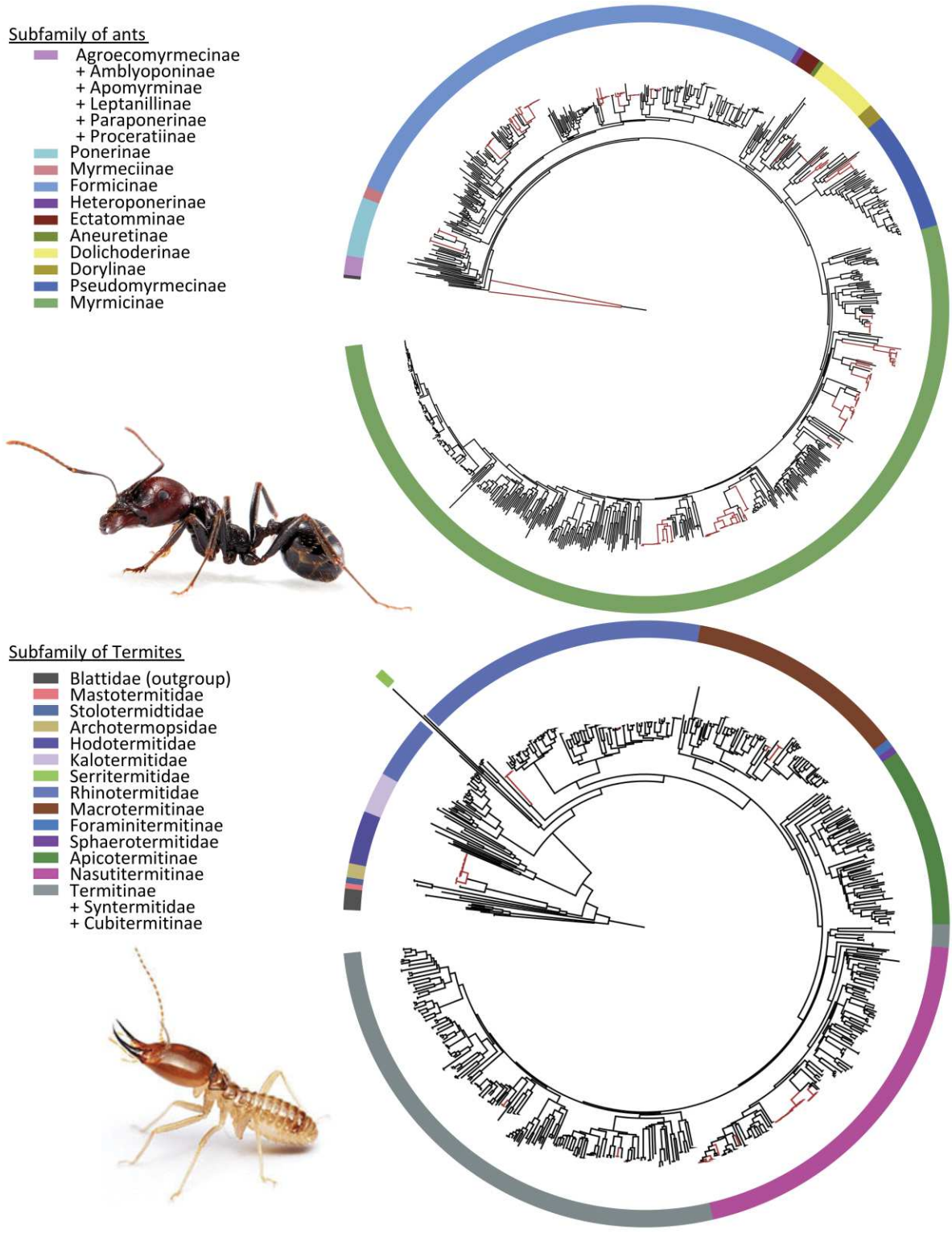


Figure 2 | Phylogenies of a) ants and b) termites based on complete reference mitogenomes and new mitogenomic data extracted from shotgun sequencing of field-collected specimens and available UCE capture libraries. Image credits for the termite: Orkin Canada.

mitogenomes of ants and termites, respectively. Then, for both groups, a mitochondrial nucleotide supermatrix consisting of the concatenation of the 13 protein-coding genes and the two rRNA genes was compiled. Finally, phylogenetic inferences were performed with Maximum Likelihood (ML) as implemented in IQ-TREE v1.6.8 (Nguyen et al. 2015) using a GTR+Γ4+I model.

1.2 - Results & discussion

Using MitoFinder, we were able to extract mitochondrial sequences from all 88 termite samples. Among these samples, 79 complete mitogenomes (including the 13 mitochondrial loci and the two rRNA loci) were retrieved. Overall, after filtering for short contigs (smaller than 2,000 bp) or contigs containing less than two annotated genes, 1.04 mitochondrial contigs and 13.88 mitochondrial loci per sample were retrieved on average. Finally, COX1 sequence was retrieved for 87 out of the 88 samples for which mitochondrial sequences were extracted. Regarding ants, we were able to extract mitochondrial sequences from 222 out of the 230 samples, including 24 complete mitogenomes, based on similarity with the 171 reference mitogenomes. Then, using the same filtering method, 1.56 mitochondrial contigs and 10.67 mitochondrial loci per sample were retrieved on average. Overall, mitochondrial sequences were extracted for 310 out of 318 samples and COX1 barcoding sequences were successfully extracted for 281 out of 310 samples for which mitochondrial sequences were retrieved.

Even if we were unable to extract mitochondrial sequences from some ant samples, these results confirm that MitoFinder is an efficient tool to automatically extract mitochondrial signal from high throughput sequencing data (Allio et al. 2020). These additional mitochondrial sequences complete the mitogenomes available for ants and termites with a specific emphasis on species found in myrmecophagous mammals' habitats. Based on the phylogenetic trees obtained from the concatenation of the thirteen protein-coding mitochondrial genes and the two rRNAs, our study permitted the addition of new mitochondrial sequences for 12-15 and 45-50 species of termites and ants, respectively (**Fig. 2a,b**). Additionally, phylogenetic inferences allowed us to identify species for which no COX1 sequences were extracted. This strategy to identify species is of particular interest in metabarcoding studies. Indeed, in our case, shotgun sequencing on fecal samples may lead to partial mitochondrial sequences including or not the COX1 barcoding region. By using a database including whole mitogenome information, we have more chances to identify the prey species found in myrmecophagous mammals' feces.

Given the low depth of sequencing coverage , no mitochondrial contig of ants or termites was reconstructed from our 119 feces shallow metagenomic data. Additionally, mammal mitochondrial contigs were successfully extracted with MitoFinder in only eight fecal samples, among which five

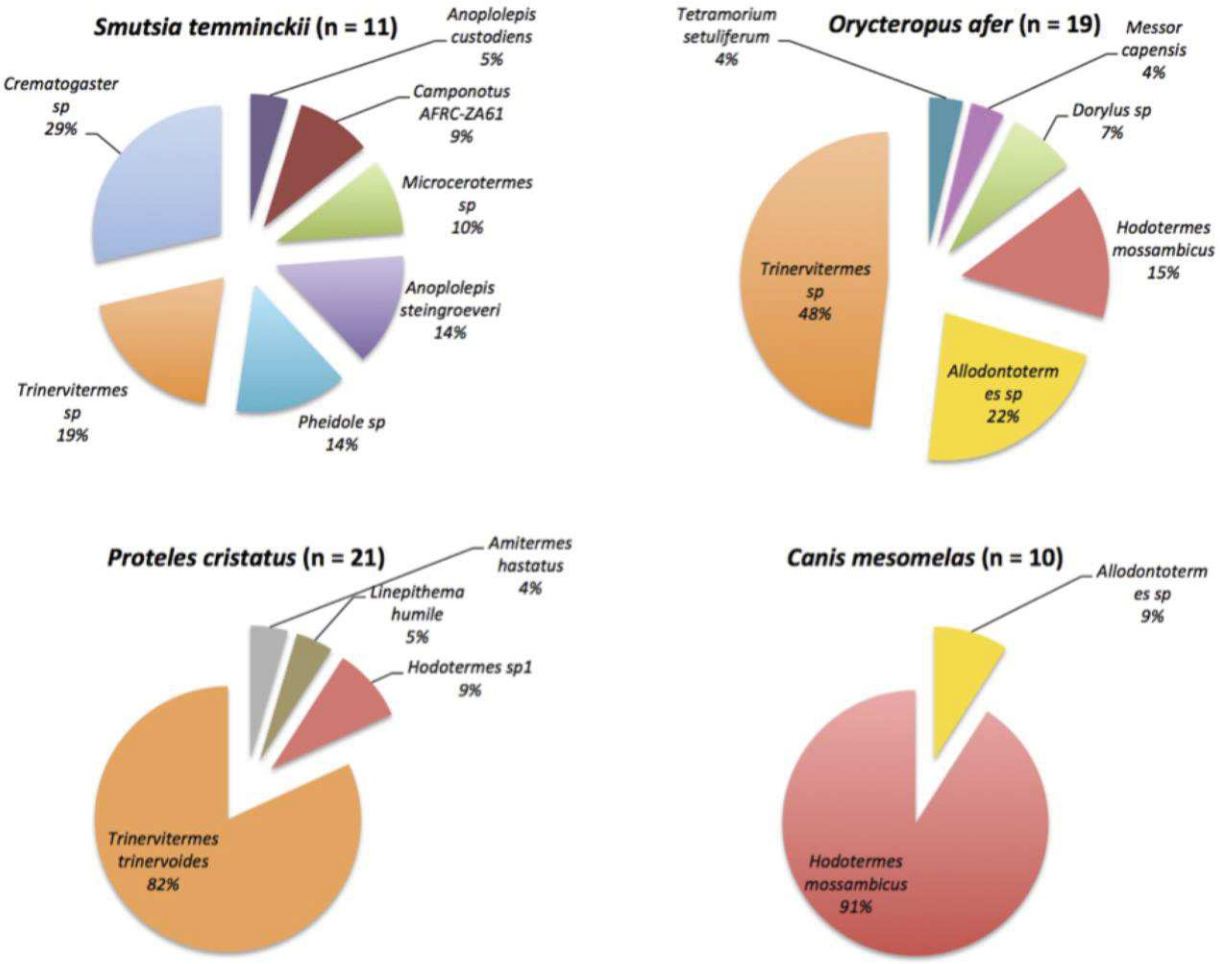


Figure 3 | Preliminary assessment of the diet of four ant-eating mammals of two south-african natural reserves using metagenomics of fecal samples. The diagrams represent the percentage of occurrence of termites and ants as detected through the mapping of sequence reads obtained from shotgun metagenomic of fecal samples from the Cape pangolin (*Smutsia temminckii*), the aardvark (*Orycteropus afer*), the aardwolf (*Proteles cristatus*), and the black-backed jackal (*Canis mesomelas*).

pangolin complete mitogenomes, allowing the confirmation of the host species associated with the sample. Direct mapping of the shotgun reads to the reference mammalian mitogenomes nevertheless allowed identifying the host species in 90/119 samples even though host reads represent only a tiny fraction of fecal metagenomic reads (0.006% on average with only 0.15% for the best pangolin sample). Importantly, this molecular signature permitted assigning a number of unidentified dry fecal samples containing a lot of termites as belonging to the black-backed jackal (*Canis mesomelas*). Similarly, mapping the metagenomic reads on the reference contigs previously assembled from ants and termites by MitoFinder retrieved reads belonging to potential preys in 50/119 fecal samples, again with very few exact matches per prey contig. This scarce metagenomic data nevertheless allowed us to get a first glimpse at the diet composition of four ant-eating mammal species for which we had more than 10 positive samples (**Fig. 3**). Interestingly, our preliminary molecular-based estimates fit very well with previous diet assessment based on macroscopic observations of fecal samples. Indeed, the Cape pangolin was found to have the most diverse diet including a number of ant (*Crematogaster*, *Pheidole* and *Anoplolepis*) and termite (*Trinervitermes* and *Microcerotermes*) genera in its diet confirming previous ecological and behavioural studies (Swart et al. 1999) including one conducted at Tswalu Kalahari reserve (Panaino 2020). The aardvark seems to predominantly preys on termites (*Trinervitermes*, *Allodontotermes*, and *Hodotermes*) with only a few ant occurrences including driver ants (*Dorylus*) as previously observed at Tswalu Kalahari reserve (Weyer 2018). At Tussen-die-Riviere reserve, our molecular results confirm the highly specialized diet of the aardwolf feeding almost exclusively on the termite species *Trinervitermes trinervoides* (De Vries et al. 2011), with a few occurrences of harvester termites (*Hodotermes*) and an interesting potential first observation of the invasive Argentine ant (*Linepithema humile*) that was indeed sampled at this reserve. Finally, the unidentified fecal samples assigned to the black-backed jackal were found to contain mostly harvester termite (*Hodotermes mossambicus*), which is known to be the most abundant and more often consumed prey item by this highly labile carnivore species during the dry season (Kaunda & Skinner 2003).

Even though our first attempt at a metagenomic analysis of the diet of ant-eating mammals provides interesting preliminary results, it is hampered by the very reduced proportion of prey reads found in fecal samples. For both prey and mammal host identification, the low number of mitochondrial sequences is explained by the overrepresentation of bacterial sequences in our extraction. This bias towards bacteria could be due in part to our extraction protocol in which only the extracellular DNA fragments of the supernatant were extracted for subsequent library preparation and sequencing, and not a mixture of the whole fecal sample. To counter this bias and be able to more specifically extract the DNA fragments belonging to the preys consumed by myrmecophagous mammals, we plan to rely on DNA sequence capture methods. Indeed, using our newly generated datasets of mitochondrial genomes of ants and termites, it will be possible to create specific capture

baits targeting specifically these two groups of social insects. Additionally, homogenising fecal samples or specifically filtering for the chitinous part of the samples before DNA extraction, could help us to obtain more DNA fragments of interest for sequencing.

To conclude, this project led to the development of a user-friendly pipeline designed to assemble, extract, and annotate mitochondrial sequences from high throughput sequencing data (see Part I - section 2 for a case study). Using this pipeline, we were able to add a significant number of reference mitochondrial sequences for ant and termite species found in myrmecophagous mammal' natural habitats to existing mitogenome datasets. Furthermore, we point out the utility of using complete mitochondrial genome information to identify species when COX1 region was not extracted. Due to overrepresentation of bacterial DNA fragments in our DNA extractions from fecal samples, only a few ant or termite mitochondrial sequence reads were retrieved in our fecal samples. Nevertheless, by using our newly generated ant and termite mitogenomes, we were able to provide a first molecular assessment of the diet of four ant-eating mammal species confirming previous studies based on macroscopic observations of feces content. Developing capture baits based on our newly created mitochondrial reference datasets for ants and termites may help us to specifically extract DNA fragments from the ingested preys to better characterize the diet of myrmecophagous mammals (Gauthier et al. 2020).

References

⇧Back to summary⇧

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 Resour* 1755–0998.13160. doi:10.1111/1755-0998.13160

AntCat. 2014. AntCat. An online catalog of the ants of the world. https://antcat.org/

Blaimer BB, Brady SG, Schultz TR, Lloyd MW, Fisher BL, Ward PS. 2015. Phylogenomic methods outperform traditional multi-locus approaches in resolving deep evolutionary history: a case study of formicine ants. *BMC Evol Biol* **15**:271. doi:10.1186/s12862-015-0552-5

Blaimer BB, LaPolla JS, Branstetter MG, Lloyd MW, Brady SG. 2016. Phylogenomics, biogeography and diversification of obligate mealybug-tending ants in the genus Acropyga. *Mol Phylogenet Evol* **102**:20–29. doi:10.1016/J.YMPEV.2016.05.030

Branstetter MG, Danforth BN, Pitts JP, Faircloth BC, Ward PS, Buffington ML, Gates MW, Kula RR, Brady SG. 2017a. Phylogenomic Insights into the Evolution of Stinging Wasps and the Origins of Ants and Bees. *Curr Biol* **27**:1019–1025. doi:10.1016/J.CUB.2017.03.027

Branstetter MG, Ješovnik A, Sosa-Calvo J, Lloyd MW, Faircloth BC, Brady SG, Schultz TR. 2017b. Dry habitats were crucibles of domestication in the evolution of agriculture in ants. *Proc R Soc B Biol Sci* **284**:20170095. doi:10.1098/rspb.2017.0095

Branstetter MG, Longino JT, Ward PS, Faircloth BC. 2017c. Enriching the ant tree of life: enhanced UCE bait set for genome-scale phylogenetics of ants and other Hymenoptera. *Methods Ecol Evol* **8**:768–776. doi:10.1111/2041-210X.12742

Chan PP, Lowe TM. 2019. tRNAscan-SE: Searching for tRNA genes in genomic sequences, Gene Prediction. Humana, New York, NY. doi:10.1007/978-1-4939-9173-0_1

Chen S, Zhou Y, Chen Y, Gu J. 2018. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**:i884–i890. doi:10.1093/bioinformatics/bty560

Chilamakuri CS, Lorenz S, Madoui M-A, Vodák D, Sun J, Hovig E, Myklebost O, Meza-Zepeda LA. 2014. Performance comparison of four exome capture systems for deep sequencing. *BMC Genomics* **15**:449. doi:10.1186/1471-2164-15-449

de Vries JL, Pirk CWW, Bateman PW, Cameron EZ, Dalerum F. 2011. Extension of the diet of an extreme foraging specialist, the aardwolf (*Proteles cristata*). *African Zool* **46**:194–196. doi:10.1080/15627020.2011.11407494

do Amaral FR, Neves LG, Resende Jr MFR, Mobili F, Miyaki CY, Pellegrino KCM, Biondo C. 2015. Ultraconserved elements sequencing as a low-cost source of complete mitochondrial genomes and microsatellite markers in non-model amniotes. *PLoS One* **10**:e0138446. doi:10.1371/journal.pone.0138446

Faircloth BC, Branstetter MG, White ND, Brady SG. 2015. Target enrichment of ultraconserved elements from arthropods provides a genomic perspective on relationships among Hymenoptera. *Mol Ecol Resour* **15**:489–501. doi:10.1111/1755-0998.12328

Galan M, Pons J-B, Tournayre O, Pierre É, Leuchtmann M, Pontier D, Charbonnel N. 2018. Metabarcoding for the parallel identification of several hundred predators and their prey: Application to bat species diet analysis. *Mol Ecol Resour* **18**:474–489. doi:10.1111/1755-0998.12749

Gauthier M, Konecny-Dupré L, Nguyen A, Elbrecht V, Datry T, Douady C, Lefébure T. 2020. Enhancing DNA metabarcoding performance and applicability with bait capture enrichment and DNA from conservative ethanol. *Mol Ecol Resour* **20**:79–96. doi:10.1111/1755-0998.13088

Ješovnik A, Sosa-Calvo J, Lloyd MW, Branstetter MG, Fernández F, Schultz TR. 2017. Phylogenomic species delimitation and host-symbiont coevolution in the fungus-farming ant genus Sericomyrmex Mayr (Hymenoptera: Formicidae): ultraconserved elements (UCEs) resolve a recent radiation. *Syst Entomol* **42**:523–542. doi:10.1111/syen.12228

Jühling F, Pütz J, Bernt M, Donath A, Middendorf M, Florentz C, Stadler PF. 2012. Improved systematic tRNA gene annotation allows new insights into the evolution of mitochondrial tRNA structures and into the mechanisms of mitochondrial genome rearrangements. *Nucleic Acids Res* **40**:2833–2845. doi:10.1093/nar/gkr1131

Kaunda SKK, Skinner JD. 2003. Black-backed jackal diet at Mokolodi Nature Reserve, Botswana. *Afr J Ecol* **41**:39–46. doi:10.1046/j.1365-2028.2003.00405.x

Kearse M, Sturrock S, Meintjes P. 2012. The Geneious 6.0. 3 read mapper. *Biomatters Ltd Auckland, New Zeal*.

Laslett D, Canback B. 2007. ARWEN: a program to detect tRNA genes in metazoan mitochondrial nucleotide sequences. *Bioinformatics* **24**:172–175. doi:10.1093/bioinformatics/btm573

Li D, Luo R, Liu C-M, Leung C-M, Ting H-F, Sadakane K, Yamashita H, Lam T-W. 2016. MEGAHIT v1.0: A fast and scalable metagenome assembler driven by advanced methodologies and community practices. *Methods* **102**:3–11. doi:10.1016/J.YMETH.2016.02.020

Miranda F, Veloso R, Superina M, Zara FJ. 2009. Food habits of wild silky anteaters (*Cyclopes didactylus*) of São Luis do Maranhão, Brazil. *Edentata* **8**–**10**:1–5. doi:10.1896/020.010.0109

Nguyen L-T, 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. doi:10.1093/molbev/msu300

Nurk S, Meleshko D, Korobeynikov A, Pevzner PA. 2017. metaSPAdes: a new versatile metagenomic assembler. *Genome Res* **27**:824–834. doi:10.1101/gr.213959.116

Panaino W. 2020. Diet, activity, and body temperature patterns of ground pangolins in a semi-arid environment. University of the Witwatersrand, Johannesburg.

Peng Y, Leung HCM, Yiu SM, Chin FYL. 2012. IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics* **28**:1420–1428. doi:10.1093/bioinformatics/bts174

Pierce MP, Branstetter MG, Longino JT. 2017. Integrative taxonomy reveals multiple cryptic species within Central American Hylomyrma FOREL, 1912 (Hymenoptera: Formicidae). *Myrmecological News* **25**:131–143. doi:10.25849/myrmecol.news_025:131

Pompanon F, Deagle BE, Symondson WOC, Brown DS, Jarman SN, Taberlet P. 2012. Who is eating what: diet assessment using next generation sequencing. *Mol Ecol* **21**:1931–1950. doi:10.1111/j.1365-294X.2011.05403.x

Prebus M. 2017. Insights into the evolution, biogeography and natural history of the acorn ants, genus Temnothorax Mayr (hymenoptera: Formicidae). *BMC Evol Biol* **17**:250. doi:10.1186/s12862-017-1095-8

Ratnasingham S, Hebert PDN. 2007. BOLD: The Barcode of Life Data System (http://www. barcodinglife. org). *Mol Ecol Notes* **7**:355–364. doi:10.1111/j.1471-8286.2007.01678.x

Shehzad W, Riaz T, Nawaz MA, Miquel C, Poillot C, Shah SA, Pompanon F, Coissac E, Taberlet P. 2012. Carnivore diet analysis based on next-generation sequencing: application to the leopard cat (*Prionailurus bengalensis*) in Pakistan. *Mol Ecol* **21**:1951–1965. doi:10.1111/j.1365-294X.2011.05424.x

Smith BT, Harvey MG, Faircloth BC, Glenn TC, Brumfield RT. 2014. Target capture and massively parallel sequencing of ultraconserved elements for comparative studies at shallow evolutionary time scales. *Syst Biol* **63**:83–95. doi:10.1093/sysbio/syt061

Sun NC, Lo FH, Chen B, Yu H, Liang C, Lin C, Chin S, Li H. 2020. Digesta retention time and recovery rates of ants and termites in Chinese pangolins (*Manis pentadactyla*). *Zoo Biol* **39**:168–175. doi:10.1002/zoo.21534

Swart JM, Richardson PRK, Ferguson JWH. 1999. Ecological factors affecting the feeding behaviour of pangolins (*Manis temminckii*). *J Zool* **247**:281–292. doi:10.1111/j.1469-7998.1999.tb00992.x

Taberlet P, Prud'homme SM, Campione E, Roy J, Miquel C, Shehzad W, Gielly L, Rioux D, Choler P, Clément J-C, Melodelima C, Pompanon F, Coissac E. 2012. Soil sampling and isolation of extracellular DNA from large amount of starting material suitable for metabarcoding studies. *Mol Ecol* **21**:1816–1820. doi:10.1111/j.1365-294X.2011.05317.x

Tilak M-K, 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. *Conserv Genet Resour* **7**:37–40. doi:10.1007/s12686-014-0338-x

Ward PS. 2014. The phylogeny and evolution of ants. *Annu Rev Ecol Evol Syst* **45**:23–43. doi:10.1146/annurev-ecolsys-120213-091824

Ward PS, Branstetter MG. 2017. The acacia ants revisited: convergent evolution and biogeographic context in an iconic ant/plant mutualism. *Proc R Soc B Biol Sci* **284**:20162569. doi:10.1098/rspb.2016.2569

Weyer NM. 2018. Physiological flexibility of free-living aardvarks (Orycteropus afer) in response to environmental fluctuations. University of the Witwatersrand, Johannesburg, South Africa.

Wu S, Liu N, Li Y, Sun R. 2005. Observation on food habits and foraging behavior of Chinese pangolin (Manis pentadactyla). *Chinese J Appl Environ Biol* **11**:337.

⇧Back to summary⇧

2 - MitoFinder: A user-friendly pipeline to assemble and annotate mitochondrial genomes

The following article is associated with the development of the MitoFinder software. In this article, we focused on the usefulness of MitoFinder to extract mitochondrial data from Ultra Conserved Elements (UCEs) sequencing libraries. Indeed, in most cases, we show that mitochondrial genes could be assembled from off target reads despite the fact that they are rarely extracted and used for further analyses in UCE-based studies. However, mitochondrial signal may offer (i) the opportunity to check for species identification (which is complex with only ultraconserved element signal), (ii) additional phylogenetic information (in particular for the more recent part of the inference), and (iii) information on potential species hybridization reflected by mito-nuclear conflicts. By developing a user-friendly pipeline to extract the mitochondrial signal from UCE libraries, in light of the value of the mitochondrial signal, we think that the extraction of this additional complementary genomic marker should be more systematically performed from UCE data.

The journal article associated with this section can be found online:
https://doi.org/10.1111/1755-0998.13160
As well as the supplementary material:
https://doi.org/10.5281/zenodo.3231389
And MitoFinder pipeline:
https://github.com/RemiAllio/MitoFinder

⇧Back to summary⇧

MitoFinder: efficient automated large-scale extraction of mitogenomic data in target enrichment phylogenomics

Rémi Allio[1,*], Alex Schomaker-Bastos[2,†], Jonathan Romiguier[1], Francisco Prosdocimi[2], Benoit Nabholz[1], and Frédéric Delsuc[1,*]

[1]*Institut des Sciences de l'Évolution de Montpellier (ISEM), CNRS, EPHE, IRD, Université de Montpellier, Montpellier, France.*
[2]*Laboratório Multidisciplinar para Análise de Dados (LAMPADA), Instituto de Bioquímica Médica Leopoldo de Meis, Universidade Federal do Rio de Janeiro, Rio de Janeiro, Brasil.*
[†] *In Memoriam (08/01/2015)*

***Correspondence:** remi.allio@umontpellier.fr; frederic.delsuc@umontpellier.fr

Abstract

Thanks to the development of high-throughput sequencing technologies, target enrichment sequencing of nuclear ultraconserved DNA elements (UCEs) now allows routinely inferring phylogenetic relationships from thousands of genomic markers. Recently, it has been shown that mitochondrial DNA (mtDNA) is frequently sequenced alongside the targeted loci in such capture experiments. Despite its broad evolutionary interest, mtDNA is rarely assembled and used in conjunction with nuclear markers in capture-based studies. Here, we developed MitoFinder, a user-friendly bioinformatic pipeline, to efficiently assemble and annotate mitogenomic data from hundreds of UCE libraries. As a case study, we used ants (Formicidae) for which 501 UCE libraries have been sequenced whereas only 29 mitogenomes are available. We compared the efficiency of four different assemblers (IDBA-UD, MEGAHIT, MetaSPAdes, and Trinity) for assembling both UCE and mtDNA loci. Using MitoFinder, we show that metagenomic assemblers, in particular MetaSPAdes, are well suited to assemble both UCEs and mtDNA. Mitogenomic signal was successfully extracted from all 501 UCE libraries allowing confirming species identification using CO1 barcoding. Moreover, our automated procedure retrieved 296 cases in which the mitochondrial genome was assembled in a single contig, thus increasing the number of available ant mitogenomes by an order of magnitude. By leveraging the power of metagenomic assemblers, MitoFinder provides an efficient tool to extract complementary mitogenomic data from UCE libraries, allowing testing for potential mito-nuclear discordance. Our approach is potentially applicable to other sequence capture methods, transcriptomic data, and whole genome shotgun sequencing in diverse taxa.

Keywords: Bioinformatics, DNA Barcoding, Invertebrates, Metagenomics, Systematics, Insects

Running head: Mitochondrial signal from UCE capture data

Introduction

Next generation phylogenomics in which phylogenetic relationships are inferred from thousands of genomic markers gathered through high-throughput sequencing (HTS) is on the rise. More specifically, targeted enrichment or DNA sequence capture methods are becoming the gold standard in phylogenetic analyses because they allow subsampling the genome efficiently at reduced cost (Lemmon & Lemmon, 2013; McCormack, Hird, Zellmer, Carstens, & Brumfield 2013). The field has witnessed the rapid parallel development of exon capture from transcriptome-derived baits (Bi *et al.* 2012), anchored hybrid enrichment techniques (Lemmon, Emme, & Lemmon 2012), and the capture of ultraconserved DNA elements (UCEs; Faircloth et al. 2012). All hybridization capture methods target a particular portion of the genome corresponding to the defined probes plus flanking regions. Prior knowledge is required to generate sequence capture probes, but ethanol preserved tissues, old DNA extractions, and museum specimens can be successfully sequenced (Faircloth *et al.* 2012; Guschanski *et al.* 2013; Blaimer *et al.* 2015). The first UCEs were identified by Bejerano *et al.* (2004) in the human genome and have been shown to be conserved in mammals, birds, and even ray-finned fish (Stephen, Pheasant, Makunin, & Mattick 2008). Thanks to their large-scale sequence conservation, UCEs are particularly well suited for sequence capture experiments and have become popular for phylogenomic reconstruction of diverse animals groups (Guschanski *et al.* 2013; Blaimer *et al.* 2015; Esselstyn, Oliveros, Swanson, & Faircloth 2017). Initially restricted to a few vertebrate groups such as mammals (McCormack *et al.* 2012) and birds (McCormack *et al.* 2013), new UCE probe sets have been designed to target thousands of loci in arthropods such as hymenopterans (Blaimer *et al.* 2015; Branstetter *et al.* 2017a; Faircloth, Branstetter, White, & Brady 2015), coleopterans (Baca, Alexander, Gustafson, & Short 2017, Faircloth 2017), and arachnids (Starrett *et al.* 2017).

It has been shown that complete mitochondrial genomes could be retrieved as by-products of sequence capture/enrichment experiments such as whole exome capture in human (Picardi & Pesole, 2012). Indeed, mitogenomes can in most cases be assembled from off-target sequences of UCE capture libraries in amniotes (do Amaral *et al.* 2015). Despite its well-acknowledged limitations (Galtier, Nabholz, Glémin, & Hurst 2009), mitochondrial DNA (mtDNA) remains a marker of choice for phylogenetic inference (e.g. Hassanin *et al.* 2012), for species identification or delimitation through barcoding (e.g. Coissac *et al.* 2016), and to reveal potential cases of mito-nuclear discordance resulting from introgression and/or hybridization events (e.g. Zarza et al. 2016, 2018; Grummer, Morando, Avila, Sites Jr, & Leaché 2018). MtDNA could also be used to taxonomically validate the specimens sequenced for UCEs using CO1 barcoding (Ratnasingham & Hebert, 2007) and to control for potential cross-contaminations in HTS experiments (Ballenghien, Faivre, & Galtier 2017). In practice, the few studies that have extracted mtDNA signal from UCEs (*e.g.* Meiklejohn *et al.* 2014; Pie *et al.* 2017; Wang, Hosner, Liang, Braun, & Kimball 2017, Zarza *et al.* 2018) and anchored phylogenomics (Caparroz *et al.* 2018) have done so manually for only a few taxa. Most studies assembling mitogenomes from UCE libraries have used contigs produced by the Trinity RNAseq assembler (Grabherr *et al.* 2011) as part of the PHYLUCE pipeline (Faircloth, 2016), which was specifically designed to extract UCE loci. Indeed, RNAseq assemblers such as Trinity allow dealing with the uneven coverage of target reads in sequence-capture libraries, but also multi-copy genes such as the ribosomal RNA cluster, and organelles (chloroplasts and mitochondria). However, this strategy is likely not scaling well with hundreds of taxa because of the high computational demand required by Trinity. A potential solution to extract

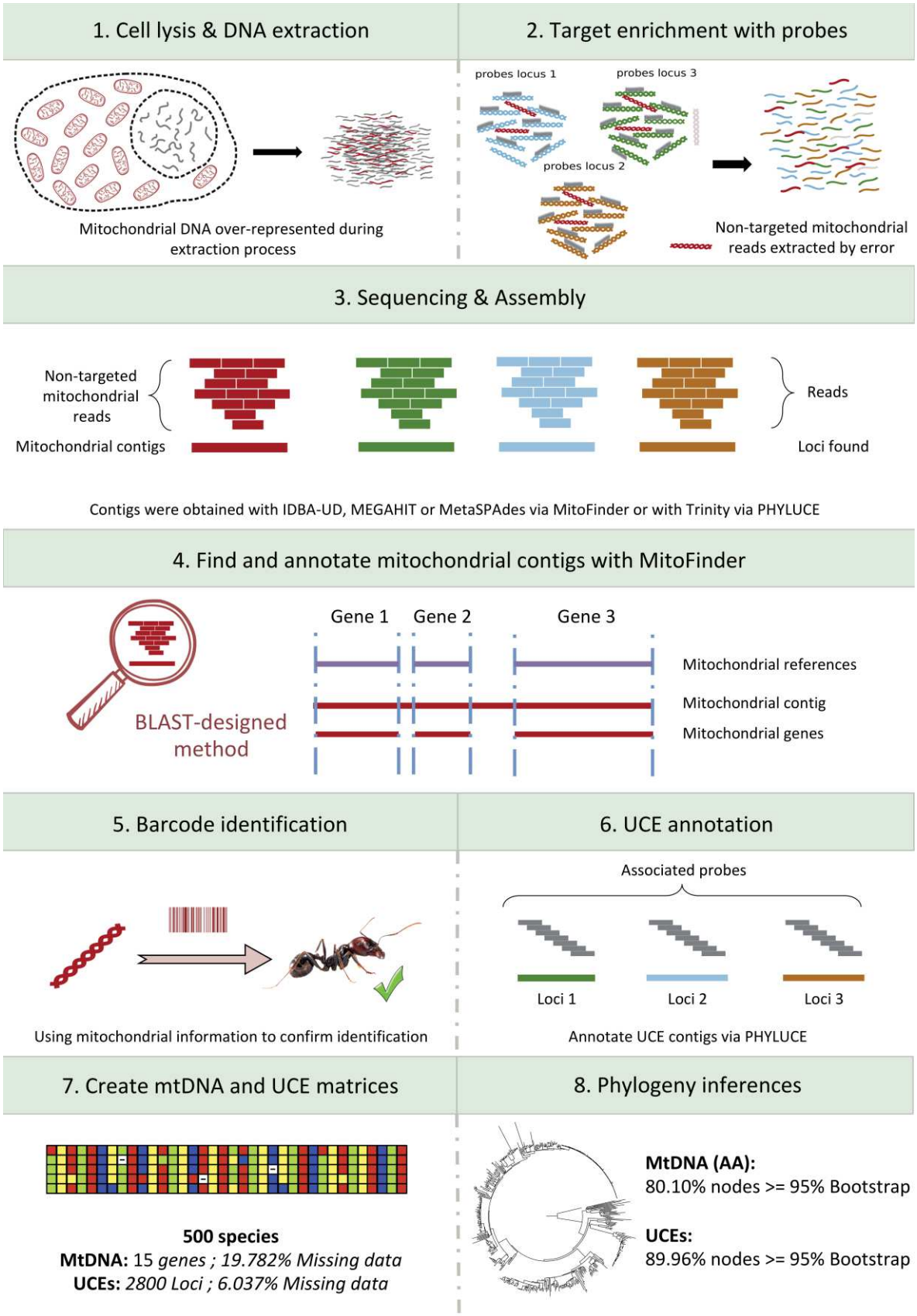


Figure 1 | Conceptualization of the pipeline used to assemble and extract UCE and mitochondrial signal from ultraconserved element sequencing data

mitochondrial signal from UCE libraries could be the use of iterative mapping against a reference mitogenome using MITObim (Hahn, Bachmann & Chevreux 2013). However, this tool requires both closely related reference mitogenomes and good coverage to perform well and also requires UCE and mtDNA assemblies to be conducted separately. Metagenomic assemblers could provide a powerful alternative to assemble both UCE loci and mtDNA simultaneously because they have been designed for an efficient *de novo* assembly of complex read populations by explicitly dealing with uneven read coverage and are computationally and memory efficient. Comparisons based on empirical bulk datasets of known composition (Vollmers, Wiegand, & Kaster 2017) have identified IDBA-UD (Peng, Leung, Yiu, & Chin 2012), MEGAHIT (Li *et al.* 2016), and MetaSPAdes (Nurk, Meleshko, Korobeynikov, & Pevzner 2017) as the most efficient current metagenomic assemblers.

As a case study, we focused on ants (Hymenoptera: Formicidae) for which only 29 mitogenomes were available on GenBank compared to 501 UCE captured libraries as of March 29th, 2018 (**Appendix S1**). This contrasts sharply with the other most speciose group of social insects, termites (Isoptera), for which almost 500 reference mitogenomes have been produced (Bourguignon *et al.* 2017) and no UCE study has been conducted so far. Sequencing and assembling difficulties stemming from both the AT-rich composition (Foster, Jermiin, & Hickey 1997) and a high rate of mitochondrial genome rearrangements in hymenopterans (Dowton, Castro, & Austin 2002) might explain the limited number of mitogenomes currently available for ants. It is only recently that a few ant mitogenomes have been assembled from UCE data (Ströher *et al.* 2017; Meza-Lázaro, Poteaux, Bayona-Vásquez, Branstetter, & Zaldívar-Riverón 2018; Vieira & Prosdocimi, 2019). Here, we built a pipeline called MitoFinder designed to automatically assemble both UCE and mtDNA from raw UCE capture libraries and to specifically extract and annotate mitogenomic contigs. Using publicly available UCE libraries for 501 ants, we show that complementary mitochondrial phylogenetic signal can be efficiently extracted using metagenome assemblers along with targeted UCE loci.

Materials and methods

Data acquisition

We used UCE raw sequencing data for 501 ants produced in 10 phylogenomic studies (Blaimer *et al.* 2015; Faircloth *et al.* 2015; Blaimer *et al.* 2016; Branstetter *et al.* 2017a,b,c; Jesovnik *et al.* 2017; Pierce *et al.* 2017; Prebus *et al.* 2017; Ward & Branstetter 2017). This dataset includes representatives of 15 of the 16 recognized subfamilies (Ward 2014) and 30 tribes. Raw sequence reads were downloaded from the NCBI Short Read Archive (SRA) on March 29th, 2018 (**Appendix S1**). For the 501 ant UCE libraries, raw reads were cleaned with Trimmomatic v0.36 (Bolger, Lohse, & Usadel 2014) using the following parameters: LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:50. A reference database with the 29 complete mitochondrial genomes available for ants on GenBank at the time was constructed.

De novo assembly of mitogenomic and UCE data with MitoFinder

To extract mitogenomic data from UCE libraries, we developed a dedicated bioinformatic pipeline called MitoFinder (**Fig. 1**). This pipeline was designed to assemble sequencing reads from target enrichment libraries, assemble, extract, and annotate mitochondrial contigs. To evaluate the impact of assembler choice, contigs were assembled with IDBA-UD v1.1.1, MEGAHIT v1.1.3, and MetaSPAdes v3.13.0 within MitoFinder, and with Trinity v2.1.1 within PHYLUCE using default parameters.

Mitochondrial contigs were then identified by similarity search using blastn with e-value ≥ 1e-06 against our ant reference mitogenomic database. Each detected mitochondrial contig was then annotated with tblastx for protein-coding genes (CDS) and blastn for 16S and 12S rRNAs taking advantage of the geneChecker module of mitoMaker (Schomaker-Bastos & Prosdocimi, 2018) that we incorporated into MitoFinder. Finally, we used ARWEN v1.2 (Laslett & Canbäck, 2007) to detect and annotate tRNA genes.

Considering possible rearrangements in ant mitogenomes, each annotated mitochondrial CDS was first aligned with MAFFT v7.271 (Katoh & Standley, 2013) algorithm FFT-NS-2 with option *--adjustdirection*. Then, to take into account potential frameshifts and stop codons, mitochondrial CDS alignments were refined with MACSE v2.03 (Ranwez *et al.* 2018) with option -prog alignSequences, which produces both nucleotide and amino acid alignments. To improve alignment accuracy and reduce calculation time, we used sequences from available ant mitogenomes as references for each CDS (option *-seq_lr*). Sequences with internal stop codons were excluded to remove incorrectly annotated fragments potentially corresponding to nuclear mitochondrial DNA segments (NUMTs) in each protein-coding gene alignment. Then, individual gene alignments were checked by eye to manually remove remaining aberant sequences. Finally, a nucleotide supermatrix was created by concatenating protein-coding and ribosomal RNA genes. Since the mitochondrial signal might be saturated for inferring deep phylogenetic relationships, an amino acid supermatrix with the 13 mitochondrial CDSs was also assembled.

Guided iterative mitogenomic data assembly with MITObim

For comparison purposes, we also ran MITObim (Hahn, Bachmann & Chevreux 2013) to extract mitochondrial sequences from the 501 UCE raw sequencing data. This software is designed to assemble mitochondrial reads using mitochondrial bait such as the CO1 sequence of a related species when available. Then, based on iterative mapping, MITObim extends as much as possible the mitochondrial contig previously obtained. For each library, given the scarcity of closely related complete mitochondrial genomes available for ants, the longest CO1 sequence available for the genus, or the most closely related genus, was used as bait for the initial step of MITObim. As there is no annotation step in MITObim, MitoFinder was used to annotate the resulting MITObim contigs.

DNA barcoding

To verify species identification of the 501 ant UCE libraries, CO1 sequences extracted by MitoFinder using MetaSPAdes (mtDNA recovered for all species) were compared with Species Level Barcode Records (3,328,881 CO1 sequences including more than 100,000 ants) through the identification server of the Barcode Of Life Data System v4 (Ratnasingham & Hebert, 2007). The same CO1 sequences were also compared against the NCBI nucleotide database using Megablast with default parameters. An identification was considered to be confirmed when the query CO1 sequence had 95% similarity with a reference sequence in BOLD or GenBank with the same identifier.

Assembly of UCEs

As recommended by Faircloth (2016), we first relied on Trinity to assemble UCE contigs using the *phyluce_assembly_assemblo_trinity* module of PHYLUCE. To assess the impact of assembler choice on UCE loci retrieval, we also used the assemblies obtained with IDBA-UD, MEGAHIT, and MetaSPAdes as implemented in MitoFinder. PHYLUCE scripts *phyluce_assembly_get_match_counts* and *phyluce_assembly_get_fastas_from_match_counts* were used to match contigs obtained for each sample to the bait set targeting 2590 UCE loci for Hymenoptera (Branstetter *et al.* 2017b). The resulting alignments were then cleaned using Gblocks (Castresana 2000) with the *phyluce_align_get_gblocks_trimmed_alignments*

_from_untrimmed script. Finally, loci found in at least 75% of species were selected to create the four corresponding UCE supermatrices using the *phyluce_align_get_only_loci_with_min_taxa* script.

Phylogenetic analyses

Phylogenetic relationships of ants were inferred from a total of 16 different supermatrices corresponding to the four supermatrices constructed from contigs obtained with each of the four assemblers (IDBA-UD, MEGAHIT, MetaSPAdes, and Trinity). The four supermatrices are as follows: (i) a UCE nucleotide supermatrix built from the concatenation of UCE loci retrieved for at least 75% of species, (ii) a mitochondrial nucleotide supermatrix consisting of the concatenation of the 13 protein-coding genes and the two rRNA genes, (iii) a mitochondrial amino-acid supermatrix of the 13 protein-coding genes, and (iv) a mixed supermatrix of UCE nucleotides and mitochondrial amino-acid protein-coding genes. For all supermatrices, phylogenetic inference was performed with Maximum Likelihood (ML) as implemented in IQ-TREE v1.6.8 (Nguyen, Schmidt, von Haeseler, & Minh 2015) using a GTR+Γ_4+I model for UCE and mitochondrial nucleotide supermatrices, a mtART+Γ_4+I model partitioned by gene for mitochondrial amino acids matrices, and a partitioned model mixing a GTR+Γ_4+I model for UCE nucleotides and a mtART+Γ_4+I model for mitochondrial amino acids for the mixed supermatrices. Statistical node support was estimated using ultrafast bootstrap (UFBS) with 1000 replicates (Hoang, Chernomor, von Haeseler, Minh, & Vinh 2018). Nodes with UFBS values higher than 95% were considered strongly supported. For all supermatrices, the congruence among the different topologies obtained with the four assemblers was evaluated by calculating quartet distances with Dquad (Ranwez, Criscuolo, & Douzery 2010).

Table 1 | Summary statistics on assembly results according to the assembler used. The values are averages over the 501 assemblies, except for the assembly time, which is a median value. The two tables report specific statistics for A) ultraconserved elements data, and B) mitochondrial data. Note that 35 CPUs were used for Trinity whereas 5 CPUs were used for other assemblers.

A) Summary statistics for UCE assemblies and supermatrices

Assembler	Assembly time	UCEs				
		Number of contigs	Number of loci	Matrix size	%Variable sites	%Missing data
IDBA-UD (5 CPUs)	0h:11m:02s	30,544	2581	132,403	43.9	6.7
MEGAHIT (5 CPUs)	0h:12m:35s	114,392	2579	147,589	43.2	12.5
MetaSPAdes (5 CPUs)	0h:25m:42s	113,303	2582	156,456	44.3	6.1
Trinity (35 CPUs)	1h:06m:22s	43,481	2579	127,803	40.5	17.8

B) Summary statistics for mtDNA assemblies and supermatrices

Assembler	Mitogenomes								
	Number of contigs	Number of species	Number of genes	AA matrix size	% missing data	% Variable sites	NT matrix size	% missing data	% Variable sites
IDBA-UD (5 CPUs)	4.2	499	13.04	3764	20.9	86.7	13635	26.1	85.8
MEGAHIT (5 CPUs)	3.9	499	13.61	3757	15.3	87.5	13718	20.6	86.4
MetaSPAdes (5 CPUs)	3.8	501	13.73	3766	14.6	88.9	13713	19.8	87.1
Trinity (35 CPUs)	4.2	500	13.37	3760	18.0	86.9	13648	26.7	86.1

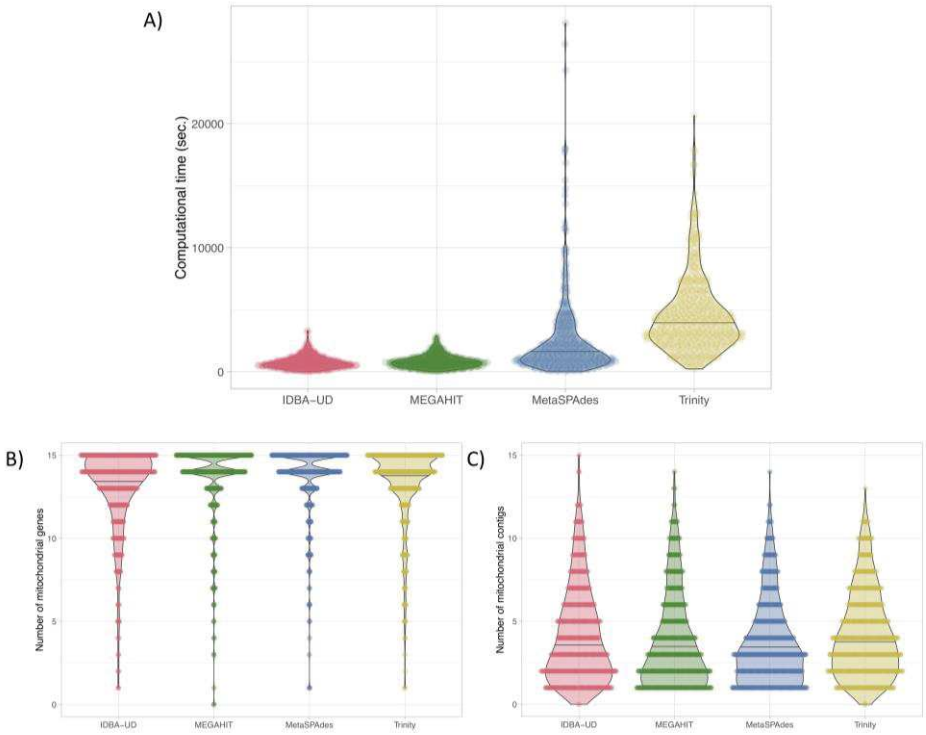


Figure 2 | Comparison of the efficiency of the assemblers in terms of: A) computational time, B) number of potentially mitochondrial contigs identified, and C) number of mitochondrial genes annotated. Violin plots reflect the data distribution with a horizontal line indicating the median. Note that for the three metagenomic assemblers, 5 CPUs were used compared to 35 CPUs for Trinity. Plots were obtained using PlotsOfData (Postma & Goedhart 2019).

Table 2 | Statistical comparison between the performances of the different assemblers. Statistical significance was estimated with a paired non parametric test (paired wilcoxon test). *** = p<0.001; ** = p<0.01; * = p<0.05; NS = p>0.05; and (+)/(-) is the result of the comparison between the row and the column.

Number of mtDNA contigs				
	IDBA-UD	MEGAHIT	MetaSPAdes	Trinity
IDBA-UD		** (+)	*** (+)	NS (−)
MEGAHIT	** (−)		* (+)	*** (−)
MetaSPAdes	*** (−)	* (−)		*** (−)
Trinity	NS (+)	*** (+)	*** (+)	

Number of coding mtDNA nucleotides				
	IDBA-UD	MEGAHIT	MetaSPAdes	Trinity
IDBA-UD		*** (−)	*** (−)	* (−)
MEGAHIT	*** (+)		* (−)	*** (−)
MetaSPAdes	*** (+)	* (+)		*** (+)
Trinity	* (+)	*** (−)	*** (−)	

Number of mtDNA genes				
	IDBA-UD	MEGAHIT	MetaSPAdes	Trinity
IDBA-UD		*** (−)	*** (−)	*** (−)
MEGAHIT	*** (+)		* (−)	*** (+)
MetaSPAdes	*** (+)	* (+)		*** (+)
Trinity	*** (+)	*** (+)	*** (−)	

Number of UCE nucleotides				
	IDBA-UD	MEGAHIT	MetaSPAdes	Trinity
IDBA-UD		*** (−)	*** (−)	*** (+)
MEGAHIT	*** (+)		*** (−)	*** (+)
MetaSPAdes	*** (+)	*** (+)		*** (+)
Trinity	*** (−)	*** (−)	*** (−)	

Results

Assembly of UCE datasets

De novo assembly of 501 UCE capture sequencing libraries was performed with four different assemblers: IDBA-UD, MEGAHIT, and MetaSPAdes via MitoFinder and Trinity via PHYLUCE. All assemblers provided different numbers of contigs (**Table 1**) ranging from 30,544 (IDBA-UD) to 114,392 (MEGAHIT) on average. The average computational time per assembly was highly variable among assemblers with Trinity being by far the slowest (35 CPUs, median time per sample: 1h:06m:22s, total time for all samples: 26.9 days) and IDBA-UD the fastest (5 CPUs, median time per sample: 0h:11m:01s, total time for all samples: 4.4 days), MEGAHIT (5 CPUs, median time per sample: 0h:12m35s, total time for all samples: 4.9 days) being slightly slower, and MetaSPAdes (5 CPUs, median time per sample: 0h:25m:44s, total time for all samples: 14.9 days) having a median assembly time about twice as slow as the other two metagenomic assemblers (**Table 1 & Fig. 2A**).

The UCE supermatrices created by PHYLUCE for each of the four assemblers contained on average 2580 out of the 2590 UCE loci for Hymenoptera (**Table 1**). All matrices contained 501 species, but the size of the supermatrix and the percentage of missing data varied depending on the assembler (**Table 1**). Trinity, which is generally used as the default assembler in PHYLUCE, resulted in the shortest and most incomplete supermatrix with 2579 loci representing 127,803 sites (40.5% variable) and 17.8% missing data. Among metagenomic assemblers, MetaSPAdes provided the largest and most complete supermatrix with 2582 loci representing 156,456 sites (44.5% variable) and only 6.0% missing data. IDBA-UD retrieved 2581 loci representing 132,403 sites (43.9% variable) with only 6.7% missing data, and MEGAHIT resulted in a supermatrix with 2579 loci representing 147,589 sites (43.2% variable) but with 12.4% missing data. Note that less than 30 loci were retrieved for *Phalacromyrmex fugax* (between 4 and 27 loci depending on the assembler). This is congruent with the original publication in which this low-quality library was not included in phylogenetic analyses (Branstetter *et al.* 2017a). Accordingly, we removed the *Phalacromyrmex fugax* library (SRR5437956) from the dataset.

Extracting mitochondrial sequences from UCEs sequencing data

Depending on the assembler used in MitoFinder, mitochondrial reads were recovered in 499, 500, and 501 libraries out of a total of 501 (**Table 1, Fig. 2B)**. Overall, mitochondrial signal thus was detected in all libraries but only MetaSPAdes retrieved it in all species (**Appendix S2**). On average, 3.8 contigs per species were identified (**Table 1, Table 2, Fig. 2B**) and 13.7 genes were annotated with MitoFinder (**Fig. 2C, Table 2**). In 296/501 cases, MitoFinder was able to assemble a contig of more than 15,000 bp containing at least 13 annotated genes that likely represents the complete mitochondrial genome. In 52 of these cases, all 15 genes were annotated. In the remaining cases, the putative mitogenome contigs were missing one or two genes, mostly the short and divergent ATP8 (131/296), the 12S rRNA (29/296), and the 16S rRNA (10/296), which were present but not directly annotated by our BLAST-based procedure. By comparison, MITObim produced a mitochondrial contig for only 358 libraries for which an average of 3.51 genes were annotated representing 2840.24 nucleotides on average.

After alignment and cleaning, mitochondrial genes obtained with MitoFinder were used to create nucleotide and amino acid supermatrices. To be consistent with UCE analyses, and despite the recovery of some mitochondrial signal, we ignored *Phalacromyrmex fugax* in further analyses. In the nucleotide supermatrices (13 protein-coding + 12S and 16S rRNAs), we obtained 13 genes on average per species, which resulted in

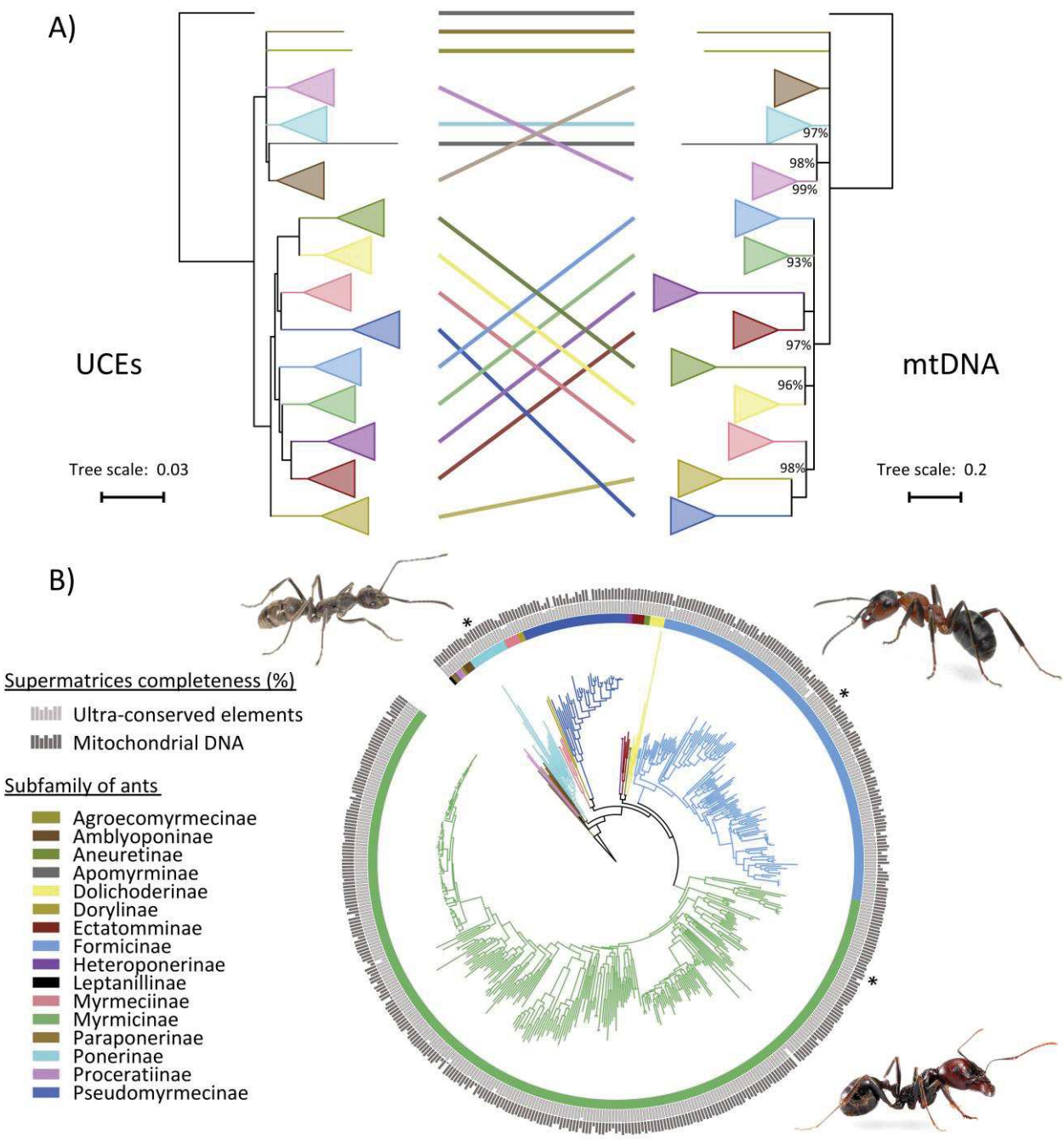


Figure 3 | Phylogenomic relationships of ants (Formicidae). A) Mito-nuclear phylogenetic differences among subfamily relationships based on the UCE and mtDNA supermatrices obtained with the assembler MetaSPAdes assembler. Clades corresponding to subfamilies were collapsed. Inter-subfamily relationships with UFBS < 95% were collapsed. Non-maximal node support values are reported. B) The topology obtained reflects the results of phylogenetic analyses based on the amino acid mitochondrial supermatrix (using MetaSPAdes as assembler). Histograms reflect the percent of UCEs (light grey) and mitochondrial genes (dark grey) recovered for each species. Illustrative pictures (*): *Diacamma sp*. (Ponerinae; top left), *Formica sp*. (Formicinae; top right), and *Messor barbarus* (Myrmicinae; bottom right).

supermatrices with 13,679 nucleotide sites (86.4% variable) and 23.3% missing data on average (**Table 1**). In the amino acid matrices (13 protein-coding genes), we obtained supermatrices with 3762 amino acid sites (87.4% variable) and 17.2% missing data on average (**Table 1**).

Barcoding analyses

A total of 534 CO1 sequences retrieved from the 501 MetaSPAdes assemblies were used to verify species identification of the UCE libraries (**Appendix S3**). Similarity searches against BOLD and Genbank allowed confirming the species identity in only 312 cases most likely because of the limited availability of CO1 barcoding data for these ant species. Moreover, in 42 cases, two or three CO1 sequence fragments were retrieved from the same UCE library. In seven of these cases, the slightly overlapping CO1 fragments most likely resulted from bad assembly or erroneous annotation. However, in the 35 remaining cases, the genuine complete CO1 sequence overlapped with shorter contigs assembled from a minority of the reads suggesting either cross-contaminations, nuclear mitochondrial DNA segments (NUMTs), endoparasites, or bacterial symbionts. For instance, in *Temnothorax* sp. mmp11 (SRR5809551), a 391-bp fragment annotated as CO1 by MitoFinder was found to be 98.2% identical to both the *Wolbachia pipientis* wAlbB and *Wolbachia* Pel strain wPip genomes, which are bacterial endosymbionts of the mosquitoes *Aedes albopictus* and *Culex quinquefasciatus*, respectively. Also, in *Sericomyrmex bondari* (SRR5044901) and *Sericomyrmex mayri* (SRR5044856) short CO1 fragments best matched with nematodes. However, in the 312 cases for which CO1 barcoding allowed to confirm the species identity of the UCE library, we did not detect any obvious cases of cross-contaminations where the CO1 extracted from a given library would have been identical to the one of another library (**Appendix S3**).

Phylogenetic results

The ML topologies inferred from the different UCE supermatrices were very similar with an average quartet distance of 0.005 among assemblers (**Appendix S4**). However, the percentage of supported nodes (UFBS > 95) differed depending on the assembler: IDBA-UD (91.37%), MetaSPAdes (89.96%), MEGAHIT (89.56%), and Trinity (85.85%). In the following, we only discuss the phylogenetic results obtained with MetaSPAdes that provides the most comprehensive assemblies for both UCE and mitochondrial data (**Table 1**). The following 12 well-established subfamilies were retrieved with maximal UFBS support (100%): Aneuretinae, Amblyoponinae, Dolichoderinae, Dorylinae, Ectatomminae, Formicinae, Heteroponerinae, Myrmeciinae, Myrmicinae, Pseudomyrmecinae, and Ponerinae (**Fig. 3A**). The two supergroups Formicoid and Poneroid were also retrieved with maximal UFBS support, as well as consensual phylogenetic relationships among Formicoid subfamilies (Ward 2014).

For mitochondrial matrices, the percentage of supported nodes (UFBS > 95) with nucleotides also differed depending on the assembler and was higher than with the amino acids: MetaSPAdes (84.5% *vs.* 80.1%), MEGAHIT (84.0% *vs*. 79.4%), Trinity (83.3% *vs*. 80.4%), and IDBA-UD (80.2% *vs.* 78.0%). However, ML mitogenomic trees inferred from amino acids were more congruent with UCE topologies than the ones inferred from the mitochondrial nucleotides (average quartet distance = 0.035 *v.s.* 0.063; **Appendix S4**). Among assemblers, the ML topologies inferred with amino acid matrices were highly congruent with an average quartet distance of 0.007 (**Appendix S4**). In the ML tree obtained with the MetaSPAdes supermatrix (**Fig. 3B**), all ant subfamilies were retrieved with maximal UFBS support values except for Myrmicinae (93%), Ponerinae (97%), and Proceratiinae (99%) (**Fig. 3A**). However, relationships among subfamilies were not congruent with UCE phylogenomic inferences except for Heteroponerinae + Ectatomminae (UFBS = 100) and

Dolichoderinae + Aneuretinae (UFBS = 96) (**Fig. 3A**).

Finally, phylogenetic inference carried on mixed supermatrices composed of UCEs and mitochondrial amino acids resulted in ML topologies that were also highly similar among assemblers with an average quartet distance of 0.006 (**Appendix S4**). The percentages of supported nodes (UFBS > 95) were: IDBA-UD (91.2%), MEGAHIT (92.8%), MetaSPAdes (92.2%), and Trinity (90.4%). As with UCE matrices, the 12 well-established subfamilies, the two supergroups Formicoid and Poneroid, and consensus Formicoid inter-subfamilies relationships (Ward 2014) were all retrieved with maximal UFBS support.

Discussion

Metagenomic assemblers as powerful tools for assembling UCEs

Currently, genomic and transcriptomic *de novo* assemblers are commonly used to assemble UCE loci from DNA capture sequencing data (Faircloth 2016). Since metagenomic assemblers such as IDBA-UD, MEGAHIT, and MetaSPAdes have been designed to account for variance in sequencing coverage, they seem to be well adapted for targeted enrichment or DNA sequence capture data. Our results show that metagenomic assemblers are indeed faster at assembling UCE loci than the classically used, but computationally intensive, Trinity transcriptomic assembler. Furthermore, they seem more effective and lead to datasets containing more variable sites, less missing data, and increased phylogenetic signal (**Table 1, Table 2**). Indeed, the topologies obtained with the metagenomic assemblers are very similar to the topology obtained with the Trinity-based supermatrix, contain a higher number of supported nodes (UFBS ≥ 95%), and are consistent with previous studies (Ward 2014). Furthermore, assemblies obtained with the three metagenomic assemblers provide variable numbers of contigs (ranging from 30,544 to 114,392) resulting in differences in the completeness of the matrices (6.0% to 17.8% of missing data for UCE matrices and 29.9% to 41.3% for mitochondrial matrices) and in numbers of variable sites (for UCE, 40.5% to 44.3%; for mtDNA, 77.2% to 79.0%). Interestingly, for both UCE matrices and mtDNA matrices, MetaSPAdes consistently provides more loci, more variable sites, and less missing data. In addition, mitochondrial signal was extracted from all libraries only using MetaSPAdes within Mitofinder. Despite a computation time on average twice that of the other two metagenomic assemblers, MetaSPAdes was the more effective assembler for ant UCEs. This software therefore provides a much-needed alternative to Trinity for efficiently assembling hundreds of UCE libraries.

MitoFinder efficiently extracts mitochondrial signal from UCE capture data

Ultraconserved elements are key loci exploited as target capture sequences in an increasing number of phylogenomic studies. DNA sequence capture methods are used to efficiently enrich targeted DNA regions in library preparation prior to sequencing, but non-targeted regions are always sequenced in the process resulting in so called “off-target reads”. Interestingly, off-target reads could represent up to 40% of the sequenced reads in exome capture experiments (Chilamakuri *et al*. 2014) and many contigs not belonging to targeted UCE loci are typically assembled from UCE capture data (e.g. Smith, Harvey, Faircloth, Glenn, & Brumfield, 2014; Faircloth *et al*. 2015). Given this high proportion of off-target reads, we can expect that mitochondrial DNA could be found as off-target sequences in many target enrichment data. Accordingly, several studies have succeeded in extracting mtDNA from UCE libraries (e.g. Smith *et al*. 2014; do Amaro *et al*. 2015). The development of MitoFinder allowed the automatic extraction of mitochondrial signal from all 501 ant UCE libraries. This maximum success rate indicates that this approach is highly efficient at least in Formicidae. However, the

success in retrieving mitochondrial sequences ultimately depends on the number of mitochondria contained in the tissue used for DNA extraction and library preparation. As expected, mitochondrial off-target reads are much more common in muscle and heart than in lung tissues in human (D'Erchia *et al.* 2015). Similarly, mitochondrial sequences are probably rare or absent in library constructed from vertebrate blood, even in birds in which nucleated red blood cells contain mitochondria, but in very low numbers (Reverter *et al.* 2016). In invertebrates, our case study with 100% success rate in ant UCEs demonstrates that mitochondrial sequences could probably be easily retrieved for many arthropod taxa as a by product of target enrichment sequencing experiments. Finally, the comparison between MitoFinder and MITObim emphasizes that the use of *de novo* assembly instead of iterative mapping is a suitable solution for recovering mitochondrial signal for groups with limited mitogenomic references.

The value of complementary mitochondrial signal

Mitochondrial sequences could provide interesting and important complementary information compared to nuclear sequences. First, mtDNA can be used to confirm the identity of the species sequenced for conserved UCE loci. Here, we were able to confirm the identification of 312 ant species out of the 501 UCE libraries using CO1 barcoding without revealing a single case of obvious species misidentification. Given that ant UCE libraries have been constructed from museum specimens, the 501 CO1 sequences we annotated could be used as reference barcoding sequences in future studies. Then, even though we did not detect such cases, the high mutation rate and the absence of heterozygous sites in mtDNA also make it well adapted for cross-contamination detection analyses (Ballenghien *et al*. 2017).

Nevertheless, mitochondrial markers also have some well-identified limitations (Galtier *et al*. 2009). First, mtDNA could be inserted in the nuclear genome in the form of NUMTs (Bensasson, Zhang, Hartl, & Hewitt 2001). NUMTs could potentially be assembled as off-target contigs in DNA capture libraries and we might have indeed extracted some fragments corresponding to NUMTs for the CO1 gene using MitoFinder (**Appendix S2**). Theoretically, NUMTs could be picked up by analysing the coverage of putative mitochondrial contigs as they are expected to have a coverage comparable to other off-targets nuclear contigs, whereas genuine mitochondrial contigs should have a higher coverage. A second limitation of mtDNA exists in arthropods where maternally inherited intra-cellular bacteria are frequent. Among those bacteria, *Wolbachia* is particularly widespread and could distort the mitochondrial genealogy when a particular strain spreads within the host species hitchhiking its linked mitochondrial haplotype (Cariou, Duret, & Charlat 2017). *Wolbachia* infection is frequent among ants and could therefore be responsible of some mito-nuclear discordance (Wenseleers *et al*. 1998). We indeed discovered such an instance with a *Wolbachia CO1* sequence identified in *Temnothorax* sp. mmp11 (SRR5809551), which was confirmed by several assembled contigs matching to *Wolbachia* strain genomes in this sample.

Beyond the methodological aspects of species identification and potential cross-contamination detection, mitochondrial sequences could also be useful to tackle fundamental evolutionary questions. UCEs have also proved to be useful genetic markers for phylogeography and for resolving shallow phylogenetic relationships (Musher & Cracraft 2018; Smith *et al*. 2014). In this context, mtDNA could also bring complementary information. In most animals, mtDNA has a maternal inheritance without recombination, which means that all mitochondrial genes behave as a single locus. This simplifies the interpretation of the phylogenetic pattern between closely related species or within subdivided populations of a species. Mito-nuclear phylogenetic discordance could also reveal interesting phenomena

involving hybridization, sex-biased dispersal, and introgression (Toews & Brelsford, 2012; Bonnet, Leblois, Rousset & Crochet 2017). In practise, hybridization events are often identified using mito-nuclear discordance (Li *et al* 2016) and in some cases, the mitochondrial introgression events have proven to be adaptive (Seixas, Boursot & Melo-Ferreira 2018). Nevertheless, in our ant case study, a detailed comparison of mitochondrial and UCE phylogenies did not reveal convincing occurrences of such discordances.

Ant phylogenetic relationships from 500 UCE and mitochondrial data

Both nuclear and mitochondrial data retrieved the most consensual phylogenetic relationships in the ant phylogeny (Ward 2014; Branstetter *et al.* 2017b; Borowiec *et al.* 2019). Twelve Formicidae subfamilies were recovered as monophyletic in all analyses, both with the nuclear and mitochondrial datasets, confirming their robustness. However, the well-defined inter-subfamily relationships within Formicoids (Ward 2014; Branstetter *et al.* 2017; Borrowiec *et al.* 2019) were only supported by the UCE dataset, but not by the mitochondrial amino acid dataset. For example, the army ant subfamily (Dorylinae) was not retrieved as the sister-group of all other Formicoids, but was the closest relative of Pseudomyrmicinae (UFBS = 100). Similarly, contradicting the classical and well-defined relationship of Heteroponerinae + Ectatomminae as the sister-group of Myrmicinae (Ward 2014; Branstetter *et al.* 2017; Borrowiec *et al.* 2019), the mitochondrial dataset supported an alternative relationship with Dolichoderinae + Aneuretinae (UFBS = 96). These differences suggest that mitochondrial data might be not well suited to resolve ancient phylogenetic relationships at the ant inter-subfamily level diverging about 100 Mya (Moreau, Bell, Vila, Archibald & Pierce 2006), even if they look suitable for more recent nodes such as intra-subfamily relationships.

Interestingly, these topological incongruences between UCEs and mitochondrial genes also featured different topologies regarding the existence of the Poneroid taxa, a controversial clade not always retrieved depending on the studies (Ward 2014), but that tends to be retrieved in the most recent studies (Branstetter *et al.* 2017; Borrowiec *et al.* 2019; UCE dataset in this study) and is not recovered by our mitochondrial amino acid dataset (**Fig. 3B**). The same applies to the phylogenetic placement of Apomyrminae, a subfamily either grouped with Leptanillinae or Amblyoponinae in past studies (Ward 2014), but that was grouped with Proceratiinae in our mitochondrial dataset (UFBS = 98; **Fig. 3B**). For such controversial nodes, our study demonstrates that the nature of the phylogenetic markers can provide different results. Such differences between nuclear and mitochondrial data might be due to the substitutional saturation of mitochondrial data even at the amino acid level. This problem may actually be exacerbated in hymenopteran mitochondria that possess high AT content translating into strongly biased codon usage potentially leading to phylogenetic reconstruction artefacts (Foster, Jermiin & Hickey 1997; Foster & Hickey 1999). Interestingly, such differences between mitochondrial and nuclear inference for ancient phylogenetic relationships, is not observed with insects with less AT-rich mitochondrial genomes such as swallowtail butterflies (Condamine, Nabholz, Clamens, Dupuis & Sperling, 2018; Allio *et al.* 2019) or tiger beetles (Vogler & Pearson 1996). This calls for additional studies on both controversial and consensual ant inter-subfamily relationships with more comprehensive genome-wide datasets.

Conclusions

In this study, we developed the MitoFinder tool to automatically extract and annotate mitogenomic data from raw sequencing data in an efficient way. For the assembly step of our pipeline, we tested four different assemblers and showed that MetaSPAdes is the most efficient and accurate assembler for both UCE and mitochondrial data. Applying MitoFinder to ants, we were able to extract mitochondrial signal from 501 UCE libraries. This demonstrates that mitochondrial DNA can be found as off-target sequences in UCEs sequencing data. Interestingly, mitochondrial DNA extracted from UCE libraries can also be used to: (i) confirm species identification with barcoding methods, (ii) highlight potential sample cross-contamination, and (iii) reveal potential cases of mito-nuclear discordance caused by hybridization events leading to mitochondrial introgression. Finally, MitoFinder was developed with UCE libraries but our approach should also work with data obtained from other capture methods in which numerous off-targets reads are sequenced, as well as with transcriptomic and whole genome sequencing data, in which mitochondrial reads are overrepresented.

Acknowledgements

This paper is dedicated to the memory of graduate student Alex Schomaker-Bastos (1992-2015) who was assassinated by the time he was writing the mitoMaker program on which we built upon for the annotation module of MitoFinder. We also thank Fabien Condamine and two anonymous reviewers for providing helpful comments on a previous version of the manuscript. This work has been supported by grants from the European Research Council (ERC-2015-CoG-683257 ConvergeAnt project) and Investissements d'Avenir of the Agence Nationale de la Recherche (CEBA: ANR-10-LABX-25-01; CEMEB: ANR-10-LABX-0004). This is contribution ISEM 2020-071 SUD of the Institut des Sciences de l'Evolution de Montpellier.

Data accessibility

The MitoFinder software is available from GitHub (https://github.com/RemiAllio/MitoFinder) and GitLab (https://gitlab.com/RemiAllio/mitofinder). Annotated mitogenomes and partial mitogenomic contigs containing at least 2 genes and 1,000 bp have been deposited in GenBank and are available in the Third Party Annotation Section of the DDBJ/ENA/GenBank databases under the accession numbers TPA: BK012118-BK012857 (**Appendix S5**). The full analytical pipeline, MitoFinder results including UCE and mtDNA contigs for all assemblers, phylogenetic datasets and corresponding trees can be retrieved from zenodo.org (DOI:10.5281/zenodo.3231390).

Authors' contributions

RA and FD conceived the ideas and designed methodology, analysed the data, and led the writing of the manuscript; RA implemented the MitoFinder software in part using code previously written by AS-B; JR, FP, and BN contributed to the writing of the manuscript. All authors contributed critically to the drafts and gave final approval for publication.

Supporting information

Appendix S1. List of the 501 UCE libraries (SRA accessions) and associated metadata.

Appendix S2. Summary statistics on mitochondrial signal recovered per species and depending on the assembler used. The table provides the number of contigs and genes recovered with MitoFinder and the size of each annotated gene.

Appendix S3. Summary statistics of barcoding analyses. Detailed results for both BOLDsystem and Megablast analyses are provided for each CO1 recovered with MitoFinder using MetaSPAdes.

Appendix S4. Detailed results of tree distance analyses realized with Dquad (Ranwez, Criscuolo, & Douzery 2010). Trees obtained with each assembler with mitochondrial amino acid supermatrix, mitochondrial nucleotide supermatrix, and UCE nucleotide supermatrix were compared with each others.

Appendix S5. List of Genbank accession numbers for newly generated mitchondrial contigs.

References

⇧Back to summary⇧

Allio, R., Scornavacca, C., Nabholz, B., Clamens, A. L., Sperling, F. A., & Condamine, F. (2019). Whole genome shotgun phylogenomics resolves the pattern and timing of swallowtail butterfly evolution. *Systematic Biology*, syz030. doi:10.1093/sysbio/syz030

Baca, S. M., Alexander, A., Gustafson, G. T., & Short, A. E. Z. (2017). Ultraconserved elements show utility in phylogenetic inference of Adephaga (Coleoptera) and suggest paraphyly of 'Hydradephaga'. *Systematic Entomology*, *42*(4), 786–795. doi:10.1111/syen.12244

Ballenghien, M., Faivre, N., & Galtier, N. (2017). Patterns of cross-contamination in a multispecies population genomic project: detection, quantification, impact, and solutions. *BMC Biology*, *15*(1), 25. doi:10.1186/s12915-017-0366-6

Bejerano, G., Pheasant, M., Makunin, I., Stephen, S., Kent, W. J., Mattick, J. S., & Haussler, D. (2004). Ultraconserved elements in the human genome. *Science*, *304*(5675), 1321–5. doi:10.1126/science.1098119

Bensasson, D., Zhang, D.-X., Hartl, D. L., & Hewitt, G. M. (2001). Mitochondrial pseudogenes: evolution's misplaced witnesses. *Trends in Ecology & Evolution*, *16*(6), 314–321. doi:10.1016/S0169-5347(01)02151-6

Bi, K., Vanderpool, D., Singhal, S., Linderoth, T., Moritz, C., & Good, J. M. (2012). Transcriptome-based exon capture enables highly cost-effective comparative genomic data collection at moderate evolutionary scales. *BMC Genomics*, *13*(1), 403. doi:10.1186/1471-2164-13-403

Blaimer, B. B., Brady, S. G., Schultz, T. R., Lloyd, M. W., Fisher, B. L., & Ward, P. S. (2015). Phylogenomic methods outperform traditional multi-locus approaches in resolving deep evolutionary history: a case study of formicine ants. *BMC Evolutionary Biology*, *15*(1), 271. doi:10.1186/s12862-015-0552-5

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, *30*(15), 2114–2120. doi:10.1093/bioinformatics/btu170

Bonnet, T., Leblois, R., Rousset, F., & Crochet, P. A. (2017). A reassessment of explanations for discordant introgressions of mitochondrial and nuclear genomes. Evolution, 71(9), 2140-2158.

Borowiec, M. L., Rabeling, C., Brady, S. G., Fisher, B. L., Schultz, T. R., & Ward, P. S. (2019). Compositional heterogeneity and outgroup choice influence the internal phylogeny of the ants. *Molecular Phylogenetics and Evolution*, *134*, 111–121. doi:10.1016/J.YMPEV.2019.01.024

Bourguignon, T., Lo, N., Šobotník, J., Ho, S. Y. W., Iqbal, N., Coissac, E., … Evans, T. A. (2016). Mitochondrial phylogenomics resolves the global spread of higher termites, ecosystem engineers of the tropics. *Molecular Biology and Evolution*, *34*(3), 589–597. doi:10.1093/molbev/msw253

Branstetter, M. G., Danforth, B. N., Pitts, J. P., Faircloth, B. C., Ward, P. S., Buffington, M. L., … Brady, S. G. (2017a). Phylogenomic Insights into the Evolution of Stinging Wasps and the Origins of Ants and Bees. *Current Biology*, *27*(7), 1019–1025. doi:10.1016/J.CUB.2017.03.027

Branstetter, M. G., Ješovnik, A., Sosa-Calvo, J., Lloyd, M. W., Faircloth, B. C., Brady, S. G., & Schultz, T. R. (2017b). Dry habitats were crucibles of domestication in the evolution of agriculture in ants. *Proceedings of the Royal Society B: Biological Sciences*, *284*(1852), 20170095. doi:10.1098/rspb.2017.0095

Branstetter, M. G., Longino, J. T., Ward, P. S., & Faircloth, B. C. (2017c). Enriching the ant tree of life: enhanced UCE bait set for genome-scale phylogenetics of ants and other Hymenoptera. *Methods in Ecology and Evolution*, *8*(6), 768–776. doi:10.1111/2041-210X.12742

Breinholt, J. W., Earl, C., Lemmon, A. R., Lemmon, E. M., Xiao, L., & Kawahara, A. Y. (2018). Resolving relationships among the megadiverse butterflies and moths with a novel pipeline for anchored phylogenomics. *Systematic Biology*, *67*(1), 78–93. doi:10.1093/sysbio/syx048

Caparroz, R., Rocha, A. V., Cabanne, G. S., Tubaro, P., Aleixo, A., Lemmon, E. M., & Lemmon, A. R. (2018). Mitogenomes of two neotropical bird species and the multiple independent origin of mitochondrial gene orders in Passeriformes. *Molecular Biology Reports*, *45*(3), 279–285. doi:10.1007/s11033-018-4160-5

Capella-Gutierrez, S., Silla-Martinez, J. M., & Gabaldon, T. (2009). trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics*, *25*(15), 1972–1973. doi:10.1093/bioinformatics/btp348

Cariou, M., Duret, L., & Charlat, S. (2017). The global impact of *Wolbachia* on mitochondrial diversity and evolution. *Journal of Evolutionary Biology*, *30*(12), 2204–2210. doi:10.1111/jeb.13186

Castresana, J. (2000). Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Molecular biology and evolution*, *17*(4), 540-552. doi:10.1093/oxfordjournals.molbev.a026334

Chilamakuri, C. S., Lorenz, S., Madoui, M.-A., Vodák, D., Sun, J., Hovig, E., … Meza-Zepeda, L. A. (2014). Performance comparison of four exome capture systems for deep sequencing. *BMC Genomics*, *15*(1), 449. doi:10.1186/1471-2164-15-449

Coissac, E., Hollingsworth, P. M., Lavergne, S., & Taberlet, P. (2016). From barcodes to genomes: extending the concept of DNA barcoding. *Molecular Ecology*, *25*(7), 1423–1428. doi:10.1111/mec.13549

Condamine, F. L., Nabholz, B., Clamens, A.-L., Dupuis, J. R., & Sperling, F. A. (2018). Mitochondrial phylogenomics, the origin of swallowtail butterflies, and the impact of the number of clocks in Bayesian molecular dating. *Systematic entomology*, *43*(3), 460-480. doi:10.1111/syen.12284

D'Erchia, A. M., Atlante, A., Gadaleta, G., Pavesi, G., Chiara, M., De Virgilio, C., … Pesole, G. (2015). Tissue-specific mtDNA abundance from exome data and its correlation with mitochondrial transcription, mass and respiratory activity. *Mitochondrion*, *20*, 13–21. doi:10.1016/J.MITO.2014.10.005

Di Franco, A., Poujol, R., Baurain, D., & Philippe, H. (2019). Evaluating the usefulness of alignment filtering methods to reduce the impact of errors on evolutionary inferences. *BMC Evolutionary Biology*, *19*(1), 21. doi:10.1186/s12862-019-1350-2

do Amaral, F. R., Neves, L. G., Resende Jr, M. F. R., Mobili, F., Miyaki, C. Y., Pellegrino, K. C. M., & Biondo, C. (2015). Ultraconserved Elements Sequencing as a Low-Cost Source of Complete Mitochondrial Genomes and Microsatellite Markers in Non-Model Amniotes. *PloS One*, *10*(9), e0138446. doi:10.1371/journal.pone.0138446

Dowton, M., & Austin, A. D. (1999). Evolutionary dynamics of a mitochondrial rearrangement 'hot spot' in the Hymenoptera. *Molecular Biology and Evolution*, *16*(2), 298–309. doi:10.1093/oxfordjournals.molbev.a026111

Dowton, M., Castro, L. R., & Austin, A. D. (2002). Mitochondrial gene rearrangements as phylogenetic characters in the invertebrates: the examination of genome 'morphology'. *Invertebrate Systematics*, *16*(3), 345. doi:10.1071/IS02003

Esselstyn, J. A., Oliveros, C. H., Swanson, M. T., & Faircloth, B. C. (2017). Investigating Difficult Nodes in the Placental Mammal Tree with Expanded Taxon Sampling and Thousands of Ultraconserved Elements. *Genome Biology and Evolution*, *9*(9), 2308–2321. doi:10.1093/gbe/evx168

Faircloth, B. C., McCormack, J. E., Crawford, N. G., Harvey, M. G., Brumfield, R. T., & Glenn, T. C. (2012). Ultraconserved Elements Anchor Thousands of Genetic Markers Spanning Multiple Evolutionary Timescales. *Systematic Biology*, *61*(5), 717–726. doi:10.1093/sysbio/sys004

Faircloth, B. C., Branstetter, M. G., White, N. D., & Brady, S. G. (2015). Target enrichment of ultraconserved elements from arthropods provides a genomic perspective on relationships among Hymenoptera. *Molecular Ecology Resources*, *15*(3), 489–501. doi:10.1111/1755-0998.12328

Faircloth, B. C. (2016). PHYLUCE is a software package for the analysis of conserved genomic loci. *Bioinformatics*, *32*(5), 786–788. doi:10.1093/bioinformatics/btv646

Faircloth, B. C. (2017). Identifying conserved genomic elements and designing universal bait sets to enrich them. Methods in Ecology and Evolution, 8(9), 1103-1112. doi:10.1111/2041-210X.12754

Foster, P. G., Jermiin, L. S., & Hickey, D. A. (1997). Nucleotide composition bias affects amino acid content in proteins coded by animal mitochondria. *Journal of Molecular Evolution*, *44*(3), 282–288. doi:10.1007/PL00006145

Foster, P. G., & Hickey, D. A. (1999). Compositional bias may affect both DNA-based and protein-based phylogenetic reconstructions. *Journal of Molecular Evolution*, *48*(3), 284-290. doi:10.1007/PL00006

Galtier, N., Nabholz, B., Glémin, S., & Hurst, G. D. D. (2009). Mitochondrial DNA as a marker of molecular diversity: a reappraisal. *Molecular Ecology*, *18*(22), 4541–4550. doi:10.1111/j.1365-294X.2009.04380.x

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., … Regev, A. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology*, *29*(7), 644–652. doi:10.1038/nbt.1883

Grummer, J. A., Morando, M. M., Avila, L. J., Sites, J. W., & Leaché, A. D. (2018). Phylogenomic evidence for a recent and rapid radiation of lizards in the Patagonian Liolaemus fitzingerii species group. *Molecular Phylogenetics and Evolution*, *125*, 243–254. doi:10.1016/J.YMPEV.2018.03.023

Guschanski, K., Krause, J., Sawyer, S., Valente, L. M., Bailey, S., Finstermeier, K., … Savolainen, V. (2013). Next-Generation Museomics Disentangles One of the Largest Primate Radiations. *Systematic Biology*, *62*(4), 539–554. doi:10.1093/sysbio/syt018

Hahn C., Bachmann L., Chevreux, B. (2013). Reconstructing mitochondrial genomes directly from genomic next-generation sequencing reads a baiting and iterative mapping approach. Nucleic Acids Research, 41(13), 129. doi:10.1093/nar/gkt371

Hassanin, A., Delsuc, F., Ropiquet, A., Hammer, C., Jansen van Vuuren, B., Matthee, C., … Couloux, A. (2012). Pattern and timing of diversification of Cetartiodactyla (Mammalia, Laurasiatheria), as revealed by a comprehensive analysis of mitochondrial genomes. *Comptes Rendus Biologies*, *335*(1), 32–50. doi:10.1016/J.CRVI.2011.11.002

Hoang, D. T., Chernomor, O., von Haeseler, A., Minh, B. Q., & Vinh, L. S. (2018). UFBoot2: Improving the Ultrafast Bootstrap Approximation. *Molecular Biology and Evolution*, *35*(2), 518–522. doi:10.1093/molbev/msx281

Ješovnik, A., Sosa-Calvo, J., Lloyd, M. W., Branstetter, M. G., Fernández, F., & Schultz, T. R. (2017). Phylogenomic species delimitation and host-symbiont coevolution in the fungus-farming ant genus Sericomyrmex Mayr (Hymenoptera: Formicidae): ultraconserved elements (UCEs) resolve a recent radiation. *Systematic Entomology*, *42*(3), 523–542. doi:10.1111/syen.12228

Katoh, K., & Standley, D. M. (2013). MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Molecular Biology and Evolution*, *30*(4), 772–780. doi:10.1093/molbev/mst010

Laslett, D., & Canback, B. (2007). ARWEN: a program to detect tRNA genes in metazoan mitochondrial nucleotide sequences. *Bioinformatics*, *24*(2), 172–175. doi:10.1093/bioinformatics/btm573

Le, S. Q., & Gascuel, O. (2008). An Improved General Amino Acid Replacement Matrix. *Molecular Biology and Evolution*, *25*(7), 1307–1320. doi:10.1093/molbev/msn067

Le, S. Q., Dang, C. C., & Gascuel, O. (2012). Modeling Protein Evolution with Several Amino Acid Replacement Matrices Depending on Site Rates. *Molecular Biology and Evolution*, *29*(10), 2921–2936. doi:10.1093/molbev/mss112

Lemmon, A. R., Emme, S. A., & Lemmon, E. M. (2012). Anchored Hybrid Enrichment for Massively High-Throughput Phylogenomics. *Systematic Biology*, *61*(5), 727–744. doi:10.1093/sysbio/sys049

Lemmon, E. M., & Lemmon, A. R. (2013). High-Throughput Genomic Data in Systematics and Phylogenetics. *Annual Review of Ecology, Evolution, and Systematics*, *44*(1), 99–121. doi:10.1146/annurev-ecolsys-110512-135822

Li, D., Luo, R., Liu, C.-M., Leung, C.-M., Ting, H.-F., Sadakane, K., … Lam, T.-W. (2016). MEGAHIT v1.0: A fast and scalable metagenome assembler driven by advanced methodologies and community practices. *Methods*, *102*, 3–11. doi:10.1016/J.YMETH.2016.02.020

McCormack, J. E., Faircloth, B. C., Crawford, N. G., Gowaty, P. A., Brumfield, R. T., & Glenn, T. C. (2012). Ultraconserved elements are novel phylogenomic markers that resolve placental mammal phylogeny when combined with species-tree analysis. *Genome Research*, *22*(4), 746–54. doi:10.1101/gr.125864.111

McCormack, J. E., Harvey, M. G., Faircloth, B. C., Crawford, N. G., Glenn, T. C., & Brumfield, R. T. (2013). A Phylogeny of Birds Based on Over 1,500 Loci Collected by Target Enrichment and High-Throughput Sequencing. *PLoS ONE*, *8*(1), e54848. doi:10.1371/journal.pone.0054848

Meiklejohn, K. A., Danielson, M. J., Faircloth, B. C., Glenn, T. C., Braun, E. L., & Kimball, R. T. (2014). Incongruence among different mitochondrial regions: A case study using complete mitogenomes. *Molecular Phylogenetics and Evolution*, *78*, 314–323. doi:10.1016/j.ympev.2014.06.003

Meza-Lázaro, R. N., Poteaux, C., Bayona-Vásquez, N. J., Branstetter, M. G., & Zaldívar-Riverón, A. (2018). Extensive mitochondrial heteroplasmy in the neotropical ants of the *Ectatomma ruidum*complex (Formicidae: Ectatomminae). *Mitochondrial DNA Part A*, *29*(8), 1203–1214. doi:10.1080/24701394.2018.1431228

Moreau, C. S., Bell, C. D., Vila, R., Archibald, S. B., & Pierce, N. E. (2006). Phylogeny of the ants: diversification in the age of angiosperms. Science, 312(5770), 101-104. doi:10.1126/science.1124891

Musher, L. J., & Cracraft, J. (2018). Phylogenomics and species delimitation of a complex radiation of Neotropical suboscine birds (Pachyramphus). *Molecular Phylogenetics and Evolution*, *118*, 204–221. doi:10.1016/j.ympev.2017.09.013

Nguyen, L.-T., Schmidt, H. A., von Haeseler, A., & Minh, B. Q. (2015). IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. *Molecular Biology and Evolution*, *32*(1), 268–274. doi:10.1093/molbev/msu300

Nurk, S., Meleshko, D., Korobeynikov, A., & Pevzner, P. A. (2017). metaSPAdes: a new versatile metagenomic assembler. *Genome Research*, *27*(5), 824–834. doi:10.1101/gr.213959.116

Peng, Y., Leung, H. C. M., Yiu, S. M., & Chin, F. Y. L. (2012). IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics*, *28*(11), 1420–1428. doi:10.1093/bioinformatics/bts174

Picardi, E., & Pesole, G. (2012). Mitochondrial genomes gleaned from human whole-exome sequencing. *Nature Methods*, *9*(6), 523–524. doi:10.1038/nmeth.2029

Pie, M. R., Ströher, P. R., Belmonte-Lopes, R., Bornschein, M. R., Ribeiro, L. F., Faircloth, B. C., & McCormack, J. E. (2017). Phylogenetic relationships of diurnal, phytotelm-breeding Melanophryniscus (Anura: Bufonidae) based on mitogenomic data. *Gene*, *628*, 194–199. doi:10.1016/J.GENE.2017.07.048

Pierce, M. P., Branstetter, M. G., & Longino, J. T. (2017). Integrative taxonomy reveals multiple cryptic species within Central American *Hylomyrma* FOREL, 1912 (Hymenoptera: Formicidae). *Myrmecological News*, *25*, 131–143. doi:10.25849/myrmecol.news_025:131

Prebus, M. (2017). Insights into the evolution, biogeography and natural history of the acorn ants, genus Temnothorax Mayr (hymenoptera: Formicidae). *BMC Evolutionary Biology*, *17*(1), 250. doi:10.1186/s12862-017-1095-8

Postma, M., & Goedhart, J. (2019). PlotsOfData—A web app for visualizing data together with their summaries. *PLoS biology*, *17*(3), e3000202.

Ranwez, V., Criscuolo, A., & Douzery, E. J. P. (2010). SuperTriplets: a triplet-based supertree approach to phylogenomics. *Bioinformatics*, *26*(12), i115-i123. doi:10.1093/bioinformatics/btq196

Ranwez, V., Douzery, E. J. P., Cambon, C., Chantret, N., & Delsuc, F. (2018). MACSE v2: toolkit for the alignment of coding sequences accounting for frameshifts and stop codons. *Molecular Biology and Evolution*, *35*(10), 2582–2584. doi:10.1093/molbev/msy159

Ratnasingham, S., & Hebert, P. D. N. (2007). BOLD: The Barcode of Life Data System (http://www. barcodinglife. org). *Molecular Ecology Notes*, *7*(3), 355–364. doi:10.1111/j.1471-8286.2007.01678.x

Reverter, A., Okimoto, R., Sapp, R., Bottje, W. G., Hawken, R., & Hudson, N. J. (2017). Chicken muscle mitochondrial content appears co-ordinately regulated and is associated with performance phenotypes. *Biology Open*, *6*(1), 50–58. doi:10.1242/bio.022772

Schomaker-Bastos, A., & Prosdocimi, F. (2018). mitoMaker: a pipeline for automatic assembly and annotation of animal mitochondria using raw NGS data. doi:10.20944/preprints201808.0423.v1

Seixas, F. A., Boursot, P., & Melo-Ferreira, J. (2018). The genomic impact of historical hybridization with massive mitochondrial DNA introgression. *Genome Biology*, *19*(1), 91. doi:10.1186/s13059-018-1471-8

Smith, B. T., Harvey, M. G., Faircloth, B. C., Glenn, T. C., & Brumfield, R. T. (2014). Target capture and massively parallel sequencing of ultraconserved elements for comparative studies at shallow evolutionary time scales. *Systematic Biology*, *63*(1), 83–95. doi:10.1093/sysbio/syt061

Starrett, J., Derkarabetian, S., Hedin, M., Bryson, R. W., McCormack, J. E., & Faircloth, B. C. (2017). High phylogenetic utility of an ultraconserved element probe set designed for Arachnida. *Molecular Ecology Resources*, *17*(4), 812–823. doi:10.1111/1755-0998.12621

Stephen, S., Pheasant, M., Makunin, I. V., & Mattick, J. S. (20087). Large-scale appearance of ultraconserved elements in tetrapod genomes and slowdown of the molecular clock. Molecular biology and evolution, 25(2), 402-408. doi:https://doi.org/10.1093/molbev/msm268

Ströher, P. R., Zarza, E., Tsai, W. L. E., McCormack, J. E., Feitosa, R. M., & Pie, M. R. (2017). The mitochondrial genome of *Octostruma stenognatha* and its phylogenetic implications. *Insectes Sociaux*, *64*(1), 149–154. doi:10.1007/s00040-016-0525-8

Tavaré, S. (1986). Some probabilistic and statistical problems in the analysis of DNA sequences. *Lectures on Mathematics in the Life Sciences*, *17*(2), 57–86.

Vieira, G. A., & Prosdocimi, F. (2019). Accessible molecular phylogenomics at no cost: obtaining 14 new mitogenomes for the ant subfamily Pseudomyrmecinae from public data. *PeerJ*, *7*, e6271. doi:10.7717/peerj.6271

Vogler, A. P., & Pearson, D. L. (1996). A molecular phylogeny of the tiger beetles (Cicindelidae): congruence of mitochondrial and nuclear rDNA data sets. *Molecular Phylogenetics and Evolution*, *6*(3), 321-338.

Vollmers, J., Wiegand, S., & Kaster, A.-K. (2017). Comparing and Evaluating Metagenome Assembly Tools from a Microbiologist's Perspective - Not Only Size Matters! *PloS One*, *12*(1), e0169662. doi:10.1371/journal.pone.0169662

Wang, N., Hosner, P. A., Liang, B., Braun, E. L., & Kimball, R. T. (2017). Historical relationships of three enigmatic phasianid genera (Aves: Galliformes) inferred using phylogenomic and mitogenomic data. *Molecular Phylogenetics and Evolution*, *109*, 217–225. doi:10.1016/J.YMPEV.2017.01.006

Ward, P. S. (2014). The Phylogeny and Evolution of Ants. *Annual Review of Ecology, Evolution, and Systematics*, *45*(1), 23–43. doi:10.1146/annurev-ecolsys-120213-091824

Ward, P. S., & Branstetter, M. G. (2017). The acacia ants revisited: convergent evolution and biogeographic context in an iconic ant/plant mutualism. *Proceedings of the Royal Society B: Biological Sciences*, *284*(1850), 20162569. doi:10.1098/rspb.2016.2569

Wenseleers, T., Ito, F., Van Borm, S., Huybrechts, R., Volckaert, F., & Billen, J. (1998). Widespread occurrence of the microorganism Wolbachia in ants. *Proceedings of the Royal Society B: Biological Sciences*, *265*(1404), 1447–1452. doi:10.1098/rspb.1998.0456

Young, A. D., Lemmon, A. R., Skevington, J. H., Mengual, X., Ståhls, G., Reemer, M., … Wiegmann, B. M. (2016). Anchored enrichment dataset for true flies (order Diptera) reveals insights into the phylogeny of flower flies (family Syrphidae). *BMC Evolutionary Biology*, *16*(1), 143. doi:10.1186/s12862-016-0714-0

Zarza, E., Faircloth, B. C., Tsai, W. L. E., Bryson, R. W., Klicka, J., & McCormack, J. E. (2016). Hidden histories of gene flow in highland birds revealed with genomic markers. *Molecular Ecology*, *25*(20), 5144–5157. doi:10.1111/mec.13813

Zarza, E., Connors, E. M., Maley, J. M., Tsai, W. L., Heimes, P., Kaplan, M., & McCormack, J. E. (2018). Combining ultraconserved elements and mtDNA data to uncover lineage diversity in a Mexican highland frog (Sarcohyla; Hylidae). *PeerJ*, *6*, e6045

⇧Back to summary⇧

– PART II –

Genomic dataset and phylogenomic approach

The first section of this chapter describes the methodological workflow developed to build a genome-scale dataset from rare or degraded DNA samples (e.g. roadkill samples) in order to study the evolutionary history of myrmecophagous mammals. Here, the main objective is to describe some of the methodological choices we made to construct the full dataset. We do not detail the complete list of softwares and pipelines used in the text, but they are represented in a schematic workflow associated with this section. The second section of this chapter illustrates one of the possible uses of this workflow with a case study on two ant-eating Carnivora species. More specifically, using non-optimal samples from roadkill specimens, we performed genome-wide species delineation to unravel phylogenetic uncertainties on the taxonomic status of the African subspecies of bat-eared fox (*Otocyon megalotis megalotis* and *Otocyon megalotis virgatus*) and aardwolf (*Proteles cristata cristata* and *Proteles cristata septentrionalis*). Finally, we constructed a phylogenomic dataset allowing us to reconstruct a robust phylogeny of Carnivora from more than 50 complete genomes.

Species	Basecalling mode	ID	Genome size (Gb)	Data Type		BUSCO (Mammalia = 4104)				Assembly stats					
				Illumina (Gb)	Nanopore Data (Gb)	Complete S %	Complete D %	Fragmented %	Missing %	num seqs	sum len	N50	avg len	max len	min len
Bradypus tridactylus	Guppy v.3.2.4 HAC Mode	V3450 (T7029)	3.4	367.9	41.7	83.7	2.2	6.6	7.5	15 237	3 244 835 702	684 256	212 958	9 093 477	1 134
Chlamyphorus truncatus	Guppy v 3.1.5 fast mode	CT1	3.3	304.3	32.9	68.3	3.1	14.9	13.7	101 658	3 213 424 224	64 202	31 610	690 676	1 046
	Guppy v.3.2.4 HAC Mode									36 203	3 231 634 670	247 690	89 264		
Cyclopes didactylus	Guppy v 3.1.5 fast mode	M2300	3.5	317	52.7	76.9	1.5	10.3	11.3	18 601	3 551 270 084	555 204	190 918	4 118 742	1 147
	Guppy v.3.2.4 HAC mode					82.6	1.7	6.9	8.8	14 869	3 578 339 820	795 956	240 658	6 147 388	1260
Myrmecophaga tridactyla	NA	BROAD	3.1		NA	83.5	0.9	6.6	9.9	1 621 408	3 547 273 050	41 255			
	Guppy v 3.1.5 fast mode	M3023		380.2	53					8 509	3 121 622 680	1 241 848	366 861	9 685 767	1 183
	Guppy v.3.2.4 HAC mode									6 911	3 133 159 297	1 482 581	453 358	16 125 692	1 264
Otocyon megalotis megalotis	Guppy v 3.1.5 fast mode	TS305	2.5	213	32.95	89.2	1.1	5.8	3.9	12 735	2 347 287 473	676 434	184 318	5 813 064	1 006
	Guppy v.3.1.5 HAC Mode									11 081	2 374 694 846	728 073	214 303		
	NA	TS306		258.4	NA	NA						NA			
Otocyon megalotis virgatus	NA	FMNH158128	2.5	198.5	NA	NA						NA			
Priodontes maximus	Guppy v.3.2.4 HAC Mode	M844	4.1	291.6	59.8					51 157	4 088 061 607	184 893	79 912		
Proteles cristatus	Guppy v 3.1.5 fast mode	TS307	2.5	215	27.5	91.5	0.3	4.4	3.8	8 874	2 388 075 386	699 305	269 209	4 198 179	1 169
	Guppy v.3.1.5 HAC Mode					90.2	0.5	5.4	3.9	5 669	2 388 965 834	1 308 801	421 408	10 217 545	2 153
	NA	TS491		179.9	NA	NA						NA			
Tamandua tetradactyla	NA	BROAD	3.2	NA	NA	54.7	1.4	22.1	21.8	2 144 204	3 849 374 830	19 789	1 794	386 785	200
	Guppy v 3.1.5 fast mode	M3075		376	51.3	80.8	2.7	6.9	9.6	7 611	3 245 460 668	1 329 139	426 417	9 778 874	1 334
	Guppy v.3.2.4 HAC Mode					84	3.2	4.8	8	6 275	3 256 447 370	1 581 541	518 956	13 146 806	2 143
	Guppy v.3.2.4 HAC Mode				100	84.7	2.8	4.7	7.8	4 309	3 318 049 110	3 102 936	770 028	33 898 748	1 283
Smutsia gigantea	Guppy v.3.2.4 HAC Mode	CAM011	2.5	200	31.7	82.7	0.9	9.9	6.5	24 429	2 463 847 862	227 038	100 857	2 852 167	1 185

Table 1 | Genomic dataset summary.

1 - Methodological workflow

As mentioned in the general introduction, when I started my PhD project 3 years ago, the objective of my thesis was to search, at the genomic scale, for molecular signatures of convergent evolution in independently evolved ant-eating mammals. This study was part of an ERC project - ERC ConvergeAnt - in which a large proportion of the fundings was dedicated to the sequencing and the assembly of nine high quality genomes for myrmecophagous mammals and close relatives (*Chlamyphorus truncatus*, *Priodontes maximus*, *Cyclopes didactylus*, *Myrmecophaga tridactyla*, *Tamandua tetradactyla, Bradypus tridactylus*, *Otocyon megalotis*, *Proteles cristata*, and *Smutsia gigantea*). Knowing the genome-size of myrmecophagous species (ranging from ~2.5 Gb to ~4 Gb) and the difficulty to obtain contiguous genomes for these species (available genomes from the BROAD presenting millions of contigs), we decided to opt for a hybrid sequencing strategy - combining short read and long read sequencing data. Unfortunately, the DNA quality of our samples was, at this time, under the DNA quality thresholds required by the different sequencing platforms specialized in long-read sequencing, such as PacBio (Rhoads & Au 2015) or 10X Genomics linked-reads (Zheng et al. 2016). In the meantime, Oxford Nanopore Technologies (ONT) was becoming popular. Originally designed to allow direct sequencing of DNA molecules with simplified library preparation procedures, ONT instruments such as the MinION (Jain et al., 2016) have been co-opted as a portable sequencing method in the field that proved useful in a diversity of environmental conditions (Blanco et al., 2019; Parker et al., 2017; Pomerantz et al., 2018; Srivathsan et al., 2018). With the objective to generate high quality genomes using long-reads sequencing and the support of our lab engineer, Marie-Ka Tilak, we decided to bring this new technology into ISEM laboratory and the first MinION sequencer was received on January 2018, just three months after the beginning of my PhD. This gave me the opportunity to follow and be closely involved in all the development of this technology from the wet lab work to the *in silico* basecalling and genome assembly. First, this section describes how we generated long-read sequencing data from roadkill samples (from which it is particularly challenging to obtain a large amount of high-quality DNA because of post-mortem degradation processes) using Oxford Nanopore Technologies directly in our lab using the MinION (and even in the field, in French Guiana). Then, it introduces the strategy - and the associated tools - used to generate high quality annotated genomes of ant-eating mammals. Finally, it presents how we generated genome-scale phylogenomic matrices from these *de novo* genome assemblies.

Figure 1 | Oxford Nanopore MinIT (left) and MinION (right) sequencing devices. This picture was taken during the sequencing of *Bradypus tridactylus* at the Paracou research station, French Guiana (2019). Screenshot of the MinKNOW GUI interface performed during a sequencing run.

1.a. Setting up long read sequencing at ISEM

The MinION is a pocket-sized sequencing device (**Fig. 1a**) which allows in-house, live sequencing. The first step of the sequencing procedure is to load the DNA library in the flow-cell (i.e. the part of the MinION device containing the nanopores). Then, the sequencing process can be launched and controlled using the MinKNOW interface on the connected computer **(Fig. 1b**). In addition, the MinKNOW software allows users to follow the sequencing process live (quantity of data generated, flowcell status, etc...). On average, a sequencing run takes two days and can generate up to 30Gb with theoretically no limit of DNA sequence fragment size. Indeed, if the DNA fragments are not degraded, the nanopore technology allows the sequencing of DNA molecules in their entirety. However, because the sequencing is done by measuring the disruption of ionic current when the DNA strand passes through a nanopore, any contaminant may lead to the dysfonction of this technology. In addition, even though the entire sequence is read, degraded DNA molecules (fragmented or containing nicks) lead to shorter sequencing reads and lower yields per sequencing run. In our case, when we decided to use ONT in our laboratory, we wanted to sequence DNA extracted from roadkill samples. Unfortunately, roadkill tissues contain necrotized cells and impurities that, combined with post mortem degradation processes, result in bad quality and degraded DNA extractions (**Fig. 2**). Hence, to be able to generate long read data from these samples, we had to develop an optimized protocol for sequencing mammalian roadkill tissues with ONT. To do so, we tried to improve each step of the Oxford Nanopore protocol, from sampling to sequencing.

First, when sampling roadkill tissues, it is important to collect the tissues as fresh as possible. Hence, given that most roadkill accidents happen by night, it is better to do the sampling early in the morning or at dusk, when the ambient temperature is still cool and DNA has not significantly degraded yet. In our experience, earlobe constitutes one of the best tissues to sample as it dries rather quickly after death. Additionally, we observed that DNA preservation was generally better in RNAlater than in 95% EtOH preserved tissues (**Fig. 2**). To better preserve the samples, it is also strongly recommended to cut the extracted samples into small pieces to allow the RNAlater to penetrate the tissues. Second, we found that physically removing necrotized and epidermal cells before DNA extraction resulted in better DNA quality and purity (**Fig. 2**). In practice, Marie-Ka Tilak used a binocular magnifier to check the tissues before and after the extraction to ensure that no hair or dust particle remnants were included in the library preparation. Finally, during library preparation, there was a step to remove contaminants such as salts, adapters, or nucleotides (dNTPs). In this step, AMPure beads were used to specifically extract the DNA fragments. Adjusting the ratio of AMPure beads used at this step allows to specifically extract the longest DNA fragments. Indeed, if 1x is the ratio used to extract the totality of the DNA in the preparation, using a ratio of 0.4x leads to the extraction of less DNA fragments. Given that the longest DNA fragments cling better to the beads than shorter fragments, they are preferentially extracted. In our case, 0.4x proved to be the best ratio

to use to specifically extract the longest fragments of the preparation without losing too much DNA (**Fig. 2**).

By implementing these optimization steps, we were able to significantly increase both read length and throughput per flow cell for roadkill samples sequenced on the ONT MinION device (**Fig. 1a**). Indeed, for example, this protocol (freely available here: https://www.protocols.io/view/an-optimized-protocol-for-sequencing-mammalian-roa-beixjcfn) increased the median size of the sequenced raw DNA fragments from 1.536 Kb to 4.857 Kb for the pygmy anteater (*Cyclopes didactylus*). Given these results, this protocol was used for the nine mammalian species of my PhD project. Finally, after 126 flow cells used for both developing the protocol and sequencing the nine mammalian genomes of the project, raw sequencing data, representing about 385 Gb, were generated in 20 months (September 2019), thanks to the huge effort of Marie-Ka Tilak.

To conclude, three years after the beginning of my PhD, several projects based on ONT sequencing have been started in our institute. We now have, just in our team, several MinIONs, the MinIT, the new MinION mk1c device and even a GridION (5 flowcells simultaneously) instrument. From my point of view, Oxford Nanopore Technologies instruments, with their flexibility and relatively affordable costs, are paving the way for ecology and evolution laboratories to sequence their favorite non-model species and join the modern genomic era.

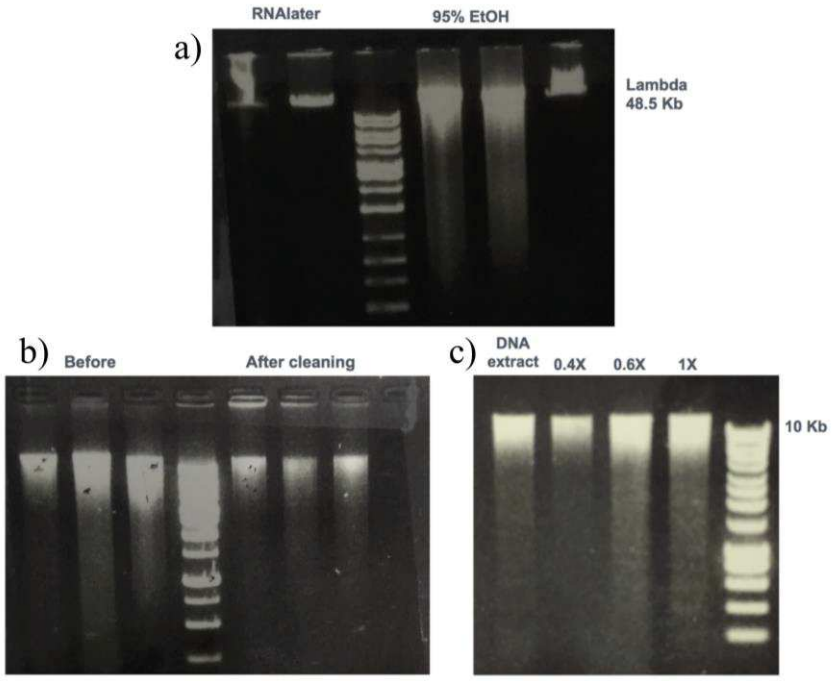


Figure 2 | Illustration of the effect of a) using either 95% EtOH or RNAlater for tissues conservation, b) removing hairs and epidermal cells before extraction, and c) using different ratios of AMPure beads for DNA size selction.

1.b. Long read processing and hybrid assembly

During sequencing with Oxford Nanopore devices, the DNA molecule is sequenced live when one of the DNA strands goes into the nanopore. The disruption of the ionic current caused by each nucleotide of the molecule is measured and saved in specific files called FAST5 files. The first bioinformatic step, called "basecalling", consists in converting this raw electronic signal into sequencing reads (i.e. in DNA sequence fragments). The most commonly used tool to perform this step is the Guppy basecaller developed by Oxford Nanopore. However, several other tools have been developed to try to improve this step such as Nanocall (David et al 2017), DeepNano (Boza et al 2017), causacall (Zeng et al. 2020), and Bonito (https://github.com/nanoporetech/bonito/) or Fast-Bonito (Xu et al. 2020). Indeed, the conversion of the electronic signal into the corresponding DNA sequence is not an easy task and leads to a relatively high level of sequencing errors (about 15% when I started my PhD). Luckily, given that the sequencing errors are essentially caused by the interpretation of FAST5 files, it is possible to basecall old FAST5 files with more recent basecallers to improve the accuracy of the resulting sequencing data. For example, two different basecallings were done for most of our genomes. Indeed, a new version of the Guppy software was released by Oxford Nanopore during my PhD, which remarkably improved basecalling by including a *high accuracy* mode. This new GPU-optimized version was first tested on the aardwolf genome and the improved results encouraged us to (re-)run the basecalling for all nine genomes of the project using GPU machines of the Montpellier Bioinformatics Biodiversity platform (https://mbb.univ-montp2.fr/MBB/). Specifically, the new high accuracy mode increased the read quality, which consequently led to better genome contiguity. Overall, during my PhD, the improvement of the basecalling increased the sequencing accuracy from 15-20% sequencing errors to only 4% in the latest version of Guppy (version 4.2.2). The developers of Bonito even talk about 2% errors with their latest update (version 0.3.0), which is scheduled to eventually replace Guppy as the default ONT basecaller. However, even if the improvement in accuracy of the basecalling may encourage continually improving the data, the computational time and the quantity of raw data to manage for rerunning the analyses make it complicated in practice. Indeed, for each genome sequenced in our lab, around two weeks were necessary to convert raw FAST5 files to FASTQ files using GPU-equipped computers.

The relatively high level of sequencing errors associated with Oxford Nanopore Technologies, at least at the beginning of my PhD project, can be compensated by sequencing at a high depth of coverage to avoid sequencing errors in *de novo* genome assembly and thus obtain reference genomes with high base accuracy, contiguity, and completeness (Koren et al., 2017; Shafin et al., 2020; Vaser et al., 2017). Otherwise, it is possible to correct errors in ONT long reads by combining them with Illumina short reads (known to be much more accurate), either to polish *de novo* long read-based genome assemblies (Batra et al., 2019a; Jain et al., 2018; Nicholls et al., 2019;

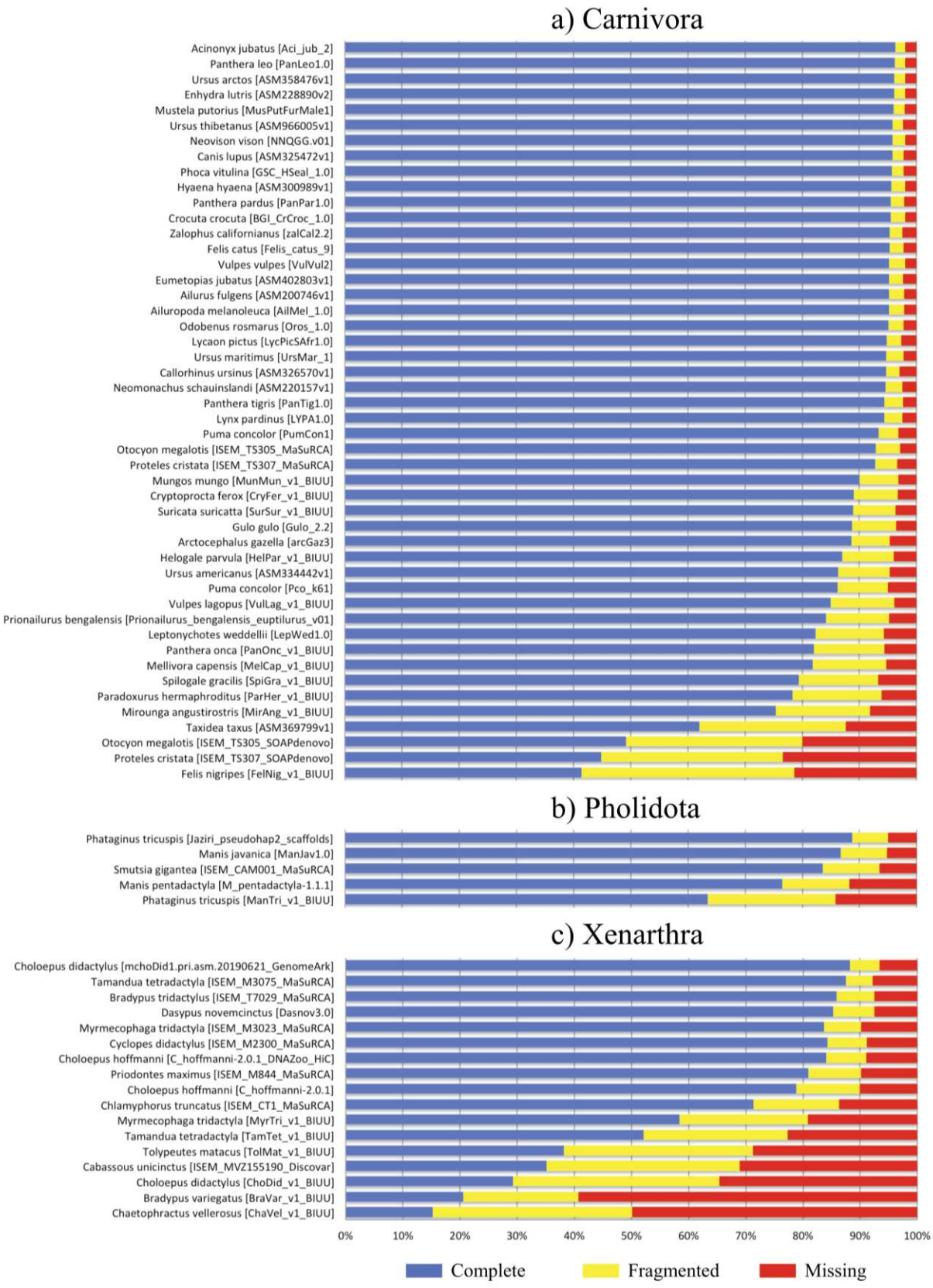


Figure 3 | BUSCO scores obtained for the genomes generated for the project.

Walker et al., 2014) or to construct hybrid assemblies (Di Genova et al., 2018; Gan et al., 2019; Tan et al., 2018; Zimin et al., 2013). In hybrid assembly approaches, the accuracy of short reads with high depth of coverage (50-100x) allows the use of long reads at lower depth of coverage (10-30x) essentially for scaffolding (Armstrong et al., 2020; Kwan et al., 2019). Because we were using roadkill samples, for which obtaining high quality DNA for long read sequencing was complicated, we naturally decided to use a hybrid assembly approach. To do so, an interesting method, developed by Zimin et al. (2017, 2019) and implemented in MaSuRCA, proposed to combine short and long reads by first, assembling the short reads in longer highly accurate "super reads" and then assemble these reads in a second step with the long reads, which are essentially used for scaffolding. Even if this method had not yet been tried in mammals when I started my PhD, it has since produced promising results in plants (Scott et al., 2020; Wang et al., 2020; Zimin et al., 2017), birds (Gan et al., 2019), and fishes (Jiang et al., 2019; Kadobianskyi et al., 2019; Tan et al., 2018). In that context, the nine genomes of the project were assembled using short and long reads using the MaSuRCA pipeline. This computationally intensive approach required having access to significant computing power over long periods and was performed at both the ABiMS platform at the Roscoff Biological Station (http://abims.sb-roscoff.fr/) and the MESO@LR supercomputing infrastructure (https://meso-lr.umontpellier.fr/) hosted by the université de Montpellier. Indeed, the assembly step for the different genomes of the project lasted from 3-4 weeks for the smallest genomes to 51 days for the giant armadillo (*Priodontes maximus*) with an estimated genome size of 4.1 Gb.

To evaluate the quality of our genomes, we assessed genome contiguity (number of contigs and N50 values) and genome completeness based on reference single-copy orthologs (Mammal database, BUSCO v3, Waterhouse et al. 2018). The newly generated genomes of our project are high quality genomes with N50 values ranging from 185 Kb to 3.1 Mb and containing between 51,157 to 4,309 contigs (**Table 1**). Regarding gene completeness, our genomes rank among the best genomes for Xenarthra (**Fig. 3a**), Pholidota (**Fig. 3b**) and Carnivora (**Fig. 3c**). Overall, we show that hybrid assembly is a suitable strategy to generate high quality mammalian genomes from degraded samples.

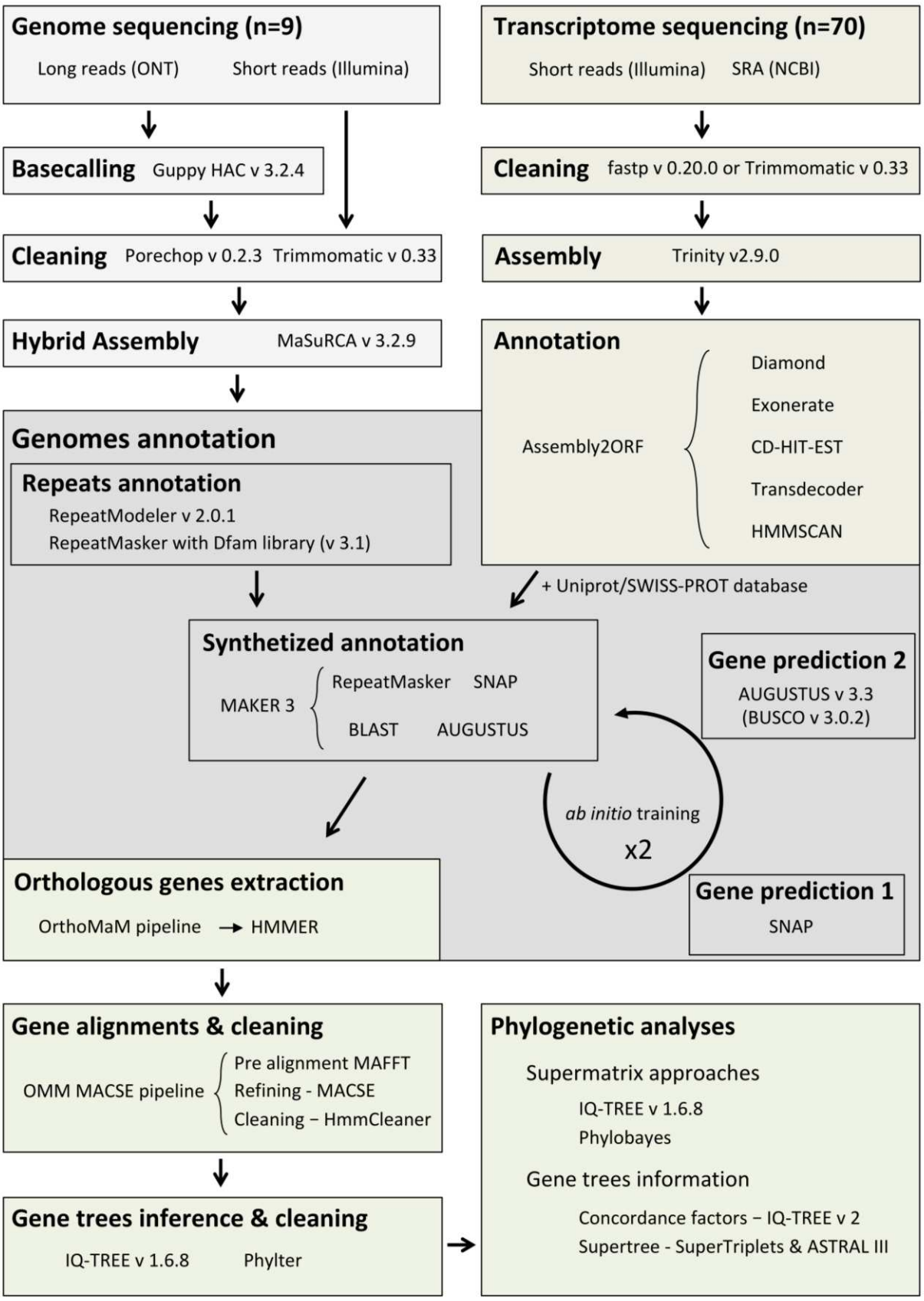


Figure 4 | Schematic workflow for the sequencing, assembly and annotation of mammal genomes.

1.c. *De novo* genome annotation

Genomic datasets allow addressing a number of evolutionary questions. In my case, the ultimate objective of my PhD project was to look for evidence of convergent evolution in orthologous genes in mammals. In that sense, after having assembled the nine focal genomes of the project, the next step consisted in annotating these genomes. Instead of limiting the annotation to orthologous genes, we decided to produce the best possible *de novo* annotation to make these genomes useful for the other parts of the ERC project (e.g. investigating gene family evolution and transposable elements composition).

Gene annotations might be done by using (i) *ab initio* gene prediction models based on general gene features (Korf 2004; Stanke et al. 2006) or (ii) homology-based gene prediction using sequence similarities with either transcriptomic information or protein databases (e.g. She et al. 2011). Recently, several approaches have been developed to combine both *ab initio* and homology-based gene prediction (Holt & Yandell 2011; Testa et al - 2015 - BMC Genomics; Hoff et al. 2016 - Bioinformatics, Kellwagen et al - 2018 - BMC bioinformatics). In our case, we decided to take advantage of every strategy thanks to the pipeline implemented in MAKER 3 (Holt & Yandell 2011). This pipeline is specifically designed to summarize the information obtained from different annotation analyses and automatically synthesize these data into gene annotations having evidence-based quality values. Additionally, MAKER is also easily trainable with the outputs of preliminary runs being usable to automatically retrain its gene prediction algorithm, producing higher quality gene-models on subsequent runs. In the next paragraphs, I will develop the strategy used to annotate our nine genomes by taking advantage of the MAKER pipeline (inspired by both DNAzoo strategy: Announcing the release of updated genome annotations; and a GitHub post by Daren C. Card: darencard/maker_genome_annotation.md).

First of all, a very important step of genome annotation is identifying repetitive contents. In fact, protein-like transposable elements have to be identified before running MAKER annotation to avoid misannotation. Left unmasked, repeats can seed millions of spurious BLAST alignments, producing false evidence for gene annotations (Yandell & Ence, 2012). Additionally, many transposon open reading frames (ORFs) look like true host genes to gene predictors, causing portions of transposon ORFs to be added as additional exons to gene predictions, completely corrupting the final gene annotations (Yandell & Ence, 2012). Even if this step can be done directly by MAKER, to optimise the annotation of repeats, we decided to do it ourselves using two complementary approaches. The first one consisted in *de novo* identifying repeats from each genome separately using RepeatModeler (Smit & Hubley 2008). Given that *de novo* identified libraries can include highly conserved protein-coding genes, such as histones and tubulins, every library was subsequently cleaned by removing protein-like sequences. To improve the accuracy of these *de novo* annotations, the libraries obtained from the different genomes were clustered for further analyses. The second step

was to identify repeat elements by similarity with publicly available existing libraries of mammalian repeats (DFAM, Wheeler et al. 2012) using RepeatMasker (Tarailo-Graovac & Chen 2009). The annotations resulting from these two steps were synthesized in a GFF file to be fed in MAKER.

Then, to improve the gene annotation, we decided to rely on transcriptomic information using publicly available and/or newly generated RNAseq data for our focal species. Indeed, RNA-seq data have the greatest potential to improve the accuracy of gene annotations, as these data provide copious evidence for better delimitation of exons, splice sites, and alternatively spliced exons (Yandell & Ence, 2012). To benefit from this advantage, 70 transcriptomes were annotated with an adapted version of assembly2ORF (https://github.com/ellefeg/assembly2orf), which is specifically designed to annotate transcriptomes. This pipeline relies on evidence-based gene predictions to extract and annotate gene CDS from transcriptome assemblies. For the species for which several tissues were sequenced, RNA-seq data were assembled individually but the CDS resulting from the annotation were concatenated and clustered by similarity to improve the efficiency of MAKER annotation. For each genome, the CDS obtained from RNA-seq were fed to MAKER to help the evidence-based gene prediction. Additionally, the manually annotated, non-redundant protein sequence database Uniprot/SWISSPROT (Bairoch & Apweiler - Nucl Acids Res - 2000, UniProt Consortium - Nucl Acids Res - 2010) was fed to MAKER for the annotation.

Once repeat elements are identified, CDS from diverse RNA-seq data extracted, and non-redundant protein sequence databases obtained, it is possible to run an initial annotation with MAKER 3. In this step, MAKER computes evidence-based gene predictions using sequence similarities with the CDS extracted from both the Swiss-prot database and the transcriptomes. Given that this annotation is mostly based on sequence similarity, some genes can be missed or improperly annotated. One of MAKER's interests is that it can be run iteratively, using the gene models from the one round to train *ab initio* software to improve the inference of gene models in the next round. To improve this annotation, it is thus recommended to use *ab initio* gene predictors such as SNAP (Korf 2004) and Augustus (Stanke et al. 2006; via BUSCO v3, Waterhouse et al. 2018) to optimize the gene models of the first annotation done by MAKER. Then, another round of MAKER was run, but this time with SNAP and Augustus running within MAKER to help create more sound gene models. Indeed, in doing this, MAKER uses the annotations from the two prediction programs in addition to the evidence-based gene predictions (similarities with reference CDS) when constructing its models. For our genome annotation, a couple of rounds of *ab initio* software training and MAKER annotation were done (**Fig. 4**).

1.d. Orthologous gene identification and extraction

Once the genomes were annotated, the next step before being able to perform phylogenetic inference was to extract mammalian single copy orthologous genes (orthologs 1:1). Even though several tools exist to identify ortholog groups from a set of genes from different species (e.g. OrthoFinder2, Emms & Kelly 2018; OrthoMCL, Li et al. 2003), we decided to rely on the Orthologous Mammalian Markers database (OrthoMaM v10; Scornavacca et al. 2019) to identify the single-copy orthologous genes from our nine genomes. To do so, we followed the orthology delineation process of the OrthoMaM database. First, for each orthologous gene alignment of OMM, a HMM profile was created via *hmmbuild* of the HMMER toolkit (Eddy, 2011) using default parameters and all HMM profiles were concatenated and summarized using *hmmpress* to construct a HMM database. Then, for each CDS newly annotated by MAKER, *hmmscan* was used on the HMM database to retrieve the best hits among the orthologous gene alignments. For each orthologous gene alignment, the most similar sequences for each species were detected via *hmmsearch*. Outputs from *hmmsearch* and *hmmscan* were discarded if the first hit score was not substantially better than the second ($hit_2 < 0.9\ hit_1$). This ensured our orthology predictions for the newly annotated CDSs to be robust.

1.e. Orthologous gene alignments and filtering

To infer the phylogeny of a group, the first thing to do after having extracted the single copy orthologous genes is to align these genes independently. This step is crucial and should be carried-out in the best possible way because of its many possible impacts on downstream phylogenetic inferences (Morrison 2006; Ranwez & Chantret 2020). In that context, to create the most accurate alignments as possible, we rely on the procedure of the OrthoMaM database implemented in the OMM_MACSE pipeline (Ranwez et al. 2020). This procedure consists in (i) aligning orthologous genes using both MAFFT (Katoh and Standley, 2013) and MACSE v2 (Ranwez et al. 2018), and (ii) filtering non homologous sequences, and masking erroneous/dubious part of gene sequences with HMMcleaner (Di Franco et al., 2019). One particular interest of this approach is the implementation of MACSE v2. This software is specifically designed to handle frameshifts in protein-coding nucleotide alignments. Finally, to exclude misidentified orthologous genes, gene tree inferences were conducted for each orthologous gene alignments using IQ-TREE v1.6.8 (Nguyen et al. 2015) and sequences leading to abnormally long branches were detected and iteratively removed using PhylteR (https://github.com/damiendevienne/phylter).

1.f. Phylogenetic inferences

Once the single-copy orthologous genes are extracted, aligned and cleaned, phylogenetic inference from orthologous alignments can be done either by concatenating the multiple alignments together into one super-alignment, and then estimating a tree on the super-alignment (called the supermatrix approach), or using information from each alignment independently and summarising gene tree inferences in one final phylogenetic tree (called gene tree/species tree reconciliation approach). In supermatrix approaches the quantity of homologous sites allows parameter-rich substitutions models that could be more realistic (e.g., the site-heterogeneous models, Lartillot et al. 2004, 2007; Lartillot 2020) to be optimized thanks to the large amount of information considered. Indeed, some key elements such as the heterogeneity in nucleotide composition or in the evolutionary rate of the different genes must be taken into account to avoid tree reconstruction artifacts such as long-branch attraction (Lartillot et al. 2007; Philippe et al. 2017; Simion et al. 2017). However, this approach considers that all orthologous genes share the same evolutionary history leading to the same global topology. This could not be the case if hybridization events, horizontal gene transfers, or incomplete lineage sorting processes occur (Edwards 2008). In practice, with the burgeoning of genome-scale supermatrices, containing thousands of genes, several studies (e.g. Jeffroy et al. 2006; Kumar et al 2012) have pointed out that high node support values obtained with supermatrix approaches can hide statistically significant incongruences at the gene level. Also, concatenation can be statistically inconsistent with respect to incomplete lineage sorting (ILS, Roch and Steel 2015). In that context, using gene-tree/species-tree inference approaches based on the multispecies coalescent (Bryant & Hahn 2020; Rannala et al. 2020) allows to study the evolution of genes at a finer scale and take into account the independent evolution of genes in the evaluation of the robustness of a species tree (Edwards et al. 2007, Mirarab et al. 2014). However, in some cases, phylogenetic errors either due to bioinformatics errors, phylogenetic information weakness of independent loci can exaggerate the importance of ILS by inflating gene tree versus species tree discordance (Gatesy and Springer, 2014; Springer and Gatesy, 2016). In this context, a particularly interesting approach consists in measuring the respective concordance between alignment sites and gene trees with the species tree inferred from the supermatrix approach (Minh et al. 2020). By using information from both species tree and gene tree inferences, this approach allows to point out well-supported versus poorly-supported nodes in phylogenetic trees. During my PhD, I unfortunately didn't have the chance to apply this strategy on a dataset including the new genomes of myrmecophagous species due to limited time, but this approach was conducted in two others phylogenetic studies and in both cases, it allows to identify interesting nodes of the phylogeny at which underlying processes created incongruent gene trees (see Part 2.2 and Annexe 1 for more details).

References

⇧Back to summary⇧

Armstrong EE, Taylor RW, Miller DE, Kaelin CB, Barsh GS, Hadly EA, Petrov D. 2020. Long live the king: chromosome-level assembly of the lion (Panthera leo) using linked-read, Hi-C, and long-read data. *BMC Biol* **18**:3. doi:10.1186/s12915-019-0734-5

Bairoch A, Apweiler R. 2000. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res* **28**:45–48. doi:10.1093/nar/28.1.45

Batra SS, Levy-Sakin M, Robinson J, Guillory J, Durinck S, Kwok P-Y, Cox LA, Seshagiri S, Song YS, Wall JD. 2019. Accurate assembly of the olive baboon (*Papio anubis*) genome using long--read and Hi-C data. *bioRxiv* 678771. doi:10.1101/678771

Blanco MB, Greene LK, Williams RC, Andrianandrasana L, Yoder AD, Larsen PA. 2019. Next-generation in situ conservation and educational outreach in Madagascar using a mobile genetics lab. *bioRxiv* 650614. doi:10.1101/650614

Boža V, Brejová B, Vinař T. 2017. DeepNano: Deep recurrent neural networks for base calling in MinION nanopore reads. *PLoS One* **12**:e0178751. doi:10.1371/journal.pone.0178751

Consortium TU. 2010. The universal protein rResource (UniProt) in 2010. *Nucleic Acids Res* **38**:D142–D148. doi:10.1093/nar/gkp846

David M, Dursi LJ, Yao D, Boutros PC, Simpson JT. 2017. Nanocall: an open source basecaller for Oxford Nanopore sequencing data. *Bioinformatics* **33**:49–55. doi:10.1093/bioinformatics/btw569

Di Genova A, Ruz GA, Sagot M-F, Maass A. 2018. Fast-SG: an alignment-free algorithm for hybrid assembly. *Gigascience* **7**. doi:10.1093/gigascience/giy048

Di Franco A, Poujol R, Baurain D, Philippe H. 2019. Evaluating the usefulness of alignment filtering methods to reduce the impact of errors on evolutionary inferences. *BMC Evol Biol* **19**:21. doi:10.1186/s12862-019-1350-2

Eddy SR. 2011. Accelerated profile HMM searches. *PLoS Comput Biol* **7**:e1002195. doi:10.1371/journal.pcbi.1002195

Edwards S V. 2009. Is a new and general theory of molecular systematics emerging? *Evolution (N Y)* **63**:1–19. doi:10.1111/j.1558-5646.2008.00549.x

Edwards S V., Liu L, Pearl DK. 2007. High-resolution species trees without concatenation. *Proc Natl Acad Sci* **104**:5936–5941. doi:10.1073/PNAS.0607004104

Emms DM, Kelly S. 2019. OrthoFinder: Phylogenetic orthology inference for comparative genomics. *Genome Biol* **20**:238. doi:10.1186/s13059-019-1832-y

Gan HM, Falk S, Morales HE, Austin CM, Sunnucks P, Pavlova A. 2019. Genomic evidence of neo-sex chromosomes in the eastern yellow robin. *Gigascience* **8**. doi:10.1093/gigascience/giz111

Gatesy J, Springer MS. 2014. Phylogenetic analysis at deep timescales: Unreliable gene trees, bypassed hidden support, and the coalescence/concatalescence conundrum. *Mol Phylogenet Evol* **80**:231–266. doi:10.1016/J.YMPEV.2014.08.013

Hoff KJ, Lange S, Lomsadze A, Borodovsky M, Stanke M. 2016. BRAKER1: Unsupervised RNA-seq-based genome annotation with GeneMark-ET and AUGUSTUS. *Bioinformatics* **32**:767–769. doi:10.1093/bioinformatics/btv661

Holt C, Yandell M. 2011. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics* **12**:491. doi:10.1186/1471-2105-12-491

Jain M, Koren S, Miga KH, Quick J, Rand AC, Sasani TA, Tyson JR, Beggs AD, Dilthey AT, Fiddes IT, Malla S, Marriott H, Nieto T, O'Grady J, Olsen HE, Pedersen BS, Rhie A, Richardson H, Quinlan AR, Snutch TP, Tee L, Paten B, Phillippy AM, Simpson JT, Loman NJ, Loose M. 2018. Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nat Biotechnol* **36**:338–345. doi:10.1038/nbt.4060

Jain M, Olsen HE, Paten B, Akeson M. 2016. The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. *Genome Biol* **17**:1–11. doi:10.1186/s13059-016-1103-0

Jeffroy O, Brinkmann H, Delsuc F, Philippe H. 2006. Phylogenomics: the beginning of incongruence? *Trends Genet* **22**:225–231. doi:10.1016/J.TIG.2006.02.003

Jiang JB, Quattrini AM, Francis WR, Ryan JF, Rodríguez E, McFadden CS. 2019. A hybrid de novo assembly of the sea pansy (Renilla muelleri) genome. *Gigascience* **8**. doi:10.1093/gigascience/giz026

Kadobianskyi M, Schulze L, Schuelke M, Judkewitz B. 2019. Hybrid genome assembly and annotation of Danionella translucida, a transparent fish with the smallest known vertebrate brain. *bioRxiv* 539692. doi:10.1101/539692

Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Mol Biol Evol* **30**:772–780. doi:10.1093/molbev/mst010

Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. 2017. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res* **27**:722–736. doi:10.1101/GR.215087.116

Korf I. 2004. Gene finding in novel genomes. *BMC Bioinformatics* **5**:59. doi:10.1186/1471-2105-5-59

Kumar S, Filipski AJ, Battistuzzi FU, Kosakovsky Pond SL, Tamura K. 2012. Statistics and truth in phylogenomics. *Mol Biol Evol* **29**:457–472. doi:10.1093/molbev/msr202

Kwan HH, Culibrk L, Taylor GA, Leelakumari S, Tan R, Jackman SD, Tse K, MacLeod T, Cheng D, Chuah E, Kirk H, Pandoh P, Carlsen R, Zhao Y, Mungall AJ, Moore R, Birol I, Marra MA, Rosen DAS, Haulena M, Jones SJM, Kwan HH, Culibrk L, Taylor GA, Leelakumari S, Tan R, Jackman SD, Tse K, MacLeod T, Cheng D, Chuah E, Kirk H, Pandoh P, Carlsen R, Zhao Y, Mungall AJ, Moore R, Birol I, Marra MA, Rosen DAS, Haulena M, Jones SJM. 2019. The Genome of the Steller Sea Lion (Eumetopias jubatus). *Genes (Basel)* **10**:486. doi:10.3390/genes10070486

Lartillot N. 2020. Phylobayes: Bayesian phylogenetics using site-heterogeneous models. *Phylogenetics genomic era*.

Lartillot N, Brinkmann H, Philippe H. 2007. Suppression of long-branch attraction artefacts in the animal phylogeny using a site-heterogeneous model. *BMC Evol Biol* **7**:S4. doi:10.1186/1471-2148-7-S1-S4

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. doi:10.1093/molbev/msh112

Li L, Stoeckert CJ, Roos DS. 2003. OrthoMCL: Identification of ortholog groups for eukaryotic genomes. *Genome Res* **13**:2178–2189. doi:10.1101/GR.1224503

Minh BQ, Hahn MW, Lanfear R. 2020. New methods to calculate concordance factors for phylogenomic datasets. *Mol Biol Evol*. doi:10.1093/molbev/msaa106

Mirarab S, Reaz R, Bayzid MS, Zimmermann T, Swenson MS, Warnow T. 2014. ASTRAL: genome-scale coalescent-based species tree estimation. *Bioinformatics* **30**:i541–i548. doi:10.1093/bioinformatics/btu462

Morrison DA. 2006. Multiple sequence alignment for phylogenetic purposes. *Aust Syst Bot* **19**:479–539.

Nguyen L-T, 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. doi:10.1093/molbev/msu300

Nicholls SM, Quick JC, Tang S, Loman NJ. 2019. Ultra-deep, long-read nanopore sequencing of mock microbial community standards. *Gigascience* **8**. doi:10.1093/gigascience/giz043

Palmer J, Stajich JE. 2017. Funaannotate: eukaryotic genome annotation pipeline.

Parker J, Helmstetter AJ, Devey D, Wilkinson T, Papadopulos AST. 2017. Field-based species identification of closely-related plants using real-time nanopore sequencing. *Sci Rep* **7**:8345. doi:10.1038/s41598-017-08461-5

Philippe H, Vienne DM de, Ranwez V, Roure B, Baurain D, Delsuc F. 2017. Pitfalls in supermatrix phylogenomics. *Eur J Taxon* **0**. doi:10.5852/ejt.2017.283

Pomerantz A, Peñafiel N, Arteaga A, Bustamante L, Pichardo F, Coloma LA, Barrio-Amorós CL, Salazar-Valenzuela D, Prost S. 2018. Real-time DNA barcoding in a rainforest using nanopore sequencing: opportunities for rapid biodiversity assessments and local capacity building. *Gigascience* **7**. doi:10.1093/gigascience/giy033

Ranwez V, Chantret N. 2020. Strengths and limits of multiple sequence alignment and filtering methods strengths and limits of multiple sequence alignment and filtering methods. *Phylogenetics Genomic Era* 2.2:1-2.2:36.

Ranwez V, Chantret N, Delsuc F. 2020. Aligning protein-coding nucleotide sequences with MACSE. *Methods Mol Biol* **In press**.

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. doi:10.1093/molbev/msy159

Rhoads A, Au KF. 2015. PacBio sequencing and its applications. *Genomics Proteomics Bioinformatics* **13**:278–289. doi:10.1016/J.GPB.2015.08.002

Roch S, Steel M. 2015. Likelihood-based tree reconstruction on a concatenation of aligned sequence data sets can be statistically inconsistent. *Theor Popul Biol* **100**:56–62. doi:10.1016/J.TPB.2014.12.005

Scornavacca C, Belkhir K, Lopez J, Dernat R, Delsuc F, Douzery EJP, Ranwez V. 2019. OrthoMaM v10: Scaling-up orthologous coding sequence and exon alignments with more than one hundred mammalian genomes. *Mol Biol Evol* **36**:861–862. doi:10.1093/molbev/msz015

Scott AD, Zimin A V., Puiu D, Workman R, Britton M, Zaman S, Caballero M, Read AC, Bogdanove AJ, Burns E, Wegrzyn J, Timp W, Salzberg SL, Neale DB. 2020. The giant sequoia genome and proliferation of disease resistance genes. *bioRxiv* 2020.03.17.995944. doi:10.1101/2020.03.17.995944

Shafin K, Pesout T, Lorig-Roach R, Haukness M, Olsen HE, Bosworth C, Armstrong J, Tigyi K, Maurer N, Koren S, Sedlazeck FJ, Marschall T, Mayes S, Costa V, Zook JM, Liu KJ, Kilburn D, Sorensen M, Munson KM, Vollger MR, Monlong J, Garrison E, Eichler EE, Salama S, Haussler D, Green RE, Akeson M, Phillippy A, Miga KH, Carnevali P, Jain M, Paten B. 2020. Nanopore sequencing and the Shasta toolkit enable efficient de novo assembly of eleven human genomes. *Nat Biotechnol* 1–10. doi:10.1038/s41587-020-0503-6

She R, Chu JS-C, Uyar B, Wang J, Wang K, Chen N. 2011. genBlastG: using BLAST searches to build homologous gene models. *Bioinformatics* **27**:2141–2143. doi:10.1093/bioinformatics/btr342

Simion P, Philippe H, Baurain D, Jager M, Richter DJ, Di Franco A, Roure B, Satoh N, Quéinnec É, Ereskovsky A, Lapébie P, Corre E, Delsuc F, King N, Wörheide G, Manuel M. 2017. A large and consistent phylogenomic dataset supports sponges as the sister group to all other animals. *Curr Biol* **27**:958–967. doi:10.1016/j.cub.2017.02.031

Smit AF, Hubley R. 2008. RepeatModeler Open-1.0. *Available fom http//www repeatmasker org*.

Springer MS, Gatesy J. 2016. The gene tree delusion. *Mol Phylogenet Evol* **94**:1–33. doi:10.1016/J.YMPEV.2015.07.018

Srivathsan A, Baloğlu B, Wang W, Tan WX, Bertrand D, Ng AHQ, Boey EJH, Koh JJY, Nagarajan N, Meier R. 2018. A MinIONTM-based pipeline for fast and cost-effective DNA barcoding. *Mol Ecol Resour* **18**:1035–1049. doi:10.1111/1755-0998.12890

Stanke M, Keller O, Gunduz I, Hayes A, Waack S, Morgenstern B. 2006. AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Res* **34**:W435–W439. doi:10.1093/nar/gkl200

Tan MH, Austin CM, Hammer MP, Lee YP, Croft LJ, Gan HM. 2018. Finding Nemo: hybrid assembly with Oxford Nanopore and Illumina reads greatly improves the clownfish (Amphiprion ocellaris) genome assembly. *Gigascience* **7**. doi:10.1093/gigascience/gix137

Tarailo-Graovac M, Chen N. 2009. Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr Protoc Bioinforma* **25**:4.10.1-4.10.14. doi:10.1002/0471250953.bi0410s25

Testa AC, Hane JK, Ellwood SR, Oliver RP. 2015. CodingQuarry: highly accurate hidden Markov model gene prediction in fungal genomes using RNA-seq transcripts. *BMC Genomics* **16**:170. doi:10.1186/s12864-015-1344-4

Vaser R, Sović I, Nagarajan N, Šikić M. 2017. Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Res* **27**:737–746. doi:10.1101/GR.214270.116

Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, Cuomo CA, Zeng Q, Wortman J, Young SK, Earl AM. 2014. Pilon: an Integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One* **9**:e112963. doi:10.1371/journal.pone.0112963

Wang W, Das A, Kainer D, Schalamun M, Morales-Suarez A, Schwessinger B, Lanfear R. 2019. The draft nuclear genome assembly of Eucalyptus pauciflora: new approaches to comparing de novo assemblies. *bioRxiv* 678730. doi:10.1101/678730

Waterhouse RM, Seppey M, Simão FA, Manni M, Ioannidis P, Klioutchnikov G, Kriventseva E V, Zdobnov EM. 2018. BUSCO applications from quality assessments to gene prediction and phylogenomics. *Mol Biol Evol* **35**:543–548. doi:10.1093/molbev/msx319

Wheeler TJ, Clements J, Eddy SR, Hubley R, Jones TA, Jurka J, Smit AFA, Finn RD. 2012. Dfam: a database of repetitive DNA based on profile hidden Markov models. *Nucleic Acids Res* **41**:D70–D82. doi:10.1093/nar/gks1265

Xu Z, Mai Y, Liu D, He W, Lin X, Xu C, Zhang L, Meng X, Mafofo J, Zaher WA, Li Y, Qiao N. 2020. Fast-Bonito: A faster basecaller for nanopore sequencing. *bioRxiv* 2020.10.08.318535. doi:10.1101/2020.10.08.318535

Yandell M, Ence D. 2012. A beginner's guide to eukaryotic genome annotation. *Nat Rev Genet* **13**:329–342. doi:10.1038/nrg3174

Zeng J, Cai H, Peng H, Wang H, Zhang Y, Akutsu T. 2020. Causalcall: Nanopore basecalling using a temporal convolutional network. *Front Genet* **10**:1332. doi:10.3389/fgene.2019.01332

Zheng GXY, Lau BT, Schnall-Levin M, Jarosz M, Bell JM, Hindson CM, Kyriazopoulou-Panagiotopoulou S, Masquelier DA, Merrill L, Terry JM, Mudivarti PA, Wyatt PW, Bharadwaj R, Makarewicz AJ, Li Y, Belgrader P, Price AD, Lowe AJ, Marks P, Vurens GM, Hardenbol P, Montesclaros L, Luo M, Greenfield L, Wong A, Birch DE, Short SW, Bjornson KP, Patel P, Hopmans ES, Wood C, Kaur S, Lockwood GK, Stafford D, Delaney JP, Wu I, Ordonez HS, Grimes SM, Greer S, Lee JY, Belhocine K, Giorda KM, Heaton WH, McDermott GP, Bent ZW, Meschi F, Kondov NO, Wilson R, Bernate JA, Gauby S, Kindwall A, Bermejo C, Fehr AN, Chan A, Saxonov S, Ness KD, Hindson BJ, Ji HP. 2016. Haplotyping germline and cancer genomes with high-throughput linked-read sequencing. *Nat Biotechnol* **34**:303–311. doi:10.1038/nbt.3432

Zimin A V., Marçais G, Puiu D, Roberts M, Salzberg SL, Yorke JA. 2013. The MaSuRCA genome assembler. *Bioinformatics* **29**:2669–2677. doi:10.1093/bioinformatics/btt476

Zimin A V., Puiu D, Luo M-C, Zhu T, Koren S, Marçais G, Yorke JA, Dvořák J, Salzberg SL. 2017. Hybrid assembly of the large and highly repetitive genome of Aegilops tauschii, a progenitor of bread wheat, with the MaSuRCA mega-reads algorithm. *Genome Res* **27**:787–792. doi:10.1101/GR.213405.116

⇧Back to summary⇧

2 - High-quality carnivore genomes from roadkill samples enable species delimitation in aardwolf and bat-eared fox

The second article of my thesis is an illustration of the use of this methodological approach (with only one run for MAKER annotation and without RNA-seq data). In this study, we were particularly interested in the subspecies of bat-eared fox (*Otocyon megalotis*) and aardwolf (*Proteles cristata*), which are ant-/termite-eating mammals presenting similar non-continuous distributions in Eastern and Southern Africa. The characterization of the taxonomic status of these sub-species potentially has consequences for further convergent-oriented phylogenetic analyses. Given that our population genomics analyses suggested that the two subspecies of *P. cristata* warrant full species status, a new carnivoran phylogeny was inferred from 14,307 orthologous genes extracted following the methodological approach described above.

The preprint associated with this section is available on BioRxiv:
https://doi.org/10.1101/2020.09.15.297622
As well as the supplementay material:
https://www.biorxiv.org/content/10.1101/2020.09.15.297622v1.supplementary-material

⇧Back to summary⇧

High-quality carnivore genomes from roadkill samples enable species delimitation in aardwolf and bat-eared fox

Rémi Allio[1]*, Marie-Ka Tilak[1], Céline Scornavacca[1], Nico L. Avenant[2], Erwan Corre[3], Benoit Nabholz[1], and Frédéric Delsuc[1]*

[1]Institut des Sciences de l'Evolution de Montpellier (ISEM), CNRS, IRD, EPHE, Université de Montpellier, France remi.allio@umontpellier.fr marie-ka.tilak@umontpellier.fr celine.scornavacca@umontpellier.fr benoit.nabholz@umontpellier.fr frederic.delsuc@umontpellier.fr
[2]National Museum and Centre for Environmental Management, University of the Free State, Bloemfontein, South Africa navenant@nasmus.co.za
[3]CNRS, Sorbonne Université, FR2424, ABiMS, Station Biologique de Roscoff, 29680 Roscoff, France corre@sb-roscoff.fr

***Correspondence:** remi.allio@umontpellier.fr, frederic.delsuc@umontpellier.fr

Abstract

In a context of ongoing biodiversity erosion, obtaining genomic resources from wildlife is becoming essential for conservation. The thousands of yearly mammalian roadkill could potentially provide a useful source material for genomic surveys. To illustrate the potential of this underexploited resource, we used roadkill samples to sequence reference genomes and study the genomic diversity of the bat-eared fox (Otocyon megalotis) and the aardwolf (Proteles cristata) for which subspecies have been defined based on similar disjunct distributions in Eastern and Southern Africa. By developing an optimized DNA extraction protocol, we successfully obtained long reads using the Oxford Nanopore Technologies (ONT) MinION device. For the first time in mammals, we obtained two reference genomes with high contiguity and gene completeness by combining ONT long reads with Illumina short reads using hybrid assembly. Based on re-sequencing data from few other roakill samples, the comparison of the genetic differentiation between our two pairs of subspecies to that of pairs of well-defined species across Carnivora showed that the two subspecies of aardwolf might warrant species status (P. cristata and P. septentrionalis), whereas the two subspecies of bat-eared fox might not. Moreover, using these data, we conducted demographic analyses that revealed similar trajectories between Eastern and Southern populations of both species, suggesting that their population sizes have been shaped by similar environmental fluctuations. Finally, we obtained a well resolved genome-scale phylogeny for Carnivora with evidence for incomplete lineage sorting among the three main arctoid lineages. Overall, our cost-effective strategy opens the way for large-scale population genomic studies and phylogenomics of mammalian wildlife using roadkill.

Keywords: Roadkill, Genomics, Population genomics, Phylogenomics, Species delimitation, Carnivora, Systematics, Genetic differentiation, Mitogenomes, Africa.

Introduction

In the context of worldwide biodiversity erosion, obtaining large-scale genomic resources from wildlife is essential for biodiversity assessment and species conservation. An underexploited but potentially useful source of material for genomics is the thousands of annual wildlife fatalities due to collisions with cars. Mammalian roadkill in particular are unfortunately so frequent that several citizen science surveys have been implemented on this subject in recent decades (Périquet et al., 2018; Shilling et al., 2015). For example, in South Africa alone, over 12,000 wildlife road mortality incidents were recorded by The Endangered Wildlife Trust's Wildlife and Roads Project from 1949 to 2017 (Endangered Wildlife Trust 2017). Initially developed to measure the impact of roads on wildlife, these web-based systems highlight the amount of car-wildlife collision. The possibility of retrieving DNA from roadkill tissue samples (Etherington et al., 2020; Maigret, 2019) could provide new opportunities in genomics by giving access not only to a large number of specimens of commonly encountered species but also to more elusive species that might be difficult to sample otherwise.

Recent advances in the development of high-throughput sequencing technologies have made the sequencing of hundreds or thousands of genetic loci cost efficient and have offered the possibility of using ethanol-preserved tissues, old DNA extracts, and museum specimens (Blaimer et al., 2016; Guschanski et al., 2013). This method combined with third generation long read sequencing technologies such as Pacific Biosciences (PacBio) and Oxford Nanopore Technologies (ONT) sequencing have increased the sizes of the sequenced molecules from several kilobases to several megabases. The relatively high level of sequencing errors (10-15%) associated with these technologies can be compensated by sequencing at a high depth of coverage to avoid sequencing errors in de novo genome assembly and thus obtain reference genomes with high base accuracy, contiguity, and completeness (Koren et al., 2017; Shafin et al., 2020; Vaser et al., 2017). Originally designed to provide a portable sequencing method in the field, ONT instruments such as the MinION (Jain et al., 2016) allow direct sequencing of DNA molecules with simplified library preparation procedures even in tropical environments with elevated temperature and humidity conditions (Blanco et al., 2019; Parker et al., 2017; Pomerantz et al., 2018; Srivathsan et al., 2018). This approach is particularly suitable for sequencing roadkill specimens for which it is notoriously difficult to obtain a large amount of high-quality DNA because of post-mortem DNA degradation processes. Furthermore, it is possible to correct errors in ONT long reads by combining them with Illumina short reads, either to polish de novo long read-based genome assemblies (Batra et al., 2019a; Jain et al., 2018; Nicholls et al., 2019; Walker et al., 2014) or to construct hybrid assemblies (Di Genova et al., 2018; Gan et al., 2019; Tan et al., 2018; Zimin et al., 2013). In hybrid assembly approaches, the accuracy of short reads with high depth of coverage (50-100x) allows the use of long reads at lower depth of coverage (10-30x) essentially for scaffolding (Armstrong et al., 2020; Kwan et al., 2019). A promising hybrid assembly approach combining short and long read sequencing data has been implemented in MaSuRCA software (Zimin et al., 2017, 2013). This approach consists of transforming large numbers of short reads into a much smaller number of longer highly accurate "super reads" allowing the use of a mixture of read lengths. Furthermore, this method is designed to tolerate a significant level of sequencing error. Initially developed to address short reads from Sanger sequencing and longer reads from 454 Life Sciences instruments, this method has already shown promising results for combining Illumina and ONT/PacBio sequencing data in several taxonomic groups, such as plants (Scott et al., 2020; Wang et al., 2020; Zimin et al., 2017), birds (Gan et al., 2019), and fishes (Jiang et al., 2019; Kadobianskyi et al., 2019; Tan et al., 2018) but not yet in mammals.

To illustrate the potential of roadkill as a useful resource for whole genome sequencing

and assembly, we studied two of the most frequently encountered mammalian roadkill species in South Africa (Périquet et al., 2018): the bat-eared fox (Otocyon megalotis, Canidae) and the aardwolf (Proteles cristata, Hyaenidae). These two species are among several African vertebrate taxa presenting disjunct distributions between Southern and Eastern African populations that are separated by more than a thousand kilometres (e.g. Ostrich (Miller et al., 2011), Ungulates Lorenzen et al. 2012). Diverse biogeographical scenarios involving the survival and divergence of populations in isolated savannah refugia during the climatic oscillations of the Pleistocene have been proposed to explain these disjunct distributions in ungulates (Lorenzen et al., 2012). Among Carnivora, subspecies have been defined based on this peculiar allopatric distribution not only for the black-backed jackal (Canis mesomelas; Walton and Joly 2003) but also for both the bat-eared fox (Clark, 2005) and the aardwolf (Koehler and Richardson, 1990) (**Fig. 1**). The bat-eared fox is divided into the Southern bat-eared fox (O. megalotis megalotis) and the Eastern bat-eared fox (O. megalotis virgatus) (Clark, 2005), and the aardwolf is devided into the Southern aardwolf (P. cristata cristata) and the Eastern aardwolf (P. cristata septentrionalis) (Koehler and Richardson, 1990). However, despite known differences in behaviour between the subspecies of both species groups (Wilson et al., 2009), no genetic or genomic assessment of population differentiation has been conducted to date. In other taxa, similar allopatric distributions have led to genetic differences between populations and several studies reported substantial intraspecific genetic structuration between Eastern and Southern populations (Atickem et al., 2018; Barnett et al., 2006; Dehghani et al., 2008; Lorenzen et al., 2012; Miller et al., 2011; Rohland et al., 2005). Here, with a novel approach based on a few individuals, we investigate whether similar genetic structuration and population differentiation have occurred between subspecies of bat-eared fox and aardwolf using whole genome data.

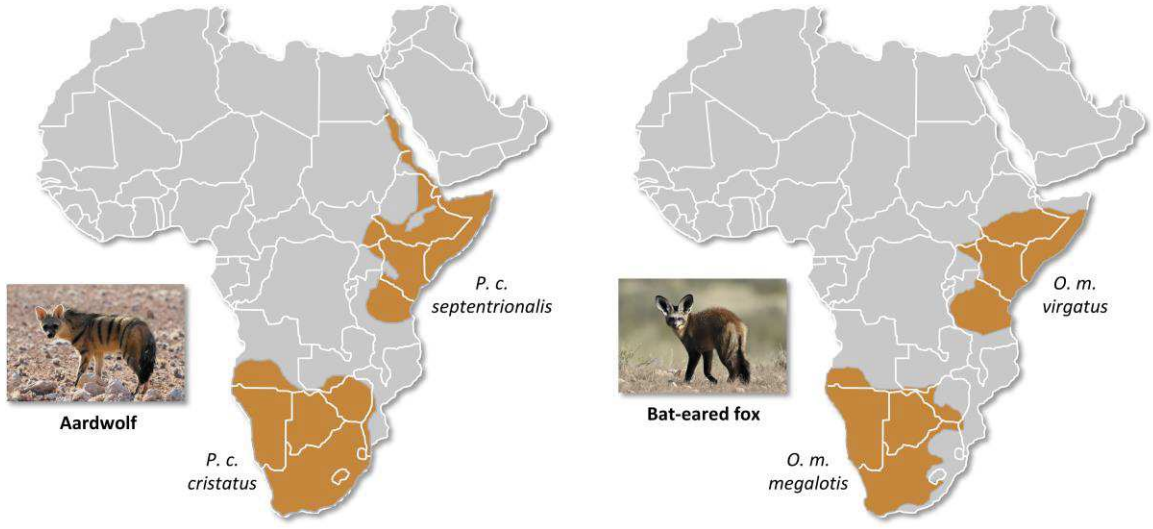


Figure 1 | Disjunct distributions of the aardwolf (Proteles cristata) and the bat-eared fox (Otocyon megalotis) in Eastern and Southern Africa. Within each species, two subspecies have been recognized based on their distributions and morphological differences (Clark, 2005; Koehler and Richardson, 1990).

To evaluate the taxonomic status of the proposed subspecies within both O. megalotis and P. cristata, we first sequenced and assembled two reference genomes from roadkill samples by combining ONT long reads and Illumina short reads using the MaSuRCA hybrid assembler. The quality of our genome assemblies was assessed by comparison to available mammalian genome assemblies. Then, to estimate the genetic diversity of these species and to perform genome-scale species delimitation analyses, two additional individuals from the disjunct South African and Tanzanian populations of both species were resequenced at high depth of coverage using Illumina short reads. Using these additional individuals, we estimated the genetic diversity and differentiation of each subspecies pair via an FST-like measure, which we called the genetic differentiation index, and compared the results with the genetic differentiation among pairs of well-established carnivoran species. Based on genetic differentiation measures, we find that the two subspecies of P. cristata warrant potential species delineation, whereas the subspecies of O. megalotis are likely allocated properly. Our results show that high-quality reference mammalian genomes could be obtained through a combination of short- and long-read sequencing methods providing opportunities for large-scale population genomic studies of mammalian wildlife using (re)sequencing of samples collected from roadkill.

Results

Mitochondrial diversity within Carnivora

The first dataset, composed of complete carnivoran mitogenomes available in GenBank and the newly generated sequences of the two subspecies of P. cristata, the two subspecies of O. megalotis, Parahyaena brunnea, Speothos venaticus and Vulpes vulpes, plus the sequences extracted from UCE libraries for Bdeogale nigripes, Fossa fossana, and Viverra tangalunga, consists of 142 species or subspecies representing all families of Carnivora, including 5 *O. megalotis* and 10 *P. cristata* individuals. Maximum likelihood (ML) analyses reconstructed a robust mitogenomic phylogeny, with 91.4% of the nodes (128 out of 140) recovered with bootstrap support higher than 95% (**Fig. 2a**). The patristic distances based on complete mitogenomes between the allopatric subspecies of aardwolf and bat-eared fox were 0.045 and 0.020 substitutions per site, respectively (**Table S1**). These genetic distances are comparable to those observed between different well-defined species of Carnivora such as the red fox (Vulpes vulpes) and the fennec (Vulpes zerda) (0.029) or the Steppe polecat (Mustela eversmannii) and the Siberian weasel (Mustela sibirica) (0.034) (see **Table S1**).

To further assess the genetic distances between the two pairs of subspecies and compare them to both polymorphism and divergence values observed across Carnivora, two supplemental datasets including at least two individuals per species were assembled by retrieving all COX1 and CYTB sequences, which are the two widely sequenced mitochondrial markers for carnivores, available on GenBank. These datasets include 3,657 COX1 sequences for 150 species and 6,159 CYTB sequences for 203 species of Carnivora. After adding the corresponding sequences from the newly assembled mitogenomes, ML phylogenetic inference was conducted on each dataset (Supplementary materials). The patristic distances between all tips of the resulting phylogenetic trees were measured and classified into two categories: (i) intraspecific variation (polymorphism) for distances inferred among individuals of the same species and (ii) interspecific divergence for distances inferred among individuals of different species. Despite an overlap between polymorphism and divergence in both mitochondrial genes, this analysis revealed a threshold between polymorphism and divergence of approximately 0.02 substitutions per site for Carnivora (**Fig. 2b**). With a nucleotide distance of 0.054 for both COX1 and CYTB, the genetic distance observed between the two subspecies of aardwolf (Proteles ssp.) was higher than the

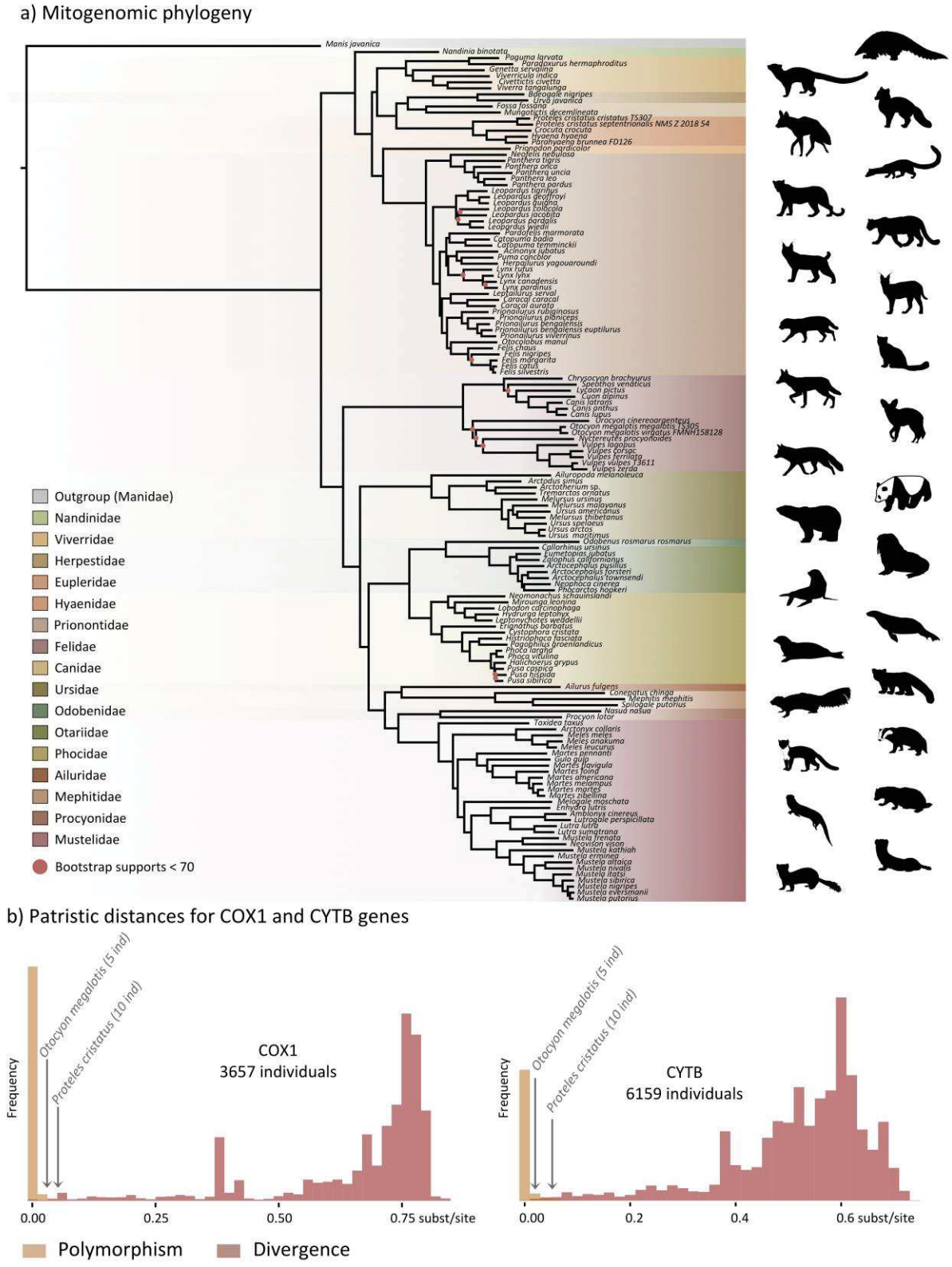


Figure 2 | Representation of the mitochondrial genetic diversity within Carnivora with a) the mitogenomic phylogeny inferred from 142 complete Carnivora mitogenomes including those of the two populations of aardwolf (Proteles cristata) and bat-eared fox (Otocyon megalotis) and b) intraspecific (orange) and the interspecific (red) genetic diversities observed for the two mitochondrial markers COX1 and CYTB.

majority of the intraspecific distances observed across Carnivora. However, with a nucleotide distances of 0.020 for COX1 and 0.032 for CYTB, the genetic distance observed between the two subspecies of bat-eared fox (Otocyon ssp.) was clearly in the ambiguous zone and did not provide a clear indication of the specific taxonomic status of these populations.

Finally, to test whether the two pairs of allopatric subspecies diverged synchronously or in two different time periods, Bayesian molecular dating inferences were performed on the 142-taxon ML mitogenomic tree. The resulting divergence times were slightly different depending on the clock model used (strict clock [CL], autocorrelated [LN or TK02] and uncorrelated [UGAM or UCLM]) (Supplementary materials) despite the convergence of the MCMC chains for all models. Cross-validation analyses resulted in the selection of the LN and UGAM models as the models with the best fit based on a higher cross-likelihood score than that of CL (LN and UGAM versus CL mean scores = 35 ± 8). Unfortunately, these two statistically indistinguishable models provided different divergence times for the two pairs of subspecies, with LN favouring a synchronous divergence (approximately 1 Mya [95% credibility interval (CI): 6.72 - 0.43]; **Table S2**), while UGAM favoured an asynchronous divergence (~0.6 [CI: 0.83 - 0.39] Mya for O. megalotis ssp. and ~1.3 [CI: 1.88 - 0.93] Mya for P. cristata ssp.; **Table S2**). However, the three chains performed with the UGAM model recovered highly similar ages for the two nodes of interest with low CI 95% values whereas the three chains performed with the LN model recovered less similar ages between chains and high CI 95% values (**Table 1**).

Assembling reference genomes from roadkill

Considering the DNA quality and purity required to perform single-molecule sequencing with ONT, a specific protocol to extract DNA from roadkill was developed (Tilak et al., 2020). This protocol was designed to specifically select the longest DNA fragments present in the extract also containing short degraded fragments. This protocol increased the median size of the sequenced raw DNA fragments three-fold in the case of aardwolf (Tilak et al., 2020). In total, after high-accuracy basecalling, adapter trimming, and quality filtering, 27.3 Gb of raw Nanopore long reads were sequenced using 16 MinION flow cells for the Southern aardwolf (P. c. cristata) and 33.0 Gb using 13 flow cells for the Southern bat-eared fox (O. m. megalotis) (**Table 1**). Due to quality differences among the extracted tissues for both species, the N50 of the DNA fragment size for P. cristata (9,175 bp) was about twice higher than the N50 of the DNA fragment size obtained for O. megalotis (4,393 bp). The quality of the reads basecalled with the high accuracy option of Guppy was significantly higher than the quality of those translated with the fast option, which led to better assemblies (see **Fig. S1**). Complementary Illumina sequencing returned 522.8 and 584.4 million quality-filtered reads per species corresponding to 129.5 Gb (expected coverage = 51.8x) and 154.8 Gb (expected coverage = 61.6x) for P. c. cristata and O. m. megalotis, respectively. Regarding the resequenced individuals of each species, on average 153.5 Gb were obtained with Illumina resequencing (**Table 1**). The two reference genomes were assembled using MinION long reads and Illumina short reads in combination with MaSuRCA v3.2.9 (Zimin et al., 2013). Hybrid assemblies for both species were obtained with a high degree of contiguity with only 5,669 scaffolds and an N50 of 1.3 Mb for the aardwolf (P. cristata) and 11,081 scaffolds and an N50 of 728 kb for the bat-eared fox (O. megalotis) (Table 1). Exhaustive comparisons with 503 available mammalian assemblies revealed a large heterogeneity among taxonomic groups and a wide variance within groups in terms of both number of scaffolds and N50 values (**Fig. 3**, **Table S3**). Xenarthra was the group with the lowest quality genome assemblies, with a median number of scaffolds of more than one million and a median N50 of

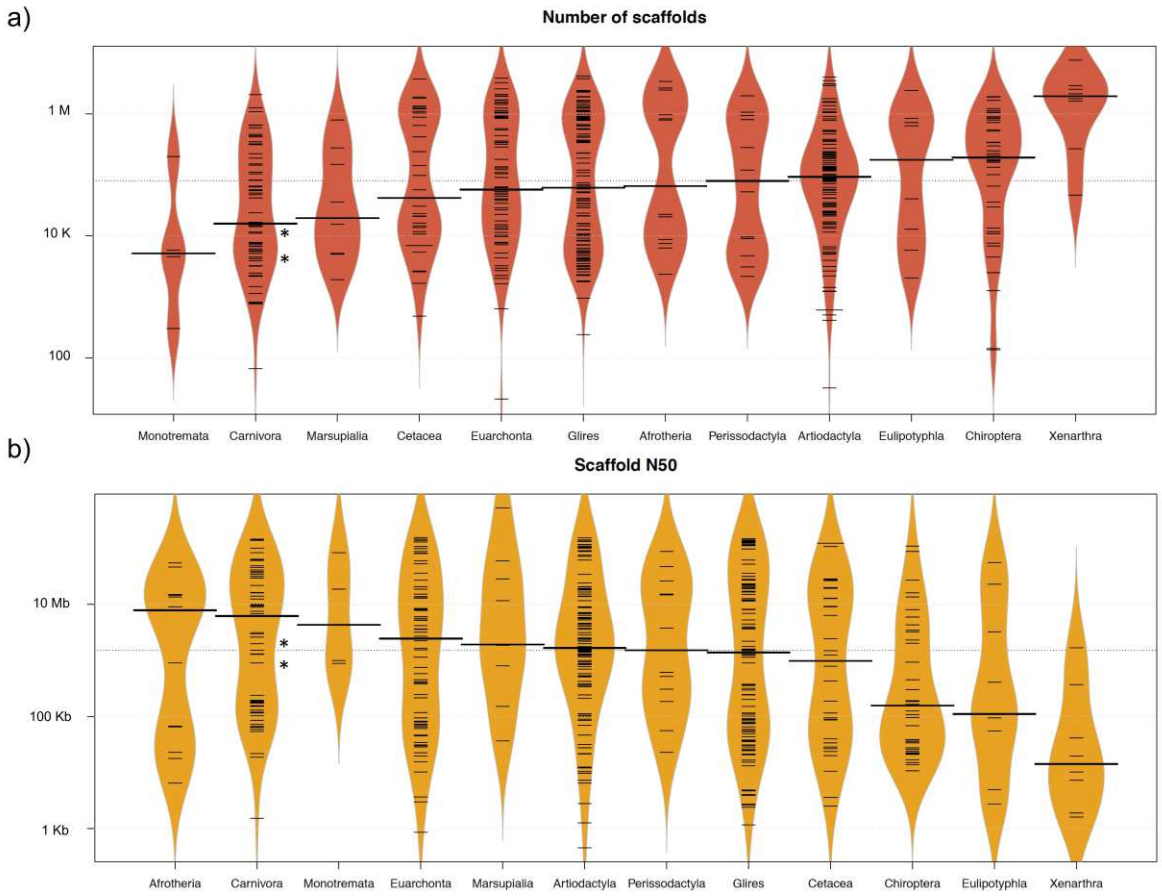


Figure 3 | Comparison of 503 mammalian genome assemblies from 12 taxonomic groups using bean plots of the a) number of scaffolds, and b) scaffold N50 values ranked by median values. Thick black lines show the medians, dashed black lines represent individual data points, and polygons represent the estimated density of the data. Note the log scale of the Y axes. The bat-eared fox (Otocyon megalotis) and aardwolf (Proteles cristata) assemblies produced in this study using SOAPdenovo and MaSuRCA are indicated by asterisks. Bean plots were computed using BoxPlotR (Spitzer et al., 2014).

Table 1 | Summary of sequencing and assembly statistics of the genomes generated in this study.

Individuals			Illumina				Oxford Nanopore Sequencing					Assembly statistics					
Species	Subspecies	Voucher	Raw reads (M)	Cleaned reads	Nbr of gigabases	Estimated coverage	Nbr of flowcells	Nbr of bases (Gb)	N50	Average size	Estimated coverage	Genome size (Gb)	Nbr of scaff.	N50 (kb)	Busco score	OMM genes	Missing data (%)
Proteles cristatus	cristatus	TS307	716.7	522.8	129.50	51.8	16	27.3	9,175	5,555	10.9	2.39	5,669	1,309	92.8	12,062	22.43
Proteles cristatus	cristatus	TS491	663.8	526.1	140.73	56.3	NA					NA				NA	
Proteles cristatus	septentrionalis	NMSZ201854	750.9	516.2	132.44	53.0										12,050	22.96
Otocyon megalotis	megalotis	TS305	710.2	584.4	154.81	61.6	13	33	4,393	3,092	13.2	2.75	11,081	728	92.9	11,981	22.02
Otocyon megalotis	megalotis	TS306	861.2	820	240.71	96.3	NA					NA					
Otocyon megalotis	virgatus	FMNH158128	661.7	554.1	100.30	40.1											

only 15 kb. Conversely, Carnivora contained genome assemblies of much better quality, with a median number of scaffolds of 15,872 and a median N50 of 4.6 Mb, although a large variance was observed among assemblies for both metrics (**Fig. 3**, **Table S3**). Our two new genomes compared favourably with the available carnivoran genome assemblies in terms of contiguity showing slightly less than the median N50 and a lower number of scaffolds than the majority of the other assemblies (**Fig. 3**, **Table S3**). Comparison of two hybrid assemblies with Illumina-only assemblies obtained with SOAPdenovo illustrated the positive effect of introducing Nanopore long reads even at moderate coverage by reducing the number of scaffolds from 409,724 to 5,669 (aardwolf) and from 433,209 to 11,081 (bat-eared fox) while increasing the N50 from 17.3 kb to 1.3 Mb (aardwolf) and from 22.3 kb to 728 kb (bat-eared fox). With regard to completeness based on 4,104 single-copy mammalian BUSCO orthologues, our two hybrid assemblies are among the best assemblies with more than 90% complete BUSCO genes and less than 4% missing genes (**Fig. 4**, **Table S4**). As expected, the two corresponding Illumina-only assemblies were much more fragmented and had globally much lower BUSCO scores (**Fig. 4**, **Table S4**).

Genome-wide analyses of population structure

To evaluate the population structure between the subspecies of P. cristata and O. megalotis, the number of shared heterozygous sites, unique heterozygous sites, and homozygous sites between individuals was computed to estimate an FST-like statistic (hereafter called the genetic differentiation index or GDI). Since we were in possession of two individuals for the Southern subspecies and only one for the Eastern subspecies of both species, the genetic differentiation between the two individuals within the Southern subspecies and between the Southern and Eastern subspecies was computed. To account for the variation across the genome, 10 replicates of 100 regions with a length of 100 kb were randomly chosen to estimate genetic differentiation. Interestingly, in both species, the mean heterozygosity was higher in the Southern subspecies than in the Eastern subspecies. For aardwolf, the mean heterozygosity was 0.189 per kb (sd = 0.010) in the Southern population and 0.121 per kb (sd = 0.008) in the Eastern population. For the bat-eared fox, the mean heterozygosity was 0.209 per kb (sd = 0.013) in the Southern population and 0.127 per kb (sd = 0.003) in the Eastern population. This heterozygosity level is low compared to those of other large mammals (Diez-del-Molino et al 2018) and is comparable to that of the Iberian lynx, the cheetah or the brown hyena, which have notoriously low genetic diversity (Abascal et al., 2016; Casas-Marce et al., 2013; Westbury et al., 2018). Since we had very limited power to fit the evolution of the genetic differentiation statistics with a hypothetical demographic scenario because of our limited sample size, we chose a comparative approach and applied the same analyses to four well-defined species pairs of carnivorans for which similar individual sampling was available. The genetic differentiation estimates between the two individuals belonging to the same subspecies (Southern populations in both cases) were on average equal to 0.005 and 0.014 for P. c. cristata and O. m. megalotis, respectively. This indicated that the polymorphism observed in the two individuals within the Southern subspecies of each species was comparable (genetic differentiation index close to 0) and thus that these two subpopulations are likely panmictic (**Fig. 5**). In contrast, the genetic differentiation estimates for the two pairs of individuals belonging to the different subspecies were respectively equal to 0.533 and 0.294 on average for P. cristata ssp. and O. megalotis ssp., indicating that the two disjunct populations are genetically structured. To contextualize these results, the same genetic differentiation measures were estimated for four other well-defined species pairs (**Fig. 5**). First, the comparison of the polymorphism of two individuals of the same species led to intraspecific GDIs ranging from 0.029 on average for polar bear (Ursus

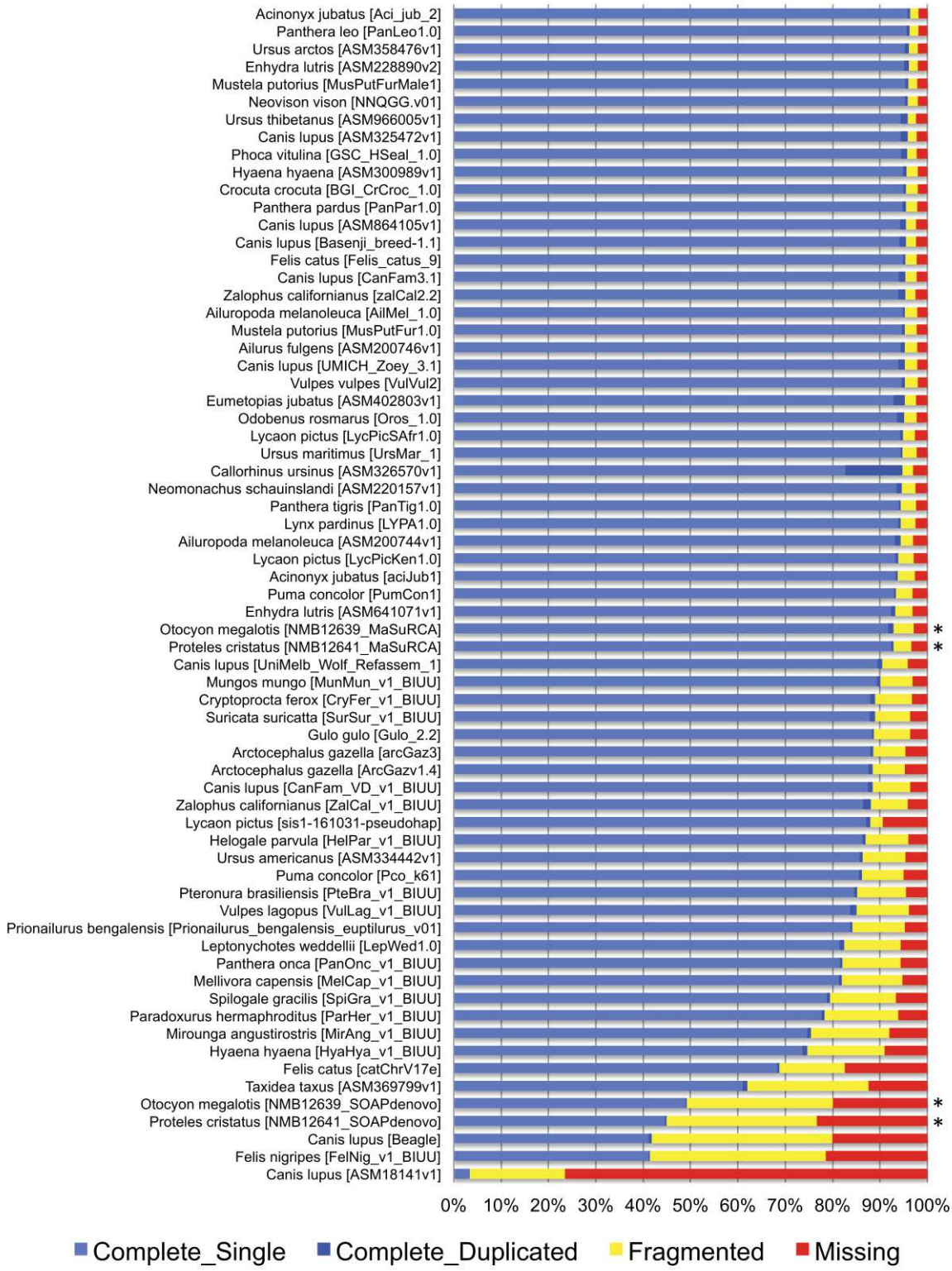


Figure 4 | BUSCO completeness assessment of 67 Carnivora genome assemblies visualized as bar charts representing percentages of complete single-copy (light blue), complete duplicated (dark blue), fragmented (yellow), and missing (red) genes ordered by increasing percentage of total complete genes. The bat-eared fox (Otocyon megalotis) and aardwolf (Proteles cristata) assemblies produced in this study using MaSuRCA and SOAPdenovo are indicated by asterisks.

maritimus) to 0.137 for lion (Panthera leo). As expected, comparing the polymorphisms of two individuals between closely related species led to a higher interspecific GDI ranging from 0.437 on average for the wolf/golden jackal (Canis lupus/Canis aureus) pair to 0.760 for the lion/leopard (P. leo/Panthera pardus) pair (**Fig. 5**). The genetic differentiation indices between the grey wolf (C. lupus) and the golden jackal (C. aureus) averaged 0.44, indicating that the two subspecies of aardwolf (GDI = 0.533) are genetically more differentiated than these two well-defined species, and only slightly less differentiated than the brown bear (Ursus arctos) and the polar bear (U. maritimus). Conversely, the genetic differentiation obtained between the bat-eared fox subspecies (GDI = 0.294) were lower than the genetic differentiation estimates obtained for any of the four reference species pairs evaluated here (**Fig. 5**).

Effective population size reconstructions

We used the pairwise sequential Markovian coalescent (PSMC) model to estimate the ancestral effective population size (Ne) trajectory over time for each sequenced individual. For both the aardwolf and the bat-eared fox, the individual from Eastern African populations showed a continuous decrease in Ne over time, leading to the recent Ne being lower than that in Southern African populations (**Fig. 6**). This is in agreement with the lower heterozygosity observed in the Eastern individuals of both species. For the bat-eared fox, the trajectories of the three sampled individuals were synchronised approximately 200 kya ago (**Fig. 6a**), which could correspond to the time of divergence between the Southern and Eastern populations. In contrast, the Ne trajectories for the aardwolf populations did not synchronise over the whole period (~2 Myrs). Interestingly, the Southern populations of both species showed a marked increase in population size between ~10-30 kya before sharply decreasing in more recent times (**Fig. 6**).

Phylogenomics of Carnivora

Phylogenetic relationships within Carnivora were inferred from a phylogenomic dataset comprising 52 carnivoran species (including the likely new Proteles septentrionalis species) representing all but two families of Carnivora (Nandiniidae and Prionodontidae). The non-annotated genome assemblies of these different species were annotated with a median of 18,131 functional protein-coding genes recovered for each species. Then, single-copy orthologous gene identification resulted in a median of 12,062 out of the 14,509 single-copy orthologues extracted from the OrthoMaM database for each species, ranging from a minimum of 6,305 genes for the California sea lion (Zalophus californianus) and a maximum of 13,808 for the dog (Canis lupus familiaris) (**Table S5**). Our new hybrid assemblies allowed the recovery of 12,062 genes for the Southern aardwolf (P. c. cristata), 12,050 for the Eastern aardwolf (P. c. septentrionalis), and 11,981 for the Southern bat-eared fox (O. m. megalotis) (**Table 1**). These gene sets were used to create a supermatrix consisting of 14,307 genes representing a total of 24,041,987 nucleotide sites with 6,495,611 distinct patterns (27.0%) and 22.8% gaps or undetermined nucleotides. Phylogenomic inference was first performed on the whole supermatrix using ML. The resulting phylogenetic tree was highly supported, with all but one node being supported by maximum bootstrap (UFBS) values (**Fig. 7**). To further dissect the phylogenetic signal underlying this ML concatenated topology, we measured gene concordance (gCF) and site concordance (sCF) factors to complement traditional bootstrap node-support values. For each node, the proportion of genes (gCF) or sites (sCF) that supported the node inferred with the whole supermatrix was compared to the proportion of the genes (gDF) or sites (sDF) that supported an alternative resolution of the node (**Fig. 7,** Supplementary materials). Finally, a coalescent-based approximate species tree inference was performed using ASTRAL-III based on individual gene trees (Supplementary materials).

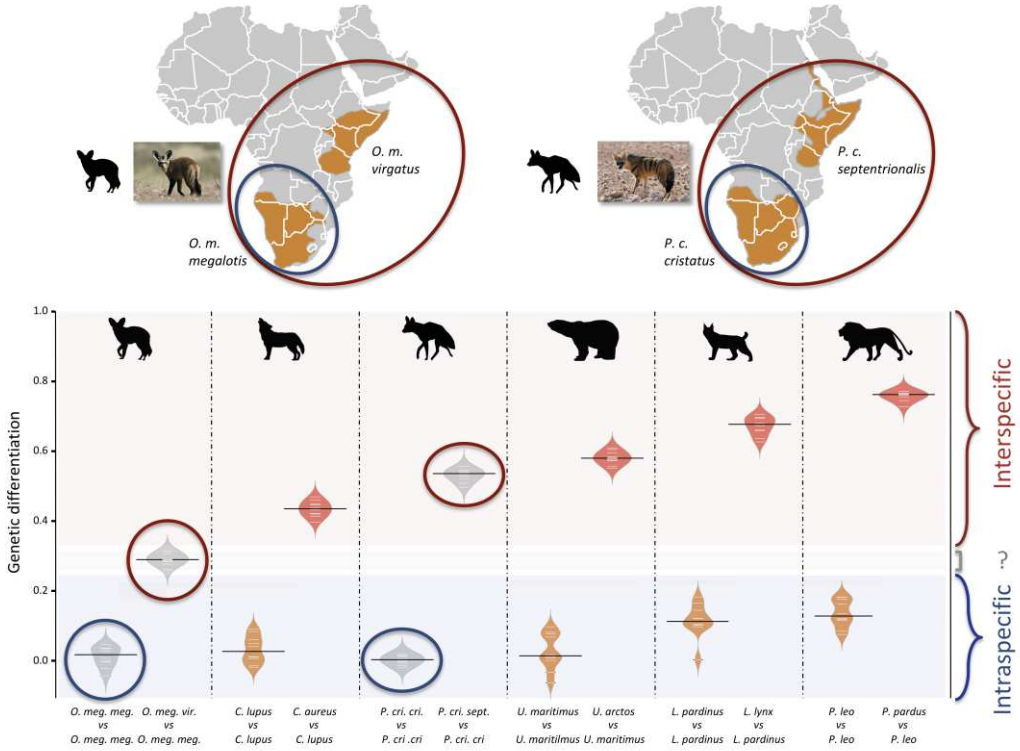


Figure 5 | Genetic differentiation indices obtained from the comparison of intraspecific (orange) and interspecific (red) polymorphisms in four pairs of well-defined Carnivora species and for the subspecies of aardwolf (Proteles cristata) and bat-eared fox (Otocyon megalotis) (grey).

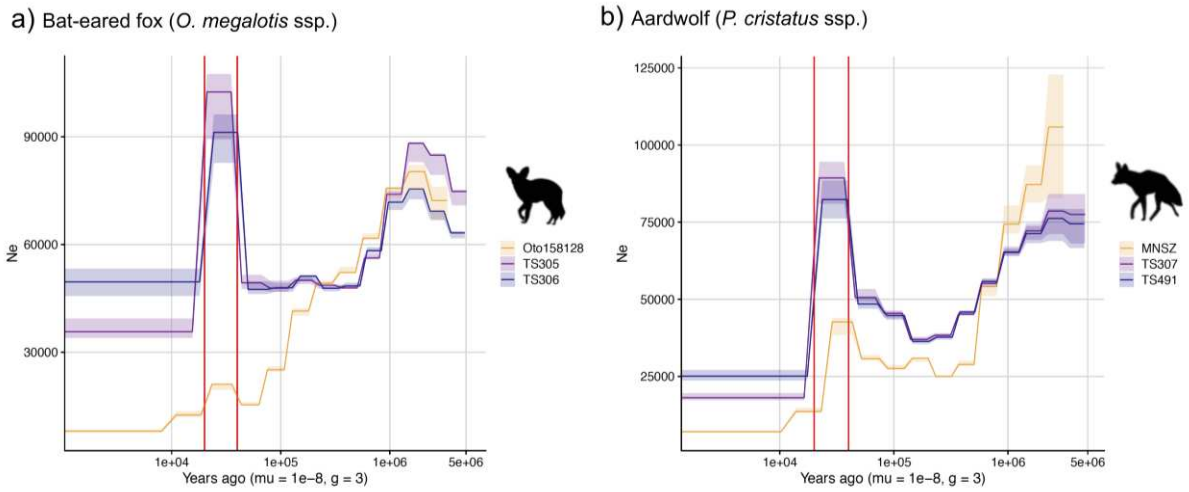


Figure 6 | PSMC estimates of the change in effective population size over time for the Eastern (orange) and Southern (blue and purple) populations of a) bat-eared fox and) aardwolf. mu = mutation rate of 10-8 mutations per site per generation and g = generation time of 2 years. Vertical red lines indicate 20kyrs and 40kyrs.

Overall, the three different analyses provided well-supported and almost identical results (**Fig. 7**). The order Carnivora was divided into two distinct suborders: a cat-related clade (Feliformia) and a dog-related clade (Caniformia). Within Feliformia, the first split separated Felidae (felids) from Viverroidea, a clade composed of the four families Viverridae (civets and genets), Eupleridae (fossa), Herpestidae (mongooses), and Hyaenidae (hyaenas). In hyaenids, the two species of termite-eating aardwolves (P. cristata and P. septentrionalis) were the sister-group of a clade composed of the carnivorous spotted (Crocuta crocuta) and striped (Hyaena hyaena) hyenas. Congruent phylogenetic relationships among Feliformia families and within hyaenids were also retrieved with the mitogenomic data set (**Fig. 2a**). The short internal nodes of Felidae were the principal source of incongruence among the three different analyses with concordance factor analyses pointing to three nodes for which many sites and genes support alternative topologies (**Fig. 7**) including one node for which the coalescent-based approximate species tree inference supported an alternative topology (Supplementary materials) to the one obtained with ML on the concatenated supermatrix. In Viverroidea, Viverridae split early from Herpestoidea regrouping Hyaenidae, Herpestidae, and Eupleridae, within which Herpestidae and Eupleridae formed a sister clade to Hyaenidae. Within Caniformia, Canidae (canids) was recovered as a sister group to Arctoidea. Within Canidae, in accordance with the mitogenomic phylogeny, the Vulpini tribe, represented by O. megalotis and V. vulpes, was recovered as the sister clade of the Canini tribe, represented here by Lycaon pictus and C. l. familiaris. The Arctoidea were recovered as a major clade composed of eight families grouped into three subclades: Ursoidea (Ursidae), Pinnipedia (Otariidae, Odobedinae, and Phocidae), and Musteloidea, composed of Ailuridae (red pandas), Mephitidae (skunks), Procyonidae (raccoons), and Mustelidae (badgers, martens, weasels, and otters). Within Arctoidea, the ML phylogenetic inference on the concatenation provided support for grouping Pinnipedia and Musteloidea to the exclusion of Ursidae (bears) with maximum bootstrap support (**Fig. 7**), as in the mitogenomic tree (**Fig. 2a**). However, the concordance factor analyses revealed that many sites and many genes actually supported alternative topological conformations for this node characterized by a very short branch length (sCF=34.1, SDF1=29.2, sDF2=36.7, gCF=46.9, gDF1=18.6, gDF2=18.2, gDFP=16.3) (**Fig. 7**). In Pinnipedia, the clade Odobenidae (walruses) plus Otariidae (eared seals) was recovered to the exclusion of Phocidae (true seals), which was also in agreement with the mitogenomic scenario (**Fig. 2a**). Finally, within Musteloidea, Mephitidae represented the first offshoot, followed by Ailuridae, and a clade grouping Procyonidae and Mustelidae. Phylogenetic relationships within Musteloidea were incongruent with the mitogenomic tree, which alternatively supported the grouping of Ailuridae and Mephitidae (**Fig. 2a**).

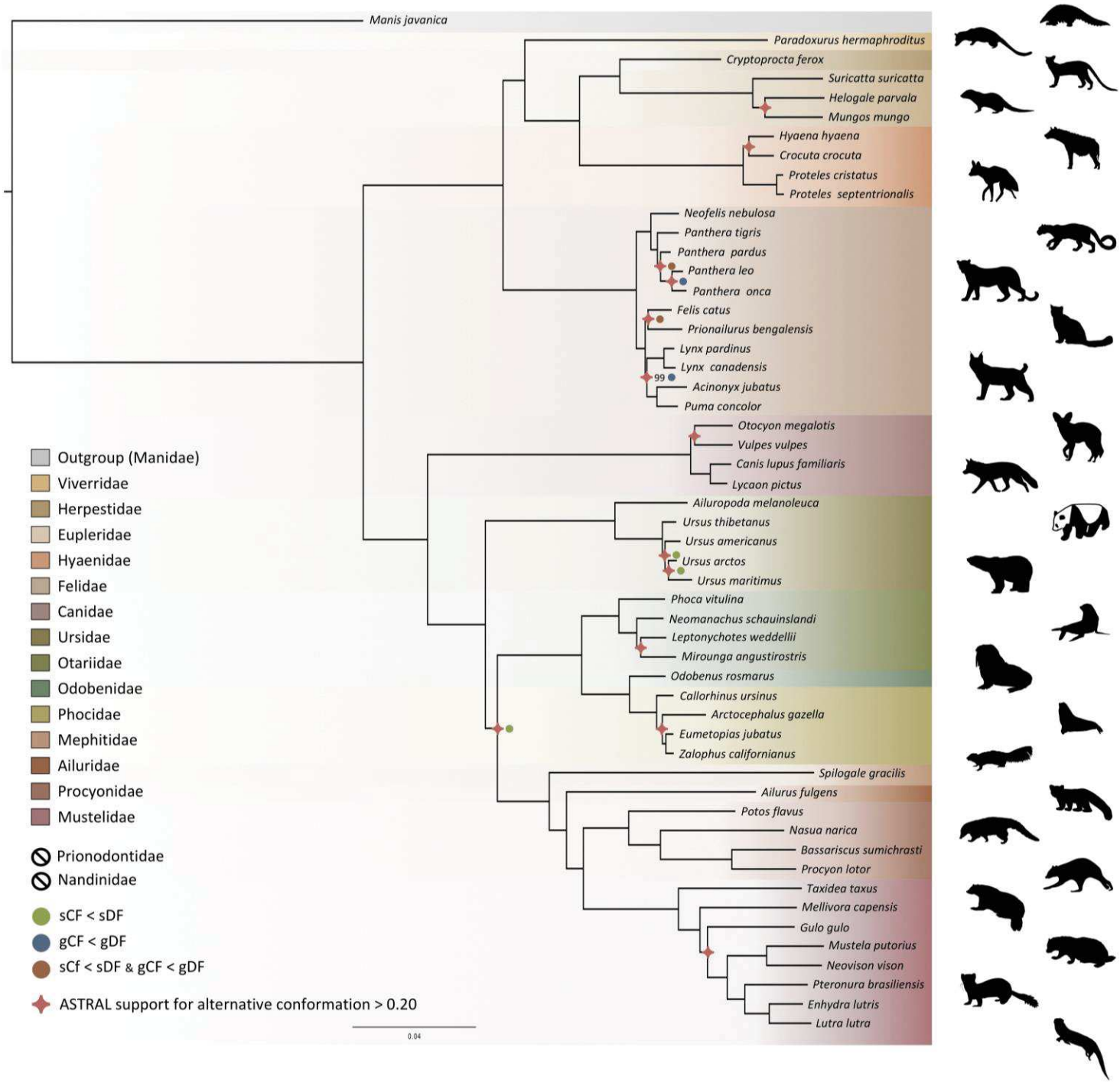


Figure 7 | Phylogenomic tree reconstructed from the nucleotide supermatrix composed of 14,307 single-copy orthologous genes for 52 species of Carnivora plus one outgroup (Manis javanica). The family names in the legend are ordered as in the phylogeny.

Discussion

High-quality mammalian genomes from roadkill using MaSuRCA hybrid assembly

Long-read sequencing technologies and associated bioinformatic tools hold promise for making chromosome-length genome assemblies the gold standard (Dudchenko et al., 2017; Koepfli et al., 2015; Rice and Green, 2019). However, obtaining relatively large mammalian genomes of high quality remains a challenging and costly task for researchers working outside of large genome sequencing consortia (Li et al., 2010; Lindblad-Toh et al., 2011). Despite the accuracy of short-read sequencing technologies, the use of PCR amplification is needed to increase the depth of coverage, creating uneven genomic representation and leading to sequencing biases such as GC-rich regions being less well sequenced than AT-rich ones in classical Illumina libraries (Aird et al., 2011; Tilak et al., 2018). Moreover, the use of short reads involves difficulties in the assembly of repeated regions or transposable elements longer than the sequencing read length. The use of less GC-biased long reads of single DNA molecules and ultra-long reads spanning repeated genomic regions provides a powerful solution for obtaining assemblies with high contiguity and completeness, although long-read sequencing has limited accuracy (10-20% errors). Long reads can indeed be used alone at a high depth of coverage permitting autocorrection (Koren et al., 2017; Shafin et al., 2020) or in combination with short reads for (1) scaffolding short-read contigs (Armstrong et al., 2020; Kwan et al., 2019), (2) using short reads to polish long-read contigs (Batra et al., 2019b; Datema et al., 2016; Jansen et al., 2017; Michael et al., 2018), or (3) optimizing the assembly process by using information from both long and short reads (Díaz-Viraqué et al., 2019; Gan et al., 2019; Jiang et al., 2019; Kadobianskyi et al., 2019; Tan et al., 2018; Wang et al., 2020; Zimin et al., 2017). Given the previously demonstrated efficiency of the MaSuRCA tool for the assembly of large genomes (Scott et al., 2020; Wang et al., 2020; Zimin et al., 2017), we decided to rely on hybrid sequencing data combining the advantages of Illumina short-read and Nanopore long-read sequencing technologies.

With an increasing number of species being threatened worldwide, obtaining genomic resources from mammalian wildlife can be difficult. We decided to test the potential of using roadkill samples, a currently underexploited source material for genomics (Etherington et al., 2020; Maigret, 2019). Despite limited knowledge and difficulties associated with de novo assembly of non-model species (Etherington et al., 2020), we designed a protocol to produce DNA extracts of suitable quality for Nanopore long-read sequencing from roadkill (Tilak et al., 2020). Additionally, we tested the impact of the accuracy of the MinION base calling step on the quality of the resulting MaSuRCA hybrid assemblies. In line with previous studies (Wenger et al., 2019; Wick et al., 2019), we found that using the high accuracy option rather than the fast option of Guppy 3.1.5 leads to more contiguous assemblies by increasing the N50 value. By relying on this protocol, we were able to generate two hybrid assemblies by combining Illumina reads at relatively high coverage (80x) and MinION long reads at relatively moderate coverage (12x) which provides genomes with high contiguity and completeness. These represent the first two mammalian genomes obtained with such a hybrid Illumina/Nanopore approach using the MaSuRCA assembler for non-model carnivoran species: the aardwolf (P. cristata) and the bat-eared fox (O. megalotis). Despite the use of roadkill samples, our assemblies compare favourably, in terms of both contiguity and completeness, with the best carnivoran genomes obtained so far from classical genome sequencing approaches that do not rely on complementary optical mapping or chromatin conformation approaches. Overall, our carnivoran hybrid assemblies are fairly comparable to those obtained using the classic Illumina-based genome sequencing protocol

involving the sequencing of both paired-end and mate-paired libraries (Li et al., 2010). The benefit of adding Nanopore long reads is demonstrated by the fact that our hybrid assemblies are of better quality than all the draft genome assemblies generated using the DISCOVAR de novo protocol based on a PCR-free single Illumina 250 bp paired-end library (Weisenfeld et al. 2014; DISCOVAR) used in the 200 Mammals Project of the Broad Genome Institute (The 200 mammals project).These results confirm the capacity of the MaSuRCA hybrid assembler to produce quality assemblies for large and complex genomes by leveraging the power of long Nanopore reads (Wang et al., 2020). Moreover, these two hybrid assemblies could form the basis for future chromosome-length assemblies by adding complementary HiC data (van Berkum et al., 2010) as proposed in initiatives such as the Vertebrate Genome Project (Koepfli et al., 2015) and DNAZoo (Dudchenko et al., 2017). Our results demonstrate the feasibility of producing high-quality mammalian genome assemblies at moderate cost using roadkill and should encourage genome sequencing of non-model mammalian species in ecology and evolution laboratories.

Genomic evidence for two distinct species of aardwolves

The mitogenomic distances inferred between the subspecies of O. megalotis and P. cristata were comparable to those observed for other well-defined species within Carnivora. Furthermore, by comparing the genetic diversity between several well-defined species (divergence) and several individuals of the same species (polymorphism) based on the COX1 and CYTB genes across Carnivora, we were able to pinpoint a threshold of approximately 0.02 substitutions per base separating divergence from polymorphism, which is in accordance with a recent study of naturally occurring hybrids in Carnivora (Allen et al., 2020). This method, also known as the barcoding gap method (Meyer and Paulay, 2005), allowed us to show that the two subspecies of P. cristata present a genetic divergence greater than the threshold, whereas the divergence is slightly lower for the two subspecies of O. megalotis. These results seem to indicate that the subspecies P. c. septentrionalis might have to be elevated to the species level (P. septentrionalis). Conversely, for O. megalotis, this first genetic indicator seems to confirm the distinction at the subspecies level. However, mitochondrial markers have some well-identified limitations (Galtier et al., 2009), and it is difficult to properly determine a threshold between polymorphism and divergence across Carnivora. The measure of mtDNA sequence distances can thus be seen only as a first useful indicator for species delineation. The examination of variation at multiple genomic loci in a phylogenetic context, combined with morphological, behavioural and ecological data, is required to establish accurate species boundaries.

The newly generated reference genomes allowed us to perform genome-wide evaluation of the genetic differentiation between subspecies using short-read resequencing data of a few additional individuals of both species. Traditionally, the reduction in polymorphism in two subdivided populations (p within) compared to the population at large (p between) is measured with several individuals per population (FST; Hudson et al. 1992). However, given that the two alleles of one individual are the results of the combination of two a priori non-related individuals of the population (i.e., the parents), with a large number of SNPs, the measurement of heterozygosity can be extended to estimation of the (sub)population polymorphism. Furthermore, in a panmictic population with recombination along the genome, different chromosomal regions can be considered to be independent and can be used as replicates for heterozygosity estimation. In this way, genome-wide analyses of heterozygosity provide a way to assess the level of polymorphism in a population and a way to compare genetic differentiation between two populations. If we hypothesize that the two compared populations are panmictic, picking one individual or another of the

population has no effect (i.e., there is no individual with excess homozygous alleles due to mating preference across the population), and the population structure can be assessed by comparing the heterozygosity of the individuals of each population compared to the heterozygosity observed for two individuals of the same population (see Methods). Such an index of genetic differentiation, by measuring the level of population structure, could provide support to establish accurate species boundaries. In fact, delineating species has been and still is a complex task in evolutionary biology (Galtier, 2019; Ravinet et al., 2016; Roux et al., 2016). Given that accurately defining the species taxonomic level is essential for a number of research fields, such as macroevolution (Faurby et al., 2016) or conservation (Frankham et al., 2012), defining thresholds to discriminate between populations or subspecies in different species is an important challenge in biology. However, due to the disagreement on the definition of species, the different routes of speciation observed in natura and the different amount of data available among taxa, adapting a standardized procedure for species delineation seems complicated (Galtier, 2019).

As proposed by Galtier (Galtier, 2019), we decided to test the taxonomic level of the P. cristata and O. megalotis subspecies by comparing the genetic differentiation observed between Eastern and Southern populations within these species to the genetic differentiation measured for well-defined Carnivora species. Indeed, estimation of the genetic differentiation either within well-defined species (polymorphism) or between two closely related species (divergence) allowed us to define a threshold between genetic polymorphism and genetic divergence across Carnivora (**Fig. 5**). Given these estimates, and in accordance with mitochondrial data, the two subspecies of P. cristata (1) present more genetic differentiation between each other than the two well-defined species of golden jackal (Canis aureus) and wolf (Canis lupus), and (2) present more genetic differentiation than the more polymorphic species of the dataset, the lion (P. leo). Despite known cases of natural hybridization reported between C. aureus and C. lupus (Galov et al., 2015; Gopalakrishnan et al., 2018), the taxonomic rank of these two species is well accepted. In that sense, given the species used as a reference, the two subspecies of P. cristata seem to deserve to be elevated to the species level. The situation is less clear regarding the subspecies of O. megalotis. Indeed, while the genetic differentiation observed between the two subspecies is significantly higher than the polymorphic distances observed for all the well-defined species of the dataset, there is no species in our dataset that exhibits equivalent or lower genetic divergence than a closely related species. This illustrates the limits of delineating closely related species due to the continuous nature of the divergence process (De Queiroz, 2007). The subspecies of O. megalotis fall into the “grey zone” of the speciation continuum (De Queiroz, 2007; Roux et al., 2016) and are likely undergoing speciation due to their vicariant distributions. To be congruent with the genetic divergence observed across closely related species of Carnivora (according to our dataset), we thus propose that (1) the taxonomic level of the P. cristata subspecies be reconsidered by elevating the two subspecies P. c. cristata and P. c. septentrionalis to the species level, and (2) the taxonomic level for the two subspecies of O. megalotis be maintained. These new taxonomic results should prompt a deeper investigation of morphological and behavioural differences that have been reported between the two proposed subspecies of aardwolf to formally validate our newly proposed taxonomic arrangement. They also have conservation implications, as the status of the two distinct aardwolf species will have to be re-evaluated separately in the International Union for Conservation of Nature (IUCN) Red List of Threatened Species (IUCN 2020, 2020).

Population size variation and environmental change

The Pairwise Sequentially Markovian Coalescent (PSMC) analyses revealed that the Southern and Eastern African populations have

different effective population size estimates over time, confirming that they have been genetically isolated for several thousand years, which is more so for the aardwolf than for the bat-eared fox. This supports the hypothesis of two separate events leading to the same disjunct repartitions for the two taxa, in accordance with mitochondrial dating. Nevertheless, the population trends are rather similar and are characterized by continuous declines between 1 Mya and 100-200 kya that are followed by an increase that is much more pronounced in the Southern populations of both species between 30-10 kya. The similar trajectories exhibited by both species suggest that they were under the influence of similar environmental factors, such as climate and vegetation variations.

Aardwolves and bat-eared foxes live in open environments including short-grass plains, shrubland, and open and tree savannahs, and both are highly dependent on herbivorous termites for their diet. Therefore, the fluctuation of their populations could reflect the evolution of these semi-arid ecosystems determining prey abundance during the last million years. However, the global long-term Plio-Pleistocene African climate is still debated. For Eastern Africa, some studies have suggested an evolution towards increased aridity (deMenocal, 2004, 1995) whereas others have proposed the opposite (Grant et al., 2017; Maslin et al., 2014; Trauth et al., 2009). Our data therefore support the latter hypothesis, as a global long-term tendency towards a wetter climate in East Africa could have been less favourable for species living in open environments.

Southern populations exhibit a similar decreasing trend between 1 Mya and 100 kya. Once again, the relevant records appear contradictory. This could be the result of regional variation across South Africa, with aridification in the Southwestern part and wetter conditions in the Southeast (Caley et al., 2018; Johnson et al., 2016). Finally, the 30-10 kya period appears to have been more humid (Chase et al., 2019; Chevalier and Chase, 2015; Lim et al., 2016). This seems inconsistent with the large population increase detected in Southern populations of both species; however, the large regions of the Namib Desert that are currently unsuitable could have been more favorable in wetter conditions.

The global decrease in population size detected in the Southern and Eastern populations could also reflect the fragmentation of a continuous ancestral range. The global trend towards a wetter climate may have favoured the development of the tropical rainforest in central Africa creating a belt of unsuitable habitat. This is in line with previous studies describing diverse biogeographical scenarios involving the survival and divergence of ungulate populations in isolated savannah refuges during Pleistocene climate oscillations (Lorenzen et al., 2012). In this respect, it could be interesting to study population trends in other species living in semi-arid environments and having a similar range as disconnected populations. Interestingly, several bird species also have similar distributions including the Orange River francolin (Scleroptila gutturalis), the greater kestrel (Falco rupicoloides), the double-banded courser (Smutsornis africanus), the red-fronted tinkerbird (Pogoniulus pusillus), the cape crow (Corvus capensis) and the black-faced waxbill (Estrilda erythronotos), supporting the role of the environment in the appearance of these disjunct repartitions. Finally, these new demographic results showing recent population size declines in both regions in both species might be taken into account when assessing the conservation status of the two distinct aardwolf species and bat-eared fox subspecies.

Genome-scale phylogeny of Carnivora

In this study, we provide a new phylogeny of Carnivora including the newly recognized species of aardwolf (P. septentrionalis). The resulting phylogeny is fully resolved with all nodes supported with UFBS values greater than 95% and is congruent with previous studies (Doronina et al., 2015; Eizirik et al., 2010) (**Fig. 5**). Across Carnivora, the monophyly of all superfamilies described are strongly supported (Flynn et al., 2010) and are divided into two

distinct suborders: a cat-related clade (Feliformia) and a dog-related clade (Caniformia). On the one hand, within Feliformia, the different families and their relative relationships are well supported and are in accordance with previous studies (Eizirik et al., 2010). There is one interesting point regarding the Felidae family. While almost all the nodes of the phylogeny were recovered as strongly supported from the three phylogenetic inference analyses (ML inferences, concordance factor analyses and coalescent-based inferences), one third of the nodes (3 out of 9) within Felidae show controversial node supports. This result is not surprising and is consistent with previous studies arguing for ancient hybridization among Felidae (Li et al., 2016). Another interesting point regarding Feliformia and particularly Hyaenidae is the relationship of the two aardwolves. The two species, P. cristasta and P. septentrionalis form a sister clade to the clade composed of the striped hyena (H. hyaena) and the spotted hyena (C. crocuta), in accordance with previous studies (Koepfli et al., 2006; Westbury et al., 2018) and the two subfamilies Protelinae and Hyaeninae that have been proposed for these two clades, respectively. However, although the phylogenetic inferences based on the supermatrix of 14,307 single-copy orthologues led to a robust resolution of this node according to the bootstrap supports, both concordance factors and coalescent-based analyses revealed conflicting signals with support for alternative topologies. In this sense, the description and acceptance of the Hyaninae and Protelinae families still require further analyses, and including genomic data for the brown hyena (Parahyena brunnea) seems essential (Westbury et al., 2018).

On the other hand, within Caniformia, the first split separates Canidae from the Arctoidea. Within Canidae, the bat-eared fox (O. megalotis) is grouped with the red fox (Vulpes vulpes), the other representative of the tribe Vulpini, but with a very short branch and concordance analyses indicating conflicting signals on this node. Regarding Arctoidea, historically, the relationships between the three superfamilies of arctoids have been contradictory and debated. The least supported scenario from the litterature is that in which the clade Ursoidea/Musteloidea is a sister group of Pinnipedia (Flynn and Nedbal, 1998). Based on different types of phylogenetic characters, previous studies found support for both the clade Ursoidea/Pinnipedia (Agnarsson et al., 2010; Meredith et al., 2011; Rybczynski et al., 2009) and the clade Pinnipedia/Musteloidea (Arnason et al., 2007; Eizirik et al., 2010; Flynn et al., 2005; Sato et al., 2009, 2006; Schröder et al., 2009). However, investigations of the insertion patterns of retroposed elements revealed the occurrence of incomplete lineage sorting (ILS) at this node (Doronina et al., 2015). With a phylogeny inferred from 14,307 single-copy orthologous genes, our study, based on both gene trees and supermatrix approaches, gives support to the variant Pinnipedia/Musteloidea excluding Ursoidea as the best supported conformation for the Arctoidea tree (Doronina et al., 2015; Eizirik et al., 2010; Sato et al., 2006). Interestingly, in agreement with Doronina et al. (Doronina et al., 2015), our concordance factor analysis supports the idea that the different conformations of the Arctoidea tree are probably due to incomplete sorting of the lineage by finding almost the same number of sites supporting each of the three conformations (34.11%, 29.61% and 36.73%). However, although trifurcation of this node is supported by these proportions of sites, a majority of genes taken independently (gene concordance factors: 6,624 out of 14,307 genes) and the coalescent-based species tree approach (quartet posterior probabilities q1 = 0.53, q2 = 0.24, q3 = 0.24) support the clade Pinnipedia/Musteloidea excluding Ursoidea. Considering these results, the difficulty of resolving this trifurcation among Carnivora (Delisle and Strobeck, 2005) has likely been contradictory due to the ILS observed among these three subfamilies (Doronina et al., 2015), which led to different phylogenetic scenarios depending on the methods (Peng et al., 2007) or markers (L and YP, 2006) used. Another controversial point, likely due to incomplete lineage sorting

(Doronina et al., 2015) within the Carnivora phylogeny, is the question regarding which of Ailuridae and Mephitidae is the most basal family of the Musteloidea (Doronina et al., 2015; Eizirik et al., 2010; Flynn et al., 2005; Sato et al., 2009). Interestingly, our phylogenetic reconstruction based on mitogenomic data recovered the clade Ailuridae/Mephitidae as a sister clade of all other Musteloidea families. The phylogenomic inferences based on the genome-scale supermatrix recovered the Mephitidae family as the most basal family of Musteloidea. This result is supported by both coalescent-based inferences and concordance factors. In that sense, despite incomplete lineage sorting (Doronina et al., 2015), at the genomic level, it seems that the Mephitidae family would be the most basal family of Musteloidea.

Overall, the phylogenomic inference based on 14,307 single-copy orthologous genes provides a new vision of the evolution of Carnivora. The addition of information from both concordance factor analyses (Minh et al., 2020) and coalescent-based inference (Zhang et al., 2018) supports previous analyses showing controversial nodes in the Carnivora phylogeny. Indeed, this additional information seems essential in phylogenomic analyses based on thousands of markers, which can lead to highly resolved and well-supported phylogenies despite support for alternative topological conformations for controversial nodes (Allio et al., 2020b; Jeffroy et al., 2006; Kumar et al., 2012).

Conclusions

The protocol developed here to extract the best part of the DNA from roadkill samples provides a good way to obtain genomic data from wildlife. Combining Illumina sequencing data and Oxford Nanopore long-read sequencing data using the MaSuRCA hybrid assembler allowed us to generate high-quality reference genomes for the Southern aardwolf (P. cristata) and the Southern bat-eared fox (O. megalotis megalotis). This cost-effective strategy provides opportunities for large-scale population genomic studies of mammalian wildlife using resequencing of samples collected from roadkill. Indeed, by defining a genetic differentiation index based on only three individuals, we illustrate the potential of the approach for genome-scale species delineation in both species for which subspecies have been defined based on disjunct distributions and morphological differences. Our results, based on both mitochondrial and nuclear genome analyses, indicate that the two subspecies of P. cristata warrant elevation to the species taxonomic level; the O. megalotis subspecies do not warrant this status, but are likely ongoing species. Hence, by generating reference genomes with high contiguity and completeness, this study shows a concrete application for genomics of roadkill samples.

Methods

Biological samples

We conducted fieldwork in the Free State province of South Africa in October 2016 and October 2018. While driving along the roads, we opportunistically collected tissue samples from four roadkill specimens from which we sampled ear necropsies preserved in 95% Ethanol: two bat-eared foxes (O. megalotis NMB TS305, GPS: 29°1’52”S, 25°9’38”E and NMB TS306, GPS: 29°2’33”S, 25°10’26”E), and two aardwolves (P. cristata NMB TS307, GPS: 29°48’45”S, 26°15’0”E and NMB TS491, GPS: 29°8’42”S, 25°39’4”E). As aardwolf specimen NMB TS307 was still very fresh, we also sampled muscle and salivary gland necropsies preserved in RNAlater™ stabilization solution (Thermo Fisher Scientific). These roadkill specimens have been sampled under standing collecting permit number S03016 issued by the Department of National Affairs in Pretoria (South Africa) granted to the National Museum, Bloemfontein. These samples have been sent to France under export permits (JM 3007/2017 and JM 5043/2018) issued by the Free State Department of Economic, Small Business Development, Tourism and Environmental Affairs (DESTEA) in Bloemfontein (Free State,

South Africa) and import permits issued by the Direction régionale de l'environnement, de l'aménagement et du logement (DREAL) Occitanie in Toulouse (France). All tissue samples collected in this study have been deposited in the mammalian tissue collection of the National Museum, Bloemfontein (Free State, South Africa).

Mitochondrial barcoding and phylogenetics

Mitogenomic dataset construction

In order to assemble a mitogenomic data set for assessing mitochondrial diversity among P. cristata and O. megalotis subspecies, we generated seven new Carnivora mitogenomes using Illumina shotgun sequencing (**Table S6**). Briefly, we extracted total genomic DNA total using the DNeasy Blood and Tissue Kit (Qiagen) for P. c. cristata (NMB TS307), P. c. septentrionalis (NMS Z.2018.54), O. m. megalotis (NMB TS305), O. m. virgatus (FMNH 158128), Speothos venaticus (ISEM T1624), Vulpes vulpes (ISEM T3611), and Parahyaena brunnea (ISEM FD126), prepared Illumina libraries following the protocol of Tilak et al. (Tilak et al., 2015), and sent libraries to the Montpellier GenomiX platform for single-end 100 bp sequencing on a Illumina HiSeq 2500 instrument to obtain about 5 to 10 million reads per sample. We then assembled and annotated mitogenomes from these single-read shotgun sequencing data with MitoFinder v1.0.2 (Allio et al., 2020a) using default parameters. We also used MitoFinder to extract three additional mitogenomes from paired-end Illumina capture libraries of ultra-conserved elements (UCEs) and available from the Short Read Archive (SRA) of NCBI for Viverra tangalunga, Bdeogale nigripes, and Fossa fossana Additional read mappings were done with Geneious (Kearse et al., 2012) to close gaps when the mitochondrial genome was fragmented. Finally, we downloaded all RefSeq carnivoran mitogenomes available in Genbank (135 species as of July 1st, 2019) and the mitogenome of the Malayan pangolin (Manis javanica) to use as outgroup.

Mitogenomic phylogenetics and dating

Mitochondrial protein coding genes were individually aligned using MACSE v2 (Ranwez et al., 2018) with default parameters, and ribosomal RNA genes using MAFFT (Katoh and Standley, 2013) algorithm FFT-NS-2 with option --adjustdirection. A nucleotide supermatrix was created by concatenating protein-coding and ribosomal RNA genes for the 142 taxa (140 species and 2 subspecies). Phylogenetic inferences were performed with Maximum likelihood (ML) as implemented in IQ-TREE 1.6.8 (Nguyen et al., 2014) with the GTR+G4+F model. Using the resulting topology, divergence time estimation was performed using Phylobayes v4.1c (Lartillot et al., 2013) with strict clock (CL), autocorrelated (LN or TK02), and uncorrelated (UGAM or UCLM) models combined with 18 fossil calibrations (**Table S7**). Three independent Markov chains Monte Carlo (MCMC) analyses starting from a random tree were run until 10,000 generated cycles with trees and associated model parameters sampled every cycle. A burn-in of 25% was applied before constructing the majority-rule Bayesian consensus tree with the readdiv subprogram. Finally, to determine the best-fitting clock model, cross-validation analyses were performed with Phylobayes by splitting the dataset randomly into two parts. Then, parameters of one model were estimated on the first part of the dataset (here representing 90%) and the parameter values were used to compute the likelihood of the second part of the dataset (10%). This procedure was repeated ten times for each model. Finally, the likelihood of each repeated test was computed and summed for each model with the readcv and sumcv subprograms, respectively. The molecular clock model with the highest cross-likelihood scores was considered as the best fitting.

Mitochondrial diversity and barcoding gap analyses

To check if a threshold between intraspecific variation and interspecific divergence could be determined across Carnivora (Meyer and Paulay,

2005), two mitochondrial barcoding datasets were assembled from all COX1 and CYTB sequences available for Carnivora plus the corresponding sequences for the two subspecies of O. megalotis and P. cristata, respectively. After aligning each barcoding dataset with MACSE v2, ML phylogenetic inferences were performed with IQ-TREE 1.6.6 using the optimal substitution model as determined by ModelFinder (Kalyaanamoorthy et al., 2017). Then, pairwise patristic distances between all individuals were calculated from the resulting ML phylogram. Finally, based on the actual taxonomic assignment, patristic distances were considered as intraspecific variation between two individuals belonging to the same species and as interspecific divergence between individuals of different species.

Short reads and long reads hybrid assembly of reference genomes

Sampling

To construct reference assemblies with high contiguity for the two focal species we selected the best-preserved roadkill samples: NMB TS305 for O. megalotis and NMB TS307 for P. cristata (**Table 1**). Total genomic DNA extractions were performed separately for Illumina short-read sequencing and MinION long-read sequencing.

Illumina short-read sequencing

Total genomic DNA extractions were performed from ear necropsies for the two sampled individuals using the DNeasy Blood and Tissue Kit (Qiagen) following manufacturer's instructions. A total amount of 1.0µg DNA per sample was sent as input material for Illumina library preparation and sequencing to Novogene Europe (Cambridge, UK). Sequencing 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 350bp 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 analysed for size distribution by Agilent 2100 Bioanalyzer and quantified using real-time PCR. Since the genome sizes for these two species was estimated to be about 2.5 Gb, Illumina paired-end 250 bp sequencing was run on HiSeqX10 and NovaSeq instruments to obtain about 200 Gb per sample corresponding to a genome depth of coverage of about 80x.

MinION long-read sequencing

Considering the DNA quality required to perform sequencing with Oxford Nanopore Technologies (ONT), a specific protocol to extract DNA from roadkill was designed (Tilak et al., 2020). First, genomic DNA was extracted by using the classical Phenol-chloroform method. Then, we evaluated the cleanliness of the extractions by using (1) a binocular magnifying glass to check the absence of suspended particles (e.g. hairpieces), and (2) both Nanodrop and Qubit/Nanodrop ratio. To select the longest DNA fragments, we applied a specific ratio of 0.4x of AMPure beads applied (Tilak et al., 2020). Extracted-DNA size was then homogenized using covaris G-tubes. Finally, long-read ONT sequencing was performed through MinION flowcells (FLO-MIN-106) using libraries prepared with the ONT Ligation Sequencing kit SQK-LSK109. For both species, we run MinION sequencing until about 30 Gb per sample were obtained to reach a genome depth of coverage of about 12x.

Hybrid assembly of short and long reads

Short reads were cleaned using Trimmomatic 0.33 (Bolger et al., 2014) by removing low quality bases from their beginning (LEADING:3) and end (TRAILING:3), by removing reads shorter than 50 bp (MINLEN:50). Quality was measured for sliding windows of four base pairs and had to be greater than 15 on average (SLIDINGWINDOW:4:15). For MinION sequencing, base calling of fast5

files were performed using Guppy v3.1.5 (developed by ONT) with the high accuracy option, which is longer but more accurate than the standard fast model (**Fig. S1**). Long read adapters were removed using Porechop v0.2.3 (https://github.com/rrwick/Porechop). To take advantage of both the high accuracy of Illumina short reads sequencing and the size of MinION long reads, assemblies were performed using the MaSuRCA hybrid genome assembler (Zimin et al., 2013). This method transforms large numbers of paired-end reads into a much smaller number of longer 'super-reads' and permits assembling Illumina reads of differing lengths together with longer ONT reads. To illustrate the advantage of using short reads and long reads conjointly, assemblies were also performed with short reads only using SOAP-denovo (Luo et al., 2012) (kmer size=31, default parameters) and gaps between contigs were closed using the abundant paired relationships of short reads with GapCloser 1.12 (Luo et al., 2012). To evaluate genome quality, traditional measures like the number of contigs, the N50, the mean and maximum length were evaluated for 503 mammalian genome assemblies retrieved from NCBI (https://www.ncbi.nlm.nih.gov/assembly) on August 13th, 2019 with filters: "Exclude derived from surveillance project", "Exclude anomalous", "Exclude partial", and using only the RefSeq assembly for Homo sapiens. Finally, we assessed the gene completeness of our assemblies by comparison with the 63 carnivoran assemblies available at NCBI on August 13th, 2019 using Benchmarking Universal Single-Copy Orthologs (BUSCO) v3 (Waterhouse et al., 2018) with the Mammalia OrthoDB 9 BUSCO gene set (Zdobnov et al., 2017) through the gVolante web server (Nishimura et al., 2017).

Species delimitation based on genomic data

Sampling and resequencing

To assess the genetic diversity in P. cristata, we sampled an additional roadkill individual of the South African subspecies P. c. cristata (NMB TS491) and an individual of the East African subspecies P. c. septentrionalis (NMS Z.2018.54) from Tanzania (**Table 1**; **Table S6**). A similar sampling was done for O. megalotis, with an additional roadkill individual of the South African subspecies O. m. megalotis (NMB TS306) and an individual of the East African subspecies O. m. virgatus (FMNH 158128) from Tanzania (**Table 1**; **Table S6**). DNA extractions were performed with the DNeasy Blood and Tissue Kit (Qiagen), following manufacturer's instructions and a total amount of 1.0µg DNA per sample was outsourced to Novogene Europe (Cambridge, UK) for Illumina library preparation and Illumina paired-end 250 bp sequencing on HiSeqX10 and NovaSeq instruments to obtain about 200 Gb per sample (genome depth of coverage of about 80x). The resulting reads were cleaned using Trimmomatic 0.33 with the same parameters as described above.

Heterozygosity and genetic differentiation estimation

In a panmictic population, alleles observed in one individual are shared randomly with other individuals of the same population and the frequencies of homozygous and heterozygous alleles should follow Hardy-Weinberg expectations. However, a structuration in subpopulations leads to a deficiency of heterozygotes (relative to Hardy-Weinberg expectations) in these subpopulations due to inbreeding (Holsinger and Weir, 2009; Walhund, 2010) and thus decreases the polymorphism within the inbred subpopulations with respect to the polymorphism of the global population. Given that, Hudson et al. (Hudson et al., 1992) defined the FST as a measure of polymorphism reduction in two subdivided populations (p within) compared to the population at large (p between).

To assess the p within and p between of the two subspecies of each species (P. cristata and O. megalotis), we compared the heterozygous alleles (SNPs) of two individuals of the same subspecies and the SNPs of two individuals of different subspecies by computing a FST-like

statistic (hereafter called Genetic Differentiation Index: GDI) (**Fig. S2**). In fact, polymorphic sites can be discriminated in four categories: (1) fixed in one individual (e.g. AA/TT); (2) shared with both individuals (e.g. AT/AT); (3) specific to individual 1 (e.g. AT/AA); and (4) specific to individual 2 (e.g. AA/AT). Using these four categories, it is possible to estimate the polymorphism of each individual 1 and 2 and thus estimate a GDI between two individuals of the same population A and the GDI between two individuals of different populations A and B as follows:

$$GDI_{intra\ A} = 1 - \frac{(\pi_{A1} + \pi_{A2})/2}{\pi_{totA}}$$

$$GDI_{intra\ B} = 1 - \frac{(\pi_{B1} + \pi_{B2})/2}{\pi_{totB}}$$

For each species, cleaned short reads of all individuals (the one used to construct the reference genome and the two resequenced from each population) were aligned with their reference genome using BWA-MEM (Li, 2013). BAM files were created and merged using SAMtools (Li et al., 2009). Likely contaminant contigs identified using BlobTools (Laetsch and Blaxter, 2017) (**Fig. S3**, **Tables S8-S9**) and contigs belonging to the X chromosome following BLASTN annotation (-perc_identity 80%, -evalue 10e-20) were removed. Then, 100 regions of 100,000 bp were randomly sampled among contigs longer than 100,000 bp and 10 replicates of this sampling were performed (i.e. 10 x 100 x 100,000 bp = 100 Mb) to assess statistical variance in the estimates. Genotyping of these regions was performed with freebayes v1.3.1-16 (git commit id: g85d7bfc) (Garrison and Marth, 2012) using the parallel mode (Tange, 2011). Only SNPs with freebayes-estimated quality higher than 10 were considered for further analyses. A first GDI estimation comparing the average of the private polymorphisms of the two southern individuals (p within A) and the total polymorphism of the two individuals (p between A) was estimated to control that no genetic structure was observed in the Southern subspecies. Then a global GDI comparing the private polymorphisms of individuals from the two populations (p within AB) and the total polymorphism of the species (the two populations, p between AB) was estimated with one individual from each population (**Fig. S2**). Finally, the two GDI were compared to check if the Southern populations were more structured than the entire populations. To contextualize these results, the same GDI measures were estimated for well-defined species of Carnivora. The species pairs used to make the comparison and thus help gauging the taxonomic status of the bat-eared fox and aardwolf subspecies were selected according to the following criteria: (1) the two species had to be as close as possible, (2) they had both reference genomes and short reads available, (3) their estimated coverage for the two species had to be greater than 20x, and (4) short read sequencing data had to be available for two individuals for one species of the pair. Given that, four species pairs were selected: (1) Canis lupus / Canis aureus (SRR8926747, SRR8926748, SRR7976426; vonHoldt et al. 2016); (2) Ursus maritimus / Ursus arctos (PB43: SRR942203, SRR942290, SRR942298; PB28: SRR942211, SRR942287, SRR942295; Brown Bear: SRR935591, SRR935625, SRR935627; Liu et al. 2014); (3) Lynx pardinus / Lynx lynx (Lynx pardinus LYNX11: ERR1255591-ERR1255594; Lynx lynx LYNX8: ERR1255579-ERR1255582; Lynx lynx LYNX23: ERR1255540-ERR1255549; Abascal et al. 2016); and (4) Panthera leo / Panthera pardus (SRR10009886, SRR836361, SRR3041424; Kim et al. 2016). The exact same GDI estimation protocol was applied to each species pair.

Demographic analyses

Historical demographic variations in effective population size were estimated using the Pairwise Sequentially Markovian Coalescent (PSMC) model implemented in the software PSMC (https://github.com/lh3/psmc) (Li and Durbin, 2011). As described above, cleaned short reads were mapped against the

corresponding reference genome using BWA-MEM (Li, 2013) and genotyping was performed using Freebayes v1.3.1-16 (git commit id: g85d7bfc) (Garrison and Marth, 2012) for the three individuals of each species. VCF files were converted to fasta format using a custom python script, excluding positions with quality below 20 and a depth of coverage below 10x or higher than 200x. Diploid sequences in fasta format were converted into PSMC fasta format using a C++ program written using the BIO++ library (Guéguen et al., 2013) with a block length of 100bp and excluding blocks containing more than 20% missing data as implemented in "fq2psmcfa" (https://github.com/lh3/psmc).

PSMC analyses were run for all other populations testing several -t and -p parameters including -p "4+30*2+4+6+10" (Nadachowska-Brzyska et al., 2013) and -p "4+25*2+4+6" (Kim et al., 2016) but also -p "4+10*3+4", -p "4+20*2+4" and -p "4+20*3+4". Overall, the tendencies were similar but some parameters led to unrealistic differences between the two individuals from the South African population of Otocyon megalotis. We chose to present the results obtained using the parameters -t15 -r4 -p "4+10*3+4". For this parameter setting, the variance in ancestral effective population size was estimated by bootstrapping the scaffolds 100 times. To scale PSMC results, based on several previous studies on large mammals, a mutation rate of 10-8 mutation/site/generation (Ekblom et al., 2018; Gopalakrishnan et al., 2017) and a generation time of two years (Clark, 2005; Koehler and Richardson, 1990; van Jaarsveld, 1993) were selected. Results were plotted in R v3.63 (Team, 2020) using the function "psmc.results" (https://doi.org/10.5061/dryad.0618v/4) (Liu and Hansen, 2017) modifed using ggplot2 (Wickham, 2016) and cowplot (Wilke, 2016).

Phylogenomic inferences

To infer the Carnivora phylogenetic relationships, all carnivoran genomes available on Genbank, the DNAZoo website (https://www.dnazoo.org), and the OrthoMaM database (Scornavacca et al., 2019) as of February 11th, 2020 were downloaded (**Table S10**). In cases where more than one genome was available per species, the assembly with the best BUSCO scores was selected. Then, we annotated our two reference genome assemblies and the other unannotated assemblies using MAKER2 (Holt and Yandell, 2011) following the recommendations of the DNAZoo (https://www.dnazoo.org/post/the-first-million-genes-are-the-hardest-to-make-r). In the absence of available transcriptomic data, this method allowed to leverage the power of homology combined with the thorough knowledge accumulated on the gene content of mammalian genomes. As advised, a mammal-specific subset of UniProtKB/Swiss-Prot, a manually annotated, non-redundant protein sequence database, was used as a reference for this annotation step (Boutet et al., 2016). Finally, the annotated coding sequences (CDSs) recovered for the Southern aardwolf (P. c. cristata) were used to assembled those of the Eastern aardwolf (P. c. septentrionalis) by mapping the resequenced Illumina reads using BWA-MEM (Li, 2013).

Orthologous genes were extracted following the orthology delineation process of the OrthoMaM database (OMM) (Scornavacca et al., 2019). First, for each orthologous gene alignment of OMM, a HMM profile was created via hmmbuild using default parameters of the HMMER toolkit (Eddy, 2011) and all HMM profiles were concatenated and summarized using hmmpress to construct a HMM database. Then, for each CDS newly annotated by MAKER, hmmscan was used on the HMM database to retrieve the best hits among the orthologous gene alignments. For each orthologous gene alignment, the most similar sequences for each species were detected via hmmsearch. Outputs from hmmsearch and hmmscan were discarded if the first hit score was not substantially better than the second (hit2 < 0.9 hit1). This ensures our orthology predictions for the newly annotated CDSs to be robust. Then, the cleaning procedure of the OrthoMaM database was applied to the set of orthologous genes obtained. This process,

implemented in a singularity image (Kurtzer et al., 2017) named OMM_MACSE.sif (Ranwez et al., 2020) is composed of several steps including nucleotide sequence alignment at the amino acid level with MAFFT (Katoh and Standley, 2013), refining alignments to handle frameshifts with MACSE v2 (Ranwez et al., 2018), cleaning of non homologous sequences, and masking of erroneous/dubious part of gene sequences with HMMcleaner (Di Franco et al., 2019). Finally, the last step of the cleaning process was to remove sequences that generated abnormally long branches during gene tree inferences. This was done by reconstructing gene trees using IQ-TREE v1.6.8 (Nguyen et al., 2014) with the MFP option to select the best fitting model for each gene. Then, the sequences generating abnormally long branches were identified and removed by PhylteR (https://github.com/damiendevienne/phylter).
This software allows detecting and removing outliers in phylogenomic datasets by iteratively removing taxa in genes and optimizing a concordance score between individual distance matrices.
Phylogenomic analyses were performed using maximum likelihood (ML) using IQ-TREE 1.6.8 (Nguyen et al., 2014) on the supermatrix resulting from the concatenation of all orthologous genes previously recovered with the TESTNEW option to select the best fitting model for each partition. Two partitions per gene were defined to separate the first two codon positions from the third codon positions. Node supports were estimated with 100 non-parametric bootstrap replicates. Furthermore, gene concordant (gCF) and site concordant (sCF) factors were measured to complement traditional bootstrap node-support measures as recommended in Minh et al. (Minh et al., 2020). For each orthologous gene alignment a gene tree was inferred using IQ-TREE with a model selection and gCF and sCF were calculated using the specific option -scf and -gcf in IQ-TREE (Minh et al., 2020). The gene trees obtained with this analysis were also used to perform a coalescent-based species tree inference using ASTRAL-III (Zhang et al., 2018).

Data access

Genome assemblies, associated SRA data and mitogenomes have been submited to genbank and will be available after publication. The full analytical pipeline, phylogenetic datasets (mitogenomic and genomic), corresponding trees and other supplementary materials will be available from zenodo.org.

Disclosure declaration

The authors declare that they have no competing interests.

Funding

This work was supported by grants from the European Research Council (ERC-2015-CoG-683257 ConvergeAnt project), Investissements d'Avenir of the Agence Nationale de la Recherche (CEMEB: ANR-10-LABX-0004; ReNaBi-IFB: ANR-11-INBS-0013; MGX: ANR-10-INBS-09).

Acknowledgements

We would like to thank Rachid Koual and Amandine Magdeleine for technical help with DNA extractions and library preprations, Aude Caizergues and Nathalie Delsuc for fieldwork assistance, Christian Fontaine, Jean-Christophe Vié (Faune Sauvage, French Guiana), Corine Esser (Fauverie du Mont Faron, Toulon, France), François Catzeflis (ISEM Mammalian Tissue Collection), Adam Ferguson and Bruce Patterson (Field Museum of Natural History, Chicago, USA), Lily Crowley (Hamerton Zoo Park, UK), and Andrew Kitchener (National Museum of Scotland, Edinburgh, UK) for access to tissue samples. We also acknowledge Pierre-Alexandre Gagnaire for helpful discussion on genetic differentiation index and Brian Chase for providing references on African paleoclimate. We thank the Montpellier GenomiX Plateform (MGX) part of the France Génomique National Infrastructure for sequencing data generation. Computational analyses benefited from the Montpellier Bioinformatics Biodiversity platform. We are also grateful to the Institut

Français de Bioinformatique and the Roscoff Bioinformatics platform ABiMS (http://abims.sb-roscoff.fr) for providing help for computing and storage resources. This is contribution ISEM 2020-XXX-SUD of the Institut des Sciences de l'Evolution de Montpellier.

Authors' contributions

RA, BN and FD conceived the ideas and designed methodology, analysed the data, and led the writing of the manuscript; FD, NLA and RA performed fieldwork sampling; MK, RA and FD developed the protocol and performed DNA long-read sequencing; MK performed molecular biology experiments; RA, CS, EC and BN performed the bioinformatic analyses; FD and EC provided access to computational resources. All authors contributed critically to the drafts and gave final approval for publication.

References

⇧Back to summary⇧

200 mammals project. n.d. Comparative Genomics – 200 Mammals.

Abascal F, Corvelo A, Cruz F, Villanueva-Cañas JL, Vlasova A, Marcet-Houben M, Martínez-Cruz B, Cheng JY, Prieto P, Quesada V, Quilez J, Li G, García F, Rubio-Camarillo M, Frias L, Ribeca P, Capella-Gutiérrez S, Rodríguez JM, Câmara F, Lowy E, Cozzuto L, Erb I, Tress ML, Rodriguez-Ales JL, Ruiz-Orera J, Reverter F, Casas-Marce M, Soriano L, Arango JR, Derdak S, Galán B, Blanc J, Gut M, Lorente-Galdos B, Andrés-Nieto M, López-Otín C, Valencia A, Gut I, García JL, Guigó R, Murphy WJ, Ruiz-Herrera A, Marques-Bonet T, Roma G, Notredame C, Mailund T, Albà MM, Gabaldón T, Alioto T, Godoy JA. 2016. Extreme genomic erosion after recurrent demographic bottlenecks in the highly endangered Iberian lynx. *Genome Biol* **17**:251. doi:10.1186/s13059-016-1090-1

Agnarsson I, Kuntner M, May-Collado LJ. 2010. Dogs, cats, and kin: A molecular species-level phylogeny of Carnivora. *Mol Phylogenet Evol* **54**:726–745. doi:10.1016/J.YMPEV.2009.10.033

Aird D, Ross MG, Chen WS, Danielsson M, Fennell T, Russ C, Jaffe DB, Nusbaum C, Gnirke A. 2011. Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biol* **12**:R18. doi:10.1186/gb-2011-12-2-r18

Allen R, Ryan H, Davis BW, King C, Frantz L, Irving-Pease E, Barnett R, Linderholm A, Loog L, Haile J, Lebrasseur O, White M, Kitchener AC, Murphy WJ, Larson G. 2020. A mitochondrial genetic divergence proxy predicts the reproductive compatibility of mammalian hybrids. *Proc R Soc B Biol Sci* **287**:20200690. doi:10.1098/rspb.2020.0690

Allio R, Schomaker-Bastos A, Romiguier J, Prosdocimi F, Nabholz B, Delsuc F. 2020a. MitoFinder: Efficient automated large-scale extraction of mitogenomic data in target enrichment phylogenomics. *Mol Ecol Resour* 1755–0998.13160. doi:10.1111/1755-0998.13160

Allio R, Scornavacca C, Nabholz B, Clamens A-L, Sperling FA, Condamine FL. 2020b. Whole genome shotgun phylogenomics resolves the pattern and timing of swallowtail butterfly evolution. *Syst Biol* **69**:38–60. doi:10.1093/sysbio/syz030

Armstrong EE, Taylor RW, Miller DE, Kaelin CB, Barsh GS, Hadly EA, Petrov D. 2020. Long live the king: chromosome-level assembly of the lion (Panthera leo) using linked-read, Hi-C, and long-read data. *BMC Biol* **18**:3. doi:10.1186/s12915-019-0734-5

Arnason U, Gullberg A, Janke A, Kullberg M. 2007. Mitogenomic analyses of caniform relationships. *Mol Phylogenet Evol* **45**:863–874. doi:10.1016/J.YMPEV.2007.06.019

Atickem A, Stenseth NC, Drouilly M, Bock S, Roos C, Zinner D. 2018. Deep divergence among mitochondrial lineages in African jackals. *Zool Scr* **47**:1–8. doi:10.1111/zsc.12257

Barnett R, Yamaguchi N, Barnes I, Cooper A. 2006. The origin, current diversity and future conservation of the modern lion (*Panthera leo*). *Proc R Soc B Biol Sci* **273**:2119–2125. doi:10.1098/rspb.2006.3555

Batra SS, Levy-Sakin M, Robinson J, Guillory J, Durinck S, Kwok P-Y, Cox LA, Seshagiri S, Song YS, Wall JD. 2019a. Accurate assembly of the olive baboon (*Papio anubis*) genome using long--read and Hi-C data. *bioRxiv* 678771. doi:10.1101/678771

Batra SS, Levy-Sakin M, Robinson J, Guillory J, Durinck S, Kwok P-Y, Cox LA, Seshagiri S, Song YS, Wall JD. 2019b. Accurate assembly of the olive baboon (Papio anubis) genome using long--read and Hi-C data. *bioRxiv* 678771. doi:10.1101/678771

Blaimer BB, LaPolla JS, Branstetter MG, Lloyd MW, Brady SG. 2016. Phylogenomics, biogeography and diversification of obligate mealybug-tending ants in the genus Acropyga. *Mol Phylogenet Evol* **102**:20–29. doi:10.1016/J.YMPEV.2016.05.030

Blanco MB, Greene LK, Williams RC, Andrianandrasana L, Yoder AD, Larsen PA. 2019. Next-generation in situ conservation and educational outreach in Madagascar using a mobile genetics lab. *bioRxiv* 650614. doi:10.1101/650614

Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* **30**:2114–2120. doi:10.1093/bioinformatics/btu170

Boutet E, Lieberherr D, Tognolli M, Schneider M, Bansal P, Bridge AJ, Poux S, Bougueleret L, Xenarios I. 2016. UniProtKB/Swiss-Prot, the manually annotated section of the UniProt KnowledgeBase: How to use the entry view. Humana Press, New York, NY. pp. 23–54. doi:10.1007/978-1-4939-3167-5_2

Caley T, Extier T, Collins JA, Schefuß E, Dupont L, Malaizé B, Rossignol L, Souron A, McClymont EL, Jimenez-Espejo FJ, García-Comas C, Eynaud F, Martinez P, Roche DM, Jorry SJ, Charlier K, Wary M, Gourves PY, Billy I, Giraudeau J. 2018. A two-million-year-long hydroclimatic context for hominin evolution in southeastern Africa. *Nature* **560**:76–79. doi:10.1038/s41586-018-0309-6

Casas-Marce M, Soriano L, López-Bao J V., Godoy JA. 2013. Genetics at the verge of extinction: insights from the Iberian lynx. *Mol Ecol* **22**:5503–5515. doi:10.1111/mec.12498

Chase BM, Niedermeyer EM, Boom A, Carr AS, Chevalier M, He F, Meadows ME, Ogle N, Reimer PJ. 2019. Orbital controls on Namib Desert hydroclimate over the past 50,000 years. *Geology* **47**:867–871. doi:10.1130/G46334.1

Chevalier M, Chase BM. 2015. Southeast African records reveal a coherent shift from high- to low-latitude forcing mechanisms along the east African margin across last glacial-interglacial transition. *Quat Sci Rev* **125**:117–130. doi:10.1016/j.quascirev.2015.07.009

Clark HO. 2005. Otocyon megalotis. *Mamm Species* 1–5. doi:10.1644/1545-1410(2005)766[0001:OM]2.0.CO;2

Datema E, Hulzink RJM, Blommers L, Valle-Inclan JE, Orsouw N van, Wittenberg AHJ, Vos M de. 2016. The megabase-sized fungal genome of Rhizoctonia solani assembled from nanopore reads only. *bioRxiv* 084772. doi:10.1101/084772

De Queiroz K. 2007. Species concepts and species delimitation. *Syst Biol* **56**:879–886. doi:10.1080/10635150701701083

Dehghani R, Wanntorp L, Pagani P, Källersjö M, Werdelin L, Veron G. 2008. Phylogeography of the white-tailed mongoose (Herpestidae, Carnivora, Mammalia) based on partial sequences of the mtDNA control region. *J Zool* **276**:385–393. doi:10.1111/j.1469-7998.2008.00502.x

Delisle I, Strobeck C. 2005. A phylogeny of the Caniformia (order Carnivora) based on 12 complete protein-coding mitochondrial genes. *Mol Phylogenet Evol* **37**:192–201. doi:10.1016/J.YMPEV.2005.04.025

deMenocal PB. 2004. African climate change and faunal evolution during the Pliocene-Pleistocene. *Earth Planet Sci Lett* **220**:3–24. doi:10.1016/S0012-821X(04)00003-2

deMenocal PB. 1995. Plio-Pleistocene African climate. *Science (80-)*. doi:10.1126/science.270.5233.53

Di Genova A, Ruz GA, Sagot M-F, Maass A. 2018. Fast-SG: an alignment-free algorithm for hybrid assembly. *Gigascience* **7**. doi:10.1093/gigascience/giy048

Di Franco A, Poujol R, Baurain D, Philippe H. 2019. Evaluating the usefulness of alignment filtering methods to reduce the impact of errors on evolutionary inferences. *BMC Evol Biol* **19**:21. doi:10.1186/s12862-019-1350-2

Díaz-Viraqué F, Pita S, Greif G, de Souza R de CM, Iraola G, Robello C. 2019. Nanopore sequencing significantly improves genome assembly of the protozoan parasite *Trypanosoma cruzi*. *Genome Biol Evol* **11**:1952–1957. doi:10.1093/gbe/evz129

DISCOVAR | Assemble genomes, find variants. n.d. https://software.broadinstitute.org/software/discovar/blog/

Doronina L, Churakov G, Shi J, Brosius J, Baertsch R, Clawson H, Schmitz J. 2015. Exploring massive incomplete lineage sorting in Arctoids (Laurasiatheria, Carnivora). *Mol Biol Evol* **32**:msv188. doi:10.1093/molbev/msv188

Dudchenko O, Batra SS, Omer AD, Nyquist SK, Hoeger M, Durand NC, Shamim MS, Machol I, Lander ES, Aiden AP, Aiden EL. 2017. De novo assembly of the Aedes aegypti genome using Hi-C yields chromosome-length scaffolds. *Science (80-)* **356**:92–95. doi:10.1126/science.aal3327

Eddy SR. 2011. Accelerated profile HMM searches. *PLoS Comput Biol* **7**. doi:10.1371/journal.pcbi.1002195

Eizirik E, Murphy WJ, Koepfli K-P, Johnson WE, Dragoo JW, Wayne RK, O'Brien SJ. 2010. Pattern and timing of diversification of the mammalian order Carnivora inferred from multiple nuclear gene sequences. *Mol Phylogenet Evol* **56**:49–63. doi:10.1016/J.YMPEV.2010.01.033

Ekblom R, Brechlin B, Persson J, Smeds L, Johansson M, Magnusson J, Flagstad Ø, Ellegren H. 2018. Genome sequencing and conservation genomics in the Scandinavian wolverine population. *Conserv Biol* **32**:1301–1312. doi:10.1111/cobi.13157

Etherington GJ, Heavens D, Baker D, Lister A, McNelly R, Garcia G, Clavijo B, Macaulay I, Haerty W, Di Palma F. 2020. Sequencing smart: *De novo* sequencing and assembly approaches for a non-model mammal. *Gigascience* **9**. doi:10.1093/GIGASCIENCE/GIAA045

Faurby S, Eiserhardt WL, Svenning J. 2016. Strong effects of variation in taxonomic opinion on diversification analyses. *Methods Ecol Evol* **7**:4–13. doi:10.1111/2041-210X.12449

Flynn JJ, Finarelli JA, Spaulding M. 2010. Phylogeny of the Carnivora and Carnivoramorpha, and the use of the fossil record to enhance understanding of evolutionary transformations In: Goswami A, Friscia A, editors. Carnivoran Evolution. Cambridge: Cambridge University Press. pp. 25–63. doi:10.1017/CBO9781139193436.003

Flynn JJ, Finarelli JA, Zehr S, Hsu J, Nedbal MA. 2005. Molecular phylogeny of the Carnivora (Mammalia): assessing the impact of increased sampling on resolving enigmatic relationships. *Syst Biol* **54**:317–337. doi:10.1080/10635150590923326

Flynn JJ, Nedbal MA. 1998. Phylogeny of the Carnivora (Mammalia): congruence vs incompatibility among multiple data sets. *Mol Phylogenet Evol* **9**:414–426. doi:10.1006/MPEV.1998.0504

Frankham R, Ballou JD, Dudash MR, Eldridge MDB, Fenster CB, Lacy RC, Mendelson JR, Porton IJ, Ralls K, Ryder OA. 2012. Implications of different species concepts for conserving biodiversity. *Biol Conserv* **153**:25–31. doi:10.1016/J.BIOCON.2012.04.034

Galov A, Fabbri E, Caniglia R, Arbanasić H, Lapalombella S, Florijančić T, Bošković I, Galaverni M, Randi E. 2015. First evidence of hybridization between golden jackal (*Canis aureus*) and domestic dog (*Canis familiaris*) as revealed by genetic markers. *R Soc Open Sci* **2**:150450. doi:10.1098/rsos.150450

Galtier N. 2019. Delineating species in the speciation continuum: A proposal. *Evol Appl* **12**:657–663. doi:10.1111/eva.12748

Galtier N, Nabholz B, Glémin S, Hurst GDD. 2009. Mitochondrial DNA as a marker of molecular diversity: a reappraisal. *Mol Ecol* **18**:4541–4550. doi:10.1111/j.1365-294X.2009.04380.x

Gan HM, Falk S, Morales HE, Austin CM, Sunnucks P, Pavlova A. 2019. Genomic evidence of neo-sex chromosomes in the eastern yellow robin. *Gigascience* **8**. doi:10.1093/gigascience/giz111

Garrison E, Marth G. 2012. Haplotype-based variant detection from short-read sequencing.

Gopalakrishnan S, Samaniego Castruita JA, Sinding M-HS, Kuderna LFK, Räikkönen J, Petersen B, Sicheritz-Ponten T, Larson G, Orlando L, Marques-Bonet T, Hansen AJ, Dalén L, Gilbert MTP. 2017. The wolf reference genome sequence (Canis lupus lupus) and its implications for Canis spp. population genomics. *BMC Genomics* **18**:495. doi:10.1186/s12864-017-3883-3

Gopalakrishnan S, Sinding M-HS, Ramos-Madrigal J, Niemann J, Samaniego Castruita JA, Vieira FG, Carøe C, Montero M de M, Kuderna L, Serres A, González-Basallote VM, Liu Y-H, Wang G-

D, Marques-Bonet T, Mirarab S, Fernandes C, Gaubert P, Koepfli K-P, Budd J, Rueness EK, Sillero C, Heide-Jørgensen MP, Petersen B, Sicheritz-Ponten T, Bachmann L, Wiig Ø, Hansen AJ, Gilbert MTP. 2018. Interspecific gene flow shaped the evolution of the genus Canis. *Curr Biol* **28**:3441–3449.e5. doi:10.1016/J.CUB.2018.08.041

Grant KM, Rohling EJ, Westerhold T, Zabel M, Heslop D, Konijnendijk T, Lourens L. 2017. A 3 million year index for North African humidity/aridity and the implication of potential pan-African Humid periods. *Quat Sci Rev* **171**:100–118. doi:10.1016/j.quascirev.2017.07.005

Guéguen L, Gaillard S, Boussau B, Gouy M, Groussin M, Rochette NC, Bigot T, Fournier D, Pouyet F, Cahais V, Bernard A, Scornavacca C, Nabholz B, Haudry A, Dachary L, Galtier N, Belkhir K, Dutheil JY. 2013. Bio++: Efficient extensible libraries and tools for computational molecular evolution. *Mol Biol Evol* **30**:1745–1750. doi:10.1093/molbev/mst097

Guschanski K, Krause J, Sawyer S, Valente LM, Bailey S, Finstermeier K, Sabin R, Gilissen E, Sonet G, Nagy ZT, Lenglet G, Mayer F, Savolainen V. 2013. Next-generation museomics disentangles one of the largest primate radiations. *Syst Biol* **62**:539–554. doi:10.1093/sysbio/syt018

Holsinger KE, Weir BS. 2009. Genetics in geographically structured populations: defining, estimating and interpreting FST. *Nat Rev Genet* **10**:639–650. doi:10.1038/nrg2611

Holt C, Yandell M. 2011. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics* **12**:491. doi:10.1186/1471-2105-12-491

Hudson RR, Slatkin M, Maddison WP. 1992. Estimation of levels of gene flow from DNA sequence data. *Genetics* **132**.

IUCN 2020. 2020. The IUCN Red List of Threatened Species. Version 2020-1. https://www.iucnredlist.org.

Jain M, Koren S, Miga KH, Quick J, Rand AC, Sasani TA, Tyson JR, Beggs AD, Dilthey AT, Fiddes IT, Malla S, Marriott H, Nieto T, O'Grady J, Olsen HE, Pedersen BS, Rhie A, Richardson H, Quinlan AR, Snutch TP, Tee L, Paten B, Phillippy AM, Simpson JT, Loman NJ, Loose M. 2018. Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nat Biotechnol* **36**:338–345. doi:10.1038/nbt.4060

Jain M, Olsen HE, Paten B, Akeson M. 2016. The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. *Genome Biol* **17**:239. doi:10.1186/s13059-016-1103-0

Jansen HJ, Liem M, Jong-Raadsen SA, Dufour S, Weltzien F-A, Swinkels W, Koelewijn A, Palstra AP, Pelster B, Spaink HP, Thillart GE van den, Dirks RP, Henkel C V. 2017. Rapid de novo assembly of the European eel genome from nanopore sequencing reads. *Sci Rep* **7**:7213. doi:10.1038/s41598-017-07650-6

Jeffroy O, Brinkmann H, Delsuc F, Philippe H. 2006. Phylogenomics: the beginning of incongruence? *Trends Genet* **22**:225–231. doi:10.1016/J.TIG.2006.02.003

Jiang JB, Quattrini AM, Francis WR, Ryan JF, Rodríguez E, McFadden CS. 2019. A hybrid de novo assembly of the sea pansy (Renilla muelleri) genome. *Gigascience* **8**. doi:10.1093/gigascience/giz026

Johnson TC, Werne JP, Brown ET, Abbott A, Berke M, Steinman BA, Halbur J, Contreras S, Grosshuesch S, Deino A, Scholz CA, Lyons RP, Schouten S, Damsté JSS. 2016. A progressively wetter climate in southern East Africa over the past 1.3 million years. *Nature* **537**:220–224. doi:10.1038/nature19065

Kadobianskyi M, Schulze L, Schuelke M, Judkewitz B. 2019. Hybrid genome assembly and annotation of Danionella translucida, a transparent fish with the smallest known vertebrate brain. *bioRxiv* 539692. doi:10.1101/539692

Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermiin LS. 2017. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat Methods* **14**:587–589. doi:10.1038/nmeth.4285

Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment oftware version 7: Improvements in performance and usability. *Mol Biol Evol* **30**:772–780. doi:10.1093/molbev/mst010

Kearse M, Moir R, Wilson A, Stones-Havas S, Cheung M, Sturrock S, Buxton S, Cooper A, Markowitz S, Duran C, Thierer T, Ashton B, Meintjes P, Drummond A. 2012. Geneious Basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* **28**:1647–1649. doi:10.1093/bioinformatics/bts199

Kim S, Cho YS, Kim H-M, Chung O, Kim H, Jho S, Seomun H, Kim J, Bang WY, Kim C, An J, Bae CH, Bhak Y, Jeon S, Yoon H, Kim Y, Jun J, Lee H, Cho S, Uphyrkina O, Kostyria A, Goodrich J, Miquelle D, Roelke M, Lewis J, Yurchenko A, Bankevich A, Cho J, Lee S, Edwards JS, Weber JA, Cook J, Kim S, Lee H, Manica A, Lee I, O'Brien SJ, Bhak J, Yeo J-H. 2016. Comparison of carnivore, omnivore, and herbivore mammalian genomes with a new leopard assembly. *Genome Biol* **17**:211. doi:10.1186/s13059-016-1071-4

Koehler CE, Richardson PRK. 1990. Proteles cristatus. *Mamm Species* 1–6. doi:10.2307/3504197

Koepfli K-P, Jenks SM, Eizirik E, Zahirpour T, Valkenburgh B Van, Wayne RK. 2006. Molecular systematics of the Hyaenidae: Relationships of a relictual lineage resolved by a molecular supermatrix. *Mol Phylogenet Evol* **38**:603–620. doi:10.1016/J.YMPEV.2005.10.017

Koepfli K-P, Paten B, O'Brien SJ. 2015. The Genome 10K Project: A Way Forward. *Annu Rev Anim Biosci* **3**:57–111. doi:10.1146/annurev-animal-090414-014900

Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. 2017. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res* **27**:722–736. doi:10.1101/GR.215087.116

Kumar S, Filipski AJ, Battistuzzi FU, Kosakovsky Pond SL, Tamura K. 2012. Statistics and truth in phylogenomics. *Mol Biol Evol* **29**:457–472. doi:10.1093/molbev/msr202

Kurtzer GM, Sochat V, Bauer MW. 2017. Singularity: Scientific containers for mobility of compute. *PLoS One* **12**:e0177459. doi:10.1371/journal.pone.0177459

Kwan HH, Culibrk L, Taylor GA, Leelakumari S, Tan R, Jackman SD, Tse K, MacLeod T, Cheng D, Chuah E, Kirk H, Pandoh P, Carlsen R, Zhao Y, Mungall AJ, Moore R, Birol I, Marra MA, Rosen DAS, Haulena M, Jones SJM, Kwan HH, Culibrk L, Taylor GA, Leelakumari S, Tan R, Jackman SD, Tse K, MacLeod T, Cheng D, Chuah E, Kirk H, Pandoh P, Carlsen R, Zhao Y, Mungall AJ, Moore R, Birol I, Marra MA, Rosen DAS, Haulena M, Jones SJM. 2019. The Genome of the Steller Sea Lion (Eumetopias jubatus). *Genes (Basel)* **10**:486. doi:10.3390/genes10070486

L Y, YP Z. 2006. Phylogeny of the caniform Carnivora: evidence from multiple genes. *Genetica* **127**. doi:10.1007/S10709-005-2482-4

Laetsch DR, Blaxter ML. 2017. BlobTools: Interrogation of genome assemblies. *F1000Research* **6**:1287. doi:10.12688/f1000research.12232.1

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. doi:10.1093/sysbio/syt022

Li G, Davis BW, Eizirik E, Murphy WJ. 2016. Phylogenomic evidence for ancient hybridization in the genomes of living cats (Felidae). *Genome Res* **26**:1–11. doi:10.1101/GR.186668.114

Li H. 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM.

Li H, Durbin R. 2011. Inference of human population history from individual whole-genome sequences. *Nature* **475**:493–496. doi:10.1038/nature10231

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. doi:10.1093/bioinformatics/btp352

Li R, Fan W, Tian G, Zhu H, He L, Cai J, Huang Q, Cai Q, Li B, Bai Y, Zhang Z, Zhang Y, Wang W, Li J, Wei F, Li H, Jian M, Li J, Zhang Z, Nielsen R, Li D, Gu W, Yang Z, Xuan Z, Ryder OA, Leung FCC, Zhou Y, Cao J, Sun X, Fu Y, Fang X, Guo X, Wang B, Hou R, Shen F, Mu B, Ni P, Lin R, Qian W, Wang G, Yu C, Nie W, Wang J, Wu Z, Liang H, Min J, Wu Q, Cheng S, Ruan J, Wang M, Shi Z, Wen M, Liu B, Ren X, Zheng H, Dong D, Cook K, Shan G, Zhang H, Kosiol C, Xie X, Lu Z, Zheng H, Li Y, Steiner CC, Lam TTY, Lin S, Zhang Q, Li G, Tian J, Gong T, Liu H, Zhang D, Fang L, Ye C, Zhang J, Hu W, Xu A, Ren Y, Zhang G, Bruford MW, Li Q, Ma L, Guo Y, An N, Hu Y, Zheng Y, Shi Y, Li Z, Liu Q, Chen Y, Zhao J, Qu N, Zhao S, Tian F, Wang X, Wang H, Xu L, Liu X, Vinar T, Wang Y, Lam TW, Yiu SM, Liu S, Zhang H, Li D, Huang Y, Wang X, Yang G, Jiang Z, Wang J, Qin N, Li L, Li J, Bolund L, Kristiansen K, Wong GKS, Olson M, Zhang X, Li S, Yang H, Wang J, Wang J. 2010. The sequence and de novo assembly of the giant panda genome. *Nature* **463**:311–317. doi:10.1038/nature08696

Lim S, Chase BM, Chevalier M, Reimer PJ. 2016. 50,000 years of vegetation and climate change in the southern Namib Desert, Pella, South Africa. *Palaeogeogr Palaeoclimatol Palaeoecol* **451**:197–209. doi:10.1016/j.palaeo.2016.03.001

Lindblad-Toh K, Garber M, Zuk O, Lin MF, Parker BJ, Washietl S, Kheradpour P, Ernst J, Jordan G, Mauceli E, Ward LD, Lowe CB, Holloway AK, Clamp M, Gnerre S, Alföldi J, Beal K, Chang J, Clawson H, Cuff J, Di Palma F, Fitzgerald S, Flicek P, Guttman M, Hubisz MJ, Jaffe DB, Jungreis I, Kent WJ, Kostka D, Lara M, Martins AL, Massingham T, Moltke I, Raney BJ, Rasmussen MD, Robinson J, Stark A, Vilella AJ, Wen J, Xie X, Zody MC, Worley KC, Kovar CL, Muzny DM, Gibbs RA, Warren WC, Mardis ER, Weinstock GM, Wilson RK, Birney E, Margulies EH, Herrero J, Green ED, Haussler D, Siepel A, Goldman N, Pollard KS, Pedersen JS, Lander ES, Kellis M, Baldwin J, Bloom T, Chin CW, Heiman D, Nicol R, Nusbaum C, Young S, Wilkinson J, Cree A, Dihn HH, Fowler G, Jhangiani S, Joshi V, Lee S, Lewis LR, Nazareth L V., Okwuonu G, Santibanez J, Delehaunty K, Dooling D, Fronik C, Fulton L, Fulton B, Graves T, Minx P, Sodergren E. 2011. A high-resolution map of human evolutionary constraint using 29 mammals. *Nature* **478**:476–482. doi:10.1038/nature10530

Liu S, Hansen MM. 2017. PSMC (pairwise sequentially Markovian coalescent) analysis of RAD (restriction site associated DNA) sequencing data. *Mol Ecol Resour* **17**:631–641. doi:10.1111/1755-0998.12606

Liu S, Lorenzen ED, Fumagalli M, Li B, Harris K, Xiong Z, Zhou L, Korneliussen TS, Somel M, Babbitt C, Wray G, Li J, He W, Wang Z, Fu W, Xiang X, Morgan CC, Doherty A, O'Connell MJ, McInerney JO, Born EW, Dalén L, Dietz R, Orlando L, Sonne C, Zhang G, Nielsen R, Willerslev E, Wang J. 2014. Population genomics reveal recent speciation and rapid evolutionary adaptation in polar bears. *Cell* **157**:785–794. doi:10.1016/J.CELL.2014.03.054

Lorenzen ED, Heller R, Siegismund HR. 2012. Comparative phylogeography of African savannah ungulates 1. *Mol Ecol* **21**:3656–3670. doi:10.1111/j.1365-294X.2012.05650.x

Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J, He G, Chen Y, Pan Q, Liu Y, Tang J, Wu G, Zhang H, Shi Y, Liu Y, Yu C, Wang B, Lu Y, Han C, Cheung DW, Yiu S-M, Peng S, Xiaoqian Z, Liu G, Liao X, Li Y, Yang H, Wang J, Lam T-W, Wang J. 2012. SOAPdenovo2: An empirically improved memory-efficient short-read de novo assembler. *Gigascience* **1**:18. doi:10.1186/2047-217X-1-18

Maigret TA. 2019. Snake scale clips as a source of high quality DNA suitable for RAD sequencing. *Conserv Genet Resour* **11**:373–375. doi:10.1007/s12686-018-1019-y

Maslin MA, Brierley CM, Milner AM, Shultz S, Trauth MH, Wilson KE. 2014. East african climate pulses and early human evolution. *Quat Sci Rev*. doi:10.1016/j.quascirev.2014.06.012

Meredith RW, Janečka JE, Gatesy J, Ryder OA, Fisher CA, Teeling EC, Goodbla A, Eizirik E, Simão TLL, Stadler T, Rabosky DL, Honeycutt RL, Flynn JJ, Ingram CM, Steiner C, Williams TL, Robinson TJ, Burk-Herrick A, Westerman M, Ayoub NA, Springer MS, Murphy WJ. 2011. Impacts of the cretaceous terrestrial revolution and KPg extinction on mammal diversification. *Science (80-)* **334**:521–524. doi:10.1126/SCIENCE.1211028

Meyer CP, Paulay G. 2005. DNA barcoding: error rates based on comprehensive sampling. *PLoS Biol* **3**:e422. doi:10.1371/journal.pbio.0030422

Michael TP, Jupe F, Bemm F, Motley ST, Sandoval JP, Lanz C, Loudet O, Weigel D, Ecker JR. 2018. High contiguity Arabidopsis thaliana genome assembly with a single nanopore flow cell. *Nat Commun* **9**:541. doi:10.1038/s41467-018-03016-2

Miller JM, Hallager S, Monfort SL, Newby J, Bishop K, Tidmus SA, Black P, Houston B, Matthee CA, Fleischer RC, Hallager S, Monfort SL, Newby J, Bishop ÁK, Tidmus SA, Black P, Houston ÁB, Matthee CA. 2011. Phylogeographic analysis of nuclear and mtDNA supports subspecies designations in the ostrich (Struthio camelus). *Conserv Genet* **12**:423–431. doi:10.1007/s10592-010-0149-x

Minh BQ, Hahn MW, Lanfear R. 2020. New methods to calculate concordance factors for phylogenomic datasets. *Mol Biol Evol*. doi:10.1093/molbev/msaa106

Nadachowska-Brzyska K, Burri R, Olason PI, Kawakami T, Smeds L, Ellegren H. 2013. Demographic divergence history of pied flycatcher and collared flycatcher inferred from whole-genome re-sequencing data. *PLoS Genet* **9**:e1003942. doi:10.1371/journal.pgen.1003942

Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ. 2014. IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol Biol Evol* **32**:268–274. doi:10.1093/molbev/msu300

Nicholls SM, Quick JC, Tang S, Loman NJ. 2019. Ultra-deep, long-read nanopore sequencing of mock microbial community standards. *Gigascience* **8**. doi:10.1093/gigascience/giz043

Nishimura O, Hara Y, Kuraku S. 2017. gVolante for standardizing completeness assessment of genome and transcriptome assemblies. *Bioinformatics* **33**:3635–3637. doi:10.1093/bioinformatics/btx445

Parker J, Helmstetter AJ, Devey D, Wilkinson T, Papadopulos AST. 2017. Field-based species identification of closely-related plants using real-time nanopore sequencing. *Sci Rep* **7**:8345. doi:10.1038/s41598-017-08461-5

Peng R, Zeng B, Meng X, Yue B, Zhang Z, Zou F. 2007. The complete mitochondrial genome and phylogenetic analysis of the giant panda (Ailuropoda melanoleuca). *Gene* **397**:76–83. doi:10.1016/J.GENE.2007.04.009

Périquet S, Roxburgh L, le Roux A, Collinson WJ. 2018. Testing the value of citizen science for roadkill studies: A case study from South Africa. *Front Ecol Evol* **6**:15. doi:10.3389/fevo.2018.00015

Pomerantz A, Peñafiel N, Arteaga A, Bustamante L, Pichardo F, Coloma LA, Barrio-Amorós CL, Salazar-Valenzuela D, Prost S. 2018. Real-time DNA barcoding in a rainforest using nanopore sequencing: opportunities for rapid biodiversity assessments and local capacity building. *Gigascience* **7**. doi:10.1093/gigascience/giy033

Ranwez V, Chantret N, Delsuc F. 2020. Aligning protein-coding nucleotide sequences with MACSE. *Methods Mol Biol* **In press**.

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. doi:10.1093/molbev/msy159

Ravinet M, Westram A, Johannesson K, Butlin R, André C, Panova M. 2016. Shared and nonshared genomic divergence in parallel ecotypes of *Littorina saxatilis* at a local scale. *Mol Ecol* **25**:287–305. doi:10.1111/mec.13332

Rice ES, Green RE. 2019. New Approaches for Genome Assembly and Scaffolding. *Annu Rev Anim Biosci* **7**:17–40. doi:10.1146/annurev-animal-020518-115344

Rohland N, Pollack JL, Nagel D, Beauval C, Airvaux J, Pääbo S, Hofreiter M. 2005. The population history of extant and extinct hyenas. *Mol Biol Evol* **22**:2435–2443. doi:10.1093/molbev/msi244

Roux C, Fraïsse C, Romiguier J, Anciaux Y, Galtier N, Bierne N. 2016. Shedding light on the grey zone of speciation along a continuum of genomic divergence. *PLOS Biol* **14**:e2000234. doi:10.1371/journal.pbio.2000234

Rybczynski N, Dawson MR, Tedford RH. 2009. A semi-aquatic Arctic mammalian carnivore from the Miocene epoch and origin of Pinnipedia. *Nature* **458**:1021–1024. doi:10.1038/nature07985

Sato JJ, Wolsan M, Minami S, Hosoda T, Sinaga MH, Hiyama K, Yamaguchi Y, Suzuki H. 2009. Deciphering and dating the red panda's ancestry and early adaptive radiation of Musteloidea. *Mol Phylogenet Evol* **53**:907–922. doi:10.1016/J.YMPEV.2009.08.019

Sato JJ, Wolsan M, Suzuki H, Hosoda T, Yamaguchi Y, Hiyama K, Kobayashi M, Minami S. 2006. Evidence from nuclear DNA sequences sheds light on the phylogenetic relationships of Pinnipedia: single origin with affinity to Musteloidea. *Zoolog Sci* **23**:125–146. doi:10.2108/zsj.23.125

Schröder C, Bleidorn C, Hartmann S, Tiedemann R. 2009. Occurrence of Can-SINEs and intron sequence evolution supports robust phylogeny of pinniped carnivores and their terrestrial relatives. *Gene* **448**:221–226. doi:10.1016/J.GENE.2009.06.012

Scornavacca C, Belkhir K, Lopez J, Dernat R, Delsuc F, Douzery EJP, Ranwez V. 2019. OrthoMaM v10: Scaling-up orthologous coding sequence and exon alignments with more than one hundred mammalian genomes. *Mol Biol Evol* **36**:861–862. doi:10.1093/molbev/msz015

Scott AD, Zimin A V., Puiu D, Workman R, Britton M, Zaman S, Caballero M, Read AC, Bogdanove AJ, Burns E, Wegrzyn J, Timp W, Salzberg SL, Neale DB. 2020. The giant sequoia genome and proliferation of disease resistance genes. *bioRxiv* 2020.03.17.995944. doi:10.1101/2020.03.17.995944

Shafin K, Pesout T, Lorig-Roach R, Haukness M, Olsen HE, Bosworth C, Armstrong J, Tigyi K, Maurer N, Koren S, Sedlazeck FJ, Marschall T, Mayes S, Costa V, Zook JM, Liu KJ, Kilburn D, Sorensen M, Munson KM, Vollger MR, Monlong J, Garrison E, Eichler EE, Salama S, Haussler D, Green RE, Akeson M, Phillippy A, Miga KH, Carnevali P, Jain M, Paten B. 2020. Nanopore sequencing and the Shasta toolkit enable efficient de novo assembly of eleven human genomes. *Nat Biotechnol* 1–10. doi:10.1038/s41587-020-0503-6

Shilling F, Perkins SE, Collinson W. 2015. Wildlife/Roadkill Observation and Reporting SystemsHandbook of Road Ecology. Chichester, UK: John Wiley & Sons, Ltd. pp. 492–501. doi:10.1002/9781118568170.ch62

Spitzer M, Wildenhain J, Rappsilber J, Tyers M. 2014. BoxPlotR: a web tool for generation of box plots. *Nat Methods* **11**:121–122. doi:10.1038/nmeth.2811

Srivathsan A, Baloğlu B, Wang W, Tan WX, Bertrand D, Ng AHQ, Boey EJH, Koh JJY, Nagarajan N, Meier R. 2018. A MinION^{TM}-based pipeline for fast and cost-effective DNA barcoding. *Mol Ecol Resour* **18**:1035–1049. doi:10.1111/1755-0998.12890

Tan MH, Austin CM, Hammer MP, Lee YP, Croft LJ, Gan HM. 2018. Finding Nemo: hybrid assembly with Oxford Nanopore and Illumina reads greatly improves the clownfish (Amphiprion ocellaris) genome assembly. *Gigascience* **7**. doi:10.1093/gigascience/gix137

Tange O. 2011. Gnu parallel-the command-line power tool. *USENIX Mag* **36**:42–47.

Team R core. 2020. R: A language and environment for statistical computing.

Tilak M-K, Allio R, Delsuc F. 2020. An optimized protocol for sequencing mammalian roadkill tissues with Oxford Nanopore Technology (ONT). doi:10.17504/PROTOCOLS.IO.BEIXJCFN

Tilak M-K, Botero-Castro F, Galtier N, Nabholz B. 2018. Illumina library preparation for sequencing the GC-rich fraction of heterogeneous genomic DNA. *Genome Biol Evol* **10**:616–622. doi:10.1093/gbe/evy022

Tilak M-K, 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. *Conserv Genet Resour* **7**:37–40. doi:10.1007/s12686-014-0338-x

Trauth MH, Larrasoaña JC, Mudelsee M. 2009. Trends, rhythms and events in Plio-Pleistocene African climate. *Quat Sci Rev* **28**:399–411. doi:10.1016/j.quascirev.2008.11.003

van Berkum NL, Lieberman-Aiden E, Williams L, Imakaev M, Gnirke A, Mirny LA, Dekker J, Lander ES. 2010. Hi-C: A method to study the three-dimensional architecture of genomes. *J Vis Exp* e1869. doi:10.3791/1869

van Jaarsveld AS. 1993. A comparative investigation of hyaena and aardwolf life-histories, with notes on spotted hyaena mortality patterns. *Trans R Soc South Africa* **48**:219–232. doi:10.1080/00359199309520272

Vaser R, Sović I, Nagarajan N, Šikić M. 2017. Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Res* **27**:737–746. doi:10.1101/GR.214270.116

vonHoldt BM, Kays R, Pollinger JP, Wayne RK. 2016. Admixture mapping identifies introgressed genomic regions in North American canids. *Mol Ecol* **25**:2443–2453. doi:10.1111/mec.13667

Walhund S. 2010. Zusammensetzung von populationen und korrelationserscheinungen vom standpunkt der vererbungslehre aus betrachtet. *Hereditas* **11**:65–106. doi:10.1111/j.1601-5223.1928.tb02483.x

Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, Cuomo CA, Zeng Q, Wortman J, Young SK, Earl AM. 2014. Pilon: an Integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One* **9**:e112963. doi:10.1371/journal.pone.0112963

Walton LR, Joly DO. 2003. Canis mesomelas. *Mamm Species* **715**:1–9. doi:10.1644/715

Wang W, Das A, Kainer D, Schalamun M, Morales-Suarez A, Schwessinger B, Lanfear R. 2020. The draft nuclear genome assembly of Eucalyptus pauciflora: a pipeline for comparing de novo assemblies. *Gigascience* **9**. doi:10.1093/gigascience/giz160

Waterhouse RM, Seppey M, Simão FA, Manni M, Ioannidis P, Klioutchnikov G, Kriventseva E V, Zdobnov EM. 2018. BUSCO applications from quality assessments to gene prediction and phylogenomics. *Mol Biol Evol* **35**:543–548. doi:10.1093/molbev/msx319

Weisenfeld NI, Yin S, Sharpe T, Lau B, Hegarty R, Holmes L, Sogoloff B, Tabbaa D, Williams L, Russ C, Nusbaum C, Lander ES, Maccallum I, Jaffe DB. 2014. Comprehensive variation discovery in single human genomes. *Nat Genet* **46**:1350–1355. doi:10.1038/ng.3121

Wenger AM, Peluso P, Rowell WJ, Chang P-C, Hall RJ, Concepcion GT, Ebler J, Fungtammasan A, Kolesnikov A, Olson ND, Töpfer A, Alonge M, Mahmoud M, Qian Y, Chin C-S, Phillippy AM, Schatz MC, Myers G, DePristo MA, Ruan J, Marschall T, Sedlazeck FJ, Zook JM, Li H, Koren S, Carroll A, Rank DR, Hunkapiller MW. 2019. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nat Biotechnol* **37**:1155–1162. doi:10.1038/s41587-019-0217-9

Westbury M V, Hartmann S, Barlow A, Wiesel I, Leo V, Welch R, Parker DM, Sicks F, Ludwig A, Dalén L, Hofreiter M. 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. doi:10.1093/molbev/msy037

Wick RR, Judd LM, Holt KE. 2019. Performance of neural network basecalling tools for Oxford Nanopore sequencing. *Genome Biol* **20**:129. doi:10.1186/s13059-019-1727-y

Wickham H. 2016. Ggplot2 : elegant graphics for data analysis. Springer.

Wilke CO. 2016. cowplot: Streamlined plot theme and plot annotations for "ggplot2." *CRAN Repos*.

Wilson DE, Mittermeier RA, Cavallini P. 2009. Handbook of the mammals of the world, Vol. 1. ed. Barcelona: Lynx Edicions.

Zdobnov EM, Tegenfeldt F, Kuznetsov D, Waterhouse RM, Simão FA, Ioannidis P, Seppey M, Loetscher A, Kriventseva E V. 2017. OrthoDB v9.1: cataloging evolutionary and functional annotations for animal, fungal, plant, archaeal, bacterial and viral orthologs. *Nucleic Acids Res* **45**:D744–D749. doi:10.1093/nar/gkw1119

Zhang C, Rabiee M, Sayyari E, Mirarab S. 2018. ASTRAL-III: polynomial time species tree reconstruction from partially resolved gene trees. *BMC Bioinforma 2018 196* **19**:15–30. doi:10.1186/s12859-018-2129-y

Zimin A V., Marçais G, Puiu D, Roberts M, Salzberg SL, Yorke JA. 2013. The MaSuRCA genome assembler. *Bioinformatics* **29**:2669–2677. doi:10.1093/bioinformatics/btt476

Zimin A V., Puiu D, Luo M-C, Zhu T, Koren S, Marçais G, Yorke JA, Dvořák J, Salzberg SL. 2017. Hybrid assembly of the large and highly repetitive genome of Aegilops tauschii, a progenitor of bread wheat, with the MaSuRCA mega-reads algorithm. *Genome Res* **27**:787–792. doi:10.1101/GR.213405.116

⇧Back to summary⇧

– PART III –

Comparative transcriptomic analysis of chitinase gene family in mammals

In this chapter, I present the results of two Master students, Dave Lutgen (M2, Université de Montpellier) and Sophie Teullet (M2, Université de Toulouse) whom I had the opportunity to co-supervise with Frédéric Delsuc during my PhD project. These two Master students were interested in understanding the relative role of both historical contingency and evolutionary convergence in shaping (i) the global gene expression in salivary glands of myrmecophagous mammals, and (ii) the expression of candidate genes for insect chitin digestion (paralogous genes of the Chitinase family in salivary glands and in additional organs). A particular effort has been made by Sophie Teullet to reconstruct the evolution of genes belonging to the chitinase family at the vertebrate and mammalian taxonomic scales. Given that some analyses have been or still need to be repeated, the manuscript presented here is only a first draft with preliminary results and discussions. It nevertheless already provides interesting insights into the molecular mechanisms underlying convergent evolution towards myrmecophagy in anteaters and pangolins specifically.

Comparative transcriptomics reveals divergent paths of chitinase evolution underlying dietary convergence in ant-eating mammals

Rémi Allio[1,□,*], Sophie Teullet[1,□], Dave Lutgen[1,2,□], Amandine Magdeleine[1], Rachid Koual[1], Marie-Ka Tilak[1], Christopher A. Emerling[1], Tristan Lefébure[3], and Frédéric Delsuc[1,*]

[1]*Institut des Sciences de l'Evolution de Montpellier (ISEM), CNRS, IRD, EPHE, Université de Montpellier, Montpellier, France*
[2]*Department of Population Ecology, Institute of Ecology and Evolution, Friedrich Schiller University Jena, Jena, Germany*
[3]*Laboratoire d'Ecologie des Hydrosystèmes Naturels et Anthropisés (LEHNA), Université de Lyon, Université Claude Bernard Lyon 1, CNRS UMR 5023, ENTPE, Villeurbanne, France*
[□]*Equal contribution*

***Correspondence:** remi.allio@umontpellier.fr; frederic.delsuc@umontpellier.fr

Abstract

Ant-eating mammals represent a textbook example of convergent morphological evolution. Among them, anteaters and pangolins exhibit the most extreme convergent phenotypes with complete tooth loss, elongated skulls, protrusive tongues, and powerful claws to rip open ant and termite nests. These two placental mammal lineages also possess hypertrophied salivary glands, which produce large quantities of saliva to capture and digest their social insect prey. Despite this remarkable convergence, comparative genomic analyses have shown that anteaters and pangolins differ in their chitinase gene (CHIA) repertoires. While the lesser anteater (*Tamandua tetradactyla*) harbours four functional CHIA paralogs (CHIA1, CHIA2, CHIA3, and CHIA4), Asian pangolins (*Manis* spp.) have only one functional paralog (CHIA5). A recent transcriptomic analysis has shown that CHIA5 is highly expressed in all major digestive organs (stomach, pancreas, large intestine, and liver) of the Malayan pangolin (*Manis javanica*), including its tongue and salivary glands. Here, we present the first comparative transcriptomic analysis of salivary glands in 23 species of placental mammals, including new ant-eating species and close relatives, together with complementary RNAseq data for the major digestive organs of the lesser anteater. Our results on digestive enzyme gene expression show that salivary glands play a major role in the adaptation to the myrmecophagous diet. A detailed analysis of nine paralogous chitinase genes revealed that convergently evolved pangolins and anteaters express different chitinases in their hypertrophied salivary glands and other digestive organs. Indeed, we confirm that CHIA5 is overexpressed in Malayan pangolin salivary glands and other digestive organs, whereas the lesser anteater exhibits high levels of CHIA3 and CHIA4 expression in salivary glands and other digestive organs. Overall, our results demonstrate that divergent molecular mechanisms underlie convergent adaptation to the ant-eating diet in pangolins and anteaters, highlighting the role of historical contingency and molecular tinkering of the chitin-digestive enzyme toolkit in this classical example of convergent evolution.

⇧Back to summary⇧

Introduction

The phenomenon of evolutionary convergence is a fascinating process in which distantly related species independently acquire similar characteristics in response to the same selection pressures. An historical question illustrated by the debate between Stephen Jay Gould (Gould 2002) and Simon Conway Morris (Conway Morris 1998), resides in the relative contribution of historical contingency and evolutionary convergence in the evolution of current biodiversity. While Gould (1990, 2002) argued that the evolution of species strongly depends on the characteristics inherited from their ancestors (historical contingency), Conway Morris (1998) retorqued that convergent evolution is one of the dominant processes leading to biodiversity evolution. Despite the huge diversity of organisms found on Earth and the numerous potential possibilities to adapt to similar conditions, the strong deterministic force of natural selection led to numerous cases of recurrent phenotypic adaptations (Losos, 2011, 2017; McGhee 2011). However, the role of historical contingency and evolutionary tinkering in convergent evolution has long been recognized, evolution proceeding from available material through natural selection often leading to structural and functional imperfections (Jacob 1977). As first pointed out by François Jacob (1977), molecular tinkering seems to be particularly frequent and seems to have shaped the evolutionary history of a number of protein families (Pillai et al. 2020; Xie et al. 2020). Indeed, if in some cases, convergent phenotypes can be associated with similar or identical mutations in the same genes occurring in independent lineages (Arendt and Reznick, 2008), in other cases, they appear to arise by diverse molecular paths (e.g. Christin et al., 2010). Hence, both historical contingency and evolutionary convergence seems to have impacted the evolution of the current biodiversity (Blount et al. 2018) and the major question relies on evaluating the relative impact of these two evolutionary processes.

A notable example of convergent evolution is the specialized ant- and/or termite-eating diet (i.e. myrmecophagy) in placental mammals (Reiss, 2001). Within placental mammals, over 200 species include ants and termites in their regime, but only 22 of them can be considered as specialized myrmecophagous mammals eating more than 90% of social insects (Redford 1987). Historically, based on morphological characters, ant-eating mammals were considered to be monophyletic (i.e. Edentata, Novacek, 1992; O'Leary et al., 2013), but molecular phylogenetic evidence now strongly supports their polyphyly (e.g. Delsuc et al., 2002; Springer et al., 2003; Meredith et al., 2011). This highly-specialized diet has independently evolved in five placental orders, armadillos (Cingulata), anteaters (Pilosa), aardvarks (Tubulidentata), pangolins (Pholidota), and aardwolves (Carnivora). As a consequence of foraging for small sized preys (Redford, 1987), these animals have evolved similar but convergent morphological adaptations such as powerful claws used to dig into ant and termite nests, tooth reduction culminating in complete tooth loss in anteaters and pangolins (Ferreira-Cardoso et al. 2019), an elongated muzzle with an extensible tongue (Ferreira-Cardoso et al. 2020), and viscous saliva produced by hypertrophied salivary glands (Reiss, 2001). Due to strong energetic

constraints imposed by a nutritionally poor diet, myrmecophagous mammals also share relatively low metabolic rates and might thus require specific adaptations to extract proteins from the chitinous exoskeletons of their preys (McNab, 1984). Previous studies have shown that chitinase genes are present in the mammalian genome and may play an important digestive function in insectivorous species (Jeuniaux, 1971; Bussink et al., 2007; Janiak et al. 2018). Elevated levels of digestive enzyme gene expression have notably been observed in placental mammal salivary glands. For instance in bat salivary glands, studies have shown that dietary adaptations can be associated with elevated expression levels in carbohydrase, lipase, and protease genes (Francischetti et al., 2013; Phillips et al., 2014; Vandewege at al. 2020).

In placental mammals, the salivary glands are composed of three major gland pairs (parotid, sublingual, and submandibular) and hundreds of minor salivary glands (Tucker, 1958). In most myrmecophagous placental lineages, it has been shown that hypertrophied cervical salivary glands are the primary source of salivary production (Kingdon, 1971). These enlarged horseshoe-shaped glands that extend posteriorly along the side of the neck and ventrally over the chest are homologous to the human parotid glands. The only notable exception is the aardwolf (*Proteles cristatus*), a species in which the mandibular gland is enlarged but not the parotid gland (Kingdon, 1971). In the Malayan pangolin (*Manis javanica*), recent transcriptomic studies have shown that genes associated with digestive enzymes are highly expressed in their salivary glands, which supports the hypothesis that the enlarged cervical gland plays an important functional role in social insect digestion (Ma et al., 2017, 2019). This result also finds support in a study on the molecular evolution of the chitinase genes across 107 placental mammals that revealed the likely existence of a repertoire of five functional paralogous chitinase (CHIA) genes in the placental ancestor that was subsequently shaped through multiple pseudogenization events associated with dietary adaptation during the placental radiation (Emerling et al., 2018). The widespread gene loss observed in carnivorous and herbivorous lineages resulted in a general correlation between the number of functional CHIA paralogs and the percentage of invertebrates in the diet across placentals (Emerling et al. 2018). Pangolins nevertheless appear as an exception as the two investigated species (*M. javanica* and *Manis pentadactyla*) possess only one functional CHIA paralog (CHIA5) whereas other myrmecophagous species such as the lesser anteater (*Tamandua tetradactyla*) and the aardvark (*Orycteropus afer*) possess respectively four (CHIA1-4) and five (CHIA1-5) functional paralogs. The presence of the sole CHIA5 in pangolins was interpreted as the consequence of historical contingency with the probable loss of CHIA1-4 functionality in the last common ancestor of Pholidota and Carnivora (Emerling et al. 2018). The fact that the functional CHIA5 paralog was found to be highly expressed in the main digestive organs of the Malayan pangolin (Ma et al., 2017, 2019) suggests that pangolins compensate their reduced chitinase repertoire by an increased pleiotropic expression of their only remaining functional paralog in multiple organs.

To test this hypothesis, we first reconstructed the detailed evolutionary history of the chitinase gene family in mammals. Then, we conducted a comparative transcriptomic analysis of salivary glands in 27 mammal species including 17 newly generated transcriptomes from myrmecophagous mammals and other species. Finally, we compared the expression of chitinase paralogs in different organs between the nine-banded armadillo (*Dasypus novemcinctus*), the Malayan pangolin (*M. javanica*), and the lesser anteater (*T. tetradactyla*) for which we produced 12 new transcriptomes from eight additional organs.

Material and Methods

Transcriptomic dataset assemblies

Salivary gland transcriptomes - Biopsies of salivary glands preserved in RNAlater were obtained from euthanized animals, deceased zoo animals, and fresh roadkill for 17 individuals representing 12 placental mammal species (**Table 1**). Total RNA was extracted from individual salivary gland tissue samples using the RNeasy extraction kit (Qiagen, Germany). Then, RNA-seq library construction and Illumina sequencing on a HiSeq 2500 system using paired-end 2x125bp reads were conducted by the Montpellier GenomiX platform (MGX) resulting in 17 newly produced salivary gland transcriptomes. This sampling was completed with the 13 mammalian salivary gland transcriptomes available as paired-end sequencing reads in the Short Read Archive (SRA) of the National Center for Biotechnology Information (NCBI) as of April 15th, 2019 representing an additional 11 species (**Table 1**). This taxon sampling includes representatives from all major mammal superorders Afrotheria (n = 3), Xenarthra (n = 4), Euarchontoglires (n = 3), and Laurasiatheria (n = 13) and covers six different diet categories: carnivory (n = 4), frugivory (n = 1), herbivory (n = 2), insectivory (n = 4), myrmecophagy (n = 6), and omnivory (n = 6). Four of the five lineages in which myrmecophagous mammals evolved are represented: aardwolf (*P. cristatus*, Carnivora), Malayan pangolin (*M. javanica*, Pholidota), southern naked-tailed armadillo (*Cabassous unicinctus*, Cingulata), giant anteater (*Myrmecophaga tridactyla*, Pilosa), and lesser anteater (*T. tetradactyla*, Pilosa). Species replicates in the form of different individuals were collected for the lesser anteater (*T. tetradactyla*; n = 3), and the nine-banded armadillo (*D. novemcinctus*; n = 3). We unfortunately were not able to obtain fresh salivary gland samples from the aardvark (*O. afer*, Tubulidentata), the only missing myrmecophagous lineage in our sampling.

Transcriptomes from additional organs - Tissue biopsies from eight additional organs (testis, lungs, heart, spleen, tongue, stomach, liver, and intestine) were sampled during dissections of three roadkill individuals of lesser anteater (*T. tetradactyla*; **Table 2**). Total RNA extractions from these RNAlater-preserved tissues, RNA-seq library construction, and sequencing were conducted as described above resulting in 12 newly generated transcriptomes. For comparative purposes, 24 additional

transcriptomes of Malayan pangolin (*M. javanica*) representing 16 organs, and 21 transcriptomes of nine-banded armadillo (*D. novemcinctus*) representing eight organs were downloaded from SRA (**Table 2**).

Transcriptome assemblies and quality control - Adapters and low quality reads were removed from raw sequencing data using FASTP v0.19.6 (Chen et al., 2018) using default parameters except for the PHRED score which was defined as "`--qualified_quality_phred ≥ 15`", as suggested by MacManes (2014). This approach has proven to be most effective for *de novo* transcriptome assemblies, because low expression transcripts are not disproportionately removed. Then, de novo assembly was performed on each individual transcriptome sample using Trinity v2.8.4 (Grabherr et al. 2011) using default parameters. For each of the 27 salivary gland transcriptomes, completeness was assessed by the presence of Benchmark Universal Single Copy Orthologs (BUSCOs) based on a dataset of 4,104 single-copy orthologs conserved in over 90% of mammal species (Waterhouse et al., 2018). This pipeline evaluates the percentage of complete, duplicated, fragmented and missing single-copy orthologs within each transcriptome.

Comparative transcriptomics of salivary glands

Transcriptome annotation and orthogroup inference - The 27 salivary gland transcriptome assemblies were annotated following the pipeline implemented in assembly2ORF (https://github.com/ellefeg/assembly2orf). This pipeline combines evidence-based and gene-model-based predictions. First, potential transcripts of protein-coding genes are extracted based on similarity searches (BLAST) against the peptides of Metazoa found in Ensembl (Yates et al. 2020). Then, using both protein similarity and exonerate functions (Slater & Birney 2005), a frameshift correction is applied to candidate transcripts. Based on homology information inferred from both BLAST and Hmmscan searches, candidate open reading frames (ORFs) are annotated and selected using TransDecoder (https://github.com/TransDecoder/TransDecoder). Finally, to be able to compare the transcriptomes obtained from all species, we relied on the inference of gene orthogroups. The orthogroup inference for the translated candidate ORFs was performed using OrthoFinder v2 (Emms & Kelly 2019) using IQ-TREE (Nguyen et al. 2015) for gene tree reconstructions. For expression analyses, orthogroups containing more than 20 copies for at least one species were discarded. Finally, we constructed two distinct datasets: in the first dataset, only orthogroups containing at least one species per taxonomic order and per diet category were conserved. In the second dataset, only the orthogroups containing at least one sequence per species were conserved.

Gene expression analyses - Quantification of transcript expression was performed on Trinity assemblies with Kallisto v.0.46.1 (Bray et al., 2016) using the *align_and_estimate_abundance.pl* script provided in the Trinity suite (Grabherr et al., 2011). Kallisto relies on pseudo-alignments of the

Table 1 | Salivary gland tissues sequenced for the project

Species	Origin	Sample	Country
Cabassous unicinctus	CABuniCU04	M2757	French Guiana
Canis lupus familiaris	SRA	SRR5889344	USA
Dasypus novemcinctus	DASnovFK06A	FK06	USA
Dasypus novemcinctus	DASnovFK06C	FK06	USA
Dasypus novemcinctus	DASnovFK08A	FK08	USA
Desmodus rotundus	SRA	SRR606902	Brazil
Desmodus rotundus	SRA	SRR606908	Brazil
Desmodus rotundus	SRA	SRR606911	Brazil
Elephantulus myurus	ELEmyuNA02	TDR	South Africa
Erinaceus europaeus	ERIeurRA02	RA03	France
Felis catus	SRA	SRR3218717	USA
Genetta genetta	GENgenRA01	RA02	France
Homo sapiens	SRA	SRR1957200	USA
Macrotus californicus	SRA	SRR1023040	USA
Manis javanica	SRA	SRR5337837	China
Meles meles	MELmelRA01	RA01	France
Microgale brevicaudata	MICspMV01	MV03	Madagascar
Mus musculus	SRA	SRR5878900	USA
Myocastor coypus	MYOcoyPH03	Myo2	France
Myrmecophaga tridactyla	MYRtriCAY01	M3023	French Guiana
Ovis aries	SRA	ERR2076303	USA
Proteles cristatus	PROcri01_S29	TS307	South Africa
Proteles cristatus	PROcri01_S2	TS307	South Africa
Rattus norvegicus	SRA	SRR3056926	USA
Sus scrofa	SRA	SRR5802558	China
Tamandua tetradactyla	TAMtetB01	M2813	French Guiana
Tamandua tetradactyla	TAMtetFC01	G2525	French Guiana
Tamandua tetradactyla	TAMtetTT49	M3075	French Guiana
Tenrec ecaudatus	SETsetMV01	MV01	Madagascar
Uroderma bilobatum	SRA	SRR1663490	Uruguay

Table 2 | Information about the additional tissues sequenced for the nine-banded armadillo (*Dasypus novemcinctus*), the Malayan pangolin (*Manis javanica*) and the lesser anteater (*Tamandua tetradactyla*).

Species	Organ	Sample	Individual	Sex	Source	Country	Sequencing instrument	Study
Dasypus novemcinctus	Liver	SRR494766	0986	Male	SRA NCBI	USA	Illumina HiSeq 2000	Unpublished
Dasypus novemcinctus	Spleen	SRR494767	0986	Male	SRA NCBI	USA	Illumina HiSeq 2000	Unpublished
Dasypus novemcinctus	Spleen	SRR494768	0986	Male	SRA NCBI	USA	Illumina HiSeq 2000	Unpublished
Dasypus novemcinctus	Heart	SRR494769	0986	Male	SRA NCBI	USA	Illumina HiSeq 2000	Unpublished
Dasypus novemcinctus	Muscle	SRR494770	0986	Male	SRA NCBI	USA	Illumina HiSeq 2000	Unpublished
Dasypus novemcinctus	Muscle	SRR494771	0986	Male	SRA NCBI	USA	Illumina HiSeq 2000	Unpublished
Dasypus novemcinctus	Colon	SRR494772	0986	Male	SRA NCBI	USA	Illumina HiSeq 2000	Unpublished
Dasypus novemcinctus	Heart	SRR494773	0986	Male	SRA NCBI	USA	Illumina HiSeq 2000	Unpublished
Dasypus novemcinctus	Colon	SRR494774	0986	Male	SRA NCBI	USA	Illumina HiSeq 2000	Unpublished
Dasypus novemcinctus	Kidney	SRR494775	0986	Male	SRA NCBI	USA	Illumina HiSeq 2000	Unpublished
Dasypus novemcinctus	Lung	SRR494776	0986	Male	SRA NCBI	USA	Illumina HiSeq 2000	Unpublished
Dasypus novemcinctus	Cerebellum	SRR494777	0986	Male	SRA NCBI	USA	Illumina HiSeq 2000	Unpublished
Dasypus novemcinctus	Liver	SRR494778	0986	Male	SRA NCBI	USA	Illumina HiSeq 2000	Unpublished
Dasypus novemcinctus	Kidney	SRR494779	0986	Male	SRA NCBI	USA	Illumina HiSeq 2000	Unpublished
Dasypus novemcinctus	Cerebellum	SRR494780	0986	Male	SRA NCBI	USA	Illumina HiSeq 2000	Unpublished
Dasypus novemcinctus	Lung	SRR494781	0986	Male	SRA NCBI	USA	Illumina HiSeq 2000	Unpublished
Dasypus novemcinctus	Heart	SRR6206903	NA	NA	SRA NCBI	USA	Illumina HiSeq 2000	Chen et al., 2019
Dasypus novemcinctus	Kidney	SRR6206908	NA	NA	SRA NCBI	USA	Illumina HiSeq 2000	Chen et al., 2019
Dasypus novemcinctus	Liver	SRR6206913	NA	NA	SRA NCBI	USA	Illumina HiSeq 2000	Chen et al., 2019
Dasypus novemcinctus	Lung	SRR6206918	NA	NA	SRA NCBI	USA	Illumina HiSeq 2000	Chen et al., 2019
Dasypus novemcinctus	Muscle	SRR6206923	NA	NA	SRA NCBI	USA	Illumina HiSeq 2000	Chen et al., 2019
Manis javanica	Cerebellum	SRR2547558	NA	Female	SRA NCBI	Malysia	Illumina HiSeq 2000	Yusoff et al., 2016
Manis javanica	Brain	SRR2561209	NA	Female	SRA NCBI	Malysia	Illumina HiSeq 2000	Yusoff et al., 2016
Manis javanica	Heart	SRR2561211	NA	Female	SRA NCBI	Malysia	Illumina HiSeq 2000	Yusoff et al., 2016
Manis javanica	Kidney	SRR2561212	NA	Female	SRA NCBI	Malysia	Illumina HiSeq 2000	Yusoff et al., 2016
Manis javanica	Liver	SRR2561213	NA	Female	SRA NCBI	Malysia	Illumina HiSeq 2000	Yusoff et al., 2016
Manis javanica	Lung	SRR2561214	NA	Female	SRA NCBI	Malysia	Illumina HiSeq 2000	Yusoff et al., 2016
Manis javanica	Spleen	SRR2561215	NA	Female	SRA NCBI	Malysia	Illumina HiSeq 2000	Yusoff et al., 2016
Manis javanica	Thymus	SRR2561216	NA	Female	SRA NCBI	Malysia	Illumina HiSeq 2000	Yusoff et al., 2016
Manis javanica	Skin	SRR3923846	NA	Female	SRA NCBI	Malysia	Illumina MiSeq	Yusoff et al., 2016
Manis javanica	Liver	SRR5341161	NA	Female	SRA NCBI	China	Illumina HiSeq 2000	Ma et al., 2017
Manis javanica	Small intestine	SRR5328124	NA	Female	SRA NCBI	China	Illumina HiSeq 2000	Ma et al., 2017
Manis javanica	Muscle	SRR5837767	NA	NA	SRA NCBI	China	Illumina HiSeq 2000	Unpublished
Manis javanica	Stomach	SRR7641085	NA	Female	SRA NCBI	China	Illumina HiSeq X Ten	Ma et al., 2019
Manis javanica	Salivary gland	SRR7641084	NA	Female	SRA NCBI	China	Illumina HiSeq X Ten	Ma et al., 2019
Manis javanica	Liver	SRR7641087	NA	Female	SRA NCBI	China	Illumina HiSeq X Ten	Ma et al., 2019
Manis javanica	Tongue	SRR7641083	NA	Female	SRA NCBI	China	Illumina HiSeq X Ten	Ma et al., 2019
Manis javanica	Small intestine	SRR7641090	NA	Female	SRA NCBI	China	Illumina HiSeq X Ten	Ma et al., 2019
Manis javanica	Pancreas	SRR7641082	NA	Female	SRA NCBI	China	Illumina HiSeq X Ten	Ma et al., 2019
Manis javanica	Liver	SRR7641080	NA	Female	SRA NCBI	China	Illumina HiSeq X Ten	Ma et al., 2019
Manis javanica	Large intestine	SRR7641089	NA	Female	SRA NCBI	China	Illumina HiSeq X Ten	Ma et al., 2019
Manis javanica	Stomach	SRR7641086	NA	Female	SRA NCBI	China	Illumina HiSeq X Ten	Ma et al., 2019
Manis javanica	Liver	SRR7641081	NA	Female	SRA NCBI	China	Illumina HiSeq X Ten	Ma et al., 2019
Manis javanica	Pancreas	SRR7641079	NA	Female	SRA NCBI	China	Illumina HiSeq X Ten	Ma et al., 2019
Manis javanica	Lung	SRR7641088	NA	Female	SRA NCBI	China	Illumina HiSeq X Ten	Ma et al., 2019
Tamandua tetradactyla	Tongue	TAMtetTT55	M3075	Male	ISEM	French Guiana	Illumina HiSeq 2500	This study
Tamandua tetradactyla	Liver	TAMtetTT59	M3075	Male	ISEM	French Guiana	Illumina HiSeq 2500	This study
Tamandua tetradactyla	Testis	TAMtetTT70	M3075	Male	ISEM	French Guiana	Illumina HiSeq 2500	This study
Tamandua tetradactyla	Lung	TAMtetTT73	M3075	Male	ISEM	French Guiana	Illumina HiSeq 2500	This study
Tamandua tetradactyla	Heart	TAMtetTT75	M3075	Male	ISEM	French Guiana	Illumina HiSeq 2500	This study
Tamandua tetradactyla	Glandular stomach	TAMtetTT78	M3075	Male	ISEM	French Guiana	Illumina HiSeq 2500	This study
Tamandua tetradactyla	Mucular stomach	TAMtetTT79	M3075	Male	ISEM	French Guiana	Illumina HiSeq 2500	This study
Tamandua tetradactyla	Small intestine	TAMtetTT99	M3075	Male	ISEM	French Guiana	Illumina HiSeq 2500	This study
Tamandua tetradactyla	Spleen	TAMterTT62	M3075	Male	ISEM	French Guiana	Illumina HiSeq 2500	This study
Tamandua tetradactyla	Spleen	TAMtetFC04	G2525	Male	ISEM	French Guiana	Illumina HiSeq 2500	This study
Tamandua tetradactyla	Testis	TAMtetR05	M2813	Male	ISEM	French Guiana	Illumina HiSeq 2500	This study
Tamandua tetradactyla	Tongue	TAMtetB07	M2813	Male	ISEM	French Guiana	Illumina HiSeq 2500	This study

reads to search for the original transcript of a read without looking for a perfect alignment (as opposed to classical quantification by counting the reads aligned on the assembled transcriptome; Wolf, 2013). Pseudo-alignments are performed on Trinity assemblies to quantify the expression of the transcripts (Bray et al., 2016). Counts (raw number of mapping reads) and the Transcripts Per kilobase Million (TPM; gene length and sequencing depth normalization) are reported. Based on the previously inferred orthogroups, orthogroup-level abundance estimates were imported and summarized using tximport (Soneson et al. 2016). To minimize variance between samples, orthogroup-level abundance estimates were standardized using DESeq2 (Love et al., 2014) taking into account the following conditions: diet, taxonomic order, and BUSCO scores. Orthogroups presenting too few counts were discarded for further analyses (sum of counts < 10). Finally, a variance stabilizing transformation (VST, Tibshirani 1988; Huber et al. 2003; Anders and Huber 2010) was performed before visualizing the data using Principal Component Analysis (PCA) to summarise the global expression of genes per species. The number of retained principal components (PC) was determined using a broken stick method. To identify factors potentially explaining differences in the global pattern of gene expression between species, coordinates of species on the selected PC axes were included in a MANOVA analyses with taxonomic order, diet and fragmentation BUSCO score (as categorical variable: low = <10%, medium = 10-15%, high = 15-20%, and very high = >30%) as explanatory variables. Effects of variables were considered significant when the p-value was below the threshold of 0.05. Finally, linear models with orthogroups as random effects were performed to test which conditions are necessary to explain orthogroups expression. Models were compared with ANOVAs. In order to explore the role of diet and phylogeny on gene expression levels, we performed linear mixed models using the lme4 R package (Bates et al. 2018). The full model included taxonomic order, diet category, and BUSCO score (same categorical variable as defined above) as fixed effect, and orthogroup as random effect to account for non-independence between the expression of the same gene across species. The best model was selected using a backward stepwise procedure starting from the full model, and the significance of each fixed effect was tested using a F-test (R function anova()). This procedure was used to eliminate non-significant variables: at each step, the effect presenting the highest p-value was removed and the model ran again without it.

Chitinase gene family tree analysis

Reconstruction of chitinase gene family evolution - The chitinase family is composed of nine paralogs (CHIA1-5, CHIT1, CHI3L1, CHI3L2, OVGP1). Mammalian sequences similar to the protein sequence of the human chitinase gene (NP_970615.2) were searched in the NCBI non-redundant protein database using BLASTP (E-value < 10). The protein sequences identified by BLASTP were then imported into Geneious Prime (Kearse et al., 2012) and aligned using MAFFT v7.450 (Katoh and Standley, 2013) used with default parameters. Preliminary gene trees were then reconstructed with maximum likelihood using RAxML v8.2.11 (Stamatakis, 2014) under the LG+G4 model (Le and Gascuel, 2008) as implemented in Geneious Prime. From the reconstructed trees, the sequences were filtered according to the following criteria: (1) fast-evolving sequences with an E-value greater than zero and not belonging to the chitinase family were excluded; (2) in cases of multiples isoforms, only the longest was retained; (3) sequences whose length represented less than at least 50% of the total alignment length were removed; (4) in case of identical sequences from the same species the longest was kept; and (5) sequences labelled as "Hypothetical protein" and "Predicted: low quality protein" were discarded. This procedure resulted in a dataset containing 528 mammalian sequences that were realigned using MAFFT. This alignment was then cleaned up by removing sites not present in at least 50% of the sequences resulting in a total length of 460 amino acid sites. A maximum likelihood tree was then reconstructed with RAxML-NG v0.9.0 (Kozlov et al., 2019) using 10 tree searches starting from maximum parsimony trees under the LG+G8+F model. To determine the optimal rooting scheme, a rapid reconciliation between the resulting gene tree and the species tree of the 143 mammal species represented in our dataset was performed using the Treerecs reconciliation algorithm based on maximum parsimony (Comte et al., 2019) as implemented in SeaView v5.0.2 (Gouy et al., 2010). The final chitinase gene family tree was produced using the maximum likelihood gene family tree reconciliation approach implemented in GeneRax v.1.1.0 (Morel et al., 2019) using the TreeRecs reconciled tree as input. GeneRax can reconstruct duplications, losses and horizontal gene transfer events but since the latter are negligible in mammals, only gene duplications and losses have been modeled here (--rec-model UndatedDL) and the LG+G model was used.

Ancestral sequence reconstructions - Ancestral sequences of the different paralogues were reconstructed from the reconciled tree using the RAxML-NG program (Kozlov et al., 2019) (--ancestral function, --model LG+G8+F). The sequences were then aligned in Geneious Prime with MAFFT. Given that active chitinases are characterized by a catalytic site with a conserved amino acid motif (DXXDXDXE; Olland et al., 2009; Hamid et al., 2013), this motif was compared among all available species. Additionally, the six conserved cysteine residues responsible for chitin binding (Tjoelker et al., 2000; Olland et al., 2009) were also investigated.

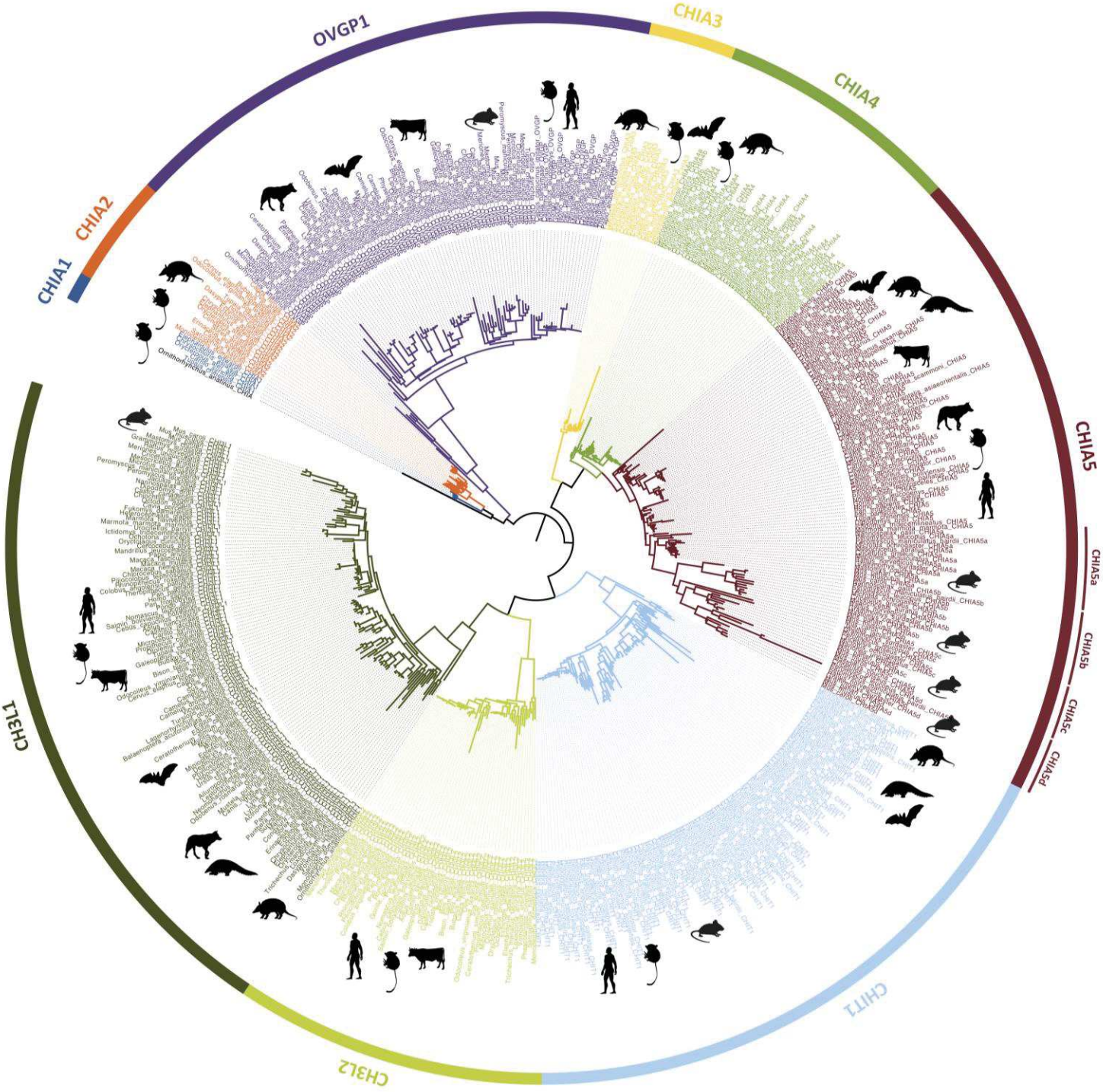


Figure 1 | Chitinase gene family tree in mammals reconstructed using a maximum likelihood gene tree species tree reconciliation approach. The nine chitinase paralogs are indicated in the outer circle.

Chitinase expression in salivary glands - The chitinase orthogroup inferred by OrthoFinder2 in previous analyses (see above) was extracted using BLASTX with the reference chitinase database previously created. First, orthogroup-level abundance estimates were compared between each separate diet, and then by grouping insectivores and non insectivores. Second, the chitinase orthogroup was divided into sub-orthogroups for each chitinase paralog (CHIA1-5, CHIT1, CHI3L1, CHI3L2, OVGP1). To take advantage of global expression information for the standardization steps, these new orthogroups were included in the previous orthogroup-level abundance matrix estimates and the same normalization approach (DESeq2) was conducted. Finally, gene-level abundance estimates for all chitinase paralogs were extracted and compared after a log2 transformation.

Chitinase expression in additional organs - The contigs containing the different chitinase paralogues genes (CHIA1-5, CHIT1, CHI3L1, CHI3L2, OVGP1) were identified by mapping the sequences of these genes for each of the three species (*T. tetradactyla*, *D. novemcinctus*, and *M. javanica*) on the transcriptome assemblies using Geneious Prime (with "medium" mapping sensitivity). Expressions were then estimated as for salivary glands (see above), and compared for different tissues between the three focal myrmecophagous species using normalized expression data (to minimize variance between samples) using the DESeq2 package v1.22.2 (Love et al., 2014) of the Bioconductor suite v3.8 (Gentleman et al., 2004).

Preliminary results

Chitinase gene family evolution

The reconciled tree of mammalian chitinase genes is presented in **Figure 1**. The evolution of this gene family is characterized by the presence of numerous gene losses with an estimated gene loss rate of 0.33765 and 384 speciation followed by gene loss. The gene duplication rate was 0.0582458 for a total number of estimated gene duplications of 48. At the base of the reconciled gene tree, we found the clade OVGP1/CHIA1-2 (optimal root inferred by the reconciliation performed with TreeRecs) then a duplication separating the CHIT1/CHI3L1-2 and CHIA3-5 paralogues. Within the CHIT1/CHI3L clade two duplications gave rise to CHIT1, then CHI3L1 and CHI3L2 (**Fig. 1**). In the CHIA3-5 clade, a first duplication allowed the separated CHIA3 from CHIA4 and CHIA5 which duplicated subsequently. Marsupial CHIA4 sequences were located at the base of the CHIA4-5 clade suggesting that this duplication is specific to placentals. The CHIA5 sequences of chiropterans were found at the base of the CHIA5 clade. The duplication that gave rise to the CHIA4 and CHIA5 genes is recent and specific to eutherians (marsupials and placentals) since no other taxon was found within these clades. Within the CHIA5 specific to Muroidea (Spalacidae, Cricetidae and Muridae), we found the four clades identified in the unreconciled gene tree: from the CHIA5a paralog, two duplications gave rise to the three CHIA5b-d paralogues represented by long branches characterizing rapidly

evolving sequences. The duplication giving rise to the CHIA5c and CHIA5d paralogs concerns only the Cricetidae and Muridae; Nannospalax galili (Spalacidae) being present only in the clade of the CHIA5b paralogous gene.

Comparison of ancestral sequences

The ancestral sequences of the different paralogues have been reconstructed from the reconciled mammalian gene tree (**Fig. 1**) and compared (**Fig. 2**) to gain insight into the potential function of the enzymes they encode. The alignment of predicted amino acid sequences allowed to find the chitinolytic domain between positions 133 and 140 at the preserved pattern DXXDXDXE and showed that the ancestral sequences of CHI3L1, CHI3L2 and CHIA5b-d have a mutated chitinolytic domain with absence of a glutamic acid in position 140 (active proton-donor site necessary for chitin hydrolysis; Olland et al, 2009; Hamid et al., 2013) and therefore cannot degrade chitin (**Fig. 2A**). The evolution of the different CHIAs therefore seems to be related to changes in their active site. The six cysteine residues allowing the binding to chitin (Tjoelker et al., 2000) are found at positions 371, 418, 445, 455, 457 and 458 (**Fig. 2B**). The mutation of one of these cysteines prevents binding to chitin (Tjoelker et al., 2000) as this is the case in the ancestral OVGP1 protein where the last four cysteine residues are changed (**Fig. 2B**). The other ancestral sequences present the six conserved cysteine residues and thus encode for enzymes that can bind to chitin (**Fig. 2B**).

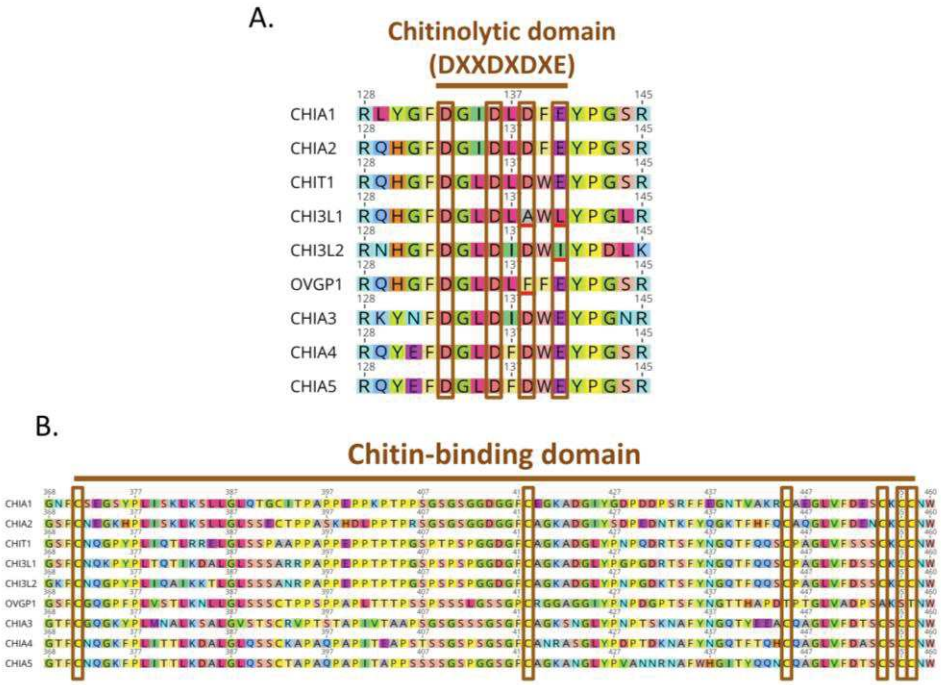


Figure 2 | Comparison of predicted ancestral sequences of mammalian chitinase paralogs. A. Conserved residues of the chitinolytic domain active site. B. Conserved cysteine sites of the chitin-binding domain.

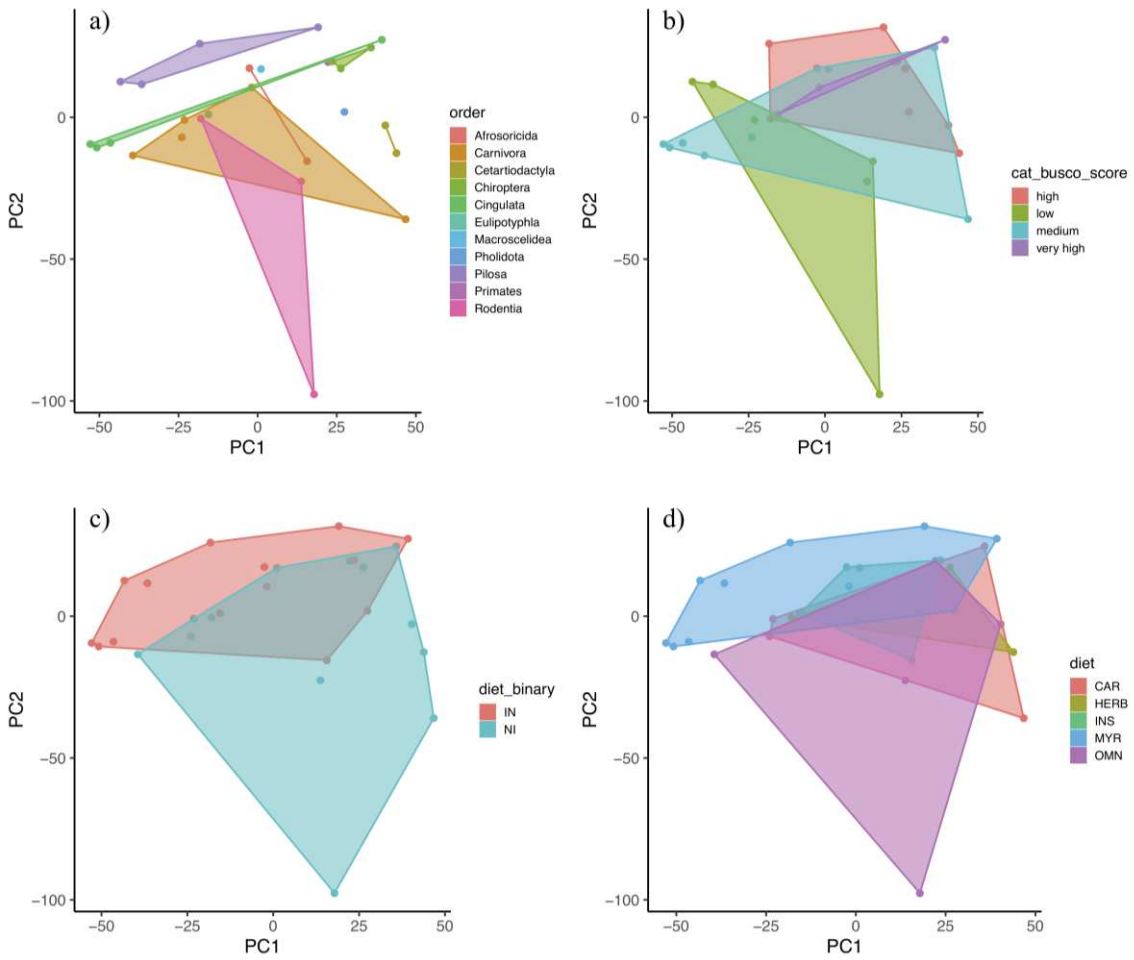
Gene expression in mammalian salivary glands

A comparative analysis of gene expression across all 27 samples, was performed on the normalized DESseq2 counts matrix. To visualize whether global expression for each species was clustered either by taxonomic groups (orders), by diet (insectivores versus non-insectivores or precise diet: carnivores, herbivores, omnivores, frugivores, myrmecophagous, and insectivores) or by transcriptome quality category (BUSCO score: percentage of fragmented genes: low = < 10%, medium = 10-15%, high = 15-20%, very high = > 25%), we performed a PCA (**Fig. 3**). This analysis was performed on the two orthogroup datasets (with and without NA in the matrix) but given the high similarity between the two resulting PCAs, only the more conservative dataset was used for further analyses (normality obtained thanks to the elimination of NA in orthogroups). The first two principal components (PC1 and PC2 explain 12.27% and 8.04% of the variance, respectively), weakly discriminate species by the taxonomic order to which they belong (Afrosoricida, Carnivora, Cetartiodactyla, Chiroptera, Cingulata, Eulipotyphla, Macroscelidea, Pholidota, Pilosa, Primates, Rodentia), and according to their diet when considering insectivores versus non-insectivores but not when considering all precise diets (**Fig. 3**). To evaluate statistical support for the observed individual clusters, we performed MANOVAs correlating diet, taxonomy (Order), and the transcriptome quality (BUSCO categories) to the first three principal components explaining 26.12% of the total variance. The overall variance was primarily explained by taxonomic order and thus phylogenetic relationships (**Fig. 3**). Diet and transcriptome quality, do not seem to explain much of the variance across the first three principal components.

To further study the role of each effect (explanatory variables: taxonomic order, diet, transcriptome quality), we performed linear model analyses. While the MANOVA performed with Principal Component (summarizing expression patterns), tests the effect of each effect at the species level, the linear mixed model was used to explain the observed variations between the abundance of orthogroups as a function of conditions. An ANOVA was performed to test the significance of each effect. All conditions significatively explained the variance observed between orthogroup abundances: order (p-value: 0.002), diet (p-value 5.7 e-05), BUSCO score categories (p-value: 0.004).

Chitinase gene expression in mammalian salivary glands

To test the hypothesis that salivary glands play an important functional role for the digestion of ants and termites in ant-eating mammals, we analysed gene expression profiles in the nine chitinase paralogs revealed by the gene family tree reconstruction (**Fig. 4**). CHIA1 was only expressed in elephant shrew (*Elephantulus myurus*, 15.70 normalized counts). CHIA2 was expressed in the wild boar (*Sus scrofa*, 47.62 normalized counts). CHIA3 was expressed in the California leaf-nosed bat (*Macrotus californicus*, 27.80 normalized counts) and in all lesser anteater individuals (*T. tetradactyla*; 39.24, 32.60 and 12.46 normalized counts). CHIA4 was also highly expressed in all lesser anteater individuals (717.81, 280.01, 239.90 normalized counts), in the giant anteater



e)

	Df	Pillai	Approx F	Num df	Den df	Pr(>F)	
Order	10	2.14046	2.24122	30	27	0.01858	*
Diet	4	0.78263	0.79414	12	27	0.65278	
Busco score	3	1.06445	1.64983	9	27	0.15104	

f)

	Sum Sq	Mean Sq	Num DF	Den DF	F value	Pr(>F)	
Order	54.118	5.4118	10	109375	2.7419	0.002237	**
Diet	48.838	12.2096	4	109375	6.1861	5.67e-05	***
Busco score	26.616	18.8721	3	109375	4.4951	0.003698	**

Figure 3 | Principal component analysis (PCA) performed on the log2 normalized counts matrix obtained from 4,207 single-copy orthologs. The 27 salivary gland samples are coloured according to (a) taxonomic orders, (b) BUSCO scores (fragmented genes: low = < 10%, medium = 10-15%, high = 15-20%, very high = > 25%), c) insectivores versus non-insectivores diet, and d) diet categories. The first two principal components (PC1 = 12.27%, PC2 = 8.04%) explain a total variance of 20.31%. e) The MANOVA analysis including PC1-3 axes and taxonomic order, diet and fragmentation BUSCO score as explanatory effects suggests an effect of the phylogeny (order) in global expression patterns of species. f) Results of the ANOVA performed to measure the significance of each effect in the best linear model (LM = counts ~ order + diet + BUSCO).

(*M. tridactyla*; 72.33 normalized counts), in the domestic mouse (*Mus musculus*; 51.45 normalized counts), and in the California leaf-nosed bat (*Macrotus californicus*; 22,093.65 normalized counts). The level of gene expression in CHIA5 was much higher in the Malayan pangolin (*Manis javanica*; 8,411.20 normalized counts) than in the two other species in which we detected expression of this gene: the common genet (*Genetta genetta*; 247.70 normalized counts), and the wild boar (*Sus scrofa*; 282.10 normalized counts). OVGP1 was expressed in the domestic dog (*Canis lupus familiaris*; 7.52 normalized counts), human (*Homo sapiens*; 16.82 normalized counts), and the wild boar (19.76 normalized counts). CHI3L2 was expressed in human (1832.54 normalized counts), the wild boar (341.85 normalized counts), the common tenrec (*Tenrec ecaudatus*; 102.79 normalized counts), and the elephant shrew (*E. myurus*; 129.11). CHI3L1 was expressed in most species (18 out of 27) with normalized counts values ranging from 90.92 for the giant anteater to 1,575.33 in one nine-banded armadillo (*D. novemcinctus*) individual. CHIT1 was expressed in many species (11 out of 27) with normalized counts values ranging from 117.25 in one nine-banded armadillo (*D. novemcinctus*) individual to 165689.6 in the California leaf-nosed bat (*M. californicus*). Finally, the aardwolf (*P. cristatus*), the Norway rat (*Rattus norvegicus*), the common vampire bat (*Desmodus rotundus*), and the tent-making bat (*Uroderma bilobatum*) did not appear to express any chitinase paralog genes.

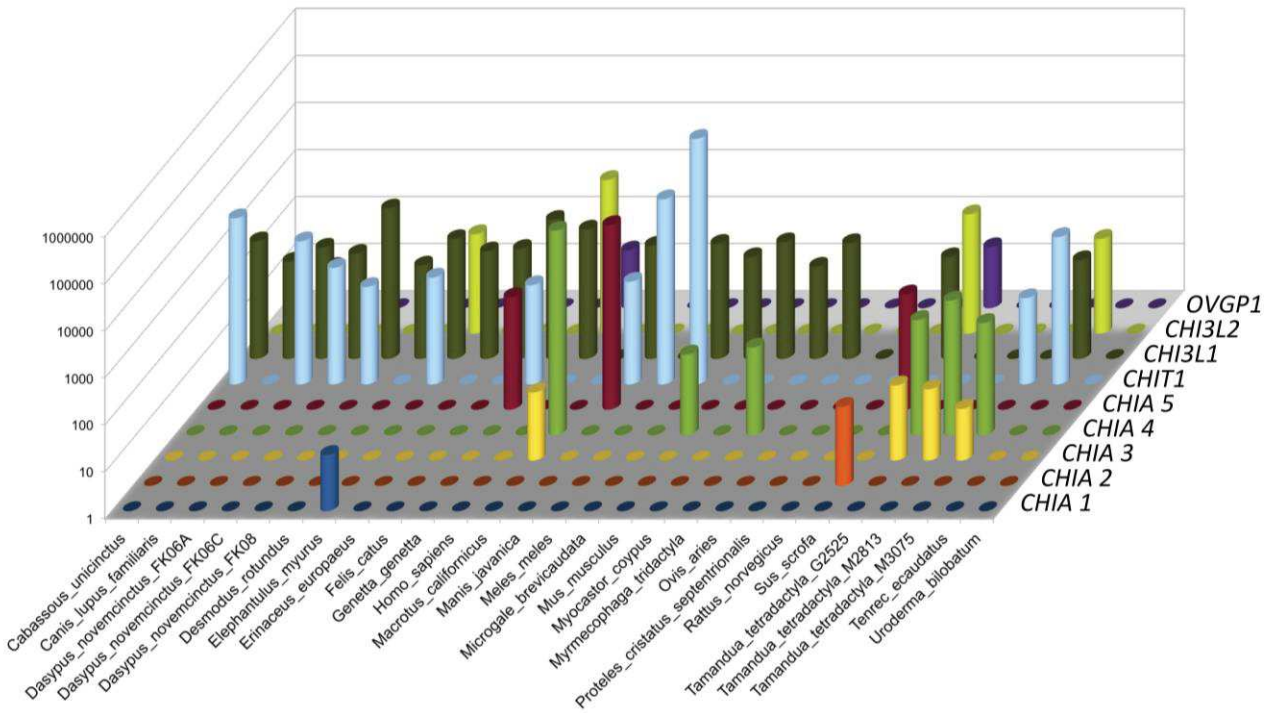


Figure 4 | Chitinase paralogous gene expression in 27 mammalian salivary gland transcriptomes.

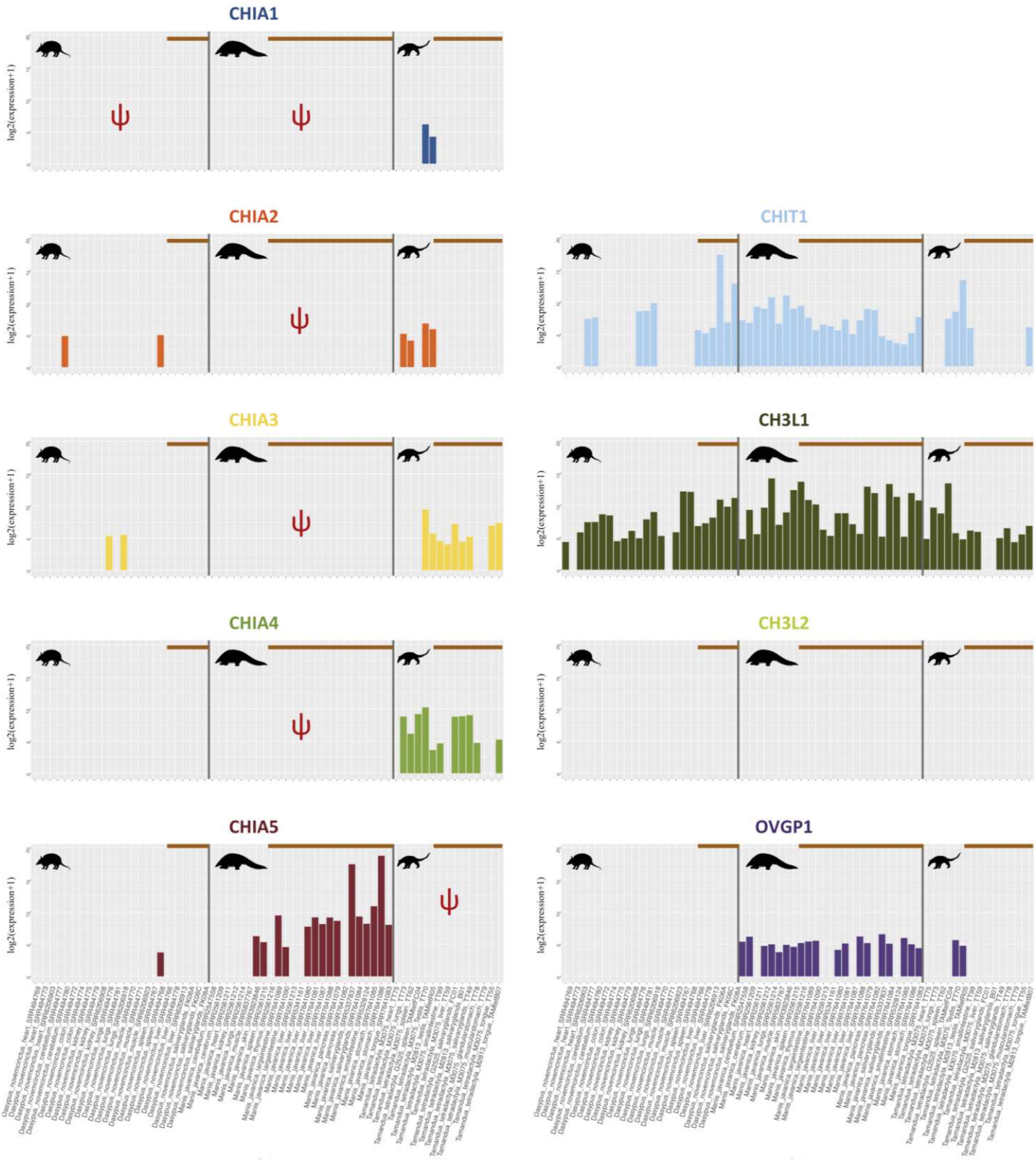


Figure 5 | Comparative expression of CHIA1-5 in 64 transcriptomes from different organs in three mammalian species: the nine-banded armadillo (*Dasypus novemcinctus*), the Malayan pangolin (*Manis javanica*), and the lesser anteater (*Tamandua tetradactyla*). Pseudogenized genes are symbolized by a Ψ and the horizontal bars indicate the digestive organs.

Chitinase gene expression in other digestive and non-digestive organs

The expression of the nine chitinase paralogs in several organs was compared between three species including two highly myrmecophagous species (*T. tetradactyla* and *M. javanica*) and an insectivorous xenarthran (*D. novemcinctus*), with the exception of the CHI3L2 gene that appears to be non-functional or even absent in these three species (**Fig. 5**). This analysis revealed differences among organs and in the expression levels of these genes in the three species. CHI3L1 was found to be expressed in the majority of tissues in all three species. CHIT1 was expressed in all tissues in *M. javanica* and only in the spleen, testes, tongue and small intestine in *T. tetradactyla*, and in the cerebellum, liver, lungs and salivary glands in *D. novemcinctus*. OVGP1 was not expressed in any of the tissues studied here in *D. novemcinctus* but in *T. tetradactyla* it is found expressed only weakly in the testes whereas in *M. javanica* it is expressed in all tissues except the heart and small intestine.

CHIA1-4 are non-functional in *M. javanica* in which only the CHIA5 gene is functional and was found to be highly expressed in the salivary glands and stomach (1,112,393.1 and 232,111.3 normalized counts on average respectively) but also in the large intestine, liver and pancreas (382.2, 285.1 and 208.2 normalized counts on average respectively) (**Fig. 5**). In the nine-banded armadillo (*D. novemcinctus*), CHIA1 is pseudogenized. CHIA2 was found expressed only in the cerebellum and spleen, CHIA3 was expressed only in the lungs, CHIA4 expression was not observed in any of the tissues studied here, and CHIA5 was only weakly expressed in the spleen. In the lesser anteater (*T. tetradactyla*), it is the CHIA5 gene that is pseudogenized. CHIA1 was found weakly expressed in the testes, and CHIA2 also weakly expressed in the testes, lungs and spleen (**Fig. 5**). CHIA3 was expressed most strongly in salivary glands and tongue (128.7 and 195.5 mean normalized counts respectively) and less strongly in liver and small intestine (29.2 and 24.3 normalized counts respectively) (**Fig. 5**). CHIA4 was expressed strongly in salivary glands but more weakly in the tongue (mean 728.5 and 20.1 normalized counts respectively), it is also expressed although weakly in the small intestine (26.4 normalized counts), and it is strongly expressed in the lungs (684.3 normalized counts) and spleen (mean 362.2 normalized counts respectively) (**Fig. 5**).

Preliminary discussion

Evolution of chitinase paralogs toward different functions

The study of the ancestral sequences of the different chitinase paralogs revealed differences in their ability to bind and degrade chitin, suggesting that these paralogues have evolved towards different functional specializations. Thus, the evolution of chitinase-like-proteins was accompanied by a loss of chitin hydrolysis enzymatic activity that occurred several times independently (Bussink et al., 2007; Funkhouser and Aronson, 2007 and Hussain and Wilson, 2013). In mammals, OVGP1 has a role in fertilization and embryonic development (Buhi, 2002; Saint-Dizier et al., 2014; Algarra et al., 2016; Laheri et al, 2018) and CHI3L1 and CHI3L2, found expressed in various cell types including

macrophages and synovial cells, play roles in cell proliferation and immune response (Recklies et al., 2002; Lee et al., 2011; Areshkov et al., 2012). CHIA genes specific to Muroidea (rodents) and characterized by rapidly evolving sequences have also been described as chitinase-like rodents-specific (CHILrs) enzymes (Bussink et al. 2007; Hussain & Wilson 2013). These enzymes also appear to have evolved for functions in the immune response (Lee et al., 2011; Hussain and Wilson, 2013). CHIA5b cannot bind to chitin unlike CHIA5c and CHIA5d, suggesting different roles between these three paralogues. The role of chitinase-3-like proteins in non-mammalian species (named here CHI3L0a and CHI3L0b) remains to be determined. These paralogues probably assume different functions in the species that carry them, the study of their ancestral sequences (results not presented here) has shown that CHI3L0a has a catalytic site active in contrast to CHI3L0b. Contrary to chitinase-like proteins, CHIT1 and CHIAs are able to degrade chitin. In humans, CHIT1 is expressed in macrophages and neutrophils and is suspected to be involved in the defense against chitin-containing pathogens such as fungi (Gordon-Thomson et al., 2009; Lee et al., 2011). In addition to their role in chitin digestion (Boot et al., 2001), CHIAs are also suspected to play a role in the inflammatory response (Lee et al., 2011) and are found expressed in non-digestive tissues, in agreement with our comparative transcriptomics results. It has been proposed that the expansion of the chitinase gene family is linked to the emergence of the innate and adaptive immune systems in vertebrates (Funkhouser and Aronson, 2007).

The evolution of the different CHIA1-5 genes seems to be related to changes in their catalytic sites that could have consequences on the secondary structure of enzymes affecting their optimal pH or function and may explain the differences observed between CHIAs in terms of adaptation to a rather acidic or alkaline environment. Testing the chitin degradation activity on different substrates and at different pH of enzymes produced from the ancestral sequences reconstructed for each of the paralogues would allow a better understanding of their role and, complemented by transcriptomic data, to determine their expression sites. Finally, studying the potential binding of these enzymes with other substrates would shed more light on their roles; the change of a cysteine in the chitin binding domain prevents binding to this substrate but not to tri-N-acetyl-chitotriose (Tjoelker et al., 2000), a compound derived from chitin with antioxidant properties (Chen et al., 2003; Salgaonkar et al., 2015).

Digestive enzymes and chitinase gene expression

Chitinase genes have previously been suggested to play an important role in insect digestion (Jeuniaux, 1971; Bussink et al., 2007). Maximum likelihood phylogenetic analyses, recovered nine paralogous chitinase gene sequences that we found expressed in placental mammal salivary glands. In addition to the five CHIA paralogs identified by Emerling et al. (2018), we were able to detect an additional gene OVGP1 that is closely related to the previously characterized CHIA genes. Different aliases for OVGP1 include Mucin 9 and CHIT5 (www.genecards.org), which suggests a possible digestive function. This result was further confirmed by synteny analysis suggesting a common origin

for all five CHIA genes and OVGP1 (data not shown). The study of Emerling et al. (2018) also showed that the lesser anteater (*T. tetradactyla*) has four functional CHIA genes, whereas the Malayan pangolin (*M. javanica*) only one (CHIA5). This raised the question whether the Malayan pangolin potentially compensates for the paucity of functional chitinase genes by overexpressing CHIA5. We were able to confirm this hypothesis because expression profiles for the CHIA5 gene in *M. javanica* were significantly higher than in the two other species in which we detected gene expression. Interestingly, the importance of CHIA5 for the Malayan pangolin is further supported by its high expression profiles in all major digestive tissue types (tongue, stomach, pancreas, large intestine, and liver) (Ma et al., 2019 and **Fig. 5**). For the lesser anteater (*T. tetradactyla*) on the other hand, no expression of CHIA5 was detected in any organ in agreement with its likely loss of function through pseudogenization (Emerling et al., 2018). Nevertheless, the lesser anteater showed high levels of CHIA3 and CHIA4 gene expression.

Specifically, in ant-eating mammals, the Malayan pangolin (*M. javanica*), the lesser anteater (*T. tetradactyla*), the giant anteater (*M. tridactyla*) and the southern naked-tailed armadillo (*C. unicinctus*) all express one or more chitinase genes. Interestingly, the aardwolf (*P. cristatus*) does not seem to express any chitinase gene. A possible explanation could be that chitinase genes might not be functional (pseudogenized) due to phylogenetic constraints associated with a carnivorous ancestor. For instance, Emerling et al. (2018) have shown that CHIA1, CHIA2, CHIA3, and CHIA4 are pseudogenes in numerous members of Carnivora. Bearing this idea in mind, the possible presence of frameshift mutations and stop codons were inspected in all nine chitinase genes in the aardwolf genome assembly (Allio et al. 2020). CHIA1, CHIA2, CHIA3, CHIA4 were indeed found to be non functional, and CHI3L2 seems to be absent from the genome of the aardwolf as in most other carnivorans. However, no pseudogenization events could be detected for CHIA5, CHI3L1, CHIT1 and OVGP1 that seem to be fully functional. The observation that the aardwolf does not express any of the chitinase paralogous genes might be the consequence of technical difficulties in transcriptome characterization (tissue quality and fragmented assembly) or might have a biological explanation. Possible biological explanations could be that the aardwolf expresses these genes in other digestive tissues including the enlarged submandibular glands, since our transcriptome was obtained from a sublingual salivary gland. Alternatively, the gut microbiome might be implicated in the digestion of the chitinous exoskeleton of termites that the aardwolf specifically preys on. Consequently, the lack of expression of the chitinase paralogous genes due to either pseudogenization or downregulation might not affect the aardwolf's ability to digest termites. The adaptation of the aardwolf to myrmecophagy is relatively recent (<4 Myr) and there are no clear signs of dietary adaptation in its genome (Westbury et al. 2020) further suggesting that the gut microbiome might play a key role for termite digestion in this species.

Interestingly, in the California leaf-nosed bat (*M. californicus*), which is a 12g insectivore, chitinase gene expression is particularly elevated in CHIA3 and CHIA4. This result, together with the previous observations made on myrmecophagous mammals, strongly supports the hypothesis that salivary glands play a primordial adaptive role in placental mammal evolution towards insectivory. Indeed, in the blood-feeding common vampire bat (*D. rotundus*) and in the frugivorous tent-making bat (*U. bilobatum*), none of the chitinase genes was expressed. As in this study, analyses of additional tissues in bats may help to better understand this pattern. The most likely explanation is that these genes have been pseudogenized in both species, which would be concordant with the findings of Emerling et al. (2018) who showed that CHIA1-5 could be pseudogenized across multiple non-insectivorous bat species. Accordingly, a recent study of 10 bat genomes found evidence for widespread CHIA gene losses and pseudogenization in frugivorous species and reported complete loss of CHIA1-5 function in *D. rotundus* (Wang et al. 2020).

Overall, digestive enzyme gene expression results suggest a primary role for salivary glands in placental mammal food digestion. Here, we showed that overall chitinase gene expression is particularly elevated in insectivorous and particularly in ant- and/or termite-eating mammals. This result is concordant with studies in other organisms. For instance, Chen and Zhao (2019) have recently shown the adaptive nature of CHIA and CHIT1 genes across different bird taxa. Gene expression across the 4,207 single-copy orthologs suggests that the primary drivers for similarity in gene expression are historical contingency and environmental pressures associated with the myrmecophagous life history trait. Whereas the majority of digestive enzyme genes are under stabilizing selection (Perry et al., 2012; Chen et al., 2019), suggesting an important role for historical contingencies, arguments in favour of the role of environmental pressures can be obtained from most chitinase genes. The extremely high expression profile of CHIA5 in the Malayan pangolin and the fact that different molecular and morphological pathways have been employed in order to adapt to a similar environmental constraint, invokes the occurrence of positive selection leading towards a physiological adaptation to nutritionally poor and chitinous rich diet.

Different molecular adaptations to the myrmecophagous diet in placentals

In the specific case of adaptation to myrmecophagy, comparative genomic and transcriptomic analyses of these chitinase genes, in particular chitin-degrading CHIAs, have led to a better understanding of how convergent adaptation to myrmecophagy in placentals occurs at the molecular level. These analyses have highlighted different pseudogenization events between the myrmecophagous species studied as well as differences in the expression profiles of these genes between species.

In myrmecophagous carnivores (*P. cristatus* and *O. megalotis*; Carnivora) and pangolins (Pholidota), CHIA5 is functional while CHIA1-4 are pseudogenized. Similar inactivating mutations are observed in the CHIA1 gene in carnivores and pangolins and are dated to at least 67 Mya, well

before the origin of carnivores (46.2 Ma) and pangolins (26.5 Ma) (Emerling et al., 2018). Thus, in spite of a 100% invertebrate diet, pangolins have only one functional CHIA gene, probably due to their common inheritance with carnivores (historical contingency, Emerling et al., 2018). CHIA5 is found highly expressed in several organs of the digestive tract and is over-expressed in the salivary glands and stomach most probably to compensate for the presence of a single CHIA gene allowing chitin degradation in this species. Ma et al. (2017) further identified several metabolic pathways (of sugars, amino acids and lipids) potentially involved in adaptation to the myrmecophagous diet in *M. javanica*. OVGP1 is found expressed in several organs of the digestive tract of *M. javanica* whereas it is not expressed in *T. tetradactyla* and *D. novemcinctus*. The catalytic site of this enzyme is inactive (absence of glutamic acid) in this species suggesting that this enzyme does not hydrolyze chitin, however its role in *M. javanica* remains to be determined. Finally, no chitinase paralogs were expressed in the salivary glands of the aardwolf. Additionally, the global gene expression profile in salivary glands seems to be more similar within Carnivora species (historical contingency) than within myrmecophagous placentals (diet convergence) suggesting significant phylogenetic constraints. These results should be confirmed by studying the expression profiles of these genes in other digestive organs of aardwolf but also of other Carnivora species, including the bat-eared fox (*Otocyon megalotis*), to compare the expression of these genes in these two carnivore species specialized in invertebrate consumption with other carnivores to discriminate the impact of the historical contingency and the diet in their gene expression.

Anteaters (Pilosa; Vermilingua) inherited the CHIA genes from an insectivorous ancestor (Emerling et al., 2018) and then the CHIA5 gene was lost. In *T. tetradactyla*, the inactivating mutations of CHIA5 have been identified and the estimated inactivation time of this gene was 6.8 Ma, subsequent to the origin of Vermilingua (34.2 Ma) and after the divergence with *M. tridactyla* (11.3 Ma) suggesting a loss specific to the genus *Tamandua* (Emerling et al., 2018). Our study did not find this gene expressed in *M. tridactyla*. Although CHIA1 and CHIA2 are functional in *T. tetradactyla*, they were not found to be expressed in the organs of the digestive tract studied here, so these genes do not seem to be involved in digestion. On the other hand, CHIA3 and CHIA4 were found expressed in several digestive organs including salivary glands, tongue, liver and intestine. Thus, in the case of the lesser anteater (*T. tetradactyla*) and the Malayan pangolin (*M. javanica*), two myrmecophagous species that diverged about 100 Ma ago (Meredith et al., 2011; Irisarri et al., 2017), convergent adaptation to myrmecophagy has been achieved by using paralogs of different chitinase genes to digest chitin, probably due to phylogenetic constraints leading to the loss of CHIA1-4 in the ancestor of the Ferae (Carnivora and Pholidota). These two taxa present extreme morphological adaptations (including total loss of dentition) that also did not involve the same mechanisms (Ferreira-Cardoso et al., 2019). These results thus remind us that the evolution of convergent phenotypes does not systematically imply similar mechanisms.

Acknowledgments

We would like to thank Hugues Parrinello (Montpellier GenomiX platform) for advice on RNAseq. We are also indebted to Frank Knight, Mark Scherz, Miguel Vences, Andolalao Rakotoarison, Nico Avenant, Pierre-Henri Fabre, Quentin Martinez, Nathalie Delsuc, Aude Caizergues, Roxanne Schaub, Benoit de Thoisy, Lionel Hautier, Fabien Condamine, Sérgio Ferreira-Cardoso, and François Catzeflis for their help with tissue sampling. We also thank Marie Sémon for providing useful advice on RNAseq statistical analyses. Computational analyses benefited from the Montpellier Bioinformatics Biodiversity (MBB) platform. This work has been supported by grants from the European Research Council (ConvergeAnt project: ERC-2015-CoG-683257) and Investissements d'Avenir of the Agence Nationale de la Recherche (CEBA: ANR-10-LABX-25-01; CEMEB: ANR-10-LABX-0004).

References

⇧Back to summary⇧

Algarra B, Han L, Soriano-Úbeda C, Avilés M, Coy P, Jovine L, Jiménez-Movilla M. 2016. The C-terminal region of OVGP1 remodels the zona pellucida and modifies fertility parameters. *Sci Rep* **6**:32556. doi:10.1038/srep32556

Allio R, Tilak M-K, Scornavacca C, Avenant NL, Corre E, Nabholz B, Delsuc F. 2020. High-quality carnivore genomes from roadkill samples enable species delimitation in aardwolf and bat-eared fox. *bioRxiv* 2020.09.15.297622. doi:10.1101/2020.09.15.297622

Anders S, Huber W. 2010. Differential expression analysis for sequence count data. *Genome Biol* **11**:R106. doi:10.1186/gb-2010-11-10-r106

Arendt J, Reznick D. 2008. Convergence and parallelism reconsidered: what have we learned about the genetics of adaptation? *Trends Ecol Evol* **23**:26–32. doi:10.1016/J.TREE.2007.09.011

Areshkov PO, Avdieiev SS, Balynska O V., LeRoith D, Kavsan VM. 2012. Two closely related human members of chitinase-like family, CHI3L1 and CHI3L2, activate ERK1/2 in 293 and U373 cells but have the different Influence on cell proliferation. *Int J Biol Sci* **8**:39–48. doi:10.7150/ijbs.8.39

Bates D, Maechler M, Bolker B, Walker S. 2018. Package "lme4."

Blount ZD, Lenski RE, Losos JB. 2018. Contingency and determinism in evolution: Replaying life's tape. *Science (80-)* **362**. doi:10.1126/SCIENCE.AAM5979

Boot RG, Blommaart EFC, Swart E, Vlugt KG der, Bijl N, Moe C, Place A, Aerts JMFG. 2001. Identification of a novel acidic mammalian chitinase distinct from chitotriosidase. *J Biol Chem* **276**:6770–6778. doi:10.1074/JBC.M009886200

Bray NL, Pimentel H, Melsted P, Pachter L. 2016. Near-optimal probabilistic RNA-seq quantification. *Nat Biotechnol* **34**:525–527. doi:10.1038/nbt.3519

Buhi WC. 2002. Characterization and biological roles of oviduct-specific, oestrogen-dependent glycoprotein. *Reproduction*. doi:10.1530/rep.0.1230355

Bussink AP, Speijer D, Aerts JMFG, Boot RG. 2007. Evolution of mammalian chitinase(-like) members of family 18 glycosyl hydrolases. *Genetics* **177**:959–970. doi:10.1534/GENETICS.107.075846

Chen A, Taguchi T, Sakai K, Kikuchi K, Wang MW, Miwa I. 2003. Antioxidant activities of chitobiose and chitotriose. *Biol Pharm Bull* **26**:1326–1330.

Chen J, Swofford R, Johnson J, Cummings BB, Rogel N, Lindblad-Toh K, Haerty W, Palma F di, Regev A. 2019. A quantitative framework for characterizing the evolutionary history of mammalian gene expression. *Genome Res* **29**:53–63. doi:10.1101/GR.237636.118

Chen S, Zhou Y, Chen Y, Gu J. 2018. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**:i884–i890. doi:10.1093/bioinformatics/bty560

Chen Y-H, Zhao H. 2019. Evolution of digestive enzymes and dietary diversification in birds. *PeerJ* **7**:e6840. doi:10.7717/peerj.6840

Christin P-A, Weinreich DM, Besnard G. 2010. Causes and evolutionary significance of genetic convergence. *Trends Genet* **26**:400–405. doi:10.1016/J.TIG.2010.06.005

Comte N, Morel B, Hasić D, Guéguen L, Boussau B, Daubin V, Penel S, Scornavacca C, Gouy M, Stamatakis A, Tannier E, Parsons DP. 2020. Treerecs: an integrated phylogenetic tool, from sequences to reconciliations. *Bioinformatics* **36**:4822–4824. doi:10.1093/bioinformatics/btaa615

Delsuc F, Scally M, Madsen O, Stanhope MJ, de Jong WW, Catzeflis FM, Springer MS, Douzery EJP. 2002. Molecular phylogeny of living xenarthrans and the impact of character and taxon sampling on the placental tree rooting. *Mol Biol Evol* **19**:1656–1671. doi:10.1093/oxfordjournals.molbev.a003989

Emerling CA, Delsuc F, Nachman MW. 2018. Chitinase genes (*CHIA* s) provide genomic footprints of a post-Cretaceous dietary radiation in placental mammals. *Sci Adv* **4**:eaar6478. doi:10.1126/sciadv.aar6478

Emms DM, Kelly S. 2019. OrthoFinder: Phylogenetic orthology inference for comparative genomics. *Genome Biol* **20**:238. doi:10.1186/s13059-019-1832-y

Ferreira-Cardoso S, Delsuc F, Hautier L. 2019. Evolutionary tinkering of the mandibular canal linked to convergent regression of teeth in placental mammals. *Curr Biol* **29**:468–475.e3. doi:10.1016/J.CUB.2018.12.023

Ferreira-Cardoso S, Fabre P-H, de Thoisy B, Delsuc F, Hautier L. 2020. Comparative masticatory myology in anteaters and its implications for interpreting morphological convergence in myrmecophagous placentals. *PeerJ* **8**:e9690. doi:10.7717/peerj.9690

Francischetti IMB, Assumpção TCF, Ma D, Li Y, Vicente EC, Uieda W, Ribeiro JMC. 2013. The "Vampirome": Transcriptome and proteome analysis of the principal and accessory submaxillary glands of the vampire bat *Desmodus rotundus* , a vector of human rabies. *J Proteomics* **82**:288–319. doi:10.1016/J.JPROT.2013.01.009

Funkhouser JD, Aronson NN. 2007. Chitinase family GH18: evolutionary insights from the genomic history of a diverse protein family. *BMC Evol Biol* **7**:96. doi:10.1186/1471-2148-7-96

Gordon-Thomson C, Kumari A, sciences LT-… molecular life, 2009 undefined. 2009. Chitotriosidase and gene therapy for fungal infections. *Cell Mol life Sci* **66**:1116–1125.

Gould SJ. 2002. The structure of evolutionary theory. Harvard University Press.

Gould SJ. 1990. Wonderfull life: the Burgess Shale and the nature of history. WW Norton & Company.

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. doi:10.1093/molbev/msp259

Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q, Chen Z, Mauceli E, Hacohen N, Gnirke A, Rhind N, di Palma F, Birren BW, Nusbaum C, Lindblad-Toh K, Friedman N, Regev A. 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol* **29**:644–652. doi:10.1038/nbt.1883

Hamid R, Khan MA, Ahmad M, Ahmad MM, Abdin MZ, Musarrat J, Javed S. 2013. Chitinases: An update. *J Pharm Bioallied Sci* **5**:21. doi:10.4103/0975-7406.106559

Huber W, von Heydebreck A, Sueltmann H, Poustka A, Vingron M. 2003. Parameter estimation for the calibration and variance stabilization of microarray data. *Stat Appl Genet Mol Biol* **2**. doi:10.2202/1544-6115.1008

Hussain M, Wilson J. 2013. New paralogues and revised time line in the expansion of the vertebrate GH18 family. *J Mol Evol* **76**:240–260. doi:10.1007/S00239-013-9553-4

Irisarri I, Baurain D, Brinkmann H, Delsuc F, Sire J-Y, Kupfer A, Petersen J, Jarek M, Meyer A, Vences M, Philippe H. 2017. Phylotranscriptomic consolidation of the jawed vertebrate timetree. *Nat Ecol Evol* **1**:1370–1378. doi:10.1038/s41559-017-0240-5

Jacob F. 1977. Evolution and tinkering. *Science (80-)* **196**:1161–1166.

Janiak MC, Chaney ME, Tosi AJ. 2018. Evolution of acidic mammalian chitinase genes (CHIA) is related to body mass and insectivory in primates. *Mol Biol Evol* **35**:607–622. doi:10.1093/molbev/msx312

Jeuniaux CC. 1971. Chitinous structures. *Compr Biochem* **26**:595–632.

Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment oftware version 7: Improvements in performance and usability. *Mol Biol Evol* **30**:772–780. doi:10.1093/molbev/mst010

Kingdon J. 2014. Mammals of Africa: Volume V: Carnivores, Pangolins, Equids and Rhinoceroses.

A&C Black.

Kozlov AM, Darriba D, Flouri T, Morel B, Stamatakis A. 2019. RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. *Bioinformatics* **35**:4453–4455. doi:10.1093/bioinformatics/btz305

Laheri S, Ashary N, Bhatt P, Modi D. 2018. Oviductal glycoprotein 1 (OVGP1) is expressed by endometrial epithelium that regulates receptivity and trophoblast adhesion. *J Assist Reprod Genet* **35**:1419–1429. doi:10.1007/s10815-018-1231-4

Le SQ, Gascuel O. 2008. An improved general amino acid replacement matrix. *Mol Biol Evol* **25**:1307–1320. doi:10.1093/molbev/msn067

Lee CG, Da Silva CA, Dela Cruz CS, Ahangari F, Ma B, Kang M-J, He C-H, Takyar S, Elias JA. 2011. Role of chitin and chitinase/chitinase-like proteins in inflammation, tissue remodeling, and injury. *Annu Rev Physiol* **73**:479–501. doi:10.1146/annurev-physiol-012110-142250

Losos JB. 2017. Improbable destinies: Fate, chance, and the future of evolution. Penguin.

Losos JB. 2011. Convergence, adaptation, and constraint. *Evolution (N Y)* **65**:1827–1840. doi:10.1111/j.1558-5646.2011.01289.x

Love MI, Huber W, Anders S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* **15**:550. doi:10.1186/s13059-014-0550-8

Ma J-E, Jiang H-Y, Li L-M, Zhang X-J, Li H-M, Li G-Y, Mo D-Y, Chen J-P. 2019. SMRT sequencing of the full-length transcriptome of the Sunda pangolin (*Manis javanica*). *Gene* **692**:208–216. doi:10.1016/J.GENE.2019.01.008

Ma J-E, Li L-M, Jiang H-Y, Zhang X-J, Li J, Li G-Y, Yuan L-H, Wu J, Chen J-P. 2017. Transcriptomic analysis identifies genes and pathways related to myrmecophagy in the Malayan pangolin (*Manis javanica*). *PeerJ* **5**:e4140. doi:10.7717/peerj.4140

MacManes MD. 2014. On the optimal trimming of high-throughput mRNA sequence data. *Front Genet* **5**:13. doi:10.3389/fgene.2014.00013

McGhee GR. 2011. Convergent evolution: limited forms most beautiful. MIT Press.

McNab BK. 1984. Physiological convergence amongst ant-eating and termite-eating mammals. *J Zool* **203**:485–510. doi:10.1111/j.1469-7998.1984.tb02345.x

Meredith RW, Janečka JE, Gatesy J, Ryder OA, Fisher CA, Teeling EC, Goodbla A, Eizirik E, Simão TLL, Stadler T, Rabosky DL, Honeycutt RL, Flynn JJ, Ingram CM, Steiner C, Williams TL, Robinson TJ, Burk-Herrick A, Westerman M, Ayoub NA, Springer MS, Murphy WJ. 2011. Impacts of the cretaceous terrestrial revolution and KPg extinction on mammal diversification. *Science (80-)* **334**:521–524. doi:10.1126/SCIENCE.1211028

Morel B, Kozlov AM, Stamatakis A, Szöllősi GJ. 2020. GeneRax: A tool for species-tree-aware maximum likelihood-based gene family tree Inference under gene duplication, transfer, and loss. *Mol Biol Evol* **37**:2763–2774. doi:10.1093/molbev/msaa141

Morris SC. 1998. The crucible of creation: the Burgess Shale and the rise of animals. Peterson's.

Nguyen L-T, 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. doi:10.1093/molbev/msu300

Novacek MJ. 1992. Mammalian phytogeny: shaking the tree. *Nature* **356**:121–125. doi:10.1038/356121a0

O'Leary MA, Bloch JI, Flynn JJ, Gaudin TJ, Giallombardo A, Giannini NP, Goldberg SL, Kraatz BP, Luo Z-X, Meng J, Ni X, Novacek MJ, Perini FA, Randall ZS, Rougier GW, Sargis EJ, Silcox MT, Simmons NB, Spaulding M, Velazco PM, Weksler M, Wible JR, Cirranello AL. 2013. The placental mammal ancestor and the post–K-Pg radiation of placentals. *Science (80-)* **339**:662–667. doi:10.1126/SCIENCE.1229237

Olland AM, Strand J, Presman E, Czerwinski R, Joseph-McCarthy D, Krykbaev R, Schlingmann G, Chopra R, Lin L, Fleming M, Kriz R, Stahl M, Somers W, Fitz L, Mosyak L. 2009. Triad of polar residues implicated in pH specificity of acidic mammalian chitinase. *Protein Sci* **18**:NA-NA. doi:10.1002/pro.63

Perry GH, Melsted P, Marioni JC, Wang Y, Bainer R, Pickrell JK, Michelini K, Zehr S, Yoder AD, Stephens M, Pritchard JK, Gilad Y. 2012. Comparative RNA sequencing reveals substantial genetic variation in endangered primates. *Genome Res* **22**:602–610. doi:10.1101/GR.130468.111

Phillips CJ, Phillips CD, Goecks J, Lessa EP, Sotero-Caio CG, Tandler B, Gannon MR, Baker RJ. 2014. Dietary and flight energetic adaptations in a salivary gland transcriptome of an insectivorous bat. *PLoS One* **9**:e83512. doi:10.1371/journal.pone.0083512

Pillai AS, Chandler SA, Liu Y, Signore A V., Cortez-Romero CR, Benesch JLP, Laganowsky A, Storz JF, Hochberg GKA, Thornton JW. 2020. Origin of complexity in haemoglobin evolution. *Nature* **581**:480–485. doi:10.1038/s41586-020-2292-y

Recklies AD, White C, Ling H. 2002. The chitinase 3-like protein human cartilage glycoprotein 39 (HC-gp39) stimulates proliferation of human connective-tissue cells and activates both extracellular signal-regulated kinase- and protein kinase B-mediated signalling pathways. *Biochem J* **365**:119–126. doi:10.1042/BJ20020075

Redford KH. 1987. Ants and termites as foodCurrent Mammalogy. Boston, MA: Springer US. pp. 349–399. doi:10.1007/978-1-4757-9909-5_9

Reiss KZ. 2001. Using phylogenies to study convergence: The case of the ant-eating mammals. *Am Zool* **41**:507–525. doi:10.1093/icb/41.3.507

Saint-Dizier M, Marnier C, Tahir MZ, Grimard B, Thoumire S, Chastant-Maillard S, Reynaud K. 2014. *OVGP1* is expressed in the canine oviduct at the time and place of oocyte maturation and fertilization. *Mol Reprod Dev* **81**:972–982. doi:10.1002/mrd.22417

Salgaonkar N, Prakash D, Nawani NN, Kapadnis BP. 2015. Comparative studies on ability of N-acetylated chitooligosaccharides to scavenge reactive oxygen species and protect DNA from oxidative damage. *Indian J Biotechnol* 186–192.

Slater G, Birney E. 2005. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* **6**:31. doi:10.1186/1471-2105-6-31

Soneson C, Love MI, Robinson MD. 2015. Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. *F1000Research* **4**:1521. doi:10.12688/f1000research.7563.1

Springer MS, Murphy WJ, Eizirik E, O'Brien SJ. 2003. Placental mammal diversification and the Cretaceous–Tertiary boundary. *Proc Natl Acad Sci* **100**:1056–1061. doi:10.1073/PNAS.0334222100

Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**:1312–1313. doi:10.1093/bioinformatics/btu033

Tibshirani R. 1988. Estimating transformations for regression via additivity and variance stabilization. *J Am Stat Assoc* **83**:394–405. doi:10.1080/01621459.1988.10478610

Tjoelker LW, Gosting L, Frey S, Hunter CL, Trong H Le, Steiner B, Brammer H, Gray PW. 2000. Structural and functional definition of the human chitinase chitin-binding domain. *J Biol Chem* **275**:514–520. doi:10.1074/JBC.275.1.514

Tucker R. 1958. Taxonomy of the salivary glands of vertebrates. *Syst Zool* **7**:74. doi:10.2307/2411794

Vandewege MW, Sotero-Caio CG, Phillips CD. 2020. Positive selection and gene expression analyses from salivary glands reveal discrete adaptations within the ecologically diverse bat family phyllostomidae. *Genome Biol Evol* **12**:1419–1428. doi:10.1093/gbe/evaa151

Wang K, Tian S, Galindo-González J, Dávalos LM, Zhang Y, Zhao H. 2020. Molecular adaptation and convergent evolution of frugivory in Old World and neotropical fruit bats. *Mol Ecol* **29**:4366–

4381. doi:10.1111/mec.15542

Waterhouse RM, Seppey M, Simão FA, Manni M, Ioannidis P, Klioutchnikov G, Kriventseva E V, Zdobnov EM. 2018. BUSCO applications from quality assessments to gene prediction and phylogenomics. *Mol Biol Evol* **35**:543–548. doi:10.1093/molbev/msx319

Wolf JBW. 2013. Principles of transcriptome analysis and gene expression quantification: an RNA-seq tutorial. *Mol Ecol Resour* **13**:559–572. doi:10.1111/1755-0998.12109

Xie VC, Pu J, Metzger BPH, Thornton JW, Dickinson BC. 2020. Chance, contingency, and necessity in the experimental evolution of ancestral proteins. *bioRxiv* 2020.08.29.273581. doi:10.1101/2020.08.29.273581

Yates AD, Achuthan P, Akanni W, Allen J, Allen J, Alvarez-Jarreta J, Amode MR, Armean IM, Azov AG, Bennett R, Bhai J, Billis K, Boddu S, Marugán JC, Cummins C, Davidson C, Dodiya K, Fatima R, Gall A, Giron CG, Gil L, Grego T, Haggerty L, Haskell E, Hourlier T, Izuogu OG, Janacek SH, Juettemann T, Kay M, Lavidas I, Le T, Lemos D, Martinez JG, Maurel T, McDowall M, McMahon A, Mohanan S, Moore B, Nuhn M, Oheh DN, Parker A, Parton A, Patricio M, Sakthivel MP, Abdul Salam AI, Schmitt BM, Schuilenburg H, Sheppard D, Sycheva M, Szuba M, Taylor K, Thormann A, Threadgold G, Vullo A, Walts B, Winterbottom A, Zadissa A, Chakiachvili M, Flint B, Frankish A, Hunt SE, IIsley G, Kostadima M, Langridge N, Loveland JE, Martin FJ, Morales J, Mudge JM, Muffato M, Perry E, Ruffier M, Trevanion SJ, Cunningham F, Howe KL, Zerbino DR, Flicek P. 2019. Ensembl 2020. *Nucleic Acids Res* **48**:D682–D688. doi:10.1093/nar/gkz966

⇧Back to summary⇧

– CONCLUSIONS & PERSPECTIVES –

The diet specialization of ant-eating mammals (i.e. myrmecophagy) is among the most famous examples of evolutionary convergence. This particular lifestyle evolved in five distinct lineages of placental mammals: the aardvark, the aardwolf, the anteaters, the giant armadillo, and the pangolins. The selective pressures associated with the high specialization of their diet towards ants and termites consumption led to extreme morphological convergences through time. In that context, using phylogenomics and comparative genomic approaches, the objective of my PhD project was to understand the molecular processes associated with this peculiar evolutionary history.

In the first chapter of my thesis, I presented a strategy to take advantage of metagenomic data extracted from fecal samples of myrmecophagous mammals for diet characterization, as an alternative to existing methods based on metabarcoding (e.g. Pompanon et al. 2012; Shehzad et al. 2012; Alberti et al. 2018; Galan et al. 2018, Gauthier et al. 2020). The first step consisted in collecting both fecal samples and potential preys (ants and termites) in the native home ranges of myrmecophagous mammals. Our fieldwork efforts for this study were focused on two specific reserves of South Africa, in which we tried to exhaustively collect every ant and termite species encountered. The objective of this sampling was to construct a specific database against which mitochondrial sequences extracted from metagenomic fecal samples could be compared. To create these mitochondrial databases, we developed MitoFinder, a user-friendly pipeline to efficiently assemble, extract and annotate mitochondrial sequences from high throughput sequencing data (Allio et al. 2020). This pipeline proved its efficiency in successfully extracting mitogenomic sequences for most ant and termite samples and allowed us to create two databases, including 87 and 222 individuals of termites and ants, respectively. Then, using MitoFinder on the metagenomic data extracted from fecal samples we were unable to extract mitochondrial signals corresponding to either ant or termite species. However, using the ant and termite mitochondrial contigs generated with MitoFinder as references for mapping fecal metagenomic reads, we were able to detect few reads corresponding to the preys consumed by myrmecophagous species (according to previous studies: Weyer 2018, Panaino 2020). Given that only a few reads were retrieved/mapped for each species (ranging from 2 to 30 reads), this strategy allowed a first preliminary molecular assessment of the diet of these species but further analyses are needed. The few number of reads recovered from fecal samples is likely due to the vast overrepresentation of bacterial DNA fragments in fecal sample extractions (Yang et al. 2020). In that context, we plan to take advantage of the database of ants and termites created during this thesis to design specific baits for preferentially sequencing prey DNA fragments in myrmecophagous mammals' fecal samples (Gauthier et al. 2020).

The second chapter of this thesis presented the development of experimental and bioinformatic approaches to generate high quality genomes of myrmecophagous mammals from roadkill samples. Myrmecophagous mammal species present relatively large genomes, ranging from 2.5 Gb in pangolins to 4.5 Gb in xenarthrans. To be able to generate high quality assemblies, both in terms of contiguity and completeness, we decided to rely on a hybrid assembling strategy. This strategy consists in taking advantages from both the high accuracy of short reads generated by next generation sequencing methods and the size of long reads generated with third generation sequencing strategies. Because of the non-optimal preservation of our tissues, the resulting extractions for our species were of too low quality to be accepted, three years ago, by long-read sequencing platforms. In that context, we decided to develop an optimized protocol for sequencing mammalian roadkill tissues with the MinION portable sequencer developed by Oxford Nanopore Technologies (ONT). Briefly, this protocol (available here from Protocols.io: dx.doi.org/10.17504/protocols.io.beixjcfn) consists (i) in preserving tissues in RNAlater instead of traditional 95% Ethanol, (ii) preferentially extract the most well-preserved parts of the tissue and remove all perceptible impurities, and (iii) adjusting the ratio of AMPure beads to 0.4x to optimize size selection during ONT library construction. Applying this optimized protocol, we were able to generate good quality long read sequencing data for all nine focal species of the project. Then, Illumina short reads and ONT long reads were used conjointly through an hybrid assembly approach (implemented in MaSuRCA) that resulted in high quality assemblies. These assemblies are highly contiguous with a number of contigs ranging from 51,157 to 4,309, which is much less than previously available assemblies for myrmecophagous mammals, especially xenarthrans (Zoonomia Consortium 2020). Similarly, BUSCO analyses (Waterhouse et al. 2018), estimating the completeness in previously-defined orthologous genes, suggest a high level of completeness for these genomes. Interestingly, despite their high quality, xenarthran genomes do not exceed BUSCO scores of 90% of complete genes. This result may be due to the difficulty in assembling the genomic regions containing these genes or might signify that those genes are simply absent from xenarthran genomes.

Once the genomes were sequenced and assembled, the next step was to annotate them. First, repeat elements were identified and masked for further annotation steps. Then, we decided to rely on evidence-based and *ab initio* gene predictions (Yandell & Ence 2012). In that sense, for evidence-based gene prediction, we took advantage of both transcriptomic data assembled and annotated for this purpose (70 transcriptomes) and available manually-curated genes reference databases (uniprot/SWISSPROT) to annotate our genomes. The information obtained from evidence-based gene prediction and *ab initio* training gene prediction were summarized with the pipeline implemented in MAKER 3 (Yandell 2011). To improve the accuracy of the annotations, this pipeline was run three times iteratively by intercalating an *ab initio* training step based on previous annotation results (Korf 2004, Stanke et al. 2006). Given that sequencing, basecalling, assembly and genome annotation steps are relatively long processes, only two genomes generated during my PhD project, the genomes of

Smutsia gigantea and *Myrmecophaga tridactyla*, were fully annotated. Indeed, the sequencing of all genomes with ONT took a total of 22 months. Then, the conversion of raw sequencing information to sequencing reads took around two weeks per genome followed by the hybrid assembly step for which about 3-4 additional weeks were needed. Finally, two to three weeks were necessary to run the annotation pipeline on each genome assembly (considering all external data as available, e.g. annotated transcriptome assemblies).

Although this pipeline represents a lengthy process, the resulting annotated genomes provide an inestimable resource to study the evolutionary convergence of myrmecophagous mammals. Indeed, using the annotations produced by our analyses, it will be much easier to accurately extract orthologous genes. In our case, we plan to use the same pipeline as the one developed to create the OrthoMaM database (Scornavacca et al. 2019) to assemble a genomic dataset including both myrmecophagous and non-myrmecophagous species. The combination of our newly generated genomes and the ones selected from OrthoMaM will provide an excellent dataset to detect potential traces of molecular convergence associated with the adaptation to myrmecophagy in single-copy orthologous genes. After having inferred a solid phylogenetic backbone using the approach described in this thesis (Part II - section 1), different approaches will be used to investigate molecular convergence in myrmecophagous mammals. Indeed, this pipeline has been established during my PhD project by Mathilde Barthe, a master student that I had the chance to co-supervised with Frédéric Delsuc. Mathilde worked on a dataset composed of 12 Carnivora species, including the aardwolf (*Proteles cristatus*) and the bat-eared fox (*Otocyon megalotis*), both having a diet composed of more than 70% of ants and termites. She designed a pipeline to search for molecular convergences in genes based on (i) convergent amino acid substitutions (using PCOC, Rey et al. 2018), (ii) convergent traces of adaptive selection (dN/dS analysis, Yang 2007), and (iii) similarity in evolutionary rates between convergent lineages (RERconverge, Kowalczyk et al. 2018). Although this master project was conducted on a reduced dataset, it allowed the development of a pipeline investigating molecular convergence at different scales. Furthermore, numerous genes were pointed out conjointly by the three different approaches implemented in the pipeline, but a large proportion of these genes were false positives resulting from the bad quality of the genome annotations (previously annotated with genBlastG, She et al. 2011). These results encouraged us to implement the sophisticated annotation pipeline presented in the second chapter of this thesis. Now that we have more accurate annotations, the objective will be to apply the pipeline previously implemented by Mathilde on our newly-generated dataset. Finally, I would like to look at molecular convergence with another interesting approach developed by Wu et al. (2017). Briefly, this approach consists in inferring ancestral states (historical traits such as the diet) based on the associations observed between the states of extant species and the evolution of their genes (resumed as gene profiles, corresponding to the evolutionary rate of the genes at the tips of the phylogeny). Specific gene profiles are associated with every extant

state. Then, for each node of the phylogeny, the evolutionary rates inferred are compared to each profile and the ancestral state selected during the inference is the one with the most similar profile. By searching for a specific gene profile associated with a given state (in our case, specialization in eating ants and/or termites), this approach could help us to point out convergent evolution at the scale of entire genes or even set of genes.

The full annotations of our genomes will also allow us to study the evolutionary history of other potentially interesting genomic portions besides single-copy orthologous genes. For instance, it has been shown that conserved non-coding regions involved in the regulation of gene expression, have convergently evolved in flightless paleognathous birds (Sackton et al. 2019). Using a genome-scale dataset, Sackton et al. (2019), extracted about 280,000 conserved non-exonic elements (CNEEs) having a potential regulatory role in birds and other taxa. Among them, they found a large number of portions having convergently accelerated in flightless paleognaths. These results suggest that convergent evolutionary processes may involve regulatory regions instead of gene evolution in some cases. In this light, Mathilde Barthe just started her PhD project to study convergent evolution in non-coding conserved regions in myrmecophagous mammal genomes.

Another point of particular interest is the evolution of gene families. Indeed, many comparative genomic studies focus on single-copy orthologous genes. However, these genes represent only a small portion of the coding genes found in genomes. Interestingly, candidate gene families have already been reported as having a central role in dietary adaptation, especially when evolving detoxification mechanisms (e.g. Berenbaum et al. 1996). Additionally, in mammals the large gene families of taste receptors (TR) and olfactory receptors (OR) have long been studied and seem to be involved in social interactions, feeding and mating (Dulac & Torello 2003; Shi et al. 2003; Bachmanov & Beauchamp 2007; Hayden & Teeling 2014; Rymer 2020). Overall, the development of recent sequencing methods allow now studying more precisely the evolution of gene families (e.g. Hayden et al. 2014; Edger et al. 2015; Yohe et al. 2019; Thomas et al. 2020). Hence, the full annotation of our genomes will help us reconstruct the evolution of gene families in myrmecophagous mammals. In that context, Sophie Teullet just started her PhD project in which she will focus on the evolution of the TR and OR gene families in mammals with a particular interest in myrmecophagous species convergence.

Finally, in the third chapter of this thesis, I presented comparative transcriptomic analyses of salivary glands and other organs in myrmecophagous mammals. Indeed, among the numerous morphological convergences observed in myrmecophagous mammals, the hypertophysis of salivary glands is particularly remarkable. Salivary glands of myrmecophagous mammals likely play an important role in insect digestion as suggested by the high expression of digestive enzymes in this organ in the Malayan pangolin (Ma et al. 2017). In this third chapter, I first presented a global analysis of salivary gland transcriptomes of 23 species (28 individuals) including both non-myrmecophagous and myrmecophagous mammals. The transcriptomes were annotated and expression level of transcript orthogroups were compared between taxonomic and dietary groups. Overall, preliminary analyses suggest that global gene expression in mammalian salivary glands is mainly driven by the evolutionary history of species (phylogeny). Indeed, closely related species have more similar expression profiles than more distantly related species. This result is expected and has already been reported in previous studies on multiple mammalian organs (e.g. Brawand et al., 2011). This suggests an important impact of the historical contingency on global gene expression. Interestingly, an example of the effect of contingency history resides on the evolution of digestive enzymes genes belonging to the Chitinase family. This family is composed of five paralogous chitinase genes (CHIA1-5) and Emerling et al. (2018) found a positive correlation between the number of functional (non-pseudogenized) gene copies of chitinase genes and the percent of the diet consisting of invertebrates in placental mammals. Interestingly, although the lesser anteater (*Tamandua tetradactyla*) and the aardvark (*Orycteropus afer*) have four and five functional chitinase copies, the Malayan pangolin (*Manis javanica*) has only one functional copy. Indeed, the Malayan pangolin has only one functional chitinase gene (CHIA5) likely because the common ancestor of the Pholidata and the Carnivora had already lost the four other chitinase genes. In this light, we decided to focus on the expression of the different copies of chitinase, first in mammal salivary glands, and then in myrmecophagous non-digestive and digestive organs. We found that despite different chitinase repertoires, the myrmecophagous species (*Manis javanica* and *Tamandua tetradactyla*) highly expressed their chitinase genes in digestive organs. In particular, we were able to show that the Malayan pangolin potentially compensates for the paucity of functional chitinase genes by overexpressing CHIA5 in all major digestive organs (salivary glands, tongue, stomach, pancreas, large intestine, and liver, Ma et al., 2019). These results show the importance of the historical contingency in shaping the evolution of organisms through molecular thinkering. Nevertheless, the overexpression of its last available chitinase gene by the Malayan pangolin provides an excellent example of adaptive evolution to counter the effect of historical contingency on dietary adaptation.

As a general conclusion, the different approaches developed during my PhD project, from the sequencing of degraded tissues to the assembly and the annotation of genomes of non-model organisms, allow us to generate nine high quality mammal genomes. These genomes provide an inestimable resource to study the evolutionary convergence of myrmecophagous mammals. By combining the genes extracted from these genomes with available mammal gene databases, we will be able to conduct molecular convergence detection at different levels. Additionally, the example of the evolution of the chitinase gene family presented here joins the many examples showing the impact of historical contingency in the evolution of organisms, suggesting that different evolutionary paths might be followed to adapt to similar conditions. In this light, we plan to take advantage of the full annotation of the genomes generated during the project to combine different detection approaches to study the evolutionary convergence towards myrmecophagy.

References

⇧Back to summary⇧

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 Resour* 1755–0998.13160. doi:10.1111/1755-0998.13160

Bachmanov AA, Beauchamp GK. 2007. Taste receptor genes. *Annu Rev Nutr* **27**:389–414. doi:10.1146/annurev.nutr.26.061505.111329

Berenbaum MR, Favret C, Schuler MA. 1996. On defining "Key Innovations" in an adaptive radiation: Cytochrome P450S and Papilionidae. *Am Nat* **148**:S139–S155. doi:10.1086/285907

Brawand D, Soumillon M, Necsulea A, Julien P, Csárdi G, Harrigan P, Weier M, Liechti A, Aximu-Petri A, Kircher M, Albert FW, Zeller U, Khaitovich P, Grützner F, Bergmann S, Nielsen R, Pääbo S, Kaessmann H. 2011. The evolution of gene expression levels in mammalian organs. *Nature* **478**:343–348. doi:10.1038/nature10532

Dulac C, Torello AT. 2003. Molecular detection of pheromone signals in mammals: from genes to behaviour. *Nat Rev Neurosci* **4**:551–562. doi:10.1038/nrn1140

Emerling CA, Delsuc F, Nachman MW. 2018. Chitinase genes (*CHIA* s) provide genomic footprints of a post-Cretaceous dietary radiation in placental mammals. *Sci Adv* **4**:eaar6478. doi:10.1126/sciadv.aar6478

Galan M, Pons J-B, Tournayre O, Pierre É, Leuchtmann M, Pontier D, Charbonnel N. 2018. Metabarcoding for the parallel identification of several hundred predators and their prey: Application to bat species diet analysis. *Mol Ecol Resour* **18**:474–489. doi:10.1111/1755-0998.12749

Gauthier M, Konecny-Dupré L, Nguyen A, Elbrecht V, Datry T, Douady C, Lefébure T. 2020. Enhancing DNA metabarcoding performance and applicability with bait capture enrichment and DNA from conservative ethanol. *Mol Ecol Resour* **20**:79–96. doi:10.1111/1755-0998.13088

Hayden S, Bekaert M, Goodbla A, Murphy WJ, Dávalos LM, Teeling EC. 2014. A cluster of olfactory receptor genes linked to frugivory in bats. *Mol Biol Evol* **31**:917–927. doi:10.1093/molbev/msu043

Holt C, Yandell M. 2011. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics* **12**:491. doi:10.1186/1471-2105-12-491

Kowalczyk A, Meyer WK, Partha R, Mao W, Clark NL, Chikina M. 2019. RERconverge: an R

package for associating evolutionary rates with convergent traits. *Bioinformatics* **35**:4815–4817. doi:10.1093/bioinformatics/btz468

Ma J-E, Jiang H-Y, Li L-M, Zhang X-J, Li H-M, Li G-Y, Mo D-Y, Chen J-P. 2019. SMRT sequencing of the full-length transcriptome of the Sunda pangolin (*Manis javanica*). *Gene* **692**:208–216. doi:10.1016/J.GENE.2019.01.008

Ma J-E, Li L-M, Jiang H-Y, Zhang X-J, Li J, Li G-Y, Yuan L-H, Wu J, Chen J-P. 2017. Transcriptomic analysis identifies genes and pathways related to myrmecophagy in the Malayan pangolin (*Manis javanica*). *PeerJ* **5**:e4140. doi:10.7717/peerj.4140

Panaino W. 2020. Diet, activity, and body temperature patterns of ground pangolins in a semi-arid environment. University of the Witwatersrand, Johannesburg.

Pompanon F, Deagle BE, Symondson WOC, Brown DS, Jarman SN, Taberlet P. 2012. Who is eating what: diet assessment using next generation sequencing. *Mol Ecol* **21**:1931–1950. doi:10.1111/j.1365-294X.2011.05403.x

Rey C, Guéguen L, Sémon M, Boussau B. 2018. Accurate detection of convergent amino-acid evolution with PCOC. *Mol Biol Evol* **35**:2296–2306. doi:10.1093/molbev/msy114

Rymer TL. 2020. The role of olfactory genes in the expression of rodent paternal care behavior. *Genes (Basel)* **11**:292. doi:10.3390/genes11030292

Sackton TB, Clark N. 2019. Convergent evolution in the genomics era: new insights and directions. *Philos Trans R Soc B Biol Sci* **374**:20190102. doi:10.1098/rstb.2019.0102

Scornavacca C, Belkhir K, Lopez J, Dernat R, Delsuc F, Douzery EJP, Ranwez V. 2019. OrthoMaM v10: Scaling-up orthologous coding sequence and exon alignments with more than one hundred mammalian genomes. *Mol Biol Evol* **36**:861–862. doi:10.1093/molbev/msz015

She R, Chu JS-C, Uyar B, Wang J, Wang K, Chen N. 2011. genBlastG: using BLAST searches to build homologous gene models. *Bioinformatics* **27**:2141–2143. doi:10.1093/bioinformatics/btr342

Shehzad W, Riaz T, Nawaz MA, Miquel C, Poillot C, Shah SA, Pompanon F, Coissac E, Taberlet P. 2012. Carnivore diet analysis based on next-generation sequencing: application to the leopard cat (*Prionailurus bengalensis*) in Pakistan. *Mol Ecol* **21**:1951–1965. doi:10.1111/j.1365-294X.2011.05424.x

Shi P, Zhang J, Yang H, Zhang Y. 2003. Adaptive diversification of bitter taste receptor genes in mammalian evolution. *Mol Biol Evol* **20**:805–814. doi:10.1093/molbev/msg083

Waterhouse RM, Seppey M, Simão FA, Manni M, Ioannidis P, Klioutchnikov G, Kriventseva E V, Zdobnov EM. 2018. BUSCO applications from quality assessments to gene prediction and phylogenomics. *Mol Biol Evol* **35**:543–548. doi:10.1093/molbev/msx319

Weyer NM. 2018. Physiological flexibility of free-living aardvarks (Orycteropus afer) in response to environmental fluctuations. University of the Witwatersrand, Johannesburg, South Africa.

Wu J, Yonezawa T, Kishino H. 2017. Rates of molecular evolution suggest natural history of life history traits and a Post-K-Pg nocturnal bottleneck of placentals. *Curr Biol* **27**:3025–3033. doi:10.1016/J.CUB.2017.08.043

Yandell M, Ence D. 2012. A beginner's guide to eukaryotic genome annotation. *Nat Rev Genet* **13**:329–342. doi:10.1038/nrg3174

Yang F, Sun J, Luo H, Ren H, Zhou H, Lin Y, Han M, Chen B, Liao H, Brix S, Li J, Yang H, Kristiansen K, Zhong H. 2020. Assessment of fecal DNA extraction protocols for metagenomic studies. *Gigascience* **9**. doi:10.1093/gigascience/giaa071

Yang Z. 2007. PAML 4: Phylogenetic analysis by maximum likelihood. *Mol Biol Evol* **24**:1586–1591. doi:10.1093/molbev/msm088

Zoonomia consortium. 2020. A comparative genomics multitool for scientific discovery and conservation. *Nature* **587**:240–245. doi:10.1038/s41586-020-2876-6

⇧Back to summary⇧

– APPENDICES –

During my PhD project, I had the opportunity to continue and publish my Master project supervised by Fabien Condamine and Benoit Nabholz. This project aimed at studying the phylogenetic relationships within swallowtail butterflies (Papilionidae) and the role of their host plants in their evolutionary diversification. First, based on a molecular matrix composed of 6621 genes extracted from about 60 genomes of butterflies, we generated a robust time-calibrated phylogeny for this family that confirmed previous genus-level relationships but also unveiled new relationships (**Appendix 1**). Second, we used this dataset and investigated genome-wide macroevolutionary signatures of butterflies' adaptation when colonizing new host plants (**Appendix 2**). Interestingly, we found that more genes were positively selected in phylogenetic branches leading to host-plant shifts than in branches without host-plant shifts. Additionally, host-plant shifts were generally associated with bursts of speciation rates. Overall, these results support the importance of host plants in the evolution and diversification of butterflies and encourage the use of genomic datasets to better understand the evolution of organisms, especially when they evolved in interaction.

Appendix 1 – Whole genome shotgun phylogenomics resolves the pattern and timing of swallowtail butterfly evolution

The journal article associated with this appendix can be found online:
https://doi.org/10.1093/sysbio/syz030
As well as the supplementary material:
https://doi.org/10.5061/dryad.ff18q9d

⇧Back to summary⇧

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

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

[1]*Institut des Sciences de l'Evolution de Montpellier (Université de Montpellier | CNRS | IRD | EPHE), Place Eugène Bataillon, 34095 Montpellier, France;*
[2]*Institut de Biologie Computationnelle (IBC), Montpellier, France;*
[3]*INRA, UMR 1062 Centre de Biologie pour la Gestion des Populations (INRA, IRD, CIRAD, Montpellier SupAgro), 755 avenue du Campus Agropolis, 34988 Montferrier-sur-Lez, France;*
[4]*University of Alberta, Department of Biological Sciences, Edmonton T6G 2E9, AB, Canada.*

***Correspondence:** remi.allio@umontpellier.fr; fabien.condamine@gmail.com

Abstract

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

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

ID		Data		Genome assembly				Cross-contamination analyses			Final *Data set 1*		Final *Data set 2*	
Species name	Voucher	Number of raw reads (M)	Reads after cleaning (M)	Assembly size (M)	N50	Number of scaffolds	Average length	Mean %GC	Number of genes before CroCo: Data set 1	Number of genes before CroCo: Data set 2	% contamination	Number of genes	% missing data	Number of genes	% missing data	SRA number
Allancastria cerisyi	2081	73.342	60.072	146.973	150	905,391	162.3	36.30	328	485	0.169	104	93.84	543	95.52	SRR8954514
Archon apollinus	2124	67.361	62.930	787.111	398	2,752,266	286	35.69	954	1840	0.242	748	18.29	4189	40.17	SRR8954513
Atrophaneura dixoni	2008	59.715	55.536	369.571	1616	785,140	470.7	39.77	957	1913	0.327	758	5.22	4708	25.3	SRR8954516
Baronia brevicornis	167	77.182	72.038	488.028	1169	973,148	501.5	35.21	959	1757	0.472	760	7.72	4344	33.39	SRR8954515
Battus polydamas	308	76.666	69.456	453.655	427	1,499,801	302.5	33.23	957	1902	0.010	758	12.21	4506	33.2	SRR8954510
Bhutanitis mansfieldi	2097	44.295	33.719	317.424	200	1,637,993	193.8	36.28	890	1526	0.011	675	44.12	2746	68.7	SRR8954509
Byasa alcinous	93	94.005	86.018	319.164	2081	443,395	719.8	33.82	959	1942	0.094	760	4.72	4760	24.3	SRR8954512
Cressida cressida	1827	66.593	59.979	459.866	558	1,414,336	325.1	32.48	954	1876	0.231	754	10.18	4515	31.6	SRR8954511
Eurypides corethrus	3716	51.927	45.178	343.287	301	1,372,876	250	33.15	957	1912	0.010	758	9.66	4595	30.42	SRR8954508
Eurytides dolicaon	11002	76.265	68.143	589.135	354	2,185,856	269.5	34.77	957	1873	9.303	720	18.22	3885	44.7	SRR8954507
Graphium androcles	1904	62.412	56.862	933.096	495	3,089,220	302	33.14	957	1833	2.464	751	13.4	4262	37.67	SRR8954549
Graphium sarpedon	2021	36.692	34.228	590.915	221	2,775,869	212.9	33.04	938	1718	0.122	748	27.6	3719	52.04	SRR8954548
Hypermnestra helios	Hyp.01-01a	17.743	10.851	164.657	150	1,040,790	158.2	36.27	887	1540	0.000	645	53.52	2500	75.51	SRR8954546
Hypermnestra helios	2069	33.169	29.511	292.277	161	1,741,883	167.8	37.74			4.278					SRR8954547
Iphiclides podalirius	6	38.605	34.373	390.909	599	1,041,389	375.4	35.92	959	1876	0.105	759	14.08	4378	36.37	SRR8954545
Lamproptera meges	1955	59.278	53.733	512.606	275	2,153,366	238	33.64	957	1824	0.650	755	16.52	4232	40.68	SRR8954544
Losaria neptunus	3706	31.248	27.061	248.117	449	785,959	315.7	36.02	939	1794	0.054	751	17.01	4164	41.5	SRR8954543
Luehdorfia japonica	356	82.816	74.878	543.466	360	1,930,389	281.5	37.07	957	1901	0.182	757	13.31	4455	33.91	SRR8954542
Meandrusa sciron	1896	42.681	39.647	509.369	479	1,699,899	299.6	38.02	959	1996	0.009	759	10.48	4740	28.87	SRR8954541
Mimoides lysithous	3789	92.328	83.586	541.013	730	1,377,484	392.8	35.65	958	1889	0.057	759	7.13	4595	28.92	SRR8954540
Ornithoptera priamus	1919	62.251	57.565	419.015	1305	762,773	549.3	36.78	958	1931	0.009	760	4.66	4770	23.97	SRR8954527
Ornithoptera richmondia	1938	70.462	60.269	480.793	315	1,827,473	263.1	39.20	957	1891	1.979	749	15.41	4326	38.12	SRR8954528
Pachliopta kotzebuea	1920	94.744	87.123	304.109	5273	161,975	1877.5	35.38	958	1920	8.394	747	6.3	4372	30.55	SRR8954525
Papilio alexanor	Alx.01-01a	161.124	128.084	409.954	3716	565,709	724.7	36.11	959	2238	0.034	759	4.77	5252	18.32	SRR8954526
Papilio antimachus	Dru.01-01a	195.146	144.268	388.176	2842	646,388	600.5	40.50	959	2270	1.083	750	9.18	5146	23.63	SRR8954523
Papilio glaucus (GenBank)	3	—	—	374.816	230,841	60,470	6198.4	35.47	959	2224	—	760	4.6	5206	19.08	—
Papilio joanae	295	61.752	57.302	408.782	643	1,010,689	404.5	32.72	958	2382	0.123	760	6.43	5465	19.01	SRR8954524
Papilio machaon (GenBank)	1	—	—	278.436	1,174,287	63,187	4406.5	33.77	959	2413	—	759	3.2	5376	14.87	—
Papilio polytes (GenBank)	2	—	—	227.021	3,672,263	3874	58,601.2	33.77	946	2233	—	744	8.67	5098	22.16	—
Papilio slateri	1796	60.549	54.703	356.588	855	811,786	439.3	36.67	959	2213	0.377	758	7.81	5094	23.68	SRR8954521
Papilio thoas	Her.01-01a	153.287	122.570	339.705	1002	748,457	453.9	33.34	959	2192	2.229	744	11.24	4991	27.04	SRR8954522
Papilio xuthus (GenBank)	—	—	—	—	—	—	—	—	959	2482	—	760	1.16	5730	9.76	—
Papilio zelicaon	Pap.01-01a	199.531	164.682	399.147	764	947,128	421.4	33.57	959	2351	1.982	755	13.48	5151	26.89	SRR8954529
Parides photinus	149	67.923	63.876	380.543	835	871,292	436.8	33.64	959	1916	1.010	757	5.61	4684	26.28	SRR8954530
Parnassius honrathi	Parn.01-01a	178.536	150.484	994.537	955	2,130,202	466.9	36.56	959	1872	0.106	758	6.43	4578	28.51	SRR8954539
Parnassius imperator	Kai.01-01a	35.528	21.955	483.372	171	2,712,024	178.2	37.86	841	1307	26.714	394	72.94	1368	86.98	SRR8954538
Parnassius orleans	Drio.01-01a	156.174	120.544	814.564	485	2,954,078	275.7	35.38	955	1908	8.955	708	14.37	4145	38.93	SRR8954532
Parnassius smintheus	1706	103.125	95.149	846.707	379	3,198,623	264.7	35.92	957	1866	0.260	759	11.45	4417	33.94	SRR8954531
Pharmacophagus antenor	1646	120.153	107.409	387.987	4329	399,998	970	40.99	958	1874	1.034	760	5.32	4612	26.79	SRR8954535
Protesilaus protesilaus	3711	79.486	72.676	901.946	234	4,111,085	219.4	33.97	954	1774	10.172	706	27.2	3434	55.15	SRR8954534
Protographium marcellus	7	64.339	58.553	636.959	320	2,472,899	257.6	35.25	955	1872	0.560	757	14.06	4408	35.48	SRR8954537
Sericinus montela	Ser.01-01a	137.724	109.471	526.009	1843	750,961	700.4	35.74	959	1887	2.622	750	7.82	4535	29.83	SRR8954536
Teinopalpus imperialis	1566	64.583	58.453	530.299	848	1,161,826	456.4	35.12	958	1956	0.158	759	6.92	4737	26.84	SRR8954519
Trogonoptera brookiana	2023	35.422	30.401	282.281	423	897,158	314.6	34.22	956	1888	10.976	699	25.74	3865	47.93	SRR8954533
Troides plateni	Tr.20-01h	168.313	134.462	449.338	2905	693,452	648	42.36	959	1904	0.892	758	5.63	4664	25.82	SRR8954518
Zerynthia polyxena	2122	32.378	29.101	461.127	255	1,969,767	234.1	34.39	939	1751	0.091	738	30.92	3638	54.68	SRR8954517
Agraulis vanillae	NA	—	—	390.758	21,413	45,022	8679.3	31.80	946	1547	—	754	12.05	3702	44.52	—
Bicyclus anynana	NA	—	—	—	—	—	—	—	941	2273	—	754	9.87	5157	27.72	—
Bombyx mori	NA	—	—	—	—	—	—	—	938	2191	—	739	16.39	5044	31.44	—
Calycopis cecrops	NA	—	—	—	—	—	—	—	875	1826	—	714	17.17	4365	38.84	—
Choristoneura fumiferana	399	155.399	139.886	570.774	606	1,574,167	362.6	36.59	946	1548	0.000	752	15.63	3652	47.38	SRR8954520
Danaus plexippus	NA	—	—	—	—	—	—	—	952	2480	—	757	10.52	5564	22.97	—
Dryas iulia	NA	—	—	566.959	21,916	53,067	10,683.8	34.64	950	1555	—	757	12.06	3725	44.17	—
Euploea tales	NA	—	—	539.115	32,552	49,376	10,918.6	33.40	951	1541	—	751	13.03	3669	45.16	—
Heliconius melpomene	NA	—	—	—	—	—	—	—	957	2507	—	756	6.45	5703	18.64	—
Junonia coenia	NA	—	—	—	—	—	—	—	883	2135	—	739	12.78	5012	31.28	—
Laparus doris	NA	—	—	404.741	46,843	31,556	12,826.1	35.41	948	1533	—	758	11.95	3690	45.02	—
Leremia accius	NA	—	—	—	—	—	—	—	905	2239	—	732	13.3	5066	30.09	—
Melitaea cinxia	NA	—	—	—	—	—	—	—	881	2169	—	712	21.81	4660	41.37	—
Pararge aegeria	NA	—	—	406.226	5576	118,444	3429.7	36.24	943	1483	—	754	14.46	3449	49.35	—
Phoebis sennae	NA	—	—	—	—	—	—	—	946	2241	—	754	8.55	5124	25.99	—
Polygonia calbum	NA	—	—	345.882	2829	177,472	1948.9	32.85	940	1512	—	741	19.71	3442	50.67	—
Median		67.642	60.269	409.954	643	1,010,689	294	35.43	957	1891	0.260	754	12.05	4515	31.6	—

Butterfly and moth outgroups are included along with a new low-coverage genome for *Choristoneura fumiferana*. All voucher specimens are deposited at the University of Montpellier in the Institut des Sciences de l'Evolution de Montpellier or at the Sperling lab of the University of Alberta.

Introduction

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

Two sequencing methods have risen to the forefront of phylogenomics: transcriptomics (Oakley et al. 2012; Misof et al. 2014; Garrison et al. 2016) and hybrid enrichment (Faircloth et al. 2012; Lemmon et al. 2012; Lemmon and Lemmon 2013), and a third, shotgun sequencing, has recently become attractive (Allen et al. 2017). Transcriptomics relies on sequencing of expressed RNAs, and no knowledge of targeted gene regions is required. However, the availability of fresh or properly stored tissues limits the number of taxa included in such phylogenetic studies (Lemmon and Lemmon 2013; McCormack et al. 2013). In contrast, hybrid enrichment uses DNA probes to hybridize and selectively capture targets from a genome, which requires prior knowledge of the desired targets (Lemmon and Lemmon 2013; McCormack et al. 2013). An advantage of hybrid enrichment techniques is the ease of using ethanol-preserved tissues, old DNA extractions, and in some cases, old museum specimens (e.g. Guschanski et al. 2013; Blaimer et al. 2016). This can greatly increase the number of taxa in a phylogenomic study. However, later studies mining the original data are limited to the conserved regions of the hybrid enrichment. The third sequencing method - shotgun sequencing - can readily provide similar amounts of genomic data as the two other methods (Staden 1979; Anderson 1981; Gardner et al. 1981; Fuentes-Pardo and Ruzzante 2017). This method breaks up template DNA sequences across the genome into many small fragments before sequencing them, which has been used for both high-level and low-divergence phylogenomic analyses (Harkins et al. 2016; Allen et al. 2017; Pouchon et al. 2018; Zhang et al. 2019). Three main approaches for reconstructing phylogenetic relationships from whole genome shotgun sequencing have recently been developed (Allen et al. 2015; Schwartz et al. 2015; Hughes and Teeling 2018; Pouchon et al. 2018; Zhang et al. 2019). The first involves a search for shared conserved sequences in different species without focus on coding sequences (Schwartz et al. 2015; Pouchon et al. 2018). Both Schwartz et al. (2015) and Pouchon et al. (2018) rely on selecting reads with high similarity with respect to reference contigs to create a *de novo* sequence (i.e. mapping methods). This method is more suitable for low divergence datasets, since mapping to more divergent datasets can result in difficulties when identifying homologous data (Schwartz et al 2015). The second approach is to extract sequences from *de novo* assemblies via a set of predefined orthologous gene clusters (Hughes and Teeling 2018; Zhang et al. 2019). This approach allows focusing on genes of interest while avoiding difficulties in orthology detection, but its use is confined to groups with suitable genomic resources that provide an adequate initial set of orthologous genes. However, orthologous datasets are not available for some groups. Therefore, to make better use of less suitable genomic resources, a third approach was developed by Allen et al. (2015). The advantage of this approach lies in the assembly of predefined targeted genes by selecting reads with an optimized BLAST search step (a standard all-to-all BLAST search would have been impractical due to the number of reads in shotgun sequencing). Extending the rationale of Allen et al. (2015), we used a custom-designed BLAST method to directly annotate *de novo* assemblies of highly fragmented genomes instead of selecting reads. Additionally, rather than using predefined orthologous genes to annotate *de novo* genomes (Allen et al. 2017), we used all genes available from the reference genome. Orthology detection was then

performed specifically on our dataset, which is likely to generate more specific data (and potentially a larger amount of data) than from a restricted focus on a predefined list of genes. This approach allows annotation of divergent and highly fragmented genomes, with the potential to resolve complex phylogenomic relationships and contribute to analyses like molecular dating.

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

Swallowtail butterflies (Papilionidae) represent a charismatic and well-known family of butterflies, with colorful wing patterns and extensive morphological diversity - such as wingspans ranging from 2-3 cm (the tiny dragontail butterflies, *Lamproptera*) to 20 cm (the world's largest butterflies, *Ornithoptera*). Their global distribution currently includes 32 genera comprising at least 550 described species (Collins and Morris 1985; Tyler et al. 1994; Scriber et al. 1995). Most species are found in tropical regions, where they reach their greatest species richness within the true swallowtails (*Papilio*, Wallace 1865; Condamine et al. 2012), while mountain-adapted apollo butterflies occur on temperate and cold climates (*Parnassius*, Condamine et al. 2018a). Papilionidae include model organisms that have contributed to fundamental studies in biogeography (Wallace 1865; Condamine et al. 2013), insect-plant interactions (Ehrlich and Raven 1964; Berenbaum and Feeny 2008), speciation (Dupuis and Sperling 2015, 2016), and other areas of evolution and ecology (Scriber et al. 1995; Kunte 2009; Condamine et al. 2012; Kunte et al. 2014). Although numerous studies have investigated the phylogeny of this group (Munroe 1961; Hancock 1983; Igarashi 1984; Miller 1987; Tyler et al. 1994; Caterino et al. 2001; Zakharov et al. 2004; Nazari et al. 2007; Simonsen et al. 2011; Condamine et al. 2012, 2018b), the phylogenetic backbone of Papilionidae has not been resolved, potentially constraining our understanding of global biogeographic processes like those affecting the divergence of key clades of swallowtail butterflies in the Southern Hemisphere (Condamine et al. 2013).

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

relied on Parnassiinae and Papilioninae being monophyletic sister groups (Condamine et al. 2012). As for other groups, the lack of resolution of phylogenetic relationships within the swallowtail butterflies with molecular and morphological data can be attributed to (i) evolutionary processes like ancient and rapid diversification of lineages (e.g. birds: Jarvis et al. 2014; Prum et al. 2015; Suh 2016) or ancient hybridization (e.g. living cats: Li et al. 2016), and/or (ii) methodological and sampling artifacts such as missing data, low taxon sampling, or long branch attraction (Nabhan and Sarkar 2012; Roure et al. 2013). Phylogenetic patterns that are not due to artifacts can be important signatures of patterns of diversification, revealing links to events that were responsible for the current diversity of butterflies.

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

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

Materials and Methods

Taxon Sampling

In order to be phylogenetically informative about the most ancient relationships, our taxon sampling incorporates all described genera in the family Papilionidae (32 genera *sensu* Scriber et al. 1995; Simonsen et al. 2011; Condamine et al. 2012, 2018b). We sampled 41 species representing all subfamilies and all genera of Papilionidae (**Table 1**). We also included four genomes in the analyses that were already available for swallowtail butterflies (*Papilio glaucus*, Cong et al. 2015a; *P. machaon*, Li et al. 2015; *P. polytes*, Nishikawa et al. 2015; *P. xuthus*, Li et al. 2015). In our taxon sampling, we also included *Papilio joanae* (from the USA), a species of the *machaon* group (Dupuis and Sperling 2015), which we compare to the available *P. machaon* (from China, Li et al. 2015) as a control for our approach. Based on the latest phylogenies of Papilionoidea (Heikkilä et al. 2012; Kawahara and Breinholt 2014; Breinholt et al. 2018), we selected 16 outgroups, of which 14 are families closely related to Papilionidae including: one Hesperiidae (*Lerema accius*, Cong et al. 2015b), one Pieridae (*Phoebis sennae*, Cong et al. 2016a), one Lycaenidae (*Calycopis cecrops*, Cong et al. 2016b), and 11 Nymphalidae (*Heliconius melpomene*, Davey et al. 2016; *Laparus doris*; *Eueides tales*; *Agraulis vanillae*; *Dryas iulia*; *Junonia coenia*; *Melitaea cinxia*, Ahola et al. 2014; *Polygonia c-album*, de la Paz Celorio-Mancera et al. 2013; *Bicyclus anynana*, Nowell et al. 2017; *Pararge aegeria*, Carter et al. 2013*; Danaus plexippus*, Zhan et al. 2011); in addition, two moth species in the families Bombycidae (*Bombyx mori*, Mita et al. 2004), and Tortricidae (*Choristoneura fumiferana*, *de-novo* sequencing) were used to root the phylogeny as these families are distant outgroups of the Papilionoidea (Wahlberg et al. 2013). The lepidopteran data was recovered from Lepbase (http://lepbase.org/). In total, the taxon sampling represents 61 taxa (45 ingroup and 16 outgroup species).

DNA Extractions, Library Preparation and Shotgun Sequencing

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

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

For library sequencing, we relied on the NextSeq® series of sequencing systems, which are fast, flexible, high-throughput desktop sequencers. They support a broad range of sequencing applications, with fast turnaround time and moderate output compared to the MiSeq and HiSeq platforms (generating up to 800 million reads pair-ended, 100-120 Gb of data in less than 30 hours). Since prior work

showed genome size of swallowtail butterflies to be about 300 Mb (Cong et al. 2015), we multiplexed between 11 and 15 butterfly samples per NextSeq run to give about 10 Gb DNA sequence per sample and obtain low-coverage whole genomes at a sequence depth of about 30x. We used the NextSeq 500/550 High Output v2 kit (300 cycles, 2 x 150 bp) for a total of four NextSeq sequencing runs. We also added several negative controls for each sequencing run, including sham DNA extractions and library preparations, to allow potential removal of reads belonging to laboratory contaminants from analyses and facilitate assemblies of genomes. The choice of Nextera and NextSeq technology is based on the need to generate numerous mid-size DNA fragments at an affordable cost (compared to HiSeq).

Assembly of Low-Coverage Whole Genomes

The full analytical pipeline is illustrated in **Fig. 1**, and the scripts necessary to reproduce the study are available in the Supplementary Material that accompanies this article, as well as at https://doi.org/10.5061/dryad.ff18q9d.

From reads to coding DNA sequences. Using NGS technology (Illumina© NextSeq, paired-end reads with an averaged insert size of 500 bp), we sequenced and assembled 41 new low-coverage whole genomes of Papilionidae (added to four genomes on GenBank). In addition, we sequenced and assembled a new low-coverage whole genome for *Choristoneura fumiferana*, and assembled five outgroup genomes from raw reads available on the Lepbase database (added to ten genomes on GenBank). For these 47 genomes, raw reads were cleaned using Trimmomatic 0.33 (Bolger et al. 2014) by removing low quality bases from their beginning (LEADING:3) and the end (TRAILING:3), by removing reads below 50 bp (MINLEN:50), and by evaluating read quality with a sliding window approach (SLIDINGWINDOW:4:15). Quality was measured for sliding windows of 4 base pairs and had to be greater than 15 on average. A plethora of methods now exists for *de novo* genome assembly (e.g. ALLPATHS-LG, Gnerre et al. 2011; SOAPdenovo, Luo et al. 2012; MaSuRCA, Zimin et al. 2013; Platanus, Kajitani et al. 2014). Here we assembled the genomes using SOAPdenovo-63mer 2.04 (Luo et al. 2012). Several kmer size values (between 27 and 39) were tested for ten genome assemblies, which lead to no substantial difference for the N50 of our assemblies (median of 96 bp of difference between the lowest and highest N50). Kmer size of 31 was selected for further analysis. Then, we closed gaps emerging during the scaffolding process with SOAPdenovo, using the abundant pair relationships of short reads with GapCloser 1.12 (Bolger et al. 2014) (**Fig. 1**). *Papilio* genomes have recently been successfully assembled using Platanus (Cong et al. 2015), a tool designed to handle highly heterozygous genomes. In fact, when heterozygosity is too elevated, some assemblers split homologous haplotypes into different contigs. We quantified the impact of heterozygosity on our assemblies with a BLAST (Basic Local Alignment Search Tool) search of our contigs against themselves (96% similarity or higher). We found that duplicated portions of the genomes (found in two or more contigs) amount to only about 1% of the genome on average (including repeated elements); this indicates that the level of heterozygosity did not cause abundant artifactual contig duplications in our assemblies. Nonetheless, to deal with potential alleles still present in separate contigs in our assembly (due to heterozygosity, for example), our annotation approach makes a consensus sequence for ambiguous sites (see below and consensus step in **Fig. 1**). Duplicated contigs could also be the result of recent real duplications but we opted for a more conservative approach since our focus is on the deeper phylogeny of the family.

To annotate the sequences of all genomes, we performed a BLAST search using all available proteins for *Papilio xuthus* (**Fig. 1**). We used the tblastn function to annotate nucleotide sequences with reference protein

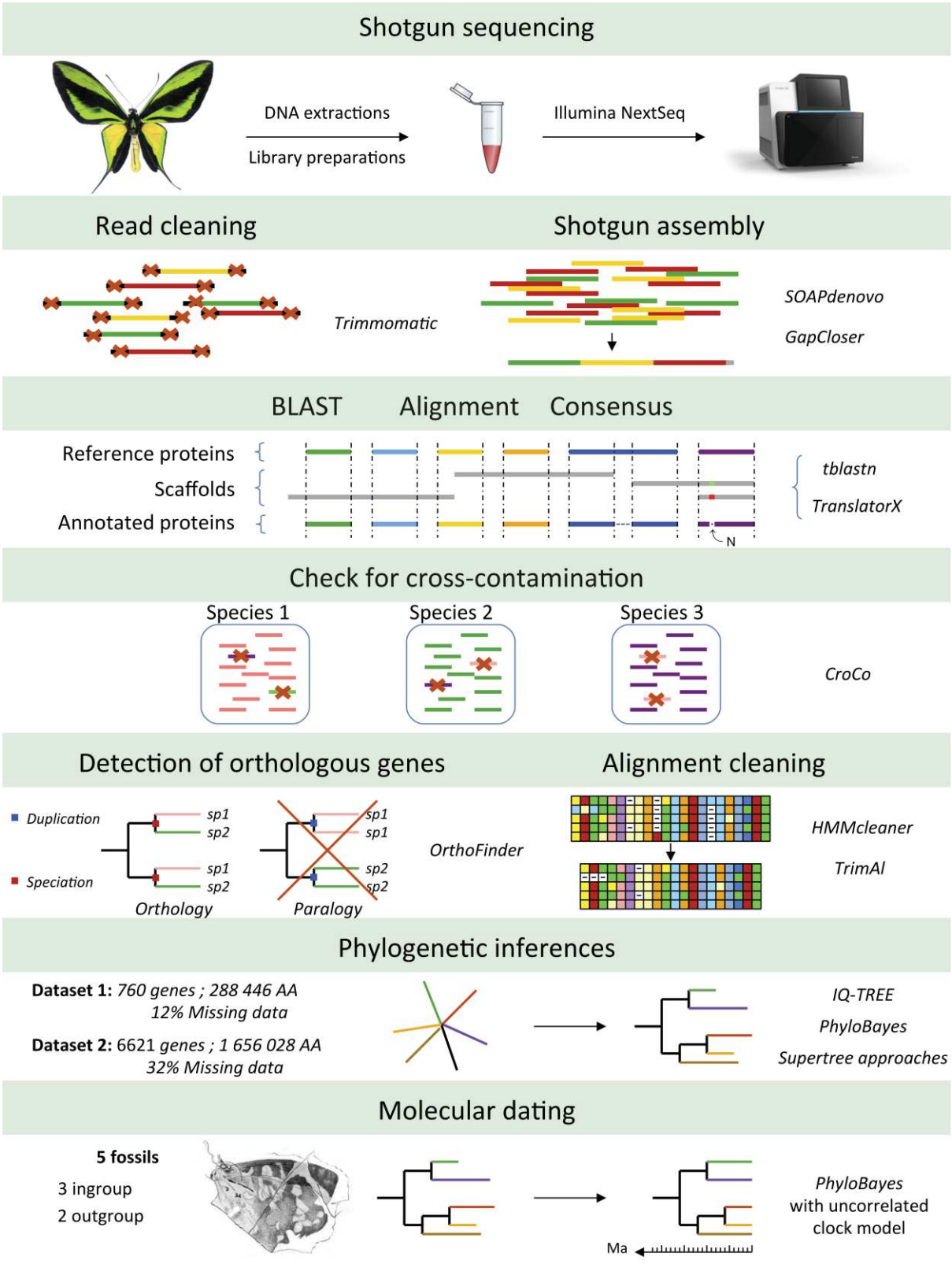


Figure 1. Conceptualization of the shotgun sequencing pipeline used to construct and analyze the *Dataset 1* (760 genes in amino acids), the *Dataset 2* (6,621 genes in amino acids), the *Dataset 3* (760 genes in nucleotides) and the *Dataset 4* (6,407 genes in nucleotides).

sequences of *Papilio xuthus* (Altschul et al. 2010). Only scaffolds with 60% or more similarity with the reference protein were selected. Several thresholds were tested for our dataset, and we retained 60% because this threshold provided the best trade-off between missing too many nucleotides versus including spurious nucleotides in the sequences. For example, for a threshold of 80% only highly-conserved regions (with less phylogenetic signal) were generally kept, while for a threshold of 40%, a larger proportion of presumably non-orthologous nucleotides were included. For each species, all scaffolds selected for a single coding DNA sequence (CDS) were aligned with *Papilio xuthus* with TranslatorX (Abascal et al. 2010) to generate a consensus (**Fig. 1**). This approach relies on amino acid translations to generate multiple alignments of nucleotides. All sites showing intraspecific variation were set to N, to conservatively avoid false informative sites. For example, recently duplicated genes could match (BLAST step) the same reference protein-coding gene. In this case, all divergent sites between the two copies of genes are replaced by N in the consensus, which avoids creating false informative sites due to a recent duplication event.

Check for cross-contaminations. Cross-contamination is a known but largely neglected issue (Ballenghien et al. 2017). Using shotgun sequencing, we were particularly exposed to the risk of cross-contamination since we multiplexed between 11 and 15 butterfly samples per sequencing run. Before creating the datasets (**Fig. 1**), we checked the cross-contamination level in our different sequencing runs using CroCo 0.1 (Simion et al. 2018), which was developed for identifying and removing cross contaminants from assembled transcriptomes. For any given focal species, CroCo identifies CDS that have significantly higher coverage (number of reads mapped to the CDS) in another species than the focal one, with each species of the dataset successively considered as focal. To measure relative coverage between two species, CroCo implements a metric, called Fragments per Kilobase Million (FPKM; Mortazavi et al. 2008), that is used to estimate relative coverage for each gene and is directly comparable between genes because the value is normalized by sequencing depth and size of each gene. Originally developed for transcriptomic data, this method can also be applied to CDS annotated in whole genome sequences. CroCo is thereby used to estimate relative coverage for each CDS of each species and to identify CDS that are suspiciously similar among species. CroCo was set to default parameters, i.e. the option -R to use the tool RapMap for mapping (Srivastava et al. 2016), with values between 0.2 and 300 for minimum and maximum coverage. Any contigs suspected of being contaminated were then discarded in subsequent analyses.

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

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

We used HMMCleaner 1.8 (Di Franco et al. 2019) to clean CDS alignments from misaligned sequences (gene by gene). This method cleans an alignment by first building a Hidden Markov Model profile of the alignment, and then measuring the score of the different sequence regions along this profile. After that, the sites present in at least two thirds of the sampled species were selected for the phylogenomic dataset. Finally, we performed a last cleaning step using trimAl 1.2rev59 (Capella-Gutiérrez et al. 2009), which is

designed to trim alignments for large-scale phylogenomic analyses. We adopted a stringent approach by selecting all CDS for each species that have at least 30% of sites overlapping with 75% of the rest of the sequences (-seqoverlap 30 and -resoverlap 0.75 options).

After these steps, we built two amino-acid phylogenomic datasets to test the impact of missing data (Roure et al. 2013). In *Dataset 1*, we kept all genes present in at least 95% of species. For the *Dataset 2*, we selected all genes present in at least four species. The two amino acid matrices concatenated hundreds (*Dataset 1*) or thousands (*Dataset 2*) of selected orthologous genes. In addition, since phylogenomic incongruences between amino-acid and nucleotide datasets have been observed (e.g. in spider flies, Gillung et al. 2018), we also created two nucleotide-based versions of *Dataset 1* and *2* (*Datasets 3* and *4*, respectively). Final alignments are available on Dryad (at http://dx.doi.org/10.5061/[NNNN], **Appendices S2, S3, S4, and S5**).

Phylogenomic Analyses with a Supermatrix Approach

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

For *Dataset 1*, a ML analysis with IQ-TREE was first performed using a single LG model for amino acids (Le and Gascuel 2008) including four matrices, each corresponding to one discrete gamma rate category (+Γ_4 option; Le et al. 2012), and empirical amino acid frequencies estimated from the data (+F option). Node supports were calculated with 100 non-parametric bootstrap (BS) replicates. To compare node supports, a second ML analysis with IQ-TREE was carried out under the same conditions but with 1,000 ultrafast bootstrap (UFBS) replicates (Minh et al. 2013; Hoang et al. 2018). BS values and UFBS values were considered strong when higher than 70% and 95%, respectively. These ML analyses assumed a single rate matrix for the whole dataset; however, rate heterogeneity is widespread in phylogenomic datasets (Yang 1996; Jia et al. 2014) and must be taken into account. IQ-TREE provides a number of site specific frequency models such as the posterior mean site frequency (PMSF) model as a rapid approximation to the time- and memory-consuming profile mixture models C10 to C60 (Le et al. 2008; a variant of the CAT model in PhyloBayes, Lartillot and Philippe 2004). PMSF is the amino-acid profile for each alignment site computed from an input mixture model and a guide tree, and the PMSF model is much faster and requires much less memory than C10 to C60 models (Wang et al. 2018), regardless of the number of mixture classes. Moreover, simulations and empirical phylogenomic data analyses have shown that PMSF models can be effective against long branch attraction artefacts (Wang et al. 2018). We performed IQ-TREE analyses with the C50 model as well as the PMSF model. The C50 analysis required 466 Gb of memory and more than five days to infer the ML tree, so we did not perform bootstrap analysis. However, we ran 1,000 UFBS replicates for the PMSF analysis. For all IQ-TREE analyses, we estimated the most likely tree with 100 separate ML searches, as well as 100 searches using the –t RANDOM option, which after initial model optimization on a parsimony tree uses 100 random tree topologies as starting trees for each search.

Bayesian phylogenetic reconstruction was conducted using PhyloBayes MPI (Lartillot et al. 2013) under the CAT+F81+Γ_4 mixture model (Lartillot and Philippe 2004). The CAT model allowed us to take into account the across-site heterogeneities in the amino-acid replacement process (Lartillot and Philippe 2004), and has proven to perform well on large molecular datasets (e.g. Chiari et al. 2012). PhyloBayes MPI has been run as follows: two independent Markov chains Monte Carlo (MCMC) analyses starting from a random tree were run until we generated at least 5,000 cycles

after convergence (maximum allowed 10,000 cycles), with trees and associated model parameters sampled every cycle. After checking for convergence in both likelihood and model parameters (*tracecomp* subprogram), the trees sampled in each MCMC run before reaching convergence were discarded as burn-in. The 50% majority-rule Bayesian consensus tree and associated posterior probabilities (PP) were then computed from the remaining trees (*bpcomp* subprogram). We consider node support with PP ≥ 0.95 to be robust.

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

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

Phylogenomic Analyses with a Supertree Approach

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

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

Estimation of Gene and Site Concordance Factors

As noted in the previous section, concatenation analyses can return fully-resolved and well-supported trees even when the level of gene incongruence is high (e.g., Jeffroy et al. 2006; Kumar et al. 2012). As recommended in Minh et al. (2018), we measured gene concordant (gCF) and site concordant (sCF) factors to complement

traditional bootstrap node-support measures for *Datasets 1* and *3* (760 loci). First, using the concatenation of all 760 loci, a reference tree was inferred with IQ-TREE with a search for substitution partition for each locus via ModelFinder (Kalyaanamoorthy et al. 2017). Second, we inferred a gene tree for each locus alignment using IQ-TREE with a model selection. Finally, gCF and sCF were calculated using the specific option -scf and -gcf in IQ-TREE (Minh et al. 2018).

Estimation of Divergence Times

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

Constraints on swallowtail clade ages were enforced by fossil calibrations with systematic position assessed using phylogenetic analyses (Condamine et al. 2018a). Four unambiguous and informative fossils belong to Papilionidae, two of which are Parnassiinae (Nazari et al. 2007). The first is †*Thaites ruminiana* (Scudder 1875), a compression fossil from limestone in the Niveau du gypse d'Aix Formation of France (Bouches-du-Rhône, Aix-en-Provence) within the Chattian (23.03–28.1 Ma) of the late Oligocene (Sohn et al. 2012). †*Thaites* was often recovered as sister to Parnassiini, and occasionally as sister to Luehdorfiini + Zerynthiini. Thus, we constrained the crown age of Parnassiinae with a uniform distribution bounded by a minimum age of 23.03 Ma. The second is †*Doritites bosniaskii* (Rebel 1898), an exoskeleton and compression fossil from Italy (Tuscany) from the Messinian (5.33–7.25 Ma, late Miocene; Sohn et al. 2012). †*Doritites* was reconstructed as sister to *Archon* (Luehdorfiini), in agreement with Carpenter (1992). The crown of Luehdorfiini was thus constrained for divergence time estimation using a uniform distribution bounded with 5.33 Ma. Third is the genus †*Praepapilio,* with two fossil species †*P. colorado* and †*P. gracilis* (Durden and Rose 1978) from the early Lutetian (Eocene) of the Green River Formation (Colorado, U.S.A.). This fossil was used to constrain the crown age of Papilionidae with a uniform distribution bounded by a minimum age of 47.8 Ma (Smith et al. 2003; de Jong 2007).

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

We were unable to use other fossil calibrations, although suitable butterfly fossils exist for other families (e.g. Wahlberg et al. 2009; Sohn et al. 2012), because the corresponding nodes to which the fossil calibrations could be assigned were not present

in our phylogeny. In particular, the families Lycaenidae and Riodinidae have few representatives. Moreover, four fossils have been used to date the phylogeny of Pieridae (Braby et al. 2006) but their identification and phylogenetic assignment is doubtful (de Jong 2007, 2016).

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

All PhyloBayes calculations were conducted by running three independent MCMC until we generated at least 5,000 cycles after convergence (maximum allowed 10,000 cycles), with sampling posterior rates and dates collected every cycle. After checking for convergence in both likelihood and model parameters (*tracecomp*), posterior estimates of divergence times were then retrieved from the sampled trees of each chain after the burn-in period to compute the Bayesian time-calibrated tree and associated 95% CI (*readdiv* subprogram). As recommended by Brown and Smith (2018), we compared prior and posterior distributions to determine whether signal is coming from the data or the prior.

Results

Low-Coverage Whole Genomes and Phylogenomic Datasets

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

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

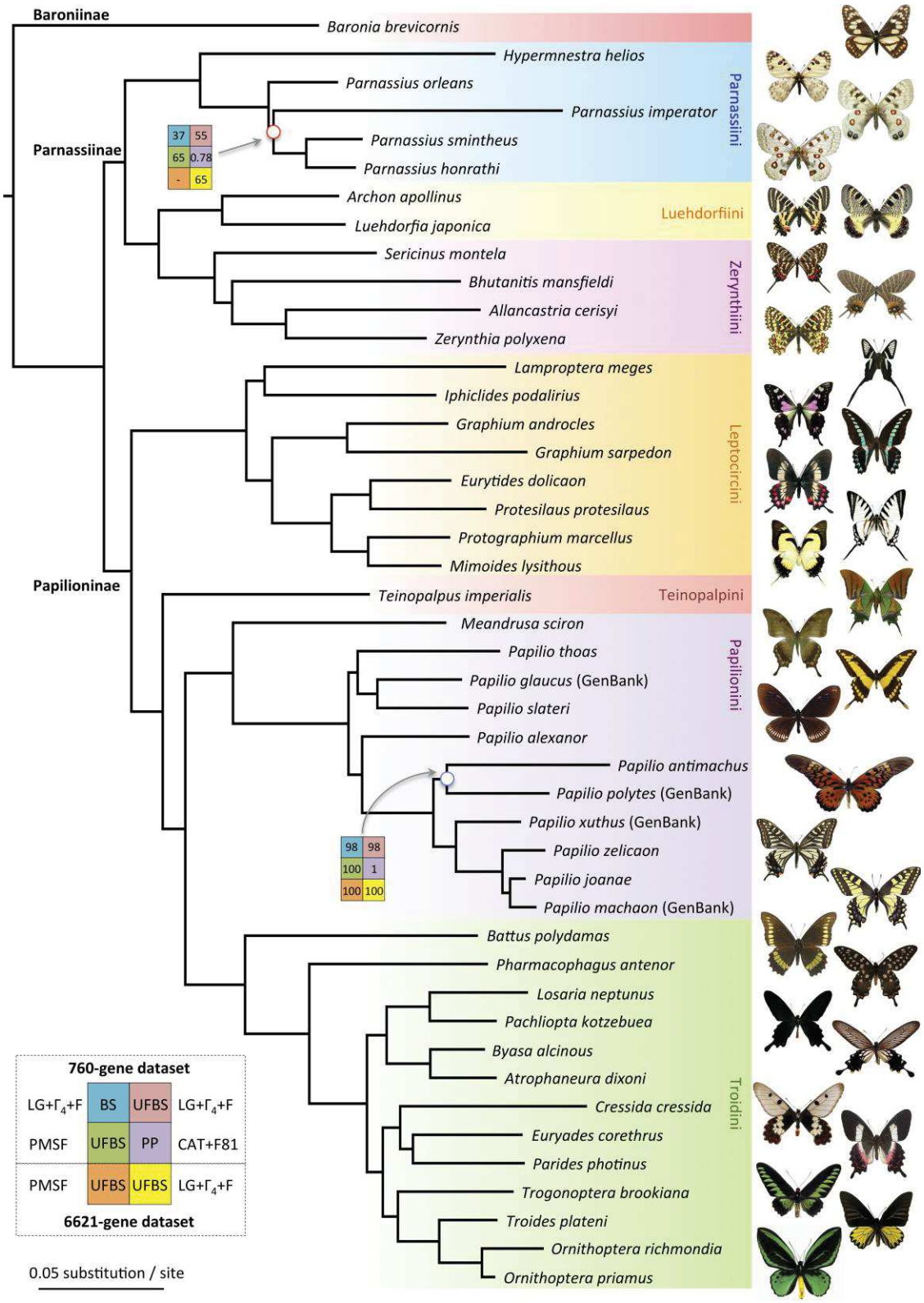


Figure 2 | Phylogenomic relationships of Papilionidae based on supermatrix analyses. All nodes have maximal BS, UFBS and PP support, except for two nodes with circles and support values in colored boxes, explained in the lower left corner legend. The topology reflects the results of all phylogenetic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

contaminations were removed for downstream analyses.

OrthoFinder was used to find 30,043 orthogroups, where an orthogroup is a set of genes originating by speciation of a gene present in the last common ancestor. Among these orthogroups, we selected those having only one copy per species. The selected groups were then filtered again, with the genes present in at least 95% of the species comprising *Dataset 1*, while the orthologous genes present in at least four species formed *Dataset 2*. These sets of genes were used to create both nucleotide-based and amino-acid-based matrices. In the smallest matrix, we obtained 760 genes, which represent 288,446 amino acids, 162,859 variable sites (56.5%), and 100,994 parsimony-informative sites (35%). We found an average of 96% of genes per species and a median of 12% missing data per species (gaps and undetermined sites in the supermatrix). In the largest matrix, we obtained 6,621 genes, which represent 1,656,028 amino acids, 1,020,365 variable sites (61.6%), and 608,399 parsimony-informative sites (36.7%). Here we found an average of 65% of genes per species and a median of 31.6% missing data per species.

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

Supermatrix Phylogenomics

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

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

For *Dataset 2* (6,621 CDS, 1,656,028 amino acids, 61 species), we performed only ML reconstructions with IQ-TREE and tested the effect of the protein model (LG+Γ_4+F vs PMSF). The ML analyses with the LG+Γ_4+F model (total CPU time = 1066h:37m, and memory = 57 Gb) yielded the same topology as obtained with the analyses of *Dataset 1*. ML analyses with the

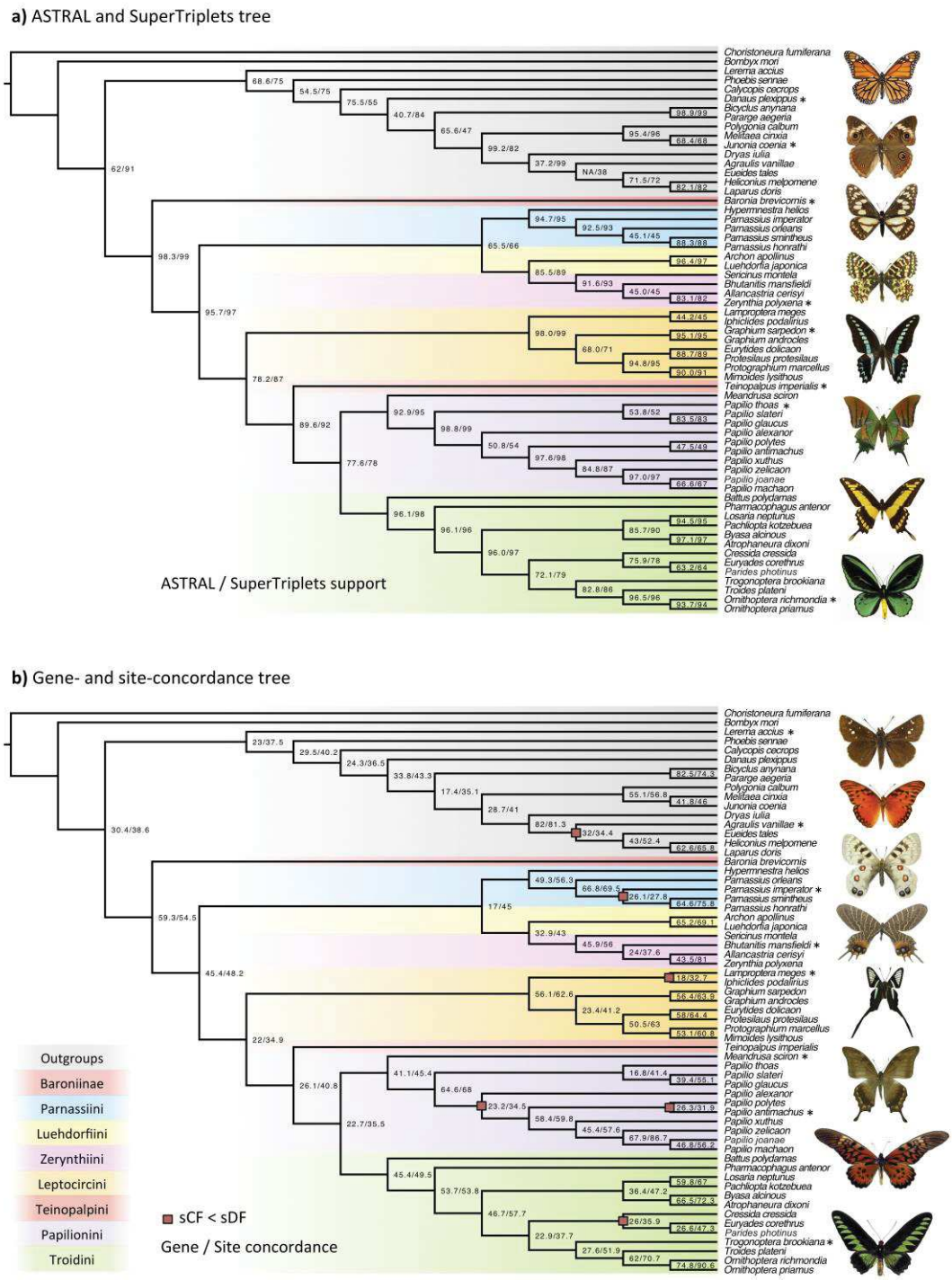


Figure 3 | Phylogenomic relationships of Papilionidae based on a) supertree analyses and b) gene and site concordance of supermatrix analyses. The supertree topology is inferred by ASTRAL and SuperTriplets with 6,621 genes and 5,367 rooted gene trees, respectively. For those analyses, nodes from source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

PMSF model (total CPU time = 4016h:00m, and memory = 70 Gb) provided a very similar topology, except for the branching of *Parnassius imperator*, which was retrieved as sister to *P. orleans* (**Appendices S6** and **S7**).

For *Dataset 3* (760 CDS, 889,191 nucleotides, 61 species), and *Dataset 4* (6,407 CDS, 5,267,461 nucleotides, 61 species), the best substitution model for each gene was selected with ModelFinder followed by ML analyses (total CPU time = 127h:26m, and memory = 15 Gb for *Dataset 3*, and total CPU time = 884h:33m, and memory = 76 Gb for *Dataset 4*). The ML analyses provided the same topology as the one obtained with *Datasets 1* and *2*, except for the relationships of *Iphiclides* and *Lamproptera* and the relationships of *Papilio antimachus* and *Papilio polytes*, which were not recovered as sister taxa in the nucleotide-based analyses (**Appendices S6** and **S7**).

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

Supertree Phylogenomics

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

Gene and Site Concordance Factors

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

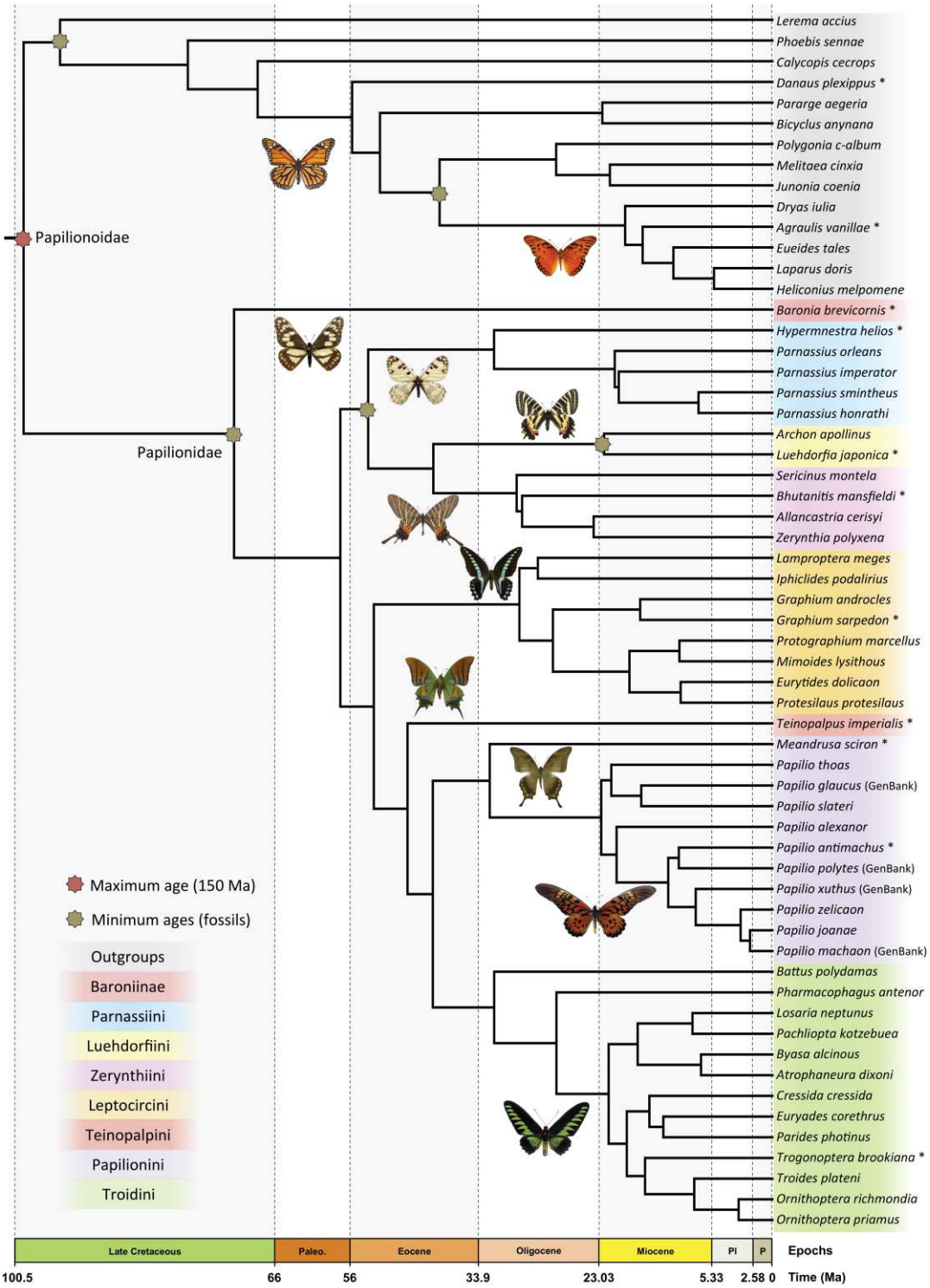


Figure 4 | Bayesian time-calibrated phylogeny of butterflies. The dated tree was obtained with PhyloBayes analyses of *Dataset 1* (excluding *Bombyx mori* and *Choristoneura fumiferana*) using the CAT-GTR model, a birth-death model, and an uncorrelated clock model constrained with five fossil calibrations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

Molecular Dating

Bayesian analyses of divergence times performed with the CAT-GTR model in PhyloBayes reached convergence between 1,500-2,000 cycles (total CPU time [1 thread per chain; 3 chains] = between 6 and 8 months). For a conservative estimate of posterior node ages, 1,500 cycles were discarded as burn-in (**Appendices S9** and **S10** available on Dryad). Dating analysis results for swallowtails and outgroups are shown in **Fig. 4**. The crown group of butterflies (Papilionoidea) began diversifying in the Late Cretaceous at 99.2 Ma (95% CI: 68.6-142.7 Ma), and swallowtails (Papilionidae) originated in the end of the Late Cretaceous at 71.4 Ma (95% CI: 49.8-103.6 Ma). Subfamilies Papilioninae and Parnassiinae began to diversify at 52.9 Ma (95% CI: 36.7-77.4 Ma) and at 53.6 Ma (95% CI: 36.9-79.2 Ma), respectively. We recovered early Oligocene to mid-Miocene origins for the species-rich genera: *Papilio* at 22.8 Ma (95% CI: 14.9-34.6 Ma), *Graphium* at 17.5 Ma (95% CI: 9.9-28.7 Ma), and *Parnassius* at 21.2 Ma (95% CI: 12.4-35.2 Ma). Comparison of the prior (uniform) distributions and the posterior (normal) distributions of node ages indicates that the priors did not influence the posteriors (**Appendix S11**).

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

Cross-Contamination Issues

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

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

Sampling	Papilionoidea		Papilionidae		Parnassiinae		Parnassiini		*Parnassius*		Luehdorfiini		Zerynthiini	
	Median age	95% CI	Median age	95% CI	Median age	95% CI	Median age	95% CI	Median age	95% CI	Median age	95% CI	Median age	95% CI
59 species	99.2	68.6–142.7	71.4	49.8–103.6	53.6	36.9–79.2	36.9	22.8–57.9	21.2	12.4–35.2	22.4	9.9–39.9	34.0	21.3–52.1
58 species	100.4	70.6–142.5	72.0	50.4–103.5	53.7	37.0–77.2	35.3	21.1–55.0	14.5	8.0–25.4	22.5	11.0–37.4	34.6	21.7–52.9
45 species	NA	NA	71.9	43.7–139.8	58.1	31.7–115.6	40.2	19.9–83.4	23.3	11.2–49.6	26.0	10.6–56.3	36.8	18.3–74.8
44 species	NA	NA	71.7	43.9–139.9	56.5	31.2–111.1	37.0	17.8–76.8	15.8	6.9–34.7	24.6	10.8–52.8	36.1	18.4–73.1
Sampling	Papilioninae		Leptocircini		*Graphium*		Teinopalpini (stem)		Papilionini		*Papilio*		Troidini	
	Median age	95% CI	Median age	95% CI	Median age	95% CI	Median age	95% CI	Median age	95% CI	Median age	95% CI	Median age	95% CI
59 species	52.9	36.7–77.4	33.4	21.1–50.9	17.5	9.9–28.7	48.4	33.6–71.3	37.5	25.3–56.3	22.8	14.9–34.6	37.0	25.2–55.1
58 species	52.6	36.7–75.3	33.2	21.7–49.6	17.5	10.2–27.9	48.3	33.5–69.5	37.4	25.3–54.8	22.4	14.7–33.4	36.8	25.3–53.4
45 species	57.8	31.7–114.9	36.7	18.6–75.1	19.1	8.6–40.4	53.0	29.0–105.6	40.1	21.8–82.2	24.9	13.1–50.7	40.6	21.8–81.4
44 species	55.6	31.0–109.6	35.4	18.7–72.0	18.4	8.6–38.4	51.0	28.4–100.9	39.5	21.6–78.4	23.8	12.6–47.9	39.0	21.2–77.8

Discussion

Using Shotgun Sequencing for Phylogenomics

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

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

Here, we meet the challenge of phylogenomic reconstruction by orthologous CDS identification from contigs obtained with whole-genome shotgun sequencing. The method is designed for highly fragmented and low-coverage genomes and requires the availability of a single (related) reference genome. Despite the low-coverage nature of our data, we were able to cost-effectively identify more than 10,000 CDS for 41 newly sequenced species (plus *Choristoneura fumiferana*). Applying a rigorous cleaning procedure, we extracted 6,621 orthologous genes and assembled four genomic datasets including 100,994 (35%) and 608,399 (36.7%) informative amino-acid sites for *Datasets 1* and *2*, respectively; and 449,010 (50.5%) and 2,581,850 (49%) informative nucleotide sites for *Datasets 3* and *4*, respectively. This amount of informative data for phylogenomic analyses is comparable to sequence-capture datasets like UCEs (854 UCE loci for stinging wasps including 143,608 [70.7%] informative nucleotide sites, Branstetter et al. 2017), which now constitute the most widely used approach in phylogenomics (McCormack et al. 2013). Our BLAST-based annotation and orthologous detection was validated because two closely-related species of the *machaon* group were consistently found as sister lineages and had short branch lengths. In addition, both ML and Bayesian phylogenies agreed with several established studies (Simonsen et al. 2011; Condamine et al. 2012) and uncover new relationships (see below). Remarkably, even with poor-quality libraries (*Allancastria cerisyi, Hypermnestra helios* and *Parnassius imperator*), our approach correctly places these species in the same position as in a fully sampled tree of Parnassiinae (Condamine et al. 2018a), although with low support for *Parnassius imperator*.

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

The Importance of Controlling for Cross Contamination

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

Using Shotgun Sequencing for Dating

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

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

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

Computational Limitations for Phylogenomics

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

Confirming and Uncovering Phylogenomic Relationships within Papilionidae

Using shotgun sequencing of whole genomes, we have provided genomic data for all genera of Papilionidae, a dataset that is potentially useful for more diverse evolutionary questions than those normally encompassed by a family tree. Despite the fragmented nature of the genomes, we obtained a resolved and strongly supported phylogeny displaying the relationships of all extant swallowtail genera. The tree is noteworthy for its node support, with only one node not supported, and that being partly due to the poor quality library of the species *Parnassius imperator*. Supertree methods (SuperTriplets and ASTRAL) gave the same topology as supermatrix methods, indicating that this topology is robust (**Fig. 3**), which is also confirmed by gene concordance factors (**Appendix S8**). All phylogenomic analyses showed that *Baronia* is sister to all remaining Papilionidae with maximal node support in Bayesian and ML analyses (**Fig. 2**, **Appendix S6**). In previous studies *Baronia* has not always been recovered as sister to other Papilionidae, but our result benefits from the largest molecular dataset ever assembled for swallowtail genera and also agrees with the latest Sanger-based phylogenies (Simonsen et al. 2011; Condamine et al. 2012) and a mitogenomic study (Condamine et al. 2018b).

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

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

factors (**Figs. 2 and 3**, **Appendices S6** and **S13**), in agreement with previous studies (e.g. Simonsen et al. 2011; Condamine et al. 2012, 2018b). It is possible that the non-monophyly of Papilioninae in Espeland et al. (2018) arose from their limited taxon sampling in Papilionidae. Indeed, Leptocircini contain 140 species and seven genera, and Parnassiinae comprise 85 species and eight genera. We sampled all genera while Espeland et al. (2018) sampled only two genera for Leptocircini and three genera for Parnassiinae. The lack of key genera that diverged early in Leptocircini (*Iphiclides* and *Lamproptera*) or Parnassiinae (*Hypermnestra* and *Sericinus*) may have led to the apparent non-monophyly of Papilioninae based on exon-capture data. Alternatively, it is possible that our analyses recovered the monophyly of Papilioninae because our datasets rely on two- (*Dataset 1*) and eighteen-fold (*Dataset 2*) more genes than Espeland et al. (2018). Also, previous studies relying on few genes always recovered Papilioninae as monophyletic (e.g. Simonsen et al. 2011; Condamine et al. 2012), and the same is true for studies with morphological characters (e.g. Munroe 1961; Hancock 1983; Miller 1987; Parsons 1996). This suggests that dense taxon sampling is essential to phylogenomic tree reconstruction, since insufficient sampling may lead to highly supported clade relationships that are wrong.

Systematic debates have surrounded the phylogenetic positions of enigmatic genera like *Meandrusa* and *Teinopalpus*, *Cressida* and *Euryades*, or *Iphiclides* and *Lamproptera* (Ford 1944; Hancock 1983; Miller 1987; Tyler et al. 1994; Parsons 1996; Simonsen et al. 2011; Condamine et al. 2012). In the first case, we found strong support for *Teinopalpus* as the sister group of Troidini and Papilionini, with *Meandrusa* as the sister group of *Papilio* and both together forming the tribe Papilionini. This result was suggested by a mitogenome study (Condamine et al. 2018b), but not recovered with Sanger-based phylogenies (Simonsen et al. 2011; Condamine et al. 2012). In the second case, *Cressida* was recovered as sister to *Parides* and *Euryades* with supermatrix and supertree analyses but not with a mitogenomic study, which showed *Cressida* as sister to *Euryades* (Condamine et al. 2018b). This latter result seems unlikely given that *Cressida* is an Australasian genus, while *Euryades* and *Parides* are both Neotropical, and the divergence of these three lineages dates back to the early-middle Miocene (**Fig. 4**). This combined with the fact that both low node supports are obtained with supertree approaches and site concordance factor is lower than site discordance factors may indicate effects of ILS or hybridization in these parts of the tree, or the effect of model misspecification when reconstructing gene trees or even hidden paralogy. For the third case, *Iphiclides* is found as sister to *Lamproptera* with amino-acid-based phylogenomic analyses, with both being sister to all Leptocircini (**Fig. 2**), but we found *Lamproptera* as sister to all Leptocircini in nucleotide-based phylogenomic analyses (**Appendix S6**). Gene-tree analyses provide insights into this supermatrix-driven discrepancy with supertree methods showing low node support for the sister relationship (**Fig. 3**), and site concordance factors are lower than site discordance factors despite gene concordance factors being higher than gene discordance factors (although all factors for these branches have low values, **Appendix S8**), which means that this relationship remains unclear even using the information provided by this large genomic dataset. Our study demonstrates the need for more specific studies to clarify the phylogeny of Leptocircini, which represents a phylogenetic impediment within Papilionidae. Interestingly, two similar topological issues within the genus *Papilio* are revealed for the placement of *P. alexanor*, and for the relationship between *P. antimachus* and *P. polytes,* but may be an artefact of low taxon sampling (10 out of 200 species are sampled in *Papilio*; Nabhan and Sarkar 2012). Further work with more comprehensive taxon sampling is needed to identify the causes of these low supports and is beyond the scope of this study. Such a comprehensive topology will have important evolutionary implications in terms of

trait evolution like host-plant associations or historical biogeography.

Cretaceous Origin of Papilionoidea and Paleogene Diversification of Papilionidae

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

Within butterflies, most extant lineages diverged after the K-Pg boundary (**Fig. 4**, **Appendix S9**), suggesting that this event had a major impact on the evolutionary history of butterflies, with lineages possibly going extinct (Wahlberg et al. 2009; Heikkilä et al. 2012). We infer that the most recent common ancestor of the Papilionidae lived in the Late Cretaceous ca. 71.4 Ma, but the divergence of ancestors of all other extant lineages lagged 10 million years behind the end-Cretaceous catastrophe (**Fig. 4**), and likely survived in Northern Hemisphere regions (Condamine et al. 2012, 2013). Such a pattern of diversification suggests clade extinctions at the K-Pg boundary and subsequent diversification of extant clades in the Cenozoic (52.9 Ma for Papilioninae and 53.6 Ma for Parnassiinae, **Fig. 4, Table 2**). Subsequent diversification within the two subfamilies occurred in the Eocene, with almost all lineages leading to currently recognized tribes originating in the early Oligocene at 33.5 Ma on average (ranging from 37.5 Ma for Papilionini to 22.4 Ma for Luehdorfiini) and most genera diverging from sister genera in the Miocene (**Fig. 4, Table 2**). This diversification pattern is similar to that shown in Nymphalidae (Wahlberg et al. 2009), Riodinidae (Espeland et al. 2015) and Hesperiidae (Sahoo et al. 2017), suggesting that common drivers or causes have shaped butterfly diversification dynamics through time.

Conclusion

The utility of whole genomes for building and dating phylogenies has never been more auspicious than today. The successful development of powerful analytical tools, in conjunction with the rapid and massive increase in the availability of genomic data (Fuentes-Pardo and Ruzzante 2017), allows us to resolve and understand evolutionary histories that are more and more complex. We still face important limitations in data accessibility (too few genomes are available) and methodological shortcomings (orthology assessment, running time). However, our approach (and analytical pipeline) has empowered the use of low-coverage and highly fragmented whole genomes, providing productive perspectives for future investigations of other model groups. Applied to an insect radiation, we were able to produce a much-needed stable backbone for a revised classification of swallowtail butterflies through a fully resolved phylogenomic framework unveiling novel relationships and confirming previous hypotheses. The resulting time-calibrated tree also permits a much better understanding of the major events of Papilionidae diversification for interpreting future comparative studies ranging from ecology to genome evolution.

Acknowledgements

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

Supplementary Material

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

Funding

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

⇧Back to summary⇧

Supplementary Information

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

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

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

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

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

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

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

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

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

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

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

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

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

⇧Back to summary⇧

References

⇧Back to summary⇧

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Gnerre S., Maccallum I., Przybylski D., Ribeiro F.J., Burton J.N., Walker B.J., Sharpe T., Hall G., Shea TP., Sykes S., Berlin A.M., Aird D., Costello M., Daza R., Williams L., Nicol R., Gnirke A., Nusbaum C., Lander E.S., Jaffe D.B. 2011. High-quality draft assemblies of mammalian genomes from massively parallel sequence data. Proc. Natl Acad. Sci. USA 108:1513-1518.

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

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

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

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

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

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

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

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

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

Warnow T., Jun W., Gilbert M.T., Zhang G. 2014. Whole-genome analyses resolve early branches in the tree of life of modern birds. Science 346:1320–1331.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Nazari V., Zakharov E.V., Sperling F.A.H. 2007. Phylogeny, historical biogeography, and taxonomic ranking of Parnassiinae (Lepidoptera: Papilionidae) based on morphology and seven genes. Mol. Phylogenet. Evol. 42:131–156.

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

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

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

Nowell R.W., Elsworth B., Oostra V., Zwaan B.J., Wheat C.W., Saastamoinen M., Saccheri I.J., van't Hof A.E., Wasik B.R., Connahs H., Aslam M.L., Kumar S., Challis R.J., Monteiro A., Brakefield P.M., Blaxter M. 2017. A high-coverage draft genome of the mycalesine butterfly *Bicyclus anynana*. GigaScience 6:1-7.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Simonsen T.J., Zakharov E.V., Djernaes M., Cotton A.M., Vane-Wright R.I., Sperling F.A.H. 2011. Phylogenetics and divergence times of Papilioninae (Lepidoptera) with special reference to the enigmatic genera *Teinopalpus* and *Meandrusa*. Cladistics 27:113–137.

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

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

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

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

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

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

Talla V., Suh A., Kalsoom F., Dincă V., Vila R., Friberg M., Wiklund C., Backström N. 2017. Rapid increase in genome size as a consequence of transposable element hyperactivity in wood-white (*Leptidea*) butterflies. Genome Biol. Evol. 9:2491-2505.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

⇧Back to summary⇧

Appendix 2 – Genome-wide macroevolutionary signatures of key innovations in butterflies colonizing new host plants

The preprint associated with this section is available on BioRxiv:
https://doi.org/10.1101/2020.07.08.193086
As well as the supplementay material:
https://doi.org/10.6084/m9.figshare.12278402.v1

⇧Back to summary⇧

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

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

[1]*Institut des Sciences de l'Evolution de Montpellier (Université de Montpellier | CNRS | IRD | EPHE), Place Eugène Bataillon, 34095 Montpellier, France.*
[2]*Institut für Botanik, Technische Universität Dresden, Zellescher Weg 20b, 01062, Dresden, Germany.*
[3]*Department of Bioscience, Durham University, Stockton Rd, Durham DH1 3LE, UK.*
[4]*Royal Botanic Gardens, Kew, TW9 3AB, Surrey, UK.*
[5]*86/2 Moo 5, Tambon Nong Kwai, Hang Dong, Chiang Mai, Thailand.*
[6]*CBGP, INRAE, CIRAD, IRD, Montpellier SupAgro, Univ. Montpellier, Montpellier, France.*
[7]*University of Alberta, Department of Biological Sciences, Edmonton T6G 2E9, AB, Canada.*

***Correspondence:** rem.allio@umontpellier.fr; fabien.condamine@gmail.com

Abstract

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

Introduction

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

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

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

These recent results have illustrated the need for a multidisciplinary approach to studying the evolution of insect-plant interactions within a macroevolutionary and genomic framework. However, a major knowledge gap lies in our understanding of the evolutionary links and drivers of host-plant shifts, genome-wide signatures of adaptations, and processes of species diversification[28]. As noted by Hembry and Weber[29], this implies that the questions of if, when, and how coevolution has an impact on macroevolutionary dynamics remain open challenges. Here we address this gap with an emblematic group that was instrumental in Ehrlich & Raven's model - the swallowtail

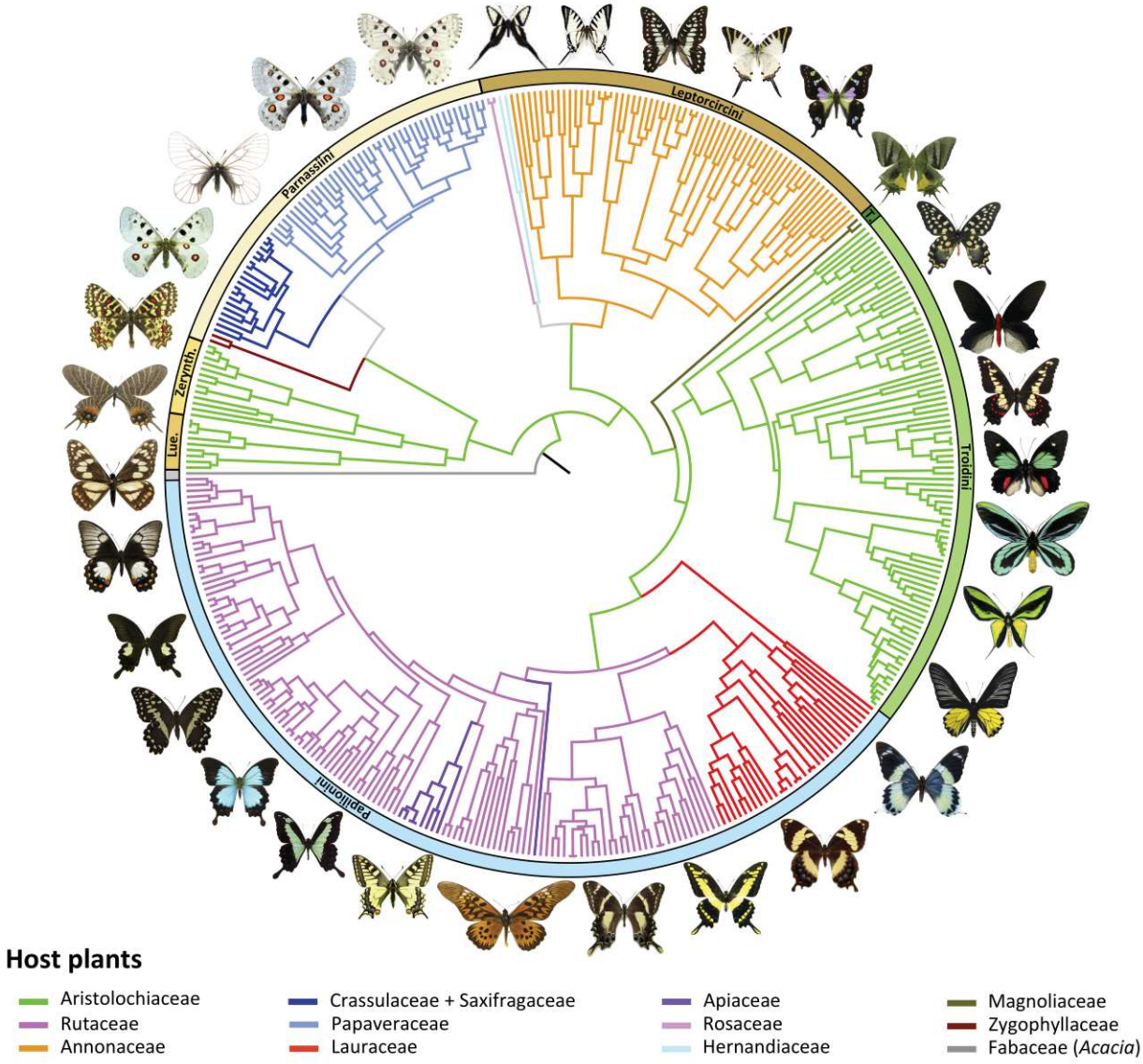


Figure 1 | Evolution of host-plant association through time shows strong host-plant conservatism across swallowtail butterflies. Phylogenetic relationships of swallowtail butterflies, with coloured branches mapping the evolution of host-plant association, as inferred by a maximum-likelihood model (**Supplementary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

butterflies (Lepidoptera: Papilionidae). Swallowtail caterpillars feed on a range of different flowering families[30] but a third of all species, including the tribes Zerynthiini (Parnassiinae), Luehdorfiini (Parnassiinae) and Troidini (Papilioninae), feeds exclusively on the birthwort family (Aristolochiaceae), which is one of the most toxic plant groups[31]. The Aristolochiaceae notoriously contain toxic aristolochic acids, which are known to be carcinogenic to many organisms, and Papilionidae are among the few that can feed on these plants[32,33]. By eating these toxic plants, the caterpillars sequester aristolochic acids that render both the caterpillars and the adults unpalatable for predators[31]. Interestingly, previous phylogenetic estimations of ancestral states indicated either that Aristolochiaceae was the ancestral host plant of Papilionidae[34] or that Aristolochiaceae was colonized twice[35], suggesting that the host-plant shifts have ancient origins and seem to be highly constrained as shown by the high level of host conservatism. Moreover, the arms race between Papilionidae and their host plants has been demonstrated at the molecular level with the evolution of a cytochrome P450 gene that plays a role in the detoxification of secondary plant compounds[36]. Some mutations can bypass the toxic defences of certain plants, providing survival and diversification on certain plants (and not others). Further studies have shown how changes in the use of host plants are associated with changes in the sequence, structure and function of P450. Results provide evidence that new P450 copies can appear for herbivores that colonize new hosts, supporting the hypothesis that interaction between herbivores and their host plants contributed to P450-gene diversification[37].

These studies provide convincing examples of host-plant shifts that may result in increased net diversification rate[18,34] and specific changes in key genes that confer new abilities to feed on toxic plants[36–38].

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

Results and Discussion

Co-phylogenetic history of an insect-plant antagonistic interaction.

First, we created an extensive phylogenetic dataset including seven genetic markers for 71% of swallowtail species diversity (408 of ~570 described species, *Methods*). This dataset leads to the assembly of the most complete and well-resolved dated phylogeny of swallowtail butterflies (79% of nodes with strong bootstrap support defined as greater or equal to 95%; **Supplementary Figs. 1-3**). Both tribe- and genus-level relationships are mostly consistent with previous results using multilocus datasets[18,34,42–46]. However, our species tree benefits from a phylogenomic backbone that we recently inferred at the genus level for the Papilionidae using genome scale data[47]. Second, we compiled host-plant preferences for each swallowtail species in the dataset, and we performed ancestral state estimations (Methods). Phylogenetic estimates of ancestral host-plant preferences indicate that Aristolochiaceae were either the food plant of ancestral Papilionidae[34] or were colonized twice[35], suggesting an ancient and highly conserved association with Aristolochiaceae throughout swallowtail butterflies evolution. Using this robust time-calibrated phylogeny (**Supplementary Figs. 1-3**), we have traced the evolutionary history of food-plant use and infer that the family Aristolochiaceae was the ancestral host for

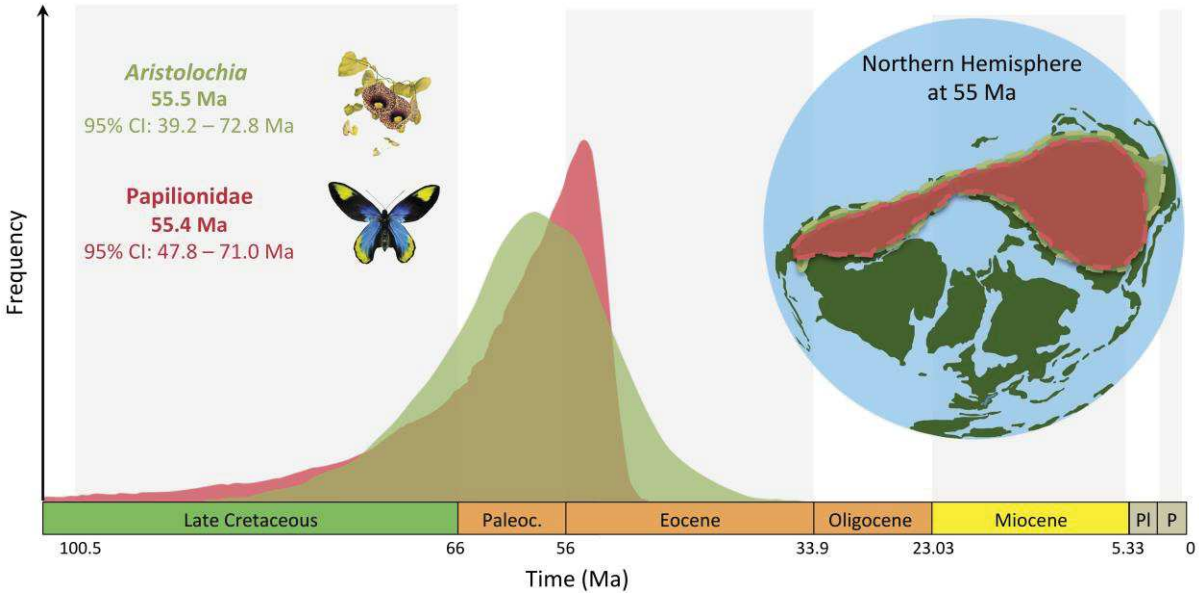


Figure 2 | Synchronous temporal and geographic origin for swallowtails and birthworts. Bayesian molecular divergence times with exponential priors estimate an early Eocene origin (~55 Ma) for both swallowtails and *Aristolochia* (alternatively, analyses with uniform prior estimated an origin around 67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

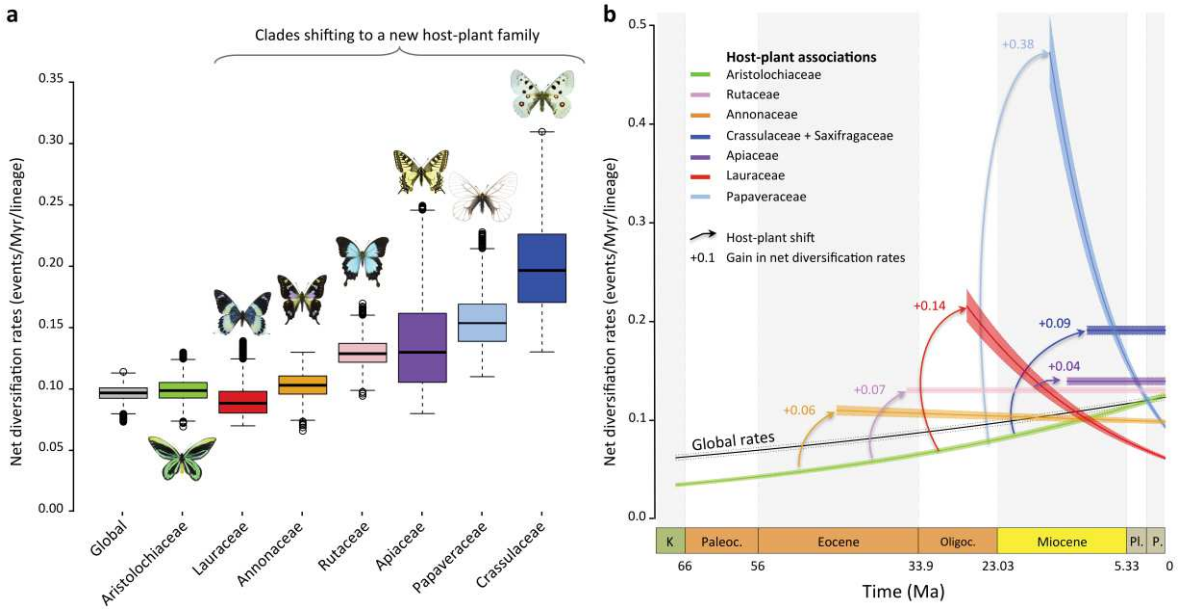


Figure 3 | Host-plant shifts lead to repeated bursts in diversification rates and a sustained overall increase in diversification through time. a, Diversification tends to be higher for clades shifting to new host plants, as estimated by trait-dependent diversification models. Boxplots represent Bayesian estimates of net diversification rates for clades feeding on particular host plants (see also **Supplementary Fig. 12**). **b,** A global increase in diversification is recovered with birth-death models estimating time-dependent diversification (see also **Supplementary Figs. 14, 15**). Taking into account rate heterogeneity by estimating

Papilionidae (**Fig. 1**; relative probabilities = 0.915, 0.789, and 0.787 with three models, **Supplementary Figs. 4, 5**). We further show that the genus *Aristolochia* was the ancestral host-plant, as almost all Aristolochiaceae-associated swallowtails feed on *Aristolochia* (**Supplementary Fig. 6**). Across the swallowtail phylogeny, we recover only 14 host-plant shifts at the plant family level (14 nodes out of 407; **Supplementary Figs. 4, 5**), suggesting strong evolutionary host-plant conservatism.

With the ancestor of swallowtails feeding on birthworts, evidence for synchronous temporal and geographical origins further links the genus *Aristolochia* and the family Papilionidae and supports the escape-and-radiate model. Reconstructions of co-phylogenetic history for other insect-plant antagonistic interactions have shown either synchronous diversification[11] or herbivore diversification lagging behind that of their host plants[10,48]. We assembled a molecular dataset for ~49% of the species diversity of Aristolochiaceae (247 of ~502 described species; *Methods*) and reconstructed their phylogeny (**Supplementary Fig. 7**), which is in agreement with previous works[49–53]. Divergence time estimates strongly suggest synchronous radiations of Papilionidae (55.4 million years ago [Ma], 95% credibility intervals: 47.8-71.0 Ma) and *Aristolochia* (55.5 Ma, 95% credibility intervals: 39.2-72.8 Ma) since the early Eocene (**Fig. 2**; **Supplementary Figs. 3, 8, 9**). This result is robust to known biases in inferring divergence times, with slightly older ages inferred for both groups when using more conservative priors on clade ages (**Supplementary Fig. 9**). Such temporal congruence between *Aristolochia* and Papilionidae raises the question of whether both clades had similar geographical origins and dispersal routes. To characterize the macroevolutionary patterns of the *Aristolochia*/Papilionidae arms race in space, we assembled two datasets of current geographic distributions for all species included in the phylogenies of both Aristolochiaceae and Papilionidae. We reconstructed the historical biogeography of both groups, taking into account palaeogeographical events throughout the Cenozoic (*Methods*). Along with the known fossil record of both groups[54–58], these results suggest that both Papilionidae and *Aristolochia* were ancestrally co-distributed throughout a region including West Nearctic, East Palearctic, and Central America in the early Eocene, when Asia and North America were connected by the Bering land bridge (**Fig. 2**, **Supplementary Figs. 10, 11**). This combination of close temporal and spatial congruence provides strong evidence that Papilionidae and *Aristolochia* diversified concurrently through time and space until several swallowtail lineages shifted to new host-plant families in the middle Eocene.

Host-plant shifts confer higher rates of diversification.

Our ancestral state estimates and biogeographic analyses are consistent with a sustained arms race between *Aristolochia* and Papilionidae in the past 55 million years. According to the escape-and-radiate model, a host-plant shift should confer higher rates of species diversification for herbivores through the acquisition of novel resources to radiate into[5,6] and/or the lack of competitors (Aristolochiaceae-feeder swallowtails have almost no competitors[31]). We tested the hypothesis that increases of diversification rates occurred in swallowtail lineages that shifted to new host-plants. Given the uncertainty surrounding the inferences of macroevolutionary rates from phylogenies of extant species, we applied a suite of birth-death models to cross-validate the estimated rates of diversification (LASER, MuSSE, RPANDA, BAMM, CoMET, and RevBayes; see Methods). We find evidence for (1) increases of diversification at host-plant shifts with trait-dependent birth-death models (as inferred with: MuSSE, **Fig. 3a**, **Supplementary Fig. 12**; RPANDA, **Supplementary Fig. 13**; and LASER, **Supplementary Table 1**), and (2) host-plant shifts contributing to a global increase through time with clade- and time-dependent birth-death models (as inferred with: RPANDA, **Fig. 3b**, **Supplementary Fig. 13**; BAMM,

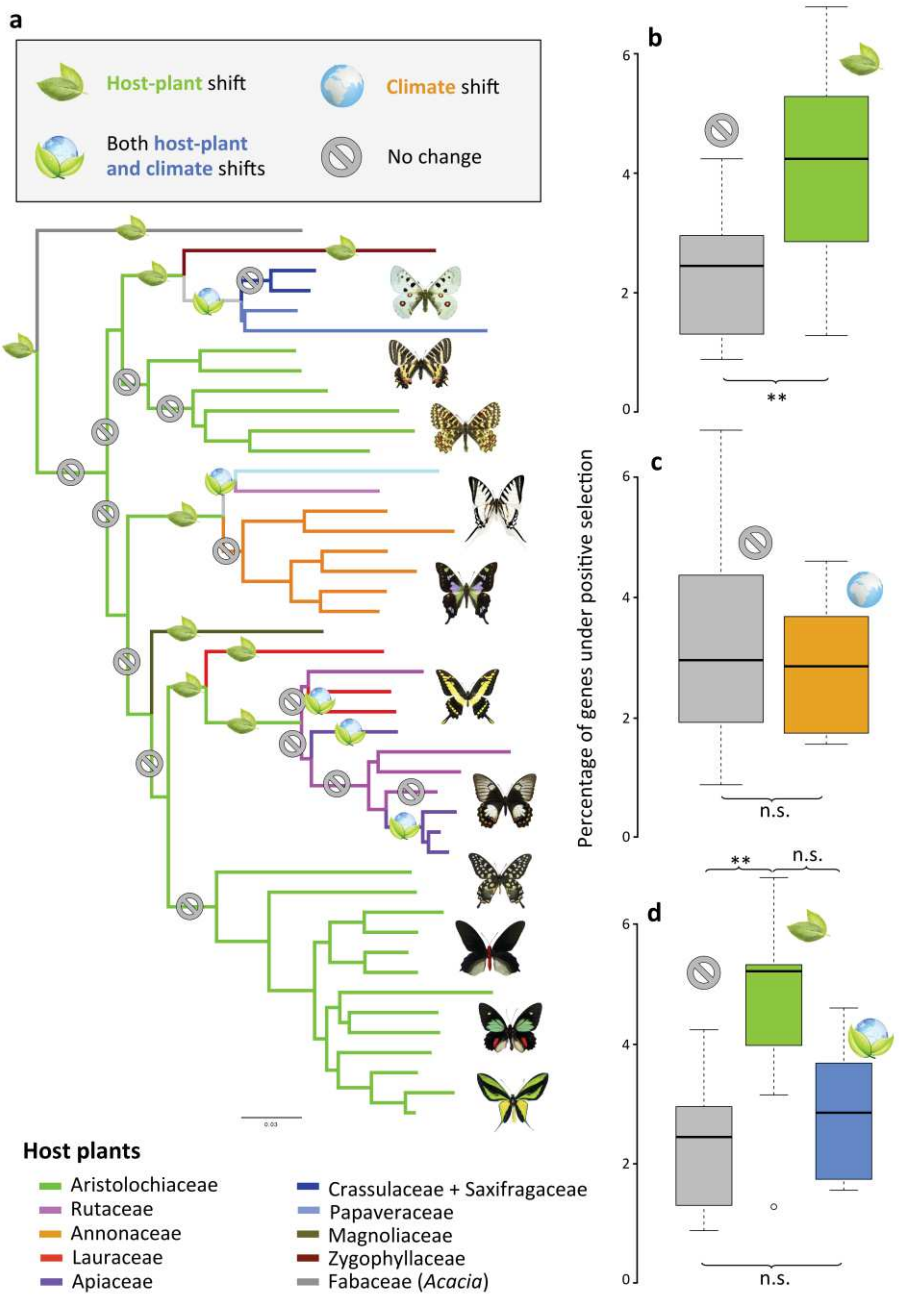


Fig. 4. Host-plant shifts promote higher molecular adaptations. a, Genus-level phylogenomic tree displaying branches with and without host-plant shifts, on which genome-wide analyses of molecular evolution are performed. **b,** Number of genes under positive selection (dN/dS > 1) for swallowtail lineages shifting to new host-plant families (*n*=14, green) or not (*n*=14, grey). **c,** Number of genes under positive selection for swallowtail lineages undergoing climate shifts (*n*=5, orange) or not (*n*=23, grey). **d,** Number of genes under positive selection for swallowtail lineages shifting to new host plants (*n*=9, green), shifting both host-plant and climate (*n*=5, blue) or not (*n*=14, grey). The proportion of genes was estimated with *Dataset 2* (1,533 genes, see **Supplementary Fig. 19** for the results with *Dataset 1* and 520 genes). This demonstrates genome-wide signatures of adaptations in swallowtail lineages shifting to new host-plant families. Genes under positive selection did not contain over- or under-represented functional GO categories (Supplementary Data 2). Wilcoxon rank-sum test: n.s. = not significant ($P > 0.05$), * = $P \leq 0.05$, ** = $P \leq 0.01$. Pictures and icons made by Fabien Condamine.

Supplementary Fig. 14; RevBayes, **Supplementary Fig. 15**; and CoMET, **Supplementary Fig. 16**). Although we should be cautious about the estimations of macroevolutionary rates[59–64],all models concur that diversification rates increase through time either globally or due to recurrent host-plant shifts. Interestingly, this results contrast with the slowdown of diversification that is classically recovered in most phylogenies, often attributed to ecological limits and niche filling processes[64]. This sustained and increasing diversification during the Cenozoic may be explained by ecological opportunities not decreasing, due to a steady increase in host breadth for Papilionidae with new host-plant families colonized through time (**Supplementary Fig. 17**). Opening up new niches, which can also expand due to diversification increases of the host-plant families through time[65–67], would allow continuous increase in diversification rates through time in a dynamic biotic environment, lending support to the primary role of ecological interactions in clade diversification over long timescales – a long-contentious issue[29]. Nonetheless, when taking into account the possibility that rates may have been heterogeneous across the phylogeny, we find that the diversification of three lineages (those feeding on Annonaceae, Lauraceae, and Papaveraceae) had early rates of speciation that are higher than the ancestral rates, but slowed down through time.

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

Genome-wide adaptations to host-plant shifts.

Key innovations are often considered to underlie ecological opportunities and/or evolutionary success[68], particularly in the case of chemically mediated interactions between butterflies and their host plants[13]. Studies on Papilionidae have provided strong examples of specific changes in key genes that confer new abilities to feed on toxic plants and allow host-plant shifts[36,37]. Adaptations of swallowtails to their hosts have particularly been assessed through the study of cytochrome P450 monooxygenases (P450s), which have a major role in detoxifying secondary plant compounds. New P450s appear to arise in swallowtails that colonize new hosts to bypass toxic defences, providing survival and diversification on some but not all plants[15,36,37]. This supports the hypothesis that insect-plant interactions contributed to P450-gene family diversification, with P450s being key innovations that explain the evolutionary and ecological success of phytophagous insects[14,15,36,38,69,70]. However, host-plant shifts not only alter single genes but may also influence unlinked genes[40]. Moreover, host-plant shifts can accompany changes of abiotic environment, which may in turn require further biotic adaptation (new predators and/or competitors). But the macroevolutionary and genomic consequences of the evolutionary dynamics of host-plant shifts have not yet been demonstrated.

Relying on a genomic dataset comprising 45 genomes covering all swallowtail genera[47,71–73], we constructed two specific datasets (Dataset 1: 520 genes & Dataset 2: 1533 genes; mean gene coverage = 26.7x; Methods and Supplementary Data 1). To test whether there are any genomic signatures of positive selection caused by host-plant shifts within swallowtails, we performed a comparative genomic survey of molecular adaptation between

swallowtail lineages that shifted to new host plants compared to non-shifting lineages (*Methods*). We selected 14 phylogenetic branches representing a host-plant shift and 14 phylogenetic branches with no change as negative controls[74,75] (**Fig. 4a**). For a fair molecular comparison, each branch selected as a negative control was chosen to be as close as possible to a test branch representing a host-plant shift (i.e. sister groups, Supplementary **Fig. 18**). Among branches with host-plant shifts, five branches also had a shift in climate preference (represented by distributional changes from tropical to temperate conditions). Using a maximum-likelihood method, we estimated the ratio of non-synonymous substitutions (dN) other synonymous substitutions (dS) in all branches where a host-plant shift was identified relative to branches with no host-plant shift[76,77] (*Methods*). The dN/dS analyses on branches with host-plant shifts (combined or not with environmental shifts) showed more genes with a subset of codons evolving under positive selection (dN/dS > 1) in lineages shifting to a new plant family, although the difference was marginally non-significant for the smallest dataset and highly significant for the second dataset containing more genes (**Fig. 4b**; **Supplementary Fig. S19**; **Supplementary Table 2**, P = 0.0501 / 0.0079 for the two datasets, respectively, Wilcoxon rank-sum test, see *Methods* for the definition of the datasets). However, dN/dS analyses on branches with environmental shifts indicated a balanced number of genes under positive selection (**Fig. 4c**; **Supplementary Fig. S19**; **Supplementary Table 2**, P = 0.336 / 0.8162 for the two datasets, respectively, Wilcoxon rank-sum test), suggesting a lower impact of environmental shifts than host-plant shifts. We then performed dN/dS analyses for branches with host-plant shifts only (not followed by environmental shifts) and found that swallowtail lineages shifting to a new host-plant family had significantly more genes under positive selection (4.41% / 3.98% of genes under positive selection for the two datasets, respectively; **Supplementary Table 2**) than non-shifting lineages (3.02% / 2.43% of genes under positive selection for the two datasets, respectively, **Fig. 4d**; **Supplementary Fig. S19**; Supplementary Table 2, P = 0.0071 / 0.00156 for the two datasets, respectively, Wilcoxon rank-sum test). Surprisingly, the dual changes in environment and host-plant preferences did not spur molecular adaptation across swallowtail lineages compared to control branches (P = 1 / 0.4439 for the two datasets, respectively, Wilcoxon rank-sum test; **Fig. 4d**; **Supplementary Fig. S19**; **Supplementary Table 2**). Comparing the proportion of genes under positive selection between the branches with dual changes and branches with host-plant shifts only shows a marginally significant difference with *Dataset 1* and no difference with *Dataset 2* (P = 0.0327 / 0.1471 for the two datasets, respectively, Wilcoxon rank-sum test; **Fig. 4d**; **Supplementary Figure S19**). However, this result might be an artefact due to the use of a few branches to perform the statistical comparison. Although we did not control for the effect of multi-nucleotide mutations[78], which should affect dN/dS analyses equally for control and host-plant shift branches, we checked individually the gene alignments and performed sensitivity analyses that showed our results are not driven either by an excess of misaligned regions, nor missing data and GC-content variations among species (Methods; **Supplementary Figs. 20-26**). Finally, given that fixing the topology for CodeML (Methods) can spuriously inflate substitution rates on some branches[79], we computed the proportion of genes under positive selection by selecting the gene trees from the largest dataset (*Dataset 2*) for which the focal branches were recovered (in agreement with the species tree). These analyses confirmed the previous results suggesting more genes under positive selection during host-plant shifts (P = 0.0444, Wilcoxon-rank test; **Supplementary Table 2**).

We further studied the functional categories of positively selected genes by using gene ontology (GO) analyses (PANTHER and EggNOG; *Methods*). Applied to the high-quality genomes of *Papilio xuthus*[72] and *Heliconius*

melpomene[80], we found that ~70% of the genes could be associated with a gene function and ~30% lacked annotation, which suggests a gap of knowledge in the current insect database of gene function. Among the annotated genes, we found that genes under positive selection along branches with host shifts did not contain over- or under-represented functional GO categories: 252 out of 1,213 GO categories represented by genes under positive selection ($P > 0.05$, Fisher's exact test after false discovery rate correction; Supplementary Data 2). These results support the hypothesis that genome-wide signatures of adaptations are associated with host-plant shifts, and encourage enlarging the hypothesis that changes in only one or a few candidate gene families could be enough to act as key innovations for adaptation to new resources[13,17]. Despite a weak signal, it is striking that host-plant shifts left stronger genome-wide signatures than were associated with changing climate preferences. This result further suggests that the success of phytophagous insects involved widespread adaptation to biotic interactions than for shifts in the abiotic environment.

To conclude, establishing evolutionary links between ecological adaptations, genomic changes, and species diversification over geological timescales remains a tremendous challenge[28,81,82] with, for instance, important limitations due to the lack of knowledge in functional gene annotations in insects. However, the successful development of powerful analytical tools in conjunction with the increasing availability of insect genomes and improvements in genomic analyses[83] have allowed detection of more genes than those already known to be involved in detoxification pathways playing a role in long-term relationships between plants and insects. Our genome-wide analyses have also generated a list of candidates potentially involved in plant-insect interactions. This opens new research avenues for finding the functionality of genes involved in the adaptation and diversification of phytophagous insects. We hope that our study will help movement in that direction, and that it will provide perspectives for future investigations of other model groups.

Over a half century ago, Ehrlich and Raven[5] proposed that insect-plant interactions driven by diffuse coevolution over long evolutionary periods can be a major source of terrestrial biodiversity. Applied to a widely appreciated case in the insect-plant interactions theory, our study has been able to investigate genome-wide adaptive processes and corresponding macroevolutionary consequences in a comprehensive framework, suggesting that more genes could be involved in host-plant shifts than previously studied in the diversification of herbivorous insects. This result confirms the general belief in the insect-plant community that host-plant shifts are complex and would thus require a number of adaptations, which likely affect various genes beyond those directly involved in detoxification of the plant compounds[36,39,40]. By expanding the possible genes and gene families and identifying more adaptations than those gene families in detoxification pathways that were detected through antagonist interactions[39], we show genomically wide-ranging co-evolutionary consequences[40,84] for close relationships between insects and their larval host plants. Hence, genome-wide macroevolutionary consequences of key adaptations in new insect-plant interactions may be a general feature of the coevolutionary interactions that have generated Earth's diversity.

Acknowledgements

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

Author contributions

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

Competing interests

The authors declare no competing interests.

⇧Back to summary⇧

Methods

Time-calibrated phylogeny of Papilionidae.

We assembled a supermatrix dataset with available data extracted from GenBank as of May 2017 (most of which has been generated by our research group), using five mitochondrial genes (*COI*, *COII*, *ND1*, *ND5* and *rRNA 16S*) and two nuclear markers (*EF-1a* and *Wg*) for 408 Papilionidae species (~71% of the total species diversity) and 20 outgroup species. We aligned the DNA sequences for each gene using MAFFT 7.110[85] with default settings (E-INS-i algorithm), and the alignments were checked for codon stops and eventually refined by eye with Mesquite 3.1 (available at: www.mesquiteproject.org). The best-fit partitioning schemes and substitution models for phylogenetic analyses were determined with PartitionFinder 2.1.1[86] using the *greedy* search algorithm and the Bayesian Information Criterion. All gene alignments were concatenated in a supermatrix, which is available in Figshare (see Data availability).

Phylogenetic relationships were estimated with both maximum likelihood (ML) and Bayesian inference. ML analyses were carried out with IQ-TREE 1.6.8[87]. We set the best-fit partitioning scheme (-ssp option) and used ModelFinder to determine the best-fit substitution model for each partition[88] and then estimated model parameters separately for every partition[89] such that all partitions shared the same set of branch lengths, but we allowed each partition to have its own evolution rate (-m TESTNEW option). For tree search parameters, we relied on a more thorough and slower Nearest-Neighbor-Interchange search to consider all possible Nearest-Neighbor-Interchanges instead of only those in the vicinity previously applied (-allnni option). Following recommendation of IQ-TREE developers, we also set smaller perturbation strength (-pers 0.2) and larger number of stop iterations (-nstop 500) to avoid local optima. We performed 2,000 ultrafast bootstrap replicates to investigate nodal support across the topology, considering values $\geq$ 95 as strongly supported nodes[90].

Estimating phylogenetic relationships for such a dataset is computationally intensive with Bayesian inference. The ML tree inferred with IQ-TREE was used as a starting tree for Bayesian inference as implemented in MrBayes 3.2.6[91]. Rather than using a single substitution model per molecular partition, we sampled across the entire substitution-model space[92] using reversible-jump Markov Chain Monte Carlo (rj-MCMC). Two independent analyses with one cold chain and seven heated chains, each run for 50 million generations, sampled every 5,000 generations. Convergence and performance of Bayesian runs were evaluated using Tracer 1.7.1[93], the average deviation of split frequencies (ADSF) between runs, the effective sample size (ESS) and the potential scale reduction factor (PSRF) values for each parameter. The runs had to have values of ADSF approaching zero, PSRF close to 1.0 and ESS above 200 to be considered convergent. A 50% majority-rule consensus tree was built after conservatively discarding 25% of sampled trees as burn-in. Node support was evaluated with posterior probability considering values $\geq$ 0.95 as strong support[94]. All analyses were performed on the CIPRES Science Gateway computer cluster[95], using BEAGLE[96].

Dating inferences were performed using Bayesian relaxed-clock methods accounting for rate variation across lineages[97]. MCMC analyses implemented in BEAST 1.8.4[98] were employed to approximate the posterior distribution of rates and divergences times and infer their credibility intervals. Estimation of divergence times relied on constraining clade ages through fossil calibrations. Swallowtail fossils are scarce, but five can unambiguously be attributed to the family. The oldest fossil occurrences of Papilionidae are the fossils †*Praepapilio colorado* and †*Praepapilio gracilis*[54], both from the Green River Formation (Colorado, USA). The Green River Formation encompasses a 5 million-years period between ~48.5 and 53.5 Ma, which falls within the Ypresian (47.8-56 Ma) in the early Eocene[99]. These fossils can be

phylogenetically placed at the crown of the family as they share synapomorphies with all extant subfamilies[56,100], and have proven to be reliable calibration points for the crown group[18,34,47]. Two other fossils belong to Parnassiinae, whose systematic position was assessed using phylogenetic analyses based on both morphological and molecular data in a total-evidence approach[18]. The first is †*Thaites ruminiana*[101], a compression fossil from limestone in the Niveau du gypse d'Aix Formation of France (Bouches-du-Rhône, Aix-en-Provence, France) within the Chattian (23.03-28.1 Ma) of the late Oligocene[55,102]. †*Thaites* is sister to Parnassiini, and occasionally sister to Luehdorfiini + Zerynthiini[18]. 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*[103], an exoskeleton and compression fossil from Italy (Tuscany) from the Messinian (5.33-7.25 Ma, late Miocene)[55]. †*Doritites* is sister to *Archon* (Luehdorfiini[18]), in agreement with Carpenter[104]. The crown of Luehdorfiini was thus constrained for divergence time estimation using a uniform distribution bounded with 5.33 Ma. Absolute ages of geological formations were taken from the latest update of the geological time scale.

We used a conservative approach to apply calibration priors with the selected fossil constraints by setting uniform priors bounded with a minimum age equal to the youngest age of the geological formation where each fossil was found. All uniform calibration priors were set with an upper bound equal to the estimated age of angiosperms (150 Ma[105]), which is more than three times older than the oldest Papilionidae fossil. This upper age is intentionally set as ancient to allow exploration of potentially old ages for the clade. Since the fossil record of butterflies is incomplete and biased[106], caution is needed in using these fossil calibrations (effect shown in burying beetles[107]).

After enforcing the fossil calibrations, we set the following settings and priors: a partitioned dataset (after the best-fitting PartitionFinder scheme) was analysed using the uncorrelated lognormal distribution clock model, with the mean set to a uniform prior between 0 and 1, and an exponential prior (lambda = 0.333) for the standard deviation. The branching process prior was set to a birth–death[108] process, using the following uniform priors: the birth–death mean growth rate ranged between 0 and 10 with a starting value at 0.1, and the birth–death relative death rate ranged between 0 and 1 (starting value = 0.5). We performed four independent BEAST analyses for 100 million generations, sampled every 10,000th, resulting in 10,000 samples in the posterior distribution of which the first 2,500 samples were discarded as burn-in. All analyses were performed on the CIPRES Science Gateway computer cluster[95], using BEAGLE[96]. Convergence and performance of each MCMC run were evaluated using Tracer 1.7.1[93] and the ESS for each parameter (ESS > 200). We combined the four runs using LogCombiner 1.8.4[98]. A maximum-clade credibility (MCC) tree was reconstructed, with median ages and 95% credibility intervals (CI). The BEAST files generated for this study are available in Figshare (see Data availability).

Estimating ancestral host-plant association.

We inferred the temporal evolution of host-plant association up to the ancestral host plant(s) at the root of Papilionidae using three approaches: the ML implementation of the Markov k-state (Mk) model[109], the ML Dispersal-Extinction-Cladogenesis (DEC) model[110], and the Bayesian approach in BayesTraits[111]. These approaches require a time-calibrated tree and a matrix of character states (current host-plant preference) for each species in the tree. An extensive bibliographic survey was conducted to obtain primary larval host-plants at the family level[5,30,112–114]. The host associations of species were categorized using the following twelve character states: (1) Annonaceae, (2) Apiaceae, (3) Aristolochiaceae, (4) Crassulaceae or Saxifragaceae (core Saxifragales), (5) Fabaceae, (6) Hernandiaceae, (7) Lauraceae; (8) Magnoliaceae, (9) Papaveraceae, (10) Rosaceae, (11) Rutaceae, and (12) Zygophyllaceae. The

host-plant matrix of Papilionidae is available in Figshare (see Data availability).

Ancestral states for host-plant association were first reconstructed using the Mk model (one rate for all transitions between states) allowing any host shift to be equally probable. The Mk model does not allow multiple states for a species. The few species that use multiple host families were thus scored with the most frequent host association. The Mk model was performed with Mesquite 3.1 (available at: www.mesquiteproject.org). To estimate the support of any one character state over another, the most likely state was selected according to a decision threshold, such that if the log likelihoods between two states differ by two log-likelihood units, the one with lower likelihood is rejected[109].

The DEC model was also used to reconstruct ancestral host-plant states[110,115]. As with the Mk model, we assumed that host-plant shifts occurred at equivalent probabilities between plant families and through time, which may not be true given that the host-plant families of Papilionidae did not originate at the same time (e.g. Aristolochiaceae originated around 108.07 Ma [95% credibility intervals: 81.01-132.66 Ma][116], and Annonaceae originated about 98.94 Ma [95% credibility intervals: 84.78-113.70 Ma][116]). We used the estimated molecular ages of the different host-plant groups to constrain our inferences of ancestral host plants *a posteriori*. We preferred such an approach compared to a more constrained one in which the DEC model is informed with a matrix of host-plant appearances based on their estimated ages by implementing matrices of presence/absence of the character states through time (equivalent to the time-stratified paleogeographic model, see below for inference of biogeographical history).

Finally, the Bayesian approach implemented in BayesTraits 3.0.1[111] was performed to provide a cross-validation of ML analyses. This approach automatically detects shifts in rates of evolution for multistate data using rj-MCMC. Numbers of parameters and priors were set by default. We ran the rj-MCMC for 10 million generations and sampled states and parameters every 1,000 generations (burn-in of 10,000 generations). We specifically estimated ancestral states at 21 nodes as well as at the root of Papilionidae. For this analysis, we used a set of 100 trees randomly taken from the dating analysis to probe the robustness of our ancestral state estimation across topological uncertainty.

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

The Mk and BayesTraits models always inferred with high support (relative probability = 0.915 and 0.789, respectively) that Aristolochiaceae is the ancestral host plant at the crown of Papilionidae. With the unconstrained DEC model, we found that the ancestral host-plant preference for Papilionidae was always composed of Aristolochiaceae, but also included another family (either Fabaceae, Hernandiaceae or Zygophyllaceae, which are only fed upon by *Baronia*, *Lamproptera* and *Hypermnestra*, respectively). As the sister lineage to all other Papilionidae, *Baronia* is the only species that feeds on Fabaceae. More precisely, only one species of Fabaceae is consumed: *Vachellia cochliacantha* (formerly *Acacia cochliacantha*; recent changes in *Acacia* taxonomy[117]). However, *Vachellia* diverged from its sister clade in the Eocene, approximately 50 Ma, and diversified in the Miocene between 13 and 17 Ma[118], which substantially postdate the origin of Papilionidae. Therefore this result suggests that the family Aristolochiaceae represents the most likely candidate as the ancestral host-plant of Papilionidae. Hernandiaceae are consumed by *Lamproptera* (occasionally by *Papilio homerus*, *Graphium codrus*, *G. doson* and *G. empedovana*[114]). More precisely, the host plants of *Lamproptera* belong to the genus *Illigera*. This plant genus diverged from its sister genus

48 Ma[116] and started diversifying 27 Ma[119]. The derived phylogenetic position of *Lamproptera* and the age of its use as a host plant make it very unlikely that Hernandiaceae could constitute the ancestral host plant for Papilionidae. Similarly, the family Zygophyllaceae is consumed by *Hypermnestra*, most specifically it feeds on the genus *Zygophyllum* in Central Asia. The genus *Zygophyllum* is not monophyletic, but Asian *Zygophyllum* appeared 19.6 Ma[120]. Applying the same rationale, we are able to discard Zygophyllaceae as a candidate ancestral host plant for Papilionidae. To further refine our ancestral host-plant estimates, we built a presence-absence matrix of plant families based on clade origins estimated in molecular dating studies. Thereby, the age of the different plants can be used to constrain the inference of ancestral host plants. Under such a constrained model, Aristolochiaceae is always recovered as the most likely ancestral host-plant for Papilionidae. It is also interesting that almost all Aristolochiaceae feeders have *Aristolochia* as host plants, and tests to determine which genus of Aristolochiaceae was originally consumed by Papilionidae showed that it was *Aristolochia*.

Time-calibrated phylogeny of the ancestral host: the Aristolochiaceae.

Estimation of ancestral host-plant relationships indicated that the family Aristolochiaceae was the ancestral host for Papilionidae. We refer to Aristolochiaceae in its traditional circumscription including the genera *Asarum*, *Saruma*, *Thottea* and *Aristolochia*. The Angiosperm Phylogeny Group[121] proposes that Aristolochiaceae also includes the holoparasitic genera *Hydnora* and *Prosopanche* (Hydnoraceae), as well as the monotypic family Lactoridaceae from the Juan Fernandez Islands of Chile (*Lactoris fernandeziana*). The conclusion of APG[121] is based on an online survey[122] rather than on primary data and this is why we disagree with their argumentation as well as the resulting conclusion of APG given available resilient primary molecular phylogenomic data. However, arguments based on morphology and anatomy[123–126], genetics[50,127–131], molecular divergence time[116,131], and conservation considerations (Tod Stuessy, pers. comm. with S.W., July 2019) favour splitting them into four families: Aristolochiaceae (*Aristolochia* and *Thottea*), Asaraceae (*Asarum* and *Saruma*), Hydnoraceae (*Hydnora* and *Prosopanche*), and Lactoridaceae (*Lactoris*), collectively called the perianth-bearing Piperales. Therefore we extracted and assembled a supermatrix dataset with available data from GenBank for the perianth-bearing Piperales and its sister lineage, the perianth-less Piperales including Saururaceae and Piperaceae (as of May 2017, most of which has been generated by our research group). We obtained four chloroplast genes (*matK*, *rbcl*, *trnL*, *trnL-trnF*) and one nuclear marker (*ITS*) for 247 species of perianth-bearing Piperales (~49% of the total species diversity[132]) and six outgroups from perianth-less Piperales. We could not include the two genera *Hydnora* and *Prosopanche* (Hydnoraceae) because available genetic data do not overlap those of perianth-bearing Piperales[127,130,133,134]. We applied the same analytical procedure that we did for Papilionidae. DNA sequences for each gene were aligned using MAFFT 7.110[85] with default settings (E-INS-i algorithm and Q-INS-I to take into account secondary structure). Resulting alignments were checked for codon stops and eventually refined by eye with Mesquite 3.1 (available at: www.mesquiteproject.org). The best-fit partitioning schemes and substitution models for phylogenetic analyses were determined with PartitionFinder 2.1.1[86]. All gene alignments were concatenated into a supermatrix; the final dataset is available in Figshare (see Data availability).

Phylogenetic relationships were estimated with Bayesian inference as implemented in MrBayes 3.2.6[91]. Rather than using a single substitution model per molecular partition, we sampled across the entire substitution-model space[92] using rj-MCMC. Two independent analyses with one cold chain and seven heated chains, each were run for 50 million generations, sampled every 5,000 generations. Convergence and performance of

Bayesian runs were evaluated using Tracer 1.7.1[93] and the ESS, ADSF and PSRF criteria. Once convergence was achieved, a 50% majority-rule consensus tree was built after discarding 25% of the sampled trees as burn-in.

Bayesian relaxed-clock methods were used that accounted for rate variation across lineages[97]. MCMC analyses implemented in BEAST 1.8.4[98] were employed to approximate the posterior distribution of rates and divergences times and infer their credibility intervals. Estimation of divergence times relied on constraining clade ages through fossil calibrations. Three unambiguous fossils from perianth-bearing Piperales (Aristolochiaceae *sensu lato*), and one corresponding to the family Saururaceae were used. First, we relied on the fossil record of the monotypic family Lactoridaceae (*Lactoris fernandeziana*)[127,131], a shrub endemic to cloud forest of the Juan Fernández Islands archipelago of Chile. The oldest pollen fossil for the group is †*Lactoripollenites africanus*[135,136] from the Turonian/Campanian (72.1-89.8 Ma) of the Orange Basin in South Africa. This fossil confers a minimum age of 72.1 Ma for the stem node of *Lactoris fernandeziana*. Second, the oldest and only pollen record of the Aristolochiaceae was recently described from Late Cretaceous sediments of Siberia: †*Aristolochiacidites viluiensis*[57] from the Timerdyakh Formation of the latest Campanian to earliest Maastrichtian (66-72.1 Ma) in the Vilui Basin (Russia). Because inaperturate pollen grains in combination with this unique exine configuration and fitting size can be observed in extant members of Aristolochiaceae, this fossil provides a minimum age of 66 Ma for the family. The third fossil belongs to the genus *Aristolochia* and described as †*Aristolochia austriaca*[58] from the Pannonian (late Miocene) in the Hollabrunn-Mistelbach Formation (Austria). Based on a thorough morphological leaf comparison, this fossil is assigned to a species group including *Aristolochia baetica* and *Aristolochia rotunda*, which then confers a minimum age of 7.25 Ma for the clade. Finally, we used the fossil †*Saururus tuckerae*[137] from the Princeton Chert of Princeton in British Columbia (Canada), which is part of the Princeton Group, Allenby Formation dated with stable isotopes to the middle Eocene[138]. This fossil has been phylogenetically placed as sister to extant *Saururus* species[138], hence providing a minimum age of 44.3 Ma for the stem node of *Saururus*. Absolute ages of geological formations were taken from the latest update of the geological time scale.

We set the following settings and priors: a partitioned dataset (after the best-fitting PartitionFinder scheme) was analysed using the uncorrelated lognormal distribution clock model, with the mean set to a uniform prior between 0 and 1, and an exponential prior (lambda = 0.333) for the standard deviation. The branching process prior was set to a birth–death[108] process, using the following uniform priors: the birth–death mean growth rate ranged between 0 and 10 with a starting value at 0.1, and the birth–death relative death rate ranged between 0 and 1 (starting value = 0.5). We performed four independent BEAST analyses for 100 million generations, sampled every 10,000th, resulting in 10,000 samples in the posterior distribution of which the first 2500 samples were discarded as burn-in. All analyses were performed on the CIPRES Science Gateway computer cluster[95], using BEAGLE[96]. Convergence and performance of each MCMC run were evaluated using Tracer 1.7.1[93] and the ESS for each parameter. We combined the four runs using LogCombiner 1.8.4[98]. The MCC tree was reconstructed with median age and 95% CI. The BEAST files generated for this study are available in Figshare (see Data availability).

Dual biogeographic history of Papilionidae and Aristolochiaceae.

We estimated the ancestral area of origin and geographic range evolution for both clades using the ML approach of DEC model[110] as implemented in the C++ version[139,140] that is available at: https://github.com/champost/DECX. To infer the biogeographic history of a clade, DEC requires a

time-calibrated tree, the current distribution of each species for a set of geographic areas, and a time-stratified geographic model that is represented by connectivity matrices for specified time intervals spanning the entire evolutionary history of the group.

The geographic distribution for each species in Papilionidae[30,113,114] and Aristolochiaceae was categorized as present or absent in each of the following areas: (1) West Nearctic [WN], (2) East Nearctic [EN], (3) Central America [CA], (4) South America [SA], (5) West Palearctic [WP], (6) East Palearctic [EP], (7) Madagascar [MD], (8) Indonesia and Wallacea [WA], (9) India [IN], (10) Africa [AF], and (11) Australasia [AU]. The resulting matrices of species distribution for the two groups are available in Figshare (see Data availability).

A time-stratified geographic model was built using connectivity matrices that take into account paleogeographic changes through time, with time slices indicating the possibility or not for a species to access a new area[140]. Based on palaeogeographical reconstructions[141–143], we created a connectivity matrix for each geological epoch that represented a period bounded by major changes in tectonic and climatic conditions thought to have affected the distribution of organisms. The following geological epochs were selected: (1) 0 to 5.33 Ma (Pliocene to present), (2) 5.33 to 23.03 Ma (Miocene), (3) 23.03 to 33.9 Ma (Oligocene), (4) 33.9 to 56 Ma (Eocene), and (5) 56 Ma to the origin of the clade (Palaeocene to Late Cretaceous). For each of these five time intervals, we specified constraints on area connectivity by coding 0 if any two areas are not connected or 1 if they are connected in a given time interval. We assumed a conservative dispersal matrix with equal dispersal rates between areas through time[144].

Impact of host-plant shifts on swallowtail diversification.

We tested the effect of host-plant association on diversification by estimating speciation and extinction rates with five methods to cross-test hypotheses and corroborate results. Analyses were performed on 100 dated trees randomly sampled from the Bayesian dating analyses to take into account the uncertainty in age estimates. We used the following approaches: (1) ML-based trait-dependent diversification[145,146]; (2) ML-based time-dependent diversification[147]; (3) Bayesian analysis of macroevolutionary mixture[148]; (4) Bayesian branch-specific diversification rates[149]; and (5) Bayesian episodic birth-death model[150]. It is worth mentioning that each method differs at several points in their estimation of speciation and extinction rates. For instance, trait-dependent birth-death models estimate constant speciation and extinction rates[146], whereas time-dependent birth-death models estimate clade-specific speciation and extinction rates and their variation through time[147,149]. Therefore, we expect some differences in the values of estimated diversification rates that are inherent to each approach. Our diversification analyses should be seen as complementary to the inferred diversification trend rather than corroborating the values and magnitude of speciation and extinction rates.

First, we computed the probability of obtaining a clade as large as size *n*, given the crown age of origin, the overall net diversification rate of the family, and an extinction rate as a fraction of speciation rate following the approach in Condamine et al.[34] relying on the method of moments[151]. We used the R-package *LASER* 2.3[152] to estimate the net diversification rates of Papilionidae and six clades shifting to new host-plants with the *bd.ms* function (providing crown age and total species diversity). Then, we used the *crown.limits* function to estimate the mean expected clade size for each clade shifting to new host-plants given clades' crown age and overall net diversification rates, and we finally computed the probability to observe such clade size using the *crown.p* function. All rate estimates were calculated with three ε values (ε=0/0.5/0.9), knowing that the extinction rate in swallowtails

is usually low[34] (supported by the results of this study).

Second we relied on the state-dependent speciation and extinction (SSE) model, in which speciation and extinction rates are associated with phenotypic evolution of a trait along a phylogeny[145]. In particular, we used the Multiple State Speciation Extinction model (MuSSE[146]) implemented in the R-package *diversitree* 0.9–10[153], which allows multiple character states to be studied. Larval host-plant data were taken from previous works[5,18,30,34,113,114,154]. The following 10 host-plant character states and corresponding ratios of sampled species in the tree of all known species for each character (sampling fractions) were used: 1 = Aristolochiaceae (110/152), 2 = Annonaceae (69/138), 3 = Lauraceae (33/39), 4 = Apiaceae (9/10), 5 = Rutaceae (119/163), 6 = Crassulaceae (19/19), 7 = Papaveraceae (44/44), 8 = Fabaceae (1/1), 9 = Zygophyllaceae (2/2), and 10 = Magnoliaceae (2/2). Data at a lower taxonomic level than plant family were not used because of the large number of multiple associations exhibited by genera that could alter the phylogenetic signal. We assigned a single state to each species by selecting the food plant with the maximum number of collections for each species. We did not employ multiple states per species, which represents a lesser problem because (1) few swallowtail species feed on multiple plant families, (2) current shared-state models can only model two states, and (3) the addition of multi-plant states to the MuSSE analysis would have greatly increased the number of parameters. We performed both ML and Bayesian MCMC analyses (10,000 steps) performed using an exponential (1/(2 × net diversification rate)) prior with starting parameter values obtained from the best-fitting ML model and resulting speciation, extinction and transition rates. After a burn-in of 500 steps, we estimated posterior density distribution for speciation, extinction and transition rates. There have been concerns about the power of SSE models to infer diversification dynamics from a distribution of species traits[155–157], hence other birth-death models were used to corroborate the results obtained with SSE models.

Third, to provide an independent assessment of the relationship between diversification rates and host specificity, we used the ML approach of Morlon et al.[147] implemented in the R-package *RPANDA* 1.3[158]. This is a birth–death method in which speciation and/or extinction rates may change continuously through time. This method has the advantage of not assuming constant extinction rate over time (unlike BAMM[148]), and allows clades to have declining diversity since extinction can exceed speciation, meaning that diversification rates can be negative[147]. For each clade that shifted to a new host family, we designed and fitted six diversification models: (1) a Yule model, where speciation is constant and extinction is null; (2) a constant birth-death model, where speciation and extinction rates are constant; (3) a variable speciation rate model without extinction; (4) a variable speciation rate model with constant extinction; (5) a rate-constant speciation and variable extinction rate model; and (6) a model in which both speciation and extinction rates vary. Models were compared by computing the ML estimate of each model and the resulting Akaike information criterion corrected by sample size (AICc) We then plotted rates through time with the best fit model for each clade, and the rates for the family as a whole for comparison purpose.

Fourth, we performed models that allow diversification rates to vary among clades across the whole phylogeny. BAMM 2.5[148,159] was used to explore for differential diversification dynamic regimes among clades differing in their host-plant feeding. BAMM can automatically detect rate shifts and sample distinct evolutionary dynamics that explain the diversification dynamics of a clade without *a priori* hypotheses on how many and where these shifts might occur. Evolutionary dynamics can involve time-variable diversification rates; in BAMM, speciation is allowed to vary exponentially through time while extinction is maintained constant: subclades in a tree may diversify faster (or slower) than others. This

Bayesian approach can be useful in detecting shifts of diversification potentially associated with key innovations[159]. BAMM analyses were run with four MCMC for 20 million generations, sampling every 20,000th and with three different values (1, 5 and 10; **Supplementary Table 3**) of the compound Poisson prior (CPP) to ensure the posterior is independent of the prior[160]. We accounted for non-random incomplete taxon sampling using the implemented analytical correction; we set a sampling fraction per genus based on the known species diversity of each genus. Mixing and convergence among runs (ESS > 200 after 15% burn-in) were assessed with the R-package *BAMMtools* 2.1[161] to estimate (1) the mean global rates of diversification through time, (2) the estimated number of rate shifts evaluating alternative diversification models comparing priors and posterior probabilities, and (3) the clade-specific rates through time when a distinct macroevolutionary regime is identified.

Fifth, BAMM has been criticized for incorrectly modelling rate-shifts on extinct lineages, that is, unobserved (extinct or non-sampled) lineages inherit the ancestral diversification process and cannot experience subsequent diversification-rate shifts[160,162]. To solve this, we used a Bayesian approach implemented in RevBayes 1.0.10[163] that models rate shifts consistently on extinct lineages by using the SSE framework[149,160]. Although there is no information of rate shifts for unobserved/extinct lineages in a phylogeny including extant species only, these types of events must be accounted for in computing the likelihood. The number of rate categories is fixed in the analysis but RevBayes allows any number to be specified, thus allowing direct comparison of different macroevolutionary regimes.

Finally, we evaluated the impact of abrupt changes in diversification using the Bayesian episodic birth-death model of CoMET[150] implemented in the R-package *TESS* 2.1[164]. These models allow detection of discrete changes in speciation and extinction rates concurrently affecting all lineages in a tree, and estimate changes in diversification rates at discrete points in time, but can also infer mass extinction events (sampling events in which the extant diversity is reduced by a fraction[165]). Speciation and extinction rates can change at those points but remain constant within time intervals. In addition, TESS uses independent CPPs to simultaneously detect mass extinction events and discrete changes in speciation and extinction rates, while TreePar estimates the magnitude and timing of speciation and extinction changes independently to the occurrence of mass extinctions (i.e. the three parameters cannot be estimated simultaneously due to parameter identifiability issues[165]). We performed two independent analyses allowing and disallowing mass extinction events. Bayes factor comparisons were used to assess model fit between models with varying number and time of changes in speciation/extinction rates and mass extinctions.

Detecting genome-wide adaptations during host-plant shifts.

We analysed genomic sequence data in swallowtail butterflies that have independently shifted to new ecological (biological) traits. Similar approaches have been conducted on mammals[166,167] and birds[168], but have been rarely implemented on arthropod groups over such a long geological time scale. Here we estimated swallowtail molecular evolution with whole genome data and compared selection regimes on protein-coding genes along independent branches with or without host-plant shift and/or environmental shift.

For these analyses, we studied 45 whole genomes[47] covering all 32 genera of the family Papilionidae: 41 of which were previously generated by our research group added to four genomes already available[71–73]. In summary, raw reads (Sequence Read Archive: SRR8954507-SRR8954549) were cleaned using Trimmomatic 0.33[169], and assembled into contigs and scaffolds with SOAPdenovo-63mer 2.04[170] to obtain whole genome assemblies (30x average read depth[47]). All coding DNA sequences (CDS) were

retrieved from the high-quality annotated genome of *Papilio xuthus*[72]. To annotate the sequences of all our genomes, a BLAST search using all available CDS of *Papilio xuthus* was performed at the amino-acid level (using tblastn). For each species the recovered genes were aligned one by one with *Papilio xuthus* using TranslatorX[171]. This method performs alignment at the amino-acid level and preserves the open reading frame. All sites showing intraspecific variation were set to N, to conservatively avoid false informative sites. Any contamination was removed using CroCo 0.1[172] and orthologous proteins were identified with OrthoFinder 2.2.0[173]. Finally, CDS alignments were strongly cleaned from misaligned sequences (gene by gene) using HMMCleaner 1.8[174]. A last cleaning step was performed using trimAl 1.2rev59[175], which is designed to trim alignments for large-scale phylogenomic analyses. The resulting dataset comprised 6,621 genes in at least four sampled species (median of 32% of missing data), which was used to reconstruct a robust phylogenomic tree of Papilionidae[47] (**Supplementary Fig. 18**).

We used this genomic dataset of 45 species representing all genera in which the resulting genus-level swallowtail phylogenomic tree[47] accurately represents the evolutionary associations with host plants as estimated using the ancestral-state analyses applied to the species-level phylogeny[34] (**Fig. 1 Supplementary Figs. 4, 5**). We thus transferred the inference of ancestral host-plant shifts on the phylogenomic tree and selected the branches representing a host-plant shift and/or a shift of climate preference (in general from tropical to temperate conditions; **Supplementary Fig. 10**). We also selected branches with no change as negative controls[74]. As a result, 14 branches are selected to measure the impact of a host-plant shift and 14 branches are selected as controls (**Supplementary Fig. 18**). Within these 14 branches with an ecological change, nine branches represent host-plant shifts only, and five branches correspond to shifts in both host plant and environment (from tropical to temperate conditions). To test the impact of these different changes on the genomes, two datasets were created, *Dataset 1* and *2*. Given the low quality of the genomes of *Allancastria cerisyi* and *Parnassius imperator*, these two genomes were discarded for the downstream analyses. We first selected the genes from the 6,621-gene dataset for each focal branch using three criteria: (1) the dataset is composed only of orthologous protein-coding genes (OrthoFinder 2.2[173]), (2) the species needed to accurately define the branch were available (i.e. crown node of the clade), and (3) for each branch, one species per tribe was available, and therefore include a different number of genes per branch. Thus, for the *Dataset 1*, only the genes containing sequences for the species needed to generate all focal branches were selected. This stringent selection leads to a *Dataset 1* comprising only 520 genes but the same genes for all branches (no missing genes). For *Dataset 2*, the genes were selected for each branch independently (i.e. for a given branch, a gene was selected if the sequence needed to generate that branch was present). This second selection leads to 1,439 genes per branch on average among a total of 1,533 genes, which were selected at least once for one branch. The genomic dataset is available in Figshare (see Data availability).

We studied the ratio (ω) of nonsynonymous/synonymous substitution rate (dN/dS) to find genes under positive selection[77,176]. The dN/dS ratio is traditionally used to estimate selective pressure from protein-coding sequences. If host-plant shifts have no effect on the selection of a given gene, we expect a dN/dS = 1 and the selective regime is considered neutral. However, if host-plant shifts result in positive selection on coding genes, the ratio increases such that dN/dS > 1. Finally, it is possible that host-plant shifts lead to purifying selection, thus reducing the number of non-synonymous substitutions and resulting in dN/dS < 1. Here we focused on the adaptation of Papilionidae to host-plant shifts, i.e. outgroups are not studied. We tested if branches representing inferred host-plant shifts along the phylogeny of swallowtails have more genes with

dN/dS > 1 than lineages that did not have an inferred shift. The *branch-site* models allow ω to vary both among sites in the protein and across branches on the tree and aim to detect positive selection affecting a few sites along particular lineages. The approach described by Zhang et al.[177] was chosen to determine genome-wide selection regimes as performed with two maximum-likelihood models: (1) a null model assuming two site classes, one with dN/dS < 1 and one with dN/dS = 1 (model = 2, NSsites = 2, fix_omega = 1, omega = 1); and (2) an alternative model adding a third site class with dN/dS > 1 (model = 2, NSsites = 2, fix_omega = 0, omega = 1.5). The fit for including positive selection is tested using a likelihood ratio test comparing the null model with the alternative model with one degree of freedom[77,178]. If the alternative model is better suited to host-shift branches, it is more likely the gene was under positive selection during the host-plant shifts. For each gene and for each branch, both the null and alternative models using CodeML were implemented in PAML 4[179] with a fixed topology (as inferred with the phylogenomic dataset[47]) and the nucleotide alignment of each gene. To test the robustness of the estimations, we used a false discovery rate test to control false positives[180]. Finally, for each branch, we reported the number of genes under positive selection (i.e. for which the alternative model including the site class with dN/dS >1 have a better likelihood) on the total gene number. The proportion of genes under positive selection was compared with associated control branches for branches representing host-plant shifts, environmental shifts or both plant and environmental shifts using the non-parametric Wilcoxon rank-sum test[181].

Sensitivity analyses.

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

First, it has been shown that the choice of the tree is an important factor for the branch-site analysis of positive selection[182]. Indeed, constraining the topology for a given gene may lead to overestimating the number of substitution events for the constrained branches[79] and so could lead to overestimating the dN/dS ratio. Estimating dN/dS over thousands of gene trees would make the branch comparison not equal between control and test branches. Indeed, in a given gene tree it is likely and expected that the species topology is not always recovered, which results in a different number of branches compared to the species tree. For instance the host-plant shift to Annonaceae might disappear in certain proportions of genes. We thus decided to estimate dN/dS on a fixed species tree topology for all genes to be sure to be able to measure this ratio for each gene that must be present in the topology for the focal branches. However, given that this issue can lead to a bias in our analysis, we decided to compute the number of gene trees that did not recover the branches of interest. We then checked whether the branches leading to a host-plant shift were more often unrecovered than the control branches without shift. Overall the control branches were less often recovered than host-plant shift branches (P = 0.030, Wilcoxon rank-sum test; data presented in **Supplementary Table 4**), which suggests that if gene tree/species tree discordance leads to overestimation of positive selection, then this overestimation is higher for control branches than for host-plant shift branches. Finally, we filtered out the gene trees for which the focal branches were recovered in agreement with the species tree and used these genes to re-estimate the proportion of genes under positive selection among this new set of genes. We found that the P-value remains significant (P = 0.0444, Wilcoxon- rank test; more genes during host-plant shifts than along control branches, **Supplementary Table 2**). Then we specifically focused on missing data and GC content variation among genes known to bias dN/dS estimations. Missing data are prone to introducing misaligned regions that could create false positives in branch-site likelihood method for detecting positive selection[183–185]. Variations

in GC content are known to impact the estimation of dN/dS mainly through the process of GC-biased gene conversion (gBGC[186–188]).

The number of missing data ('N' and '-') sites and GC content at the third codon position (GC3) were computed using a home-made C++ program created with BIO++ library[189]. Mean GC content and missing data was calculated per gene and for each branch. For a given branch, mean GC3 and missing data were computed for the species of a clade for which the branch is the root. All statistics and graphical representations were performed using the R-packages *tidyverse*[190] and *cowplot*[191]. We found that genes under positive selection (PS_{genes}, $n_{Dataset1}$ = 142, $n_{Dataset2}$ = 407) have significantly more missing data and GC3 than genes not under positive selection (NS_{genes}, $n_{Dataset1}$ = 378, $n_{Dataset2}$ = 1126; P = 0.001 / 0.02 for the two datasets, respectively, Mann-Whitney test; **Supplementary Fig. 20**). This result confirms that branch-site likelihood methods for detecting positive selection are sensitive to missing data, probably because of misaligned sites[183,184], and that GC content that may be influenced by gBGC[186,187].

Missing data was, however, heterogeneously distributed among species, ranging from less than 1% in *Papilio xuthus* to 45% in *Hypermnestra helios* (**Supplementary Fig. 21**). The difference in missing data between branches with (n = 14, mean $missing_{Dataset1}$ = 13.4%, mean $missing_{Dataset2}$ = 14.1%) or without host-plant shifts (n = 14, mean $missing_{Dataset1}$ = 12.8%, mean $missing_{Dataset2}$ = 12.7%) is not significant (P = 0.83 / 1.00 for the two datasets, respectively, Mann-Whitney test; **Supplementary Fig. 22**). Additionally, there is no correlation between the number of genes under positive selection and the amount of missing data (P = 0.33 / 0.20 for the two datasets, respectively, Spearman's correlation test; **Supplementary Fig. 23**). For GC3, we also found variation between species ranging from 37% in *Parnassius smintheus* to 44% in *Papilio antimachus* (**Supplementary Fig. 24**). Similarly to missing data, we found no significant difference between plant-shift and no plant-shift branches (P = 0.63 / 0.63 for the two datasets, Mann-Whitney test; **Supplementary Fig. 25**) and there is no correlation between the number of genes under positive selection and GC3 (P = 0.20 / 0.1362 for the two datasets, respectively, Spearman's correlation test; **Supplementary Fig. 26**).

Despite the known fact that false positives can increase with the amount of missing data, our control analyses indicate that variations in missing data and GC content do not drive the signal that more genes are under positive selection in branches that have undergone a host-plant shift. Additionally to these controls, we checked by eyes all the gene alignments at the amino-acid level for genes under positive selection in branches with and without host-plant shifts using SeaView 4[192]. Misaligned regions, which could lead to biased dN/dS ratios[193], were not significantly more detected for genes under positive selection in branches with host-plant shifts. In some cases we found ourselves in complicated situations to discriminate between false and true positive selected genes.

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

Gene ontology.

To annotate proteins of our alignment, we used the two different approaches implemented in PANTHER 14[194] (available at: http://pantherdb.org/) and EggNOG 5.0[195,196] (available at: http://eggnog5.embl.de/#/app/home). We used the HMM Scoring tool to assign PANTHER family (library version 14.1[194]) to the protein of *Papilio xuthus* (assembly Pxut_1.0); similar results were obtained using another high-quality annotated genome (from *Heliconius melpomene*) as reference (assembly ASM31383v2). We

performed the statistical overrepresentation test implemented on the PANTHER online website, relying on the GO categories in the PANTHER GO-Slim annotation dataset including Molecular function, Biological process, and Cellular component. Firstly, we tested if positively selected genes have over- or under-represented functional GO categories as compared to the whole set of genes (option "PANTHER Generic Mapping"). Secondly, we tested if positively selected genes involving a host-plant shift along the 14 branches have over- or under-represented functional categories. These statistical comparisons were performed with the Fisher's exact test using the false discovery rate correction to control for false positives. Independently, we used the eggNOG-mapper v2[195] (https://github.com/eggnogdb/eggnog-mapper) and the associated Lepidoptera database (LepNOG, including the genomes of *Bombyx mori*, *Danaus plexippus* and *Heliconius melpomene*[196]) to annotate the proteins of our dataset. EggNOG uses precomputed orthologous groups and phylogenies from the database to transfer functional information from fine-grained orthologs only. We used the diamond method as recommended[195]. Finally, we reported the GO families inferred for the proteins of the *Dataset 2*.

Data availability

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

Code availability

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

⇧Back to summary⇧

References

⇧Back to summary⇧

1. Becerra, J. X. On the factors that promote the diversity of herbivorous insects and plants in tropical forests. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 6098–103 (2015).
2. Stork, N. E. How many species of insects and other terrestrial arthropods are there on earth? *Annu. Rev. Entomol.* **63**, 31–45 (2018).
3. Grimaldi, D. A. & Engel, M. S. *Evolution of the insects*. (Cambridge University Press, 2005).
4. Strong, D. R., Lawton, J. H. & Southwood, R. *Insects on plants : community patterns and mechanisms*. (Harvard University Press, 1984).
5. Ehrlich, P. R. & Raven, P. H. Butterflies and plants: a study in coevolution. *Evolution* **18**, 586–608 (1964).
6. Thompson, J. N. Concepts of coevolution. *Trends Ecol. Evol.* **4**, 179–183 (1989).
7. Mitter, C., Farrell, B. & Wiegmann, B. The phylogenetic study of adaptive zones: Has phytophagy promoted insect diversification? *Am. Nat.* **132**, 107–128 (1988).
8. Farrell, B. D. 'Inordinate fondness' explained: why are there so many beetles? *Science* **281**, 555–9 (1998).
9. Berenbaum, M. & Specialization, P. F. *Chemical mediation of host-plant specialization: the papilionid paradigm*. *Specialization, speciation, and radiation: The evolutionary biology of herbivorous insects* (University of California Press, 2008).
10. Winter, S., Friedman, A. L. L., Astrin, J. J., Gottsberger, B. & Letsch, H. Timing and host plant associations in the evolution of the weevil tribe Apionini (Apioninae, Brentidae, Curculionoidea, Coleoptera) indicate an ancient co-diversification pattern of beetles and flowering plants. *Mol. Phylogenet. Evol.* **107**, 179–190 (2017).
11. Kergoat, G. J. *et al.* Opposite macroevolutionary responses to environmental changes in grasses and insects during the Neogene grassland expansion. *Nat. Commun.* **9**, 5089 (2018).
12. Wheat, C. W. *et al.* The genetic basis of a plant–insect coevolutionary key innovation. *Proc. Natl. Acad. Sci.* **104**, 20427–20431 (2007).
13. Edger, P. P. *et al.* The butterfly plant arms-race escalated by gene and genome duplications. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 8362–8366 (2015).
14. Calla, B. *et al.* Cytochrome P450 diversification and hostplant utilization patterns in specialist and generalist moths: Birth, death and adaptation. *Mol. Ecol.* **26**, 6021–6035 (2017).
15. Nallu, S. *et al.* The molecular genetic basis of herbivory between butterflies and their host plants. *Nat. Ecol. Evol.* **2**, 1418–1427 (2018).
16. Karageorgi, M. *et al.* Genome editing retraces the evolution of toxin resistance in the monarch butterfly. *Nature* **574**, 409–412 (2019).
17. Sahoo, R. K., Warren, A. D., Collins, S. C. & Kodandaramaiah, U. Hostplant change and paleoclimatic events explain diversification shifts in skipper butterflies (Family: Hesperiidae). *BMC Evol. Biol.* **17**, 174 (2017).
18. Condamine, F. L., Rolland, J., Höhna, S., Sperling, F. A. H. & Sanmartín, I. Testing the role of the red queen and court jester as drivers of the macroevolution of apollo butterflies. *Syst. Biol.* **67**, 940–964 (2018).
19. Letsch, H. *et al.* Climate and host-plant associations shaped the evolution of ceutorhynch weevils throughout the Cenozoic. *Evolution* **72**, 1815–1828 (2018).
20. Forister, M. L. *et al.* The global distribution of diet breadth in insect herbivores. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 442–7 (2015).
21. Winkler, I. S., Mitter, C. & Scheffer, S. J. Repeated climate-linked host shifts have promoted diversification in a temperate clade of leaf-mining flies. *Proc. Natl. Acad. Sci. U. S. A.* **106**, 18103–8 (2009).
22. Chomicki, G., Weber, M., Antonelli, A., Bascompte, J. & Kiers, E. T. The impact of mutualisms on species richness. *Trends Ecol. Evol.* **34**, 698–711 (2019).
23. Janz, N. Ehrlich and Raven revisited: Mechanisms underlying codiversification of plants and enemies. *Annu. Rev. Ecol. Evol. Syst.* **42**, 71–89 (2011).
24. Suchan, T. & Alvarez, N. Fifty years after Ehrlich and Raven, is there support for plant–insect coevolution as a major driver of species diversification? *Entomol. Exp. Appl.* **157**, 98–112 (2015).

25. Endara, M.-J. *et al.* Coevolutionary arms race versus host defense chase in a tropical herbivore-plant system. *Proc. Natl. Acad. Sci. U. S. A.* **114**, E7499–E7505 (2017).
26. Simon, J.-C. *et al.* Genomics of adaptation to host-plants in herbivorous insects. *Brief. Funct. Genomics* **14**, 413–23 (2015).
27. Hammer, T. J., Janzen, D. H., Hallwachs, W., Jaffe, S. P. & Fierer, N. Caterpillars lack a resident gut microbiome. *Proc. Natl. Acad. Sci. U. S. A.* **114**, 9641–9646 (2017).
28. Hua, X. & Bromham, L. Darwinism for the genomic age: connecting mutation to diversification. *Front. Genet.* **8**, 12 (2017).
29. Hembry, D. H. & Weber, M. G. Ecological interactions and macroevolution: A new field with old roots. *Annu. Rev. Ecol. Evol. Syst.* **51**, (2020).
30. Scriber, J. M., Tsubaki, Y. & Lederhouse, R. C. *Swallowtail butterflies : their ecology and evolutionary biology*. (Scientific Publishers, 1995).
31. Nishida, R. Sequestration of defensive substances from plants by Lepidoptera. *Annu. Rev. Entomol.* **47**, 57–92 (2002).
32. Schmeiser, H. H., Stiborovà, M. & Arlt, V. M. Chemical and molecular basis of the carcinogenicity of Aristolochia plants. *Curr. Opin. Drug Discov. Devel.* **12**, 141–148 (2009).
33. Poon, S. L. *et al.* Genome-wide mutational signatures of aristolochic acid and its application as a screening tool. *Sci. Transl. Med.* **5**, 197ra101 (2013).
34. Condamine, F. L., Sperling, F. A. H., Wahlberg, N., Rasplus, J.-Y. & Kergoat, G. J. What causes latitudinal gradients in species diversity? Evolutionary processes and ecological constraints on swallowtail biodiversity. *Ecol. Lett.* **15**, 267–277 (2012).
35. Simonsen, T. J. *et al.* Phylogenetics and divergence times of Papilioninae (Lepidoptera) with special reference to the enigmatic genera *Teinopalpus* and *Meandrusa*. *Cladistics* **27**, 113–137 (2011).
36. Berenbaum, M. R., Favret, C. & Schuler, M. A. On defining 'Key Innovations' in an adaptive radiation: Cytochrome P450s and Papilionidae. *Am. Nat.* **148**, S139–S155 (1996).
37. Cohen, M. B., Schuler, M. A. & Berenbaum, M. R. A host-inducible cytochrome P-450 from a host-specific caterpillar: molecular cloning and evolution. *Proc. Natl. Acad. Sci. U. S. A.* **89**, 10920–10924 (1992).
38. Li, W., Schuler, M. A. & Berenbaum, M. R. Diversification of furanocoumarin-metabolizing cytochrome P450 monooxygenases in two papilionids: Specificity and substrate encounter rate. *Proc. Natl. Acad. Sci. U. S. A.* **100 Suppl**, 14593–14598 (2003).
39. Thompson, J. N. Variation in preference and specificity in monophagous and oligophagous swallowtail butterflies. *Evolution* **42**, 118–128 (1988).
40. Thompson, J. N., Wehling, W. & Podolsky, R. Evolutionary genetics of host use in swallowtail butterflies. *Nature* **344**, 148–150 (1990).
41. Berenbaum, M. R. & Feeny, P. P. Chemical mediation of host-plant specialization: the papilionid paradigm. in *Specialization, Speciation, and Radiation: The Evolutionary Biology of Herbivorous Insects* (ed. Tilmon, K.) 2–19 (University of California Press, 2008).
42. Zakharov, E. V., Caterino, M. S. & Sperling, F. A. H. Molecular phylogeny, historical biogeography, and divergence time estimates for swallowtail butterflies of the genus *Papilio* (Lepidoptera: Papilionidae). *Syst. Biol.* **53**, 193–215 (2004).
43. Braby, M., Trueman, J. & Eastwood, R. When and where did troidine butterflies (Lepidoptera: Papilionidae) evolve? Phylogenetic and biogeographic evidence suggests an origin in remnant Gondwana in the Late Cretaceous. *Invertebr. Syst.* **19**, 113–143 (2005).
44. Simonsen, T. J. *et al.* Phylogenetics and divergence times of Papilioninae (Lepidoptera) with special reference to the enigmatic genera *Teinopalpus* and *Meandrusa*. *Cladistics* **27**, 113–137 (2011).
45. Condamine, F. L., Silva-Brandão, K. L., Kergoat, G. J. & Sperling, F. A. Biogeographic and diversification patterns of Neotropical Troidini butterflies (Papilionidae) support a museum model of diversity dynamics for Amazonia. *BMC Evol. Biol.* **12**, 82 (2012).
46. Condamine, F. L. *et al.* Deciphering the evolution of birdwing butterflies 150 years after Alfred Russel Wallace. *Sci. Rep.* **5**, 11860 (2015).
47. Allio, R. *et al.* Whole genome shotgun phylogenomics resolves the pattern and timing of swallowtail butterfly evolution. *Syst. Biol.* **69**, 38–60 (2020).

48. McKenna, D. D., Sequeira, A. S., Marvaldi, A. E. & Farrell, B. D. Temporal lags and overlap in the diversification of weevils and flowering plants. *Proc. Natl. Acad. Sci. U. S. A.* **106**, 7083–7088 (2009).
49. Takahashi, D. & Setoguchi, H. Molecular phylogeny and taxonomic implications of *Asarum* (Aristolochiaceae) based on ITS and *matK* sequences. *Plant Species Biol.* **33**, 28–41 (2018).
50. Wanke, S. *et al.* Evolution of Piperales—matK gene and trnK intron sequence data reveal lineage specific resolution contrast. *Mol. Phylogenet. Evol.* **42**, 477–497 (2007).
51. Neinhuis, C., Wanke, S., Hilu, K. W., Müller, K. & Borsch, T. Phylogeny of Aristolochiaceae based on parsimony, likelihood, and Bayesian analyses of trnL-trnF sequences. *Plant Syst. Evol.* **250**, 7–26 (2005).
52. Wanke, S., González, F. & Neinhuis, C. Systematics of pipevines: combining morphological and fast-evolving molecular characters to investigate the relationships within subfamily Aristolochioideae (Aristolochiaceae). *Int. J. Plant Sci.* **167**, 1215–1227 (2006).
53. González, F. *et al.* Present trans-Pacific disjunct distribution of *Aristolochia* subgenus *Isotrema* (Aristolochiaceae) was shaped by dispersal, vicariance and extinction. *J. Biogeogr.* **41**, 380–391 (2014).
54. Durden, C. J. & Rose, H. *Butterflies from the middle Eocene: the earliest occurrence of fossil Papilionoidea (Lepidoptera)*. (Prarce-Sellards Ser. Tax. Mem. Mus., 1978).
55. Sohn, J., Labandeira, C., Davis, D. & Mitter, C. An annotated catalog of fossil and subfossil Lepidoptera (Insecta: Holometabola) of the world. *Zootaxa* **3286**, 1–132 (2012).
56. de Jong, R. Estimating time and space in the evolution of the Lepidoptera. *Tijdschr. voor Entomol.* **150**, 319–346 (2007).
57. Hofmann, C.-C. & Zetter, R. Upper Cretaceous sulcate pollen from the Timerdyakh Formation, Vilui Basin (Siberia). *Grana* **49**, 170–193 (2010).
58. Meller, B. The first fossil *Aristolochia* (Aristolochiaceae, Piperales) leaves from Austria. *Palaeontol. Electron.* **17**, 1–17 (2014).
59. Nee, S., May, R. M. & Harvey, P. H. The reconstructed evolutionary process. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **344**, 305–11 (1994).
60. Nee, S. Birth-death models in macroevolution. *Annu. Rev. Ecol. Evol. Syst.* **37**, 1–17 (2006).
61. Rabosky, D. L. & Lovette, I. J. Explosive evolutionary radiations: Decreasing speciation or increasing extinction through time? *Evolution* **62**, 1866–1875 (2008).
62. Crisp, M. D. & Cook, L. G. Explosive radiation or cryptic mass extinction? Interpreting signatures in molecular phylogenies. *Evolution* **63**, 2257–2265 (2009).
63. Quental, T. B. & Marshall, C. R. Diversity dynamics: molecular phylogenies need the fossil record. *Trends Ecol. Evol.* **25**, 434–441 (2010).
64. Morlon, H. Phylogenetic approaches for studying diversification. *Ecol. Lett.* **17**, 508–525 (2014).
65. Xue, B. *et al.* Accelerated diversification correlated with functional traits shapes extant diversity of the early divergent angiosperm family Annonaceae. *Mol. Phylogenet. Evol.* **142**, 106659 (2020).
66. Folk, R. A. *et al.* Rates of niche and phenotype evolution lag behind diversification in a temperate radiation. *Proc. Natl. Acad. Sci. U. S. A.* **116**, 10874–10882 (2019).
67. Sun, M. *et al.* Recent accelerated diversification in rosids occurred outside the tropics. *Nat. Commun.* **11**, 3333 (2020).
68. Losos, J. B. Adaptive radiation, ecological opportunity, and evolutionary determinism. *Am. Nat.* **175**, 623–639 (2010).
69. Cheng, T. *et al.* Genomic adaptation to polyphagy and insecticides in a major East Asian noctuid pest. *Nat. Ecol. Evol.* **1**, 1747–1756 (2017).
70. Rane, R. V *et al.* Detoxifying enzyme complements and host use phenotypes in 160 insect species. *Curr. Opin. Insect Sci.* **31**, 131–138 (2019).
71. Cong, Q., Borek, D., Otwinowski, Z. & Grishin, N. V. Tiger swallowtail genome reveals mechanisms for speciation and caterpillar chemical defense. *Cell Rep.* **10**, 910–919 (2015).
72. Li, X. *et al.* Outbred genome sequencing and CRISPR/Cas9 gene editing in butterflies. *Nat. Commun.* **6**, 8212 (2015).
73. Nishikawa, H. *et al.* A genetic mechanism for female-limited Batesian mimicry in *Papilio* butterfly. *Nat. Genet.* **47**, 405–409 (2015).
74. Thomas, G. W. C. & Hahn, M. W. Determining the null model for detecting adaptive convergence

from genomic data: a case study using echolocating mammals. *Mol. Biol. Evol.* **32**, 1232–1236 (2015).
75. Zou, Z. & Zhang, J. No genome-wide protein sequence convergence for echolocation. *Mol. Biol. Evol.* **32**, 1237–1241 (2015).
76. Kimura, M. *The Neutral Theory of Molecular Evolution*. (Cambridge University Press, 1983).
77. Yang, Z. *Computational Molecular Evolution*. (Oxford University Press, 2006).
78. Venkat, A., Hahn, M. W. & Thornton, J. W. Multinucleotide mutations cause false inferences of lineage-specific positive selection. *Nat. Ecol. Evol.* **2**, 1280–1288 (2018).
79. Mendes, F. K. & Hahn, M. W. Gene tree discordance causes apparent substitution rate variation. *Syst. Biol.* **65**, 711–21 (2016).
80. Dasmahapatra, K. K. *et al.* Butterfly genome reveals promiscuous exchange of mimicry adaptations among species. *Nature* **487**, 94–98 (2012).
81. Walden, N. *et al.* Nested whole-genome duplications coincide with diversification and high morphological disparity in Brassicaceae. *Nat. Commun.* **11**, 3795 (2020).
82. McGee, M. D. *et al.* The ecological and genomic basis of explosive adaptive radiation. *Nature* **586**, 75–79 (2020).
83. Thomas, G. W. C. *et al.* Gene content evolution in the arthropods. *Genome Biol.* **21**, 15 (2020).
84. de Medeiros, B. A. S. & Farrell, B. D. Evaluating species interactions as a driver of phytophagous insect divergence. *bioRxiv* 842153 (2019). doi:10.1101/842153
85. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).
86. Lanfear, R., Frandsen, P. B., Wright, A. M., Senfeld, T. & Calcott, B. PartitionFinder 2: New methods for selecting partitioned models of evolution for molecular and morphological phylogenetic analyses. *Mol. Biol. Evol.* **34**, 772–773 (2016).
87. Nguyen, L.-T., Schmidt, H. A., von Haeseler, A. & Minh, B. Q. IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* **32**, 268–274 (2015).
88. Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., von Haeseler, A. & Jermiin, L. S. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat. Methods* **14**, 587–589 (2017).
89. Chernomor, O., von Haeseler, A. & Minh, B. Q. Terrace aware data structure for phylogenomic inference from supermatrices. *Syst. Biol.* **65**, 997–1008 (2016).
90. Minh, B. Q., Nguyen, M. A. T. & von Haeseler, A. Ultrafast approximation for phylogenetic bootstrap. *Mol. Biol. Evol.* **30**, 1188–1195 (2013).
91. Ronquist, F. *et al.* MrBayes 3.2: Efficient bayesian phylogenetic inference and model choice across a large model space. *Syst. Biol.* **61**, 539–542 (2012).
92. Huelsenbeck, J. P., Larget, B. & Alfaro, M. E. Bayesian phylogenetic model selection using reversible jump Markov Chain Monte Carlo. *Mol. Biol. Evol.* **21**, 1123–1133 (2004).
93. Rambaut, A., Drummond, A. J., Xie, D., Baele, G. & Suchard, M. A. Posterior summarization in bayesian phylogenetics using Tracer 1.7. *Syst. Biol.* **67**, 901–904 (2018).
94. Douady, C. J., Delsuc, F., Boucher, Y., Doolittle, W. F. & Douzery, E. J. P. Comparison of bayesian and maximum likelihood bootstrap measures of phylogenetic reliability. *Mol. Biol. Evol.* **20**, 248–254 (2003).
95. Miller, M. A. *et al.* A RESTful API for access to phylogenetic tools via the CIPRES Science Gateway. *Evol. Bioinforma.* **11**, EBO.S21501 (2015).
96. Ayres, D. L. *et al.* BEAGLE: An application programming interface and high-performance computing library for statistical phylogenetics. *Syst. Biol.* **61**, 170–173 (2012).
97. Drummond, A. J., Ho, S. Y. W., Phillips, M. J. & Rambaut, A. Relaxed phylogenetics and dating with confidence. *PLoS Biol.* **4**, e88 (2006).
98. Drummond, A. J., Suchard, M. A., Xie, D. & Rambaut, A. Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Mol. Biol. Evol.* **29**, 1969–1973 (2012).
99. Smith, M. E., Singer, B. & Carroll, A. $^{40}Ar/^{39}Ar$ geochronology of the Eocene Green River Formation, Wyoming. *Geol. Soc. Am. Bull.* **115**, 549–565 (2003).
100. de Jong, R. Fossil butterflies, calibration points and the molecular clock (Lepidoptera: Papilionoidea). *Zootaxa* **4270**, 1–63 (2017).

101. Scudder, S. H. Fossil butterflies. *Mem. Am. Assoc. Adv. Sci.* **1**, 1–99 (1875).
102. Rasnitsyn, A. P. & Zherikhin, V. V. Appendix: Alphabetic list of selected insect fossil sites. in *History of Insects* 437–446 (Kluwer Academic Publishers, 2002). doi:10.1007/0-306-47577-4_4
103. Rebel, H. *Doritites bosniaskii*. Sitzungsberichte der akademie der wissenschaften. Mathematischen-Naturwissenschaftliche classe. *Abteilung 1 Mineral. Biol. Erdkd.* **1**, 734–741 (1898).
104. Carpenter, F. *Treatise on Invertebrate Paleontology: Arthropoda 4. Superclass Hexapoda*. *Geological Society of America* (1992).
105. Magallón, S., Gómez-Acevedo, S., Sánchez-Reyes, L. L. & Hernández-Hernández, T. A metacalibrated time-tree documents the early rise of flowering plant phylogenetic diversity. *New Phytol.* **207**, 437–453 (2015).
106. Sohn, J.-C., Labandeira, C. C. & Davis, D. R. The fossil record and taphonomy of butterflies and moths (Insecta, Lepidoptera): implications for evolutionary diversity and divergence-time estimates. *BMC Evol. Biol.* **15**, 12 (2015).
107. Toussaint, E. F. A. & Condamine, F. L. To what extent do new fossil discoveries change our understanding of clade evolution? A cautionary tale from burying beetles (Coleoptera: *Nicrophorus*). *Biol. J. Linn. Soc.* **117**, 686–704 (2016).
108. Gernhard, T. The conditioned reconstructed process. *J. Theor. Biol.* **253**, 769–778 (2008).
109. Lewis, P. O. A likelihood approach to estimating phylogeny from discrete morphological character data. *Syst. Biol.* **50**, 913–925 (2001).
110. Ree, R. H. & Smith, S. A. Maximum likelihood inference of geographic range evolution by dispersal, local extinction, and cladogenesis. *Syst. Biol.* **57**, 4–14 (2008).
111. Pagel, M. & Meade, A. Bayesian analysis of correlated evolution of discrete characters by reversible-jump Markov chain Monte Carlo. *Am. Nat.* **167**, 808–25 (2006).
112. Igarashi, S. The classification of the Papilionidae mainly based on the morphology of their immature stages. *Lepid. Sci.* **34**, 41–96 (1984).
113. Collins, N. M. & Morris, M. *Threatened swallowtail butterflies of the world: the IUCN red data book.* (IUCN, 1985).
114. Tyler, H. A., Brown, K. S. & Wilson, K. H. *Swallowtail Butterflies of the Americas: A Study in Biological Dynamics, Ecological Diversity, Biosystematics, and Conservation*. (Scientific Publishers, 1994).
115. Ree, R. H., Moore, B. R., Webb, C. O. & Donoghue, M. J. A likelihood framework for inferring the evolution of geographic range on phylogenetic trees. *Evolution* **59**, 2299–2311 (2005).
116. Massoni, J., Couvreur, T. L. & Sauquet, H. Five major shifts of diversification through the long evolutionary history of Magnoliidae (Angiosperms). *BMC Evol. Biol.* **15**, 49 (2015).
117. Kyalangalilwa, B., Boatwright, J. S., Daru, B. H., Maurin, O. & van der Bank, M. Phylogenetic position and revised classification of *Acacia s.l.* (Fabaceae: Mimosoideae) in Africa, including new combinations in *Vachellia* and *Senegalia*. *Bot. J. Linn. Soc.* **172**, 500–523 (2013).
118. Miller, J. T., Murphy, D. J., Ho, S. Y. W., Cantrill, D. J. & Seigler, D. Comparative dating of *Acacia*: combining fossils and multiple phylogenies to infer ages of clades with poor fossil records. *Aust. J. Bot.* **61**, 436–445 (2013).
119. Michalak, I., Zhang, L.-B. & Renner, S. S. Trans-Atlantic, trans-Pacific and trans-Indian Ocean dispersal in the small Gondwanan Laurales family Hernandiaceae. *J. Biogeogr.* **37**, 1214–1226 (2010).
120. Wu, S.-D. *et al.* Evolution of asian interior arid-zone biota: Evidence from the diversification of asian *Zygophyllum* (Zygophyllaceae). *PLoS One* **10**, e0138697 (2015).
121. Chase, M. W. *et al.* An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG IV. *Bot. J. Linn. Soc.* **181**, 1–20 (2016).
122. Christenhusz, M. J. M., Vorontsova, M. S., Fay, M. F. & Chase, M. W. Results from an online survey of family delimitation in angiosperms and ferns: recommendations to the Angiosperm Phylogeny Group for thorny problems in plant classification. *Bot. J. Linn. Soc.* **178**, 501–528 (2015).
123. Gonzáles, F., Rudall, P. J. & Furness, C. A. Microsporogenesis and systematics of Aristolochiaceae. *Bot. J. Linn. Soc.* **137**, 221–242 (2001).

124. González, F. & Rudall, P. The questionable affinities of *Lactoris*: Evidence from branching pattern, inflorescence morphology, and stipule development. *Am. J. Bot.* **88**, 2143–2150 (2001).
125. Isnard, S. *et al.* Growth form evolution in Piperales and its relevance for understanding angiosperm diversification: An integrative approach combining plant architecture, anatomy, and biomechanics. *Int. J. Plant Sci.* **173**, 610–639 (2012).
126. Wagner, S. T. *et al.* Major trends in stem anatomy and growth forms in the perianth-bearing Piperales, with special focus on *Aristolochia*. *Ann. Bot.* **113**, 1139–1154 (2014).
127. Nickrent, D. L. *et al.* Molecular data place Hydnoraceae with Aristolochiaceae. *Am. J. Bot.* **89**, 1809–1817 (2002).
128. Kelly, L. M. & González, F. Phylogenetic relationships in Aristolochiaceae. *Syst. Bot.* **28**, 236–249 (2003).
129. Neinhuis, C., Wanke, S., Hilu, K. W., Müller, K. & Borsch, T. Phylogeny of Aristolochiaceae based on parsimony, likelihood, and Bayesian analyses of trnL-trnF sequences. *Plant Syst. Evol.* **250**, 7–26 (2005).
130. Naumann, J. *et al.* Single-copy nuclear genes place haustorial Hydnoraceae within piperales and reveal a cretaceous origin of multiple parasitic angiosperm lineages. *PLoS One* **8**, e79204 (2013).
131. Salomo, K. *et al.* The emergence of earliest angiosperms may be earlier than fossil evidence indicates. *Syst. Bot.* **42**, 607–619 (2017).
132. Christenhusz, M. J. M. & Byng, J. W. The number of known plants species in the world and its annual increase. *Phytotaxa* **261**, 201–217 (2016).
133. Naumann, J. *et al.* Detecting and characterizing the highly divergent plastid genome of the nonphotosynthetic parasitic plant *Hydnora visseri* (Hydnoraceae). *Genome Biol. Evol.* **8**, 345–363 (2016).
134. Jost, M., Naumann, J., Rocamundi, N., Cocucci, A. A. & Wanke, S. The first plastid genome of the Holoparasitic genus *Prosopanche* (Hydnoraceae). *Plants* **9**, 306 (2020).
135. Zavada, M. S. & Benson, J. M. First fossil evidence for the primitive angiosperm family Lactoricidae. *Am. J. Bot.* **74**, 1590–1594 (1987).
136. Gamerro, J. C. & Barreda, V. New fossil record of Lactoridaceae in southern South America: A palaeobiogeographical approach. *Bot. J. Linn. Soc.* **158**, 41–50 (2008).
137. Smith, S. Y. & Stockey, R. A. Establishing a fossil record for the perianthless Piperales: *Saururus tuckerae* sp. nov. (Saururaceae) from the Middle Eocene Princeton Chert. *Am. J. Bot.* **94**, 1642–1657 (2007).
138. Massoni, J., Doyle, J. & Sauquet, H. Fossil calibration of Magnoliidae, an ancient lineage of angiosperms. *Palaeontol. Electron.* **18**, 1–25 (2015).
139. Smith, S. A. Taking into account phylogenetic and divergence-time uncertainty in a parametric biogeographical analysis of the Northern Hemisphere plant clade Caprifolieae. *J. Biogeogr.* **36**, 2324–2337 (2009).
140. Beeravolu, C. R. & Condamine, F. L. An extended maximum likelihood inference of geographic range evolution by dispersal, local extinction and cladogenesis. *bioRxiv* 038695 (2016). doi:10.1101/038695
141. Scotese, C. R. A continental drift flipbook. *J. Geol.* **112**, 729–741 (2004).
142. Blakey, R. C. Gondwana paleogeography from assembly to breakup — A 500 m.y. odyssey. *Geol. Soc. Am. Spec. Pap.* **441**, 1–28 (2008).
143. Seton, M. *et al.* Global continental and ocean basin reconstructions since 200 Ma. *Earth-Science Rev.* **113**, 212–270 (2012).
144. Chacón, J. & Renner, S. S. Assessing model sensitivity in ancestral area reconstruction using Lagrange: A case study using the Colchicaceae family. *J. Biogeogr.* **41**, 1414–1427 (2014).
145. Maddison, W. P., Midford, P. E. & Otto, S. P. Estimating a binary character's effect on speciation and extinction. *Syst. Biol.* **56**, 701–710 (2007).
146. FitzJohn, R. G., Maddison, W. P. & Otto, S. P. Estimating trait-dependent speciation and extinction rates from incompletely resolved phylogenies. *Syst. Biol.* **58**, 595–611 (2009).
147. Morlon, H., Parsons, T. L. & Plotkin, J. B. Reconciling molecular phylogenies with the fossil record. *Proc. Natl. Acad. Sci. U. S. A.* **108**, 16327–16332 (2011).
148. Rabosky, D. L. *et al.* Rates of speciation and morphological evolution are correlated across

the largest vertebrate radiation. *Nat. Commun.* **4**, 1958 (2013).
149. Höhna, S. *et al.* A Bayesian approach for estimating branch-specific speciation and extinction rates. *bioRxiv* 555805 (2019). doi:10.1101/555805
150. May, M. R., Höhna, S. & Moore, B. R. A Bayesian approach for detecting the impact of mass-extinction events on molecular phylogenies when rates of lineage diversification may vary. *Methods Ecol. Evol.* **7**, 947–959 (2016).
151. Magallon, S. & Sanderson, M. J. Absolute diversification rates in angiosperm clades. *Evolution* **55**, 1762–1780 (2001).
152. Rabosky, D. L. Likelihood methods for detecting temporal shifts in diversification rates. *Evolution* **60**, 1152–1164 (2006).
153. FitzJohn, R. G. Diversitree: Comparative phylogenetic analyses of diversification in R. *Methods Ecol. Evol.* **3**, 1084–1092 (2012).
154. Scriber, J. M. Host-plant suitability. in *Chemical Ecology of Insects* (eds. Bell, W. J. & Cardé, R. T.) 159–202 (Springer US, 1984).
155. Davis, M. P., Midford, P. E. & Maddison, W. Exploring power and parameter estimation of the BiSSE method for analyzing species diversification. *BMC Evol. Biol.* **13**, 38 (2013).
156. Maddison, W. P. & FitzJohn, R. G. The unsolved challenge to phylogenetic correlation tests for categorical characters. *Syst. Biol.* **64**, 127–136 (2015).
157. Rabosky, D. L. & Goldberg, E. E. Model inadequacy and mistaken inferences of trait-dependent speciation. *Syst. Biol.* **64**, 340–355 (2015).
158. Morlon, H. *et al.* RPANDA: An R package for macroevolutionary analyses on phylogenetic trees. *Methods Ecol. Evol.* **7**, 589–597 (2016).
159. Rabosky, D. L. Automatic detection of key innovations, rate shifts, and diversity-dependence on phylogenetic trees. *PLoS One* **9**, e89543 (2014).
160. Moore, B. R., Höhna, S., May, M. R., Rannala, B. & Huelsenbeck, J. P. Critically evaluating the theory and performance of Bayesian analysis of macroevolutionary mixtures. *Proc. Natl. Acad. Sci. U. S. A.* **113**, 9569–9574 (2016).
161. Rabosky, D. L. *et al.* BAMMtools: An R package for the analysis of evolutionary dynamics on phylogenetic trees. *Methods Ecol. Evol.* **5**, 701–707 (2014).
162. Rabosky, D. L., Mitchell, J. S. & Chang, J. Is BAMM flawed? Theoretical and practical concerns in the analysis of multi-rate diversification models. *Syst. Biol.* **66**, 477–498 (2017).
163. Höhna, S. *et al.* RevBayes: Bayesian phylogenetic inference using graphical models and an interactive model-specification language. *Syst. Biol.* **65**, 726–736 (2016).
164. Höhna, S., May, M. R. & Moore, B. R. TESS: An R package for efficiently simulating phylogenetic trees and performing Bayesian inference of lineage diversification rates. *Bioinformatics* **32**, 789–791 (2016).
165. Stadler, T. Mammalian phylogeny reveals recent diversification rate shifts. *Proc. Natl. Acad. Sci. U. S. A.* **108**, 6187–6192 (2011).
166. Partha, R. *et al.* Subterranean mammals show convergent regression in ocular genes and enhancers, along with adaptation to tunneling. *eLife* **6**, e25884 (2017).
167. Wu, J., Yonezawa, T. & Kishino, H. Rates of molecular evolution suggest natural history of life history traits and a Post-K-Pg nocturnal bottleneck of placentals. *Curr. Biol.* **27**, 3025–3033 (2017).
168. Zhang, G. *et al.* Comparative genomics reveals insights into avian genome evolution and adaptation. *Science* **346**, 1311–1320 (2014).
169. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
170. Luo, R. *et al.* SOAPdenovo2: An empirically improved memory-efficient short-read de novo assembler. *Gigascience* **1**, 18 (2012).
171. Abascal, F., Zardoya, R. & Telford, M. J. TranslatorX: Multiple alignment of nucleotide sequences guided by amino acid translations. *Nucleic Acids Res.* **38**, W7–W13 (2010).
172. Simion, P. *et al.* A software tool 'CroCo' detects pervasive cross-species contamination in next generation sequencing data. *BMC Biol.* **16**, 28 (2018).
173. Emms, D. M. & Kelly, S. OrthoFinder: Phylogenetic orthology inference for comparative genomics. *Genome Biol.* **20**, 238 (2019).

174. Di Franco, A., Poujol, R., Baurain, D. & Philippe, H. Evaluating the usefulness of alignment filtering methods to reduce the impact of errors on evolutionary inferences. *BMC Evol. Biol.* **19**, 21 (2019).
175. Capella-Gutierrez, S., Silla-Martinez, J. M. & Gabaldon, T. trimAl: A tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972–1973 (2009).
176. Yang, Z. & Nielsen, R. Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. *Mol. Biol. Evol.* **17**, 32–43 (2000).
177. Zhang, J., Nielsen, R. & Yang, Z. Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. *Mol. Biol. Evol.* **22**, 2472–2479 (2005).
178. Yang, Z. Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. *Mol. Biol. Evol.* **15**, 568–573 (1998).
179. Yang, Z. PAML 4: Phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24**, 1586–1591 (2007).
180. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc.* **57**, 289–300 (1995).
181. Bauer, D. F. Constructing confidence sets using rank statistics. *J. Am. Stat. Assoc.* **67**, 687–690 (1972).
182. Diekmann, Y. & Pereira-Leal, J. B. Gene tree affects inference of sites under selection by the branch-site test of positive selection. *Evol. Bioinforma.* **11**, 11–17 (2015).
183. Mallick, S., Gnerre, S., Muller, P. & Reich, D. The difficulty of avoiding false positives in genome scans for natural selection. *Genome Res.* **19**, 922–933 (2009).
184. Fletcher, W. & Yang, Z. The effect of insertions, deletions, and alignment errors on the branch-site test of positive selection. *Mol. Biol. Evol.* **27**, 2257–2267 (2010).
185. Jordan, G. & Goldman, N. The effects of alignment error and alignment filtering on the sitewise detection of positive selection. *Mol. Biol. Evol.* **29**, 1125–1139 (2012).
186. Duret, L. & Galtier, N. Biased gene conversion and the evolution of mammalian genomic landscapes. *Annu. Rev. Genomics Hum. Genet.* **10**, 285–311 (2009).
187. Galtier, N. & Duret, L. Adaptation or biased gene conversion? Extending the null hypothesis of molecular evolution. *Trends Genet.* **23**, 273–277 (2007).
188. Ratnakumar, A. *et al.* Detecting positive selection within genomes: The problem of biased gene conversion. *Philos. Trans. R. Soc. B Biol. Sci.* **365**, 2571–2580 (2010).
189. Guéguen, L. *et al.* Bio++: Efficient extensible libraries and tools for computational molecular evolution. *Mol. Biol. Evol.* **30**, 1745–1750 (2013).
190. Wickham, H. & Grolemund, G. *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data*. (O'Reilly Media, Inc. Canada, 2016).
191. Wilke, C. O. cowplot: streamlined plot theme and plot annotations for 'ggplot2.' *CRAN Repos* (2016).
192. Gouy, M., Guindon, S. & Gascuel, O. SeaView version 4: A multiplatform graphical user interface for sequence alignment and phylogenetic tree building. *Mol. Biol. Evol.* **27**, 221–224 (2010).
193. Redelings, B. Erasing errors due to alignment ambiguity when estimating positive selection. *Mol. Biol. Evol.* **31**, 1979–1993 (2014).
194. Mi, H., Muruganujan, A., Ebert, D., Huang, X. & Thomas, P. D. PANTHER version 14: More genomes, a new PANTHER GO-slim and improvements in enrichment analysis tools. *Nucleic Acids Res.* **47**, D419–D426 (2019).
195. Huerta-Cepas, J. *et al.* Fast genome-wide functional annotation through orthology assignment by eggNOG-mapper. *Mol. Biol. Evol.* **34**, 2115–2122 (2017).
196. Huerta-Cepas, J. *et al.* eggNOG 5.0: A hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res.* **47**, D309–D314 (2019).

⇧Back to summary⇧

– FRENCH SUMMARY –

La spécialisation alimentaire des mammifères qui se nourrissent exclusivement de fourmis et/ou de termites (appelés myrmecophagie) est l'un des exemples les plus célèbres de convergence évolutive. Ce mode de vie particulier est apparu dans cinq lignées distinctes de mammifères placentaires : l'oryctérope, le proteles, les fourmiliers, le tatou géant et les pangolins. Les pressions sélectives associées à la consommation de fourmis et de termites ont d'ailleurs conduit à des convergences morphologiques extrêmes au fil du temps. Dans ce contexte, en utilisant des approches de phylogénomique et de génomique comparative, l'objectif de mon projet de doctorat était de comprendre les processus moléculaires associés à la convergence vers la myrmécophagie chez les mammifères myrmécophages.

Dans le premier chapitre de ma thèse, j'ai présenté une stratégie visant a essayer, à partir données métagénomiques extraites d'échantillons fécaux, de caractériser le régime alimentaire des mammifères myrmecophages. Ceci représenté une alternative possible aux méthodes existantes basées sur des méthodes de metabarcoding (par exemple, Pompanon et al. 2012 ; Shehzad et al. 2012 ; Alberti et al. 2018 ; Galan et al. 2018, Gauthier et al. 2020). La première étape a consisté à collecter des échantillons de matières fécales et de proies potentielles (fourmis et termites) dans les aires de répartition des mammifères myrmécophages. Les expéditions de terrain pour cette étude se sont concentrées sur deux réserves spécifiques d'Afrique du Sud, dans lesquelles nous avons essayé de collecter de manière exhaustive toutes les espèces de fourmis et de termites rencontrées. L'objectif de cet échantillonnage était de construire une base de données spécifique à laquelle les séquences mitochondriales extraites des échantillons fécaux métagénomiques pourraient être comparées. Pour créer ces bases de données mitochondriales, nous avons développé MitoFinder, un pipeline facile à utiliser pour assembler, extraire et annoter efficacement les séquences mitochondriales à partir de données de séquençage à haut débit (Allio et al. 2020). Ce pipeline a prouvé son efficacité en extrayant avec succès les séquences mitochondriales de la plupart des échantillons de fourmis et de termites et nous a permis de créer deux bases de données, comprenant respectivement 87 et 222 individus de termites et de fourmis. En utilisant MitoFinder sur les données métagénomiques extraites des échantillons de féces, nous n'avons malheureusement pas pu extraire de contigs correspondant a des portions de mitogénomes de fourmis ou de termites. Cependant, en utilisant les contigs mitochondriaux de fourmis et de termites générés avec MitoFinder (à partir des fourmis et termites collectées sur le terrain) comme référence pour mapper les reads (issus du séquençage des féces), nous avons pu détecter quelques reads correspondant aux proies consommées par les espèces myrmécophages (en accord avec des études précédentes : Weyer 2018, Panaino 2020). Étant donné que seuls quelques reads ont été identifiés pour chaque espèce (de 2 à 30 reads), cette stratégie a permis d'obtenir une première évaluation moléculaire préliminaire du régime alimentaire de ces espèces, mais des analyses supplémentaires sont nécessaires. Le faible nombre de reads récupérés

dans les échantillons fécaux est probablement dû à la surreprésentation des fragments d'ADN bactériens dans les extractions d'échantillons fécaux. Dans ce contexte, nous prévoyons de profiter de la base de données de fourmis et de termites créée au cours de cette thèse pour concevoir des amorces spécifiques permettant de séquencer préférentiellement les fragments d'ADN des proies dans les échantillons fécaux de mammifères myrmécophages (Gauthier et al. 2020).

Le deuxième chapitre de cette thèse présentait le développement d'approches expérimentales et bioinformatiques pour générer des génomes de mammifères myrmécophages de bonne qualité à partir d'échantillons prélevés sur la route. Les espèces de mammifères myrmécophages présentent des génomes relativement grands, allant de 2,5 Gb pour les pangolins à 4,5 Gb pour les xénarthres. Pour pouvoir générer des assemblages de bonne qualité, tant en termes de contiguïté que de complétude, nous avons décidé de nous appuyer sur une stratégie d'assemblage hybride. Cette stratégie consiste à tirer parti à la fois de la grande précision des « short-reads » générés par les méthodes de séquençage de nouvelle génération et de la taille des « long-reads » générés avec des méthodes de séquençage de la troisième génération. En raison de la faible qualité de nos tissus, l'ADN extrait pour nos espèces était de trop faible qualité pour être accepté, il y a trois ans, par les plateformes de séquençage « long-reads ». Dans ce contexte, nous avons décidé de développer un protocole optimisé pour le séquençage de tissus prélevés sur des mammifères écrasés sur la route avec le séquenceur MinION (développé par Oxford Nanopore Technologies). En bref, ce protocole (disponible ici : dx.doi.org/10.17504/protocols.io.beixjcfn) consiste (i) à préserver les tissus dans de l'ARNlater au lieu d'utiliser de l'EtOH 95% traditionnel, (ii) à extraire de préférence les parties les mieux préservées du tissu et à éliminer toutes les impuretés perceptibles, et (iii) à ajuster le ratio de billes AMPure à 0,4x pour favoriser la sélection de fragments de grande taille lors de la construction de la librairie ONT. En appliquant ce protocole optimisé, nous avons pu générer des données de séquençage « long-reads » de bonne qualité pour les neuf espèces prévues dans le projet. Ensuite, les « short-reads » Illumina et les long-reads ONT ont été utilisés conjointement grâce à une approche d'assemblage hybride (implémentée dans MaSuRCA) qui a permis d'obtenir des assemblages de bonne qualité. Ces assemblages sont peu fragmentés avec un nombre de contigs allant de 51 157 à 4 309, ce qui est beaucoup moins que les assemblages précédemment disponibles pour les mammifères myrmécophages, et en particulier pour les xénarthres (Zoonomia Consortium 2020). De même, les analyses BUSCO (Waterhouse et al. 2018), qui estiment la complétude en gènes orthologues (selon une base de données préalableement définie), suggèrent un niveau élevé de complétude pour ces génomes. Il est intéressant de noter que, malgré leur bonne qualité, les génomes des xénarthres ne dépassent pas le score de 90 % de gènes complets. Ce résultat peut être dû à la difficulté d'assembler les régions génomiques contenant ces gènes.

Une fois les génomes séquencés et assemblés, l'étape suivante consistait à les annoter. Dans un premier temps, les éléments répétés ont été identifiés et masqués pour les étapes d'annotation suivantes. Ensuite, nous avons décidé de nous appuyer à la fois sur des prédictions de gènes basées sur la similarité des séquences avec des protéines connues et sur des predictions basées sur des modèles de gènes (Yandell & Ence 2012). Nous avons utilisé comme référence des données transcriptomiques assemblées et annotées prévues à cet effet (70 transcriptomes) et des bases de données de référence de gènes disponibles (uniprot/SWISSPROT). Les informations obtenues à partir de deux stratégies de prédiction de gènes ont été résumées avec le pipeline implémenté dans MAKER 3 (Yandell 2011). Pour améliorer la précision de l'annotation, ce pipeline a été exécuté trois fois de manière itérative en intercalant une étape d'entraineent de modèle de gène basée sur les résultats des annotations précédentes (Korf 2004, Stanke et al. 2006). Étant donné que les étapes de séquençage, de basecalling, d'assemblage et d'annotation du génome sont des processus relativement longs, seuls deux génomes générés au cours de mon projet de doctorat, les génomes de *Smutsia gigantea* et de *Myrmecophaga tridactyla*, ont été entièrement annotés. En effet, le séquençage de tous les génomes avec ONT a duré 22 mois. Ensuite, la conversion des informations brutes de séquençage en données analysable (fastq) a pris environ deux semaines par génome, suivie de l'étape d'assemblage hybride pour laquelle environ 3-4 semaines supplémentaires ont été nécessaires. Enfin, deux à trois semaines ont été nécessaires pour faire tourner le pipeline d'annotation sur chaque assemblage (en considérant toutes les données externes disponibles, par exemple les assemblages de transcriptome annotés).

Bien que ce pipeline représente un long processus, les génomes annotés qui en résultent constituent une ressource inestimable pour étudier la convergence évolutive chez les mammifères myrmécophages. En effet, grâce à l'annotation produite par nos analyses, il sera beaucoup plus facile d'extraire avec précision les gènes orthologues. Dans notre cas, nous prévoyons d'utiliser le même pipeline que celui développé pour créer la base de données OrthoMaM (Scornavacca et al. 2019) afin d'assembler un ensemble de données génomiques comprenant à la fois des espèces myrmécophages et non myrmécophages. La combinaison de nos génomes nouvellement générés et de quelques génomes sélectionnés dans OrthoMaM fournira un excellent jeu de données permettant de détecter des traces potentielles de convergences moléculaires associées à l'adaptation à la myrmécophagie dans des gènes orthologues à simple copie. Après avoir inféré un backbone phylogénétique solide en utilisant l'approche décrite dans cette thèse (Partie II - section 1), différentes approches seront utilisées pour étudier la convergence moléculaire chez les mammifères myrmécophages. En effet, un pipeline a été établi au cours de mon projet de doctorat par Mathilde Barthe, une étudiante en master que j'ai eu la chance de co-encadrer avec Frédéric Delsuc. Elle a travaillé sur un jeu de données composé de 12 espèces de Carnivora, dont le protèle (*Proteles cristatus*) et le renard à oreilles de chauve-souris (*Otocyon megalotis*), tous deux ayant un régime alimentaire composé à plus de 70% de fourmis et de termites. Elle a conçu un pipeline pour la recherche de convergences moléculaires dans les gènes basé

sur (i) des substitutions convergentes d'acides aminés (en utilisant le PCOC, Rey et al. 2018), (ii) des traces convergentes de sélection adaptative (analyse dN/dS, Yang 2007), et (iii) la similarité des taux d'évolution entre les lignées convergentes (RERconverge, Kowalczyk et al. 2018). Bien que ce projet ait été mené sur un jeu de données réduit, il a permis le développement d'un pipeline étudiant la convergence moléculaire à différentes échelles. De plus, de nombreux gènes ont été mis en évidence conjointement par les trois différentes approches mises en œuvre dans le pipeline, mais une grande partie de ces gènes étaient des faux positifs résultant de la mauvaise qualité de l'annotation des génomes (précédemment annoté avec genBlastG, She et al. 2011). Ces résultats nous ont encouragés à mettre en œuvre le pipeline d'annotation présenté dans le deuxième chapitre de cette thèse. Maintenant que nous avons des annotations plus précises, l'objectif sera d'appliquer le pipeline précédemment développé par Mathilde sur le jeu de données nouvellement généré. Enfin, je voudrais examiner la convergence moléculaire avec une autre approche intéressante développée par Wu et al. (2017). En bref, cette approche consiste à inférer les états ancestraux (traits d'histoire de vie tels que le régime alimentaire) aux nœuds des phylogénies à partir des associations observées entre les états des espèces actuelles et l'évolution de leurs gènes (résumé sous forme de profils d'évolution de gènes, correspondant à la vitesse d'évolution des gènes aux extrémités de la phylogénie). Ainsi, des profils de gènes spécifiques sont associés à chaque état existant. Ensuite, pour chaque nœud de la phylogénie, les taux d'évolution inférés sont comparés à chaque profil et l'état ancestral sélectionné lors de l'inférence est celui qui présente le profil le plus similaire à celui inféré pour le nœud de la phylogénie. En recherchant un profil génétique spécifique associé à un état donné (dans notre cas, spécialisation dans la consommation de fourmis et/ou de termites), cette approche pourrait nous aider à mettre en évidence une évolution convergente à l'échelle des gènes.

L'annotation complète de nos génomes nous permettra également d'étudier l'histoire évolutive d'autres portions potentiellement intéressantes des génomes. Par exemple, il a été démontré que des régions conservées non codantes, impliquées dans la régulation de l'expression des gènes, ont évolué de manière convergente chez des oiseaux paléognathes incapables de voler (Sackton et al. 2019). En utilisant un jeu de données génomique, Sackton et al. (2019), ont extrait environ 280 000 éléments non exoniques conservés ayant un rôle régulateur potentiel chez les oiseaux et d'autres taxons. Parmi eux, ils ont trouvé un grand nombre de portions ayant accéléré de manière convergente chez les paléognathes ayant perdu la faculté de vol. Ces résultats suggèrent que les processus d'évolution convergents peuvent impliquer des régions dans certains cas. Dans cette optique, Mathilde Barthe vient de commencer son projet de doctorat pour étudier l'évolution convergente dans les régions conservées non codantes des génomes de mammifères myrmécophages.

Un autre point intéréssant concerne l'étude de l'évolution des familles de gènes. En effet, de nombreuses études de génomique comparative se concentrent uniquement sur des gènes orthologues à simple copie. Cependant, ces gènes ne représentent qu'une petite partie des gènes codants que l'on trouve dans les génomes. Il est intéressant de noter que des familles de gènes candidats ont déjà été

signalées comme ayant un rôle central dans l'adaptation alimentaire, en particulier lors de l'évolution de mécanismes de détoxication (par exemple Berenbaum et al. 1996). En outre, chez les mammifères, les grandes familles de gènes des récepteurs gustatifs (TR) et olfactifs (OR) ont été étudiées depuis longtemps et semblent être impliquées dans les interactions sociales, l'alimentation et l'accouplement (Dulac & Torello 2003 ; Shi et al. 2003 ; Bachmanov & Beauchamp 2007 ; Hayden & Teeling 2014 ; Rymer 2020). Dans l'ensemble, le développement de méthodes de séquençage récentes permet désormais d'étudier plus précisément l'évolution des familles de gènes (par exemple Hayden et al. 2014 ; Edger et al. 2015 ; Yohe et al. 2019 ; Thomas et al. 2020). Ainsi, l'annotation complète de nos génomes nous aidera à reconstruire l'évolution des familles de gènes chez les mammifères myrmécophages. Dans ce contexte, Sophie Teullet vient de commencer son projet de doctorat dans lequel elle se concentrera sur l'évolution des familles de gènes TR et OR chez les mammifères avec un intérêt particulier pour la convergence chez les espèces myrmécophages.

Enfin, dans le troisième chapitre de cette thèse, j'ai présenté des analyses de transcriptomique comparative réalisées, entre autre, sur des glandes salivaires de mammifères myrmécophages. En effet, parmi les nombreuses convergences morphologiques observées chez les mammifères myrmécophages, l'hypertophie des glandes salivaires est particulièrement remarquable. Les glandes salivaires des mammifères myrmécophages jouent probablement un rôle important dans la digestion des insectes, comme le suggère la forte expression d'enzymes digestives dans cet organe chez le pangolin malais (Ma et al. 2017). Dans ce troisième chapitre, j'ai d'abord présenté une analyse globale des transcriptomes des glandes salivaires de 24 espèces (28 individus) comprenant à la fois des mammifères non myrmécophages et myrmécophages. Les transcriptomes ont été annotés et le niveau d'expression des orthogroupes de transcrits a été comparé entre les différents groupes taxonomiques et les differents groupes représentant les régimes alimentaires. Dans l'ensemble, les analyses préliminaires suggèrent que l'expression globale des gènes dans les glandes salivaires des mammifères est principalement déterminée par l'histoire évolutive des espèces (phylogénie). En effet, les espèces étroitement apparentées ont des profils d'expression plus similaires que les espèces plus éloignées. Ce résultat est attendu et a déjà été signalé dans des études antérieures sur de multiples organes de mammifères (par exemple, Brawand et al., 2011). Cela suggère un impact important de la contingence historique sur l'expression génétique. Il est intéressant de noter qu'un exemple de l'effet de l'histoire de la contingence réside dans l'évolution enzymes digestives appartenant à la famille des chitinases. Cette famille est composée de cinq gènes de chitinase paralogues (CHIA1-5) et Emerling et al. (2018) ont trouvé une corrélation positive entre le nombre de copies de gènes fonctionnels (non pseudogénéisés) des gènes de chitinase et le pourcentage du régime alimentaire composé d'invertébrés chez les mammifères placentaires. Il est intéressant de noter que le tamandua (*Tamandua tetradactyla*) et l'oryctérope (*Orycteropus afer*) possèdent quatre et cinq copies fonctionnelles de

chitinase alors que le pangolin (*Manis javanica*) n'en possède qu'une seule. Cet exemple illustre l'impact de la contingence historique dans l'évolution. En effet, le pangolin ne possède qu'un seul gène de chitinase fonctionnel (CHIA5) probablement parce que l'ancêtre commun des Pholidata et des Carnivora avait déjà perdu les quatre autres gènes de chitinase. C'est pourquoi nous avons décidé de nous concentrer sur l'expression des différentes copies de la chitinases, d'abord dans les glandes salivaires des mammifères, puis plus spécifiquement dans des organes digestifs et non digestifs de mammifères myrmécophages. Nous avons constaté que malgré des répertoires de gènes de chitinases différents, les espèces myrmécophages (*Manis javanica* et *Tamandua tetradactyla*) expriment fortement leurs gènes de chitinase dans les organes digestifs. En particulier, nous avons pu montrer que le pangolin malais compense potentiellement la perte de gènes de chitinase fonctionnels en surexprimant CHIA5 dans tous les principaux organes digestifs (glandes salivaires, langue, estomac, pancréas, gros intestin et foie, Ma et al., 2019). Ces résultats montrent l'importance de la contingence historique dans le façonnement de l'évolution moléculaire des organismes. Néanmoins, la surexpression de son dernier gène de chitinase disponible par le pangolin malais fournit un excellent exemple d'évolution adaptative pour contrer l'effet de la contingence historique et evoluer vers la myrmécophagie.

Pour conclure, les différentes approches développées au cours de mon projet de doctorat, depuis le séquençage des tissus dégradés à l'assemblage et l'annotation des génomes d'organismes non modèles, nous ont permis de générer neuf génomes de mammifères de grande qualité. Ces génomes constituent une ressource inestimable pour étudier la convergence évolutive des mammifères myrmécophages. En combinant les gènes extraits de ces génomes avec les bases de données de gènes de mammifères disponibles, nous serons en mesure de procéder à la détection de la convergence moléculaire à différents niveaux. De plus, l'exemple de l'évolution de la famille des chitinases présenté ici rejoint les nombreux exemples montrant l'impact de la contingence historique dans l'évolution des organismes. Tout ceci suggère que différentes voies d'évolution peuvent être suivies pour s'adapter à des conditions similaires. Dans cette optique, nous prévoyons de profiter de l'annotation complète des génomes générée au cours du projet pour combiner différentes approches de détection afin d'étudier la convergence évolutive vers la myrmécophagie.

References

⇧Back to summary⇧

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 Resour* 1755–0998.13160. doi:10.1111/1755-0998.13160

Bachmanov AA, Beauchamp GK. 2007. Taste receptor genes. *Annu Rev Nutr* **27**:389–414. doi:10.1146/annurev.nutr.26.061505.111329

Berenbaum MR, Favret C, Schuler MA. 1996. On defining "Key Innovations" in an adaptive radiation: Cytochrome P450S and Papilionidae. *Am Nat* **148**:S139–S155. doi:10.1086/285907

Brawand D, Soumillon M, Necsulea A, Julien P, Csárdi G, Harrigan P, Weier M, Liechti A, Aximu-Petri A, Kircher M, Albert FW, Zeller U, Khaitovich P, Grützner F, Bergmann S, Nielsen R, Pääbo S, Kaessmann H. 2011. The evolution of gene expression levels in mammalian organs. *Nature* **478**:343–348. doi:10.1038/nature10532

Dulac C, Torello AT. 2003. Molecular detection of pheromone signals in mammals: from genes to behaviour. *Nat Rev Neurosci* **4**:551–562. doi:10.1038/nrn1140

Emerling CA, Delsuc F, Nachman MW. 2018. Chitinase genes (*CHIA* s) provide genomic footprints of a post-Cretaceous dietary radiation in placental mammals. *Sci Adv* **4**:eaar6478. doi:10.1126/sciadv.aar6478

Galan M, Pons J-B, Tournayre O, Pierre É, Leuchtmann M, Pontier D, Charbonnel N. 2018. Metabarcoding for the parallel identification of several hundred predators and their prey: Application to bat species diet analysis. *Mol Ecol Resour* **18**:474–489. doi:10.1111/1755-0998.12749

Gauthier M, Konecny-Dupré L, Nguyen A, Elbrecht V, Datry T, Douady C, Lefébure T. 2020. Enhancing DNA metabarcoding performance and applicability with bait capture enrichment and DNA from conservative ethanol. *Mol Ecol Resour* **20**:79–96. doi:10.1111/1755-0998.13088

Hayden S, Bekaert M, Goodbla A, Murphy WJ, Dávalos LM, Teeling EC. 2014. A cluster of olfactory receptor genes linked to frugivory in bats. *Mol Biol Evol* **31**:917–927. doi:10.1093/molbev/msu043

Holt C, Yandell M. 2011. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics* **12**:491. doi:10.1186/1471-2105-12-491

Kowalczyk A, Meyer WK, Partha R, Mao W, Clark NL, Chikina M. 2019. RERconverge: an R package for associating evolutionary rates with convergent traits. *Bioinformatics* **35**:4815–4817. doi:10.1093/bioinformatics/btz468

Ma J-E, Jiang H-Y, Li L-M, Zhang X-J, Li H-M, Li G-Y, Mo D-Y, Chen J-P. 2019. SMRT sequencing of the full-length transcriptome of the Sunda pangolin (*Manis javanica*). *Gene* **692**:208–216. doi:10.1016/J.GENE.2019.01.008

Ma J-E, Li L-M, Jiang H-Y, Zhang X-J, Li J, Li G-Y, Yuan L-H, Wu J, Chen J-P. 2017. Transcriptomic analysis identifies genes and pathways related to myrmecophagy in the Malayan pangolin (*Manis javanica*). *PeerJ* **5**:e4140. doi:10.7717/peerj.4140

Panaino W. 2020. Diet, activity, and body temperature patterns of ground pangolins in a semi-arid environment. University of the Witwatersrand, Johannesburg.

Pompanon F, Deagle BE, Symondson WOC, Brown DS, Jarman SN, Taberlet P. 2012. Who is eating what: diet assessment using next generation sequencing. *Mol Ecol* **21**:1931–1950. doi:10.1111/j.1365-294X.2011.05403.x

Rey C, Guéguen L, Sémon M, Boussau B. 2018. Accurate detection of convergent amino-acid evolution with PCOC. *Mol Biol Evol* **35**:2296–2306. doi:10.1093/molbev/msy114

Rymer TL. 2020. The role of olfactory genes in the expression of rodent paternal care behavior. *Genes (Basel)* **11**:292. doi:10.3390/genes11030292

Sackton TB, Clark N. 2019. Convergent evolution in the genomics era: new insights and directions. *Philos Trans R Soc B Biol Sci* **374**:20190102. doi:10.1098/rstb.2019.0102

Scornavacca C, Belkhir K, Lopez J, Dernat R, Delsuc F, Douzery EJP, Ranwez V. 2019. OrthoMaM v10: Scaling-up orthologous coding sequence and exon alignments with more than one hundred mammalian genomes. *Mol Biol Evol* **36**:861–862. doi:10.1093/molbev/msz015

She R, Chu JS-C, Uyar B, Wang J, Wang K, Chen N. 2011. genBlastG: using BLAST searches to build homologous gene models. *Bioinformatics* **27**:2141–2143. doi:10.1093/bioinformatics/btr342

Shehzad W, Riaz T, Nawaz MA, Miquel C, Poillot C, Shah SA, Pompanon F, Coissac E, Taberlet P. 2012. Carnivore diet analysis based on next-generation sequencing: application to the leopard cat (*Prionailurus bengalensis*) in Pakistan. *Mol Ecol* **21**:1951–1965. doi:10.1111/j.1365-294X.2011.05424.x

Shi P, Zhang J, Yang H, Zhang Y. 2003. Adaptive diversification of bitter taste receptor genes in mammalian evolution. *Mol Biol Evol* **20**:805–814. doi:10.1093/molbev/msg083

Waterhouse RM, Seppey M, Simão FA, Manni M, Ioannidis P, Klioutchnikov G, Kriventseva E V, Zdobnov EM. 2018. BUSCO applications from quality assessments to gene prediction and phylogenomics. *Mol Biol Evol* **35**:543–548. doi:10.1093/molbev/msx319

Weyer NM. 2018. Physiological flexibility of free-living aardvarks (Orycteropus afer) in response to environmental fluctuations. University of the Witwatersrand, Johannesburg, South Africa.

Wu J, Yonezawa T, Kishino H. 2017. Rates of molecular evolution suggest natural history of life history traits and a Post-K-Pg nocturnal bottleneck of placentals. *Curr Biol* **27**:3025–3033. doi:10.1016/J.CUB.2017.08.043

Yandell M, Ence D. 2012. A beginner's guide to eukaryotic genome annotation. *Nat Rev Genet* **13**:329–342. doi:10.1038/nrg3174

Yang F, Sun J, Luo H, Ren H, Zhou H, Lin Y, Han M, Chen B, Liao H, Brix S, Li J, Yang H, Kristiansen K, Zhong H. 2020. Assessment of fecal DNA extraction protocols for metagenomic studies. *Gigascience* **9**. doi:10.1093/gigascience/giaa071

Yang Z. 2007. PAML 4: Phylogenetic analysis by maximum likelihood. *Mol Biol Evol* **24**:1586–1591. doi:10.1093/molbev/msm088

Zoonomia consortium. 2020. A comparative genomics multitool for scientific discovery and conservation. *Nature* **587**:240–245. doi:10.1038/s41586-020-2876-6

⇧Back to summary⇧

Abstract

The phenomenon of evolutionary convergence is a fascinating process in which distantly related species independently acquire similar characteristics in response to similar selective pressures. Ant- and termite-eating mammals are among the most famous examples of morphological convergence. Indeed, this particular lifestyle evolved in five distinct lineages of mammals: the aardvark (Tubulidentata), the aardwolf (Carnivora), the anteaters (Pilosa), the giant armadillo (Cingulata), and the pangolins (Pholidota). To better undestand the evolution of these organisms, several approaches were developed in this thesis. First, I present an original strategy to characterize the precise diet of myrmecophagous mammals taking advantage of metagenomic sequencing data generated from fecal samples and a reference mitogenomic database of termites and ants. Second, with the final objective of detecting molecular convergence at the genomic scale in ant-eating mammals, we generated nine high quality mammlian genomes using Oxford Nanopore technologies. The different strategies developed from the set-up of MinION qesuencing to annotation of the resulting assemblies are presented together with a first case study illustrating the use of two of these new reference genomes for species delineation. Finally, I present comparative transcriptomic analyses of salivary glands and other organs in ant-eating mammals suggesting that historical contingency and molecular evolutionary tinkering of chitinase genes played a major role in the convergent evolution of myrmecophagy.

Résumé

Le phénomène de convergence évolutive est un processus fascinant dans lequel des espèces phylogénétiquement éloignées acquièrent indépendamment des caractéristiques similaires en réponse à des pressions de sélection similaires. Les mammifères myrmécophages figurent parmi les exemples les plus célèbres de convergence morphologique. En effet, ce mode de vie particulier a évolué chez cinq lignées distinctes de placentaires : l'oryctérope (Tubulidentata), le protèle (Carnivora), les fourmiliers (Xenarthra), le tatou géant (Cingulata) et les pangolins (Pholidota). Pour mieux comprendre l'évolution de ces organismes, plusieurs approches ont été développées dans cette thèse. Tout d'abord, je présente une stratégie originale pour caractériser le régime alimentaire précis des mammifères myrmécophages en tirant parti des données métagénomiques générées à partir d'échantillons fécaux et d'une base de données mitogénomique de référence sur les termites et les fourmis. Ensuite, avec l'objectif final de détecter la convergence moléculaire à l'échelle génomique chez les mammifères myrmécophages, nous avons généré neuf génomes de mammifères de qualité en utilisant la technologie Oxford Nanopore. Les différentes stratégies développées depuis la mise en place du séquençage MinION jusqu'à l'annotation finale des assemblages sont présentées avec une première étude de cas illustrant l'utilisation de deux de ces génomes de référence pour la délimitation des espèces. Enfin, je présente une analyse de transcriptomique comparative des glandes salivaires et d'autres organes chez les mammifères myrmécophages qui suggère que la contingence historique et le bricolage moléculaire des gènes de chitinase ont joué un rôle majeur dans l'évolution convergente de la myrmécophagie.