Advancing Evolutionary Biology: Genomics, Bayesian Statistics, and Machine Learning

Tobias Andermann

Department of Biological and Environmental Sciences

Faculty of Science

University of Gothenburg

UNIVERSITY OF GOTHENBURG

Gothenburg 2020

Cover illustration: Types of data that can be derived from a single specimen, using the example of the critically endangered Verreaux's sifaka (*Propithecus verreauxi*). Photographed in the Kirindy reserve in Western Madagascar by Tobias Andermann.

Advancing Evolutionary Biology:

Genomics, Bayesian Statistics, and Machine Learning

tobiasandermann88@gmail.com

ISBN 978-91-8009-136-7 (PRINT)

ISBN 978-91-8009-137-4 (PDF)

Digital version available at http://hdl.handle.net/2077/66848

Printed by Stema Specialtryck AB, Borås, Sweden, 2020

SVANENMÄRKET Trycksak 3041 0234

To my wife, my parents, and to you, the reader

Abstract

During the recent decades the field of evolutionary biology has entered the era of big data, which has transformed the field into an increasingly computational discipline. In this thesis I present novel computational method developments, including their application in empirical case studies. The presented chapters are divided into three fields of computational biology: **genomics**, **Bayesian statistics**, and **machine learning**. While these are not mutually exclusive categories, they do represent different domains of methodological expertise.

Within the field of **genomics**, I focus on the computational processing and analysis of DNA data produced with target capture, a pre-sequencing enrichment method commonly used in phylogenetic studies. I demonstrate on an empirical case study how common computational processing workflows introduce biases into the phylogenetic results, and I present an improved workflow to address these issues. Next I introduce a novel computational pipeline for the processing of target capture data, intended for general use. In an in-depth review paper on the topic of target capture, I provide general guidelines and considerations for successfully carrying out a target capture project. Within the context of **Bayesian statistics**, I develop a new computer program to predict future extinctions, which utilizes custom-made Bayesian components. I apply this program in a separate chapter to model future extinctions of mammals, and contrast these predictions with estimates of past extinction rates, produced from fossil data by a set of different recently developed Bayesian algorithms. Finally, I touch upon newly emerging **machine learning** algorithms and investigate how these can be improved in their utility for biological problems, particularly by explicitly modeling uncertainty in the predictions made by these models.

The presented empirical results shed new light onto our understanding of the evolutionary dynamics of different organism groups and showcase the utility of the methods and workflows developed in this thesis. To make these methodological advancements accessible for the whole research community, I embed them into well documented open-access programs. This will hopefully foster the use of these methods in future studies, and contribute to more informed decision-making when applying computational methods to a given biological problem.

Keywords: Computational biology, bioinformatics, phylogenetics, neural networks, NGS, target capture, Illumina sequencing, fossils, IUCN conservation status, extinction rates

Svensk sammanfattning

Under de senaste årtiondena har forskningsfältet evolutionärbiologi trätt in i eran av Big data vilket har förvandlat fältet till en allt mer datordominerad disciplin. I denna avhandling presenterar jag nyutvecklade metoder samt hur de appliceras på empiriska fallstudier. De presenterade kapitlena är indelade i tre fält inom databiologi: genomik, Bayesiansk statistik och maskininlärning. Dessa fälten är inte fullständigt skilda från varandra men representerar ändå olika områden av metodologisk expertis.

Inom fältet för genomik fokuserar jag på den digital hanteringen och analysen av DNA vilken producerats med tekniken target capture, en metod för att berika mängden genetisk data vilken ofta används inom fylogenetiska studier. Jag demonstrerar med en empirisk fallstudie hur vanligt förekommande beräkningsmetoder producerar skeva fylogenetiska resultat och jag presenterar en ny arbetsgång för att motarbeta dessa problem. Därefter presenterar jag en ny beräkningsmetod och tillvägagångssätt för hanteringen av target capture data, avsett för allmän användning. I en uttömmande översiktsartikel på ämnet target capture presenterar jag generella riktlinjer och överväganden att ha i åtanke för att på ett framgångsrikt sätt utföra target capture projekt. Inom ramverket för Bayesiansk statistik utvecklar jag ett nytt program för att förutse framtida utdöenden vilket använder sig av skräddarsydda Bayesianska komponenter. Jag applicerar detta program i ett separat kapitel för att modellera framtida utdöenden av däggdjur och kontrasterar dessa uppskattningar med uppskattningar av dåtida utrotningshastigheter vilka producerats av en annan uppsättning av nyligen utvecklade Bayesianska algoritmer. Slutligen undersöker jag hur nyligen skapade maskininlärningsalgoritmer kan förbättras i syftet att användas för biologiska problemställningar, specifikt genom att uttryckligen modellera osäkerheten i uppskattningarna gjorda av dessa modeller.

De presenterade empiriska resultaten kastar nytt ljus på vår förståelse av den evolutionära dynamiken hos olika organismgrupper och påvisar hur användbara dessa utvecklade metoder och arbetsflöden är. För att göra dessa metodologiska framstegen lättillgängliga för hela forskningssamfundet har jag inkorporerat dem i väldokumenterade, fritt tillgängliga program. Detta kommer förhoppningsvis främja användningen av dessa metoder i framtida studier samt bidra till mer välinformerade beslut när dataanalytiska metoder appliceras på biologiska problemställningar.

Manuscript overview

Genomics:

1. **Andermann**, Tobias, Alexandre M. Fernandes, Urban Olsson, Mats Töpel, Bernard Pfeil, Bengt Oxelman, Alexandre Aleixo, Brant C. Faircloth, and Alexandre Antonelli. **2019**. "Allele Phasing Greatly Improves the Phylogenetic Utility of Ultraconserved Elements." *Systematic Biology* 68 (1): 32–46. https://doi.org/10.1093/sysbio/syy039.
2. **Andermann**, Tobias, Ángela Cano, Alexander Zizka, Christine D. Bacon, and Alexandre Antonelli. **2018**. "SECAPR—a Bioinformatics Pipeline for the Rapid and User-Friendly Processing of Targeted Enriched Illumina Sequences, from Raw Reads to Alignments." *PeerJ* 6 (July): e5175. https://doi.org/10.7717/peerj.5175.
3. ***Andermann**, Tobias, *****Maria Fernanda Torres Jiménez, Pável Matos-Maraví, Romina Batista, José L. Blanco-Pastor, A. Lovisa S. Gustafsson, Logan Kistler, Isabel M. Liberal, Bengt Oxelman, Christine D. Bacon, and Alexandre Antonelli. **2020**. "A Guide to Carrying Out a Phylogenomic Target Sequence Capture Project." *Frontiers in Genetics* 10. https://doi.org/10.3389/fgene.2019.01407.

Bayesian statistic

4. **Andermann**, Tobias, Søren Faurby, Robert Cooke, Daniele Silvestro, and Alexandre Antonelli. **2020**. "iucn_sim: A New Program to Simulate Future Extinctions Based on IUCN Threat Status." *Ecography (in print)*. https://doi.org/10.1111/ecog.05110.
5. **Andermann**, Tobias, Søren Faurby, Samuel T. Turvey, Alexandre Antonelli, and Daniele Silvestro. **2020**. "The Past and Future Human Impact on Mammalian Diversity." *Science Advances* 6 (36): eabb2313. https://doi.org/10.1126/sciadv.abb2313.

Machine Learning

6. Silvestro, Daniele, and Tobias **Andermann**. **2020**. "Prior Choice Affects Ability of Bayesian Neural Networks to Identify Unknowns." *ArXiv Preprint* arXiv:2005.04987. http://arxiv.org/abs/2005.04987.

***** indicates shared first authorship

Additional manuscripts, not included in this thesis

7. Batista, Romina, Urban Olsson, Tobias **Andermann**, Alexandre Aleixo, Camila Cherem Ribas, and Alexandre Antonelli. **2020**. "Phylogenomics and Biogeography of the World's Thrushes (Aves, Turdus): New Evidence for a More Parsimonious Evolutionary History." *Proceedings of the Royal Society B: Biological Sciences* 287 (1919): 20192400.

8. Zizka, Alexander, Daniele Silvestro, Tobias **Andermann**, Josué Azevedo, Camila Duarte Ritter, Daniel Edler, Harith Farooq, Andrei Herdean, María Ariza, Ruud Scharn, Sten Svantesson, Niklas Wengström, Vera Zizka, and Alexandre Antonelli. **2019**. "CoordinateCleaner: Standardized Cleaning of Occurrence Records from Biological Collection Databases." *Methods in Ecology and Evolution* 10 (5): 744–751.

9. Hagen, Oskar, Tobias **Andermann**, Tiago B. Quental, Alexandre Antonelli, and Daniele Silvestro. **2018**. "Estimating Age-Dependent Extinction: Contrasting Evidence from Fossils and Phylogenies." *Systematic Biology* 67 (3): 458–474.

10. Antonelli, Alexandre, María Ariza, James Albert, Tobias **Andermann**, Josué Azevedo, Christine Bacon, Søren Faurby, Thais Guedes, Carina Hoorn, Lúcia G. Lohmann, Pável Matos-Maraví, Camila D. Ritter, Isabel Sanmartín, Daniele Silvestro, Marcelo Tejedor, Hans ter Steege, Hanna Tuomisto, Fernanda P. Werneck, Alexander Zizka, and Scott V. Edwards. **2018**. "Conceptual and Empirical Advances in Neotropical Biodiversity Research." *PeerJ* 6: e5644.

11. Barrett, Craig F., Christine D. Bacon, Alexandre Antonelli, Ángela Cano, and Tobias **†Hofmann**. **2016**. "An Introduction to Plant Phylogenomics with a Focus on Palms." *Botanical Journal of the Linnean Society* 182 (2): 234–255.

12. Abarenkov, Kessy, Rachel I. Adams, Irinyi Laszlo, Ahto Agan, Elia Ambrosio, Alexandre Antonelli, Mohammad Bahram, Johan Bengtsson-Palme, Gunilla Bok, Patrik Cangren, Victor Coimbra, Claudia Coleine, Claes Gustafsson, Jinhong He, Tobias **†Hofmann**, Erik Kristiansson, Ellen Larsson, Tomas Larsson, Yingkui Liu, Svante Martinsson, Wieland Meyer, Marina Panova, Nuttapon Pombubpa, Camila Ritter, Martin Ryberg, Sten Svantesson, Ruud Scharn, Ola Svensson, Mats Töpel, Martin Unterseher, Cobus Visagie, Christian Wurzbacher, Andy F. S. Taylor, Urmas Kõljalg, Lynn Schriml, and R. Henrik Nilsson. **2016**. "Annotating Public Fungal ITS Sequences from the Built Environment According to the MIxS-Built Environment Standard – a Report from a May 23-24, 2016 Workshop (Gothenburg, Sweden)." *MycoKeys; Sofia* 16: 1–15.

---

† I changed my last name from Hofmann to Andermann in 2017

Data Diversity in Evolutionary Biology

The modern era of evolutionary biology is best characterized by one key term: big data. We are producing data at unprecedented speed and scale in all fields of life sciences, and this has fundamentally contributed to transforming evolutionary biology into an increasingly computational science. While the bottleneck in the past was the speed and costs of data generation, the key challenge nowadays is that of being able to store, process, and analyze the large datasets that have become common in evolutionary biology studies.

In addition to the increased speed of data generation, data traditionally stored in isolated facilities, such as museum collections and herbaria, are increasingly being digitized and organized in large centralized public databases. The large databasing efforts allow evolutionary biologists to access datasets of unprecedented size and resolution. We are finding ourselves at an exciting point in scientific history, where for the first time we can evaluate data collected over large areas and time periods and produce cross-taxonomic analyses that identify large-scale evolutionary patterns. These analyses form a crucial element in understanding the dynamics of evolution that have been shaping the diversity and distribution of life on our planet. Particularly, such analyses can substantially add to our understanding of the processes of speciation and extinction, i.e. the generation and degradation of diversity and individual lineages in a changing world. Understanding these processes has the potential to aid us in meaningfully targeting our conservation efforts in the midst of a major global extinction crisis and in a time of rapid changes in climate, rapid growth of human population sizes, and ongoing severe habitat degradation.

There are many different types and sources of data that can inform us about the evolution of organisms. In this thesis I apply several of these data types belonging to the following three categories: genetic data, fossil data, and spatial data. I demonstrate the utility of all three of these data sources for inferring evolutionary patterns and processes, and I present advances in computational methods and models that aid in extracting previously hidden information content that lies within these data.

Genetic data

Before the emergence of genetic data in the form of DNA sequence data, researchers used to define and map homologous morphological characters that would carry information about the shared evolutionary history between any pair of organisms and thus could be used to reconstruct a phylogenetic tree for a given group of organisms (Fitch and Margoliash 1967). Starting in the late 1970s, a new source of phylogenetically informative data became broadly available with the emergence of generally applicable and accurate DNA sequencing techniques (e.g. Sanger sequencing; Sanger, Nicklen, and Coulson 1977). This development was partly driven by the formulation of more data-demanding mathematical models to infer phylogenies from large character matrices (Michener and Sokal 1957; Hennig 1966). While today

morphological character matrices are still applied and are of utility for evolutionary biology studies, the availability of DNA sequence data has revolutionized the field, as it provides data-matrices of unparalleled size.

Much technological progress has been made since the early days of Sanger sequencing, and today we are finding ourselves in an era where the sequencing of whole genomes is increasingly easy, fast, and affordable (**Figure 1**). While the original human genome project, which produced the first complete human genome sequence in 2003, took 13 years with costs of approximately 3 billion USD (International Human Genome Sequencing Consortium 2004), today's costs for sequencing a complete human genome are at less than 1,000 USD and it has become merely a matter of days from sequencing to assembling a draft genome (National Human Genome Research Institute 2020).

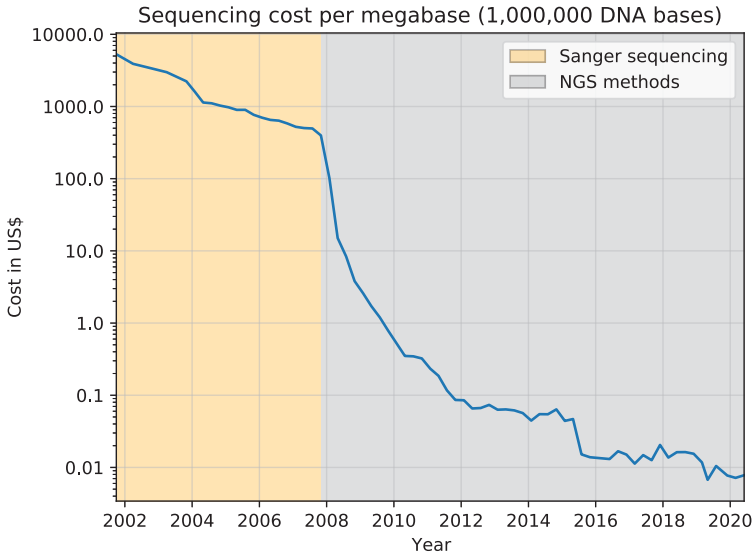

Figure 1: Development of sequencing costs through time. The data is provided by the National Human Genome Research Institute (2020) and begins with the completion of the human genome project in the year 2001. All cost information up to the end of the year 2007 is compiled from Sanger-based sequencing technology (Sanger, Nicklen, and Coulson 1977), while the costs from 2008 and beyond are based on NGS technologies. Note that the y-axis is plotted in logarithmic space, which indicates that costs have decreased more than exponentially since 2008.

This progress is mostly attributable to the advent of a new family of sequencing methods, broadly referred to as Next Generation Sequencing (NGS, see overview in Goodwin, McPherson, and McCombie 2016). These methods are being increasingly used in evolutionary biology and have become the new standard during the recent years. While there is a range of sequencing methods that are referred to as NGS, the projects in this thesis are all based on one specific method, namely sequencing by synthesis with cyclic reversible termination (Metzker 2005) as applied on the Illumina

sequencing machines (Illumina Inc., San Diego, CA, USA). From here on in this thesis, this is the sequencing method that is implied when using the term NGS without specific context.

The DNA sequence data resulting from Illumina sequencing constitute millions of short DNA reads, which are typically between 50-300 DNA base-pairs (bp) long, depending on the settings chosen on the sequencing machine. Given this limited size range of the sequencing products, these sequences are often referred to as short-read data, as opposed to the long-read sequences produced by other NGS techniques, such as the Single-Molecule Real-Time (SMRT) sequencing, applied on PacBio (Pacific Biosciences Inc., Menlo Park, CA, USA) and Nanopore (Oxford Nanopore Technologies Limited, Oxford, UK) machines, which can generate sequence lengths from several thousand up to millions of nucleotides (Amarasinghe et al. 2020).

Before sequencing a given sample on an Illumina sequencing machine, the extracted DNA is usually fragmented in the laboratory to fit the optimal fragment size range recommended for the machine (200 - 1,000 bp). All of these fragments are sequenced in parallel, starting on one end of the fragment, and in case of paired-end sequencing, followed by another sequencing round starting from the opposite end of the fragment. In the optimal case, the sequenced fragments cover the complete genome sequence and represent all areas of the genome equally, which allows to assemble the complete genome from the Illumina read sequences. With sufficient input DNA concentration and sequencing capacity of the machine, it is even possible to retrieve multiple independent sequence reads for each position on the genome, which is referred to as sequencing depth or coverage, and which leads to more confidence in the recovered sequences.

For many evolutionary studies it is not necessary to produce complete genome sequences but rather to focus sequencing efforts on a set of genetic loci that are of specific utility, for example for the purpose of inferring phylogenic trees (Faircloth et al. 2012; Lemmon, Emme, and Lemmon 2012). This locus selection is achieved by selectively amplifying DNA fragments that represent the loci of interest, while discarding all other fragments using the target capture method (Albert et al. 2007; Gnirke et al. 2009). For target capture, specific RNA bait sequences are required, which bind to the DNA fragments of interest. Each bait contains a biotin molecule, which has a high affinity to the molecule streptavidin; this relationship is utilized in a subsequent step by applying microscopic magnetic beads coated with streptavidin that consequently bind the baits; the baits at this point are still connected to the target DNA fragments (**Figure 2**). By using a magnet, the beads can be immobilized and the excess non-target DNA fragments that are still in solution (i.e. not bound to the magnetic beads) can be washed off, leaving only the target fragments behind.

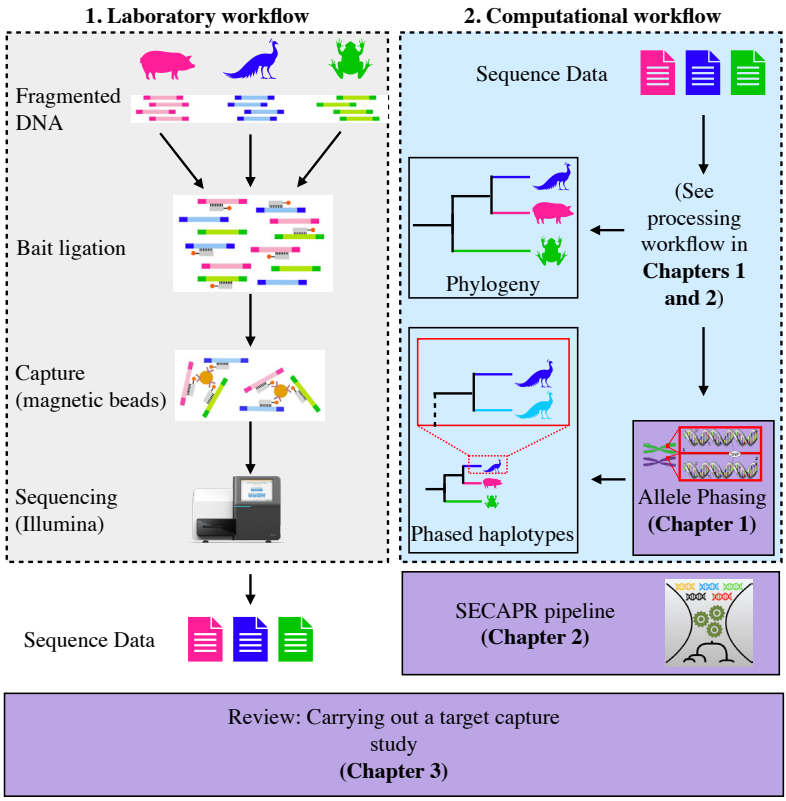

Figure 2: Simplified workflow for target capture data. The image shows a schematic overview of a target capture project, consisting of the laboratory workflow (grey box) and the bioinformatic workflow (blue box). Chapter 1 presents an addition to the bioinformatic processing workflow by implementing the phasing of allele sequences (haplotypes), which can be used for phylogenetic inference. Chapter 2 on the other hand presents a general computational pipeline, making available alternative workflows for producing multiple sequence alignments for phylogeny estimation from raw Illumina sequence data. Chapter 3 constitutes a review paper that summarizes the complete range of common laboratory and processing workflows for target capture data.

Commonly, bait sets are designed for target capture studies to capture hundreds to thousands of independent loci, each locus being between a few hundred to a few thousand bp in length. This pre-sequencing selection of target fragments has an advantage in that it drastically reduces the cumulative length of the target DNA; from essentially the whole genome (several billion bp) to a set of target loci with a cumulative length around one to several million bp. This allows pooling of more samples on the same sequencing run, while still ensuring high read coverage for the target regions of each sample. This leads to a drastic drop in sequencing costs, as hundreds of samples can be sequenced with the same sequencing effort it would otherwise take to sequence a single sample. It also leads to more manageable file-sizes per sample and to a generally simpler post-sequencing bioinformatic workflow

compared to that of assembling complete genome sequences. This is why target capture remains an increasingly popular tool for phylogenetic studies in particular.

In this thesis I apply target capture data for different organism groups, namely the hummingbird genus *Topaza* (**Chapter 1**) and the palm genus *Geonoma* (**Chapter 2**), consisting of 2,386 and 837 amplified loci, respectively. **Chapter 3**, which is a review paper, provides an overview over the application and utility of target capture in phylogenetic studies.

Fossil data

In addition to the signal of evolution that can be retrieved from the genetic code of living organisms, the evolutionary process leaves traces on a more macroscopic scale: fossil remains of organisms. Fossil data can inform us about where and when certain extant and extinct taxa occurred, provide information about morphological changes, and inform us about past diversity and its dynamics.

Recent years have seen large databasing efforts, as researchers have been collecting information about fossil occurrences in several centralized databases with different temporal and geographic focuses (e.g. Alroy, Marshall, and Miller 2004; Carrasco et al. 2007; Grimm 2008; Fortelius 2013; Rodríguez-Rey et al. 2016). The source of fossil information can include mineralized hard-tissue material (such as bones or shells), microscopic fossilized structures or cell fragments (such as pollen and phytoliths), or indirect evidence such as trace fossils (fossilized movement patterns left behind by an organism in soft substrates). There are several challenges with the inherent nature of fossil data, which can make it difficult to include such data into statistical models and large-scale analyses. These challenges are mostly related to i) taxonomic identification from morphological characters, ii) inconsistent taxonomies, iii) incomplete sampling, and iv) dating precision.

In this thesis (**Chapter 5**) I apply fossil data to estimate the times of extinction for recently extinct mammal species. In that case, the problems of morphological identification (i) and inconsistent taxonomies (ii) played a minimal role, since mammals represent the paleontologically best studied and understood taxonomic group, particularly for the rather recent time period of the Late Quaternary until today, which is the focus of that chapter. To address the issues of incomplete sampling (iii) and dating precision (iv), I apply computational methods developed and described in the program PyRate (Silvestro, Salamin, and Schnitzler 2014; Silvestro et al. 2014; 2019). I approach the issue of incomplete sampling by fully accounting for the species-specific sampling frequencies when modeling extinction dates. Regarding the issue of dating precision, I perform all analyses on 100 data replicates for each species, each based on a randomly drawn date from the dating uncertainty range. All results are summarized across these replicates, thereby fully accounting for the uncertainty in fossil dating.

Spatial data

Another important data type that is often applied in evolutionary models is spatial information about individuals, populations, and species. There are two types of spatial data that are commonly applied in evolutionary studies: occurrence data (geo-referenced point occurrences) and modeled taxon ranges. The former can for example consist of geo-referenced sightings or photographs of a taxon, while the latter consists of polygons or other geometric shapes that are inferred as a likely area for a given taxon to occur. Taxon ranges are usually modeled based on a combination of known occurrences and expert opinion, and they can be informed by additional data sources such as habitat and ecological requirements of a taxon, climatic factors, and geological information, to name a few.

The most notable and comprehensive source of point occurrence data is the Global Biodiversity Information Facility (GBIF, www.gbif.org). GBIF constitutes a centralized provider of data from many different sources, ranging from scientific inventory efforts, to citizen science projects and geotagged smartphone images from hobby naturalists. The centralized availability and data standards of GBIF enable the quick retrieval of large spatial datasets for a substantial proportion of known taxa, which can be readily applied in evolutionary studies.

There are several sources for taxon range data which can serve different purposes. For example, maps of current taxon ranges (usually on species level) are available from the International Union for the Conservation of Nature (IUCN 2020). These taxon ranges are based on expert opinion and are available for most species assessed by the IUCN Red List. While IUCN range maps exist for a large proportion of vertebrate species (subphylum: Vertebrata), most other organism groups still require substantial work and data collection before taxon ranges can be modeled.

In addition to current taxon ranges, there also exist models of potential natural taxon ranges, defined as the potential ranges of taxa if humans had not majorly interfered with their distribution (Faurby et al. 2018). This is based on the assumption that the currently observed ranges are not always representative of the actual natural habitat preferences and range extent of a given taxon. For some applications, this potential range information can be of more value than the actual current range information; for example when the aim is to infer the natural diversity of an area. For instance, the lion (*Panthera leo*) is today mainly considered an African sub-Saharan species (with a small wild population in India), but up until very recently it used to occur in wide parts of Southwest Asia and around the Mediterranean, including southern Europe (**Figure 3**). Since the current distribution of lions is heavily biased by human impact it does not represent the full range of habitats in which the species would naturally occur.

In **Chapter 5**, I apply these potential natural taxon ranges downloaded from the PHYLACINE database (Faurby et al. 2018) to determine which species are naturally endemic to specific defined bioregions. Further, in **Chapter 1** I apply point occurrence data and current range information on a much smaller scale to put into perspective the sampling locations of specimens used in that study.

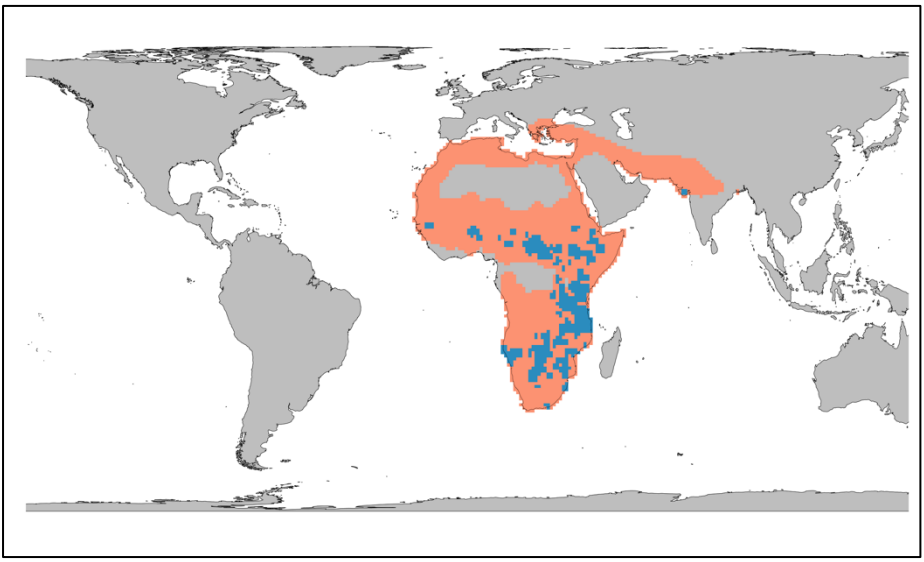

Figure 3: Current versus potential distribution of lions (*Panthera leo*). The map-area colored in orange shows the potential natural range of lions, while the area colored in blue shows the current range of the species. Range maps were downloaded from the PHYLACINE database (Faurby et al. 2018). The potential range largely reflects the historically known range of lions according to IUCN (2020). The map is plotted in cylindrical equal-area projection (CEA), standardized at 30 degrees latitude (Behrmann projection).

Computational Evolutionary Biology

Having access to the variety of data types outlined above opens up great opportunities for evolutionary biology research. With the aim to properly utilize the information content that lies within these data, a whole new discipline of research has developed during the recent decades: Computational Biology. This discipline covers a broad range of fields that all share a common thread: they utilize biological data to understand natural systems and to reconstruct biological patterns and processes. Compared to many other computational fields, the challenge is that the data computational biologists work with are notoriously imperfect; gaps in the data, biases in data collection, and many unknowns about processes that generated the data, make it impossible to develop one-shoe-fits-all software solutions. Therefore, not only the development, but also the use of existing software for processing and analyzing biological data require a thorough understanding of the mechanics of any given program or model. In this thesis I apply and develop several programs that use biological data as input. In the following I broadly describe the methodology and associated concepts used in the chapters of this thesis, divided into the sections "Genomics", "Bayesian Statistics", and "Machine Learning". Although these are not mutually exclusive categories (e.g. Bayesian statistics can be applied in genomics and machine learning), they present separate areas of methodological expertise, which I have developed during the projects represented in the chapters of this thesis.

Genomics

Genomics is the field of research that studies the whole genome or parts of the genome. In this thesis I generate and analyze DNA data representing large portions of the genome of various study organisms. The term genomics in this context is closely connected with the term bioinformatics, which describes the field of software development and application of existing software with the purpose of analyzing complex biological data, such as DNA data. The field of genomics is a rapidly evolving research field, likely due to its relevance in the medical sciences. Many computational methods and programs have been developed in the recent years for the processing of large-scale genomic data sets to keep up with the ever-increasing speed of sequence data generation. However, many of these methods are developed and streamlined for applications in the medical field, particularly for processing human DNA or that of (often haploid) pathogens. To utilize these methods in an evolutionary biology context on diploid or polyploid non-model organisms requires several alterations and new solutions for additional problems.

While any human DNA sequence can be mapped with fairly high confidence to a specific region on a reference genome, in most evolutionary biology studies we have no or very little prior genetic knowledge about the study organisms; we may not know much about the genome sequence, genome structure (number of chromosomes), and sometimes not even about the genome size or ploidy level, which severely complicates the bioinformatic processing and analysis of these data. This adds a whole range of

additional challenges that need to be tackled for genetic studies on non-model organisms, which are the majority of organisms that are being studied in evolutionary biology studies. These challenges are mainly related to i) the lack of reference sequences for sequence assembly, ii) low coverage DNA data, and iii) issues stemming from multiple paralogous gene copies, for example as a result of genome duplication.

Several computational solutions have been developed and are under current development to tackle these issues. The issue of a lack of reference sequences for non-model organisms is usually approached by using workflows that combine de novo assembly algorithms and sequence read mapping. Such workflows are implemented in several of the computational pipelines commonly used for target capture data (Faircloth 2016; Johnson et al. 2016; Allen et al. 2017), including the SECAPR pipeline of my own development (**Chapter 2**, Andermann et al. 2018).

Low coverage sequence data can be an issue, in particular when working with degraded DNA material where only small quantities of target DNA can be successfully extracted and sequenced (e.g. for ancient DNA samples). This is also a common issue with DNA samples extracted from tissues that are rich in secondary chemicals, as is commonly the case in plants tissues, since these chemicals can interfere with the buffer chemicals used during sample preparation in the lab, leading to low sequence yields (Hart et al. 2016). Besides adjusted laboratory protocols, there are some existing computational solutions that are optimized for extracting and analyzing phylogenetically informative sites from low coverage data, by combining the information of read coverage, read quality, and nucleotide identity into one joint genotype likelihood measure (Korneliussen, Albrechtsen, and Nielsen 2014). This approach avoids specific filtering thresholds that would lead to completely discarding low-coverage regions, and instead maximizes the use of all present information.

The issues of paralogy and of polyploidy at large are at this point still lacking a proper and generally applicable solution for phylogenetic NGS studies. Some partial solutions exist that can aid in filtering out sequences that show signs of paralogy (e.g. **Chapter 2**, Andermann et al. 2018), but the proper assembly of haplotype sequences for polyploid taxa remains an unresolved issue (Kyriakidou et al. 2018). Recent software developments (e.g. Moeinzadeh et al. 2020) that utilize the information in long-read datasets to assemble haplotypes of polyploid taxa may be a step toward a solution for this long-standing problem, yet the utility and general applicability of these methods still needs to be explored for most taxa.

Two specific computational methods are applied and addressed repeatedly in this thesis (**Chapters 1-3**): de novo assembly and allele phasing. I explain these two methods in more detail below.

De novo assembly

De novo assembly describes the process of assembling overlapping sequencing reads into longer sequences which are called contigs. Most assembly programs divide the sequencing reads into smaller elements, so-called kmers, which are used as the smallest

unit to identify a match between two sequences. By continuing to match overlapping kmers across the read data, the assembly algorithm builds a kmer-graph until no more overlapping kmers can be found. At that point the kmer-graph breaks and is collapsed into a single sequence. The size of the kmers (kmer length) is an important parameter in most assembly programs, which can be set by the user. This value dictates the overall length distribution of the resulting contig sequences and the number of contigs that are being assembled. The general rule is that the larger the kmer size, the longer the individual contigs and the fewer contigs are assembled. Since the choice of kmer length is non-trivial and has an effect on the resulting sequences, some assemblers, such as Spades (Bankevich et al. 2012), apply an approach that utilizes several different kmer sizes within the same assembly, allowing users to explore a whole set of kmer sizes within a single assembly run.

Some of the popular assembly programs are designed and optimized for the assembly of haploid prokaryote genomes (Bodily et al. 2015), yet they are commonly applied to assemble data from diploid and polyploid taxa. This has the result that heterozygous sites, i.e. sites where the nucleotides differ between the alleles of the sampled organism, are treated as sequencing errors. In those cases, the assembly algorithm usually decides on the most probable (i.e. most numerous among kmers) variant while discarding the alternative (Iqbal et al. 2012). This actively dismisses allelic variation and can lead to the assembly of chimeric contig sequences, i.e. sequences that are compiled from parts of different alleles, not representing actual haplotypes that exist in the population. This is an existing issue among the most commonly used assembly programs for target capture data, such as e.g. Abyss (Simpson et al. 2009), which I address in **Chapter 1** by implementing a workflow that applies allele phasing (see below). I make this improved workflow available in the SECAPR pipeline presented in **Chapter 2** (Andermann et al. 2018).

Allele phasing

Allele phasing describes the process of sorting sequencing reads into separate haplotypes. For a diploid organism, the sequencing reads that are present at a given locus stem from the two alleles at that locus, whose sequences may differ at several sites. The goal of allele phasing is to reconstruct both of these sequences. There are different types of algorithms that can aid in this task. The type of phasing that I apply in this thesis is read-connectivity based phasing (He et al. 2010), which phases reads that share the same nucleotide at a variable site into the same allele bin. This phasing approach is dependent on variable sites occurring at regular intervals at the given locus, as it will not be able to reliably connect two variants that are far apart, particularly when working with short read data. However, for most target capture datasets this limitation does not pose a problem since the targeted loci are usually of manageable length (several hundred to thousands of bp) and are often selected to be loci expected to show a decent number of variable positions. Further, the phasing algorithm can usually bridge over even sizeable non-variable gaps by utilizing the information that lies within paired-end read data (scaffolding), which is the standard for most target capture studies.

While several studies have recognized the shortcomings of contig sequences and instead compiled phased allele sequences (Lischer, Excoffier, and Heckel 2014; Potts, Hedderson, and Grimm 2014; Schrempf et al. 2016; Eriksson et al. 2018), including my own work (**Chapter 1**, Andermann et al. 2019), it is still rarely implemented in target capture studies. Further, there is ongoing discussion about how to best integrate allele sequences across many independent loci into multispecies coalescent models (Garrick, Sunnucks, and Dyer 2010; Lischer, Excoffier, and Heckel 2014; 2014; Potts, Hedderson, and Grimm 2014; Schrempf et al. 2016; Leaché and Oaks 2017). I touch upon this discussion in **Chapter 1** and demonstrate a full integration of allele sequences into phylogenetic models by applying a recently developed multispecies coalescent tree model (Jones, Aydin, and Oxelman 2015; Jones 2017), which allows species tree estimation without prior assignments of sequences to taxa. This enables treating allele sequences as independent samples from the population, rather than making strong prior assumptions by forcing monophyly on allele sequences of the same organism.

Bayesian Statistics

Bayesian statistical models have a wide range of applications, and are commonly used for modeling biological patterns and processes. One strength of Bayesian methods is that they enable the estimation of the probability distribution for all parameters of a given model. This in turn allows one to quantify the uncertainty interval surrounding these parameter estimates, which in Bayesian statistics is referred to as the credible interval (CI). Commonly a 95% CI is being reported for a given parameter estimated from the data, which translates into a 95% probability that the true parameter value lies within the reported interval.

The central Bayes theorem defining a Bayesian model is

$$\text{Posterior probability} = \frac{\text{Likelihood} \times \text{Prior probability}}{\text{Marginal likelihood}}$$

or

$$P(\alpha|x) = \frac{P(x|\alpha) \times P(\alpha)}{P(x)}$$

where α represents the parameters of the model and x are the data. The posterior probability $P(\alpha|x)$ describes the probability of the model parameters given the data, while the likelihood $P(x|\alpha)$ describes the probability of the data conditional on the parameters. The likelihood can be calculated for a given data set and a set of parameter values, based on a defined likelihood function that reflects the chosen model (e.g. **Box 1**). The prior probability $P(\alpha)$ describes the probability of the parameter values, given a function, which is defined to reflect our *a priori* knowledge about the parameter. Finally, the marginal likelihood $P(x)$ represents the posterior probability of the data integrated over the entire parameter space. This cannot be easily computed in most

cases, but as it is a normalizing constant its value is neglected for parameter estimation (Gelman et al. 2013).

Box 1: Defining a model and likelihood function.

When developing a Bayesian algorithm for a given biological problem it is crucial to define an appropriate model that accurately reflects the process that may have generated the data and that also reflects the biological question being asked. For example, let us assume that our task is to determine the distribution of body masses for a given animal population. In that case we would go into the field and collect body mass information from a limited number of individuals (usually it is not possible or feasible to measure the complete population). A reasonable assumption is that body masses (log-transformed) are normally distributed, since we find many individuals with medium body masses and only few with very large or very small body masses. In other words, we assume that our sampled body mass values are randomly drawn from a normal distribution. Therefore, in this case we could choose a normal distribution as our model of body mass distribution and use a Bayesian algorithm to estimate the model parameters μ (mean) and σ (standard deviation). Alternatively, we could also just arithmetically determine μ and σ from our collected data, but this would not account for the limited sampling and would not allow us to quantify the uncertainty in those parameter estimates.

Now where we have our data and have decided on a model (normal distribution) we can define the likelihood function to calculate the probability of the data given any set of parameter values for μ and σ. For the normal model this function is the normal probability density function

$$P(x|\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

which returns the likelihood that any given log-transformed body mass measurement (x) is drawn from a normal distribution that is defined by a given mean (μ) and standard deviation (σ).

The aim of a Bayesian algorithm is to sample from the parameter-space (the range of possible parameter values) proportionally to the posterior probability distribution. That means that we want the algorithm to sample more frequently parameter combinations that lead to high posterior probabilities but also occasionally sample parameter combinations leading to lower posterior probabilities. A common way to achieve such sampling is by using a Markov Chain Monte Carlo (MCMC) algorithm. An MCMC is an iterative process, where new parameter values are being suggested by the algorithm and are being accepted and sampled based on the posterior probability resulting from those values. Because it is an iterative process, an MCMC needs to be running for a certain number of iterations before it has reached a sufficiently representative sampling of the posterior probability distribution. Once an MCMC has sampled a sufficiently

high number of independent, non-autocorrelated samples from the posterior probability distribution (typically corresponding to an effective sample size greater than 200), the chain is considered to have converged and no further sampling is required. The sampled values for each parameter are a representative sample of the posterior probability distribution of these parameters. This allows expressing the resulting parameter estimates as credible intervals, which fully account for and scale with the limited sample size of the input data.

Estimating extinction rates

In this thesis I apply Bayesian statistics for the problem of estimating the rate of species extinction through time (**Chapter 5**). The rate of species extinction by definition is the number of extinction events over a given time frame. Therefore, in order to estimate extinction rates for a given group of taxa we need to i) reconstruct the times of extinctions of species that belonged to this group, and ii) define the width of time bins over which to summarize these extinctions (**Figure 4**). Both are non-trivial tasks, and Bayesian statistics offer a way to do both tasks jointly, while incorporating the uncertainties surrounding extinction times and the borders of the time bins.

The challenge with determining the time of extinction for a given extinct lineage is that with the exception of some extinctions of the very recent recorded history, we have no direct evidence of the disappearance of the last individual of a given species. The main evidence that we can use to model the time of extinction are fossil remains of a given species. Knowing when (and where) a certain species occurred can help to formulate some hypothesis about the likely time window of extinction. Given a fossil occurrence of a species we can say with certainty that the extinction date cannot be older than the age of the fossil (assuming correct fossil dating). Therefore, we can use the most recent fossil occurrence of a species as the oldest possible time of extinction. Another piece of information that we can use to model extinction dates is the frequency at which a given lineage occurs in the fossil record (preservation rate, sensu Silvestro, Salamin, and Schnitzler 2014). For a species that has a high frequency in the fossil record (i.e. a species with high preservation potential), we expect that the most recent fossil occurrence we know of is closer to the actual date of extinction than for a species that only rarely appears in the fossil record (**Figure 4**). In this thesis I produce estimates of extinction times based on the last occurrence of a taxon and the frequency of that taxon in the fossil record following the approach of Silvestro, Salamin, and Schnitzler (2014). To account for the stochasticity of this approach, this is done repeatedly and the resulting extinction times of all replicates are used for estimating extinction rates.

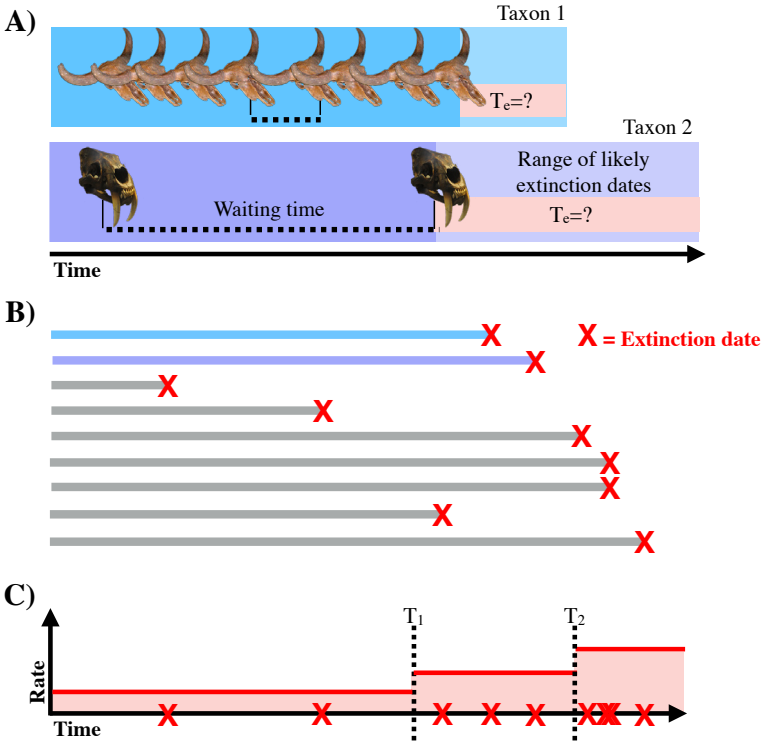

Figure 4: Estimation of extinction rates from fossil occurrences. Panel A) shows two different lineages (taxon 1 and taxon 2) with different frequencies in the fossil record; while taxon 1 has a high sampling frequency and short average waiting times between two given fossils, taxon 2 has a low frequency and long average waiting times. This leads to a larger uncertainty interval around the potential date of extinction (T_e) for taxon 2. Panel B) shows a set of different lineages and their modeled extinction times, marked by red crosses. All displayed lineages originated before the beginning of the displayed time-window (no displayed speciation events). The last panel C) shows the extinction rate estimates based on the modeled extinction times of all lineages. More extinction events in a shorter time window lead to a higher extinction rate. The time points T_1 and T_2 mark the boundaries of the time intervals that are used for rate estimation. These points are estimated jointly with the rates in the Bayesian rate-shift model applied in **Chapter 5**.

The other challenge is the definition of meaningful time intervals over which to calculate extinction rates. A simple approach would be to calculate the rate of extinction events over fixed time intervals, e.g. every 1 million years. However, this approach can be problematic, since it has been shown that the length of the chosen time interval is non-trivial as it biases the resulting rate estimates (Foote 1994; Barnosky et al. 2011). For the extinction rate estimates in this thesis (**Chapter 5**) I avoid this arbitrary definition of fixed time intervals altogether, while also fully accounting for the uncertainty in the modeled extinction dates by applying the Bayesian rate-shift model (Silvestro et al. 2019). In this model, the number and timing

of significant extinction rate shifts are estimated as parameters, and extinction rates are estimated between these dynamically adjusting time points to best fit the data (**Figure 4**). The posterior sample resulting from this model reflects the uncertainty in the number of rate-shifts, the temporal placement of these rate shifts, and the magnitude of the extinction rates for the time intervals delimited by the inferred shifts. In an alternative approach implemented in the same chapter, the extinction rate estimation is based on the trajectory of a predictor variable by estimating a correlation factor, which indicates the strength of the correlation. While the first model (rate-shift model) is a hypothesis-free approach, informed only by the extinction date data, the latter model (correlation model) can be used to determine how extinctions can best be explained by one or several predictor variables.

In **Chapter 4** I apply a Bayesian algorithm to produce future predictions of species-specific extinction rates based on the times until extinction resulting from repeated stochastic simulations for a given species. In this context, these estimated rates can be understood as extinction risks for a given species and the sampled posterior range of rate estimates represent the range of rates that could possibly produce the observed simulated outcome. This workflow of producing future extinction simulations and estimating species-specific extinction rates is available in the open-source program IUCN-SIM (Andermann et al. 2020).

Machine Learning

Machine learning algorithms represent a group of very general models that have the capability to adjust to a wide range of problems based on the observation of data. This is possible because machine learning models contain a large number of parameters, which themselves may not have meaningful interpretation to humans, but which can be optimized by the algorithm to best predict patterns in the data. This stands in contrast to conventional methods such as the Bayesian methods described above, where one needs to explicitly define a statistical model and the associated parameters, as well as the relations between these parameters. Machine learning can be very powerful in cases where we cannot define a specific statistic model and its parameters due to the sheer complexity of the problem (e.g. image recognition) or due to limited knowledge about the processes that created the data at hand.

While the field of machine learning has essentially been around and under constant development for as long as we have been using computers (Rosenblatt 1958), there has been a recent boost in the development and application of machine learning algorithms for a wide range of applications. The main drivers of the current revolution in this field are methodological progress linked to the development of deep learning methods (LeCun, Bengio, and Hinton 2015), the massively increased processing power, cloud and cluster computing, and the availability and constant production of large data (Jordan and Mitchell 2015; Dean, Patterson, and Young 2018).

Machine learning algorithms, in particular deep learning methods, are currently among the fastest growing fields of development in computer science and they are increasingly applied to a wide range of problems, including those of image and video

recognition (e.g. for personal identification or self-driving cars), medical diagnosis, speech recognition, and language translation (Dean, Patterson, and Young 2018). They also have great potential for many problems in evolutionary biology, and have already been successfully applied for tasks such as identifying gene or nucleotide function from DNA sequence data (Chen et al. 2010; Zhou and Troyanskaya 2015), species delimitation (Derkarabetian et al. 2019), species distribution modeling (Lorena et al. 2011), phylogeny estimation (Bhattacharjee and Bayzid 2020), and the estimation of extinction threats (Zizka et al. 2020).

While modern machine learning algorithms can seem like a silver bullet to apply to any problem, the successful application of them requires much work, consideration, and data engineering. Building and implementing the actual machine learning model is in most cases rather straight-forward due to well-developed code libraries that exist for this purpose, such as Tensorflow (Abadi et al. 2016), Keras (Gulli and Pal 2017), and Scikit-Learn (Pedregosa et al. 2011). The challenge usually lies in the step preceding the model implementation, which is the process of feature engineering. Feature engineering describes the transformation of the raw data into numerical features that can be used as input for the machine learning algorithm and it requires careful thought and domain knowledge. Machine learning algorithms can only function if the data contain the necessary signal, and presenting this signal to the algorithm in an unbiased numerical format constitutes the main challenge for the majority of machine learning projects.

In **Chapter 6** I apply neural networks (NNs), which are one class of machine learning algorithms. Neural Networks are inspired by the interconnectivity of neurons in the brain and process input via several layers of neurons (nodes) before producing an output. Neural networks are commonly used for one of two types of tasks: classification or regression. In this thesis I use NN implementations for classification problems, i.e. problems where we want to predict labels for our data points (e.g. image classification). Further, the implemented NNs in this thesis represent a form of supervised machine learning, which means that the algorithm is trained with a set of data for which the correct class labels are known (training set).

The nodes in one layer of an NN are connected with all or several nodes of each neighboring layer. Each of these connections is being assigned a weight, which will be optimized by the NN during the training process. Since the output labels for the training set are known, the NN can find the optimal weight configurations that map any input data point to the correct output label. A common problem is that the NN overfits towards the training set, as it optimizes the model to perfectly predict the labels for these data. Overfitting in this context is the equivalent to memorizing as opposed to actual learning. Instead of just memorizing the training set, the aim is for the NN to actively learn patterns in the data, so that it can be used to classify unknown datapoints. Several methods exist to avoid overfitting, the most common of which is to assign a separate validation set, which consists of separate data with known labels. The validation set is not used for optimizing the weights, but it is applied during the training process to evaluate how well the increasingly optimized NN predicts the labels for this set and at what point it begins to overfit towards the training set (Goodfellow et al.

2016). That is the point when training, which is an iterative process, should be terminated. Additional to the training set and validation set, it is common to evaluate the prediction accuracy of the trained model on a separate test set, which was neither used for optimization nor the determination of the point of overfitting, and is thus completely unknown to the trained model.

Once an NN is trained it can be applied to predict labels for data for which the category labels are unknown. The output comes as a list of label probabilities with one value for each of the categories on which the NN was trained. For example, if the NN was trained on images of cats, dogs, and horses, let the output for a given image be [P(cat), P(dog), P(horse)] = [0.43,0.21,0.36]. This output tells us that the animal on the image is most likely a cat, since the cat category received the highest probability. However, these output values are often incorrectly interpreted as a measure of certainty in the prediction, which they are not (Gal and Ghahramani 2016). A conventional NN makes no statement or estimate about the certainty in the prediction, but is only optimized toward making the correct prediction most of the time. This becomes evident when providing the trained network data from a class that it was not trained on; for example if we presented it an image of a cow, which does not belong to any of the three training categories, the NN would by design make a prediction for this picture to be either a cat, a dog, or a horse, and the output would make it impossible to decide that this image does not match either of the training categories. I address this issue in **Chapter 6**, using a custom implementation of a Bayesian Neural Network that provides a measure of prediction confidence and can identify data belonging to categories that are unknown to the NN.

Bayesian Neural Networks

A special case of NNs that I apply in this thesis (**Chapter 6)** are Bayesian Neural Networks (BNNs), connecting the concept of Bayesian statistics with the power of machine learning. The advantage of BNNs over regular NNs is that they enable the estimation of uncertainty in the predicted labels. The BNN implementation in **Chapter 6** is based on an MCMC, with the weights of the NN being the parameters to be estimated. A prior distribution is assigned for the parameters, which is typically a normal distribution, but see discussion in **Chapter 6**. The weights are sampled from their posterior distribution through the MCMC. Using the complete posterior sample of the weights (after discarding burn-in), predictions can be made based on the weights of each sample's iteration, leading to a slightly different set of label predictions across the whole data set. By summarizing these predictions, it is possible to calculate a credibility interval for each label prediction, reflecting the certainty of the label estimates. I demonstrate how this feature of BNNs makes it possible to identify new virus strains based on their genetic code after training the BNN on a set of known virus strains. This example shows that BNNs have great potential for the field of evolutionary biology, genomics, and beyond. For instance, BNNs could potentially be applied to detect the presence of unknown species on camera trap data or photos of herbarium specimen, or to develop new approaches of species delimitation based on the posterior probability of a given sequence to present an unknown entity to a network trained on a known set of species, to only name a few.

Objectives

The overarching aim of this thesis is to advance the computational methods in the field of evolutionary biology, to demonstrate these on empirical example data, and to make these advances available and easy to use for the research community. The work is divided into three sections, representing different fields of methodological expertise within evolutionary biology: Genomics, Bayesian Statistics, and Machine Learning. Due to the different natures of these, there are different sets of objectives for each section.

All chapters have in common that they focus on transparency and reproducibility, by being published as open access, by making the computational workflows available as free and open-source programs, and by publishing all related data and code on public repositories, such as GitHub (www.github.com). More specifically, the objectives of the individual sections are as follows.

Genomics:

- Improve the use of target capture data for reconstructing recent divergence events
- Establish a standardized and easy-to-use workflow for processing target capture data
- Centralize information about current laboratory techniques and bioinformatic tools by providing a comprehensive guide and overview of the field for future studies

Bayesian Statistics:

- Develop a framework to predict future extinctions based on species threat status information, while also accounting for ongoing trends
- Evaluate the past and expected future human impact on species extinctions, while incorporating uncertainties in the data

Machine Learning:

- Improve utility of neural networks for biological studies by modeling prediction uncertainty
- Develop method to detect data belonging to novel classes, with many potential applications in biological studies

Summary of thesis chapters

Genomics

Chapters 1-3 in this thesis fall into the field of genomics. More specifically, they all deal with NGS sequence data produced with target capture and Illumina sequencing (**Figure 2**). These chapters contribute to creating more standardized and reproducible workflows in this rapidly developing field, both by introducing and making available novel computational workflows (**Chapters 1 and 2**) and by summarizing this vast field and the different approaches that exist in a comprehensive review (**Chapter 3**).

Chapter 1 - Importance of allele phasing

Commonly, phylogenetic studies based on NGS data use de novo contig sequences assembled for their study organisms to produce phylogenetic trees. In this paper I demonstrate that this common approach can introduce bias toward older divergence times into the estimated trees, and that this bias can be avoided by phasing allele sequences and by fully integrating these sequences under the multispecies coalescent model. For diploid taxa the analysis of allele sequences as opposed to contig sequences provides more accurate and unbiased divergence-time estimates and it doubles the effective sample size by producing two sequences for a given locus for each sample, representing the two allelic variants at that locus. This is demonstrated on simulated sequence alignments representing the contig and the allele sequence approach. In this paper I also apply the introduced workflow of phasing and analyzing allele sequences to an empirical target capture data for the hummingbird genus *Topaza*. I demonstrate that the increased sample size and number of phylogenetically informative sites resulting from the process of allele phasing can aid in resolving very shallow genetic structure between recently diverged taxa. Notably this is even possible with rather conserved markers, such as the ultraconserved elements (Faircloth et al. 2012) that were targeted during target capture in this study. To enable future studies to easily apply this improved workflow for target capture data I integrate the allele phasing step into two commonly used processing pipelines for target capture data: SECAPR (**Chapter 2**, Andermann et al. 2018) and PHYLUCE (Faircloth 2016).

Chapter 2 - The SECAPR pipeline

The main challenge for many phylogenetic studies today is no longer data acquisition, but rather the proper processing and assembly of the large quantities of data produced with modern sequencing techniques. In this paper I introduce the Sequence Capture Processor (SECAPR), an open-source computational pipeline to guide researchers through the most essential parts of data processing for target capture datasets. The SECAPR pipeline is intended to make the processing of target capture data more user-friendly by establishing a well-documented workflow, also enabling the processing of these data by researchers with only limited bioinformatic backgrounds. The implemented workflow allows the user to produce different types of multiple sequence alignments for phylogenetic analyses from the raw sequencing output. One of the main

features and strengths of the pipeline is SECAPR's approach of generating sample-specific reference sequences via de novo assembly, and using these sequences in a subsequent step as a reference for read mapping. This approach is particularly useful for taxonomic groups with no available reference genomes. SECAPR also presents filtering tools to deal with loci affected by paralogy, addressing one of the major challenges in many NGS studies, particularly on plants. This is demonstrated in this paper on a challenging target capture dataset for the palm genus *Geonoma* (family: Arecaceae), for which no reference sequences are available and which harbors many potential paralogy issues. The pipeline, which can be used as a command-line program, is easy to install as a virtual environment together with all its software dependencies.

Chapter 3 - Review of target capture

The sheer number of available lab protocols, bioinformatic tools, and processing workflows for target capture data can be overwhelming. More importantly, the success of a target capture study hinges on several critical project-specific decisions that have to be made along the way. In this review paper I summarize the available tools and workflows and put them into context to help researchers make informed decisions when planning and carrying out their target capture sequencing study. The review is divided into three main sections, covering all elements of a target capture study: study design, laboratory work, and bioinformatics. The study design section mainly covers considerations regarding the choice of target loci and bait sets, and is intended to aid researchers in deciding whether to design their own bait set or to instead apply an existing publicly available set of baits. The laboratory work section summarizes common protocols and modifications of these protocols, starting with the successful extraction of DNA from different tissue types and organisms. The cleaning and amplification of DNA samples in the lab can be essential for the success of the experiment and need to be chosen in accordance with the research question and the sampled taxa. The bioinformatic section gives an overview of available processing pipelines and discusses the use cases of each of these. Further, this section summarizes the most commonly used bioinformatic tools, which are also often applied within these pipelines, to enable readers to design their own customized workflow, independently of those implemented inside the pipelines.

Bayesian statistics

Chapters 4 and 5 in this thesis both apply Bayesian statistics and contain custom-designed Bayesian algorithms for the specific problems addressed in these studies. **Chapter 5** constitutes a global case study applying novel Bayesian models to the fossil record to estimate past and current extinction rates. This chapter also contains future diversity simulations that are based on the simulation program IUCN-SIM (**Chapter 4)**, which enables the estimation of probabilities of conservation status transitions through time and future extinction rates using Bayesian algorithms.

Chapter 4 - Future extinction simulator

There are several ways in which researchers have utilized IUCN threat status information to predict future extinctions. In this paper I present a novel computational

method that is based on the current IUCN threat status of species but takes several elements into account that are usually overlooked when simulating future extinctions. We make this method available for future projects with our accompanying program IUCN-SIM, which allows for simulating future extinctions, the future threat status distribution, and species-specific extinction rates for any group of species and chosen time frame. The implemented method accounts for the recent history of a group by scanning through the last decades' record of IUCN threat status assessments. Based on these data, the program evaluates the trend inside that group, i.e. how frequently species of one threat status are assigned to a different threat status in the next IUCN evaluation. This captures possible group-specific trends of either increases in threat level, if species more frequently change from less threatened to more threatened IUCN statuses, or the success of conservation efforts, if the opposite trend is observed. The rates of status transition are estimated based on the counts of each type of transition in the IUCN history of the specified group, using a customized Bayesian algorithm modeling a Poisson process. Another novelty in our presented approach is that we account for the generation length of each given species when determining the extinction risk associated with each threat status. We demonstrate the utility of our method by simulating future extinctions for all birds (class: Aves), while also applying the IUCN status transition rates to model future changes in status. We estimate 669–809 (95% confidence interval) bird extinctions within the next 100 years, based on the current threat status and the trends in the IUCN assessment history of birds. Our program IUCN-SIM allows for the application of different methods to model the extinction risks associated with a given status, which result in significantly different estimates in the predicted number of extinctions. It is important to be aware of the effect the chosen strategy has on future estimates and to interpret the predictions in that context, as I demonstrate in this paper.

Chapter 5 - The scale of human-driven mammal extinctions

We know that human impact has negative consequences for the environment and that many species today are affected and ultimately threatened by extinction due to anthropogenic factors. This paper specifically estimates to what extent humans have historically elevated the rate of species extinction for mammals. I find that extinction rates today are currently between 1,200 to 2,300 times (95% credibility interval) higher compared to the rate at the beginning of the Late Pleistocene, i.e. before the global expansion of humans (*Homo sapiens*) out of Africa. These estimates are based on extinction events that I reconstructed from the fossil record of the last 120,000 years. I identify specific times of significant extinction rate increases locally as well as globally, which coincide with the first human arrival on several continents and island systems. Using a Bayesian correlation model, I find that the best predictor variables for these extinctions are variables related to human impact (such as global human population size and land-use). I compare these findings with an alternative factor, commonly blamed for historical extinction of mammals: global climate change, including events such as the end of last ice age and the major warming event that followed, starting about 12,000 years ago. A climate change predictor performs no better than random, suggesting that changes in climate had a minor or no effect on species extinction within the observed time frame. This is surprising given the extent of these climatic events and it demonstrates the enormous destructive potential of our

own species, by far outweighing even extreme fluctuations in climate. It also follows that under natural conditions, mammal species are unlikely to be eradicated globally in high numbers because of climatic fluctuations, even including the most severe ones in recent Earth history. However, it is important to highlight that current climate change in combination with heavily degraded habitats and already reduced population sizes has the potential to push endangered species over the edge into extinction (Woinarski et al. 2017; Díaz et al. 2019).

In this paper I also simulate future extinctions based on the current threat status of species. This is an important additional element in evaluating the full historic human impact on mammal species, since, aside from driving species to extinction we have brought many species that currently remain extant towards the brink of extinction. I find that if we allow current trends to continue, we will likely witness between 502-610 (95% confidence interval) mammal species extinctions by the year 2100. This would equate to an unprecedented escalation in the rate of extinctions and would equate to a more than 30,000-fold increase compared to the extinction rate at the beginning of the Late Pleistocene. These future simulations also show that the expected number of extinctions can be significantly reduced by increasing conservation efforts, but that many species extinctions are likely to be inevitable.

Machine Learning

The final chapter of this thesis (**Chapter 6**) applies different implementations of machine learning algorithms. More specifically in this chapter I implement and apply Bayesian neural networks (BNNs) to classification tasks and compare their performance with that of regular NNs for several different datasets.

Chapter 6 - Bayesian Neural Networks

Bayesian Neural Networks are usually implemented using a normal prior on the model parameters (weights), but little is known about the effect of this prior choice. In this chapter I investigate how the choice of the prior affects the BNN prediction accuracy as well as its ability to detect out-of-distribution data, i.e. the ability to identify datapoints that do not belong to classes included in the training set. The types of prior functions tested in this paper are uniform, normal, Cauchy (truncated), and Laplace. I find that the choice of the weight prior is not arbitrary but has a moderate impact on the prediction accuracy and, more importantly, significantly affects the ability of the BNN to detect out-of-distribution data. I demonstrate how this feature can be of great utility in biological studies, for example in identifying new virus strains based on their RNA/DNA sequences. I do not find a consistently better performance of any of the tested priors; rather, different priors perform better for different tested datasets. This suggests that it is of importance to test different priors when implementing a BNN model and to evaluate which prior performs best for the specific data at hand, rather than choosing a normal prior by default.

Another element of this paper is the comparison of BNNs with regular NNs. A standard NN has no measure of certainty in the class prediction for a given datapoint. This

precludes the detection of out-of-distribution data. However, there are existing methods to quantify uncertainties in regular NNs, such as Monte Carlo dropout (Gal and Ghahramani 2016). I compare this method with the performance of BNNs and find no significant differences in the prediction accuracy but BNNs strongly outperform NNs with Monte Carlo dropout in their ability to detect out-of-distribution data. Therefore, I conclude that although BNNs are computationally more demanding than regular NNs, they present a superior alternative in many cases and are of particular utility for many possible biological applications, where the ability to identify datapoints representing new classes (e.g. images of new species, DNA sequences of new lineages, etc.) is of particular interest.

Conclusions

Within this thesis I cover a broad range of computational methods commonly used in evolutionary biology. In the course of this work I have come to understand that it is always worthwhile to critically investigate and test existing computational tools, no matter how commonly they are applied.

I provide a case study that highlights that the common approach of phylogenetically analyzing de novo contig sequences introduces biases, and I demonstrate how the computational phasing of allele sequences properly addresses these biases (**Chapter 1**). I then make this improved workflow of allele phasing for target capture data available in the open-source program SECAPR (**Chapter 2**), together with other methodological improvements for the processing workflow of target capture data, such as generating sample-specific reference libraries for more efficient read mapping, read phasing, and SNP calling. Finally, I provide a comprehensive review of the application of target capture in phylogenetic studies to summarize the state of the art and range of resources and workflows available for target capture studies (**Chapter 3**).

These chapters are a contribution towards guiding researchers in producing phylogenetic estimates from target capture datasets. Based on my own experience, the large and complex datasets produced by NGS methods can easily overwhelm researchers that lack a comprehensive training in bioinformatics. A large number of NGS target capture studies is carried out by evolutionary biologists with a mainly taxonomic, systematic, or phylogenetic background, or by students completely new to the field. Many of those researchers lack the specific bioinformatic training necessary to implement a complete customized workflow for processing raw NGS sequence data into formats necessary for phylogenetic inference. It is therefore important to provide those researchers an overview and easy access to existing bioinformatic tools, preferably within the context of a well-documented computational pipeline, such as the SECAPR pipeline presented in this thesis. In the big picture this will hopefully lead to more informed decision-making when processing NGS data, avoiding biases and problematic data (such as e.g. sequences from potentially paralogous loci), thereby increasing the overall quality and accuracy of phylogenetic estimates from target capture data.

Another aim of my thesis was to investigate the human impact on the rate of species extinctions in mammals. The results are shocking and reveal that we have already increased the natural extinction rate by more than three orders of magnitude (**Chapter 5**). Furthermore, the large number of currently endangered species will likely lead to a further escalation of these extinction rates within the next decades, showing the full extent of our human impact on the natural world. The exact extent of expected extinctions in the near future is uncertain, but stochastic models can be used to predict likely scenarios. Preferably, such models should consider current threat levels, as well as current trends of species becoming more endangered. In this thesis I present the open-source program IUCN-SIM that models future extinctions implementing both of these sources of information (**Chapter 4**).

The program IUCN-SIM enables researchers to simulate future extinctions for any species group and time frame by utilizing the IUCN threat status information for these species (if available). One of the key challenges tackled by the program is how to translate current threat status information into numeric extinction risks. Two fundamentally different approaches have been suggested in previous studies for modeling these extinction risks and both are available to choose from in IUCN-SIM. These approaches lead to significant differences in the number of predicted future extinctions. It is therefore essential to understand the assumptions made by each model and to interpret the results in that context. I hope that by making both approaches easy to apply for any group of species through the program IUCN-SIM, I encourage future studies to evaluate these approaches in terms of accuracy of the absolute number of predicted extinctions and in terms of appropriate use cases.

In the last part of my thesis I demonstrate the utility of BNNs to estimate uncertainties in the predictions made by the trained model, as well as for detecting data belonging to classes that are unknown to the model (**Chapter 6**). This has enormous potential for many biological applications, particularly where the discovery of unknown classes could be of great interest. Many of the current challenges in biology are related to discovery, be it the discovery of new genetic virus strains, the discovery of new species, or the discovery of unknown or misclassified specimen in biological collections. For these tasks BNNs can provide an excellent solution to flag potential candidates that don't match the training categories, as demonstrated by the example of detecting virus strains that were unknown to the BNN based on their genetic code. For example, for biological collections, this has the potential to alleviate the manual work of humans having to sort through large datasets or (image) collections to find potentially problematic or new specimens; instead, these unidentifiable specimens could be automatically flagged by the BNN and then be selectively evaluated by humans.

In this thesis I present several new workflows and computational methods which I make available in the open-source programs SECAPR (18,000 downloads to date) and IUCN-SIM (400 downloads). Most researchers do not have the programming background or the time to apply complex computational workflows buried in the methods section of computational biology papers. Regardless of how relevant and innovative a proposed method is, it will be of limited use if it is not made available within an easy-to-use program, programming library, or at least documented in form of a tutorial or detailed workflow. It is key for future research in evolutionary biology and beyond that scientists ensure usability and accessibility for the whole research community, including researchers in low-income settings, by working open-source, publishing open access, and by fostering data and code transparency.

Manuscript Contributions

Chapter 1: *Allele Phasing Greatly Improves the Phylogenetic Utility of Ultraconserved Elements.* TA conceived of the project idea together with the co-authors, wrote all code, processed and analyzed the data, and wrote the manuscript with contributions from all co-authors.

Chapter 2: *SECAPR—a Bioinformatics Pipeline for the Rapid and User-Friendly Processing of Targeted Enriched Illumina Sequences, from Raw Reads to Alignments.* Tobias Andermann (TA) conceived of the project idea together with the co-authors, wrote all code, analyzed the data, and wrote the manuscript with contributions from all co-authors.

Chapter 3: *A Guide to Carrying Out a Phylogenomic Target Sequence Capture Project.* TA conducted the literature review together with all co-authors, and wrote the following elements of the manuscript with contributions from all co-authors: Abstract, Introduction, Study Design, and Bioinformatics.

Chapter 4: *iucn_sim: A New Program to Simulate Future Extinctions Based on IUCN Threat Status.* TA conceived of the project idea together with the co-authors, wrote all code with contributions by Daniele Silvestro (DS), compiled the data with contributions by co-authors, ran all analyses, and wrote the manuscript with contributions from all co-authors.

Chapter 5: *The Past and Future Human Impact on Mammalian Diversity.* TA conceived of the project idea together with the co-authors, wrote all code with contributions by DS, compiled the data with contributions by co-authors, ran all analyses, and wrote the manuscript with contributions from all co-authors.

Chapter 6: *Prior Choice Affects Ability of Bayesian Neural Networks to Identify Unknowns.* TA participated in the study design, compiled and analyzed the data with contributions by DS, produced all plots, and contributed to writing the manuscript.

References

Abadi, Martín, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, and Michael Isard. 2016. "Tensorflow: A System for Large-Scale Machine Learning." In *12th USENIX Symposium on Operating Systems Design and Implementation*, 265–283.

Albert, Thomas J., Michael N. Molla, Donna M. Muzny, Lynne Nazareth, David Wheeler, Xingzhi Song, Todd A. Richmond, Chris M. Middle, Matthew J. Rodesch, and Charles J. Packard. 2007. "Direct Selection of Human Genomic Loci by Microarray Hybridization." *Nature Methods* 4 (11): 903–905.

Allen, Julie M., Bret Boyd, Nam-Phuong Nguyen, Pranjal Vachaspati, Tandy Warnow, Daisie I. Huang, Patrick GS Grady, Kayce C. Bell, Quentin CB Cronk, and Lawrence Mugisha. 2017. "Phylogenomics from Whole Genome Sequences Using ATRAM." *Systematic Biology* 66 (5): 786–798.

Alroy, John, C Marshall, and A Miller. 2004. *Paleobiology Database*. https://paleobiodb.org/.

Amarasinghe, Shanika L., Shian Su, Xueyi Dong, Luke Zappia, Matthew E. Ritchie, and Quentin Gouil. 2020. "Opportunities and Challenges in Long-Read Sequencing Data Analysis." *Genome Biology* 21 (1): 30.

Andermann, Tobias, Ángela Cano, Alexander Zizka, Christine Bacon, and Alexandre Antonelli. 2018. "SECAPR—a Bioinformatics Pipeline for the Rapid and User-Friendly Processing of Targeted Enriched Illumina Sequences, from Raw Reads to Alignments." *PeerJ* 6 (July): e5175.

Andermann, Tobias, Søren Faurby, Robert Cooke, Daniele Silvestro, and Alexandre Antonelli. 2020. "Iucn_sim: A New Program to Simulate Future Extinctions Based on IUCN Threat Status." *Ecography* in press.

Andermann, Tobias, Alexandre M. Fernandes, Urban Olsson, Mats Töpel, Bernard Pfeil, Bengt Oxelman, Alexandre Aleixo, Brant C. Faircloth, and Alexandre Antonelli. 2019. "Allele Phasing Greatly Improves the Phylogenetic Utility of Ultraconserved Elements." *Systematic Biology* 68 (1): 32–46.

Bankevich, Anton, Sergey Nurk, Dmitry Antipov, Alexey A. Gurevich, Mikhail Dvorkin, Alexander S. Kulikov, Valery M. Lesin, Sergey I. Nikolenko, Son Pham, and Andrey D. Prjibelski. 2012. "SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing." *Journal of Computational Biology* 19 (5): 455–477.

Barnosky, Anthony D, Nicholas Matzke, Susumu Tomiya, Guinevere O U Wogan, Brian Swartz, Tiago B Quental, Charles Marshall, et al. 2011. "Has the Earth's Sixth Mass Extinction Already Arrived?" *Nature* 471 (7336): 51–57.

Bhattacharjee, Ananya, and Md. Shamsuzzoha Bayzid. 2020. “Machine Learning Based Imputation Techniques for Estimating Phylogenetic Trees from Incomplete Distance Matrices.” *BMC Genomics* 21 (1): 1–14.

Bodily, Paul M., M. Stanley Fujimoto, Cameron Ortega, Nozomu Okuda, Jared C. Price, Mark J. Clement, and Quinn Snell. 2015. “Heterozygous Genome Assembly via Binary Classification of Homologous Sequence.” *BMC Bioinformatics* 16 (S7): S5.

Carrasco, Marc A., Anthony D. Barnosky, Brian P. Kraatz, and Edward B. Davis. 2007. “The Miocene Mammal Mapping Project (Miomap): An Online Database of Arikareean Through Hemphillian Fossil Mammals.” *Bulletin of Carnegie Museum of Natural History* 2007 (39): 183–188.

Chen, Huang-Wen, Sunayan Bandyopadhyay, Dennis E. Shasha, and Kenneth D. Birnbaum. 2010. “Predicting Genome-Wide Redundancy Using Machine Learning.” *BMC Evolutionary Biology* 10 (1): 357.

Dean, Jeff, David Patterson, and Cliff Young. 2018. “A New Golden Age in Computer Architecture: Empowering the Machine-Learning Revolution.” *IEEE Micro* 38 (2): 21–29.

Derkarabetian, Shahan, Stephanie Castillo, Peter K. Koo, Sergey Ovchinnikov, and Marshal Hedin. 2019. “A Demonstration of Unsupervised Machine Learning in Species Delimitation.” *Molecular Phylogenetics and Evolution* 139 (October): 106562.

Díaz, Sandra, Josef Settele, Eduardo S. Brondízio, Hien T. Ngo, John Agard, Almut Arneth, Patricia Balvanera, et al. 2019. “Pervasive Human-Driven Decline of Life on Earth Points to the Need for Transformative Change.” *Science* 366 (6471).

Eriksson, Jonna S., Filipe de Sousa, Yann JK Bertrand, Alexandre Antonelli, Bengt Oxelman, and Bernard E. Pfeil. 2018. “Allele Phasing Is Critical to Revealing a Shared Allopolyploid Origin of Medicago Arborea and M. Strasseri (Fabaceae).” *BMC Evolutionary Biology* 18 (1): 9.

Faircloth, Brant C. 2016. “PHYLUCE Is a Software Package for the Analysis of Conserved Genomic Loci.” *Bioinformatics* 32 (5): 786–788.

Faircloth, Brant C., John E. McCormack, Nicholas G. Crawford, Michael G. Harvey, Robb T. Brumfield, and Travis C. Glenn. 2012. “Ultraconserved Elements Anchor Thousands of Genetic Markers Spanning Multiple Evolutionary Timescales.” *Systematic Biology* 61 (5): 717–726.

Faurby, Søren, Matt Davis, Rasmus Østergaard Pedersen, Simon D. Schowanek, Alexandre Antonelli, and Jens-Christian Svenning. 2018. “PHYLACINE 1.2: The Phylogenetic Atlas of Mammal Macroecology.” *Ecology* 99 (11): 2626.

Fitch, Walter M., and Emanuel Margoliash. 1967. “Construction of Phylogenetic Trees.” *Science* 155 (3760): 279–84.

Foote, Mike. 1994. “Temporal Variation in Extinction Risk and Temporal Scaling of Extinction Metrics.” *Paleobiology* 20 (4): 424–444.

Fortelius, M. 2013. *New and Old Worlds Database of Fossil Mammals (NOW)*. http://www.helsinki.fi/science/now.

Gal, Yarin, and Zoubin Ghahramani. 2016. “Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning.” In *International Conference on Machine Learning*, 1050–1059.

Garrick, Ryan C., Paul Sunnucks, and Rodney J. Dyer. 2010. “Nuclear Gene Phylogeography Using PHASE: Dealing with Unresolved Genotypes, Lost Alleles, and Systematic Bias in Parameter Estimation.” *BMC Evolutionary Biology* 10 (1): 118.

Gelman, Andrew, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari, and Donald B. Rubin. 2013. *Bayesian Data Analysis*. CRC press.

Gnirke, Andreas, Alexandre Melnikov, Jared Maguire, Peter Rogov, Emily M. LeProust, William Brockman, Timothy Fennell, Georgia Giannoukos, Sheila Fisher, and Carsten Russ. 2009. “Solution Hybrid Selection with Ultra-Long Oligonucleotides for Massively Parallel Targeted Sequencing.” *Nature Biotechnology* 27 (2): 182–189.

Goodfellow, Ian, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. 2016. *Deep Learning*. Vol. 1. 2. MIT press Cambridge.

Goodwin, Sara, John D. McPherson, and W. Richard McCombie. 2016. “Coming of Age: Ten Years of next-Generation Sequencing Technologies.” *Nature Reviews Genetics* 17 (6): 333–51.

Grimm, EC. 2008. “Neotoma: An Ecosystem Database for the Pliocene, Pleistocene, and Holocene.” *Illinois State Museum Scientific Papers E Series* 1. https://www.neotomadb.org/.

Gulli, Antonio, and Sujit Pal. 2017. *Deep Learning with Keras*. Packt Publishing Ltd.

Hart, Michelle L., Laura L. Forrest, James A. Nicholls, and Catherine A. Kidner. 2016. “Retrieval of Hundreds of Nuclear Loci from Herbarium Specimens.” *Taxon* 65 (5): 1081–1092.

He, Dan, Arthur Choi, Knot Pipatsrisawat, Adnan Darwiche, and Eleazar Eskin. 2010. “Optimal Algorithms for Haplotype Assembly from Whole-Genome Sequence Data.” *Bioinformatics* 26 (12): i183–i190.

Hennig, Willi. 1966. *Phylogenetic Systematics*. University of Illinois Press.

International Human Genome Sequencing Consortium. 2004. “Finishing the Euchromatic Sequence of the Human Genome.” *Nature* 431 (7011): 931–45.

Iqbal, Zamin, Mario Caccamo, Isaac Turner, Paul Flicek, and Gil McVean. 2012. “De Novo Assembly and Genotyping of Variants Using Colored de Bruijn Graphs.” *Nature Genetics* 44 (2): 226–232.

IUCN. 2020. “The IUCN Red List of Threatened Species.” IUCN Red List of Threatened Species. 2020. https://www.iucnredlist.org/en.

Johnson, Matthew G., Elliot M. Gardner, Yang Liu, Rafael Medina, Bernard Goffinet, A. Jonathan Shaw, Nyree JC Zerega, and Norman J. Wickett. 2016. “HybPiper: Extracting Coding Sequence and Introns for Phylogenetics from High-Throughput Sequencing Reads Using Target Enrichment.” *Applications in Plant Sciences* 4 (7): 1600016.

Jones, Graham. 2017. “Algorithmic Improvements to Species Delimitation and Phylogeny Estimation under the Multispecies Coalescent.” *Journal of Mathematical Biology* 74 (1–2): 447–467.

Jones, Graham, Zeynep Aydin, and Bengt Oxelman. 2015. “DISSECT: An Assignment-Free Bayesian Discovery Method for Species Delimitation under the Multispecies Coalescent.” *Bioinformatics* 31 (7): 991–998.

Jordan, Michael I., and Tom M. Mitchell. 2015. “Machine Learning: Trends, Perspectives, and Prospects.” *Science* 349 (6245): 255–260.

Korneliussen, Thorfinn Sand, Anders Albrechtsen, and Rasmus Nielsen. 2014. “ANGSD: Analysis of next Generation Sequencing Data.” *BMC Bioinformatics* 15 (1): 356.

Kyriakidou, Maria, Helen H. Tai, Noelle L. Anglin, David Ellis, and Martina V. Strömvik. 2018. “Current Strategies of Polyploid Plant Genome Sequence Assembly.” *Frontiers in Plant Science* 9.

Leaché, Adam D., and Jamie R. Oaks. 2017. “The Utility of Single Nucleotide Polymorphism (SNP) Data in Phylogenetics.” *Annual Review of Ecology, Evolution, and Systematics* 48: 69–84.

LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton. 2015. “Deep Learning.” *Nature* 521 (7553): 436–444.

Lemmon, Alan R., Sandra A. Emme, and Emily Moriarty Lemmon. 2012. “Anchored Hybrid Enrichment for Massively High-Throughput Phylogenomics.” *Systematic Biology* 61 (5): 727–744.

Lischer, Heidi EL, Laurent Excoffier, and Gerald Heckel. 2014. “Ignoring Heterozygous Sites Biases Phylogenomic Estimates of Divergence Times: Implications for the Evolutionary History of Microtus Voles.” *Molecular Biology and Evolution* 31 (4): 817–831.

Lorena, Ana C., Luis FO Jacintho, Marinez F. Siqueira, Renato De Giovanni, Lúcia G. Lohmann, André CPLF De Carvalho, and Missae Yamamoto. 2011. “Comparing Machine Learning Classifiers in Potential Distribution Modelling.” *Expert Systems with Applications* 38 (5): 5268–5275.

Metzker, Michael L. 2005. “Emerging Technologies in DNA Sequencing.” *Genome Research* 15 (12): 1767–76.

Michener, Charles D., and Robert R. Sokal. 1957. “A Quantitative Approach to a Problem in Classification.” *Evolution* 11 (2): 130–162.

Moeinzadeh, M.-Hossein, Jun Yang, Evgeny Muzychenko, Giuseppe Gallone, David Heller, Knut Reinert, Stefan Haas, and Martin Vingron. 2020. “Ranbow: A Fast and Accurate Method for Polyploid Haplotype Reconstruction.” *PLOS Computational Biology* 16 (5): e1007843.

National Human Genome Research Institute. 2020. “The Cost of Sequencing a Human Genome.” Genome.Gov. 2020. https://www.genome.gov/about-genomics/fact-sheets/Sequencing-Human-Genome-cost.

Pedregosa, Fabian, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, and Vincent Dubourg. 2011. “Scikit-Learn: Machine Learning in Python.” *The Journal of Machine Learning Research* 12: 2825–2830.

Potts, Alastair J., Terry A. Hedderson, and Guido W. Grimm. 2014. “Constructing Phylogenies in the Presence of Intra-Individual Site Polymorphisms (2ISPs) with a Focus on the Nuclear Ribosomal Cistron.” *Systematic Biology* 63 (1): 1–16.

Rodríguez-Rey, Marta, Salvador Herrando-Pérez, Barry W Brook, Frédérik Saltré, John Alroy, Nicholas Beeton, Michael I Bird, et al. 2016. “A Comprehensive Database of Quality-Rated Fossil Ages for Sahul’s Quaternary Vertebrates.” *Scientific Data* 3: 160053.

Rosenblatt, F. 1958. “The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain.” *Psychological Review* 65 (6): 386–408.

Sanger, F., S. Nicklen, and A. R. Coulson. 1977. “DNA Sequencing with Chain-Terminating Inhibitors.” *Proceedings of the National Academy of Sciences* 74 (12): 5463–67.

Schrempf, Dominik, Bui Quang Minh, Nicola De Maio, Arndt von Haeseler, and Carolin Kosiol. 2016. “Reversible Polymorphism-Aware Phylogenetic Models and Their Application to Tree Inference.” *Journal of Theoretical Biology* 407: 362–370.

Silvestro, Daniele, Nicolas Salamin, Alexandre Antonelli, and Xavier Meyer. 2019. “Improved Estimation of Macroevolutionary Rates from Fossil Data Using a Bayesian Framework.” *Paleobiology* 45 (4): 546–70.

Silvestro, Daniele, Nicolas Salamin, and Jan Schnitzler. 2014. “PyRate: A New Program to Estimate Speciation and Extinction Rates from Incomplete Fossil Data.” *Methods in Ecology and Evolution* 5 (10): 1126–1131.

Silvestro, Daniele, Jan Schnitzler, Lee Hsiang Liow, Alexandre Antonelli, and Nicolas Salamin. 2014. “Bayesian Estimation of Speciation and Extinction from Incomplete Fossil Occurrence Data.” *Systematic Biology* 63 (3): 349–367.

Simpson, Jared T., Kim Wong, Shaun D. Jackman, Jacqueline E. Schein, Steven JM Jones, and Inanç Birol. 2009. “ABySS: A Parallel Assembler for Short Read Sequence Data.” *Genome Research* 19 (6): 1117–1123.

Woinarski, John CZ, Stephen T. Garnett, Sarah M. Legge, and David B. Lindenmayer. 2017. “The Contribution of Policy, Law, Management, Research, and Advocacy Failings to the Recent Extinctions of Three Australian Vertebrate Species.” *Conservation Biology* 31 (1): 13–23.

Zhou, Jian, and Olga G. Troyanskaya. 2015. “Predicting Effects of Noncoding Variants with Deep Learning–Based Sequence Model.” *Nature Methods* 12 (10): 931–934.

Zizka, Alexander, Daniele Silvestro, Pati Vitt, and Tiffany M. Knight. 2020. “Automated Conservation Assessment of the Orchid Family with Deep Learning.” *Conservation Biology*.

Acknowledgements

I am thankful to:

First and foremost, I want to thank my co-supervisor **Daniele Silvestro**, who has been an incredible help and influence during my PhD. Daniele, thank you so much for always being available no matter how busy your schedule is and for all the time you have spent helping me out. In you I have found a great scientific inspiration, and a true friend. It has been a real pleasure and privilege to work with you over these years and I'm looking forward to many more fruitful collaborations in the future.

And of course, my supervisor **Alexandre Antonelli** who has given me this great opportunity to conduct my PhD within his amazing research group. I truly appreciate all the trust you have put into me and the opportunities, inspiration, and contacts you have provided me throughout my PhD education. Without you I wouldn't be where I am. Alex, thank you for everything you have provided over these years, for all the fun and productive group meetings, and for the times you invited me to your house in Gothenburg, your temporary place in Boston, or your flat in London. As I'm writing this, I'm realizing that all the attempts you have made to move away have not stopped me from following you around. I think that is a good sign.

Søren Faurby... I don't even know where to start. You just popped up one day some years ago and immediately became an integral part of several of my projects. You undoubtedly have added a lot of depth and quality to our shared projects, and you have provided essential data for making these projects possible. I was lucky that you showed up when you did, at the beginning of my PhD. Despite your repeated attempts to fire me, I have managed to cling on and bring this thing to a finish. To end this with one of your favorite ping-pong quotes, I would say the last years of working with you were "better than expected".

Christine Bacon for sharing my enthusiasm about NGS projects and for all the shared projects we have been involved in together. It has always been enjoyable to teach and work with you over the past years. Also thank you for all the fun get-togethers and parties that you have initiated. You are a huge source of inspiration and your ability to come up with new project ideas has been the beginning of many shared projects.

Allison Perrigo for basically doing everything at once: outreach, scientific writing, group management, planning group meetings, creating lots of inspiring and large-scale public events, being the director of the Gothenburg Global Biodiversity Centre (GGBC), and probably many more. Not sure how you do it, but you are always available and somehow always have a solution for things. In German we would call you "eine eierlegende Wollmilchsau" and I apologize if the literal translation may sound a little undignified, but it is meant in the most positive way. You are always there when the going gets tough. That was even true when I went on a backpacking trip in the wilderness of northern Sweden along the Kungsleden: I ascended a steep

hill with my last bit of energy and who do I run into unexpectedly right after the summit? You guessed it: Allison Perrigo, about to ask if I need help with anything.

I thank my long-term PhD companions and friends **Josue Azevedo**, **Harith Farooq**, **Camila Duarte Ritter** and **Alexander Zizka** for unforgettable times together and many fun nights out in Gothenburg. I will never forget our amazing karaoke nights, table-foosball tournaments, ping-pong matches, and legendary PhD parties. Even the time Harith dragged me to do cross-fit with him turned out to be a fun time. I couldn't think of a better crowd to have shared my PhD times with. I wish I could do it all over again.

My dear colleagues and friends at the department and beyond (in no particular order): Ferran, Rob, Adrian, Helene, Louisa, Mafe, Gydis, Nicolas, Pavel, Myriam, Jonna, Dom, Ann-Sophie, Yannick, Ntwae, Angela, Weston, Anieli, Beatriz, Matheus, Igor, Paola, Daniel E., Emke, Fernanda, Romina, Ivana, Johannes, Maria, Juan, Leon, James, Mårten, Sanna, Daniel M., Yann, Erik, Patrik, Pedro, and everybody who feels like they should be on this list but who I forgot to mention.

All other **Antonelli Lab members and visiting researchers** over the last years who have influenced and shaped my time as a PhD student, including those of you located at Kew gardens in London. There are simply too many people to list here, but this goes out to all of you: Thank you for the great times!

The **PhD community** at BioEnv and Marine, in particular my previous fellow PhD board members, for many fun get-togethers, game nights, and meetings.

Helene Aronsson for the very generous help with translating my writing into Swedish on several occasions, including the abstract of this thesis. Also thank you for the amazing work you are doing in coordinating the GGBC, including your help with coordinating the workshops I have been teaching—it was great to have your help.

Anna Ansebo for having been our lab-coordinator and for being one of the most understanding and trust-worthy people I have so far met.

Ylva Heed, who always has an open ear and heart for any issues, be they work-related or private. Thank you for being the person that you are and for spreading fun and joy wherever you go. I have truly enjoyed having you for a colleague and friend over the last years.

Matthias Obst for creating the GGBC workshops together. Thank you for putting your trust in me with creating the course materials and for giving me this exciting teaching opportunity. It has been a real pleasure working with you.

Bengt Oxelman, who has been a big influence since my Master's degree on developing my knowledge in the field of phylogenetics in particular. It has been a privilege to work, teach, and discuss with you throughout the last years. And a big thanks for

organizing the amazing South Africa fieldtrip in 2016, and for putting up with me notoriously wandering off in the field when it was time to gather again.

Urban Olsson for amazing bird-watching field trips, including teaching the bird-course together at the field station. It has been a great pleasure working and teaching with you, and I'm grateful I never hit your bumper when driving behind you and you hitting the brakes unannounced because you saw an exciting bird on the roadside.

Henrik Aronsson and **Mats Olsson** for the countless emails over the years and the swift help with any of my questions.

Mari Källersjö for being so flexible and supportive throughout all these years, from helping me to get courses registered in my curriculum to helping organize my thesis defense. Beyond that, it was a pleasure to learn from you about the astonishing Asteraceae diversity during our field trip to South Africa.

Sven Toresson for managing everything at Botan and for always helping out.

Niclas Siberg for all your help with any technology related questions.

Ingela Lyck for the efforts you had to go through with repeatedly sorting out my teaching credits and updating my work contract.

Alexander Schliep for having made time to meet with me on several occasions to discuss future career plans. Your input has been very valuable and has definitely influenced my thinking and future plans.

Hugo de Boer for enabling me to run my own course within the ForBio network, and for the legendary ForBio annual meetings.

Britt Andermann for your unconditional love and support during all these years and for putting up with me working late nights at times and teaching online courses from home. Also, huge thanks for all the proof-reading you have done for me throughout my PhD; it's a huge privilege to have full-time access to the truly professional editor and academic writer you are. You ma-ma-ma-make me happy!

ForBio and the **Adlerberska foundation** for financial support.

The **Chalmers Centre for Computational Science and Engineering** (C3SE) for the provided computational resources needed for the work presented in this thesis. In total, I ran computations worth 202,400 core-hours on the C3SE clusters during the course of this PhD. Without having access to this parallelized computing, the computations for this thesis would have taken me more than 23 years on a single computer core to complete!

This PhD was funded by a **Wallenberg Academy Fellow** grant awarded to Alexandre Antonelli by the Knut and Alice Wallenberg Foundation.