The age of museomics

How to get genomic information from museum specimens of Lepidoptera

Call, Elsa

2020

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

The age of museomics

How to get genomic information from museum specimens of Lepidoptera

ELSA CALL
DEPARTMENT OF BIOLOGY | FACULTY OF SCIENCE | LUND UNIVERSITY

The age of museomics

How to get genomic information from museum specimens of Lepidoptera

Elsa Call

LUND
UNIVERSITY

DOCTORAL DISSERTATION

by due permission of the Faculty Biology, Lund University, Sweden.
To be defended at Stora hörsalen, Pufendorf Institute, Biskopsgatan 3, Lund.
On the 29th of May 2020 at 9:00.

Faculty opponent
Dr. Sabrina Simon
Biosystematics Group, Wageningen University and Research, The Netherlands

Organization LUND UNIVERSITY Department of Biodiversity Systematic Biology Group Ecology Building SE-22362, Lund. Author: Elsa Call	**Document name:** Doctoral dissertation **Date of issue:** May 2020 **Sponsoring organization:** This project has received funding from the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No. 6422141

Title and subtitle: The age of museomics: how to get genomic information from museum samples of Lepidoptera

Abstract

In the age of museomics, the ability to sequence the genetic material from old museum specimens provides an invaluable and often untapped molecular resource. The application of the latest Next-Generation sequencing (NGS) technologies to such specimens allows us to utilise the diverse biobank that is natural history museums. These approaches provide the opportunity to study both extinct or difficult to collect taxa. The aim of this thesis is to apply NGS techniques to museum specimens of Lepidoptera, to better investigate the types of data generated and their uses.

In the first chapter, we use a targeted enrichment (TE) approach to sequence nuclear loci from museum specimens dating back to 1892 for 35 taxa across the order Lepidoptera. Loci recovery ranged from 500-1,747. The success of this technique across the specimens highlights the usefulness of such a kit to study the entire order, thereby enabling the potential to resolve both shallow and deeper nodes in the phylogeny.

In the second chapter, we applied the TE approach to three moth families belonging to the superfamily Geometroidea. Thirty-three museum specimens collected between 2001 and 1892 were sequenced from the families Epicopeiidae, Sematuridae and Pseudobistonidae. We recovered up to 1,383 raw loci per individual. Loci recovered in 20 or more specimens were carried forward for phylogenetic analysis, with a final data set consisting of 378 loci. These loci from another 19 publically available genomes and transcriptomes were combined to complete our dataset. We find strong support for the hypothesis that Sematuridae is the sister group to Epicopeiidae + Pseudobistonidae.

Further expanding on our results from the TE approach, in the third chapter we apply whole genome sequencing (WGS) to expand our dataset. We sequenced whole genomes of 30 museum specimens of Epicopeiidae and Sematuridae. Recovery of the 387 loci from the TE experiment ranged from 20 - 94%. The resulting phylogeny confirms the phylogenetic relationships within these families. Additionally, we compared the data generated between the two approaches, presenting the advantages and disadvantages of each approach.

In the final chapter, we investigated the usefulness of museum specimen WGS for population genomics studies. Data was generated for 13 specimens of Pieris napi, a common butterfly in Sweden. The availability of a recently published genome allows us to demonstrate that ~81% of the recovered DNA is from the target specimen. Average genomic coverage was 15.6X for nuclear DNA, and 1,963X for the mitochondria. We found that individuals originating from Abisko are genetically distinct from the remaining P. napi populations. This study highlights the potential of museum specimens for looking at changes in population genetic dynamics through time.

In summary, we show the usefulness of various museomics applications. In particular, we focus on the use of TE and WGS for phylogenomic studies. Additionally, we highlight that WGS sequencing can also be utilized for population genetic-based studies, opening a window to the past.

Key words: museomics, museum samples, Lepidoptera, WGS, TE, phylogenomics

Classification system and/or index terms (if any)

Supplementary bibliographical information		**Language:** English
ISSN and key title		**ISBN** 978-91-7895-506-0 (print) 978-91-7895-507-7 (pdf)
Recipient's notes	**Number of pages** 79	Price
	Security classification	

The age of museomics

How to get genomic information from museum specimens of Lepidoptera

Elsa Call

LUND
UNIVERSITY

Coverphoto by Léanie Alloing-Séguier

Faculty of Sciences
Department of Biology

ISBN
978-91-7895-506-0 (print)
978-91-7895-507-7 (pdf)

Printed in Sweden by Media-Tryck, Lund University
Lund 2020

To strive, to seek, to find, and not to yield

Alfred, Lord Tennyson – Ulysses

Table of Contents

List of papers

This thesis is based on the following papers:

I. Mayer C., Dietz, L. Call E., Kukowka S., Martin S., Espeland M. (2020). Adding to the Lepidoptera phylogeny: Capturing hundreds of nuclear genes from old museums specimens. Systematic Entomology, submitted.

II. Call E., Mayer C. Twort V., Wahlberg N., Espeland M. (2020). Museomics: phylogenomics of the moth family Epicopeiidae (Lepidoptera) using target enrichment. Insect Systematics and Diversity, submitted.

III. Call. E, Twort V., Espeland M., Wahlberg N. (2020). One Method to Sequence Them All? Comparison between Whole-Genome Sequencing (WGS) and Target Enrichment (TE) of museum specimens from the moth families Epicopeiidae and Sematuridae (Lepidoptera). Manuscript.

IV. Call E., Twort V., Wheat C. W., Wahlberg N. (2020). Rear window: population genetics using museums specimens of Pieris napi (Lepidoptera). Manuscript.

List of contributions

Contributions by Elsa Call:

I. Limited contribution to the study design. Provided a significant contribution to data collection. Limited contribution to data analyses. Limited contribution to manuscript preparation.

II. Provided a significant contribution to the study design. Performed majority of data collection. Performed majority of data analyses. Took the lead in manuscript preparation.

III. Limited contribution in study conception. Took the lead in study design. Performed majority of the data collection. Performed majority of the data analyses. Took the lead in manuscript preparation.

IV. Limited contribution in study conception. Provided a significant contribution to study design. Performed majority of data collection. Performed majority of data analyses. Took the lead in manuscript preparation.

1 Introduction

1.1 Natural history museum

Natural history museums are one of the cornerstones of western scientific culture. During the Renaissance, connoisseurs started to collect and accumulate numerous natural samples, such as animals, plants, minerals, and even some pieces of art or relics from past and different civilizations. Because of the appeal for the 'exotic', these collections often included non-European specimens and sometimes even fake ones. From what they could bring back from their exploratory journeys, to their backyards they began to collect, store, and sort out a variety of items.

At first, they stored all of these collections in cabinets of curiosities (or kunstkammer in German). Nevertheless, those wonder-rooms were private and often lacked methodology and order. In 1587 Gabriel Kaltemarckt provided guidelines for setting up an organized cabinet of curiosities, but also suggestions about the kind of qualifications required for the person in charge of these collections. According to him, a proper Kunstkammer was supposed to contain three different types of depots: sculptures, paintings, and unique items "from home and abroad." In this last category, he included "antlers, horns, claws, feathers and other things belonging to strange and curious animals, birds or fishes, including the skeletons of their anatomy" (Gutfleisch & Menzhausen, 1989). However, despite this first attempt to organize and standardize collections, Kaltemarckt also provided terrible advice by dissuading the labelling of items with collection information (Gutfleisch & Menzhausen, 1989).

Soon, collectors, merchants, but also counterfeiters and other charlatans, gave way to scholars, scientists, and naturalists who were genuinely interested in observing, studying, and indexing the living diversity on Earth. People like Ole Worm, a Danish physician and natural historian (1588–1654) and Athanasius Kircher (1602–1680), a German Jesuit scholar and mathematician, contributed with a lot of knowledge and items in these cabinets of curiosities. Gradually, these wonder-rooms became both more structured and managed, as well as more collectivized by institutions like universities. One of the first museums, as we recognize them now, was probably the Ashmolean Museum of Art and Archaeology established in 1683 from Elias Ashmole's cabinet of

curiosities offered to Oxford University (Swann, 2001). Subsequently, a lot of European natural history museums bloomed. In 1735, the Biological Museum at Lund University was initiated based on the collection of Kilian Stobaeus. Afterwards, in 1753, the British Museum was constituted predominantly based on Sir Hans Sloane's collection (Classen, 2007). The French National Museum of Natural History followed, founded during the French Revolution in 1793, and scores of other museums around Europe. With these newly formed collections and descriptions of species, museums developed and started to standardize their rules of archival. In parallel with this, new scientific disciplines emerged: systematics and taxonomy, leading to biology, as we know it nowadays.

Over the decades museums gathered a vast number of specimens either by the acquisition of natural samples or by donations from amateurs or philanthropists, leading to the immense museum collections we see around the world today. It is not straightforward to estimate the number of specimens in museums worldwide; however, the current consensus is around 2.5–3 billion specimens (Chapman, 2005; Duckworth et al., 1993; Suarez & Tsutsui, 2004).

Since their creation, natural history museums have represented a valuable reference of biological resources and knowledge for both the scientific community and the public. These arrays of natural specimens are essential for the study of systematics, biodiversity, habitat loss, and global climate change research, or biological invasion studies, as well as number of other scientific disciplines (Bradley et al., 2014).

1.2 Museomics

Until recently, scientists were using the immense amount of biological resources stored in museums only at a morphological study level. Due to the high level of damage and fragmentation of the genetic material in historical samples, researchers considered the DNA from these specimens to be too degraded to be utilized (Shapiro & Hofreiter, 2012). Consequently, DNA extractions and analysis were initially limited to freshly collected samples.

Nowadays, the scientific community in biology can face many challenges before being able to sample in the field. These issues can be monetary (*e.g.*, lack of funding), stochastic events (inaccessibility of species of interest, adverse weather conditions, *etc.*), but also administrative difficulties, in terms of the need for permits.

In an objective to, somehow, standardize this burdensome bureaucracy, the Convention on Biology Diversity (CBD) set the Nagoya Protocol (NP) on Access and

Benefit Sharing (ABS). The NP is an agreement that aims to provide a transparent legal framework for the fair and equitable sharing of benefits arising out of the utilization of genetic resources (Buck & Hamilton, 2011; "The Nagoya Protocol on access and benefit sharing of genetic resources. 1," 2010).

In 2010 the CBD met in Nagoya (Japan) to discuss how to conserve biological diversity in the fairest way possible. The purpose of the treaty they agreed on was to allow more equitable sharing of natural resources (CBD, 2012). In other words, the use of any biological material first, requires an agreement between providers, users, and countries, mainly if this trade involves the utilization of genetic resources, which concerns all biological organisms. Despite these noble intents, the adoption of the NP, however, has not been smooth. Different stakeholders lack clarification, while others doubt how the implementation of such a protocol is going to work (Neumann et al., 2018). While the intention behind the NP was to fight unjust profitable exploitation of biological resources (a.k.a biopiracy), this treaty, at the same time, is failing at its primary goal, which is to fairly compensate provider communities (Cressey, 2014; Schindel et al., 2015).

The main concerns of the scientific community were the limitations and the burden that the NP brings to studies on biodiversity (such as conservation, monitoring, treatment of infectious diseases), as well as other areas of research (like phylogeny or biogeography). However, despite these concerns, 51 countries ratified the protocol on July 14th, 2014, and the NP came into force on October 12th, 2014. Accordingly, since this date, researchers must, before sampling specimens, arrange agreements, referred to as Access Benefit Sharing (ABS) agreements, with the providing country. These preliminary permissions need to specify not only who and how will profits from the used organisms but also a fair shared definition of the study's benefits between the involved stakeholders. Both publications' co-authorship and sharing profits from products (such as vaccines or pesticides) resulting from the study of the organism are considered benefits.

Although the idea behind this is a proper and fairer distribution of resources, in practice for scientific research, it can be a deadlock situation. As an illustration, consider a phylogenetic study: in terms of evolutionary questions, *the more taxa, the merrier*. As borders do not concern species, moreover some taxa are distributed worldwide, a research team would quickly need hundreds of co-authors / agreements to be able to use all the material they require for their investigation. Often getting deals requires negotiations in person in the countries of origin, which is financially impossible for many research projects.

Under those conditions, natural history museum specimens represent a crucial resource for further expanding our understanding of the diversity we see around us. These collections could be the cornucopia for biological studies by efficiently providing millions of specimens. As previously discussed, the sampling of fresh material can be difficult and hindered by several factors. Not to mention that museums' collections can give access to not only rare but also extinct specimens, which represents a fantastic asset.

Currently, we are experiencing an unprecedented loss of biological diversity (Butchart et al., 2010). For some taxa, the rate of extinction is massive. For example, over the last 25 years, more than 75% of the total flying insect biomass has declined in protected areas in Germany (Hallmann et al., 2017). Because of this contemporary mass extinction period, some species are under protection or, in some unfortunate cases, have already disappeared. In this worrisome situation, museums collections can act as a genetic Noah's Ark, maintaining biological resources that have otherwise become rare or extinct in the wild.

One of the excellent illustrations of the use of museomics on an extinct species is the little bush moa (*Anomalopteryx didiformis*). From one mitochondrial 12S rRNA gene (A. Cooper et al., 1992), to the whole genome (Cloutier et al., 2018, 2019), DNA from museum samples of moas helps to understand the phylogenetic relationships of the New Zealand endemic flightless birds. The results of these studies strongly suggested that flightlessness evolved independently in two lineages.

1.3 Sequencing ancient and historical DNA (aDNA & hDNA)

1.3.1 Raiders of the Lost DNA

In 1984, Higuchi et al. successfully extracted DNA from a sample of quagga (*Equus quagga quagga*), a zebra subspecies that went extinct one hundred years earlier (1883). This study was the first fruitful attempt to obtain DNA from a historical specimen. However, due to technical limitations, they only managed to recover 1% of what they could get from fresh samples (Higuchi et al., 1984). One year later, another experiment focused on a 2,400 year-old Egyptian mummy (Svante Pääbo, 1985). These two studies used bacterial cloning to amplify short sequences. Nevertheless, due to the rarity of the samples and the lack of experimental reproducibility, the authenticity of these studies was questionable, and particularly the paper on the Egyptian mummy was highly controversial (Knapp et al., 2015). The probable contamination with fresh human

DNA, quite rightly pointed out, was the primary concern against the latter case. Therefore, impending ancient DNA (aDNA) studies need to be extra cautious and reduce as much as possible all the potential contamination risks associated with fragmented DNA. Despite these initial suspicions, Higuchi's research was the foundation for the ancient DNA and museomics field by showing the preservation of some endogenous DNA in those old specimens.

1.3.2 Close Encounter of the PCR Kind

The development of the polymerase chain reaction (PCR) (Saiki et al., 1985) and its popularization paved the way for ancient DNA research. Because it targets a specific segment of DNA and amplifies millions of copies of these targeted DNA fragments, PCR was a real revolution in molecular biology. With this method and the constant enhancements brought to it through time, ancient DNA and historical specimens' research could rise from dust. Indeed, several old DNA studies flourished following this discovery (S. Pääbo, 1989; Svante Pääbo & Wilson, 1988; R. H. Thomas et al., 1989).

Nevertheless, the amplification power of PCR can also lead to a lot of wrong results. Even though PCR allows bypassing the natural fragmentation of the DNA, by amplifying short fragments (<100–300 bp), the downside of this method is its tendency to target first contaminant copies of DNA. Admittedly, if there is fresh DNA contamination within samples, there are far more copies of the exogenous DNA than the endogenous DNA. Because of the abundance of contaminants, during the PCR process, these will have more chances to get amplified. Under those circumstances, false positives were – and still are – one of the principal challenges in ancient DNA research. During the early 1990s, some ambitious yet incautious studies arose, like presumed dinosaur DNA sequences (Woodward et al., 1994), which were modern human contaminations (Zischler et al., 1995). Another study with overstated conclusions was claiming to retrieve DNA sequences from millions of years (Myr) old *Magnolia* species (Golenberg et al., 1990). However, independent laboratories were not able to replicate this study (Austin et al., 1997; Hedges et al., 1995). Other studies proposed this kind of incredible results while being vague in both their methodology and lacking any replication (Cano & Borucki, 1995; DeSalle et al., 1992; Poinar et al., 1993). Because of these unsure first steps, for a long time, ancient DNA research had bad press. Therefore, the scientific community was highly skeptical regarding this kind of research.

Despite these early failures and fumbling beginnings, in the early 2000s, through advances in the methods and the understanding of the types of DNA damage, as well as how to bypass sample contaminations, ancient DNA research became a trustworthy

science (A. Cooper & Poinar, 2000; Eske Willerslev & Cooper, 2005). Increasingly studies were authenticated by independent replications, like for the brown bear (Loreille et al., 2001), the moa mitochondrial genome (Cooper et al., 2001), DNA of plants from permafrost sediments (Willerslev, 2003) and permafrost bacteria (Willerslev et al., 2004). Even so, PCR made a big difference for ancient DNA research; its inherent and specific challenges remain. PCR brings a solution against the natural fragmentation of the DNA by allowing amplification of short fragments. However, this method needs to be supplemented by overlapping, as well as specially designed primer pairs to ensure that the different PCR products originate from the same DNA source (Römpler et al., 2006). This technique alone cannot fix all kinds of damage. Indeed, chemical modification of ancient DNA can lead to PCR failures. And, even in case of success, only a small amount of endogenous DNA can be retrieved (Rowe et al., 2011).

1.3.3 Next-Generation Sequencing: A New Hope

All of these studies from the early 2000s were based on the Sanger capillary sequencing method. But, in 2005, a new significant shift occurred in the molecular field. The appearance of new sequencing technologies ushered in a new age for genomics (Margulies et al., 2005). The development of highly parallelizable and high-throughput sequencing (HTS), known as next-generation sequencing (NGS), allows the production of larger volumes of sequencing data per run. Current NGS methods generate several hundred million independent reads per run, while, in comparison, Sanger sequencing makes just a single read (Metzker, 2010; Shendure & Ji, 2008). Thereby allowing the shift from looking at genes one by one to being able to sequence and quantify all expressed genes or an organism's entire genome.

Equally important was the fact that new sequencing platforms were far more efficient, concurring with the rise of a lot of new brands. This emergence of new laboratory companies created a new market with fierce competition and drastically decreased prices. Several NGS platforms rose at that time, but the three most popular ones were Roche/454 Life Sciences pyrosequencing, Illumina, and ABI SOLiD ligation sequencing (Zhou et al., 2010). Sequencing methods are continuously improving with the development of technologies like Life Technologies Ion Torrent, Pacific Biosciences (PacBio), and Helicos Biosciences Heliscope (Egan et al., 2012; Metzker, 2010). Meaningful differences between all these platforms are the read length they produce; for platforms such as Roche/454 and Illumina, the average read length obtained is between 100 and 350 bp while technologies like PacBio can provide read lengths between 970 and 15,000 bp (Metzker, 2010; Rhoads & Au, 2015).

Initially, this sequencing enthusiasm started with the main focus on model organisms, in other words, already available genetic resources, specifically the human genome. The Human Genome Project, which was launched in 2001, notably pushed this innovations race (Mardis, 2011). Over time, the continued improvement in the technologies and methods, allowed a shift to non-model organisms (Mardis, 2008; Zhou et al., 2010).

These new methods opened the door to new possibilities. With those new sequencing platforms, it became possible to sequence short DNA fragments (<100–300 bp), and more and more ancient DNA studies arose through these new applications. The combination of sequencing technology advancement and its application to ancient DNA, has shined light on the potential of natural history museum collections for genetic use (Särkinen et al., 2012; Seguin-Orlando et al., 2013; Staats et al., 2013). These highlight the importance of museum collections, not only for the collection of rare, exotic and extinct species, but also for the genetic information contained within the specimens. Therefore, NGS technologies offer a window into the past. In 2013 the word "museomics" –meaning the use of NGS for museum specimens – started to appear in the literature, with a study on the most extensive primate radiations (Guschanski et al., 2013). At the beginning of 2020, Web of Sciences tallies 19 articles with this keyword. These studies concerning a wide diversity of species: from birds (Anmarkrud & Lifjeld, 2017; Cloutier et al., 2018, 2019) and mammals (Fabre et al., 2014; Hawkins et al., 2016), to insects (Kanda et al., 2015; Sproul & Maddison, 2017; Zhang, Shen, et al., 2019), through plants (Kadlec et al., 2017; Malakasi et al., 2019; Silva et al., 2017; Zedane et al., 2016); more studies using historical specimens DNA are also on the way. We are at the dawn of a new exciting era, and these examples are just the beginning of what is possible with museum specimens.

1.4 Why is sequencing old genomes difficult?

A museum's genomic treasure chest is not that easy to unlock, as it comes with its own specific set of difficulties. Sequencing DNA from historical specimens is troublesome. The first and foremost issue with ancient DNA is its degradation. Indeed, DNA strand breaks occur naturally in apoptotic cells (Collins et al., 1997). When cells are living, enzymatic repair processes maintain DNA integrity (Lindahl, 1993). However, when an organism dies, the enzymatic processes that were repairing DNA are no longer active, leading to an accumulation of damage. Time and preservation conditions can accelerate or slow down this degradation phenomenon. Eventually, this process leads to a tiny amount of endogenous DNA copies in the organism (Pääbo et al., 2004). Due to these

reasons, for a long time, DNA from museum specimens was seen as too degraded to be used, and therefore, scientists used museum collections primarily for morphological studies.

Regardless of how useful the NGS methods are, ancient DNA studies have to deal with specific challenges and require cautious preparation. The most threatening problem is contamination from fresh material. However, ancient DNA is also prone to particular types of damage.

1.4.1 Types of ancient DNA damage

DNA is commonly damaged during the life of an organism, although enzymes are constantly repairing this damage. But, after death, this deterioration process continues while the repairing processes are no longer functional, leading to strand breaks within the DNA. For this reason, a low amount of integral DNA tends to persist in ancient specimens. Cold, dry, and stable environments can help to regulate or restrain nuclease activities by reducing, sometimes considerably, the damage occurring after an organism's death. On the contrary, warm, humid, and variable environmental conditions will accelerate the process. Under those circumstances, DNA from old specimens contains low quantities of DNA, which have short fragment lengths (Fulton, 2012).

The second type of damage specific to ancient DNA is miscoding. The chemical mechanism causing this damage is the hydrolysis of bases, resulting in base changes. Two types of miscoding exist. The first is the deamination of cytosine to thymine or uracil (Figure 1). Similarly, the same process transforms guanine to adenine whereas the second damage type results in adenine to guanine substitutions. A warm environment (> 37 °C) can accelerate this change (Lindahl, 1993). A simple solution to deal with this is to use enzymes that remove uracil, called uracil DNA glycosylases (UDG). One should note that these enzymes generate a single nucleotide gap at the location of uracil, and this can lead to a strand break during PCR due to heat. In that case, UDG are not enough and can even be dangerous as it may lead to more fractures in the DNA. Thus, another solution against base change is to do, at least, two independent amplifications and compare the resulting products. If there is a difference between these two amplifications, it is necessary to perform at least one more amplification. These extra steps will help to ascertain which of the two sequences is reproducible (Hofreiter et al., 2001; Höss et al., 1994). Nevertheless, nowadays, UDG are efficient enough, and multiple amplification might be unnecessary (Fulton, 2012).

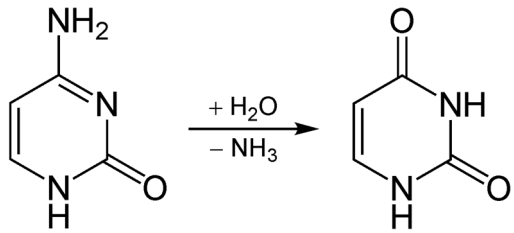

Figure 1. Deamination process. Deamination of cytosine (left) to uracil (right) as caused by hydrolysis.

Blocking and other kinds of miscoding are the third type of damage encountered in ancient DNA. This time the miscoding is also due to a chemical modification of the DNA base. This change leads to a base misincorporation of guanine to 8-oxoguanosine, which pairs with adenine instead of cytosine like it should. Under those circumstances, guanine will be misinterpreted as thymine (Figure 2) (Hofreiter et al., 2001; Le Bihan et al., 2011). These lead to base modifications, which can cause the amplification of chimeric sequences ('jumping PCR') or no amplification at all (Hofreiter et al., 2001). The solutions against these processes are to use specialized polymerases and multiple amplifications.

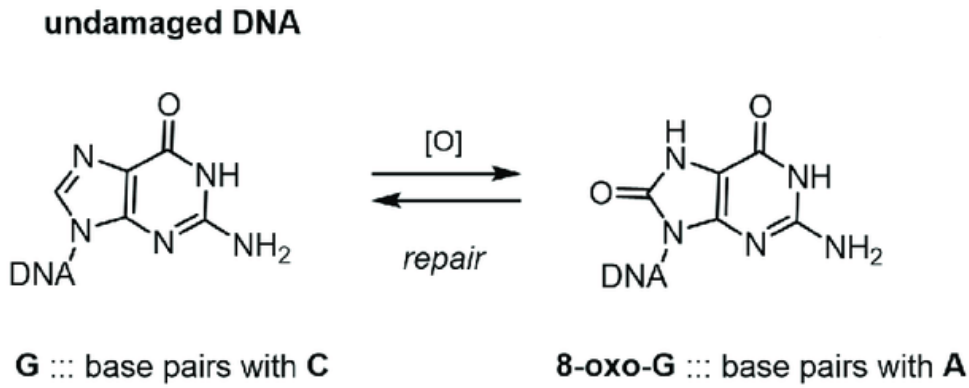

Figure 2. Guanine misincorporation. When the DNA is undamaged, guanine (G) pairs with cytosine (C). However, oxidation of G leads to the formation of 8-oxo-guanosine (8-oxo-G) that will pair with the adenine (A) and thus be read as thymine (T). The repair process consists in a reduction of the excessive oxygen.

Finally, alkylation can lead to a covalent linkage between two nucleotides that should typically not bond together, also known as crosslinking. It can occur within the same DNA strand (intrastrand) or between opposite strands (interstrand; Figure 3). The effect of these crosslinks is an absence of amplification (Hofreiter et al., 2001). The solution consists of using PTB (*N*-phenacy-lthiazolium bromide) to cleave the crosslinks. However, in a 2007 study, PTB was shown to be ineffective (Rohland &

Hofreiter, 2007). In the same survey, proteinase K alone seemed efficient enough to cleave crosslinks.

As ancient specimens endure severe fluctuations of the environment, it is difficult, if not impossible, to know all the chemical and physical reactions that have occurred to the DNA. Therefore, one has to accept that working on ancient DNA means dealing with ambiguities and unpredictability. The three previously mentioned types of damages (blocking, miscoding, and crosslinks) are more frequent in ancient DNA. For example, they have been observed for Pleistocene coprolites (Poinar et al., 1998), as well as bones from archaeological excavations (Lindahl, 1993). However, these damages can occasionally occur in museum specimens (Hofreiter et al., 2001; Lindahl, 1993). Knowing how specimens have been preserved is difficult, if not impossible. Therefore, it is crucial when doing DNA extractions from museum specimens to consider all the ways their DNA could have been altered and not discount any type of damage.

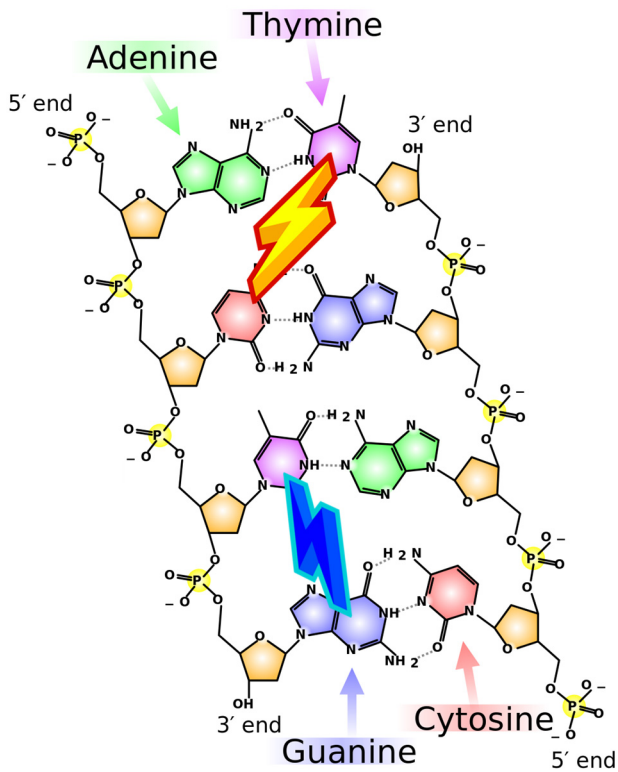

Figure 3 Schematic illustration of crosslinks. At the top, the yellow flash represents interstrand crosslinks: thymine is linking with cytosine of the opposite strand. At the bottom, the blue flash indicates intrastrand crosslink: the thymine is bonding with guanine within the same strand.

1.4.2 Alien contaminations

Our understanding of the molecular mechanisms occurring within museum specimens is continuously improving, thereby making the application of new technologies possible and more efficient. This new age of '–omics' provides us with the tools to repair ancient DNA and improves our ability to sequence genetic material from unusual specimens.

While museomics is a fantastic tool, as previously mentioned, one still needs to take care and use caution when applying these technologies. The principal obstacle is that of obtaining reliable and faithful data. In other words, are we sure there is no contamination with external DNA in our samples? Being able to trust what we are sequencing is a crucial point for both ancient DNA and museomics studies.

One of the best ways to avoid contamination is to perform the wet-lab methods in a specific lab, where no fresh genetic material has been processed, especially if fresh samples are the same species or order as your study organism. Then at each step of the sample preparation process, you need to sterilize as much as you can the tools you require to avoid cross-contamination. Another critical point is, of course, the physical isolation of the pre-PCR ancient DNA or museomics facility; that rigorously maintains the "one-way" rule of movement between pre-PCR, PCR, and post-PCR rooms. For museomics and ancient DNA studies, researchers must ensure the repeatability of their experiment and do both control extractions and PCR controls. These recommendations are the basis for proper ancient DNA and museomics research, for more guidelines on setting up an appropriate lab for these kinds of studies, see Cooper and Poinar (2000) and Fulton (2012).

1.5 How to sequence genomes?

The veteran capillary sequencing method (Sanger sequencing) is still useful nowadays, specifically for low volumes and relatively short fragments. Nevertheless, for a large number of molecules and quick turnover time, this approach can become very expensive and challenging. By parallelizing the sequencing process, and producing millions of reads all at once, High-Throughput Sequencing (HTS) technologies reduce the cost of DNA sequencing per sequenced nucleotide. HTS includes next-generation "short-reads" (NGS) and third-generation "long-reads" sequencing methods. Furthermore, HTS can generate single-nucleotide polymorphisms (SNP), which are prevalent in population genetics. In any case, when aiming for genome sequencing, HTS are the way to go. Now, different sequencing methods are available. Despite differences in the methodology, these approaches follow similar steps.

1.5.1 Transcriptomics

Each cell or populations of cells possess a set of RNA molecules, also called a transcript. These polymeric molecules are essential for gene expression. We refer to transcriptome, the complete set of transcripts in a cell. Contrary to the genome, which represents the entire genetic material and is generally constant in a given cell or population of cells, the transcriptome can vary. Transcriptome includes all transcripts as well as messenger RNA (mRNA) and thus indicates genes that are expressed at a given time (Wang et al., 2009). By analyzing transcriptomes, researchers study the level of expression of RNAs in a specific cell population. Transcriptomics also helps to understand the functional elements of the genome (Wang et al., 2009).

Unfortunately, as RNA is the core of transcriptomics, without it, this approach is not useful, particularly in the case of museomics. As previously mentioned, the DNA in both museum specimens and ancient samples is slowly degrading, making museomics already challenging by itself. But RNA degradation occurs even faster (Bushell et al., 2004; M. P. Thomas et al., 2015), making conservation of RNA elements in such specimens improbable.

1.5.2 Genome-reduction strategies

NGS allows sequencing millions of DNA molecules at the same time, generating large datasets. However, this amount of data might be overwhelming by being both time-consuming and costly. Sometimes, the less, the merrier. Some studies prefer to focus only on some parts of the genome that are relevant to them. In these cases, genome-reduction strategies are favoured. These methods still parallelize the sequencing process, and thus, produce millions of sequence short-reads; nevertheless, they concentrate on subsets of the genome. With this approach, they reduce both the volume and complexity of data to be analyzed (Andolfatto et al., 2011; Baird et al., 2008; Elshire et al., 2011; Huang et al., 2010).

While HTS are potent tools for the rapid collection of genome-wide markers, the cost and time-consumption processes for analysing these data make it still difficult (De Donato et al., 2013). In order to simplify this work, new technologies were developed to obtain subsets of genomic restrictions fragments for NGS.

GBS & RAD-sequencing

One of the most popular genome reduction methods includes restriction-site-associated DNA (RAD sequencing) (Baird et al., 2008), and, more recently, Genotyping-by-Sequencing (GBS) (Elshire et al., 2011). Both of these approaches target short DNA

fragments around specific enzymatic restriction sites (Lewis et al., 2007; Miller et al., 2007). These methods not only reduce genome complexity, but also coupled with barcoding can be a cost-effective approach to genotyping many individuals at once (Ward et al., 2013). This type of data has been used to provide insights into population dynamics, inbreeding and the study of adaptive variation (Leaché et al., 2014; Ogden et al., 2013).

UCEs

Ultraconserved elements (UCEs) are, as their name suggests, highly conserved genomic regions. Because of this, they are well shared among various taxa that are, evolutionary speaking, very distinct from each other. Their first description showed 481 fragments of ~200 bp with a perfect identity (i.e. no insertions nor deletions) between orthologous regions of the human, rat, and mouse genomes (Bejerano et al., 2004). By targeting these UCEs, we are also able to catch flanking DNA, adjacent to these markers. This data allows us to reconstruct phylogenies and understand the evolutionary history of divergent taxa (Faircloth et al., 2012, 2013, 2015; Smith et al., 2014). Because of this, the scientific community considers UCEs to be sort of universal genetic markers and are often used as probes for target capture sequencing approaches.

Target Enrichment (TE)

Target capture sequencing approaches (Target Enrichment, TE, also known as Anchored Hybrid Enrichment) is one of the multiple aspects of genome-partitioning strategies and involves the parallel enrichment of specific preselected genomic regions. One of the highly attractive advantages of this method is that only a few reference genomes are needed to design probes, which can be used for a larger group of species. Therefore, this technique can be used on non-model taxa (Jones & Good, 2016). Several alternatives exist, but here we will mainly focus on the hybrid enrichment in-solution approach. This technique uses more probes and fewer libraries, and it is designed to be better for smaller target sizes (Toussaint et al., 2018; Tsangaras et al., 2014), which is precisely what we are looking for in the case of degraded DNA (Cruz-Dávalos et al., 2017). The critical point is to use specific probes designed to target genomic regions (Lemmon & Lemmon, 2013; Mamanova et al., 2010). In the case of ancient DNA, the tiling probes are based on modern species (Gasc et al., 2016). Several examples can be found in the literature, such as on vertebrates (Lemmon et al., 2012), butterflies (Espeland et al., 2019; Toussaint et al., 2018), as well as in plants (Kadlec et al., 2017). The used protocols are described in more detail in chapters 1 and 2.

1.5.3 Whole-Genome Sequencing (WGS)

As the name says itself, Whole-Genome Sequencing (WGS) consists of determining the full genome of an organism, both coding (genes) and non-coding regions, as well as mitochondrial (and chloroplastic in the plant's case) genomes. The first successful approach to obtain an almost entire human genome was obtained by capillary sequencing. However, because this technique is too expensive and takes too much time, it has been replaced by HTS approaches.

Nowadays, different technologies exist to perform WGS, for instances, the most popular at the moment are Single-molecule real-time sequencing (SMRT) (Pacific Biosciences), ion semiconductor (Ion Torrent sequencing), Pyrosequencing (454, Roche Diagnostics), and sequencing by synthesis (Illumina). SMRT produces the longest reads with an average of 30,000 bp and a maximum read length more than 100,000 bp. In comparison, Illumina technologies provide different read sizes (from 50 bp up to 600 bp, depending on the machines used). For more details of this approach, see chapters 3 and 4.

2 Aim of thesis

Natural museum collections around the world are incredibly numerous. All of these specimens are highly valuable for a wide range of research applications, such as systematics and taxonomy, biodiversity studies, habitat loss and investigations into the history of infectious diseases or environmental contaminants. For a long time, these museum specimens were mainly used to study morphology because their DNA was seen as too degraded to be useful (Fulton, 2012). However, the emergence of the NGS methods allows sequencing of very short fragments of DNA, thereby opening the door to new possibilities, like sequencing of both ancient DNA and museum specimens. Nowadays, we have examples of successful sequencing of extinct taxa including Neanderthal (Pääbo et al., 2004), woolly mammoth (Palkopoulou et al., 2015), cave bear (Noonan et al., 2005), as well as museum specimens with a primary focus on mammals and birds (Cloutier et al., 2018; Fabre et al., 2014), and few examples of plants (Kadlec et al., 2017; Yeates et al., 2016). Studies on insects, however, are limited, as they are often considered too small and too difficult to extract DNA (Kanda et al., 2015; Sproul & Maddison, 2017; Zhang, Shen, et al., 2019).

The main ideas of the project are to develop Next-Generation sequencing methods for Lepidoptera (moths and butterflies) museum specimens and to explore some of the possibilities of museomics and their applications. One aspect is to explore the relationships among families of a superfamily of moths, Geometroidea. Within this superfamily, three families are rare and challenging to collect nowadays. Still, there are extensive collections of representative species for them in museums (such as Bonn, Copenhagen, Paris, Stockholm, Tokyo, etc.). These families, Epicopeiidae, Sematuridae and Pseudobistonidae, are quite small. Twenty-seven species are described within Epicopeiidae, forty-two in Sematuridae and only two species belong to Pseudobistonidae. Sematuridae is mainly a South American family, while the two others are Asian families. Epicopeiidae are widely distributed in the whole of Asia whereas Pseudobistonidae are primarily found in the Himalayas. Furthermore, recent phylogenetic studies have arrived at divergent conclusions regarding the position of these three families in Geometroidea (Bazinet et al., 2013; Kawahara et al., 2019; Kawahara & Breinholt, 2014; Mutanen et al., 2010; Regier et al., 2013; Wei & Yen,

2017). The idea is to try to sample as many as we can of the described species from these families to include them in our phylogenomic study.

Finally, one project will establish whether there are any general trends to be found in the degradation of DNA over time in museum specimens. This part of my work is focused on museum specimens of one species (green-veined whit butterfly or *Pieris napi*, for which there is a reference genome) for next-generation sequencing, and more specifically WGS. Here I aim to see how well the genome is recovered from museum specimens, as well as to find out how many of the recovered reads belong to the butterfly. Ultimately, we will explore the use of highly fragmented genomes for population genomic studies (chapter 4).

3 Methodology

3.1 Studied species

Mysterious taxa with uncertain phylogenetic relationships can be found in several lepidopteran lineages. Primarily based on morphological data, the relationships between and within these clades, have been clarified by the addition of more molecular data. Molecular data are becoming more and more reliable, as well as easier to obtain; it is no surprise that researchers interested in phylogenetic relationships, strenuously include molecular tools in their studies. Therefore, molecular phylogenetic data, combined with morphological studies, are powerful tools to investigate these enigmatic taxa. Here we will mainly focus on three families of moths in the superfamily Geometroidea, which are tropical or sub-tropical, and try to clarify their phylogenetic relationships. These taxa, Pseudobistonidae, Epicopeiidae and Sematuridae, all together, form a clade which is sister to the rest of Geometroidea: Uraniidae and Geometridae. Finally, we will focus on one butterfly species, *Pieris napi* (Pieridae).

3.1.1 Sematuridae

This family contains two subfamilies: the major one is Sematurinae and comprises 41 South American species, in contrast, Apoprogoninae, is a monobasic (*i.e.* one species represents the clade) southern African subfamily (Holloway et al., 2001; Minet & Scoble, 1999). This study is mainly focused on Sematurinae (Figure 4), as I was not able to find a specimen of Apoprogoninae. Among the Sematurinae sequenced for this thesis, I managed to loan eight species of four different genera (*Anurapteryx*, *Coronidia*, *Homidiana* and *Mania*). The biology of this subfamily is not well known.

Sematuridae was our big unknown, given the fact that no general study has been conducted on them. Most of the previous work on this family was exclusively morphological and heavily based on *Mania* (Cock, 2017; Cock & Lamas, 2011). Furthermore, molecular studies including Sematuridae specimens were massively relying on three species: *Mania lunus* previously *Sematura lunus* or *M. luna* (Regier et al., 2009), *Mania diana* formerly identified as *Nothus* (Breinholt et al., 2018; Kawahara

& Breinholt, 2014; Joël Minet & Scoble, 1999), and *Coronidia orithea* (Cho et al., 2011; Heikkilä et al., 2015; Rajaei et al., 2015; Sihvonen et al., 2011; H. Wang et al., 2019).

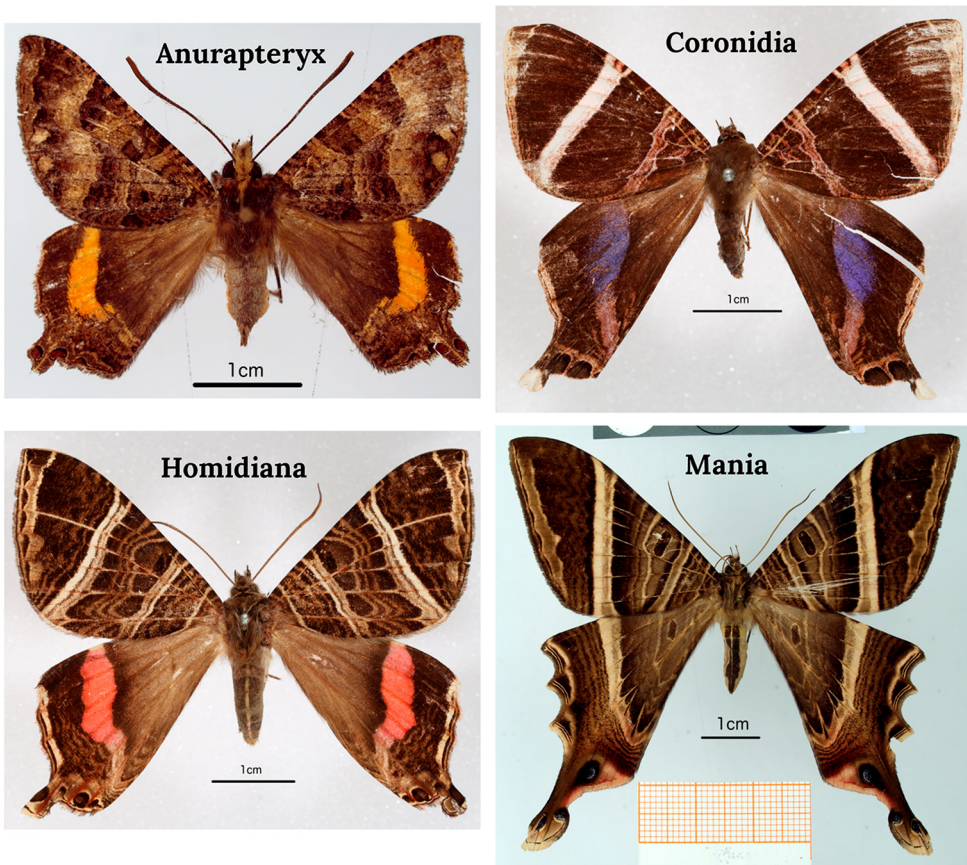

Figure 4. Pictures representing each of the genera of Sematuridae present in this study. The species displayed are *Anurapteryx interlineata*, *Coronidia erecthea*, *Homidiana egina* and *Mania lunus*.

3.1.2 Epicopeiidae

One genus, Epicopeiidae, also known as oriental swallowtail moths, are widely distributed in the whole of Asia (Palaearctic + Oriental). This family comprises ten genera broadly different in shapes and sizes for both body and wings (J. Minet, 2002). Little is known about the ecology of this family of diurnal moths.

Epicopeia, is mimicking species of butterflies in both the genera *Papilio* and *Byasa* (like *Byasa alcinous*, *Papilio protenor* and *P. helenus*) (Figure 5).

Initially, only the genus *Epicopeia* Westwood, 1841, belonged to the small Epicopeiidae family (Laithwaite & Whalley, 1975). The other species known at that time were placed in the family Epiplemidae, now considered as a subfamily of Uraniidae. When Minet added five genera to Epicopeiidae as well as regrouped all of the known species in the same family (Minet, 1983, 1986). In 2002, two new genera were added: *Deuveia* and *Burmeia*, and Epicopeiidae was placed in the superfamily Drepanoidea (Minet, 2002). Finally, the latest genus described, *Mimapora* in 2017 (Wei & Yei, 2017). At the moment when I write this thesis, this family includes ten genera. Recent molecular studies suggested that the family is in fact, related to the superfamily Geometroidea (Bazinet et al. 2013, Rajaei et al. 2015; Regier et al. 2009) and Pseudobistonidae is suggested to be the sister group of Epicopeiidae (Rajaei et al. 2015; Wang et al. 2019).

Hypotheses of the relationships within Epicopeiidae were first based on morphological characters, which placed *Deuveia* as the sister group of the rest of Epicopeiidae (Minet, 2002). However, already an uncertainty appeared in the placement of *Amana*, which was either sister to *Chatamla* + *Parabraxas*, or sister to a clade containing *Chatamla*, *Parabraxas*, *Schistomitra*, *Nossa* and *Epicopeia* (Figure 6, left). Later, the first molecular phylogeny reconstruction using three genes (*COI*, *EF-1α* and *28S*) was incongruent with Minet's hypotheses (Wei & Yen, 2017). However, it has to be noted that this study 1) mainly focuses on describing a new genus, *Mimaporia*, 2) was poorly supported and even had nodes without support at all (Figure 6, right). But already, they suspected *Epicopeia* and *Nossa* to be paraphyletic. Figure 6 displays the main differences between these two studies.

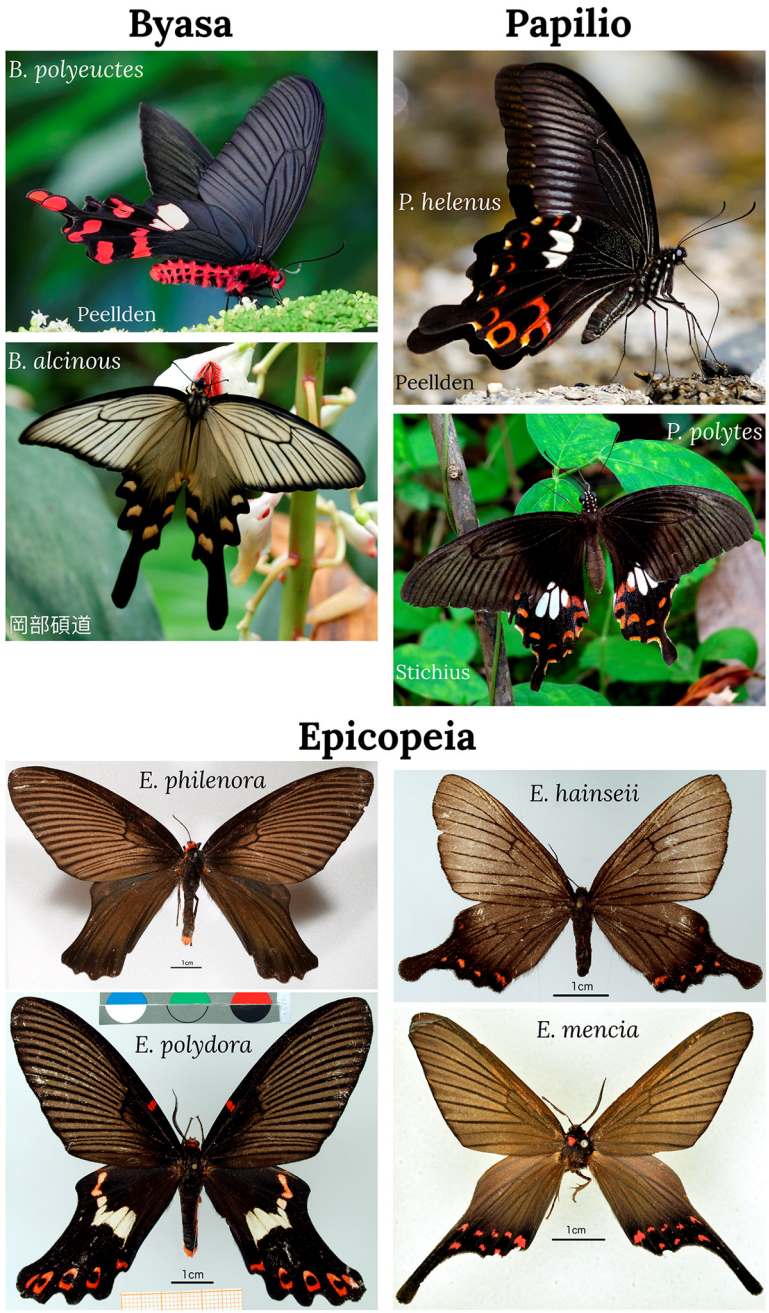

Figure 5. Pictures representing the mimetic *Epicopeia* (bottom) in regard of *Byasa* (top left) and *Papilio* (top right). Image for *Byasa* and *Papilio* are from Wikipedia, and their authors are mentioned in each picture.

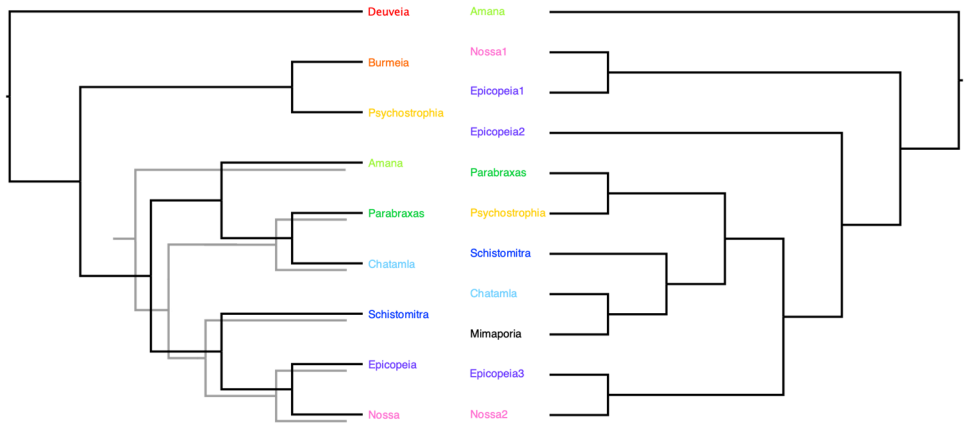

Figure 6. Simplified representation of Epicopeiidae phylogenetic relationships according to Minet, 2002 (left) and Wei and Yen, 2017 (right). Each genus has a specific colour. Minet's alternative hypothesis about the position of *Amana* is represented with grey lines.

3.1.3 Pseudobistonidae

This family is from the north of Thailand and Vietnam to Himalayan regions. Only two species represent this family: *Pseudobiston pinratanai* (Rajaei et al., 2015) and *Heracula discivitta* (H. Wang et al., 2019; Figure 7). These two are monobasic Oriental but allopatric species of Pseudobistonidae (H. Wang et al., 2019).

Pseudobiston pinratanai, the first species placed in Pseudobistonidae, was suggested to be the sister group of Epicopeiidae (Rajaei et al., 2015). This position was then later confirmed with the addition of the second species of this family: *Heracula discivitta* (H. Wang et al., 2019). However, both of these studies used the same eight genes: one mitochondrial (*COI*), and seven nuclear *EF-1α*, *Wingless*, *RpS5*, *MDH*, *GAPDH*, *CAD* and *IDH* (Rajaei et al., 2015; H. Wang et al., 2019). On top of that, they only had few representative species for Epicopeiidae (two in the case of Rajaei et al. 2015 and three for H. Wang et al. 2019) and Sematuridae (two in both cases). Finally, the support values for the nodes of these studies are quite low (Rajaei et al., 2015; H. Wang et al., 2019; chapters 2 and 3).

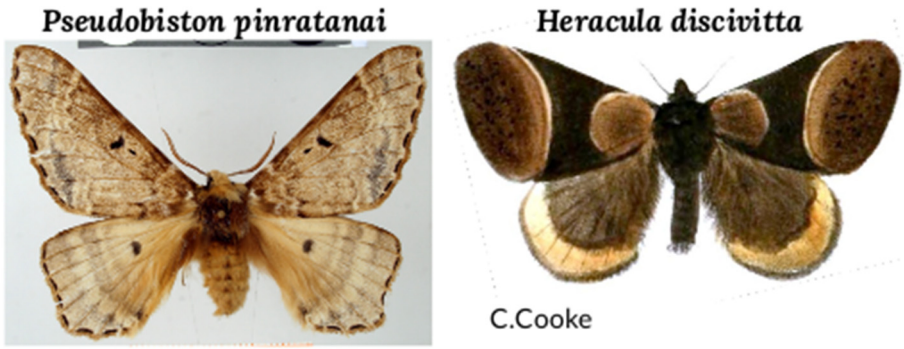

Figure 7. Pictures displaying the two species representing the Pseudobistonidae family, with Pseudobiston pinratanai (left) and Heracula discivitta (right). Image for H. discivitta is from Wikipedia, the author is mentioned in the pictures.

3.1.4 *Pieris napi*

This butterfly, also known as the green-veined white, belongs to the family Pieridae (Papilinoidae, Lepidoptera). This species is widespread across Europe and Asia. I chose to focus on this species, firstly because it is common in Sweden, meaning there is an extensive museum collection accessible in the Biological Museum, Lund University. In addition, the Lund museum possesses samples from distinct Swedish places with a broad range, from Skåne to the more northerly Abisko (Figure 8). These widespread specimens across Sweden are a gold mine for studying population genetics with a window into the past. Secondly, a reference genome is available for *P. napi* (Hill et al., 2019), which makes reference-based genome assembly possible and makes an exceptional basis for characterizing the origin of the obtained DNA.

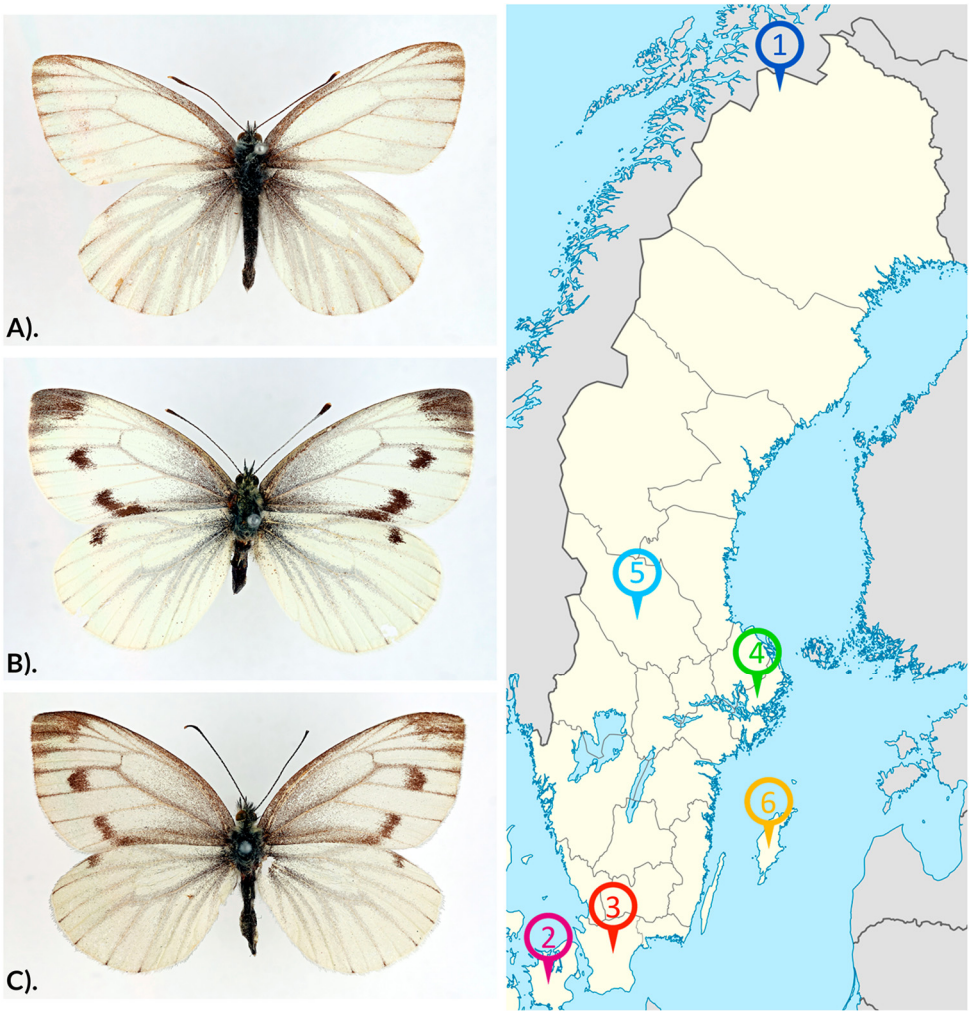

Figure 8. Picture representing dorsal view of Pieris napi (left) and the distribution of six populations across Sweden (right). On the left, A) is a specimen from Abisko (pop. 1) collected in 1906, B) from Stockholm county (pop 4) in 1885, C) from Denmark (pop. 2), in 1909. On the right, populations from Sweden (pop. 1, 3, 4, 5 and 6) and Denmark (pop.1). The number inside the point represents the population.

3.2 Laboratory protocol

3.2.1 DNA extractions

The first issue I faced was to figure out the best way to extract DNA. For this purpose, the beginning of my PhD project was dedicated to investigating different approaches to DNA extractions, with varying degrees of success. Following the literature, I started with two methods: saturated NaCl protocol (Feng et al., 2011; Rohland & Hofreiter, 2007) and phenol-chloroform (Barnett & Larson, 2012; Sambrook & Russell, 2006).

First attempts and failures with NaCl

The custom protocol I tried for my first extractions was based on previous methodologies used for ancient DNA of both cave and brown bears (Hofreiter et al., 2004; Leonard et al., 2000), as well as dried butterfly specimens (Feng et al., 2011). I chose 12 samples of Pieris napi (Pieridae) to run my first test. They were collected between 1878 and 2008. The DNA concentration of these samples was checked on NanoDropTM 2000/2000c. Their concentrations were quite low. I double-checked their DNA concentration as well as the length of the fragments on BioAnalyzer DNA 1000 Assay (Figure 9). It appeared that the DNA was either not present at all or extremely fragmented (<100 bp). The results were negative even for the most recent specimens (2008), which led me to conclude that this extraction method was not suitable for dried entomological samples.

Back to basics: Phenol-Chloroform

This DNA extraction method is widely accepted and has been used for a long time (Barnett & Larson, 2012; Sambrook & Russell, 2006). It is known to be a cheap and efficient extraction method. However, this protocol has disadvantages. First, it requires the use of a hazardous chemical reagent, and thus this approach requires careful handling and high precautions. Safety in the laboratory is a priority when using dangerous chemicals, particularly the use of chloroform requires to work under a fume hood. While phenol is volatile and can burn the skin and therefore necessitates to wear gloves and lab coat up to the wrist. Secondly, these hazardous components also demand specific handling in their storage and treatment of the wastes. Nowadays, several laboratories dropped the use of such chemical and advocate for alternatives. Finally, it is a time-consuming technique, as one batch of samples requires three days of processing. Ultimately, I compared the quality of the DNA obtained from this extraction technique to the DNA obtained with extraction kits, and found no significant difference between the two.

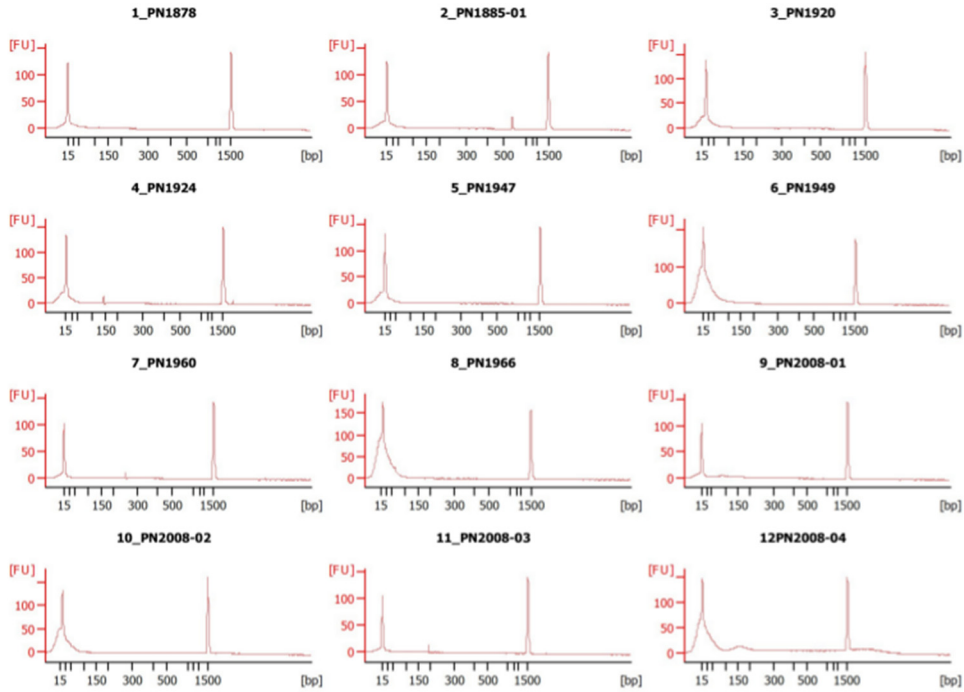

Figure 9. Display of the DNA concentration of the 12 Pieris napi samples extracted with saturated NaCl, on BioAnalyzer. Each plot represents the size (bp) versus the fluorescence intensity. The two peaks around 15 bp and 1,500 bp are the specific DNA markers, the Lower and the Upper, respectively.

Let the kits out of the bag

The methods based on commercially available kits are getting more and more popular among researchers. I found them straightforward and relatively quick to use; as simple as following a recipe. But they are not as cheap as Phenol-Chloroform. However, there is a large panel of brands and various kits; this competition leads over time to a decrease in prices. There is no significant difference in quality between different kits available and phenol-chloroform method. Especially nowadays, as the diversity of available kits make them highly specific depending on extracted samples. For example, the QIAamp DNA Micro (Qiagen) is optimized to purify DNA from small sample sizes, making it particularly suitable for degraded DNA of museum specimens. Thus, the difference resides in the trade-off between the money invested in the project and the time constraint. As the quality of the DNA extracts is similar regarding kits or phenol-chloroform approaches, these kits can be an excellent investment for a lab. In this study, I used three different kits: Dneasy Blood & Tissue kit (Qiagen, USA), QIAamp DNA Microkit (Qiagen, USA) and NucleoSpin® Tissue (Marcherey-Nagel).

3.2.2 Molecular methods and sequencing

To sequence DNA with NGS technologies, regardless of the approach (*i.e.* genome-reduction techniques such as TE or RADseq; or WGS as well), the next essential step is to construct libraries. NGS library is a collection of millions of DNA fragments that together can represent the entire genome of an organism. For fresh material, this stage is highly standardized. Also, because it profoundly depends on the sequencing platform used, usually research groups send their DNA extracts to sequencing companies, and they prepare the libraries. However, for highly fragmented genomes, especially in the case of aDNA and hDNA, the standardized protocols are not optimized. Therefore, aDNA studies use custom-built library preparation methods. Hence, I went to Stockholm to directly learn library preparation protocols for ancient DNA from Love Dalèn's team, which is working on palaeogenetics. With the help of Nicolas Dussex and Johanna von Seth, I prepared libraries from specimens of *Mania lunus* (Sematuridae).

The most common damage that can occur in hDNA is the deamination of cytosine to uracil. This process can result in a wrong interpretation of the sequence (Rowe et al. 2011). Another kind of damage that might occur in old samples is cytosine to thymine misincorporation (Brotherton et al. 2007; Briggs et al. 2010; Bi et al. 2013). One solution to bypass these issues is to repair with USER (Uracil-Specific Excision Reagent) enzyme (NEB, USA) during the first stage of the library preparation (Meyer & Krisher, 2010). This enzyme is one of the uracil DNA glycosylases (UDG; see par 1.4.1.). For more details on the library preparation protocol, see chapter 3. With this optimized protocol, the libraries were successfully prepared. Both with BioAnalyzer and after migration of the libraries on 1.2% gel (at 90 Volt for 2.5 hours), we observed they have an average length of 150 bp (Figure 10).

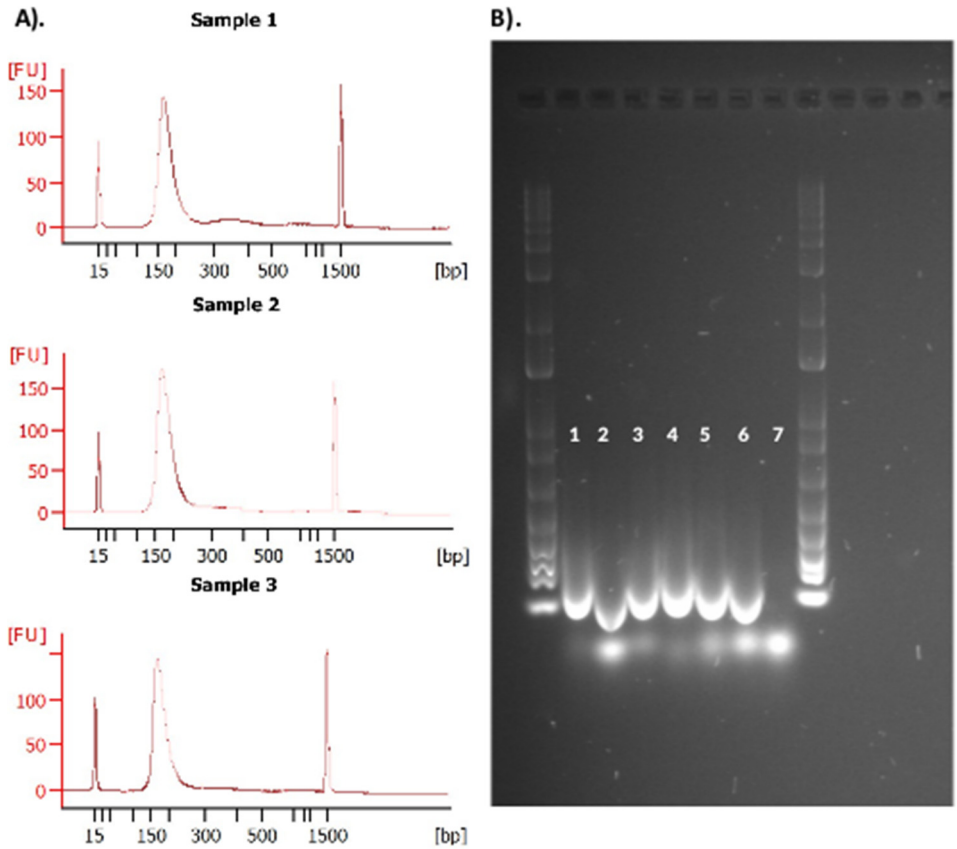

Figure 10. Example of successful library preparations. A) Display DNA concentration of three samples of Pieris napi from BioAnalyzer. B) 5µl of six libraries have been run on 1.2% gel at 90 volt for 2.5 hours. Sample 1 to 3 in the column A correspond to the same number in B.

TE

Target Enrichment (TE) is one of the numerous genome-reduction strategies. This method, specially designed for phylogenomics (Lemmon et al., 2012), targets specific segments of the genomes. Therefore, it is particularly used in phylogenetics studies (Brandley et al., 2015; Breinholt et al., 2018; Espeland et al., 2018, 2019; Gasc et al., 2016; Toussaint et al., 2018). The idea of this approach is to use tiled probes to capture not only the genes of interest but also the flanking regions (Figure 11). Detailed protocol on the library preparation for TE is available in chapters 1 and 2.

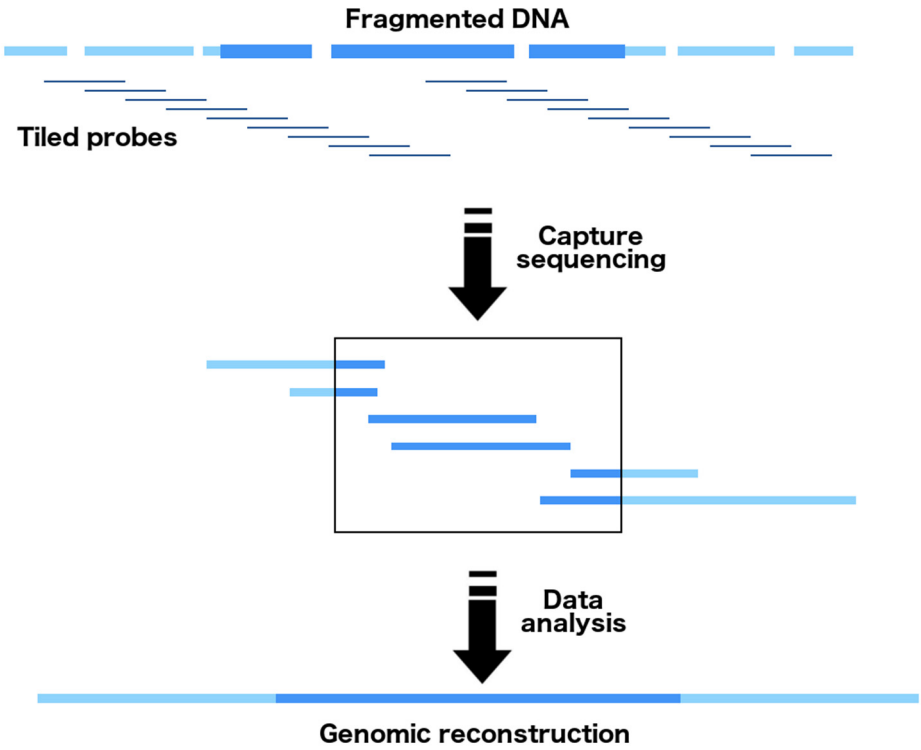

Figure 11. Workflow of Target Enrichment capture. The tiled probes are based on recent or close species and are used to enrich fragmented DNA. Once sequenced, the segment of interest (dark blue), as well as the flanking regions (light blue) can be reconstructed (based on Gasc et al. 2016).

WGS

Whole-Genome Sequencing (WGS) methods are the opposite of genome-reduction strategies. Here, the aim is to determine the full genome of an organism. The idea of this approach is to sequence everything possible and determine bioinformatically afterwards what has been sequenced (Allio et al., 2019; Burrell et al., 2015; Cloutier et al., 2019; Ekblom & Wolf, 2014; Ng & Kirkness, 2010; Zhang, Cong, et al., 2019). Detailed protocol on the library preparation for WGS, as well as data treatment, is available in chapters 3 and 4.

3.3 Bioinformatics

Following sequencing, data need to be cleaned up and handled in a certain way to make sense of them (Figure 12). When building libraries, adapters and indexes are added to DNA fragments. Indexes are assigned to distinguish each individual. Indeed, the technology used often requires pooling a group of samples together. For instance, Lepidoptera genomes tend to be smaller than 500 Mb (Triant et al., 2018). Hence for a sequencing run on HiSeq X (Illumina), it is possible to group ten specimens with the expectation of getting 20X coverage for the whole genome. Thus, we attribute to them unique indexes to be able to decide which reads correspond to which specimens. These indexes are removed as the sequencing platform demultiplexed them for us.

3.3.1 Clean up & data assembly

Adapters allow DNA pieces to attach to the sequencer, and to be read. The first step once receiving genomics data is to check the quality of the reads and clean them from adapters, ambiguous reads (with Ns), as well as over-represented sequences that might be sequencing errors or contamination.

During the sequencing process, a measure of the quality of the identification, the Phred quality score, is assigned to each nucleotide base (Ewing et al., 1998; Ewing & Green, 1998). This score, averaged for all reads, represents the logarithmic probability of error when reading a base (Ewing & Green, 1998). For example, a Phred quality score of 30 represents a probability of 1 in 1,000 that this base (or read) is incorrect, *i.e.* the base call accuracy is 99.9%. Therefore, when doing a quality check on the reads, a threshold of a Phred scores lower than 30 (or 25 in the more relaxed way) is applied.

Despite the fact our samples were sequenced in paired-end, the resulting sequencing data was carried forward as single-end. The reason behind this choice is that degraded DNA is likely to randomly ligate together during the adapter ligation step of library preparation, resulting in chimaeras of genomic regions (Eske Willerslev & Cooper, 2005). However, the sequencing information contained in the reads is still reliable, and therefore more accurate results are obtained by treating data as single-end (Rowe et al., 2011).

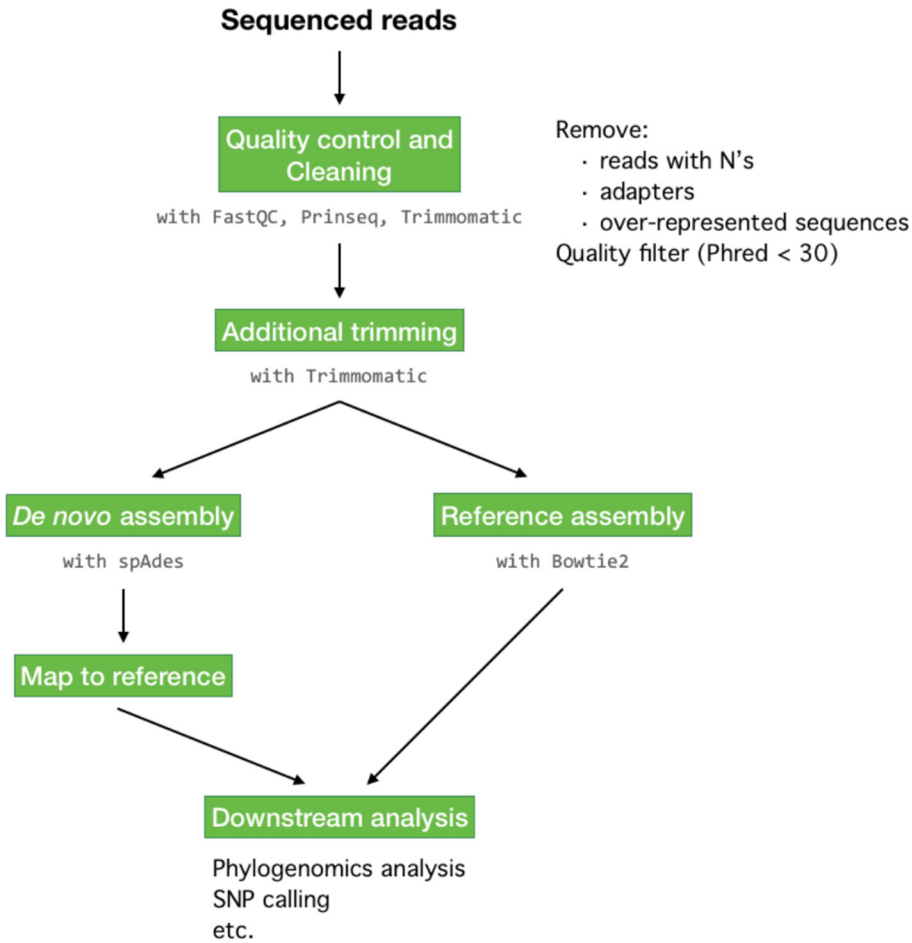

Figure 12. Bioinformatics workflow. The main steps are framed in green boxes. Examples of programs used in this thesis to accomplish each step, are listed under them.

In the case of *Pieris napi*, for example, as the complete genome of this butterfly has been sequenced (Hill et al., 2019), it is possible to map my dataset against this reference, in other words: perform a reference assembly. The reads are assembled against an existing backbone sequence, called a reference. Thus, it builds a sequence that is similar, but not necessarily identical to the reference (see chapter 4). In contrast, *de novo* assembly means to assemble reads against each other to create a full-length (sometimes novel) sequence without using a template. This method is often used when no backbone is available and was used, for instance, for Geometroidea moths (see chapter 3). However, this type of assembly is computationally demanding.

3.3.2 Phylogenomics

With phylogenomics analysis, researchers can infer evolutionary relationships among different taxonomic clades and understand better the mechanisms of molecular evolution (Philippe & Blanchette, 2007). The first mention of this concept, bringing together phylogenetics and genomics, was in the context of an "approach to the prediction of gene function" for genome-scale data (Eisen, 1998). Then, it was extended to phylogenetic inferences (O'Brien & Stanyon, 1999).

Two NGS methods were used to investigate the phylogenetic relationships of Epicopeiidae, Sematuridae and Pseudobistonidae: TE and WGS. With the TE approach, we sequenced 33 specimens (29 Epicopeiidae, 2 Sematuridae, and 2 Pseudobistonidae). The kit used in this method includes 2,745 probe regions (Breinholt et al., 2018; Espeland et al., 2018, chapter 1). We selected loci that were enriched in at least 20 specimens, and thus focus on a dataset of 378 loci, which correspond to 327 nuclear genes (chapter 2). In the WGS method, we sequenced 32 specimens (16 Epicopeiidae and 16 Sematuridae (chapter 3). Both datasets were combined in the final dataset that corresponds to 308 nuclear genes across 45 species (chapters 2 and 3).

There are two different strategies to perform model-based phylogenomics inferences, maximum likelihood (ML) or Bayesian inference (BI). These two approaches principally depend on the substitution process (i.e. the change from a nucleotide, or amino acid, to another), and involved likelihood calculations. The likelihood is the probability of observing the actual data that have been collected given the model (Harmon, 2018). We refer to the maximum likelihood (ML) as the estimates of parameters that give the highest likelihood, in other words, that provide the highest likelihood of obtaining this data, *i.e.* what is the likelihood of the data given the model (Harmon, 2018). In contrast, Bayesian inference (BI) attempts to estimate the probability that the model (*i.e.* DNA evolution model and the inferred tree) is correct given the data, and parameter values are considered as statistical distributions, with parameters assigned prior distributions that are then modified by the data to estimate posterior distributions (Yang & Rannala, 2012; Young & Gillung, 2020).

Earlier algorithms to infer phylogeny were relevant and broadly used before NGS data arose. NGS methods produce an incredible amount of data in comparison with Sanger sequencing; this leads to an incompatibility with the previous algorithms which were not designed to deal with such a vast amount of data. Therefore, there is no real consensus on the best method, and the scientific community is still questioning phylogenomics inference models, criteria and parameters, and continues to test the new ones (Young & Gillung, 2020). Hence, it is crucial to wisely select the model that will

describe the best evolutionary mechanisms of the data. A large variety of criteria are available to support this model selection. To name a few of them: the Akaike information criterion (AIC) and the corrected AIC (AICc) (Akaike, 1974; Sugiura, 1978), the Bayesian information criterion (BIC) (Schwarz, 1978), and the hierarchical and dynamic likelihood ratio tests hLRT and dLRT, respectively (Posada & Crandall, 2001). Commonly, the main model selection programs use these criteria, and are based on frequentist inference. This type of statistical inference takes the frequency (or proportion) of the data into account (Young & Gillung, 2020); its primary alternative approach is Bayesian inference (BI). Under the BI, the selection of the models depends on the Bayes factor which is a ratio quantifying the probability of a model to be more relevant, to describe the data, than the other (Goodman, 1999; Young & Gillung, 2020).

Another critical thing to keep in mind is that there are two ways to analyse protein-coding genes: as amino acids, nucleotides or codons (i.e. triplets of nucleotides that can be translated into an amino acid). Therefore, deciding on which type of data will be implemented in phylogenomics analyses is as critical as selecting the model, especially since there is such a disparity in the statistical analysis of amino acid and nucleotide data (Huelsenbeck et al., 2008). For amino acid models, most of the parameters are fixed values that are predetermined by the model (Yang & Rannala, 2012). While in this case, it is less computationally heavy to analyse, it may not be the best model selected, and improper model can lead to wrong phylogeny inferences (Buckley, 2002; E. D. Cooper, 2014). Therefore, one should be extremely careful when selecting models for phylogenomic reconstruction, particularly with amino acid models (Young & Gillung, 2020). Nucleotide models are less controversial, as their parameters are directly estimated from the data (Young & Gillung, 2020). Additionally, some studies suggest that using the model with the most parameters resulted in phylogenetic inferences as accurate as when a model has been selected for this dataset (Abadi et al., 2019; Young & Gillung, 2020).

Consequently, here, the phylogenetic relationships were inferred based on a nucleotide model. The model selection was performed using ModelFinder (Kalyaanamoorthy et al., 2017) in chapter 2 and 3. In chapter 2, the data were divided into partitions based on their rates of evolution using the RatePartitions algorithm (Rota et al., 2018). Finally, the phylogenetic relationships were inferred with IQ-TREE 1.6.10 (Chernomor et al., 2016; Nguyen et al., 2015), where ModelFinder is already implemented, under the maximum likelihood (ML) criterion.

3.3.3 SNP genotyping and population genetics approach

A change, also called a substitution, of a single nucleotide base that occurs at a specific genomic position, is called a single-nucleotide polymorphism (SNP). This variation can occur in both coding and non-coding region. The principal point that makes them relevant to be classified as SNP is if more than 1% of a population does not carry the same nucleotide at a precise location in the genome. Therefore, SNPs are particularly relevant for population genomics studies.

However, separating true polymorphisms from sequencing or alignment errors can be an issue. To avoid this, extensive filtering based on statistical models is necessary. Moreover, in the case of non-model organisms variant sites are usually not known, making the SNP detection even trickier.

This filtering method focus on non-model organisms, for which there typically are no known variant sites. There are three crucial steps to this approach (De Wit et al. 2012). First, alignment files must be processed, and the poorly mapped regions must be re-aligned. The second step consists of calling the SNP and differentiates the correct variant from the false positives. Finally, information of the variants for all sample needs to be extracted. Once all the SNPs for all sequenced specimens have been called, allele and genotype frequencies can be calculated. We can perform a Principal Components Analysis (PCA), as well. This will allow us to detect, at a larger scale, the differences between our given populations. Eventually, some population genetics statistics can be calculated, such as heterozygosity (HE) or F_{ST} (see chapter 4). However, one should keep in mind this section (chapter 4) does not have enough individuals per populations to be able to conduct such a broad population genetics study. Therefore, this chapter is meant to be a proof of concept study to demonstrate that it is possible to call SNP from museum specimens, including material that is >100 years old (hDNA). We can thus also use such data at a population genetics level.

4 Results and discussion

4.1 Phylogenetic relationships of the three families of Geometroidea

4.1.1 Across taxa

In chapter 2, we conducted a phylogenetic analysis based on 378 nuclear loci on 27 Epicopeiidae representative specimens, three Sematuridae and two *P. pinratanai*. We demonstrated strong support for the Pseudobistonidae as a sister group of Epicopeiidae, and Sematuridae as the sister group to the other two. And this hypothesis was later corroborated in chapter 3, for which this time we have 308 genes, 39 Epicopeiidae specimens, 14 Sematuridae and 2 *P. pinratanai*. Chapter 3 also supported the monophyly of Epicopeiidae, as well as the position of Pseudobistonidae, reinforcing the case for the new family (Figure 13).

4.1.2 Sematuridae

In chapter 2, we first find that *Anurapteryx* is the sister group to the rest of the family. This relationship was later confirmed in chapter 3, with strong support in both cases. We demonstrate as well in chapter 3 that *Homidiana* and *Coronidia* constitute a clade that is sister to Mania. Within Mania, we clarified the identification of several specimens that were labelled either *Mania empedocles* or *M. diana*, and now gathered under *M. lunus*. We note that the published transcriptome of a *Mania* species (Kawahara et al., 2019; Kawahara & Breinholt, 2014) belongs to the species *M. diana*, not *M. lunus* as reported. We confirm M. empedocles is sister to *M. diana* + *M. lunus* (Figure 13). Our study is only missing *M. aegisthus*, which is endemic to Jamaica, from this genus. More details are given in chapter 3.

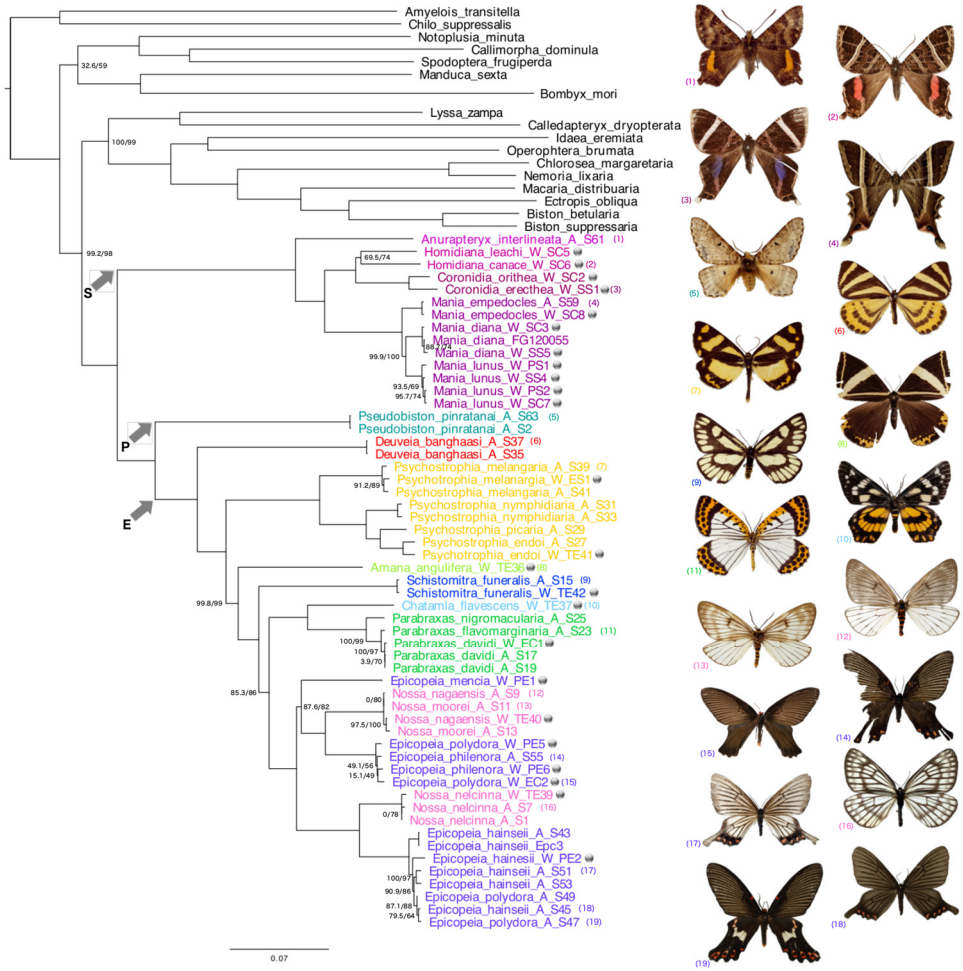

Figure 13. Phylogenetic tree from IQTREE analysis of 45 species, based on 308 genes. When displayed, numbers are SH-aLRT support (%) / ultrafast bootstrap support (%). If the support values are equal to 100/100, they are not shown on the nodes. The images are representative species (indicated with a number in parenthesis; not to scale). Each of the family is represented by an arrow and a letter, for Sematuridae (S), Pseudobistonidae (P) and Epicopeiidae (E). Specimens marked with an 'A' correspond to specimens sequenced with TE (chapter 2), while 'W' and a grey dot indicates WGS (chapter 3). See chapter 3 for more details.

4.1.3 Epicopeiidae

Within Epicopeiidae, the TE results (chapter 2) are closely corresponding with Minet's (2002) hypotheses, and therefore profoundly inconsistent with Wei and Yen (2017), although we obtained better support (Figures 6, 13). Later, the addition of our WGS data confirmed our previous hypothesis and reinforced once again the backing of our data (chapter 3, figure 12). We showed *Deuveia* as the sister group of the other Epicopeiidae genera. *Psychostrophia* is a consistent clade and is sister to the rest minus

Deuveia (chapters 2 and 3, figure 13). To continue with the disagreement with Wei and Yen's results (2017), we demonstrated that *Parabraxas*, a consistent clade, is a sister group to *Chatamla*, as Minet suggested back in 2002. These two genera are together in a clade with *Schistomitra* and the group of *Epicopeia* and *Nossa* (chapter 3). However, conflicting with Minet's hypotheses, as well as Wei and Yen, we observed *Schistomitra funeralis* to be the sister group of (*Parabraxas* + *Chatamla*) + (*Epicopeia* + *Nossa*). While these results were first not well supported in chapter 2, the inclusion of taxa, such as *Amana* and *Chatamla*, helped to stabilize this position and reinforced our hypothesis with high support (chapter 3).

Epicopeia & Nossa

One clear pattern that has been emerging since chapter 2 is that the evolutionary relationships of *Epicopeia* and *Nossa* are tangled together. There is no clear distinction between the two genera. This kind of relationship is characteristic of paraphyletic clades (see glossary). When we added WGS specimens, we found the same pattern again (chapter 3).

Nossa species morphologically look similar to each other, yet are quite different from *Epicopeia*. It appeared there are two groups of *Nossa*: *N. nagaensis* and *N. moorei* on one side, and *N. nelcinna* on the other side (Figure 13). It has to be mentioned here that some species of *Nossa nelcinna* that were misidentified as *N. palaearctica* and *N. chinensis* in chapter 2. Regarding *Nossa nagaensis* and *N. moorei*, they seem to be genetically identical and should perhaps be considered the same species. More work both on a morphological and genomics level need to be performed to identify this genus correctly.

Our analyses in chapter 3 also highlighted the fact that *Epicopeia* specimens were probably misidentified and hard to tell apart. We clarified the labeling of most of these specimens (see chapter 3 for more details), to discern three groups in *Epicopeia*: *mencia* on its own, *philenora* + *polydora* and *hainesii* + *polydora*. Within the two groups comprising *E. polydora*, we observed very short branches, meaning these specimens are genetically very close to each other. These short branches may be due to the genes we selected. Indeed, the used dataset of genes was derived from the previous TE dataset (see chapters 2 and 3), indicating a bias towards the TE dataset. In TE, probes are designed based on highly conserved genomic regions, making them able to capture loci across a large phylogenetic range (chapter 1 and 2). Therefore, more investigation on our WGS data will allow us to find less conserved regions that, hopefully, will enhance our understanding of this group (chapter 3).

4.2 Misidentification in museums collections

Genomic data revealed some misidentifications in the museum specimens used in these studies, especially on Epicopeiidae and Sematuridae (Chapter 3). Admittedly, there is no recent literature to help to identify Sematuridae specimens. However, often, simple morphological comparisons with photography, when available, or paintings from the original descriptions were enough to sort out their identification, as it was the case with *Homidiana* (Figure 14). For instance, *Coronidia erecthea* was labelled as *H. canace*, when compared with other *H. canace* and other *Homidiana* species (Boisduval & Guénée, 1836; Seitz, 1913) it is evident that it is a specimen of *C. erecthea* (Figure 14). Another unclear *Homidiana* was a specimen labelled as *H. leachi*, at first it was suspected to be *H. egina*, but it turns out to be *H. canace* instead (Figure 14). In the case of the genus *Mania*, the misidentifications were highlighted by the phylogeny (Chapter 3). Indeed, we obtain three different groups that should correspond to each of the three species but did not. After comparing them to pictures of the type specimens, it appears that most of these misidentified specimens were either *M. empedocles* or *M. diana* lumped under *M. lunus* (Figure 15). Lastly, we also notice that the published transcriptome of a *Mania* species (Kawahara et al., 2019; Kawahara & Breinholt, 2014) belongs to the species *M. diana*, instead of *M. lunus* as reported (Figure 15).

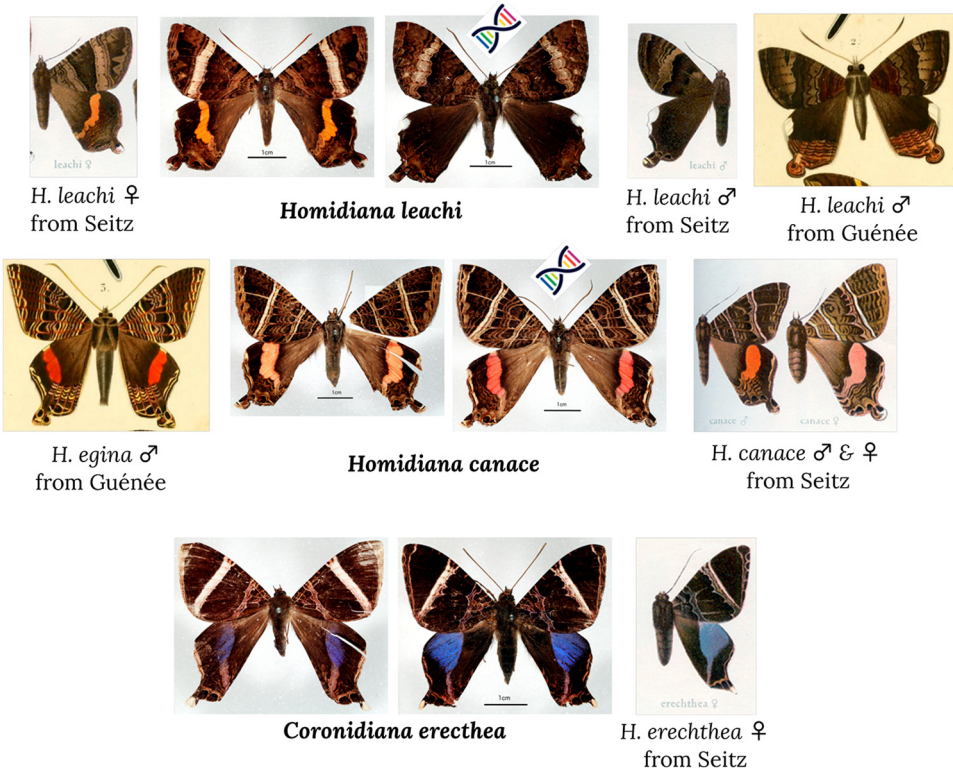

Figure 14. Figure representing the morphologies for *Homidiana* (top) and *Coronidia* (bottom) of the loaned specimens (photography) and paintings from Boisduval and Guénée, 1836 and Seitz, 1913. Specimens for which DNA was obtained are marked with a DNA logo.

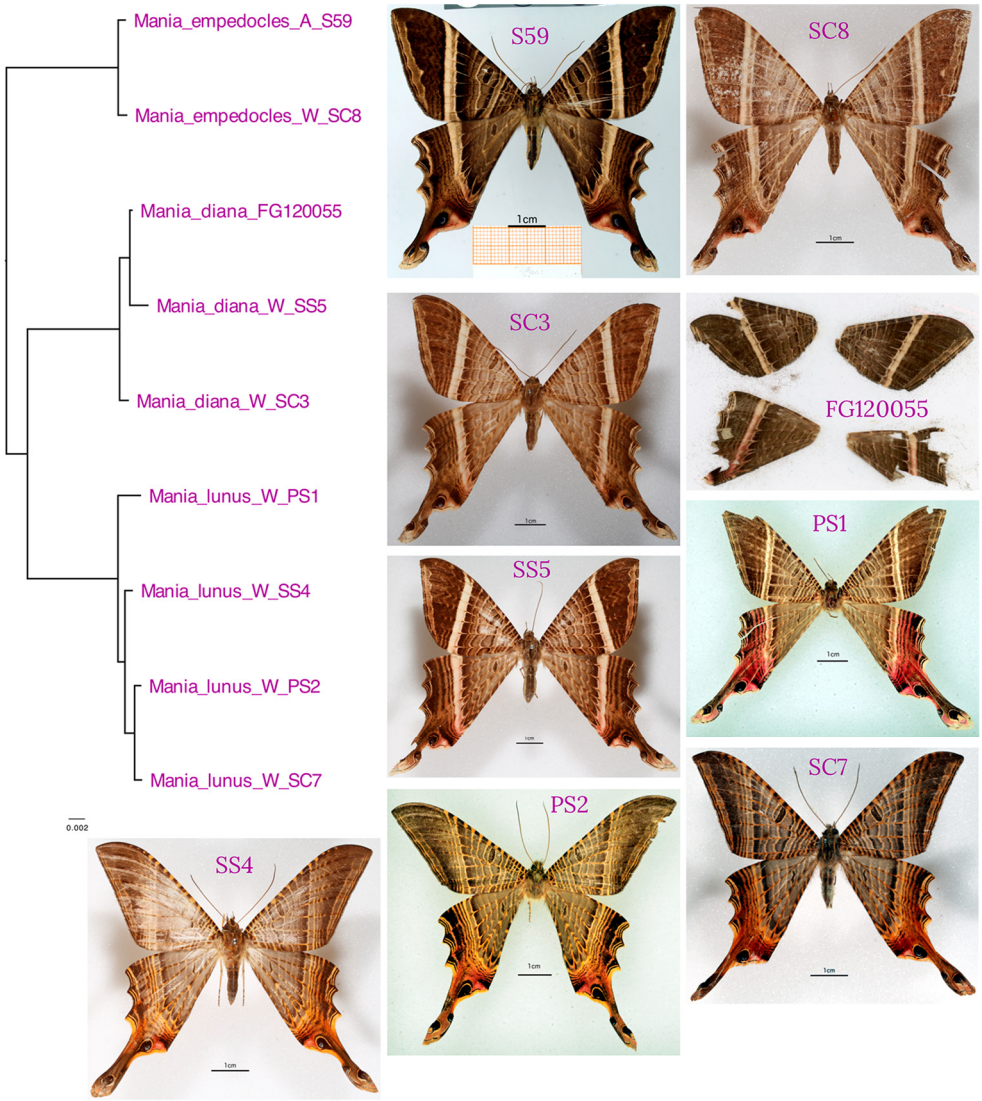

Figure 15. Morphological aspects of the six loaned *Mania* in regard to their phylogenetic relationships. Individuals sequenced with TE are marked with an 'A' while samples sequenced in WGS are 'W'. Individual with the code 'FG120055' is the published transcriptome of a *Mania* species.

Clarifying the misidentifications in Epicopeiidae was trickier, in particular regarding *Epicopeia*, as the phylogeny was less straightforward (Figure 13). *Epicopeia hainseii* and *E. mencia* are morphologically alike, displaying differences principally in the shapes of their tails (Figure 16; Chapter 3). Nevertheless, it appears that *E. mencia* possesses a long and slender tail, whereas *E. hainseii* has a short and thicker one. Additionally, the veins in the hind wings look more distinct for *E. mencia*. Accordingly, these

morphological differences suggest that two specimens, previously labelled as *E. mencia*, are *E. hainseii*. In contrast, *E. polydora* and *E. philenora* show similar morphology but are highly divergent to the former two species (Figure 16). Two of our sequenced *E. polydora* are genetically very similar to two *E. philenora*, while two other *E. polydora* are genetically similar to *E. hainseii*. The latter three are all genetically alike except *E. mencia*, which appears to be a distinct lineage not related to the other *E. hainseii* specimens (Figure 16). Finally, a specimen which is labelled *E. polydora* looks morphologically different from the other *E. polydora* presented in this study (Figure 16). Could it be possible that the differences observed in morphologies are due to sexual dimorphism, like in *Homidiana*? Or maybe it is caused by a wide range of phenotypes? The literature on these species is unfortunately too sparse; hence further work is needed to identify these specimens accurately.

4.3 TE vs. WGS

First, we successfully sequenced museum specimens using a target enrichment (TE) kit designed for Lepidoptera (chapter 1). The oldest sample, collected in 1892, generated over 500 loci while we obtained up to 1747 loci for more recent material. Here we extended the taxon sampling to other families and superfamilies of Lepidoptera, which many are not yet included in high-throughput molecular studies. Hence, we demonstrated that our kit is efficient on the entire order of Lepidoptera, providing comparable data and proper resolution at a large phylogenetic scale.

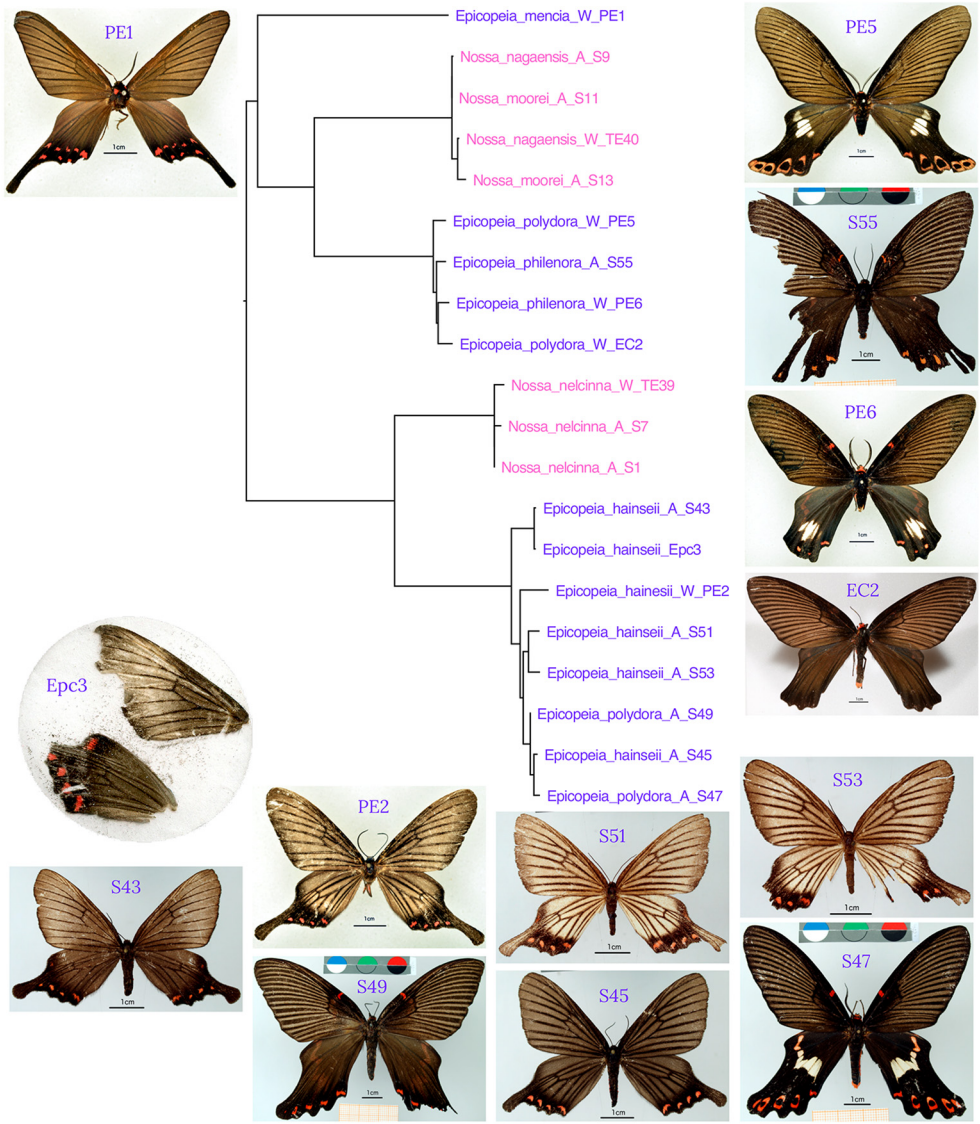

Figure 16. Morphological aspects of the thirteen loaned *Epicopeia* in regard to their phylogenetic relationships. Individuals sequenced with TE are marked with an 'A' while samples sequenced in WGS are 'W'. Individual with the code 'Epc3' is the published transcriptome of *Epicopeia hainseii*.

Our TE dataset, on Epicopeiidae and Sematuridae, included 378 nuclear loci (327 genes), for a total length of 134,881 base pairs (bp), with an average size of 336 bp (chapter 2). We used the complete references for these 327 genes to create our WGS dataset. We recovered 308 genes of the 327 used in the TE study (chapter 3). We compared the distribution of the size of the genes, to find a significant difference

between the two NGS methods (Figure 17). While with TE, we are sure to obtain a good part of the genomic regions we are looking for, the size of the aligned sequences is more than twice with WGS (chapter 3). On average, TE genes are 336 bp long (range: 6-1,752), whereas WGS genes are 786 bp (range: 27-7,167 bp). One thing to note here is that phylogenetic analyses are affected by the size of the dataset. The longer the DNA fragments are, the more robust the statistical analysis is (Chapus et al., 2005; McHardy et al., 2007).

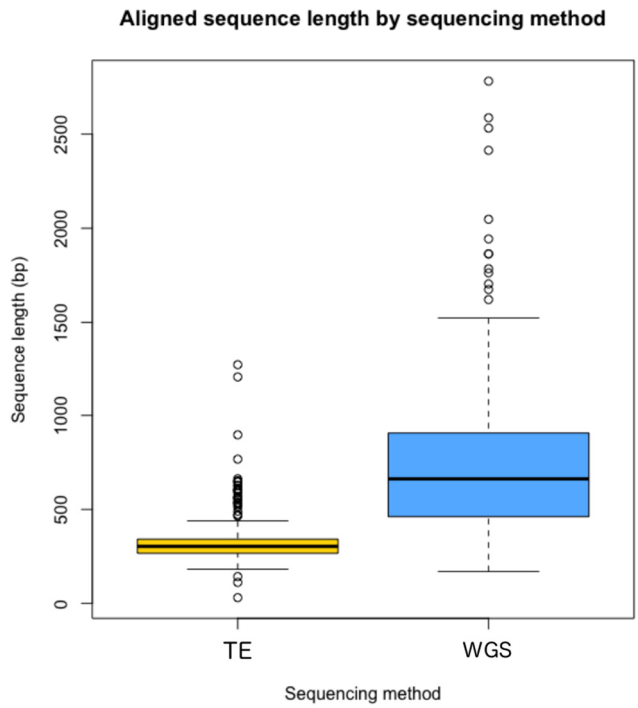

Figure 17. Distribution of the average size (bp) recovered per gene regarding the TE (yellow) and WGS (blue) methods.

In principle with WGS, the entire gene should be recovered. While with museum specimens, we do get broader parts of the genes (an average of 786 bp), the genes themselves are highly fragmented. Meaning that even if we do get in total a more extended coverage of the gene, we have more pieces. Moreover, not all genes were found with the WGS approach. Admittedly, it is less probable to get the targeted genes, but if present, they are usually longer.

One should keep in mind that this analysis is biased towards TE. Indeed, we used the successfully recovered loci of chapter 2 to create the reference dataset. Due to the time

constraint, I was not able to use the complete reference set of TE, which consists of more than 2,000 genes (chapter 1). Therefore, further analysis, including this entire set of reference, to provide a better comparison between the two approaches will be carried out in the future. Another way for this comparison would be to sequence the same specimens using TE and WGS. It is unfortunately impossible since TE method used the entire DNA extraction.

Genome reductions methods, such as TE, are an incredibly valuable asset to research and enhance our understanding the genomes of non-model organisms. However, museum specimens already have degraded DNA (Staats et al., 2013). Hence, genome reduction methods imply an extra level of data loss. Depending on the circumstances, genome reduction methods are not always the best approach, notably regarding museum specimens. Our conclusions lead us to think that WGS is more relevant, at least in the case of fragmented genomes, such as museum specimens. Equally important regarding rare specimens, is the fact that it is more desirable to obtain as much data as possible, including information that might be useful in the future.

4.4 Population genetics (*Pieris napi*)

The principal advantage in this study is that a reference genome is available for this species (Hill et al., 2019). Therefore, we can know precisely what amount of data correspond to the targeted genome. In this case, we recovered on average 81.40% of *Pieris napi* genome, which represents an average coverage of 15.6X for the nuclear genome (chapter 4). In theory, as *P. napi* has an estimated size of 350 Mb (Hill et al., 2019), for a sequencing run on HiSeq X (Illumina) when grouping ten individuals, the expected coverage is less than 20X. Regarding the mitochondrial genome, for each individual, we recovered 100% of it with an average coverage of 1963X (chapter 4). Hence, our results with museum samples, including specimens >100 years old, are thrilling and encouraging.

We recovered a positive correlation between the age and the number of cleaned reads that belongs to the butterfly (Figure 18; chapter 4). We expected that age has a negative effect on DNA preservation. However, there are still many variations, with, sometimes, younger specimens giving fewer reads than older ones (Figure 18). Therefore, there are other forces affecting DNA preservation than age only. Indeed, how samples are handled after they are collected can play an essential role in DNA preservation. For instance, it has been shown that how they were processed after their death has a crucial impact on the molecular level (Quicke et al., 1999). Propylene glycol, commonly used for trapping and storage, can also ruin the chance to get DNA (Ballare et al., 2019;

Short et al., 2018). Pest control can affect the DNA too; for example, high-temperature treatment can damage the DNA (Ackery et al., 2004), as well as sulfuryl fluoride (Su & Scheffrahn, 1990). Information on how specimens have been collected and managed through the years is not always known, especially in the case of old material (>100 years). Nevertheless, common usages for storage and pest control are freezing treatments, and seem to have a positive effect on DNA preservation (Ballare et al., 2019; Quicke et al., 1999; Short et al., 2018), and might explain why we get more DNA from some older specimens than young ones.

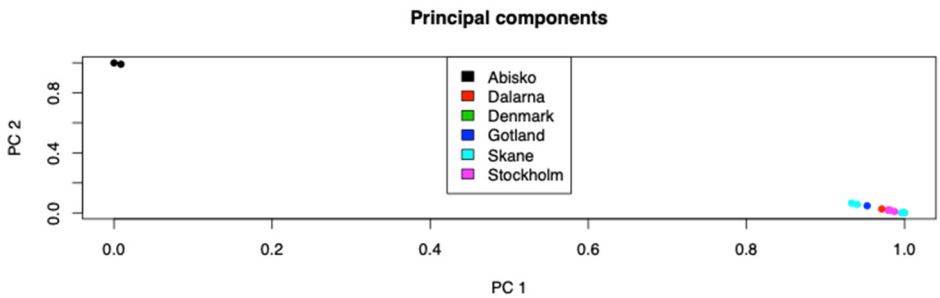

Figure 18. Principal Components Analysis (PCA) plot. Locations of each individuals are marked with colours.

The recovered genome of *Pieris napi* allowed us also to calculate genome-wide population genetics statistics (heterozygosity, F_{ST}, admixture and inbreeding), we also inferred demographic history (see chapter 4 for more details). Our results reveal a significant genetic differentiation between the individuals sampled in Abisko (Arctic Circle) and the rest of our populations. It has been suggested that Arctic Circle populations are a different subspecies (Espeland et al., 2007; Petersen, 1949; Porter, 1997; Tolman, 1997), named *Pieris napi adalwinda* (Fruhstorfer, 1909), in contrast to *P. n. napi* (Linneaus, 1758) found elsewhere in Europe. Hence, here we found genetic evidence for this subspecies separation. Subsequently, these results are promising and lead the way for future museomics projects on population genetics.

5 Conclusions & future perspectives

5.1 The long and the short of it

The age of museomics is upon us. The present sequencing methods allow us to reveal the molecular secrets dormant in museum specimens. Although age and the way these samples have been handled affect the quality and quantity of DNA, with care and rigour, Next-Generation Sequencing technologies produce suitable data from such specimens. Laboratory protocols need readjustments and optimization depending on the targeted biological model, but they showed their potential.

Here, we successfully recovered with Target Enrichment (TE) 31 genomes of Epicopeiidae, Sematuridae and Pseudobistonidae, including a >127-year-old specimen. Furthermore, we obtained 32 whole-genomes of Epicopeiidae and Sematuridae, with an additional 13 whole-genomes of *Pieris napi* (Pieridae), including in both cases >100-year-old specimens. We estimated the size of the genomes of the moth specimens (*i.e.* Epicopeiidae and Sematuridae) to 205 Mbp. In the case of the latter, as a reference genome is available (Hill et al., 2019), we measured that, on average, 81.4% of the sequencing data belongs to the targeted butterfly's genome, with an average coverage of 15.6X, while the expected coverage for fresh material was 20X. These results are encouraging by showing contaminations are not as such present as expected, but also, confirming museum specimens still hold genomic data.

We demonstrated the value of museomics data by reconstructing phylogenetic relationships based on 308 genes and using it for population genetic analysis. First, regarding our phylogenomics results, we confirmed, with strong support, that Sematuridae is the sister clade of Pseudobistonidae + Epicopeiidae. We also showed the monophyly of Epicopeiidae (*i.e.* being a consistent group where all the descendants derive from a common ancestor). Within this family, despite the position of *Schistomitra*, our results are congruent with the previous hypothesis based on morphological characters (J. Minet, 2002). We found out that *Epicopeia* and *Nossa* are paraphyletic with respect to each other. Therefore, our results lead the way for further works on these two genera. Secondly, on *Pieris napi*, we observed genetic differences between populations located in the Arctic Circle and populations in the rest of Sweden,

confirming previous suspicions of the Arctic population to be a subspecies of the former.

5.2 What else can be achieved with museomics?

One should keep in mind that the sequencing technologies are evolving and getting better quickly. We saw in the last ten years the emergence of this kind of molecular approaches to museum specimens. While the protocols are being optimized, their popularization begins and the scientific community sees the opportunity. For instance, a couple of years ago, despite the desire to, it was believed that getting suitable quality DNA from formalin-fixed museum specimens was impossible. However, recent improvements in molecular labs have made this wish achievable (Totoiu et al., 2020). Therefore, what was inconceivable some decades ago is now becoming a reachable goal.

Beyond the technology itself, museomics also open the door for answering plenty of interesting scientific questions. The potential for such approaches can have a considerable impact on phylogenomics and population genetic studies, as we saw here, but on more ecological perspectives as well, particularly by accessing data from an era before climate changes (Waldvogel et al., 2020). The limits are our imagination of thinking of the biological question we want to investigate, and our creativity to develop methods that will help us answer these questions.

Glossary

Genomics

Genomics is a relatively new scientific field, and as such, it comes with a whole new vocabulary. Therefore, the meanings behind the words can seem ambiguous and challenging to clarify for researchers not directly involved in this subject. To ensure readers correctly understand the concepts that I will use in this thesis, I chose to make a glossary. You will note that it does not follow the alphabetical order. I made this choice to keep a thread of understanding, as some concepts are referring to others.

Genome – *the entire genetic material of an organism, which includes both the genes (the coding region) and the non-coding DNA, along with mitochondrial / chloroplast DNA.*

Single Nucleotide Polymorphism (SNP) – *a substitution of a single nucleotide that occurs at a specific position in the genome.*

Next-Generation Sequencing (NGS) or High-throughput sequencing (HTS) – *sometimes also called second-generation methods. These methods are inherently different from the previous sequencing methods, like Sanger sequencing. They fragment the genome in small pieces, randomly sample for a fragment, and sequence it. By parallelizing the process, NGS methods allow the entire genome to be sequenced at once. NGS only became popular at the beginning of the 2000s.*

Library – *a collection of DNA fragments that together represent the entire genome of an organism. This set is made by cloning small fragments of DNA. Please note, we also use libraries generally no matter the application. Hence, libraries represent subsets of the genome or the transcriptome.*

Adapters – *are short sequences that ligate to the ends of other DNA or RNA molecules. They are prevalent in NGS reads, as they are used to help sequence the reads.*

Indexing – *when sequencing, we usually pool samples together. To be able to differentiate each sample, we assign them unique indexes that allow us to set them apart.*

Read – *a short fragment of DNA, resulting from shotgun sequencing of genomic DNA.*

Contig – *a set of overlapping reads that together represent a consensus region of DNA.*

Scaffold – *when creating a draft genome, individual reads of DNA are first assembled into contigs, which have gaps between them. The next step is to bridge the gaps between these contigs to create a scaffold.*

Single-End (SE) – *single-end sequencing allows us to sequence only one end of a fragment of DNA.*

Paired-Ends (PE) – *paired-end sequencing sequences both ends of a fragment and generates high-quality, alignable sequence data. This produces twice the number of reads for the same time and effort in library preparation. It also provides more accurate read alignment and detects both insertion and deletion variants, which is not possible with single-end data.*

Cleaning / Processing – *the raw reads received from sequencing have some regions that could be problematic for the rest of the downstream analysis. For example, raw data still contain their adapters. Some of the frequent problems are low quality, low complexity, contaminants, duplicates, error correction, and adapters. Therefore, raw reads need to be cleaned and processed to avoid these problems.*

Trimming – *this process deals with low-quality nucleotides by removing them, trying to eliminate only low-quality regions.*

K-mer – *all the possible nucleotides' subsequences of length k, contained in a sequence.*

Sequence assembly – *aligning and merging reads from DNA sequence to reconstruct the original sequence. DNA sequencing technologies cannot read the whole genome in one go but instead scan small pieces of between 20 and 30 000 bases, depending on the technology used.*

Mapping – *the process of aligning short reads to a reference sequence, whether the reference is a complete genome, transcriptome, or de novo assembly.*

Reference assembly – *assembling reads against an existing backbone sequence (reference), building a sequence that is similar but not necessarily identical to the reference sequence.*

***De novo* assembly** – *assembling reads to create a full-length (sometimes novel) sequence, without using a template. To assemble a genome is computationally more costly than to do a reference assembly.*

BLAST (basic local alignment search tool) – *an algorithm for comparing primary biological sequence information, such as the amino-acid sequences of proteins or the nucleotides of DNA or RNA sequences. A BLAST search allows researchers to compare a query sequence with a database of sequences and identify library sequences that resemble the query sequence above a certain threshold.*

Transcriptome – *a set of all the RNA molecules in one type cell. Transcriptome might also refer to all of the coding regions of the genome.*

Phylogenomics

Monophyletic – *a consistent group where all the descendants derive from a common ancestor (or ancestral population). Usually, monophyletic groups are defined by specific characteristic (morphological or genetic) that are shared by all the organisms present in the group.*

Paraphyletic – *conversely, a paraphyletic clade regroups all descendants of a common ancestor excluding some groups of descendants.*

Polyphyletic – *represents a set of organisms that have been grouped together while they do not share a common ancestor.*

Acknowledgements

This project has received funding from the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No. 6422141

I would like to thank the Zoological Research Museum Alexander Koenig (Bonn, Germany), the Natural History Museum of Denmark (Copenhagen, Denmark), National Museum of Nature and Science (Tokyo, Japan), the Naturhitoriska riksmuseet (Stockholm, Sweden), and the Naturalis Biodiversity Center (Leiden, Netherland) for allowing us to use their specimens for this project. Special appreciations to Marianne Espeland, Ole Karsholt, Utsugi Jinbo, Tobias Malm and Rob de Vos, for their help and advice.

Finally, I would like to thank all the wonderful people that I had the opportunity to meet during this Ph.D.

Niklas, thank you for this incredible opportunity, and for making me welcome since the beginning. I was incredibly lucky to have you as my supervisor! Because of you I discovered the fantastic world of Lepidoptera. Thank you for your guidance, and always being an inspiring researcher and person. I will always remember our exciting discussions on various topics, and the support you gave me. Thank you for your patience and reactivity, especially these last weeks! Also, thank you for introducing me to The Dragonriders of Pern!

Victoria, the Obi-Wan to my Yoda, thank you from the bottom of my heart for your crucial inputs, your good luck to balance my inner Murphy's law and for your friendship.

Hamid and **Leidys**, for the laughs, the support, our endless discussions, the parties and having each other backs! We, truly, are Niklas' Angels.

The entire Systematic Biology group notably **Jadranka**, **Nicolas**, **Anne**, **Andrea C.**, **Tobias**, **Irenka**, **Andrea S.** and all the others that came and went. Thank you all for our exchange of ideas, your input and advice.

The Lund Biological museum and especially people from the entomological collections (**Christoffer**, **Ellen**, **Rune**, **Karen** and **Christer**) for their expertise, kindness, help, and also for introducing me to Swedish fika.

To all members of the BIG4, I am glad to have met you. To the supervisors (**Alexey**, **Martin**, **Ximo**, **Frederik**, **Nesrine**, etc.), thank you for your advice, training and experience. And, to all my fellow students (**Janina**, **Trevor**, **Daniel**, **Si-Pei**, **Emmanuel**, **Matthias**, **Erik**, **Miroslav**, **Viktor**, **Josh**, **Igor**, **Ashish** and **Anne-Sarah**) for the pleasure to have shared this adventure with you and our engaging discussions, I also hope our paths will cross again!

Mikaël, thank you for being supportive, even if we did not interact much I knew I could always count on you. **Ola**, **Honor**, **Nils**, **Pål**, **Magne**, **Øystein** for our exchanges.

Jane Jönsson for her guidance in the lab and **Tomas Johansson** for his help with BioAnalyzer, PicoGreen and sequencing.

Marianne, thank you again for the fantastic experience that I had in Bonn. Thanks to your mentoring I learn a lot in a short amount of time. Thank you to **Sandra** for her help in the lab, **Christoph** for our interesting exchanges, and **Jan-Philip** for our coffee breaks.

Thank you to **Love Dalén**, for hosting me. To **Nicolas Dussex** and **Johanna von Seth** for their guidance in the lab. Thank you also to **Allison**, **Moos**, and all the people that I met in Stockholm during this time.

Simeão e **Eduardo**, nós compartilhamos o escritório por um tempo, mas a amizade é para

My dear friends and colleagues **Chon**, **Johanna**, **Romain**, **Annick**, **Maria**, **Tristan**, **Matthias**, **Dafne**, **Johan**, **Björn**, **Gróa**, **Katja**, **Ainara**, **Martin**, **Pernilla**, **Hélène**, **Paul**, **Daniel**, **Suvi**, etc. for your help, exchanges and our memorable fika. Thank as well to all my other colleagues at Lund University, I don't forget you. And to Pub Einar!

Enfin, mille mercis à famille et amis en France. **Mes parents et ma sestra**, pour leur soutien inconditionnel ! Mes loutres de cœur (**Seb**, **Flal**, **Mathieu**, **Céline**, **Sarah**, **Randy**, **Cécile**), **Léanie** (pour la couverture et tout le reste), Hodor le groupe du M2 qui tient malgré les années (**Amandine**, **PL**, **Nathan**, **Gaëtan**, **Kévin**, **Cindy**, **Titi**, **Marine**), et enfin le trio de choc (**Marine**, **Clémence** et **Ambre**), vous tous qui m'avez soutenu d'une façon ou d'un autre (et surtout supporté) pendant 4,5 ans.

Last but not least, this thesis could not have been done without Sci-Hub, R, Zotero, Grammarly, Slack and Gimp.

And Jonas for his incredible help getting this thesis in a printable format on time!

References

Abadi, S., Azouri, D., Pupko, T., & Mayrose, I. (2019). Model selection may not be a mandatory step for phylogeny reconstruction. *Nature Communications*, *10*(1), 1–11. https://doi.org/10.1038/s41467-019-08822-w

Ackery, P. R., Testa, J. M., Ready, P. D., Doyle, A. M., & Pinniger, D. B. (2004). Effects of High Temperature Pest Eradication on DNA in Entomological Collections. *Studies in Conservation*, *49*(1), 35–40. https://doi.org/10.1179/sic.2004.49.1.35

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*(6), 716–723. https://doi.org/10.1109/TAC.1974.1100705

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

Andolfatto, P., Davison, D., Erezyilmaz, D., Hu, T. T., Mast, J., Sunayama-Morita, T., & Stern, D. L. (2011). Multiplexed shotgun genotyping for rapid and efficient genetic mapping. *Genome Research*, *21*(4), 610–617. https://doi.org/10.1101/gr.115402.110

Anmarkrud, J. A., & Lifjeld, J. T. (2017). Complete mitochondrial genomes of eleven extinct or possibly extinct bird species. *Molecular Ecology Resources*, *17*(2), 334–341. https://doi.org/10.1111/1755-0998.12600

Austin, J. J., Smith, A. B., & Thomas, R. H. (1997). Palaeontology in a molecular world: The search for authentic ancient DNA. *Trends in Ecology & Evolution*, *12*(8), 303–306. https://doi.org/10.1016/S0169-5347(97)01102-6

Baird, N. A., Etter, P. D., Atwood, T. S., Currey, M. C., Shiver, A. L., Lewis, Z. A., Selker, E. U., Cresko, W. A., & Johnson, E. A. (2008). Rapid SNP Discovery and Genetic Mapping Using Sequenced RAD Markers. *PLOS ONE*, *3*(10), e3376. https://doi.org/10.1371/journal.pone.0003376

Ballare, K. M., Pope, N. S., Castilla, A. R., Cusser, S., Metz, R. P., & Jha, S. (2019). Utilizing field collected insects for next generation sequencing: Effects of sampling, storage, and DNA extraction methods. *Ecology and Evolution*, *9*(24), 13690–13705. https://doi.org/10.1002/ece3.5756

Barnett, R., & Larson, G. (2012). A Phenol–Chloroform Protocol for Extracting DNA from Ancient Samples. In B. Shapiro & M. Hofreiter (Eds.), *Ancient DNA: Methods and Protocols* (pp. 13–19). Humana Press. https://doi.org/10.1007/978-1-61779-516-9_2

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

Bejerano, G., Pheasant, M., Makunin, I., Stephen, S., Kent, W. J., Mattick, J. S., & Haussler, D. (2004). Ultraconserved Elements in the Human Genome. *Science*, *304*(5675), 1321–1325. https://doi.org/10.1126/science.1098119

Boisduval, J. A., & Guénée, A. (1836). *Histoire naturelle des insectes. Spécies général des lépidoptères.* (pp. 1–460). Roret,. https://doi.org/10.5962/bhl.title.9194

Bradley, R. D., Bradley, L. C., Garner, H. J., & Baker, R. J. (2014). Assessing the Value of Natural History Collections and Addressing Issues Regarding Long-Term Growth and Care. *BioScience*, *64*(12), 1150–1158. https://doi.org/10.1093/biosci/biu166

Brandley, M. C., Bragg, J. G., Singhal, S., Chapple, D. G., Jennings, C. K., Lemmon, A. R., Lemmon, E. M., Thompson, M. B., & Moritz, C. (2015). Evaluating the performance of anchored hybrid enrichment at the tips of the tree of life: A phylogenetic analysis of Australian Eugongylus group scincid lizards. *BMC Evolutionary Biology*, *15*. https://doi.org/10.1186/s12862-015-0318-0

Breinholt, J. W., Earl, C., Lemmon, A. R., Lemmon, E. M., Xiao, L., & Kawahara, A. Y. (2018). Resolving Relationships among the Megadiverse Butterflies and Moths with a Novel Pipeline for Anchored Phylogenomics. *Systematic Biology*, *67*(1), 78–93. https://doi.org/10.1093/sysbio/syx048

Buck, M., & Hamilton, C. (2011). The Nagoya Protocol on Access to Genetic Resources and the Fair and Equitable Sharing of Benefits Arising from their Utilization to the Convention on Biological Diversity. *Review of European Community & International Environmental Law*, *20*(1), 47–61. https://doi.org/10.1111/j.1467-9388.2011.00703.x

Buckley, T. R. (2002). Model Misspecification and Probabilistic Tests of Topology: Evidence from Empirical Data Sets. *Systematic Biology*, *51*(3), 509–523. https://doi.org/10.1080/10635150290069922

Burrell, A. S., Disotell, T. R., & Bergey, C. M. (2015). The use of museum specimens with high-throughput DNA sequencers. *Journal of Human Evolution*, *79*, 35–44. https://doi.org/10.1016/j.jhevol.2014.10.015

Bushell, M., Stoneley, M., Sarnow, P., & Willis, A. E. (2004). Translation inhibition during the induction of apoptosis: RNA or protein degradation? *Biochemical Society Transactions*, *32*(4), 606–610. https://doi.org/10.1042/BST0320606

Butchart, S. H. M., Walpole, M., Collen, B., van Strien, A., Scharlemann, J. P. W., Almond, R. E. A., Baillie, J. E. M., Bomhard, B., Brown, C., Bruno, J., Carpenter, K. E., Carr, G. M., Chanson, J., Chenery, A. M., Csirke, J., Davidson, N. C., Dentener, F., Foster, M., Galli, A., … Watson, R. (2010). Global Biodiversity: Indicators of Recent Declines. *Science*, *328*(5982), 1164–1168. https://doi.org/10.1126/science.1187512

Cano, R. J., & Borucki, M. K. (1995). Revival and identification of bacterial spores in 25- to 40-million-year-old Dominican amber. *Science*, *268*(5213), 1060–1064. https://doi.org/10.1126/science.7538699

CBD. (2012). Decisions adopted by the Conference of the Parties to the CBD. *Global Taxonomy Initiative*, *XI/29*(Capacity-building Strategy for the Global Taxonomy Initiative).

Chapman, A. D. (2005). *Uses of primary species-occurrence data*.

Chapus, C., Dufraigne, C., Edwards, S., Giron, A., Fertil, B., & Deschavanne, P. (2005). Exploration of phylogenetic data using a global sequence analysis method. *BMC Evolutionary Biology*, *5*(1), 63. https://doi.org/10.1186/1471-2148-5-63

Chernomor, O., von Haeseler, A., & Minh, B. Q. (2016). Terrace Aware Data Structure for Phylogenomic Inference from Supermatrices. *Systematic Biology*, *65*(6), 997–1008. https://doi.org/10.1093/sysbio/syw037

Cho, S., Zwick, A., Regier, J. C., Mitter, C., Cummings, M. P., Yao, J., Du, Z., Zhao, H., Kawahara, A. Y., Weller, S., Davis, D. R., Baixeras, J., Brown, J. W., & Parr, C. (2011). Can Deliberately Incomplete Gene Sample Augmentation Improve a Phylogeny Estimate for the Advanced Moths and Butterflies (Hexapoda: Lepidoptera)? *Systematic Biology*, *60*(6), 782–796. https://doi.org/10.1093/sysbio/syr079

Classen, C. (2007). Museum Manners: The Sensory Life of the Early Museum. *Journal of Social History*, *40*(4), 895–914. https://doi.org/10.1353/jsh.2007.0089

Cloutier, A., Sackton, T. B., Grayson, P., Clamp, M., Baker, A. J., & Edwards, S. V. (2019). Whole-Genome Analyses Resolve the Phylogeny of Flightless Birds (Palaeognathae) in the Presence of an Empirical Anomaly Zone. *Systematic Biology*, *68*(6), 937–955. https://doi.org/10.1093/sysbio/syz019

Cloutier, A., Sackton, T. B., Grayson, P., Edwards, S. V., & Baker, A. J. (2018). First nuclear genome assembly of an extinct moa species, the little bush moa (Anomalopteryx didiformis). *BioRxiv*, 262816. https://doi.org/10.1101/262816

Cock, M. J. W. (2017). The Corkscrew Moths (Lepidoptera, Geometroidea, Sematuridae) of Trinidad and Tobago. *Tropical Lepidoptera Research*. https://journals.flvc.org/troplep/article/view/93272

Cock, M. J. W., & Lamas, G. (2011). Case 3531 Sematura Dalman, 1825 (Insecta, Lepidoptera, sematuridae): Proposed precedence over Mania Hübner, 1821. *The Bulletin of Zoological Nomenclature*, *68*(3), 184–189. https://doi.org/10.21805/bzn.v68i3.a11

Collins, J. A., Schandl, C. A., Young, K. K., Vesely, J., & Willingham, M. C. (1997). Major DNA Fragmentation Is a Late Event in Apoptosis. *Journal of Histochemistry & Cytochemistry*, *45*(7), 923–934. https://doi.org/10.1177/002215549704500702

Cooper, A., Lalueza-Fox, C., Anderson, S., Rambaut, A., Austin, J., & Ward, R. (2001). Complete mitochondrial genome sequences of two extinct moas clarify ratite evolution. *Nature*, *409*(6821), 704–707. https://doi.org/10.1038/35055536

Cooper, A., Mourer-Chauviré, C., Chambers, G. K., Haeseler, A. von, Wilson, A. C., & Pääbo, S. (1992). Independent origins of New Zealand moas and kiwis. *Proceedings of the National Academy of Sciences*, *89*(18), 8741–8744. https://doi.org/10.1073/pnas.89.18.8741

Cooper, A., & Poinar, H. N. (2000). Ancient DNA: Do It Right or Not at All. *Science*, *289*(5482), 1139–1139. https://doi.org/10.1126/science.289.5482.1139b

Cooper, E. D. (2014). Overly simplistic substitution models obscure green plant phylogeny. *Trends in Plant Science*, *19*(9), 576–582. https://doi.org/10.1016/j.tplants.2014.06.006

Cressey, D. (2014). Biopiracy ban stirs red-tape fears. *Nature*, *514*(7520), 14–15. https://doi.org/10.1038/514014a

Cruz-Dávalos, D. I., Llamas, B., Gaunitz, C., Fages, A., Gamba, C., Soubrier, J., Librado, P., Seguin-Orlando, A., Pruvost, M., Alfarhan, A. H., Alquraishi, S. A., Al-Rasheid, K. A. S., Scheu, A., Beneke, N., Ludwig, A., Cooper, A., Willerslev, E., & Orlando, L. (2017). Experimental conditions improving in-solution target enrichment for ancient DNA. *Molecular Ecology Resources*, *17*(3), 508–522. https://doi.org/10.1111/1755-0998.12595

De Donato, M., Peters, S. O., Mitchell, S. E., Hussain, T., & Imumorin, I. G. (2013). Genotyping-by-Sequencing (GBS): A Novel, Efficient and Cost-Effective Genotyping Method for Cattle Using Next-Generation Sequencing. *PLoS ONE*, *8*(5). https://doi.org/10.1371/journal.pone.0062137

DeSalle, R., Gatesy, J., Wheeler, W., & Grimaldi, D. (1992). DNA sequences from a fossil termite in Oligo-Miocene amber and their phylogenetic implications. *Science*, *257*(5078), 1933–1936. https://doi.org/10.1126/science.1411508

Duckworth, W. D., Genoways, H. H., & Rose, C. L. (1993). Preserving Natural Science Collections: Chronicle of Our Environmental Heritage. *Mammology Papers: University of Nebraska State Museum*, *271*, 153.

Egan, A. N., Schlueter, J., & Spooner, D. M. (2012). Applications of next-generation sequencing in plant biology. *American Journal of Botany*, *99*(2), 175–185. https://doi.org/10.3732/ajb.1200020

Eisen, J. A. (1998). Phylogenomics: Improving Functional Predictions for Uncharacterized Genes by Evolutionary Analysis. *Genome Research*, *8*(3), 163–167. https://doi.org/10.1101/gr.8.3.163

Ekblom, R., & Wolf, J. B. W. (2014). A field guide to whole-genome sequencing, assembly and annotation. *Evolutionary Applications*, *7*(9), 1026–1042. https://doi.org/10.1111/eva.12178

Elshire, R. J., Glaubitz, J. C., Sun, Q., Poland, J. A., Kawamoto, K., Buckler, E. S., & Mitchell, S. E. (2011). A Robust, Simple Genotyping-by-Sequencing (GBS) Approach for High Diversity Species. *PLOS ONE*, *6*(5), e19379. https://doi.org/10.1371/journal.pone.0019379

Espeland, M., Aagaard, K., Balstad, T., & Hindar, K. (2007). Ecomorphological and genetic divergence between lowland and montane forms of the Pieris napi species complex (Pieridae, Lepidoptera). *Biological Journal of the Linnean Society*, *92*(4), 727–745. https://doi.org/10.1111/j.1095-8312.2007.00873.x

Espeland, M., Breinholt, J. W., Barbosa, E. P., Casagrande, M. M., Huertas, B., Lamas, G., Marín, M. A., Mielke, O. H. H., Miller, J. Y., Nakahara, S., Tan, D., Warren, A. D., Zacca, T., Kawahara, A. Y., Freitas, A. V. L., & Willmott, K. R. (2019). Four hundred shades of brown: Higher level phylogeny of the problematic Euptychiina (Lepidoptera, Nymphalidae, Satyrinae) based on hybrid enrichment data. *Molecular Phylogenetics and Evolution*, *131*, 116–124. https://doi.org/10.1016/j.ympev.2018.10.039

Espeland, M., Breinholt, J., Willmott, K. R., Warren, A. D., Vila, R., Toussaint, E. F. A., Maunsell, S. C., Aduse-Poku, K., Talavera, G., Eastwood, R., Jarzyna, M. A., Guralnick, R., Lohman, D. J., Pierce, N. E., & Kawahara, A. Y. (2018). A Comprehensive and Dated Phylogenomic Analysis of Butterflies. *Current Biology*, *28*(5), 770-778.e5. https://doi.org/10.1016/j.cub.2018.01.061

Ewing, B., & Green, P. (1998). Base-Calling of Automated Sequencer Traces Using Phred. II. Error Probabilities. *Genome Research*, *8*(3), 186–194. https://doi.org/10.1101/gr.8.3.186

Ewing, B., Hillier, L., Wendl, M. C., & Green, P. (1998). Base-Calling of Automated Sequencer Traces UsingPhred. I. Accuracy Assessment. *Genome Research*, *8*(3), 175–185. https://doi.org/10.1101/gr.8.3.175

Fabre, P.-H., Vilstrup, J. T., Raghavan, M., Der Sarkissian, C., Willerslev, E., Douzery, E. J. P., & Orlando, L. (2014). Rodents of the Caribbean: Origin and diversification of hutias unravelled by next-generation museomics. *Biology Letters*, *10*(7), 20140266–20140266. https://doi.org/10.1098/rsbl.2014.0266

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

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

Faircloth, B. C., Sorenson, L., Santini, F., & Alfaro, M. E. (2013). A Phylogenomic Perspective on the Radiation of Ray-Finned Fishes Based upon Targeted Sequencing of Ultraconserved Elements (UCEs). *PLOS ONE*, *8*(6), e65923. https://doi.org/10.1371/journal.pone.0065923

Feng, Q., ZhenNing, C., GuiGong, G., & SiWei, L. (2011). Study on five methods for extracting DNA from dried butterfly specimen. *Agricultural Science & Technology - Hunan*, *12*(8), 1121–1124.

Fulton, T. L. (2012). Setting Up an Ancient DNA Laboratory. In B. Shapiro & M. Hofreiter (Eds.), *Ancient DNA: Methods and Protocols* (pp. 1–11). Humana Press. https://doi.org/10.1007/978-1-61779-516-9_1

Gasc, C., Peyretaillade, E., & Peyret, P. (2016). Sequence capture by hybridization to explore modern and ancient genomic diversity in model and nonmodel organisms. *Nucleic Acids Research*, *44*(10), 4504–4518. https://doi.org/10.1093/nar/gkw309

Golenberg, E. M., Giannasi, D. E., Clegg, M. T., Smiley, C. J., Durbin, M., Henderson, D., & Zurawski, G. (1990). Chloroplast DNA sequence from a Miocene Magnolia species. *Nature*, *344*(6267), 656–658. https://doi.org/10.1038/344656a0

Goodman, S. N. (1999). Toward Evidence-Based Medical Statistics. 2: The Bayes Factor. *Annals of Internal Medicine*, *130*(12), 1005. https://doi.org/10.7326/0003-4819-130-12-199906150-00019

Guschanski, K., Krause, J., Sawyer, S., Valente, L. M., Bailey, S., Finstermeier, K., Sabin, R., Gilissen, E., Sonet, G., Nagy, Z. T., Lenglet, G., Mayer, F., & Savolainen, V. (2013). Next-Generation Museomics Disentangles One of the Largest Primate Radiations. *Systematic Biology*, *62*(4), 539–554. https://doi.org/10.1093/sysbio/syt018

Gutfleisch, B., & Menzhausen, J. (1989). "HOW A KUNSTKAMMER SHOULD BE FORMED": Gabriel Kaltemarckt's Advice to Christian I of Saxony on the Formation of an Art Collection, 1587. *Journal of the History of Collections*, *1*(1), 3–32. https://doi.org/10.1093/jhc/1.1.3

Hallmann, C. A., Sorg, M., Jongejans, E., Siepel, H., Hofland, N., Schwan, H., Stenmans, W., Müller, A., Sumser, H., Hörren, T., Goulson, D., & de Kroon, H. (2017). More than 75 percent decline over 27 years in total flying insect biomass in protected areas. *PLOS ONE*, *12*(10), e0185809. https://doi.org/10.1371/journal.pone.0185809

Harmon, L. (2018). *Phylogenetic comparative methods: Learning from trees*.

Hawkins, M. T. R., Hofman, C. A., Callicrate, T., McDonough, M. M., Tsuchiya, M. T. N., Gutiérrez, E. E., Helgen, K. M., & Maldonado, J. E. (2016). In-solution hybridization for mammalian mitogenome enrichment: Pros, cons and challenges associated with multiplexing degraded DNA. *Molecular Ecology Resources*, *16*(5), 1173–1188. https://doi.org/10.1111/1755-0998.12448

Hedges, S. B., Schweitzer, M. H., Henikoff, S., Allard, M. W., Young, D., Huyen, Y., Zischler, H., Höss, M., Handt, O., von Haeseler, A., van der Kuyl, A. C., Goudsmit, J., Pääbo, S., & Woodward, S. R. (1995). Detecting Dinosaur DNA. *Science*, *268*(5214), 1191–1194.

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), 260. https://doi.org/10.1186/s12862-015-0520-0

Higuchi, R., Bowman, B., Freiberger, M., Ryder, O. A., & Wilson, A. C. (1984). DNA sequences from the quagga, an extinct member of the horse family. *Nature*, *312*(5991), 282–284. https://doi.org/10.1038/312282a0

Hill, J., Rastas, P., Hornett, E. A., Neethiraj, R., Clark, N., Morehouse, N., Celorio-Mancera, M. de la P., Cols, J. C., Dircksen, H., Meslin, C., Keehnen, N., Pruisscher, P., Sikkink, K., Vives, M., Vogel, H., Wiklund, C., Woronik, A., Boggs, C. L., Nylin, S., & Wheat, C. W. (2019). Unprecedented reorganization of holocentric chromosomes provides insights into the enigma of lepidopteran chromosome evolution. *Science Advances*, *5*(6), eaau3648. https://doi.org/10.1126/sciadv.aau3648

Hofreiter, M., Jaenicke, V., Serre, D., Haeseler, A. von, & Pääbo, S. (2001). DNA sequences from multiple amplifications reveal artifacts induced by cytosine deamination in ancient DNA. *Nucleic Acids Research*, *29*(23), 4793–4799. https://doi.org/10.1093/nar/29.23.4793

Hofreiter, M., Rabeder, G., Jaenicke-Després, V., Withalm, G., Nagel, D., Paunovic, M., Jambrěsić, G., & Pääbo, S. (2004). Evidence for Reproductive Isolation between Cave Bear Populations. *Current Biology*, *14*(1), 40–43. https://doi.org/10.1016/j.cub.2003.12.035

Holloway, J. D., Kibby, G., & Peggie, D. (2001). *The Families of Malesian Moths and Butterflies*. BRILL.

Höss, M., Handt, O., & Pääbo, S. (1994). Recreating the Past by PCR. In K. B. Mullis, F. Ferré, & R. A. Gibbs (Eds.), *The Polymerase Chain Reaction* (pp. 257–264). Birkhäuser Boston. https://doi.org/10.1007/978-1-4612-0257-8_22

Huang, X., Wei, X., Sang, T., Zhao, Q., Feng, Q., Zhao, Y., Li, C., Zhu, C., Lu, T., Zhang, Z., Li, M., Fan, D., Guo, Y., Wang, A., Wang, L., Deng, L., Li, W., Lu, Y., Weng, Q., … Han, B. (2010). Genome-wide association studies of 14 agronomic traits in rice landraces. *Nature Genetics*, *42*(11), 961–967. https://doi.org/10.1038/ng.695

Huelsenbeck, J. P., Joyce, P., Lakner, C., & Ronquist, F. (2008). Bayesian analysis of amino acid substitution models. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *363*(1512), 3941–3953. https://doi.org/10.1098/rstb.2008.0175

Jones, M. R., & Good, J. M. (2016). Targeted capture in evolutionary and ecological genomics. *Molecular Ecology*, *25*(1), 185–202. https://doi.org/10.1111/mec.13304

Kadlec, M., Bellstedt, D. U., Le Maitre, N. C., & Pirie, M. D. (2017). Targeted NGS for species level phylogenomics: "Made to measure" or "one size fits all"? *PeerJ*, *5*. https://doi.org/10.7717/peerj.3569

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

Kanda, K., Pflug, J. M., Sproul, J. S., Dasenko, M. A., & Maddison, D. R. (2015). Successful Recovery of Nuclear Protein-Coding Genes from Small Insects in Museums Using

Illumina Sequencing. *PLOS ONE*, *10*(12), e0143929. https://doi.org/10.1371/journal.pone.0143929

Kawahara, A. Y., & Breinholt, J. W. (2014). Phylogenomics provides strong evidence for relationships of butterflies and moths. *Proc. R. Soc. B*, *281*(1788), 20140970. https://doi.org/10.1098/rspb.2014.0970

Kawahara, A. Y., Plotkin, D., Espeland, M., Meusemann, K., Toussaint, E. F. A., Donath, A., Gimnich, F., Frandsen, P. B., Zwick, A., Reis, M. dos, Barber, J. R., Peters, R. S., Liu, S., Zhou, X., Mayer, C., Podsiadlowski, L., Storer, C., Yack, J. E., Misof, B., & Breinholt, J. W. (2019). Phylogenomics reveals the evolutionary timing and pattern of butterflies and moths. *Proceedings of the National Academy of Sciences*, *116*(45), 22657–22663. https://doi.org/10.1073/pnas.1907847116

Knapp, M., Lalueza-Fox, C., & Hofreiter, M. (2015). Re-inventing ancient human DNA. *Investigative Genetics*, *6*. https://doi.org/10.1186/s13323-015-0020-4

Laithwaite, E. R. (Eric R., & Whalley, P. E. S. (1975). *Dictionary of butterflies and moths in color*. Michael Joseph. http://agris.fao.org/agris-search/search.do?recordID=US201300521786

Le Bihan, Y.-V., Angeles Izquierdo, M., Coste, F., Aller, P., Culard, F., Gehrke, T. H., Essalhi, K., Carell, T., & Castaing, B. (2011). 5-Hydroxy-5-methylhydantoin DNA lesion, a molecular trap for DNA glycosylases. *Nucleic Acids Research*, *39*(14), 6277–6290. https://doi.org/10.1093/nar/gkr215

Leaché, A. D., Fujita, M. K., Minin, V. N., & Bouckaert, R. R. (2014). Species Delimitation using Genome-Wide SNP Data. *Systematic Biology*, *63*(4), 534–542. https://doi.org/10.1093/sysbio/syu018

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

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

Leonard, J. A., Wayne, R. K., & Cooper, A. (2000). Population genetics of Ice Age brown bears. *Proceedings of the National Academy of Sciences*, *97*(4), 1651. https://doi.org/10.1073/pnas.040453097

Lewis, Z. A., Shiver, A. L., Stiffler, N., Miller, M. R., Johnson, E. A., & Selker, E. U. (2007). High-Density Detection of Restriction-Site-Associated DNA Markers for Rapid Mapping of Mutated Loci in Neurospora. *Genetics*, *177*(2), 1163–1171. https://doi.org/10.1534/genetics.107.078147

Lindahl, T. (1993). Instability and decay of the primary structure of DNA. *Nature*, *362*(6422), 709–715. https://doi.org/10.1038/362709a0

Loreille, O., Orlando, L., Patou-Mathis, M., Philippe, M., Taberlet, P., & Hänni, C. (2001). Ancient DNA analysis reveals divergence of the cave bear, Ursus spelaeus , and brown

bear, Ursus arctos , lineages. *Current Biology*, *11*(3), 200–203. https://doi.org/10.1016/S0960-9822(01)00046-X

Malakasi, P., Bellot, S., Dee, R., & Grace, O. M. (2019). Museomics Clarifies the Classification of Aloidendron (Asphodelaceae), the Iconic African Tree Aloes. *Frontiers in Plant Science*, *10*, 1227. https://doi.org/10.3389/fpls.2019.01227

Mamanova, L., Coffey, A. J., Scott, C. E., Kozarewa, I., Turner, E. H., Kumar, A., Howard, E., Shendure, J., & Turner, D. J. (2010). Target-enrichment strategies for next-generation sequencing. *Nature Methods*, *7*(2), 111–118. https://doi.org/10.1038/nmeth.1419

Mardis, E. R. (2008). The impact of next-generation sequencing technology on genetics. *Trends in Genetics*, *24*(3), 133–141. https://doi.org/10.1016/j.tig.2007.12.007

Mardis, E. R. (2011). A decade's perspective on DNA sequencing technology. *Nature*, *470*(7333), 198–203. https://doi.org/10.1038/nature09796

Margulies, M., Egholm, M., Altman, W. E., Attiya, S., Bader, J. S., Bemben, L. A., Berka, J., Braverman, M. S., Chen, Y.-J., Chen, Z., Dewell, S. B., Du, L., Fierro, J. M., Gomes, X. V., Godwin, B. C., He, W., Helgesen, S., Ho, C. H., Irzyk, G. P., … Rothberg, J. M. (2005). Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, *437*(7057), 376–380. https://doi.org/10.1038/nature03959

McHardy, A. C., Martín, H. G., Tsirigos, A., Hugenholtz, P., & Rigoutsos, I. (2007). Accurate phylogenetic classification of variable-length DNA fragments. *Nature Methods*, *4*(1), 63–72. https://doi.org/10.1038/nmeth976

Metzker, M. L. (2010). Sequencing technologies—The next generation. *Nature Reviews Genetics*, *11*(1), 31–46. https://doi.org/10.1038/nrg2626

Miller, M. R., Dunham, J. P., Amores, A., Cresko, W. A., & Johnson, E. A. (2007). Rapid and cost-effective polymorphism identification and genotyping using restriction site associated DNA (RAD) markers. *Genome Research*, *17*(2), 240–248. https://doi.org/10.1101/gr.5681207

Minet, J. (2002). The Epicopeiidae: Phylogeny and a redefinition, with the description of new taxa (Lepidoptera: Drepanoidea). *Annales de La Société Entomologique de France (N.S.)*, *38*(4), 463–487. https://doi.org/10.1080/00379271.2002.10697355

Minet, Joël. (1983). Étude morphologique et phylogénétique des organes tympaniques des Pyraloidea. 1—Généralités et homologies (Lep. Glossata). *Annales de La Société Entomologique de France*, *19*(2), 175–207. Scopus.

Minet, Joël. (1986). Ébauche d'une classification moderne de l'ordre des Lepidoptères. *Alexanor*, *14*, 291–313.

Minet, Joël, & Scoble. (1999). The Drepanoid / Geometroid assemblage. In *Lepidoptera, butterflies and moths. Vol. 1. Evolution, systematics and biogeography.: Vol. Band/Volume 4 Arthropoda: Insecta* (Kristensen, N. P. (Ed.), pp. 301–320). Walter de Gruyter Inc.

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

Nagoya Protocol on access to genetic resources and the fair and equitable sharing of benefits arising from their utilization to the convention on biological diversity. (2010). *The Nagoya Protocol on Access and Benefit Sharing of Genetic Resources, Analysis and Implementation Options for Developing Countries , Nijar, Gurdial Singh. - Geneva : South Centre*.

Neumann, D., Borisenko, A. V., Coddington, J. A., Häuser, C. L., Butler, C. R., Casino, A., Vogel, J. C., Haszprunar, G., & Giere, P. (2018). Global biodiversity research tied up by juridical interpretations of access and benefit sharing. *Organisms Diversity & Evolution*, *18*(1), 1–12. https://doi.org/10.1007/s13127-017-0347-1

Ng, P. C., & Kirkness, E. F. (2010). Whole Genome Sequencing. In M. R. Barnes & G. Breen (Eds.), *Genetic Variation: Methods and Protocols* (pp. 215–226). Humana Press. https://doi.org/10.1007/978-1-60327-367-1_12

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

Noonan, J. P., Hofreiter, M., Smith, D., Priest, J. R., Rohland, N., Rabeder, G., Krause, J., Detter, J. C., Pääbo, S., & Rubin, E. M. (2005). Genomic Sequencing of Pleistocene Cave Bears. *Science*, *309*(5734), 597–599. https://doi.org/10.1126/science.1113485

O'Brien, S. J., & Stanyon, R. (1999). Ancestral primate viewed. *Nature*, *402*(6760), 365–366. https://doi.org/10.1038/46450

Ogden, R., Gharbi, K., Mugue, N., Martinsohn, J., Senn, H., Davey, J. W., Pourkazemi, M., McEwing, R., Eland, C., Vidotto, M., Sergeev, A., & Congiu, L. (2013). Sturgeon conservation genomics: SNP discovery and validation using RAD sequencing. *Molecular Ecology*, *22*(11), 3112–3123. https://doi.org/10.1111/mec.12234

Pääbo, S. (1989). Ancient DNA: Extraction, characterization, molecular cloning, and enzymatic amplification. *Proceedings of the National Academy of Sciences*, *86*(6), 1939–1943. https://doi.org/10.1073/pnas.86.6.1939

Pääbo, Svante. (1985). Molecular cloning of Ancient Egyptian mummy DNA. *Nature*, *314*(6012), 644–645. https://doi.org/10.1038/314644a0

Pääbo, Svante, Poinar, H., Serre, D., Jaenicke-Després, V., Hebler, J., Rohland, N., Kuch, M., Krause, J., Vigilant, L., & Hofreiter, M. (2004). Genetic Analyses from Ancient DNA. *Annual Review of Genetics*, *38*(1), 645–679. https://doi.org/10.1146/annurev.genet.37.110801.143214

Pääbo, Svante, & Wilson, A. C. (1988). Polymerase chain reaction reveals cloning artefacts. *Nature*, *334*(6181), 387–388. https://doi.org/10.1038/334387b0

Palkopoulou, E., Mallick, S., Skoglund, P., Enk, J., Rohland, N., Li, H., Omrak, A., Vartanyan, S., Poinar, H., Götherström, A., Reich, D., & Dalén, L. (2015). Complete Genomes Reveal Signatures of Demographic and Genetic Declines in the Woolly Mammoth. *Current Biology*, *25*(10), 1395–1400. https://doi.org/10.1016/j.cub.2015.04.007

Petersen, B. (1949). On the Evolution of Pieris napi L. *Evolution*, *3*(4), 269–278. JSTOR. https://doi.org/10.2307/2405714

Philippe, H., & Blanchette, M. (2007). Overview of the First Phylogenomics Conference. *BMC Evolutionary Biology*, *7*(Suppl 1), S1. https://doi.org/10.1186/1471-2148-7-S1-S1

Poinar, H. N., Cano, R. J., & Jr, G. O. P. (1993). DNA from an extinct plant. *Nature*, *363*(6431), 677. https://doi.org/10.1038/363677a0

Poinar, H. N., Hofreiter, M., Spaulding, W. G., Martin, P. S., Stankiewicz, B. A., Bland, H., Evershed, R. P., Possnert, G., & Pääbo, S. (1998). Molecular Coproscopy: Dung and Diet of the Extinct Ground Sloth Nothrotheriops shastensis. *Science*, *281*(5375), 402–406. https://doi.org/10.1126/science.281.5375.402

Porter, A. (1997). The Pieris napi /bryoniae hybrid zone at Pont de Nant, Switzerland: Broad overlap in the range of suitable host plants. *Ecological Entomology*, *22*(2), 189–196. https://doi.org/10.1046/j.1365-2311.1997.00054.x

Posada, D., & Crandall, K. A. (2001). Selecting the Best-Fit Model of Nucleotide Substitution. *Systematic Biology*, *50*(4), 580–601. https://doi.org/10.1080/10635150118469

Quicke, D. L. J., Lopez-Vaamonde, C., & Belshaw, R. (1999). Preservation of hymenopteran specimens for subsequent molecular and morphological study. *Zoologica Scripta*, *28*(1–2), 261–267. https://doi.org/10.1046/j.1463-6409.1999.00004.x

Rajaei, H., Greve, C., Letsch, H., Stüning, D., Wahlberg, N., Minet, J., & Misof, B. (2015). Advances in Geometroidea phylogeny, with characterization of a new family based on Pseudobiston pinratanai (Lepidoptera, Glossata). *Zoologica Scripta*, *44*(4), 418–436. https://doi.org/10.1111/zsc.12108

Regier, J. C., Mitter, C., Zwick, A., Bazinet, A. L., Cummings, M. P., Kawahara, A. Y., Sohn, J.-C., Zwickl, D. J., Cho, S., Davis, D. R., Baixeras, J., Brown, J., Parr, C., Weller, S., Lees, D. C., & Mitter, K. T. (2013). A Large-Scale, Higher-Level, Molecular Phylogenetic Study of the Insect Order Lepidoptera (Moths and Butterflies). *PLOS ONE*, *8*(3), e58568. https://doi.org/10.1371/journal.pone.0058568

Regier, J. C., Zwick, A., Cummings, M. P., Kawahara, A. Y., Cho, S., Weller, S., Roe, A., Baixeras, J., Brown, J. W., Parr, C., Davis, D. R., Epstein, M., Hallwachs, W., Hausmann, A., Janzen, D. H., Kitching, I. J., Solis, M. A., Yen, S.-H., Bazinet, A. L., & Mitter, C. (2009). Toward reconstructing the evolution of advanced moths and butterflies (Lepidoptera: Ditrysia): an initial molecular study. *BMC Evolutionary Biology*, *9*(1), 280. https://doi.org/10.1186/1471-2148-9-280

Rhoads, A., & Au, K. F. (2015). PacBio Sequencing and Its Applications. *Genomics, Proteomics & Bioinformatics*, *13*(5), 278–289. https://doi.org/10.1016/j.gpb.2015.08.002

Rohland, N., & Hofreiter, M. (2007). Comparison and optimization of ancient DNA extraction. *BioTechniques*, *42*(3), 343–352. https://doi.org/10.2144/000112383

Römpler, H., Dear, P. H., Krause, J., Meyer, M., Rohland, N., Schöneberg, T., Spriggs, H., Stiller, M., & Hofreiter, M. (2006). Multiplex amplification of ancient DNA. *Nature Protocols*, *1*(2), 720–728. https://doi.org/10.1038/nprot.2006.84

Rota, J., Malm, T., Chazot, N., Peña, C., & Wahlberg, N. (2018). A simple method for data partitioning based on relative evolutionary rates. *PeerJ*, *6*, e5498. https://doi.org/10.7717/peerj.5498

Rowe, K. C., Singhal, S., Macmanes, M. D., Ayroles, J. F., Morelli, T. L., Rubidge, E. M., Bi, K., & Moritz, C. C. (2011). Museum genomics: Low-cost and high-accuracy genetic data from historical specimens. *Molecular Ecology Resources*, *11*(6), 1082–1092. https://doi.org/10.1111/j.1755-0998.2011.03052.x

Saiki, R. K., Scharf, S., Faloona, F., Mullis, K. B., Horn, G. T., Erlich, H. A., & Arnheim, N. (1985). Enzymatic amplification of beta-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia. *Science (New York, N.Y.)*, *230*(4732), 1350–1354.

Sambrook, J., & Russell, D. W. (2006). Purification of Nucleic Acids by Extraction with Phenol:Chloroform. *Cold Spring Harbor Protocols*, *2006*(1), pdb.prot4455. https://doi.org/10.1101/pdb.prot4455

Särkinen, T., Staats, M., Richardson, J. E., Cowan, R. S., & Bakker, F. T. (2012). How to Open the Treasure Chest? Optimising DNA Extraction from Herbarium Specimens. *PLOS ONE*, *7*(8), e43808. https://doi.org/10.1371/journal.pone.0043808

Schindel, D., Bubela, T., Rosenthal, J., Castle, D., du Plessis, P., Bye, R., & Pmcw. (2015). The New Age of the Nagoya Protocol. *Nature Conservation*, *12*, 43–56. https://doi.org/10.3897/natureconservation.12.5412

Schwarz, G. (1978). Estimating the Dimension of a Model. *Annals of Statistics*, *6*(2), 461–464. https://doi.org/10.1214/aos/1176344136

Seguin-Orlando, A., Schubert, M., Clary, J., Stagegaard, J., Alberdi, M. T., Prado, J. L., Prieto, A., Willerslev, E., & Orlando, L. (2013). Ligation Bias in Illumina Next-Generation DNA Libraries: Implications for Sequencing Ancient Genomes. *PLOS ONE*, *8*(10), e78575. https://doi.org/10.1371/journal.pone.0078575

Seitz, A. (1913). *The Macrolepidoptera of the world: A systematic account of all the known Macrolepidoptera: Vol. v.6(1913) [Text]* (pp. 1–1336). Fritz Lehmann Verlag,. https://www.biodiversitylibrary.org/item/251725

Shapiro, B., & Hofreiter, M. (2012). *Ancient DNA: Methods and Protocols*. Humana Press.

Shendure, J., & Ji, H. (2008). Next-generation DNA sequencing. *Nature Biotechnology*, *26*(10), 1135–1145. https://doi.org/10.1038/nbt1486

Short, A. E. Z., Dikow, T., & Moreau, C. S. (2018). Entomological Collections in the Age of Big Data. *Annual Review of Entomology*, *63*(1), 513–530. https://doi.org/10.1146/annurev-ento-031616-035536

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

Silva, C., Besnard, G., Piot, A., Razanatsoa, J., Oliveira, R. P., & Vorontsova, M. S. (2017). Museomics resolve the systematics of an endangered grass lineage endemic to north-western Madagascar. *Annals of Botany*, *119*(3), 339–351. https://doi.org/10.1093/aob/mcw208

Smith, B. T., Harvey, M. G., Faircloth, B. C., Glenn, T. C., & Brumfield, R. T. (2014). Target Capture and Massively Parallel Sequencing of Ultraconserved Elements for Comparative Studies at Shallow Evolutionary Time Scales. *Systematic Biology*, *63*(1), 83–95. https://doi.org/10.1093/sysbio/syt061

Sproul, J. S., & Maddison, D. R. (2017). Sequencing historical specimens: Successful preparation of small specimens with low amounts of degraded DNA. *Molecular Ecology Resources*, *17*(6), 1183–1201. https://doi.org/10.1111/1755-0998.12660

Staats, M., Erkens, R. H. J., Vossenberg, B. van de, Wieringa, J. J., Kraaijeveld, K., Stielow, B., Geml, J., Richardson, J. E., & Bakker, F. T. (2013). Genomic Treasure Troves: Complete Genome Sequencing of Herbarium and Insect Museum Specimens. *PLOS ONE*, *8*(7), e69189. https://doi.org/10.1371/journal.pone.0069189

Su, N.-Y., & Scheffrahn, R. H. (1990). Efficacy of Sulfuryl Fluoride Against Four Beetle Pests of Museums (Coleoptera: Dermestidae, Anobiidae). *Journal of Economic Entomology*, *83*(3), 879–882. https://doi.org/10.1093/jee/83.3.879

Suarez, A. V., & Tsutsui, N. D. (2004). The Value of Museum Collections for Research and Society. *BioScience*, *54*(1), 66. https://doi.org/10.1641/0006-3568(2004)054[0066:TVOMCF]2.0.CO;2

Sugiura, N. (1978). Further analysts of the data by akaike' s information criterion and the finite corrections. *Communications in Statistics - Theory and Methods*, *7*(1), 13–26. https://doi.org/10.1080/03610927808827599

Swann, M. (2001). *Curiosities and Texts: The Culture of Collecting in Early Modern England*. University of Pennsylvania Press.

Thomas, M. P., Liu, X., Whangbo, J., McCrossan, G., Sanborn, K. B., Basar, E., Walch, M., & Lieberman, J. (2015). Apoptosis Triggers Specific, Rapid, and Global mRNA Decay with 3′ Uridylated Intermediates Degraded by DIS3L2. *Cell Reports*, *11*(7), 1079–1089. https://doi.org/10.1016/j.celrep.2015.04.026

Thomas, R. H., Schaffner, W., Wilson, A. C., & Pääbo, S. (1989). DNA phylogeny of the extinct marsupial wolf. *Nature*, *340*(6233), 465–467. https://doi.org/10.1038/340465a0

Tolman, T. (1997). *Butterflies of Britain and Europe*. Harpercollins Pub Limited.

Totoiu, C. A., Phillips, J. M., Reese, A. T., Majumdar, S., Girguis, P. R., Raston, C. L., & Weiss, G. A. (2020). Vortex fluidics-mediated DNA rescue from formalin-fixed museum specimens. *PLOS ONE*, *15*(1), e0225807. https://doi.org/10.1371/journal.pone.0225807

Toussaint, E. F. A., Breinholt, J. W., Earl, C., Warren, A. D., Brower, A. V. Z., Yago, M., Dexter, K. M., Espeland, M., Pierce, N. E., Lohman, D. J., & Kawahara, A. Y. (2018). Anchored phylogenomics illuminates the skipper butterfly tree of life. *BMC Evolutionary Biology*, *18*(1), 101. https://doi.org/10.1186/s12862-018-1216-z

Triant, D. A., Cinel, S. D., & Kawahara, A. Y. (2018). Lepidoptera genomes: Current knowledge, gaps and future directions. *Current Opinion in Insect Science*, *25*, 99–105. https://doi.org/10.1016/j.cois.2017.12.004

Tsangaras, K., Wales, N., Sicheritz-Pontén, T., Rasmussen, S., Michaux, J., Ishida, Y., Morand, S., Kampmann, M.-L., Gilbert, M. T. P., & Greenwood, A. D. (2014). Hybridization Capture Using Short PCR Products Enriches Small Genomes by Capturing Flanking Sequences (CapFlank). *PLOS ONE*, *9*(10), e109101. https://doi.org/10.1371/journal.pone.0109101

Waldvogel, A.-M., Feldmeyer, B., Rolshausen, G., Exposito-Alonso, M., Rellstab, C., Kofler, R., Mock, T., Schmid, K., Schmitt, I., Bataillon, T., Savolainen, O., Bergland, A., Flatt, T., Guillaume, F., & Pfenninger, M. (2020). Evolutionary genomics can improve prediction of species' responses to climate change. *Evolution Letters*, *4*(1), 4–18. https://doi.org/10.1002/evl3.154

Wang, H., Holloway, J. D., Wahlberg, N., Wang, M., & Nylin, S. (2019). Molecular phylogenetic and morphological studies on the systematic position of Heracula discivitta reveal a new subfamily of Pseudobistonidae (Lepidoptera: Geometroidea). *Systematic Entomology*, *44*(1), 211–225. https://doi.org/10.1111/syen.12326

Wang, Z., Gerstein, M., & Snyder, M. (2009). RNA-Seq: A revolutionary tool for transcriptomics. *Nature Reviews Genetics*, *10*(1), 57–63. https://doi.org/10.1038/nrg2484

Ward, J. A., Bhangoo, J., Fernández-Fernández, F., Moore, P., Swanson, J., Viola, R., Velasco, R., Bassil, N., Weber, C. A., & Sargent, D. J. (2013). Saturated linkage map construction in Rubus idaeususing genotyping by sequencing and genome-independent imputation. *BMC Genomics*, *14*(1), 2. https://doi.org/10.1186/1471-2164-14-2

Wei, C.-H., & Yen, S.-H. (2017). Mimaporia, a new genus of Epicopeiidae (Lepidoptera), with description of a new species from Vietnam. *Zootaxa*, *4254*(5), 537. https://doi.org/10.11646/zootaxa.4254.5.3

Willerslev, E. (2003). Diverse Plant and Animal Genetic Records from Holocene and Pleistocene Sediments. *Science*, *300*(5620), 791–795. https://doi.org/10.1126/science.1084114

Willerslev, Eske, & Cooper, A. (2005). Review Paper. Ancient DNA. *Proceedings of the Royal Society of London B: Biological Sciences*, *272*(1558), 3–16. https://doi.org/10.1098/rspb.2004.2813

Willerslev, Eske, Hansen, A. J., Rønn, R., Brand, T. B., Barnes, I., Wiuf, C., Gilichinsky, D., Mitchell, D., & Cooper, A. (2004). Long-term persistence of bacterial DNA. *Current Biology*, *14*(1), R9–R10. https://doi.org/10.1016/j.cub.2003.12.012

Woodward, Weyand, N. J., & Bunnell, M. (1994). DNA sequence from Cretaceous period bone fragments. *Science*, *266*(5188), 1229–1232. https://doi.org/10.1126/science.7973705

Yang, Z., & Rannala, B. (2012). Molecular phylogenetics: Principles and practice. *Nature Reviews Genetics*, *13*(5), 303–314. https://doi.org/10.1038/nrg3186

Yeates, D. K., Zwick, A., & Mikheyev, A. S. (2016). Museums are biobanks: Unlocking the genetic potential of the three billion specimens in the world's biological collections. *Current Opinion in Insect Science*, *18*, 83–88. https://doi.org/10.1016/j.cois.2016.09.009

Young, A. D., & Gillung, J. P. (2020). Phylogenomics—Principles, opportunities and pitfalls of big-data phylogenetics. *Systematic Entomology*, *45*(2), 225–247. https://doi.org/10.1111/syen.12406

Zedane, L., Hong-Wa, C., Murienne, J., Jeziorski, C., Baldwin, B. G., & Besnard, G. (2016). Museomics illuminate the history of an extinct, paleoendemic plant lineage (Hesperelaea, Oleaceae) known from an 1875 collection from Guadalupe Island, Mexico. *Biological Journal of the Linnean Society*, *117*(1), 44–57. https://doi.org/10.1111/bij.12509

Zhang, J., Cong, Q., Shen, J., Brockmann, E., & Grishin, N. (2019). Genomes reveal drastic and recurrent phenotypic divergence in firetip skipper butterflies (Hesperiidae: Pyrrhopyginae). *Proceedings of the Royal Society B-Biological Sciences*, *286*(1903), 20190609. https://doi.org/10.1098/rspb.2019.0609

Zhang, J., Shen, J., Cong, Q., & Grishini, N. (2019). Genomic analysis of the tribe Emesidini (Lepidoptera: Riodinidae). *Zootaxa*, *4668*(4), 475–488. https://doi.org/10.11646/zootaxa.4668.4.2

Zhou, X., Ren, L., Li, Y., Zhang, M., Yu, Y., & Yu, J. (2010). The next-generation sequencing technology: A technology review and future perspective. *Science China Life Sciences*, *53*(1), 44–57. https://doi.org/10.1007/s11427-010-0023-6

Zischler, H., Hoss, M., Handt, O., Haeseler, A. von, Kuyl, A. van der, & Goudsmit, J. (1995). Detecting dinosaur DNA. *Science*, *268*(5214), 1192–1193. https://doi.org/10.1126/science.7605504

List of papers

I. Mayer C., Dietz, L. Call E., Kukowka S., Martin S., Espeland M. (2020).Adding to the Lepidoptera phylogeny: Capturing hundreds of nuclear genesfrom old museums specimens. Systematic Entomology, submitted.

II. Call E., Mayer C. Twort V., Wahlberg N., Espeland M. (2020). Museomics:phylogenomics of the moth family Epicopeiidae (Lepidoptera) using target enrichment. Insect Systematics and Diversity, submitted.

III. Call. E, Twort V., Espeland M., Wahlberg N. (2020). One Method to Sequence Them All? Comparison between Whole-Genome Sequencing (WGS) and Target Enrichment (TE) of museum specimens from the moth families Epicopeiidae and Sematuridae (Lepidoptera). Manuscript.

IV. Call E., Twort V., Wheat C. W., Wahlberg N. (2020). Rear window: population genetics using museums specimens of Pieris napi (Lepidoptera).Manuscript.

Printed by Media-Tryck, Lund 2020 NORDIC SWAN ECOLABEL 3041 0903

LUND UNIVERSITY

Faculty of Science
Department of Biology

ISBN 978-91-7895-506-0