LUND UNIVERSITY

Looking beyond the reference genomes

Integrating sparse and dense datasets for robust phylogenetic inference of species-rich groups of moths

Yapar, Etka

2025

Document Version:
Publisher's PDF, also known as Version of record

Link to publication

Citation for published version (APA):
Yapar, E. (2025). *Looking beyond the reference genomes: Integrating sparse and dense datasets for robust phylogenetic inference of species-rich groups of moths*. [Doctoral Thesis (compilation), Faculty of Science]. Media-Tryck, Lund University, Sweden.

Total number of authors:
1

General rights
Unless other specific re-use rights are stated the following general rights apply:
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.
• Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
• You may not further distribute the material or use it for any profit-making activity or commercial gain
• You may freely distribute the URL identifying the publication in the public portal

Read more about Creative commons licenses: https://creativecommons.org/licenses/

Take down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

LUND UNIVERSITY

PO Box 117
221 00 Lund
+46 46-222 00 00

Download date: 14. Jan. 2026

Looking beyond the reference genomes

Integrating sparse and dense datasets for robust phylogenetic inference of species-rich groups of moths

ETKA YAPAR
DEPARTMENT OF BIOLOGY | FACULTY OF SCIENCE | LUND UNIVERSITY

Looking beyond the reference genomes

Looking beyond the reference genomes

Integrating sparse and dense datasets for robust phylogenetic inference of species-rich groups of moths

by Etka Yapar

LUND
UNIVERSITY

Thesis for the degree of Doctor of Philosophy
Thesis advisors: Niklas Wahlberg, Jadranka Rota, Mikael Pontarp, Krzysztof Bartoszek
Faculty opponent: Martin Irestedt

To be presented, with the permission of the Faculty of Science of Lund University, for public criticism in the Blue hall (Blåhallen) at the Department of Biology on Friday, the 6th of February 2026 at 09:00.

Organization
LUND UNIVERSITY

Document name
DOCTORAL DISSERTATION

Date of disputation
2026-02-06

Sponsoring organization

Author
Etka Yapar

Title and subtitle
Looking beyond the reference genomes: Integrating sparse and dense datasets for robust phylogenetic inference of species-rich groups of moths

Abstract

Phylogenies are central to many disciplines in biology. While disciplines such as systematics are focused on phylogenetic hypotheses themselves, other disciplines depend on robust and well-supported phylogenies in order to put biological data into evolutionary context through phylogenetic comparative studies. Examples of such fields are macroevolutionary research e.g. studying the evolution of a focal character across a phylogeny; comparative genomics and transcriptomics, in which the studied phenotype is the genomic features themselves (e.g. gene copy number variation, gene expression differences); as well as other disciplines that may be interested in e.g. allometry but need to take into account the non-independent nature of the data resulting from shared ancestry. We are certainly living in the genomic era for phylogenetic inference, and this is more true for some groups of organisms than for others. Lepidoptera (butterflies and moths), with the approximately 160,000 described species, comprise about 11% of all the animal species on earth and are one of the groups with a very large number of reference genomes. There are now more than 1,200 species of Lepidoptera with such high-quality genomes. This makes them the best sampled order of animals with reference genomes and thus it presents an unprecedented opportunity for phylogenomics of Lepidoptera. When the aim is to infer the best possible phylogeny, taxon sampling is of utmost importance, and thus these reference genomes need to be complemented with other types of data to achieve that. Luckily, there are a lot of additional data available in the form of a rich collection of contig-level genome assemblies, transcriptomic raw data, and maybe the richest of them all, the tens of thousands of species that have been sequenced by low throughput methods in previous phylogenetic studies. Although promising, this data integration approach brings with it some challenges such as fragmented or wrongly identified gene sequences and errors in orthology assessment. Throughout the first four chapters of this thesis, I formulate the necessary steps to efficiently integrate these additional types of already available data with reference genomes, and apply this methodology to infer robust phylogenies for four species-rich groups (three families and a superfamily) of Lepidoptera: Gelechioidea, Geometridae, Noctuidae, and Erebidae; and develop a reproducible phylogenetic data preprocessing pipeline that makes this approach available for use. Then, in the final chapter, I demonstrate the power of such robust phylogenies in evolutionary biology by studying the macroevolution of a key adaptive trait against harsh environmental conditions, the ability to diapause (i.e. to suppress development or reproduction), across butterflies (superfamily Papilionoidea) by leveraging a robust time-calibrated phylogeny inferred earlier for the group.

Key words
Phylogenetics, Geometridae, Noctuidae, Gelechioidea, Erebidae, Papilionoidea, Butterflies, Evolution

Language
English

ISBN
978-91-8104-796-7 (print)
978-91-8104-797-4 (pdf)

Number of pages
70

I, the undersigned, being the copyright owner of the abstract of the above-mentioned dissertation, hereby grant to all reference sources the permission to publish and disseminate the abstract of the above-mentioned dissertation.

Signature _____ Date 2025-12-05

DOKUMENTDATABLAD enl SIS 61 41 21

Looking beyond the reference genomes

Integrating sparse and dense datasets for robust phylogenetic inference of species-rich groups of moths

by Etka Yapar

LUND
UNIVERSITY

Cover illustration front: A macro photo of *Junonia orithya* wing.
Cropped from the original image. Original image credits: Rudolph Steenkamp on iNaturalist, Photo ID:479228755, License: CC-BY-SA.

Funding information: The thesis work was financially supported by ELLIIT – The Excellence center in Linköping-Lund In Information Technology

Copyright pp. i–70 Etka Yapar 2026
Paper I © 2025 Royal Entomological Society (Published by Wiley)
Paper II © 2025 Royal Entomological Society (Published by Wiley)
Paper III © 2025 The authors. Published by Wiley
Paper IV © 2025 The authors (Manuscript)
Paper V © 2024 The authors. Published by Oxford University Press

Faculty of Science, Department of Biology

ISBN: 978-91-8104-796-7 (print)
ISBN: 978-91-8104-797-4 (pdf)

Printed in Sweden by Media-Tryck, Lund University, Lund 2026

Media-Tryck is a Nordic Swan Ecolabel certified provider of printed material. Read more about our environmental work at www.mediatryck.lu.se

MADE IN SWEDEN

Scientia Dux Vitae Certissimus.

Contents

Abstract

Phylogenies are central to many disciplines in biology. While disciplines such as systematics are focused on phylogenetic hypotheses themselves, other disciplines depend on robust and well-supported phylogenies in order to put biological data into evolutionary context through phylogenetic comparative studies. Examples of such fields are macroevolutionary research e.g. studying the evolution of a focal character across a phylogeny; comparative genomics and transcriptomics, in which the studied phenotype is the genomic features themselves (e.g. gene copy number variation, gene expression differences); as well as other disciplines that may be interested in e.g. allometry but need to take into account the non-independent nature of the data resulting from shared ancestry. We are certainly living in the genomic era for phylogenetic inference, and this is more true for some groups of organisms than for others. Lepidoptera (butterflies and moths), with the approximately 160,000 described species, comprise about 11% of all the animal species on earth and are one of the groups with a very large number of reference genomes. There are now more than 1,200 species of Lepidoptera with such high-quality genomes. This makes them the best sampled order of animals with reference genomes and thus it presents an unprecedented opportunity for phylogenomics of Lepidoptera. When the aim is to infer the best possible phylogeny, taxon sampling is of utmost importance, and thus these reference genomes need to be complemented with other types of data to achieve that. Luckily, there are a lot of additional data available in the form of a rich collection of contig-level genome assemblies, transcriptomic raw data, and maybe the richest of them all, the tens of thousands of species that have been sequenced by low throughput methods in previous phylogenetic studies. Although promising, this data integration approach brings with it some challenges such as fragmented or wrongly identified gene sequences and errors in orthology assessment. Throughout the first four chapters of this thesis, I formulate the necessary steps to efficiently integrate these additional types of already available data with reference genomes, and apply this methodology to infer robust phylogenies for four species-rich groups (three families and a superfamily) of Lepidoptera: Gelechioidea, Geometridae, Noctuidae, and Erebidae; and develop a reproducible phylogenetic data preprocessing pipeline that makes this approach available for use. Then, in the final chapter, I demonstrate the power of such robust phylogenies in evolutionary biology by studying the macroevolution of a key adaptive trait against harsh environmental conditions, the ability to diapause (i.e. to suppress development or reproduction), across butterflies (superfamily Papilionoidea) by leveraging a robust time-calibrated phylogeny inferred earlier for the group.

Popular summary in English

What comes to your mind when you hear the term Evolution?

I suspect that a non-negligible portion of you would think about evolutionary trees, also known as phylogenetic trees. The most common way a phylogenetic tree is inferred today is using DNA sequences. These are brought together from a group of organisms and analysed to find differences among them that can tell who is more closely related to whom, and the end result is represented as a tree. Nowadays, it is getting more common to use the publicly available reference genomes (an end-to-end sequence of an organism's DNA across all its chromosomes) to obtain the gene sequences that will be used to infer the phylogenetic tree for a group of organisms. That is, if the group of organisms you are interested in has a good number and representative distribution of reference genomes available. Fortunately, the insects I have been working with for my thesis —butterflies and moths, or Lepidoptera—are exceptionally good in this regard. However, this may not always be the case for specific subgroups within Lepidoptera, and even if the reference genomes can be somewhat adequately representative, they are usually not enough on their own as we need to include as many members of this group as possible, or in other words, they do not result in the best possible "taxon sampling". A number of available lower-quality (or less complete) genome assemblies, and other kinds of —maybe older—genomic data can be extremely useful in this case to improve the taxon sampling problem. Moreover, there are additional tens of thousands of moth or butterfly species with available gene sequences (for some tens of genes) available because they have been included in earlier phylogenetic studies.

However, this mixing of genetic data can cause some technical problems. First, when we bioinformatically identify and extract the genetic regions from less-complete genomes, sometimes we extract another gene that is roughly similar in sequence but not evolutionarily related as the true copy we were looking for would be. Mistakes like this, and more fundamental errors that can happen when we are deciding on the specific set of the gene sequences we want to look for in the first place can add up. This is important because the group of genes we select to use for inferring trees ideally should be homologous —they should be truly evolutionarily related and should have evolved from a "common ancestor" gene. In an already potentially sparse dataset, these errors can generate noise that can, at times, overpower the phylogenetic signal —the real, and reliable differences in genes that reflect how these organisms evolved from their common ancestor.

In papers I-IV of my thesis, I first come up with a bioinformatic workflow for creating reliable phylogenetic datasets from the type of "mosaic" data I explained above and then I apply this method to infer well-supported phylogenetic trees from four highly diverse groups in Lepidoptera by combining reference genomes and sparser data sources. These four groups were three moth families named Geometridae, Noctuidae, and Erebidae, and

a moth superfamily (a group of families) named Gelechioidea. In the final paper, I, together with my colleagues, go beyond obtaining these trees to using them for addressing evolutionary questions. Specifically, we focused on butterflies (a separate, fifth group in Lepidoptera) and their ability to enter a dormant phase called diapause, in which they can survive the harsh winter. Thanks to modern phylogenetic comparative methods, we modelled the evolution of this trait, if a given species enters diapause at all and if yes in which life stage (as adults, pupae, larvae, eggs), across their phylogenetic tree and we drew conclusions about the likely origins —in terms of when and in which groups of butterflies we see these different strategies first evolved.

Taken together, my thesis provides much-improved phylogenetic trees (and in many cases, the first one with genome data) for four groups of Lepidoptera; it makes the bioinformatic pipeline I used for these four groups available as a reproducible data analysis pipeline so that other researchers can also opt for this data combination strategy to improve their taxon sampling; and it demonstrates the utility of such improved and comprehensive phylogenetic trees by using a published such tree for butterflies to study the evolution of an important adaptive trait.

Populärvetenskaplig sammanfattning på svenska

Vad tänker du på när du hör termen Evolution?

Jag misstänker att en icke försumbar del av er skulle tänka på evolutionära träd, även kända som fylogenetiska träd. Det vanligaste sättet att härleda ett fylogenetiskt träd idag är att använda DNA-sekvenser. Dessa samlas in från en grupp organismer och analyseras för att hitta skillnader mellan dem som kan visa vilka som är närmare släkt med varandra, och slutresultatet representeras som ett träd. Numera blir det allt vanligare att använda offentligt tillgängliga referensgenom (en komplett sekvens av en organisms DNA över alla dess kromosomer) för att få fram de gensekvenser som ska användas för att härleda fylogenetiska träd för en grupp organismer. Detta gäller dock endast om den grupp organismer du är intresserad av har ett bra antal och en representativ fördelning av referensgenom tillgängliga. Lyckligtvis är de insekter jag har arbetat med i min avhandling — fjärilar och malar, eller Lepidoptera — exceptionellt bra i detta avseende. Detta är dock kanske inte alltid fallet för specifika undergrupper inom Lepidoptera, och även om referensgenomen kan vara någorlunda representativa är de vanligtvis inte tillräckliga i sig själva för att inkludera så många medlemmar av denna grupp som möjligt, eller med andra ord, de resulterar inte i den bästa möjliga "taxon sampling". Ett antal tillgängliga genomkarteringar av lägre kvalitet (eller mindre kompletta), och andra typer av — kanske äldre — genomiska data kan vara extremt användbara i detta fall för att förbättra problemet med taxon sampling. Dessutom finns det ytterligare tusentals arter av fjärilar eller malar med tillgängliga gensekvenser (för några tiotal gener) eftersom de har inkluderats i tidigare fylogenetiska studier.

Men denna blandning av genetiska data kan orsaka vissa tekniska problem. För det första, när vi bioinformatiskt identifierar och extraherar de genetiska regionerna från mindre kompletta genom, händer det ibland att vi extraherar en annan gen som är ungefär lik i sekvens men inte evolutionärt besläktad med den riktiga kopian vi letade efter. Misstag som detta, och mer grundläggande fel som kan uppstå när vi bestämmer vilka specifika gensekvenser vi vill leta efter från början, kan lägga ihop sig. Detta är viktigt eftersom den grupp gener vi väljer för att härleda träd idealiskt bör vara homologa — de bör vara verkligt evolutionärt besläktade och ha utvecklats från en "gemensam anfader"-gen. I ett redan potentiellt glest dataset kan dessa fel skapa brus som ibland kan överrösta den fylogenetiska signalen — de verkliga och tillförlitliga skillnaderna i gener som återspeglar hur dessa organismer utvecklats från sin gemensamma anfader.

I de första fyra artiklarna i min avhandling utvecklar jag först ett bioinformatiskt arbetsflöde för att skapa tillförlitliga fylogenetiska dataset från den typ av "mosaik"-data jag förklarat ovan och använder denna metod för att härleda välstödda fylogenetiska träd från fyra mycket diversifierade grupper inom Lepidoptera genom att kombinera referensgenom och glesare datakällor. Dessa fyra grupper var tre familjer av malar kallade Geometridae,

Noctuidae och Erebidae, samt en superfamilj av malar (en grupp av familjer) kallad Gelechioidea. I den sista artikeln går jag, tillsammans med mina kollegor, bortom att bara få fram dessa träd och använder dem för evolutionära frågor. Specifikt fokuserade vi på fjärilar (en separat, femte grupp inom Lepidoptera) och deras förmåga att gå in i en vilofas kallad *diapause* där de kan överleva den hårda vintern. Tack vare moderna fylogenetiska komparativa metoder modellerade vi utvecklingen av denna egenskap, om en given art går in i diapause över huvud taget och om ja, i vilket livsstadium (som vuxna, puppor, larver, ägg) de gör det, över deras fylogenetiska träd och drog slutsatser om de sannolika ursprungen — i termer av när och i vilka grupper av fjärilar vi först ser dessa olika strategier utvecklas.

Sammantaget tillhandahåller min avhandling mycket förbättrade fylogenetiska träd (och i många fall det första med genomdata) för fyra grupper av Lepidoptera; den gör det bioinformatiska arbetsflöde jag använde för dessa fyra grupper tillgängligt som en reproducerbar dataanalytisk pipeline så att andra forskare också kan välja denna strategi för datakombination för att förbättra sin taxon sampling; och den demonstrerar nyttan av sådana förbättrade och omfattande fylogenetiska träd genom att använda ett publicerat sådant träd för fjärilar för att studera utvecklingen av en viktig adaptiv egenskap.

Çalışmanın Türkçe popüler bilim özeti

Evrim denince aklınıza ne geliyor?

Birçoğunuzun zihninde büyük olasılıkla evrimsel ağaçlar, yani filogenetik ağaçlar canlanıyordur. Günümüzde bu ağaçlar çoğunlukla DNA dizileri kullanılarak oluşturuluyor. Farklı canlılardan elde edilen DNA dizileri bir araya getiriliyor, aralarındaki farklılıklar analiz edilerek hangi türlerin birbirine daha yakın akraba olduğu belirleniyor ve sonuç bir ağaç şeklinde ifade ediliyor. Son yıllarda, bu tür analizlerde kullanılacak gen dizilerini elde etmek için yaygınlaşan yöntemlerden biri, kamuya açık referans genomlarını kullanmak. Referans genomu, bir organizmanın tüm kromozomlarını kapsayan eksiksiz DNA dizisi anlamına geliyor. Eğer üzerinde çalıştığınız canlı grubuna ait yeterli ve temsili sayıda referans genomu mevcutsa, bu yaklaşım oldukça kullanışlı olabiliyor.

Neyse ki tez çalışmam boyunca incelemekte olduğum böcekler —kelebekler ve güveler, yani Lepidoptera— bu açıdan oldukça şanslı bir grup. Ancak Lepidoptera'nın bazı alt gruplarında durum her zaman bu kadar ideal olmayabiliyor. Dahası, referans genomları temsili olsa bile, o gruptan mümkün olduğunca çok türü incelemeye katmak için genellikle tek başlarına yeterli olmuyorlar. Başka bir ifadeyle, en iyi "takson örneklemini" her zaman sağlayamayabiliyorlar. Tam da bu noktada, daha düşük kalitede (veya daha eksik) genom dizilemeleri ve hatta —belki daha eski— farklı türde genomik veriler, örnekleme kapsamını genişletmek için son derece değerli hale geliyor. Ayrıca geçmiş filogenetik çalışmalardan kalan, az sayıda gen için de olsa dizileri bilinen binlerce ek güve ve kelebek türü de mevcut.

Fakat tüm bu farklı genetik veri türlerini bir araya getirmek bazı teknik sorunları beraberinde getirebiliyor. Örneğin, daha eksik genomlardan biyoinformatik yöntemlerle gen bölgelerini bulup çıkarmaya çalıştığımızda, bazen dizi olarak benzer olsa da aslında aradığımız genle gerçek bir evrimsel akrabalığı olmayan farklı bir gen yanlışlıkla seçilebiliyor. Bu tür hatalar, ayrıca hangi gen kümelerinin analiz için seçileceğine karar verilirken yapılabilecek temel hatalar bir araya gelerek önemli problemlere yol açabiliyor. Bu çok kritik çünkü filogenetik ağaç oluşturmak için kullandığımız genlerin gerçekten homolog olması, yani aynı "ortak ata" geninden türemiş olmaları gerekiyor. Veri eksikliği riski taşıyan bir analizde bu tür hatalar, kimi zaman gerçek evrimsel sinyali —yani türlerin ortak atalarından nasıl ayrıştığını gösteren güvenilir genetik izleri— bastıran bir gürültü oluşturabiliyor.

Tezimin ilk dört makalesinde, bu tür "karma" verilerden güvenilir filogenetik veri kümeleri oluşturmak için bir biyoinformatik çalışma akışı geliştirdim ve bu yöntemi, referans genomlarını daha seyrek veri kaynaklarıyla birleştirerek Lepidoptera'daki dört çok çeşitli gruba ait ve oldukça iyileştirilmiş filogenetik ağaçlar çıkarmak için uyguladım. Bu dört grup; Geometridae, Noctuidae ve Erebidae adlı üç güve familyasi ile Gelechioidea adlı bir güve üst familyasından oluşuyor. Tezdeki son makalede ise meslektaşlarımla birlikte bu ağaçları sadece oluşturmakla kalmayıp, evrimsel bir soruyu cevaplamak için kullanmaya

odaklandık. Bu kez Lepidoptera'nın ayrı bir grubu olan kelebekleri inceledik ve onların kış koşullarını atlatmalarını sağlayan *diapoz* adı verilen bir dormant evreye girme yeteneklerini ele aldık. Modern filogenetik karşılaştırmalı yöntemler sayesinde, bu özelliğin nasıl evrimleştiğini modelledik; türlerin diapoza girip girmediğini, giriyorsa yaşam döngüsünün hangi evresinde (ergin, pupa, larva, yumurta) bunu yaptığını filogenetik ağaç üzerinde inceleyerek, bu farklı stratejilerin kelebeklerde ilk ne zaman ve hangi gruplarda ortaya çıktığına dair çıkarımlar yaptık.

Genel olarak tez çalışmam, Lepidoptera'daki dört grup için çok daha gelişmiş ve çoğu durumda genom verisine dayalı ilk filogenetik ağaçları sunuyor; ayrıca bu analizlerde kullandığım biyoinformatik iş akışını yeniden üretilebilir bir veri analizi süreci olarak erişilebilir kılarak, diğer araştırmacıların da takson örneklemini geliştirmek için benzer veri kombinasyonlarını kullanabilmesine imkan tanıyor. Son olarak, bu tür kapsamlı ve iyileştirilmiş filogenetik ağaçların ne kadar faydalı olabileceğini, kelebekler için daha önce yayımlanmış bir ağacı kullanarak önemli bir adaptasyonun evrimini inceleyerek göstermiş oluyor.

List of publications

This thesis is based on the following publications, referred to by their Roman numerals:

I **Molecular phylogeny of north European Geometridae (Lepidoptera: Geometroidea)**
E. Őunap, V. Nedumpally, **E. Yapar**, A. R. Lemmon, T. Tammaru
Systematic Entomology (2025), 50(1), 32–67

II **Elaborating the phylogeny of Noctuidae by focusing on relationships between north European taxa**
V. Nedumpally, A. Zilli, **E. Yapar**, T. Tammaru, A. R. Lemmon, E. Őunap
Systematic Entomology (2025), e70010

III **Integrating Sanger and next generation sequencing data sheds light on phylogenetic relationships among gelechioid moths (Lepidoptera: Gelechioidea)**
E. Yapar, A. Chiocchio, M. A. Heikkilä, J. Rota, L. Kaila, N. Wahlberg
Systematic Entomology (2025), e70009

IV **Phylogenomics resolves subfamily-level relationships among Erebid moths (Lepidoptera: Noctuoidea: Erebidae)**
E. Yapar, H. R. Ghanavi, R. Zahiri, N. J. Dowdy, N. Wahlberg
Manuscript

V **Tempo and mode of winter diapause evolution in butterflies**
S. Halali, **E. Yapar**, C. W. Wheat, N. Wahlberg, K. Gotthard, N. Chazot, S. Nylin, P. Lehmann
Evolution Letters (2025), 9(1), 125–136

All papers are reproduced with permission of their respective publishers.

Author contributions

The following statements give more detail compared to the published statements in order to better reflect the author's specific contributions.

i) EŐ: Conceptualization; data curation; formal analysis; investigation; methodology; resources; validation; visualization; writing—original draft; writing—review and editing; project administration. VN: Investigation; data curation; methodology; writing—original draft; writing—review and editing. EY: Conceptualization of the overall methodological approach; supervision of VN through genomic analyses; software; writing—review and editing. ARL: Resources; software; writing—review and editing. TT: Conceptualization; funding acquisition; project administration; resources; supervision; writing—review and editing.

ii) VN: Conceptualization; investigation; data curation; methodology; visualization; formal analysis; writing—original draft; writing—review and editing. AZ: Validation; writing—original draft; writing—review and editing. EY: Conceptualization of the overall methodological approach; supervision of VN through genomic analyses; software; writing—review and editing. TT: Conceptualization; funding acquisition; project administration; resources; supervision; writing—review and editing. ARL: Resources; software; writing—review and editing. EŐ: Data curation; investigation; methodology; resources; validation; visualization; writing—original draft; writing—review and editing.

iii) EY: Conceptualization, Formal analysis, Writing—original draft, Writing—review & editing, Visualization. AC: Writing — review & editing, Data curation. MH: Writing — review & editing, Data curation. JR: Funding acquisition, Writing — review & editing. LK: Writing — review & editing, Supervision, Conceptualization. NW: Funding acquisition, Supervision, Conceptualization, Writing — review & editing.

iv) EY: Conceptualization, Formal analysis, Writing—original draft, Writing—review & editing, Visualization. HRG: Writing — review & editing, Data curation. RZ: Writing — review & editing, Data curation. NJD: Writing — review & editing, Validation. NW: Funding acquisition, Supervision, Conceptualization, Writing — review & editing.

v) All authors contributed to the study design. SN collected the diapause data, SH carried out the phylogenetic analyses with inputs from everyone, EY wrote/modified R functions for summarizing uncertainty in ancestral estimation and for calculating rate through time, SH and PL wrote the first draft of the manuscript, and everyone contributed with refining the draft.

Authors: Etka Yapar (EY), Andrea Chiocchio (AC), Allan R. Lemmon (ARL), Alberto Zilli (AZ), Erki Õunap (EŐ), Hamid R. Ghanavi (HRG), Jadranka Rota (JR), Lauri Kaila (LK), Maria Heikkilä (MH), Nicolas J. Dowdy (NJD), Niklas Wahlberg (NW), Reza Zahiri (RZ), Sridhar Halali (SH), Sören Nylin (SN), Philipp Lehmann (PL), Toomas Tammaru (TT), Vineesh Nedumpally (VN).

Papers not included in this thesis

Pre-processing of paleogenomes: Mitigating reference bias and postmortem damage in ancient genome data
D. Koptekin*, **E. Yapar***, K. B. Vural*, E. Sağlıcan, N. E. Altınışık, A.-S Malaspinas, C. Alkan, M. Somel
Genome Biology

Bayesian inference of mixed Gaussian phylogenetic models

B. Brahmantio, K. Bartoszek, **E. Yapar**
arXiv preprint

Phylogenomic insights into the relationship and the evolutionary history of planthoppers (Insecta: Hemiptera: Fulgoromorpha)
J. Deng, A. Stroiński, J. Szwedo, H. R. Ghanavi, **E. Yapar**, D. Castillo Franco, M. Prus-Frankowska, A. Michalik, N. Wahlberg, P. Łukasik
Systematic Entomology

Somatic copy number variant load in neurons of healthy controls and Alzheimer's disease patients
Z. G. Turan, V. Richter, J. Bochmann, P. Parvizi, **E. Yapar**, U. Işıldak, S.-K Waterholter, S. Leclere-Turbant, Ç. D. Son, C. Duyckaerts, İ Yet, T. Arendt, M. Somel, U. Ueberham
Acta Neuropathologica Communications

*: Equal contribution.

Introduction

Phylogenetics is a field of research that aims to infer the evolutionary history of a group of organisms. Since its dawn, phylogenetics has been producing invaluable scientific output. Today, it is common knowledge that birds are not only relatives of dinosaurs, but they are the only extant lineage of dinosaurs (Zhou, 2004). Similar to the bird example, we now know that snakes are a lineage of legless lizards (Townsend, Larson, Louis, & Macey, 2004). We would not have been able to solve these mysteries without extensive phylogenetic research. Another interesting example of applications of phylogenetics that possibly had even a higher impact on the general public lately concerns the COVID-19 pandemic. Thanks to the modern sequencing technologies and the phylogenetic inference tools, we were able to trace the newly emerging lineages of the SARS-Cov2 virus, be aware when one lineage were to dominate the available hosts, and take preemptive action. Overall, phylogenetics is an important area of research that has great descriptive and predictive power. It is nowadays at the heart of many sub-disciplines in biology as a robust and complete phylogeny is the crucial prerequisite for any comparative study.

Earlier uses of phylogenetic trees as tools for explanation of evolutionary hypotheses and even the term *phylogeny* itself goes well before there were any methods to formally infer phylogenies (e.g. Darwin, 1859; Haeckel, 1866; Lamarck, 1809). The earliest methods of phylogenetic inference relied on **parsimony**. Under Maximum-parsimony tree inference, tree topologies are evaluated by how many changes in the states of a character are needed to explain the topology, and the most parsimonious tree, which would require the fewest changes, is then accepted as the best. Parsimony methods were originally developed and used for discrete morphological characters and although their use was extended for molecular data, this line of phylogenetic inference method had limited use outside systematics and taxonomy.

Distance methods use a pairwise distance matrix to infer phylogenetic trees by clustering. While the distances on continuous characters (such as gene frequencies) could potentially be directly calculated (using simple metrics such as the Euclidean distance), set of statistical models were needed to calculate distances on discrete characters to reliably

infer the number of substitutions in a way that would be robust to repeated substitution and back-substitution events along a single branch. Jukes and Cantor (1969) introduced the first such model for DNA sequences. Remnants of the common methods from the times when distance-based methods were popular are the UPGMA, an algorithm that is nowadays mainly useful for other applications such as the hierarchical clustering of a group of genes based on their expression without regard to "evolutionary origins" per se; and the Neighbor-joining method, which is still somewhat popular as a relatively quick method of phylogenetic inference.

Maximum likelihood (**ML**) is then the natural extension by the use of available stochastic models for sequence evolution, first demonstrated by Neyman (1971) for three-leaved phylogenetic trees based on protein sequences. Under this framework, the likelihood of the data given the substitution model parameters, the tree topology and the branch lengths can be calculated for the alternative topologies; and the topology that has the maximum likelihood score then can be treated as the best tree given the data and the model parameters. Although the theoretical framework for maximum-likelihood phylogenetic methods existed as early as 1971, it was not until some practical problems regarding computational cost that prevented its use were solved by the introduction of the "pruning algorithm" (Felsenstein, 1973) and its application on DNA sequences (Felsenstein, 1981) that these methods became really feasible. Although not specifically designed for the maximum-likelihood framework, the first widely adopted method of assessing statistical confidence on the inferred phylogenies with ML was introduced by Felsenstein (1985) also, through the application of the non-parametric bootstrap resampling method to characters in a phylogenetic data matrix followed by repeated tree inferences on the bootstrap samples. In parallel, starting around mid 1980s, with the advent of the "molecular revolution"[1], molecular phylogenetics using maximum parsimony or maximum-likelihood methods were finally available to systematists as reliable and powerful tools they always wanted. However, as molecular phylogenetics became commonplace, it grew beyond its original niche in systematics and got widely adopted by many other disciplines, and is necessary when including more than two species for comparative work.

Approximately 11% of all animals inhabiting our planet are butterflies and moths — the order Lepidoptera (Bánki et al., 2025). They have economic significance in terms of both positive and negative implications. Products that we can get from some Lepidoptera, such as the silk from the silk moth *Bombyx mori* bring value to us humans while numerous forest and agricultural pests constitute some of our problems. In addition to their economic significance, Lepidoptera are trusty model systems for biologists from a diverse set of sub-disciplines. Some examples include the African butterfly species *Bicyclus anynana*

[1]This term refers to a period in history roughly between the mid-1980s and the mid-2010s, when DNA sequences were being sequenced across many fields in biology and were being used to answer multitude of scientific questions involving biomedical research, molecular genetics , ecology & evolution, systematics and many more.

being used as a model for evolutionary developmental biology (Brakefield, Beldade, & Zwaan, 2009). Lastly, arguably the most significant role of Lepidoptera in our daily lives is by providing invaluable ecosystem services by pollination, and by being tightly connected in the complex food web in that they are a food source to especially bats and birds, and consumers of plants in the caterpillar stage. Thus, studying the evolutionary history of Lepidoptera via phylogenetics is particularly important.

Lepidoptera systematics through the molecular revolution and the genomic era

Following the trends in other taxonomic levels of the tree of life, Lepidoptera have also seen improvements in the accepted classification and the proposed phylogeny as more data (both in terms of taxa and the number of characters) could be analyzed. Just before the molecular revolution took place in Lepidoptera systematics in the early 1990s, phylogenetic systematic studies on the Lepidoptera had been done with morphological characters and it was at a point in which most early divergences in Lepidoptera that constitute the non-Ditryisan groups had been figured out to be diversified in a "Hennigian comb", i.e. hierarchical gain of apomorphic characters in a stepwise manner. It was mostly the Ditrysian[2] groups, which are estimated to comprise more than 98% of the species diversity of Lepidoptera, that were unresolved apart from the early ideas that the group was diversified in three major splits. First split within Ditrysia had separated the Apoditrysians from the rest; the second had separated the clade Obtectomera from the rest of Apodityrisa; and the third and final split had separated Macrolepidoptera from the rest of Obtectomera. Within each of these nested Ditrysian clades however, superfamilies that were recognized at that time were mostly in patches of unresolved regions (Kristensen & Skalski, 1999; Minet, 1991, also see Figure 1 for a visual representation of some of these clades).

Molecular revolution reaches Lepidoptera

Early efforts in molecular phylogenetic work at or close to the order level for Lepidoptera started mainly with using ribosomal DNA sequences (Weller, Friedlander, Martin, & Pashley, 1992; Wiegmann et al., 2000) and continued with nuclear and mitochondrial protein-coding genes (Mutanen, Wahlberg, & Kaila, 2010; Regier et al., 2009). These earlier multi-locus studies tackled the question of Lepidoptera systematics at or close to the order level, the latter with considerably more comprehensive taxon sampling. Both studies pointed out the resulting poorly-supported branches in the deeper regions in their phylogenies

[2]Ditrysia: a large clade of Lepidoptera that is defined by females having two separate openings for egg-laying and mating. See Figure 1 for a visual representation of these clades.

and attributed this to the possible rapid radiation of the large clades within Lepidoptera. Interestingly, both studies called for an effort in increasing gene and taxon sampling in the hope that this would yield more robust phylogenies. A pattern common to both studies was that the three previously hypothesised large clades within Ditrysia were not recovered as monophyletic, at least not without adjustments that would require some superfamilies to be left out and some to be included in the newer, monophyletic definitions of those clades. Later studies attempted to increase their resolving power by either increasing the number of genes (Regier et al., 2013), or integrating a morphological dataset with the molecular data (Heikkilä, Mutanen, Wahlberg, Sihvonen, & Kaila, 2015), only to discover that large parts of Apoditrysia were still not resolved (**Figure 1**).

Compared with the long line of phylogenetic effort aiming to resolve the Lepidoptera tree of life, superfamilies (e.g. Heikkilä, Kaila, Mutanen, Peña, & Wahlberg, 2011; Heikkilä, Mutanen, Kekkonen, & Kaila, 2014; Kaila, Mutanen, & Nyman, 2011; Sohn et al., 2016; Wang & Li, 2020; Zahiri et al., 2011), families (e.g. Murillo-Ramos et al., 2019; Wahlberg et al., 2009; Zahiri et al., 2012), and other, lower taxonomic levels (e.g. Jiggins, Mallarino, Willmott, & Bermingham, 2006; Kodandaramaiah & Wahlberg, 2009; Peña, Nylin, & Wahlberg, 2011) have received perhaps even more attention with molecular data.

At the time many multi-locus phylogenetic studies on Lepidoptera have used the same set of eight standard or legacy genes that were proposed by Wahlberg and Wheat (2008), which were tested and worked well for Lepidoptera, and another mostly independent set of 16 genes that were widely adopted by a collaborative initiative called LepTree (Mitter et al., 2006). Now is a good moment to take a step back and talk about what were the "genes" in these single or multi-gene phylogenetic studies — what kind of properties a gene ought to have to be used as a molecular marker in a phylogenetic study? Before there were molecular phylogenetics, when researchers relied on morphological characters and parsimony methods, a character needed to — or was assumed to — be *homologous* and hereditary among the taxa studied. Any gene or a fragment of a gene would satisfy the condition to be hereditary as DNA is the essential mechanism and medium by which the living organisms pass their information to their offspring. To be homologous, however, is more complicated and a harder requirement to satisfy and verify. Several different categories of homology are needed to represent the possible discordance in evolutionary histories of genes and the species they are evolving in.

These relationships are essentially defined for any two sequences (sequences from two separate species for the sake of simplicity) and these two sequences are deemed *orthologous* if their divergence can be explained solely by a speciation event, or the main speciation event that resulting in the divergence of these two species predates the gene duplication event. In comparison, a gene duplication event is needed in addition to one or more speciation events to result in a pair of *paralogous* genes, and specifically if a gene duplication event happened before the speciation event that resulted in the divergence of these two species.

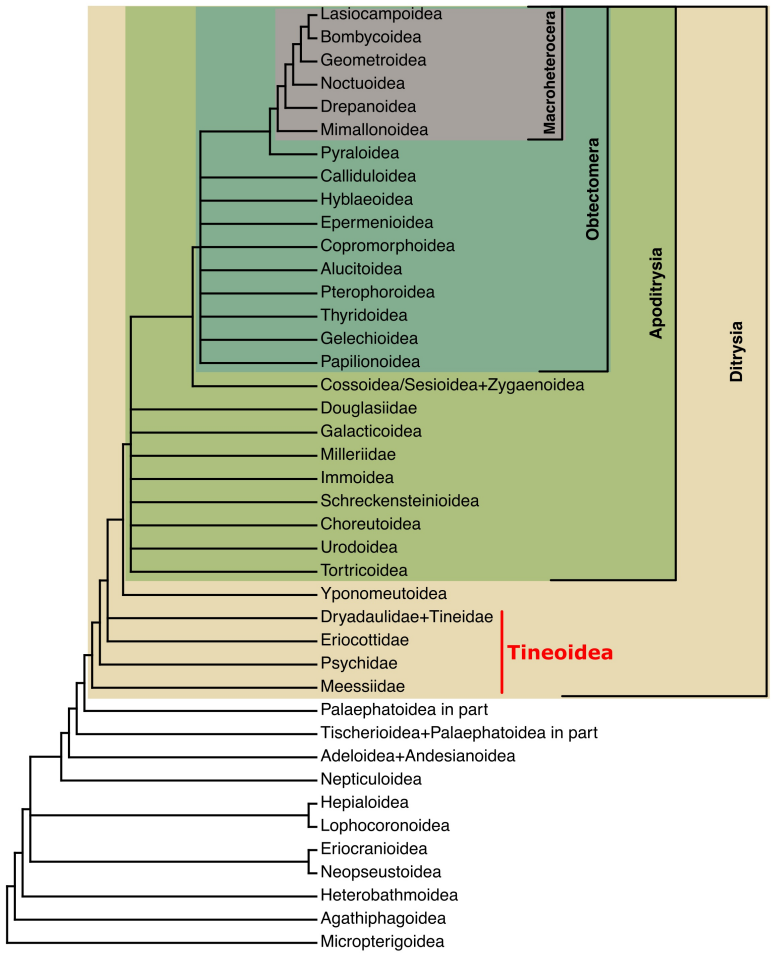

Figure 1: A summary of Lepidoptera backbone phylogeny, synthesized from different studies by Mitter, Davis, and Cummings (2017). The figure was redrawn from the same source. Each leaf in the tree represents a superfamily of Lepidoptera except for superfamilies that are not monophyletic, in which case families under the superfamily were shown by the leaves.

Figure 2a shows a simplistic example in which each terminal branch represents a gene sequence and any two sequences with the same color are orthologous to each other whereas two sequences of different colors are paralogous. As it can be seen from this simple example with only two homologous genes and three species, when not all three sequences are orthologous to each other, the resulting gene-tree topology may not reflect the true speciation history of the tree species (**Figure 2d**). This can also cause problems in concatenated maximum-likelihood analyses where phylogenetic signal in each gene contributes to the total likelihood of the entire dataset and if enough errors were made during orthology

assessment, the genes containing paralog sequences may overturn the overall likelihood in favor of a discordant topology. Therefore, when the gene sample space is increased to hundreds or thousands with phylogenomic studies, it is needed to either i) use a curated set of reliable genetic markers shown to be orthologous for a group of taxa, or ii) to start every phylogenomic study with an orthology-inference step to obtain a set of orthologous genes for the group of taxa to be studied. Although the second option would be the gold-standard approach, it may not be feasible in terms of computational cost and the expertise needed. Luckily, one can rely on the available databases for hierarchical orthologous groups defined at various taxonomic ranks such as OrthoDB (Kriventseva et al., 2019).

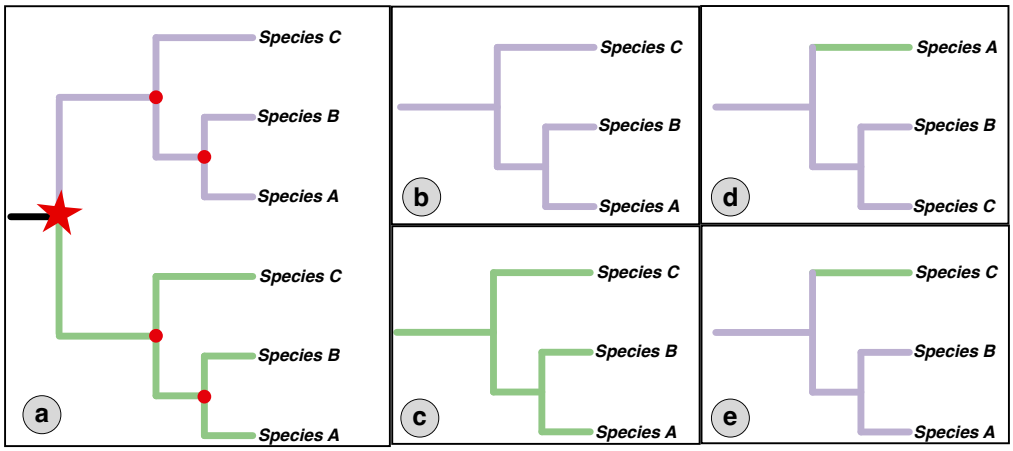

Figure 2: **A hypothetical example of a three-species and two-gene system where Species A and B are the closest relatives and the Species C is sister to this pair.** **(a)** A gene-family tree where two genes (purple and green) are homologous to each other and have diverged from each other before the speciation events. red circles represent speciation events while the red star represents a gene duplication event. **(b)** and **(c)** the expected gene-tree topologies if orthologous genes are are used in phylogenetic inference. **(d)** the resulting misleading gene-tree topology if the a gene [for Species A] which is paralogous to the remaining sequences [for Species B and C] is used in phylogenetic inference. **(e)** the resulting gene-tree topology if a gene [for Species C] which is paralogous to the remaining sequences [for Species A and B] is used in phylogenetic inference. Although the resulting gene-tree topology follows the species-tree topology in this case, some branch lengths would be overestimated.

Phylogenomics efforts towards a resolved Lepidoptera tree

Throughout a decade of phylogenomic studies focusing on Lepidoptera as a whole, some earlier hypotheses and results of previous multi-locus phylogenetic work were corroborated. Some of these large-scale phylogenomic works aimed to have a specifically increased gene sampling (Bazinet, Cummings, Mitter, & Mitter, 2013; Kawahara & Breinholt, 2014), while others aimed to have a comprehensive taxon sampling across Lepidoptera (Kawahara et al., 2019; Mayer et al., 2021; Rota et al., 2022). Perhaps the largest-scale example of genomic data confirming earlier hypotheses was the recurrent recovery of the clade Macro-

heterocera[3] across the phylogenomic studies.

Even though it seemed like the recovered relationships among some Ditrysian groups were converging to a somewhat resolved backbone with the addition of more data, there were concerning patterns emerging regarding the validity of phylogenomic results and the possible inconclusive nature of the data at times as people now had datasets that were more than a million characters in length. Overall, two main problems can be discussed here. First one is that the relationships among superfamilies or large, multi-superfamily clades at deeper parts of the tree (e.g. within Apoditrysia) received generally poor statistical support across these studies, and the second one is that the relationships that were strongly supported by these studies were in disagreement, perhaps an even more concerning problem. An example for the second problem is the seemingly "well-resolved" relationships among the macroheteroceran superfamilies Geometroidea, Noctuoidea, Bombycoidea, and Lasiocampoidea. Three out of five studies recovered Geometroidea as sister to Bombycoidea and Lasiocampoidea, and Noctuoidea as sister to this three-superfamily clade (Bazinet et al., 2013; Kawahara et al., 2019; Mayer et al., 2021, see also Figure 1). The remaining two studies were not in agreement however; in one of the studies Geometroidea and Noctuoidea swapped positions compared to the aforementioned pattern for these four superfamilies (Rota et al., 2022) whereas the other one recovered Geometroidea and Noctuoidea as sisters and this two-superfamily clade as sister to Bombycoidea and Lasiocampoidea (Kawahara & Breinholt, 2014). Other examples of large-scale discrepancies among studies included positions of species-rich superfamilies such as Gelechioidea across the tree, which was attributed partly to the problem of compositional heterogeneity after observing the incongruence between resulting topologies from the same study through nucleotide- versus amino acid-based phylogenetic inference by both Rota et al. (2022) and Mayer et al. (2021).

Phylogenetic incongruence

Exemplified by the unresolved phylogenomic tree of Lepidoptera in the previous section, phylogenetic incongruence is a general term used to refer to situations when the inferred trees differ. This may be between different studies, or between results from the analysis of different versions of essentially the same dataset. When phylogenomic studies started to become more common, phylogenetic incongruence also gained traction (Jeffroy, Brinkmann, Delsuc, & Philippe, 2006; Philippe et al., 2017; Steenwyk, Li, Zhou, Shen, & Rokas, 2023).

A diverse set of biological or technical factors may contribute to phylogenetic incongruence.

[3]Macroheterocera was defined to accommodate most of the superfamilies placed earlier in Macrolepidoptera, with the exclusion of butterflies (Papilionoidea), which were unmistakably shown to be outside of Macrolepidoptera. See van Nieukerken et al. (2011) for more details.

Biological factors include **incomplete lineage sorting (ILS)** which can be a result of rapid radiations, as well as continued **gene flow**. All of these factors could contribute to a portion of gene-tree topologies that are discordant with the "true" species tree.

The technical factors include **alignment errors**, which refers to suboptimally or spuriously aligned regions in the multiple sequence alignments (MSAs). These can be the result of random selection of equally good alignments within a region because of difficult to align regions in the gene fragments. To deal with such a problem, a plethora of so called alignment filtering software are developed to remove either gap rich/not well conserved (Castresana, 2000), or unreliably/randomly aligned columns (Kück et al., 2010; Sela, Ashkenazy, Katoh, & Pupko, 2015) from the alignment matrices. One of the most commonly used tools for such purpose is TrimAl (Capella-Gutierrez, Silla-Martinez, & Gabaldon, 2009), which is able to optimize on both aspects mentioned and also offers two heuristic methods to optimally filter columns to retain most of the phylogenetic signal. Although implementing one or another kind of alignment filtering into phylogenetic workflows is commonplace nowadays, there is evidence against their usefulness, at least in terms of increased rates of topological discordance and the number of unresolved branches in recovered single gene tree topologies treated with such workflows (Tan et al., 2015). Errors in **orthology inference** can also possibly contribute to incongruence specifically by introducing noise as opposed to increasing the phylogenetic signal in a dataset (see **Figure 2** for an example).

Taxon sampling is another factor that can strongly affect the accuracy of phylogenetic inference (Sanderson, McMahon, & Steel, 2010; Steenwyk et al., 2023), even when its comes at the cost of highly incomplete gene sampling (Cho et al., 2011). One way to increase taxon sampling for phylogenomic datasets would be then to combine the data from different genomic or pre-genomic datasets, enabling researchers to maximize the utility of their own data and the sea of publicly available sequence data that may happen to include their taxon of interest. As a rule of thumb, the data with the most reduced representation of the entire genome would dictate the completeness of the combined data matrix. For instance, after sequencing several hundred ingroup specimens with Anchored-Hybrid Enrichment — a genomic data source with a reduced representation of the genome— one could bioinformatically identify the same loci among additional dozens of published genomes from the same group to broaden the taxon sampling at critical levels in the group of interest. It would also be possible to incorporate low throughput Sanger sequencing data to a phylogenomic dataset of any kind to dramatically increase the number of taxa for instance, because this is expected to improve the accuracy of the inference, but also because this combination approach may be the only way to include some key taxa that otherwise would not be represented (see e.g. Espeland et al., 2023; Fonseca & Lohmann, 2018, for examples of this data integration approach)

Even though it may not be possible to avoid the effects of the biological factors listed above on incongruence, it may be possible to do so for some of the technical problems. This

would not only make the resulting phylogenies more robust but would also make it possible to quantify the actual biological discordance —through the established methods such as concordance factors— and investigate potential events that may have resulted in these discordant patterns without the added noise from the technical factors.

Using phylogenies in evolutionary biology

As briefly introduced before, phylogenies are adopted by a wide array of disciplines such as evolutionary genomics and transcriptomics (Brawand et al., 2011; Gillard et al., 2021; Hoile, Holland, & Mulhair, 2025; Hoile et al., 2025), for studies on macroevolution of a key morphological or a behavioral character (Chiarenza et al., 2021; Larouche, Gartner, Westneat, & Evans, 2023; Yoshida & Kitano, 2021), historical biogeography studies (Chazot et al., 2021; Condamine, Sperling, & Kergoat, 2013) and many more applications which require a robust phylogeny that is as taxon-complete as possible.

Evolution of discrete characters across a phylogeny can be studied with modern comparative methods that rely on the available specific Continuous-Time Markov chain models of character evolution (e.g. the Mk model; Lewis, 2001). These models are similar to the models of nucleotide evolution used for phylogenetic inference (e.g. The Jukes-Cantor model introduced earlier; Jukes & Cantor, 1969) in terms of the statistical properties of both kinds of models and the way in which the evolutionary process is modeled with the difference being that these models are not used for inferring phylogenies in comparative studies. Rather they are used to estimate the parameters of the underlying model, such as the stationary frequencies of the character states, and the evolutionary rate of change between them on a fixed phylogenetic tree. Then, insights regarding the likely ancestral states of the character of interest can be gained by e.g. through the calculation of marginal probabilities of each character state existing at the internal nodes of the phylogenies.

Study system

Given all the problems introduced above regarding phylogenetic incongruence and the opportunity to improve taxon sampling by integrating reference genomes with other types of data, I selected four species-rich groups of Lepidoptera with rich available molecular data as case studies to infer robust phylogenetic hypotheses by this data combination/integration strategy in a phylogenomic workflow. These are the ditrysian superfamily **Gelechioidea** and three families of Macroheterocera: **Geometridae**, **Noctuidae**, and **Erebidae**. Then, I focused on another group, butterflies (**Papilionoidea**), because of the availability of global-scale time-calibrated phylogenies and the abundance of trait data to apply modern phylo-

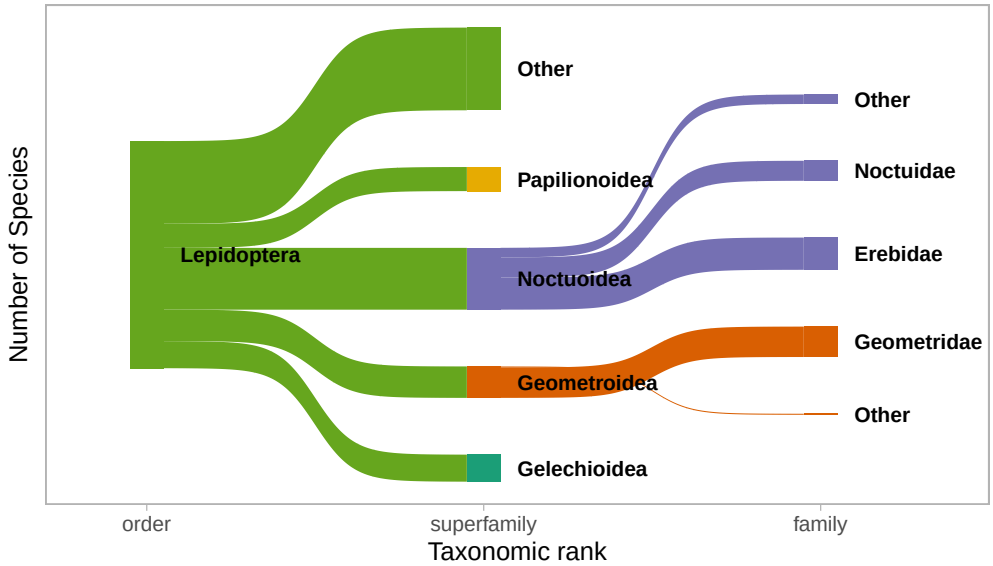

Figure 3: A diagram showing the breakdown of number of Lepidoptera species (thickness of ribbons) at the order, superfamily and the family level for the study species. Underlying data for species numbers are retrieved from The Catalogue of Life (Bánki et al., 2025).

genetic comparative methods in order to investigate the evolution of a key adaptation, winter diapause.

Geometridae (**Paper I**), are characterised by their usually broad and delicate wings and the characteristic "looping" movement of their larvae. With more than 24,000 described species, they are the second largest family of Lepidoptera after Erebidae (**Figure 3**). The higher-level phylogeny of Geometridae is yet to be resolved as the previous multi-locus phylogenetic studies (Abraham et al., 2001; Murillo-Ramos et al., 2019; Sihvonen et al., 2011; Young, 2006), and the only phylogenomic study with limited taxon sampling (Murillo-Ramos, Twort, Wahlberg, & Sihvonen, 2023) show disagreements e.g. the sister group to the rest of Geometridae.

Noctuidae (**Paper II**) have perhaps been one of the most controversial families in Lepidoptera, with an everchanging delimitation. Many subgroups that used be included under Noctuidae have later been transferred to **Erebidae** and to other families under the superfamily **Noctuoidea** (Zahiri et al., 2012, 2011). Currently, there are more than 12,000 described noctuid species (van Nieukerken et al., 2011, see also **Figure 3**). This family is also famous for its pest species (Mitchell, Mitter, & Regier, 2006), and the long-distance migration of the Bogong moth (Warrant et al., 2016). Moreover, the range of host plants a given taxon is able to feed on varies greatly across the family. Inferring a robust phylogeny with comprehensive taxon sampling is required to study the evolution of such traits in a

comparative framework.

Gelechioidea (**Paper III**) constitute another one of the largest groups under Lepidoptera with more than 18,500 described species (van Nieukerken et al., 2011, see also **Figure 3**). Some members of this superfamily are pests of economically significant crops such as tomato, wheat, cotton, and blueberry. Despite the overwhelming number of species and the economic significance, Gelechioidea had been disproportionately understudied when compared to the other groups of Lepidoptera and is thought to be the most species-rich superfamily when the estimated number of undescribed species is taken into account. Previous phylogenetic work focusing on the group as a whole presented conflicting results regarding the extent of the families and the phylogenetic relationships among those families (Heikkilä et al., 2014; Kaila et al., 2011; Sohn et al., 2016; Wang & Li, 2020).

Erebidae (**Paper IV**) have gone through major revisions in the past several decades. Shortly after being returned to family status by Fibiger and Lafontaine (2005), this family was soon "dissolved" into **Noctuidae** (Mitchell et al., 2006) until it was re-defined to include more subgroups by Zahiri et al. (2012, 2011). It is currently the largest family of Lepidoptera with approximately 25,000 described species (Sisson et al., 2025), and harbors striking morphological an behavioral diversity such as the differences in coloration that result in some Erebidae species resembling certain wasps or other moth families with aposematic colours; as well as the "vampire moth" with the adaptation to pierce animal skin to feed on blood (Bänziger, 2007). Internal relationships of the major lineages within Erebidae are still unresolved (Ghanavi, Twort, Hartman, Zahiri, & Wahlberg, 2022; Sisson et al., 2025).

Butterflies, the superfamily **Papilionoidea** (**Paper V**), are one one the largest superfamilies in Lepidoptera with more than 18,700 described species (van Nieukerken et al., 2011, see also **Figure 3**). They are also the best-studied superfamily in Lepidoptera in terms of phylogenetic research at various scales within the superfamily (Chazot et al., 2019; Heikkilä et al., 2011; Kawahara et al., 2023; Peña et al., 2011; Wahlberg et al., 2009), and in terms of collective trait databases that document a variety of essential traits (Middleton-Welling et al., 2020; Shirey et al., 2022). With their near global distribution, butterfly lineages in different parts of the world face different challenges. For example, the temperate are to survive in highly seasonal environments. Entering hibernation diapause —suspending or delaying their development or reproduction— is a highly effective strategy against for this. Therefore in addition to the four groups of Lepidoptera I selected for phylogenomic inference, I focused also on the butterflies to study this complex life-history adaptation in a phylogenetic comparative framework thanks to an existing, comprehensive, time-calibrated butterfly phylogeny (Chazot et al., 2019).

Aims

Lepidoptera are an outlier among other groups of animals in that there are more than a thousand reference-quality genome assemblies available for public use (Wright et al., 2025), from which thousands of orthologous genes could possibly be extracted. In addition to this set of gold-standard data, there are additional genomic-scale data such as transcriptomic raw data, low-coverage whole-genome sequencing data from fresh or museum specimens, and data generated for earlier phylogenomic work by enrichment-based methods targeting protein-coding genes (e.g. Anchored-hybrid enrichment [AHE]) or those targeting the so-called ultraconserved elements (UCEs). The majority of data that span the most diverse set of taxa however, is probably the Sanger-sequenced genes which accumulated across Lepidoptera from the earlier multi-locus studies throughout the past several decades.

Mirroring this abundance of data, there is a high demand for reliable phylogenies not just for taxonomic work but also for the wide array of comparative fields engaging with questions such as comparative genomics and transcriptomics, historical biogeography, the dynamics of the macroevolution of a key biological trait — questions which rely on the inferred phylogenetic hypothesis as a prerequisite. Given the great effect of comprehensive taxon sampling on the accuracy of phylogenetic inference (Sanderson et al., 2010; Steenwyk et al., 2023), it is in our best interest to approach this challenge of inferring robust phylogenetic hypotheses with the best possible taxon sampling via combining the gene-rich genomic data with taxon-rich "legacy data". Although this approach would not be attainable for fields that essentially model features of the genomes themselves as evolving traits (e.g. chromosome evolution, gene copy number evolution), for the majority of the remaining comparative questions this data integration approach, which dramatically improves taxon sampling, could potentially help in inferring well-resolved phylogenies. It is crucial however to mitigate problems I introduced earlier (e.g. problems in orthology inference, presence of outlier sequences, true biological discordance between the evolutionary histories of the species and the genes) throughout this data integration approach. While some of these problems are inherent to the biological system at hand, others could be exacerbated particularly because of the methodological approach taken. These could arise, for instance, because of the increased risk of falsely-identified target genes from transcriptomes

or fragmented genome assemblies which in turn could mask the phylogenetic signal otherwise present in the data. Another source of problem mainly concerns the inflated missing data in case of integrating "legacy data" with genomic data, which could affect the efficacy of phylogenetic inference through displaying a "bimodal" phylogenetic signal across the dataset whereby the taxon- and gene-rich portions of the integrated dataset support conflicting topologies compared to each other.

Throughout this thesis, I employ this data integration approach to address the need for robust phylogenetic hypotheses, focusing on four major species-rich groups that together comprise **49% of the species diversity** in Lepidoptera: **Noctuidae, Erebidae, Geometridae, and Gelechioidea**. The reasoning behind focusing on these specific four groups was that inferring well-supported phylogenetic hypotheses for Gelechioidea, Geometridae, and Noctuidae have proven to be difficult, illustrated by the **inconsistent results across different studies** for the same group; and that the previously suggested phylogenetic relationships for Erebidae have not been tested through another, independent study or with phylogenomics.

The overarching aims of this thesis are:

1. To formulate the bioinformatic steps needed for data integration in large-scale phylogenomic studies by addressing the methodological challenges such as false orthologs and the effects of missing data introduced above,
2. Through the application of the framework I formulate, to infer robust, well-resolved phylogenies for four species-rich groups of moths,
3. To apply modern phylogenetic comparative methods to study macroevolution of an important life-history trait in another major group of Lepidoptera, butterflies, thereby demonstrating the power of robust phylogenies in addressing broad-scale evolutionary questions.

The details regarding the individual aims and the specific case of combining reference genomes with other data sources for the five papers are as follows.

Papers I & II both utilised available public data in addition to the data which were sequenced for the respective studies using a targeted sequencing protocol, Anchored-Hybrid Enrichment. In **Paper I**, I used the available rich Sanger-sequencing data together with the genomic (AHE) data resulting in a dramatic improvement in taxon sampling that led to a better phylogeny for the family **Geometridae**. In **Paper II**, I analysed published reference genomes together with the AHE data for the same purpose of improving taxon sampling and inferring a well-resolved phylogeny for **Noctuidae**.

In **Paper III**, I studied the phylogenetic relationships among families of **Gelechioidea** using a two-step approach in which I analysed a genomic-only (gene-rich) dataset with the

best genomic taxon sampling to date to have a well-supported backbone phylogeny (i.e. placement of individual families). I then used this genomic tree as a topological constraint during the phylogenetic inference of the a larger, taxon-rich dataset I created by integrating the available legacy data.

In **Paper IV**, I investigated the phylogenetic relationships among the subfamilies of the moth family Erebidae through a phylogenomic dataset I compiled from reference genomes, transcriptome assemblies, and lower quality novel genome assemblies which sampled all the major lineages of Erebidae. In addition to the phylogenetic results, I present a reproducible bioinformatic pipeline for automated compilation of phylogenomic datasets, which I developed by implementing the necessary steps of the methodological framework I formulated earlier. This pipeline is specifically well-suited for integrating reference genomes and lower quality genomic data.

Finally, in **Paper V**, I applied phylogenetic comparative methods to study the evolution of the ability to go into winter diapause as well as the rates of change between different diapause strategies, i.e. going into diapause at various life stages, as a larva, egg, or as an adult individual, across the superfamily **Papilionoidea** (butterflies). While the previous chapters focus on inferring well-supported phylogenetic hypotheses through data integration, in this paper I utilised a previously published time-calibrated phylogeny for butterflies with impressively broad taxon sampling for the comparative analyses, which demonstrates the type of in-depth macroevolutionary research that the robust phylogenies **Papers I–IV** aim to deliver.

Methods

Raw data and alignment preprocessing

In addition to the publicly available reference genome assemblies, this thesis utilised the available public or in-house raw whole-genome or RNA sequencing data, which needed to be assembled *de novo* into genome or transcriptome assemblies. A common step prior to the *de novo* assembly was cleaning of the raw reads from sequencing adapters, and low-quality or low-complexity regions. For adapter trimming, both Trimmomatic (Bolger, Lohse, & Usadel, 2014) and Cutadapt (Martin, 2011) were used. For cleaning of low-complexity or low-quality regions, PRINSEQ++ (Cantu, Sadural, & Edwards, 2019) was used. The genome assemblies were then created using the SPAdes assembler (Prjibelski, Antipov, Meleshko, Lapidus, & Korobeynikov, 2020) with the cleaned genomic reads, and the transcriptome assemblies were created with Trinity (Grabherr et al., 2011) using the cleaned transcriptomic reads.

Since **Papers I & II** mainly relied on target-enrichment sequencing through a custom Anchored-Hybrid enrichment probeset, the assembly of targeted loci from the sequenced samples needed a procedure that operates directly on the trimmed and cleaned raw reads. For this locus assembling purpose, the software introduced by Prum et al. (2015) was used in **Paper I**. In **Paper II**, another software, HybPiper (Johnson et al., 2016), was used to assemble raw reads into the targeted set of AHE loci. Within the HybPiper protocol, the raw reads are first distributed into bins corresponding to the target loci based on sequence similarity determined by the DIAMOND aligner (i.e. a BLAST alternative; Buchfink, Reuter, & Drost, 2021), then reads for each bin are used for the *de novo* assembly of the target loci only using the reads mapped to each respective locus.

For **Papers III & IV**, main geneset was based on the Lepidoptera single-copy orthologs as defined by OrthoDB v10 (Kriventseva et al., 2019). These genes were searched within the public reference genomes, the generated in-house genome assemblies, and the assembled transcriptomes from public RNA-Seq data using the "Benchmarking Universal Single Copy Orthologs" (BUSCO, v5.2.2) software (Simão, Waterhouse, Ioannidis, Kriventseva, &

Zdobnov, 2015). The identified hits were extracted from the respective assemblies using a custom Python script that processes the output folders from BUSCO, as the version used throughout this thesis only outputs amino-acid sequences of found genes.

The nucleotide or amino-acid sequences were aligned using MAFFT (Katoh & Standley, 2013) with the L-INS-i algorithm. Resulting alignments were trimmed with TrimAl (Capella-Gutierrez et al., 2009) to remove gap-rich and nonconserved regions with the `automated1` option.

To filter out outlier sequences and paralogs, at least one form of gene tree-based outlier removal process was applied in all papers. The method that was common to all papers was a custom R script that recursively bisects the gene-trees on abnormally long branches until the tree did not contain outlier branches. The abnormally long branches were identified heuristically by exploring the distribution of relative branch lengths over all gene trees in a trade-off between data loss and integrity. For **Papers I–III**, this R script was the sole form of outlier and paralog removal and therefore applied to both terminal (i.e. likely singleton outliers) and internal branches, whose descendants likely are paralogous to the rest of the sequences in the alignment. For **Paper IV**, a two-step outlier removal strategy was adopted to mitigate pruning of outgroup branches in case of the use of few distant outgroups. This improved strategy included the processing of the gene trees first using TreeShrink (Mai & Mirarab, 2018) and then by the modified version of the R script explained above so that the script operated only on internal branches to better deal with paralogs, after observing the prevalence of long internal branches among gene trees that were only filtered with TreeShrink. See **Figure 4** for examples of outliers and paralogs shown on gene trees.

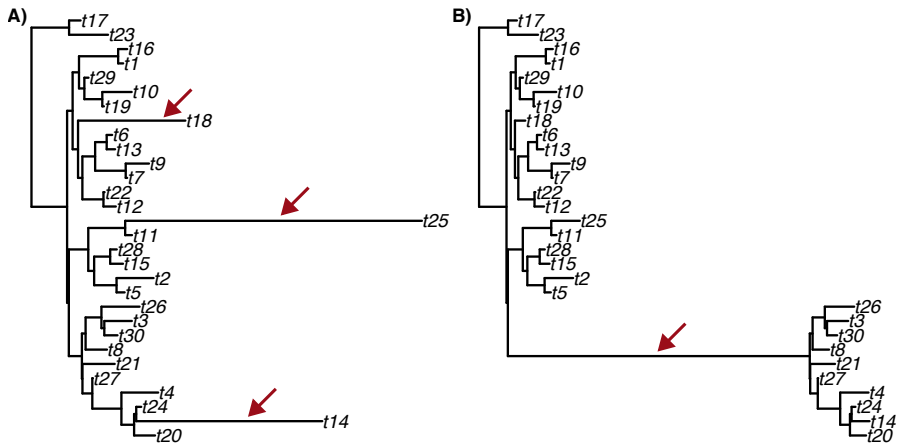

Figure 4: **Theoretical examples of outlier and paralog sequences being detected on gene trees. A)** An example tree of a gene alignment in which there were no paralogs but three potential outlier sequences. **B)** Another example tree of a gene alignment that included a set of sequences that (the smaller subtree when when the tree is bisected on the marked branch) were likely paralogous to the rest of the sequences (the larger subtree)

Phylogenetic analyses

Phylogenetic inference, assessing branch support, and hypothesis testing

In **Papers I and III**, Sanger-sequencing data were combined with data generated with genomic methods (AHE and WGS/RNA-Seq, respectively), with the taxa represented with only Sanger sequencing data being the majority in both cases. This presents a problem in phylogenetic inference in that these combined datasets with broader taxon sampling might not have adequate phylogenetic information to resolve deeper splits in the tree. To overcome this problem, different alternative approaches were taken between the two papers, although both alternatives included using the maximum-likelihood topology obtained by analysing the more complete genomic data-only datasets as topological constraints when performing phylogenetic inference using the integrated datasets. In **Paper I**, the taxon-rich dataset that was used to infer the broader tree only included the legacy loci and the fully dichotomous genomic tree was used as the (strict) topological constraint. In comparison, a majority-missing merged dataset was created for **Paper III** in which the legacy loci were the only possibly taxon-complete portion and the rest (the majority) of the matrix was missing data caused by the added taxa lacking genomic loci (**Figure 5**). Moreover, the topological constraint used for **Paper III** was not as strict in that i) the branches on the obtained genomic tree where the alternative phylogenomic analyses did not agree with each other were collapsed to polytomies, and ii) the recovered within-family topologies were not used as a constraint. This ensured that the "backbone" of the tree is dictated by the rich genomic data while shallower phylogenetic relationships that were the focus of the broader dataset are resolved using the Sanger sequencing data with a more limited number of genes but much better taxon sampling.

IQ-TREE (Minh et al., 2020) versions 2.2.0 and 2.4.0 were used in all papers for maximum-likelihood phylogenetic inference. Different partitioning strategies were used both within and across papers. For the analyses that yielded the main topology of many papers, automated partition-merging over the original partitioning according to genes, and subsequent model-fitting was implemented using the ModelFinder2 algorithm (Kalyaanamoorthy, Minh, Wong, von Haeseler, & Jermiin, 2017) provided with IQ-TREE. Moreover, the GHOST model (Crotty et al., 2020) was used in **Papers II and IV** to account for rate heterogeneity across taxa, a mixture model which does not require the *a priori* partitioning of the data matrix. Ultrafast Bootstraps (UFBS) (Hoang, Chernomor, von Haeseler, Minh, & Vinh, 2018) and the SH-aLRT metric were used in **Papers I–IV** to measure statistical branch support. Branches with at least 95% UFBS and 80% SH-aLRT values were considered strongly supported.

Moreover, a different partitioning strategy and the phylogenetic inference software was implemented for the majority-missing taxon-rich dataset of **Paper III**. The dataset was

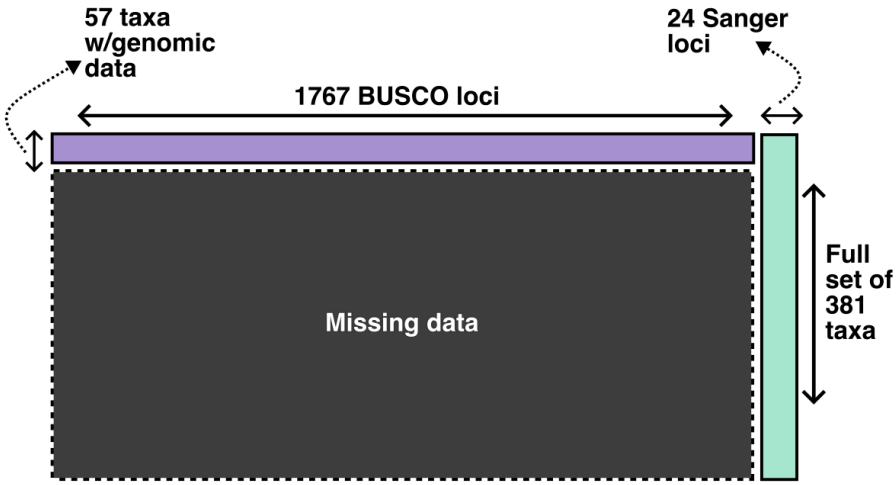

Figure 5: **The schematic representation of the combined data matrix created and analysed** in **Paper III**. Height of each rectangle represents number of taxa and the width represents the number of loci. The figure depicts the missing data problem where taxon sampling can be extended greatly by incorporating Sanger-sequencing data into a taxon-poor but locus-rich phylogenomic dataset.

partitioned into three partitions corresponding to the one mitochondrial gene (*COI*), the legacy nuclear genes —a combination of Wahlberg and Wheat (2008), and the LepTree (Mitter et al., 2006) genes— and the genomic nuclear genes. For phylogenetic inference, RAxML-NG (Kozlov, Darriba, Flouri, Morel, & Stamatakis, 2019) was used for its superior performance in topology search for taxon-rich datasets, and for its more efficient parallelisation strategy compared to IQ-TREE in cases with large datasets with few partitions.

One of the included families, which was represented with only four taxa, was recovered as non-monophyletic. These taxa were scattered across distant positions in the tree, one of those being as sister to the rest of the ingroup taxa, a highly unlikely placement. The tree topology tests provided with IQ-TREE (Kishino & Hasegawa, 1989; Kishino, Miyata, & Hasegawa, 1990; Shimodaira, 2002; Shimodaira & Hasegawa, 1999; Strimmer & Rambaut, 2002) were used to investigated this unusual, possibly technical, problem caused by missing data by testing 8 topologies, 7 possible alternatives based on previous multi-locus studies and the remaining one being the original ML topology (**Figure 6**)

Two lines of methods were used to better investigate the possible discordance between the gene and the species trees; gene-tree reconciliation methods and concordance factors. As a gene tree reconciliation-based inference of the species tree, wASTRAL-h (Zhang & Mirarab, 2022) was used in **Papers III and IV** —which is a newer implementation of the ASTRAL (Zhang, Rabiee, Sayyari, & Mirarab, 2018) algorithm that can take into account the uncertainty in the gene trees by incorporating the branch lengths and branch support

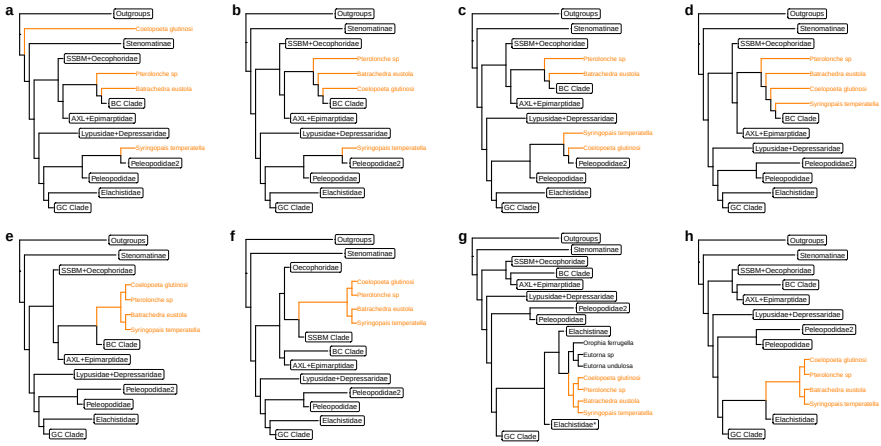

Figure 6: **Cladograms showing the alternative placements for the four members of the problematic family.** a) ML topology, b–h) alternative topologies tested. Most of the alternative topologies were selected based on the sister groups for this family recovered from previous phylogenetic work. Figure from **Paper III**.

metrics when inferring the species tree. Following prior observations on the number of independent maximum-likelihood searches needed to reliably find the "best" topology for gene trees based on simulated and empirical gene alignments (Liu et al., 2024), at least 10 (**Paper IV**) and 40 (**Paper III**) independent tree searches were implemented for each gene alignment; and UFBS and SH-aLRT branch support metrics were calculated with 1000 samples to fully leverage the weighted ASTRAL algorithm. Gene and site concordance factors were calculated using IQ-TREE and quartet concordance factors were calculated by using the ML topology as the fixed user tree for ASTRAL-III. The three step concordance vector approach highlighted by Lanfear and Hahn (2024) were adopted to better visualise and inspect the discordance at specific nodes.

Phylogenetic comparative methods

In **Paper V**, several different methods and models for discrete character evolution were used to study the evolution of diapause across the butterfly phylogeny. The original dataset that was collected with an extensive literature survey for the diapause trait from each taxon coded if a taxon enters winter diapause, and if so, in what life stage this occurs (either adult, pupa, larva or egg). This original categorisation had highly rare frequencies for some states among all the taxa. However, models of character evolution are sensitive to such cases and therefore the tip states were recoded by merging. In the end, two such alternative recodings were created: a three-state coding (adult, juvenile, no diapause), and a binary-state coding (diapausing or non-diapausing species). The Mk model (Lewis, 2001; Pagel, 1994) as implemented in the *phytools* R package (Revell, 2024) was used to fit different models with

respect to the number of parameters or with respect to different root priors, flat prior or fitzjohn prior. These models were the equal rates (ER), symmetric rates (SYM), and all rates different (ARD) models. Additional, more complex models were also fitted to mainly take the possible rate heterogeneity across the phylogeny into account. Such were the hidden-rate markov models provided by the R package *corHMM* (Beaulieu, O'Meara, Oliver, & Boyko, 2022) and the gamma-distributed rate heterogeneity model from the *phytools* package. To quantify and visualise the inconsistency among this wide array of alternative models fitted independently for each character coding strategy, the overall difference among the six alternative ancestral reconstructions at each node was summarised by calculating the total sum of squares across reconstructions.

To sample character histories from posterior distributions (i.e. sampling actual singular events happening along the branches of the phylogeny), the stochastic mapping approach provided by *phytools* with the function *make.simmap* was used. These sampled histories were then used to generate rate through time plots to visualise the rate of change between the character states (gains or losses of diapause, or switching the diapausing life stage) across time, i.e. across cross-sections of the phylogeny, thus normalised by the number of branches within a given time section or bin. Although a useful strategy, there is an inherent uncertainty especially towards the root of the tree where there are fewer branches, and a relatively few change events can have a relatively large effect on the inferred rate through time. For visualising the rate through time data while also showing how much uncertainty there was, R functions from Hughes, Berv, Chester, Sargis, and Field (2021), were modified to also show uncertainty by either plotting confidence intervals or interquartile ranges across the sampled time sections.

Contributions to the field

Genomic data corroborate earlier results and resolve old conundrums

Taken together, phylogenetic results from **Papers I–IV** showcase the utility of "lesser" molecular data when combined with the available reference genomes by providing highly supported phylogenetic hypotheses for Geometridae, Noctuidae, Gelechioidea, and Erebidae. In cases where phylogenomic incongruence was apparent, the methods employed allowed us to carefully address the parts of resulting phylogenetic trees that remained unresolved despite high amounts of genomic data. In addition to resulting in highly supported phylogenetic relationships that provide insights at larger-scales — such as the relationships of families within a superfamily — our methods and data integration approaches allowed us to tackle more finer-scale problems, such as the validity and extent of the accepted tribes[4] in a family. The specific high-level outcomes from each paper are briefly presented below. In all cases, special attention was made to the sister group to the rest of the studied group since a change in this information alone in light of genomic data can potentially alter the previous knowledge and understanding of behavioral or morphological characters based on the previously established consensus on the phylogenetic relationships within the group.

The previous multi-locus molecular work on the family **Geometridae** utilised mostly low-throughput methods such as Sanger sequencing. There have been two competing hypotheses regarding which subfamily is sister to the rest of the family: either the subfamily Larentiinae was recovered as the sole sister to the rest (Abraham et al., 2001; Young, 2006), or a clade consisting of subfamilies Sterrhinae and Larentiinae was recovered as the sister group to the rest (Ban, Jiang, Cheng, Xue, & Han, 2018; Sihvonen et al., 2011; Wahlberg, Snäll, Viidalepp, Ruohomäki, & Tammaru, 2010; Yamamoto & Sota, 2007). The only phylogenomic work aimed to resolve higher level relationships in Geometridae (Murillo-Ramos et al., 2023) supports another topology, in which the subfamily Sterrhinae is sister to the rest of the family while cautiously warning that these higher level relationships are

[4]A tribe is a taxonomic rank that is lower than a subfamily, but higher than a genus. i.e. a group of genera defined based on overall morphological similarity, or molecular evidence.

not robust against varying levels of taxon sampling, even with genomic data — highlighting the importance of adequate taxon sampling in genomic studies of Geometridae. With the most comprehensive taxon sampling to date, our results from **Paper I** support the findings of Murillo-Ramos et al. (2023) with maximal branch support (**Figures 7,8**).

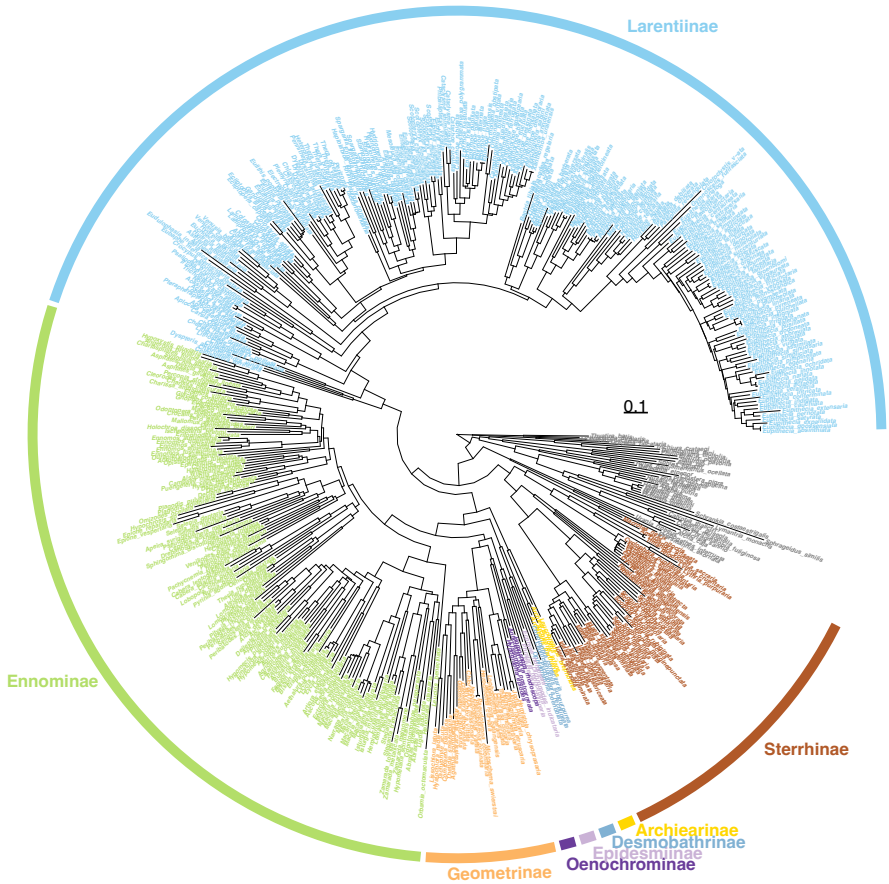

Figure 7: The maximum-likelihood tree from the expanded taxon sampling dataset for Geometridae, from Paper I. 70% all the taxa shown were included thanks to the data integration approach using "legacy" data.

Apart from the question around the sister subfamily to the rest of the family, the concept of the subfamilies and the phylogenetic relationships among them had been resolved to a great degree before **Paper I**, but there had not been many molecular studies to re-evaluate the extent and the phylogenetic relationships among tribes within subfamilies, which had been based mostly on morphology. With our integrated tree with 474 geometrid taxa, we were able to critically re-evaluate the accepted tribes prior to publication of this work to **propose two new tribes in the subfamily Larentiinae**, and to **transfer four taxa to a different tribe** in light of the molecular evidence. In summary, the extensive taxon sampling thanks to

the integration of Sanger-sequenced legacy markers with the genomic data allowed us to resolve both a long-standing, larger-scale question regarding the sister group and fine-grain problems such as the validity of the tribes.

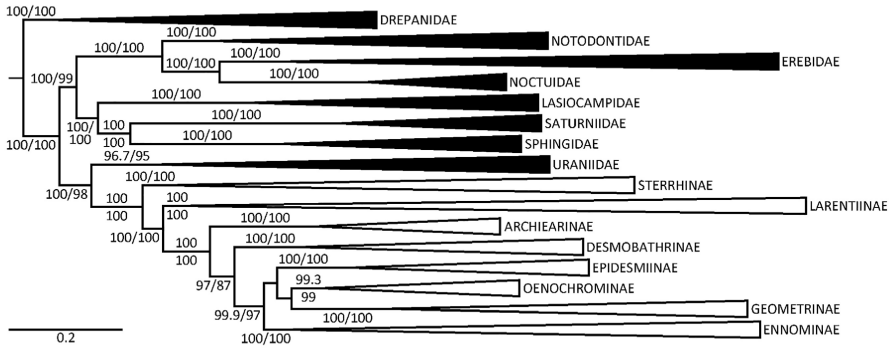

Figure 8: Maximum-likelihood phylogeny showing the relationships among the outgroups (filled in cones) subfamilies of Geometridae (empty cones). Branch support is shown as SH-aLRT/UFBS. Figure from **Paper I**.

Noctuidae has been a family with an everchanging delimitation mostly because of problems regarding the delimitations under the superfamily Noctuoidea. Before Erebidae was revived as a distinct family, taxa that are now included in Erebidae were part of Noctuidae s.l. (Zahiri et al., 2011). No published phylogenomic study was conducted for this group before. There has been conflicting evidence regarding the subfamily-level topology for Noctuidae, especially regarding the sister group to the rest of the family: it has been suggested that Plusiinae was the sister group to the rest (Keegan et al., 2021; Mitchell et al., 2006). We instead found a clade consisting of Acontiinae and Eustrotiinae to be the sister to the rest of Noctuidae (**Figure 9**). Similar to **Paper I**, we also focused on more fine-grained relationships of **Noctuidae** concerning taxonomic ranks lower than subfamily level in **Paper II** such as **raising a tribe to subfamily rank**.

It has previously been challenging to infer a robust family-level phylogeny for Gelechioidea using molecular data from only a few gene sequences. In **Paper III**, we present a much better resolved family level topology for the 18 out of 20 families of Gelechioidea by utilising genomic-scale data. Even when the disagreements among the alternative analyses are taken into account, the consensus of all analyses shows a clear structure in which the Gelechiodiea have diversified into four main lineages, compared to the synthesis of previous molecular studies focused on this superfamily (**Figure 10**). The sister group to the rest of Gelechioidea had varied among previous molecular work, such as Autostichidae (Kaila et al., 2011) or Epimarptidae (Wang & Li, 2020). We instead found the subfamily Stenomatinae from the family Depressariidae to be the sister to the rest (**Figure 11**). Integrating Sanger-sequenced markers allowed us to re-evaluate family delimitations in Gelechioidea and resulted in raising the subfamily Stenomatinae as it was not grouping with the rest of the family Depressariidae. Moreover, we showed that Elachistidae, one of the most diverse

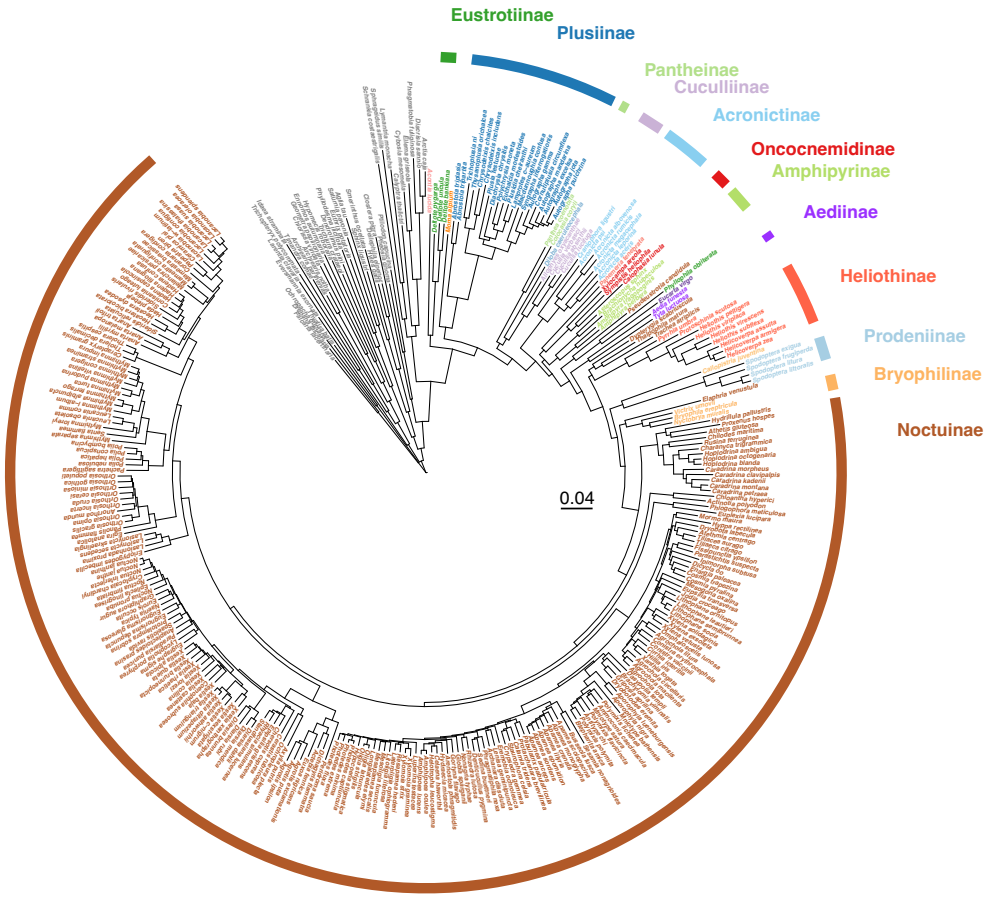

Figure 9: The maximum-likelihood tree for Noctuidae, from **Paper II**. About 16% of the taxa shown here were included thanks to a data integration strategy between AHE data and reference genomes.

families of Gelechioidea, was not monophyletic and proposed a new delimitation.

Erebidae (the focus of **Paper IV**) has also been difficult to deal with phylogenetically as there had been only one other molecular study focusing on the family as a whole in its current delimitation with somewhat adequate taxon sampling. There had been another study that sampled Erebidae although the focus was the entire superfamily Noctuoidea and it had not sampled many subfamilies of Erebidae. We utilised raw whole-genome sequencing data generated by Ghanavi et al. (2022) who used older DNA extracts among the erebid samples from Zahiri et al. (2012) to generate the new genomic data. Previously, it had been proposed that either the subfamily Scoliopteryginae or the subfamily Lymantriinae should be the sister group to the rest of the family (Li et al., 2024; Zahiri et al., 2012). By combining fragmented genome assemblies, reference genomes, and published transcrip-

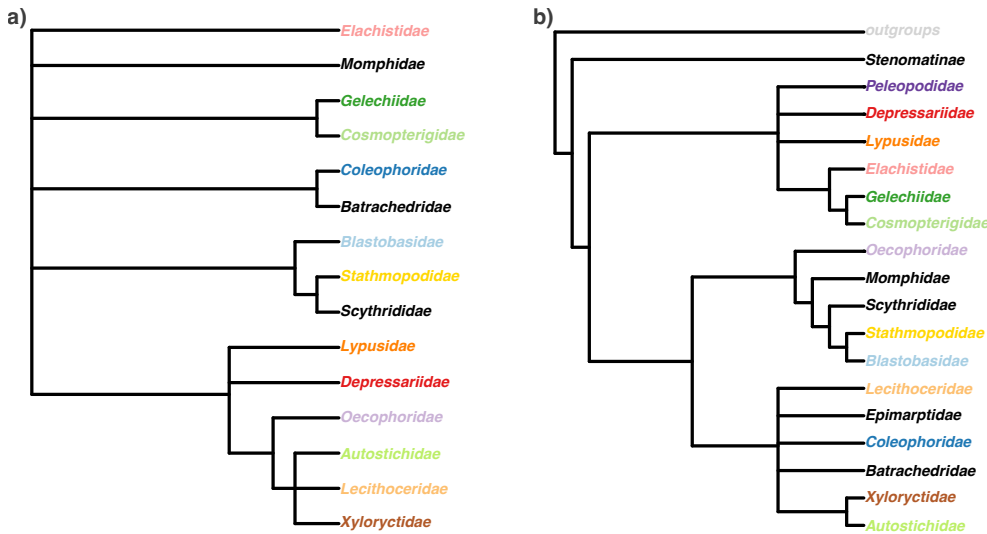

Figure 10: Comparison between a) the synthesis of three previous molecular phylogenetic studies on Gelechioidea; b) the family-level consensus tree of the results of alternative phylogenomic analyses performed in Paper III.

tomic data across 120 Erebidae taxa we show that: the subfamily Eulepidotinae instead is the sister group to the rest of the family. Large parts of the backbone are resolved in terms of subfamily-level phylogenetic relationships concerning the "Arctiinae-associated" and the "Erebinae-associated" subfamilies (**Figure 12**).

Gene-tree species-tree discordance

Gene concordance factors (gCF) were calculated for **Papers III and IV** and in some cases were useful in identifying regions in the phylogenetic tree that might have biological discordance. To put simply, gCF is calculated for each branch in a species tree, which could be either from ML or summary methods, and it shows the percentage of gene trees that supported the same branch as the species tree. Around a branch there are four possible groups (can be a terminal branch, or a clade), and depending on the way this branch is resolved, the arrangement of these four groups would differ. These metrics are retrieved from IQ-TREE in the following way, gCF is the percentage of gene trees with the exact same branch (same arrangement of four groups), gDF1 and gDF2 are the two possible alternative resolutions of this branch, and gDFP is essentially all the other gene-tree topologies. The distributions of these four values can be seen from **Figure 13a**

In both papers I argue that although the gene concordance factors are useful in principle, in practice it can be difficult to use them to their full potential. The uncertainty in the gene-tree topologies hinders the ability of these measures to give meaningful insights and this can

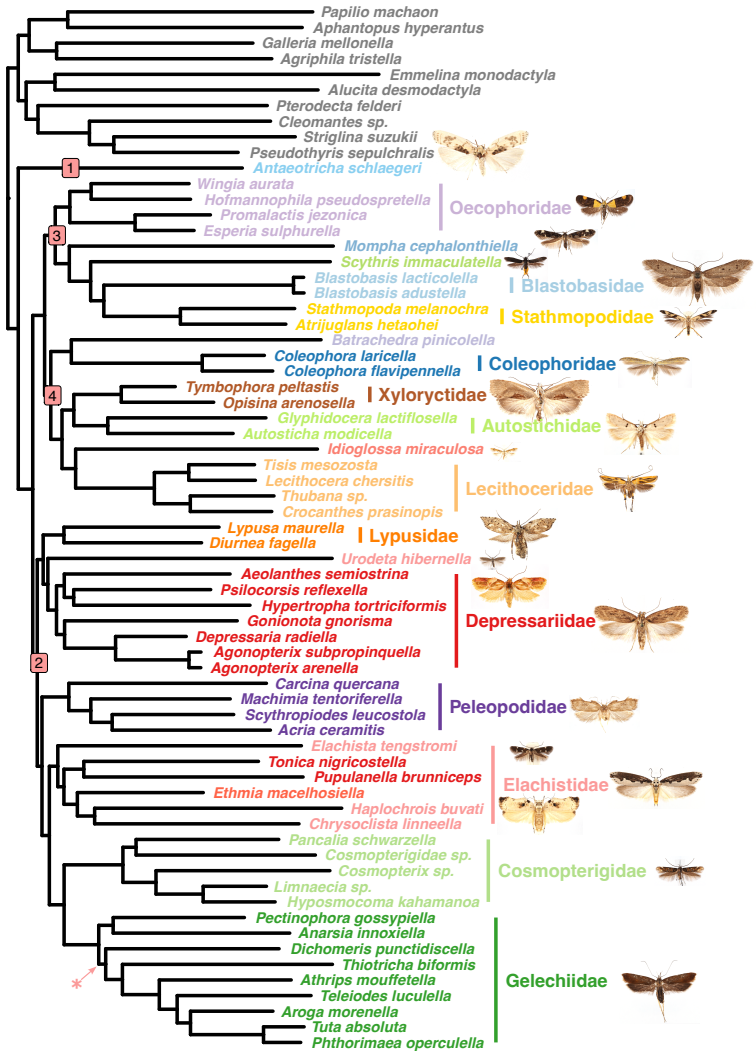

Figure 11: The maximum-likelihood phylogenomic tree of Gelechioidea. Figure from **Paper III**.

be seen in **Figure 13b**, where the gDFP values for the Gelechioidea dataset show a highly left-skewed distribution whereas this problem does not exist in the Erebidae dataset. It has been discussed before that sometimes the insufficient size of the alignments (number of site patterns) might not be enough to properly resolve the phylogenetic relationships based on a single gene (Rota et al., 2022). If this was the only source of gene-tree estimation error (GTEE) in these two datasets (giving rise to overall high gDFP values), one might expect to observe generally higher gDFP values in the more taxon-rich dataset (i.e. Erebidae) among these two datasets. On the other hand, assuming that the only, or the major, source of

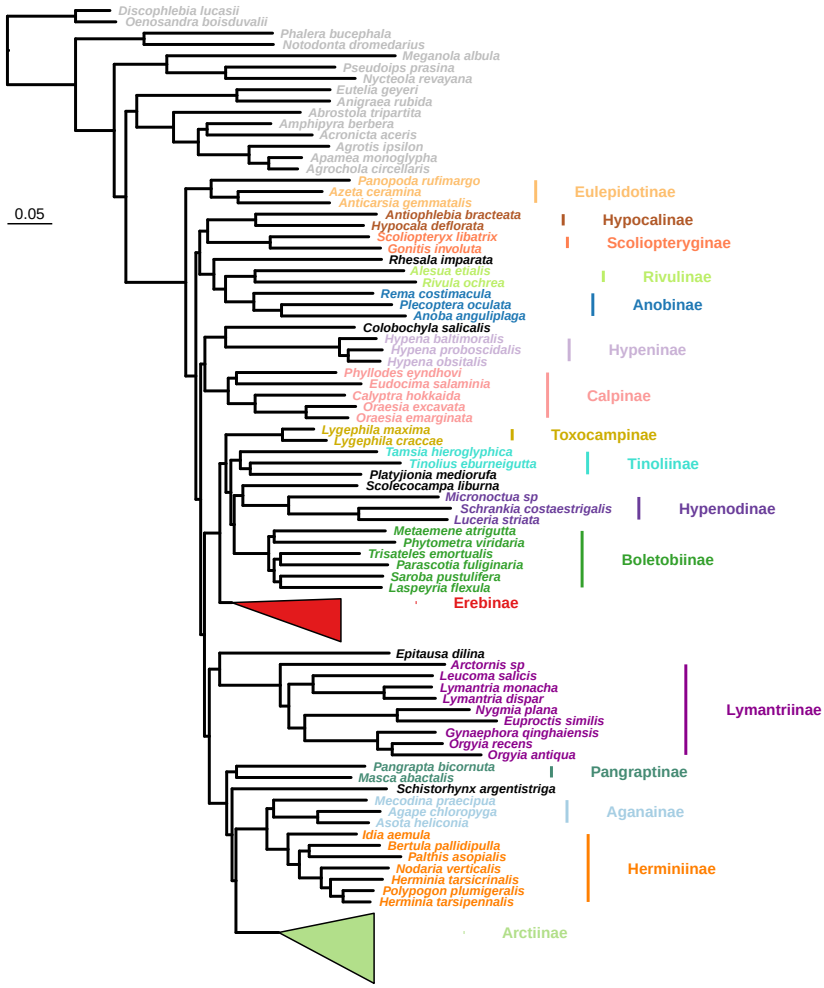

Figure 12: **The maximum-likelihood phylogeny of Erebidae** with subfamilies Arctiinae and Erebinae collapsed. Figure from Paper IV.

GTEE in these datasets should be the taxa/sites ratios of genes is highly unrealistic and another explanation may be more likely; that is the estimated age of the MRCA of the respective groups. Based on the previous work (Kawahara et al., 2019; Li et al., 2024; Wahlberg, Wheat, & Peña, 2013), Gelechioidea started diversifying approximately 80–100 MYA or earlier, whereas the estimated age for Erebidae is much younger, around 50 MYA or slightly older.

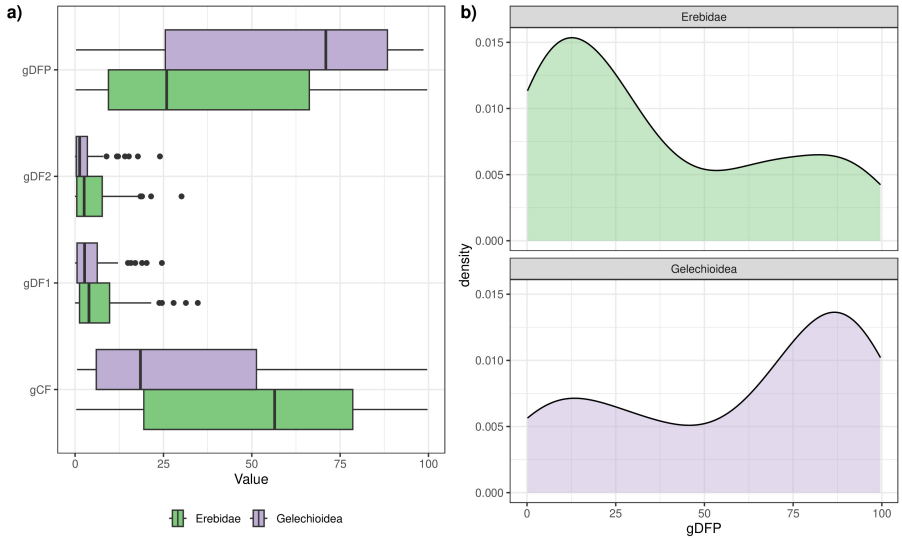

Figure 13: Comparison of gene concordance (and discordance) values of Erebidae (Paper IV) and Gelechioidea (Paper III) datasets. **a)** Boxplots of gCF, gDF1, gDF2, and gDFP values across the two datasets; **b)** Density plots of the gDFP values from the two datasets.

Outlier and paralogous sequences are prevalent in phylogenomic data of multiple sources

Filtering singleton outlier sequences and potential paralogs was necessary due to inevitability of false-positive hits when searching for genes in a given assembly bioinformatically. For **Papers I & II**, the gene set was an *a priori* defined probe set whereas for **Papers III and IV**, the Lepidoptera orthologous gene set as defined by OrthoDB v10 (Kriventseva et al., 2019) was used. OrthoDB v10 orthologs for Lepidoptera was originally defined based on only 16 Lepidoptera genomes out of which eight were butterflies (a single superfamily). Because of the limited taxon sampling when defining the ortholog set, it is highly plausible that there were gene losses and duplications that were not accounted for. Across all four papers, 2–5% of loci were removed completely from each of the respective datasets, due to them not satisfying minimum taxon coverage thresholds after the removal of outlier or paralog sequences. Overall, up to 60% of genes in a given dataset contained at least one outlier sequence. Based on the number of genes modified with the gene tree-based outlier and paralog filtering approaches, it is clear that one or another strategy for outlier filtering was needed to reliably compile phylogenetic datasets, especially when near-complete genomes and those that were generated by whole-genome resequencing (or genome skimming) techniques which result in highly fragmented genomes.

A reproducible phylogenetic dataset pipeline facilitates combining genomic data

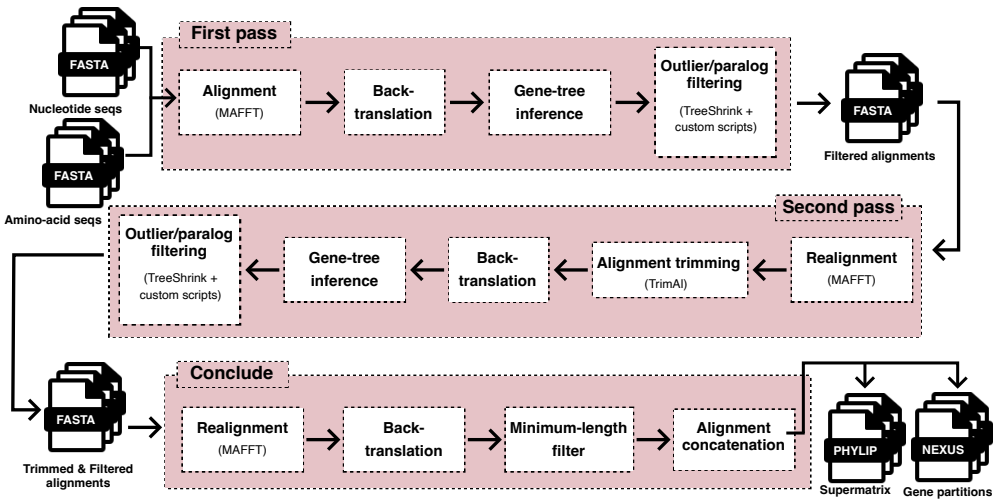

Figure 14: A workflow diagram describing the PCC pipeline. The three main phases of the pipeline are shown in non-overlapping colored rectangles each of which include several steps of data processing, which are shown in white rectangles. Multiple sequence alignment, gene-tree inference and outlier/paralog filtering steps are applied to each alignment separately in two passes: before and after the alignment-trimming step using TrimAl, i.e. removing gap-rich and highly variable columns from the alignments.

Given the added value of using non-reference genomic data and Sanger-sequencing data together with the available reference genomes, and the need for cautious outlier and paralog filtering in case of this data integration approach, I developed a phylogenetic data pre-processing pipeline called **PCC** to automate this process in a reproducible and customisable way. The pipeline essentially allows for using any kind of protein-coding gene and user-specified settings for stringency to be applied during the outlier filtering process. The end result is a concatenated phylogenetic dataset together with gene coordinates that is readily usable in a maximum-likelihood phylogenetic inference program of choice, such as IQ-TREE or RAxML-NG.

The pipeline expects extracted in-frame coding sequences as input and keeps the sequences in frame by performing alignments on the translated amino-acid sequences and then back-translating the amino-acid alignments into nucleotide sequences at each step that requires alignment. The task of outlier and paralog filtering is performed through two passes on each gene alignment before and after alignment filtering by TrimAl, which is one of the most popular tools for the tasks of removing gap-rich and overly variable regions in the alignment to help clarify the phylogenetic signal in each alignment. A workflow diagram is provided for the pipeline in **Figure 14**.

Available robust phylogenies enable in-depth exploration of evolution of winter diapause across butterflies

In **Paper V**, we gathered data on presence and the life stage of diapause (adult, pupa, larva, egg) for selected butterfly species through an extensive literature survey. Then, relying on a previously published global time-calibrated phylogenetic tree of butterflies (Chazot et al., 2019), we employed modern phylogenetic comparative methods to study the evolution of this complex adaptation. We modeled this trait with the five original states (four different life stages for entering diapause, and the state of no diapause), as well as by recoding these states into three- (adult, juvenile, no diapause), and two-state representations to reduce the burden of having rare states that could cause biases with the Mk models. Our analysis addressed key questions about the evolutionary history of diapause in butterflies, specifically focusing on when the trait first appeared and whether its current distribution is better explained by an ancestral origin followed by losses, or by multiple independent gains and losses throughout the butterfly phylogeny.

Although ancestral reconstructions at deeper nodes of the phylogeny highly varied across the six different analyses performed within each diapause state coding strategy —different root priors (Fitzjohn or flat prior), and the three different models Mk, hidden-rates, gamma-distributed rates— reconstructions were consistent regarding the first appearances of diapause at some life stages across the phylogeny, e.g. the pupal diapause strategy seems to have first appeared at the root of Papilionidae (ca. 68 MYA, see **Figure 15 A,D**)

Moreover, the rates of change across different models for each diapause state coding showed that, in general the loss of diapause was more common than its gain across the phylogeny; and that the rates of change from one diapause strategy to another were generally lower, meaning that the majority of gaining e.g. larval diapause events were from a no-diapause state and not from other life stages of diapausing. We argued that this might indicate that the underlying machinery to enter diapause at different life stages involves highly specialised strategies. Finally, we expected the overall rate of gain of diapause would increase across taxa as hypothesised before as a cold adaptation around the Eocene-Oligocene glacial maximum (EOGM), starting approximately 35 MYA. However, we could not find strong patterns indicating an increase in the gain of diapause that would coincide with the EOGM event.

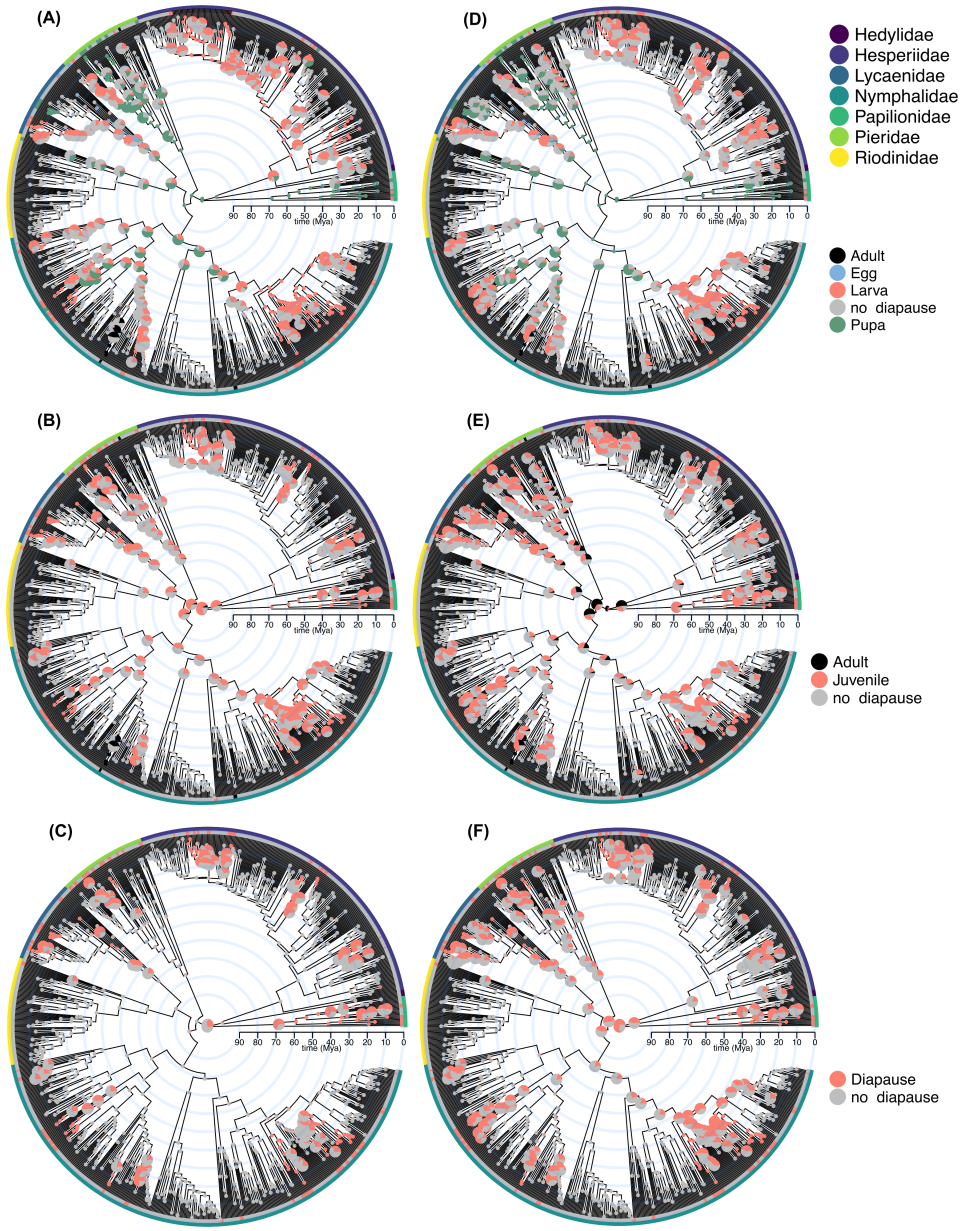

Figure 15: **Ancestral state reconstructions for the fitzjohn prior** with five-(top), three-(middle), and two-state (bottom row) modeling of winter diapause using stochastic mapping (left column) and the hidden-rates model (right column) across the butterfly phylogeny. Figure from **Paper V**.

Conclusions and Outlook

Throughout the first four papers that constitute this thesis, I have inferred robust phylogenies for four extremely species-rich groups of Lepidoptera: superfamily Gelechioidea, and families Geometridae, Noctuidae, and Erebidae. Taken as a whole, the phylogenetic results underline the usefulness of genomic scale data in phylogenetic studies, even if the data at hand comes from different methods/assays such as reference genome assemblies, transcriptomic or genomic raw data, or data generated with the Anchored-Hybrid Enrichment method. Moreover, two out of four papers highlight that the data generated in earlier studies with low-throughput methods can still provide immense value in improving taxon sampling where there is a lack of genomic data for those taxa.

With the emergence of collective sequencing projects that are at country or regional scales such as the Darwin Tree of Life project (The Darwin Tree of Life Project Consortium, 2022), which have generated a high number of Lepidoptera genome assemblies even though the project's focus was not on Lepidoptera; as well as the Psyche project (Wright et al., 2025) with the ultimate goal of sequencing and generating reference genomes from all 11,000 Lepidoptera species in Europe, currently we are living in the golden era for phylogenomic studies of Lepidoptera at various scales —from genera to families to superfamilies— or even such studies that can realistically produce better resolved backbone topologies and provide more reliable estimates of times of divergence for the entire Lepidoptera. The total number of Lepidoptera reference genomes has recently surpassed the 1000 mark and as the Psyche community, we are together in the first steps of this exciting journey. Even though there are more than 1000 Lepidoptera reference genomes at the chromosome scale right now, some groups are still poorly sampled relative to others. The existing sea of available transcriptomic data comes into play at this point to at least modestly improve the sampling for such relatively poorly sampled groups of Lepidoptera.

Although these sets of reference genomes are a remarkable resource, they do not solve the taxon sampling problem, especially for studies focusing on a family or a superfamily of Lepidoptera. The need for combining these reference genomes with alternative genomic/transcriptomic and legacy data is therefore greater for those types of studies. Across

the first four papers in this thesis, I provide a valuable resource for this purpose through not only applying this data integration approach I formulated with various combinations of data sources, but also through the reproducible data compilation pipeline I introduce in **Paper IV** which can be used to create phylogenetic datasets that will yield robust phylogenies. The pipeline has been available on GitHub under https://github.com/etkayapar/pcc for a while and it has already helped to infer well-support phylogenies of various groups of insects which were being studied in the Systematic Biology Group here at Lund University. These include a Master's thesis project on the evolutionary history of the sloth moths (Pyralidae: Chrysauginae) by a student I co-supervised; a PhD and a Master's project on the systematics of a group of wasps; as well as the phylogenomic study of a previously unstudied family of moths, Thyrididae. All of these projects utilised single copy orthologous genes from majority whole-genome skimming data.

However, this data integration approach is not without limitations. The greatest limitation is that depending on the proportion of non-reference genome or transcriptome being used, this approach can result in loss of too many sequences during the process of filtering outliers and paralogs. It is established, however, that the associated pipeline I introduce is effective at removing noise from the phylogenetic datasets that would otherwise be amplified, when the input data is a true mix of reference genomes and other genomic resources. Another aspect where there is still room for improvement is that through benchmarking on empirical or simulated datasets I can offer sane defaults for the future users of this pipeline.

Finally, I want to underline once again that a well-supported and sampled phylogeny is a must in order to tackle challenging evolutionary questions. This is true not only for macroevolutionary questions in the way I demonstrated in **Paper V**, but also for the kind of comparative genomics works that are starting to be more popular and feasible thanks to the increasingly more available reference genomes. For Lepidoptera, specifically, as we get increasingly closer to inferring an order-level phylogeny with more than 1000 reference genomes, to use the same data point for both phylogenetic inference and as the collection of evolving traits themselves in the comparative sense is no longer a mere possibility, but a necessity that the —very near— future will bring.

It is truly an exciting future for those of us who have a special place for phylogenies in their (scientific) heart!

Acknowledgements

Let's get real: I won't be fooling any one of you if I wrote an acknowledgements sections with references full of funny memories, or inside jokes with many people. I, and those of you who know me, know that I am no social butterfly :) But many of you touched my life during these four years, offered a helping hand even when I didn't know I needed help, or when I did not know how to ask help. I want to thank you here, thank you for being a part of my life, and I hope that my introvertedness did not give you the wrong idea that I did not value spending time with you, the opposite, you kept me going and sane throughout!

I would like start with **Niklas**. I know everyone says "this would not have been possible without my supervisor" and it is true for most of those cases when people say it. But I want to stress that I truly mean this. Thank you for introducing me to the world of Lepidoptera when I was a primate testicle person, for times when your full confidence in me made me realize what I fool I was doubting whether I will be enough, for always being available when I needed to discuss some overly technical things about tree inference programs, and helping me at times when I did not realise how much I needed it. Thank you also for promptly directing me when I felt lost by showing me I was almost already at the exit! **Jadranka**, I don't know how you do it but every time when I have a meeting with you I leave your office as a completely optimistic and ambitious person who is ready to tackle everthing ahead of him! Also thank you for the great chats on different languages in the world, I think this is one of the non-professional passions we deeply share :) Thank you both for everything that I would not be able to list comprehensively here, but especially for your patience during the last several weeks of this writing process when I was probably not the happiest or easiest-to-deal-with person :)

Thank you **Helena** for keeping an eye on this young researcher throughout the years, and your trust in me! I also would like to thank **Mikael** for the opportunity to peek into the modeling world of things and being such a calm supervisor. Also, **Krzysztof** and **Staffan**, I am sad that I did not actively try learning more from you and spending more time. It was my mistake to not the take the opportunity. But thank you all of my supervisors and my scientific mentor and my IR for all the occasions in which you helped me in ways you

cannot imagine.

Thank you **Hamid** for being there every time I needed — even before I came here you were there helping me :) Without you guiding me through living in Sweden, I would probably have done something very stupid along the way. And I would tell you that you are one of the funniest people I met, but with that many failed jokes in between, you are bound to have some really great ones I think :P But seriously, I am deeply thankful that I got to spend this much time with you over the years! **Sridhar**, thank you for all the trips we had together, long chats along the way about the meaning of life, about doing research so far away from our home countries and all the times you encouraged me to socialize in occasions where I would not be able to on my own. And thank you so much for providing the opportunity for getting my foot back in the PCM world after my masters, I think at least half of the fun is using these phylogenies. I will never be able to forget that one adventurous flight back home from Paris we experienced! **Emma**, my first office mate, thank you for always being a good friend, and also keeping me busy with involving me in side projects! I think there were times I really needed to not think about my own things for a while. Oh also for bringing this festive energy by encouraging people to celebrate their wins (small and big) in life :) **Zach** if for nothing I should thank you for talking so much about Snakemake around us that it made me want to take my unordered scripts that were ductaped together and turn them into a Snakemake pipeline of my own :) But more seriously, thank you for answering my countless questions on a lot of topics and for my first Julbord we had together during my first Christmas here! Thank you **Jöran** for all the baked goods, and cultural and botanical information I got to learn about, our chats about film photography, and of course being an awesome office mate. I hope you will finally able to nicely sort out your polyploid problems :) **Vineesh**, I must say I truly had a blast working together and being your friend!

My longest-time office mate together with Jöran, **Josefin**: Thank you for being such a wonderful office mate all these years and encouraging me to contact **Pedro** to take part in teaching R. Now, **Pedro** and **Iain**, thank you both for the times we had together on both courses and all the interesting chats about the Advent of Code! **Simon**, thank you so much that you always had deep and hard to answer scientific questions on the lunch table, it kept my scientific brain alive at times. And thank you **Robin** for helping me with my thesis questions, and for all the days we were office mates!

Thank you to all the nice people I wouldn't be able to meet elsewhere: **Azin, Arash, Parminder, Ruben, Alex, Lila, Lucia, Violeta, Sofie, Yedra, Arrian, Ciara, Zsófia, Quentin** and many more people I probably forgot and will get mad at myself because of it as soon as I send this thesis in...

Thank you so much!

To start doing a PhD, I had to have other degrees :) And there are more people I want to

thank that I can write in detail here from those times. But I need to mention at least one person, **Mehmet**: Bu satırları yazdığım dakikalardan yaklaşık iki hafta önce Stockholmde buluşmamız doktoramın sonuna doğru sönmeye başlayan araştırma hevesini yeniden alevlendirmeye yetti de arttı. Belki tanışalı 10 yıldan fazla oluyor, beraber çalışmaya başlayalı da neredeyse 10 yıl olacak. CompEvo'daki o bazen kaotik olabilse de cıvıl cıvıl sosyal ve bilimsel atmosferi ve labdaki herkesi her gün daha da çok özlüyorum. Umarım beraber bilim üretmeye devam ederiz!

Anne, Baba, Esra, Enes, Nahide, İyi ki varsınız <3

There are no words that could truly describe the love I have for my wife **Ece**. Senin de her zaman dediğin gibi, nasıl oluyor bilmiyorum ama sanki hiç bunca yıl geçmemiş gibi, sanki hala dün tanışmışız gibi hissediyorum. Yabancı bir ülkede, bu inişli çıkışlı dört yıl boyunca sensiz ne yapardım hiç düşünmek bile istemiyorum. Gerçekten hayatımın en büyük neşesi ve onsuz yapamayacağım biriciğimsin <3 İnsanın önünü bile göremeyecek halde dibe batmış hissederken onu derin çukurlardan çıkarıp alacak birine sahip olmasının ne kadar güven verici olduğunu sanıyorum benim hissetiğim kadar derinden hissetmemişsindir. Ne diyeyim ben de seni bu kadar mutlu edebiliyorsam kendimi çok şanslı addederim. Bugün bu tezi bitirebiliyorsam, başka yönlerden hayattan belki de pes etmediysem, bunları ve çok daha fazlasını sana borçluyum, hepsi senin sayende aşkım. Bu hayatta en büyük dileğim seninle birlikte yaşlanmak ve bu günlere dönüp baktığımda yaşadığımız bu irili ufaklı zorlukları görüp ne kadar uzaklarda kaldıklarını görüp gülümseyebilmek. İyi ki varsın ve iyi ki benimlesin <3 <3

References

Abraham, D., Ryrholm, N., Wittzell, H., Holloway, J. D., Scoble, M. J., & Löfstedt, C. (2001, July). Molecular Phylogeny of the Subfamilies in Geometridae (Geometroidea: Lepidoptera). *Molecular Phylogenetics and Evolution*, *20*(1), 65–77. doi: 10.1006/mpev.2001.0949

Ban, X., Jiang, N., Cheng, R., Xue, D., & Han, H. (2018, October). Tribal classification and phylogeny of Geometrinae (Lepidoptera: Geometridae) inferred from seven gene regions. *Zoological Journal of the Linnean Society*, *184*(3), 653–672. doi: 10.1093/zoolinnean/zly013

Bánki, O., Roskov, Y., Döring, M., Ower, G., Hernández Robles, D. R., Plata Corredor, C. A., … World Flora Online (2025). *Catalogue of Life.* Catalogue of Life Foundation. doi: 10.48580/DGTPL

Bänziger, H. (2007). Skin-piercing blood-sucking moths VI : Fruit-piercing habits in Calyptra (Noctuidae) and notes on the feeding strategies of zoophilous and frugivorous adult Lepidoptera.
doi: 10.5169/SEALS-402952

Bazinet, A. L., Cummings, M. P., Mitter, K. T., & Mitter, C. W. (2013, December). Can RNA-Seq Resolve the Rapid Radiation of Advanced Moths and Butterflies (Hexapoda: Lepidoptera: Apoditrysia)? An Exploratory Study. *PLOS ONE*, *8*(12), e82615. doi: 10.1371/journal.pone.0082615

Beaulieu, J., O'Meara, B., Oliver, J., & Boyko, J. (2022, June). *corHMM: Hidden Markov Models of Character Evolution.*

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

Brakefield, P. M., Beldade, P., & Zwaan, B. J. (2009, January). The African Butterfly Bicyclus anynana: A Model for Evolutionary Genetics and Evolutionary Developmental Biology. *Cold Spring Harbor Protocols*, *2009*(5), pdb.emo122. doi: 10.1101/pdb.emo122

Brawand, D., Soumillon, M., Necsulea, A., Julien, P., Csárdi, G., Harrigan, P., … Kaessmann, H. (2011, October). The evolution of gene expression levels in mammalian

organs. *Nature*, *478*(7369), 343–348. doi: 10.1038/nature10532

Buchfink, B., Reuter, K., & Drost, H.-G. (2021, April). Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nature Methods*, *18*(4), 366–368. doi: 10 .1038/s41592-021-01101-x

Cantu, V. A., Sadural, J., & Edwards, R. (2019, February). *PRINSEQ++, a multi-threaded tool for fast and efficient quality control and preprocessing of sequencing datasets* (Preprint). PeerJ Preprints. doi: 10.7287/peerj.preprints.27553v1

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

Castresana, J. (2000, April). 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

Chazot, N., Condamine, F. L., Dudas, G., Peña, C., Kodandaramaiah, U., Matos-Maraví, P., … Wahlberg, N. (2021, September). Conserved ancestral tropical niche but different continental histories explain the latitudinal diversity gradient in brush-footed butterflies. *Nature Communications*, *12*(1), 5717. doi: 10.1038/s41467-021-25906-8

Chazot, N., Wahlberg, N., Freitas, A. V. L., Mitter, C., Labandeira, C., Sohn, J.-C., … Heikkilä, M. (2019, September). Priors and Posteriors in Bayesian Timing of Divergence Analyses: The Age of Butterflies Revisited. *Systematic Biology*, *68*(5), 797–813. doi: 10.1093/sysbio/syz002

Chiarenza, A. A., Fabbri, M., Consorti, L., Muscioni, M., Evans, D. C., Cantalapiedra, J. L., & Fanti, F. (2021, December). An Italian dinosaur Lagerstätte reveals the tempo and mode of hadrosauriform body size evolution. *Scientific Reports*, *11*(1), 23295. doi: 10.1038/s41598-021-02490-x

Cho, S., Zwick, A., Regier, J. C., Mitter, C., Cummings, M. P., Yao, J., … Parr, C. (2011, December). Can Deliberately Incomplete Gene Sample Augmentation Improve a Phylogeny Estimate for the Advanced Moths and Butterflies (Hexapoda: Lepidoptera)? *Systematic Biology*, *60*(6), 782–796. doi: 10.1093/sysbio/syr079

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. *Journal of Biogeography*, *40*(1), 9–23. doi: 10.1111/j.1365-2699.2012.02787.x

Crotty, S. M., Minh, B. Q., Bean, N. G., Holland, B. R., Tuke, J., Jermiin, L. S., & Haeseler, A. V. (2020, March). GHOST: Recovering Historical Signal from Heterotachously Evolved Sequence Alignments. *Systematic Biology*, *69*(2), 249–264. doi: 10.1093/sysbio/syz051

Darwin, C. (1859). *On the Origin of Species by Means of Natural Selection*. London: John Murray.

Espeland, M., Nakahara, S., Zacca, T., Barbosa, E. P., Huertas, B., Marín, M. A., …

Willmott, K. R. (2023). Combining target enrichment and Sanger sequencing data to clarify the systematics of the diverse Neotropical butterfly subtribe Euptychiina (Nymphalidae, Satyrinae). *Systematic Entomology*, *48*(4), 498–570. doi: 10.1111/syen.12590

Felsenstein, J. (1973, September). Maximum-likelihood estimation of evolutionary trees from continuous characters. *American Journal of Human Genetics*, *25*(5), 471–492.

Felsenstein, J. (1981, November). Evolutionary trees from DNA sequences: A maximum likelihood approach. *Journal of Molecular Evolution*, *17*(6), 368–376. doi: 10.1007/BF01734359

Felsenstein, J. (1985, July). Confidence limits on phylogenies: An approach using the bootstrap. *Evolution*, *39*(4), 783–791. doi: 10.1111/j.1558-5646.1985.tb00420.x

Fibiger, M., & Lafontaine, J. D. (2005). A review of the higher classification of the Noctuoidea (Lepidoptera) with special reference to the Holarctic fauna. *Esperiana*, *11*, 7–92.

Fonseca, L. H. M., & Lohmann, L. G. (2018, June). Combining high-throughput sequencing and targeted loci data to infer the phylogeny of the "*Adenocalymma-Neojobertia*" clade (Bignonieae, Bignoniaceae). *Molecular Phylogenetics and Evolution*, *123*, 1–15. doi: 10.1016/j.ympev.2018.01.023

Ghanavi, H. R., Twort, V., Hartman, T. J., Zahiri, R., & Wahlberg, N. (2022). The (non) accuracy of mitochondrial genomes for family-level phylogenetics in Erebidae (Lepidoptera). *Zoologica Scripta*, *51*(6), 695–707. doi: 10.1111/zsc.12559

Gillard, G. B., Grønvold, L., Røsæg, L. L., Holen, M. M., Monsen, Ø., Koop, B. F., … Hvidsten, T. R. (2021, April). Comparative regulomics supports pervasive selection on gene dosage following whole genome duplication. *Genome Biology*, *22*(1), 103. doi: 10.1186/s13059-021-02323-0

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., … Regev, A. (2011, July). Trinity: Reconstructing a full-length transcriptome without a genome from RNA-Seq data. *Nature biotechnology*, *29*(7), 644. doi: 10.1038/nbt.1883

Haeckel, E. (1866). *Generelle Morphologie der Organismen*. Berlin: G. Reimer.

Heikkilä, M., Kaila, L., Mutanen, M., Peña, C., & Wahlberg, N. (2011, September). Cretaceous origin and repeated tertiary diversification of the redefined butterflies. *Proceedings of the Royal Society B: Biological Sciences*, *279*(1731), 1093–1099. doi: 10.1098/rspb.2011.1430

Heikkilä, M., Mutanen, M., Kekkonen, M., & Kaila, L. (2014). Morphology reinforces proposed molecular phylogenetic affinities: A revised classification for Gelechioidea (Lepidoptera). *Cladistics*, *30*(6), 563–589. doi: 10.1111/cla.12064

Heikkilä, M., Mutanen, M., Wahlberg, N., Sihvonen, P., & Kaila, L. (2015). Elusive ditrysian phylogeny: An account of combining systematized morphology with molecular data (Lepidoptera). *BMC Evolutionary Biology*, *15*(1), 1–27. doi: 10.1186/s12862-015-0520-0

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

Hoile, A. E., Holland, P. W. H., & Mulhair, P. O. (2025, February). Gene novelty and gene family expansion in the early evolution of Lepidoptera. *BMC Genomics*, *26*(1), 161. doi: 10.1186/s12864-025-11338-x

Hughes, J. J., Berv, J. S., Chester, S. G. B., Sargis, E. J., & Field, D. J. (2021). Ecological selectivity and the evolution of mammalian substrate preference across the K–Pg boundary. *Ecology and Evolution*, *11*(21), 14540–14554. doi: 10.1002/ece3.8114

Jeffroy, O., Brinkmann, H., Delsuc, F., & Philippe, H. (2006, April). Phylogenomics: The beginning of incongruence? *Trends in Genetics*, *22*(4), 225–231. doi: 10.1016/j.tig.2006.02.003

Jiggins, C. D., Mallarino, R., Willmott, K. R., & Bermingham, E. (2006). The Phylogenetic Pattern of Speciation and Wing Pattern Change in Neotropical Ithomia Butterflies (lepidoptera: Nymphalidae). *Evolution*, *60*(7), 1454–1466. doi: 10.1111/j.0014-3820.2006.tb01224.x

Johnson, M. G., Gardner, E. M., Liu, Y., Medina, R., Goffinet, B., Shaw, A. J., … Wickett, N. J. (2016). HybPiper: Extracting coding sequence and introns for phylogenetics from high-throughput sequencing reads using target enrichment. *Applications in Plant Sciences*, *4*(7), 1600016. doi: 10.3732/apps.1600016

Jukes, TH., & Cantor, CR. (1969). Evolution of protein molecules. In HM. Munro (Ed.), *Mammalian protein metabolism* (pp. 21–120). Academic Press, New York.

Kaila, L., Mutanen, M., & Nyman, T. (2011, December). Phylogeny of the mega-diverse Gelechioidea (Lepidoptera): Adaptations and determinants of success. *Molecular Phylogenetics and Evolution*, *61*(3), 801–809. doi: 10.1016/j.ympev.2011.08.016

Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., von Haeseler, A., & Jermiin, L. S. (2017, June). ModelFinder: Fast model selection for accurate phylogenetic estimates. *Nature Methods*, *14*(6), 587–589. doi: 10.1038/nmeth.4285

Katoh, K., & Standley, D. M. (2013, April). 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

Kawahara, A. Y., & Breinholt, J. W. (2014, August). Phylogenomics provides strong evidence for relationships of butterflies and moths. *Proceedings of the Royal Society B: Biological Sciences*, *281*(1788), 20140970. doi: 10.1098/rspb.2014.0970

Kawahara, A. Y., Plotkin, D., Espeland, M., Meusemann, K., Toussaint, E. F., Donath, A., … Breinholt, J. W. (2019). Phylogenomics reveals the evolutionary timing and pattern of butterflies and moths. *Proceedings of the National Academy of Sciences of the United States of America*, *116*(45), 22657–22663. doi: 10.1073/pnas.1907847116

Kawahara, A. Y., Storer, C., Carvalho, A. P. S., Plotkin, D. M., Condamine, F. L., Braga, M. P., … Lohman, D. J. (2023, June). A global phylogeny of butterflies reveals their evolutionary history, ancestral hosts and biogeographic origins. *Nature Ecology*

& Evolution, *7*(6), 903–913. doi: 10.1038/s41559-023-02041-9
Keegan, K. L., Rota, J., Zahiri, R., Zilli, A., Wahlberg, N., Schmidt, B. C., … Wagner, D. L. (2021, May). Toward a Stable Global Noctuidae (Lepidoptera) Taxonomy. *Insect Systematics and Diversity*, *5*(3), 1. doi: 10.1093/isd/ixab005
Kishino, H., & Hasegawa, M. (1989, August). Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in hominoidea. *Journal of Molecular Evolution*, *29*(2), 170–179. doi: 10.1007/BF02100115
Kishino, H., Miyata, T., & Hasegawa, M. (1990, August). Maximum likelihood inference of protein phylogeny and the origin of chloroplasts. *Journal of Molecular Evolution*, *31*(2), 151–160. doi: 10.1007/BF02109483
Kodandaramaiah, U., & Wahlberg, N. (2009). Phylogeny and biogeography of Coenonympha butterflies (Nymphalidae: Satyrinae) – patterns of colonization in the Holarctic. *Systematic Entomology*, *34*(2), 315–323. doi: 10.1111/j.1365-3113.2008.00453.x
Kozlov, A. M., Darriba, D., Flouri, T., Morel, B., & Stamatakis, A. (2019, November). RAxML-NG: A fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. *Bioinformatics*, *35*(21), 4453–4455. doi: 10.1093/bioinformatics/btz305
Kristensen, N. P., & Skalski, A. W. (1999). Phylogeny and Palentology. In N. P. Kristensen (Ed.), *Lepidoptera: Moths and butterflies. Handbook of Zoology/Handbuch der Zoologie 35* (Vol. 1, pp. 7–25). Berlin: Walter de Gruyter.
Kriventseva, E. V., Kuznetsov, D., Tegenfeldt, F., Manni, M., Dias, R., Simão, F. A., & Zdobnov, E. M. (2019, January). OrthoDB v10: Sampling the diversity of animal, plant, fungal, protist, bacterial and viral genomes for evolutionary and functional annotations of orthologs. *Nucleic Acids Research*, *47*(D1), D807-D811. doi: 10.1093/nar/gky1053
Kück, P., Meusemann, K., Dambach, J., Thormann, B., von Reumont, B. M., Wägele, J. W., & Misof, B. (2010, March). Parametric and non-parametric masking of randomness in sequence alignments can be improved and leads to better resolved trees. *Frontiers in Zoology*, *7*(1), 10. doi: 10.1186/1742-9994-7-10
Lamarck, JBM. (1809). *Philosophie zoologique*. Paris: Dentu.
Lanfear, R., & Hahn, M. W. (2024, November). The Meaning and Measure of Concordance Factors in Phylogenomics. *Molecular Biology and Evolution*, *41*(11), msae214. doi: 10.1093/molbev/msae214
Larouche, O., Gartner, S. M., Westneat, M. W., & Evans, K. M. (2023, March). Mosaic Evolution of the Skull in Labrid Fishes Involves Differences in Both Tempo and Mode of Morphological Change. *Systematic Biology*, *72*(2), 419–432. doi: 10.1093/sysbio/syac061
Lewis, P. O. (2001, November). A Likelihood Approach to Estimating Phylogeny from Discrete Morphological Character Data. *Systematic Biology*, *50*(6), 913–925. doi:

10.1080/106351501753462876
Li, X., Breinholt, J. W., Martinez, J. I., Keegan, K., Ellis, E. A., Homziak, N. T., … Kawahara, A. Y. (2024). Large-scale genomic data reveal the phylogeny and evolution of owlet moths (Noctuoidea). *Cladistics*, *40*(1), 21–33. doi: 10.1111/cla.12559
Liu, C., Zhou, X., Li, Y., Hittinger, C. T., Pan, R., Huang, J., … Shen, X.-X. (2024, October). The Influence of the Number of Tree Searches on Maximum Likelihood Inference in Phylogenomics. *Systematic Biology*, *73*(5), 807–822. doi: 10.1093/sysbio/syae031
Mai, U., & Mirarab, S. (2018, May). TreeShrink: Fast and accurate detection of outlier long branches in collections of phylogenetic trees. *BMC Genomics*, *19*(5), 272. doi: 10.1186/s12864-018-4620-2
Martin, M. (2011, May). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*, *17*(1), 10–12. doi: 10.14806/ej.17.1.200
Mayer, C., Dietz, L., Call, E., Kukowka, S., Martin, S., & Espeland, M. (2021). Adding leaves to the Lepidoptera tree: Capturing hundreds of nuclear genes from old museum specimens. *Systematic Entomology*, *46*(3), 649–671. doi: 10.1111/syen.12481
Middleton-Welling, J., Dapporto, L., García-Barros, E., Wiemers, M., Nowicki, P., Plazio, E., … Shreeve, T. (2020, October). A new comprehensive trait database of European and Maghreb butterflies, Papilionoidea. *Scientific Data*, *7*(1), 351. doi: 10.1038/s41597-020-00697-7
Minet, J. (1991, January). Tentative reconstruction of the ditrysian phylogeny (Lepidoptera: Glossata). *Insect Systematics & Evolution*, *22*(1), 69–95. doi: 10.1163/187631291X00327
Minh, B. Q., Schmidt, H. A., Chernomor, O., Schrempf, D., Woodhams, M. D., von Haeseler, A., & Lanfear, R. (2020, May). IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Molecular Biology and Evolution*, *37*(5), 1530–1534. doi: 10.1093/molbev/msaa015
Mitchell, A., Mitter, C., & Regier, J. C. (2006). Systematics and evolution of the cutworm moths (Lepidoptera: Noctuidae): Evidence from two protein-coding nuclear genes. *Systematic Entomology*, *31*(1), 21–46. doi: 10.1111/j.1365-3113.2005.00306.x
Mitter, C., Baixeras, J., Brown, J., Cho, S., Cummings, M., Davis, D., … others (2006). LepTree. net, a genomics-inspired community collaboration. In *The 2006 ESA annual meeting, december 10-13, 2006.*
Mitter, C., Davis, D. R., & Cummings, M. P. (2017). Phylogeny and Evolution of Lepidoptera. *Annual Review of Entomology*, *62*, 265–283. doi: 10.1146/annurev-ento-031616-035125
Murillo-Ramos, L., Brehm, G., Sihvonen, P., Hausmann, A., Holm, S., Ghanavi, H. R., … Wahlberg, N. (2019, August). A comprehensive molecular phylogeny of Geometridae (Lepidoptera) with a focus on enigmatic small subfamilies. *PeerJ*, *7*, e7386. doi: 10.7717/peerj.7386

Murillo-Ramos, L., Twort, V., Wahlberg, N., & Sihvonen, P. (2023). A phylogenomic perspective on the relationships of subfamilies in the family Geometridae (Lepidoptera). *Systematic Entomology*, *48*(4), 618–632. doi: 10.1111/syen.12594

Mutanen, M., Wahlberg, N., & Kaila, L. (2010, September). Comprehensive gene and taxon coverage elucidates radiation patterns in moths and butterflies. *Proceedings of the Royal Society B: Biological Sciences*, *277*(1695), 2839–2848. doi: 10.1098/rspb.2010.0392

Neyman, J. (1971, January). Molecular studies of evolution: A source of novel statistical problems. In S. S. Gupta & J. Yackel (Eds.), *Statistical Decision Theory and Related Topics* (pp. 1–27). Academic Press. doi: 10.1016/B978-0-12-307550-5.50005-8

Pagel, M. (1994, January). Detecting correlated evolution on phylogenies: A general method for the comparative analysis of discrete characters. *Proceedings of the Royal Society B: Biological Sciences*, *255*(1342), 37–45. doi: 10.1098/rspb.1994.0006

Peña, C., Nylin, S., & Wahlberg, N. (2011, January). The radiation of Satyrini butterflies (Nymphalidae: Satyrinae): A challenge for phylogenetic methods. *Zoological Journal of the Linnean Society*, *161*(1), 64–87. doi: 10.1111/j.1096-3642.2009.00627.x

Philippe, H., de Vienne, D. M., Ranwez, V., Roure, B., Baurain, D., & Delsuc, F. (2017). Pitfalls in supermatrix phylogenomics. *European Journal of Taxonomy*, *283*, 1. doi: 10.5852/ejt.2017.283

Prjibelski, A., Antipov, D., Meleshko, D., Lapidus, A., & Korobeynikov, A. (2020). Using SPAdes De Novo Assembler. *Current Protocols in Bioinformatics*, *70*(1), e102. doi: 10.1002/cpbi.102

Prum, R. O., Berv, J. S., Dornburg, A., Field, D. J., Townsend, J. P., Lemmon, E. M., & Lemmon, A. R. (2015, October). A comprehensive phylogeny of birds (Aves) using targeted next-generation DNA sequencing. *Nature*, *526*(7574), 569–573. doi: 10.1038/nature15697

Regier, J. C., Mitter, C., Zwick, A., Bazinet, A. L., Cummings, M. P., Kawahara, A. Y., … Mitter, K. T. (2013, March). A Large-Scale, Higher-Level, Molecular Phylogenetic Study of the Insect Order Lepidoptera (Moths and Butterflies). *PLOS ONE*, *8*(3), e58568. doi: 10.1371/journal.pone.0058568

Regier, J. C., Zwick, A., Cummings, M. P., Kawahara, A. Y., Cho, S., Weller, S., … Mitter, C. (2009, December). Toward reconstructing the evolution of advanced moths and butterflies (Lepidoptera: Ditrysia): An initial molecular study. *BMC Evolutionary Biology*, *9*(1), 280. doi: 10.1186/1471-2148-9-280

Revell, L. J. (2024, January). Phytools 2.0: An updated R ecosystem for phylogenetic comparative methods (and other things). *PeerJ*, *12*, e16505. doi: 10.7717/peerj.16505

Rota, J., Twort, V., Chiocchio, A., Peña, C., Wheat, C. W., Kaila, L., & Wahlberg, N. (2022). The unresolved phylogenomic tree of butterflies and moths (Lepidoptera): Assessing the potential causes and consequences. *Systematic Entomology*(October 2021), 1–20. doi: 10.1111/syen.12545

Sanderson, M. J., McMahon, M. M., & Steel, M. (2010, May). Phylogenomics with incomplete taxon coverage: The limits to inference. *BMC Evolutionary Biology*, *10*(1), 155. doi: 10.1186/1471-2148-10-155

Sela, I., Ashkenazy, H., Katoh, K., & Pupko, T. (2015, July). GUIDANCE2: Accurate detection of unreliable alignment regions accounting for the uncertainty of multiple parameters. *Nucleic Acids Research*, *43*(W1), W7-W14. doi: 10.1093/nar/gkv318

Shimodaira, H. (2002, May). An Approximately Unbiased Test of Phylogenetic Tree Selection. *Systematic Biology*, *51*(3), 492–508. doi: 10.1080/10635150290069913

Shimodaira, H., & Hasegawa, M. (1999, August). Multiple Comparisons of Log-Likelihoods with Applications to Phylogenetic Inference. *Molecular Biology and Evolution*, *16*(8), 1114. doi: 10.1093/oxfordjournals.molbev.a026201

Shirey, V., Larsen, E., Doherty, A., Kim, C. A., Al-Sulaiman, F. T., Hinolan, J. D., … Ries, L. (2022, July). LepTraits 1.0 A globally comprehensive dataset of butterfly traits. *Scientific Data*, *9*(1), 382. doi: 10.1038/s41597-022-01473-5

Sihvonen, P., Mutanen, M., Kaila, L., Brehm, G., Hausmann, A., & Staude, H. S. (2011, June). Comprehensive Molecular Sampling Yields a Robust Phylogeny for Geometrid Moths (Lepidoptera: Geometridae). *PLOS ONE*, *6*(6), e20356. doi: 10.1371/journal.pone.0020356

Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., & Zdobnov, E. M. (2015, October). BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, *31*(19), 3210–3212. doi: 10.1093/bioinformatics/btv351

Sisson, M. S., Dowdy, N. J., Fisher, M. L., Gall, L. F., Goldstein, P. Z., Homziak, N. T., … Zilli, A. (2025, May). Erebidae systematics: Past, present, and future—progress in understanding a diverse lepidopteran lineage. *Insect Systematics and Diversity*, *9*(3), 5. doi: 10.1093/isd/ixaf018

Sohn, J.-C., Regier, J. C., Mitter, C., Adamski, D., Landry, J.-F., Heikkilä, M., … Schmitz, P. (2016). Phylogeny and feeding trait evolution of the mega-diverse Gelechioidea (Lepidoptera: Obtectomera): New insight from 19 nuclear genes. *Systematic Entomology*, *41*(1), 112–132. doi: 10.1111/syen.12143

Steenwyk, J. L., Li, Y., Zhou, X., Shen, X.-X., & Rokas, A. (2023, June). Incongruence in the phylogenomics era. *Nature Reviews Genetics*, 1–17. doi: 10.1038/s41576-023-00620-x

Strimmer, K., & Rambaut, A. (2002, January). Inferring confidence sets of possibly misspecified gene trees. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, *269*(1487), 137–142. doi: 10.1098/rspb.2001.1862

Tan, G., Muffato, M., Ledergerber, C., Herrero, J., Goldman, N., Gil, M., & Dessimoz, C. (2015, September). Current Methods for Automated Filtering of Multiple Sequence Alignments Frequently Worsen Single-Gene Phylogenetic Inference. *Systematic Biology*, *64*(5), 778–791. doi: 10.1093/sysbio/syv033

The Darwin Tree of Life Project Consortium. (2022, January). Sequence locally, think

globally: The Darwin Tree of Life Project. *Proceedings of the National Academy of Sciences*, *119*(4), e2115642118. doi: 10.1073/pnas.2115642118

Townsend, T. M., Larson, A., Louis, E., & Macey, J. R. (2004, October). Molecular Phylogenetics of Squamata: The Position of Snakes, Amphisbaenians, and Dibamids, and the Root of the Squamate Tree. *Systematic Biology*, *53*(5), 735–757. doi: 10.1080/10635150490522340

van Nieukerken, E. J., Kaila, L., Kitching, I. J., Kristensen, N. P., Lees, D. C., Minet, J., … Zwick, A. (2011, December). Order Lepidoptera Linnaeus, 1758. In: Zhang, Z.-Q. (Ed.) Animal biodiversity: An outline of higher-level classification and survey of taxonomic richness. *Zootaxa*, *3148*(1), 212–221. doi: 10.11646/zootaxa.3148.1.41

Wahlberg, N., Leneveu, J., Kodandaramaiah, U., Peña, C., Nylin, S., Freitas, A. V. L., & Brower, A. V. Z. (2009, December). Nymphalid butterflies diversify following near demise at the Cretaceous/Tertiary boundary. *Proceedings of the Royal Society B: Biological Sciences*, *276*(1677), 4295–4302. doi: 10.1098/rspb.2009.1303

Wahlberg, N., Snäll, N., Viidalepp, J., Ruohomäki, K., & Tammaru, T. (2010, June). The evolution of female flightlessness among Ennominae of the Holarctic forest zone (Lepidoptera, Geometridae). *Molecular Phylogenetics and Evolution*, *55*(3), 929–938. doi: 10.1016/j.ympev.2010.01.025

Wahlberg, N., & Wheat, C. W. (2008, April). Genomic Outposts Serve the Phylogenomic Pioneers: Designing Novel Nuclear Markers for Genomic DNA Extractions of Lepidoptera. *Systematic Biology*, *57*(2), 231–242. doi: 10.1080/10635150802033006

Wahlberg, N., Wheat, C. W., & Peña, C. (2013, November). Timing and Patterns in the Taxonomic Diversification of Lepidoptera (Butterflies and Moths). *PLOS ONE*, *8*(11), e80875. doi: 10.1371/journal.pone.0080875

Wang, Q.-Y., & Li, H.-H. (2020). Phylogeny of the superfamily Gelechioidea (Lepidoptera: Obtectomera), with an exploratory application on geometric morphometrics. *Zoologica Scripta*, *49*(3), 307–328. doi: 10.1111/zsc.12407

Warrant, E., Frost, B., Green, K., Mouritsen, H., Dreyer, D., Adden, A., … Heinze, S. (2016, April). The Australian Bogong Moth Agrotis infusa: A Long-Distance Nocturnal Navigator. *Frontiers in Behavioral Neuroscience*, *10*. doi: 10.3389/fnbeh.2016.00077

Weller, S. J., Friedlander, T. P., Martin, J. A., & Pashley, D. P. (1992, December). Phylogenetic studies of ribosomal RNA variation in higher moths and butterflies (Lepidoptera: Ditrysia). *Molecular Phylogenetics and Evolution*, *1*(4), 312–337. doi: 10.1016/1055-7903(92)90007-4

Wiegmann, B. M., Mitter, C., Regier, J. C., Friedlander, T. P., Wagner, D. M., & Nielsen, E. S. (2000, May). Nuclear Genes Resolve Mesozoic-Aged Divergences in the Insect Order Lepidoptera. *Molecular Phylogenetics and Evolution*, *15*(2), 242–259. doi: 10.1006/mpev.1999.0746

Wright, C. J., Wahlberg, N., Vila, R., Mutanen, M., Matos-Maraví, P., Lucek, K., … Meier, J. I. (2025, December). Project Psyche: Reference genomes for all Lepidoptera in Europe. *Trends in Ecology & Evolution*, *40*(12), 1234–1250. doi: 10.1016/j.tree.2025.10.007

Yamamoto, S., & Sota, T. (2007, August). Phylogeny of the Geometridae and the evolution of winter moths inferred from a simultaneous analysis of mitochondrial and nuclear genes. *Molecular Phylogenetics and Evolution*, *44*(2), 711–723. doi: 10.1016/j.ympev.2006.12.027

Yoshida, K., & Kitano, J. (2021, April). Tempo and mode in karyotype evolution revealed by a probabilistic model incorporating both chromosome number and morphology. *PLOS Genetics*, *17*(4), e1009502. doi: 10.1371/journal.pgen.1009502

Young, C. J. (2006, July). Molecular relationships of the Australian Ennominae (Lepidoptera: Geometridae) and implications for the phylogeny of the Geometridae from molecular and morphological data. *Zootaxa*, *1264*(1), 1–147. doi: 10.11646/zootaxa.1264.1.1

Zahiri, R., Holloway, J. D., Kitching, I. J., Lafontaine, J. D., Mutanen, M., & Wahlberg, N. (2012). Molecular phylogenetics of Erebidae (Lepidoptera, Noctuoidea). *Systematic Entomology*, *37*(1), 102–124. doi: 10.1111/j.1365-3113.2011.00607.x

Zahiri, R., Kitching, I. J., Lafontaine, J. D., Mutanen, M., Kaila, L., Holloway, J. D., & Wahlberg, N. (2011). A new molecular phylogeny offers hope for a stable family level classification of the Noctuoidea (Lepidoptera). *Zoologica Scripta*, *40*(2), 158–173. doi: 10.1111/j.1463-6409.2010.00459.x

Zhang, C., & Mirarab, S. (2022, December). Weighting by Gene Tree Uncertainty Improves Accuracy of Quartet-based Species Trees. *Molecular Biology and Evolution*, *39*(12), msac215. doi: 10.1093/molbev/msac215

Zhang, C., Rabiee, M., Sayyari, E., & Mirarab, S. (2018, May). ASTRAL-III: Polynomial time species tree reconstruction from partially resolved gene trees. *BMC Bioinformatics*, *19*(6), 153. doi: 10.1186/s12859-018-2129-y

Zhou, Z. (2004, October). The origin and early evolution of birds: Discoveries, disputes, and perspectives from fossil evidence. *Naturwissenschaften*, *91*(10), 455–471. doi: 10.1007/s00114-004-0570-4

List of papers

I **Molecular phylogeny of north European Geometridae (Lepidoptera: Geometroidea)** E. Õunap, V. Nedumpally, E. Yapar, A. R. Lemmon, T. Tammaru. *Systematic Entomology* (2025), 50(1), 32–67

II **Elaborating the phylogeny of Noctuidae by focusing on relationships between north European taxa** V. Nedumpally, A. Zilli, E. Yapar, T. Tammaru, A. R. Lemmon, E. Õunap. *Systematic Entomology* (2025), e70010

III **Integrating Sanger and next generation sequencing data sheds light on phylogenetic relationships among gelechioid moths (Lepidoptera: Gelechioidea)** E. Yapar, A. Chiocchio, M. A. Heikkilä, J. Rota, L. Kaila, N. Wahlberg. *Systematic Entomology* (2025), e70009

IV **Phylogenomics resolves subfamily-level relationships among Erebid moths (Lepidoptera: Noctuoidea: Erebidae)** E. Yapar, H. R. Ghanavi, R. Zahiri, N. J. Dowdy, N. Wahlberg. Manuscript.

V **Tempo and mode of winter diapause evolution in butterflies** S. Halali, E. Yapar, C. W. Wheat, N. Wahlberg, K. Gotthard, N. Chazot, S. Nylin, P. Lehmann. *Evolution Letters* (2025), 9(1), 125–136

Printed by Media-Tryck, Lund 2026 NORDIC SWAN ECOLABEL 3041 0903

LUND UNIVERSITY

Faculty of Science
Department of Biology

ISBN 978-91-8104-796-7