frontiers | Research Topics

Recent advances in museomics: Revolutionizing biodiversity research

Edited by
Jonathan J. Fong, Anchalee Aowphol, Jimmy McGuire, Chirasak Sutcharit, Mozes Blom and Pamela Soltis

Published in
Frontiers in Ecology and Evolution

May 2023

ISSN 1664-8714
ISBN 978-2-8325-2433-6
DOI 10.3389/978-2-8325-2433-6

About Frontiers

Frontiers is more than just an open access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers journal series

The Frontiers journal series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the *Frontiers journal series* operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews. Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the *Frontiers journals series*: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area.

Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers editorial office: frontiersin.org/about/contact

Recent advances in museomics: Revolutionizing biodiversity research

Topic editors
Jonathan J. Fong — Lingnan University, Tuen Mun, SAR China
Anchalee Aowphol — Kasetsart University, Thailand
Jimmy McGuire — University of California, Berkeley, United States
Chirasak Sutcharit — Chulalongkorn University, Thailand
Mozes Blom — Museum of Natural History Berlin (MfN), Germany
Pamela Soltis — University of Florida, United States

Citation
Fong, J. J., Aowphol, A., McGuire, J., Sutcharit, C., Blom, M., Soltis, P., eds. (2023). *Recent advances in museomics: Revolutionizing biodiversity research*. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-8325-2433-6

Table of contents

frontiers | Frontiers in Ecology and Evolution

TYPE Editorial
PUBLISHED 03 May 2023
DOI 10.3389/fevo.2023.1188172

Check for updates

OPEN ACCESS

EDITED AND REVIEWED BY
Mark A. Elgar,
The University of Melbourne, Australia

*CORRESPONDENCE
Jonathan J. Fong
✉ jonfong@ln.edu.hk

RECEIVED 17 March 2023
ACCEPTED 03 April 2023
PUBLISHED 03 May 2023

CITATION
Fong JJ, Blom MPK, Aowphol A, McGuire JA, Sutcharit C and Soltis PS (2023) Editorial: Recent advances in museomics: revolutionizing biodiversity research.
Front. Ecol. Evol. 11:1188172.
doi: 10.3389/fevo.2023.1188172

Editorial: Recent advances in museomics: revolutionizing biodiversity research

Jonathan J. Fong[1]*, Mozes P. K. Blom[2], Anchalee Aowphol[3], Jimmy A. McGuire[4], Chirasak Sutcharit[5] and Pamela S. Soltis[6]

[1]Science Unit, Lingnan University, Tuen Mun, Hong Kong SAR, China, [2]Museum für Naturkunde, Leibniz Institut für Evolutions- und Biodiversitätsforschung, Berlin, Germany, [3]Department of Zoology, Faculty of Science, Kasetsart University, Bangkok, Thailand, [4]Museum of Vertebrate Zoology, Department of Integrative Biology, University of California, Berkeley, Berkeley, CA, United States, [5]Animal Systematics Research Unit, Department of Biology, Faculty of Sciences, Chulalongkorn University, Bangkok, Thailand, [6]Florida Museum of Natural History, University of Florida, Gainesville, FL, United States

KEYWORDS

museomics, natural history collection (NHC), historical DNA, archival DNA, DNA barcoding, formalin extraction, target capture, herbaria

Editorial on the Research Topic
Recent advances in museomics: revolutionizing biodiversity research

Introduction

Museomics, a term coined by Drs. Stephan Schuster and Webb Miller in ∼2009, refers to "the large-scale analysis of the DNA content of museum collections" (http://mammoth.psu.edu/museomics.html). Although such DNA studies existed before the term was first used, "museomics" highlighted the importance of specimens in biological studies.

Specimens in natural history collections (NHCs) have been collected for hundreds of years to document the spatial and temporal occurrences of species. It is estimated that NHCs worldwide house 3 billion specimens (Soberon, 1999). These specimens preserve a wealth of information, such as morphological and genetic data on the identity and phylogenetics of species, biogeographic and ecological data, and even biographical information of the collectors, and the contributions of NHCs extend well-beyond organismal biology research to fields such as public health (Suarez and Tsutsui, 2004; Cook et al., 2020) and education (Ellwood et al., 2020; Lendemer et al., 2020; National Academies of Sciences Engineering and Medicine, 2020). NHCs are valuable resources with unknown future potential, and there are countless examples of research made possible that was not the goal of the original collector (Heberling et al., 2019; Miller et al., 2020). We provide three examples. First, Moritz et al. (2008) compared modern specimens of small mammals to those collected ∼100 years prior to document how climate change caused the distributions of some species to shift in elevation. Second, bird egg collections in museums were instrumental in showing the role of DDT in causing egg-shell thinning that adversely affected raptor and pelican populations (Ratcliffe, 1967; Hickey and Anderson, 1968). Lastly, Freelance et al. (2022) stress the importance of properly designing captive breeding programs, since the sensory organs of the endangered Lord Howe Island stick insect (*Dryococelus australis*) differed between wild specimens (>100 years old) and individuals bred in captivity. Given the accelerated rate of biodiversity loss, the role of NHCs will increase in prominence by being an archive of genetic and phenotypic diversity across space and time for many species that have gone extinct or where populations have vanished.

Similarly in terms of unexpected potential, the advent of DNA sequencing technology opened up new avenues for specimen-based research. Modern specimen preparation now includes special steps to preserve DNA/RNA in tissues (e.g., freezing or placing tissues in ethanol or other storage media) for genetic studies, while previously there were no special efforts to preserve the DNA. There are challenges working with these materials, such as DNA naturally degrading over time and the DNA of formalin-fixed specimens being cross-linked with proteins and other DNA (Raxworthy and Smith, 2021). Advances in laboratory methods and new sequencing technologies (e.g., high throughput short-read sequencing) have facilitated improvements in our ability to recover and sequence DNA from museum specimens.

There are four primary sources of DNA that we discuss here: ancient DNA (aDNA), historical DNA (hDNA), modern DNA, and archival DNA (Raxworthy and Smith, 2021). DNA extracted from samples that died under natural circumstances and were later recovered from the field are referred to as aDNA. Familiar examples of aDNA include samples obtained from species such as mammoths and cave bears, which can be quite old and are often >200 years in age. In contrast, DNA extracted from formalin-fixed or ethanol-fixed specimens that were preserved and stored in museum collections is referred to as hDNA (these specimens are usually <200 years old). DNA extracted from tissue samples specifically prepared with genetic analysis in mind is referred to as modern DNA and is usually <40 years old. Archival DNA refers to hDNA and modern DNA stored in museum specimens. The first studies from researchers using the word "museomics" sequenced mitochondrial genomes from the aDNA in hair of the extinct Siberian mammoth (Gilbert et al., 2008) and Tasmanian tiger (Miller et al., 2009).

This Research Topic is a collection of studies highlighting advances in museomics, both in demonstrating applications and refining methodologies. Some applications demonstrated in this Research Topic include using DNA barcoding of a degraded whale sample to identify it to subspecies (Ren et al.), obtaining data from a holotype to verify the existence of an undescribed rodent genus (Castañeda-Rico et al.), obtaining DNA from hundreds of herbarium specimens to elucidate the phylogeography of the genus *Dalbergia* (Sotuyo et al.), and using target capture to understand the phylogenetic placement of two rare shark species (Agne, Naylor et al.). These studies are diverse in the DNA type used (hDNA and modern DNA), taxa studied, objectives, and approaches. A variety of factors have been identified that affect the performance of sequencing DNA from specimens, and a major goal of museomics is to develop a set of best practices to maximize success (Raxworthy and Smith, 2021). Efforts are being made to document and understand these factors (e.g., Irestedt et al., 2022), and this Research Topic was initiated to further this cause. As an overview of this Research Topic, we identify several factors being addressed across the articles (Figure 1). Following the terminology of Roycroft et al., we organize these factors temporally in the research process as pre-sequencing and post-sequencing (Figure 1). This list of factors is not exhaustive, but rather highlights those that are addressed in this Research Topic. We note that findings in different studies may contradict each other, highlighting the dynamic state of the field and the need for more exhaustive research on this topic.

Pre-sequencing

Pre-sequencing factors dealt with in these studies are either related to the specimen or methodological advances to improve our ability to obtain DNA from historical collections.

Specimen-related factors

Four specimen-related factors are addressed: taxa, tissue type, age, and preservation history. A diversity of taxa was targeted across studies (mammals, insects, gastropods, bony fish, cartilaginous fish, reptiles, sponges, polychaetes, crustaceans, amphibians, plants, arachnids, birds), with mammals being the most frequent focal group (six studies). Agne, Preick et al. included samples from nine classes of animals and found lower success with crustaceans, insects, and cartilaginous fish, and higher success with sponges, gastropods, polychaetes, and amphibians. Another study on gastropods (Clewing et al.) noted that mollusks can be difficult to work with because their tissues are high in mucopolysaccharides, which can hinder DNA extraction.

Several studies compared the performance of different tissue types. In a study of wolf specimens comparing tissue types (jaw bone, nasal bone, skin), skin had the best performance and should be preferred because it is less destructive to the specimen (Pacheco et al.). In contrast, Roycroft et al. found in their mammal study that DNA extraction from toe pad and bone tissue performed better than with skin.

The importance of the age of specimens was commonly explored in these studies, with both types of archival DNA (hDNA and modern DNA) investigated across studies. The oldest specimen included was 192 years old (Agne, Preick et al.). Some studies found a negative correlation between age and DNA yield (Bernstein and Ruane; Hawkins et al.; Roycroft et al.), while others found no relationship (Nunes et al.; Pacheco et al.; Pavlek et al.).

Preservation history is an important factor that can be difficult to evaluate because the entire preservation process is usually not fully documented. Frozen tissue, as expected, preserves DNA better than other methods (Speer et al.). Agne, Preick et al. found that dry specimens performed better than wet across a variety of taxa, while Nunes et al. found the opposite for insects where ethanol-preserved specimens performed better than dry papered and pinned specimens. Variation within preservation types, obscuring trends, is potentially confounded by the time between euthanization and preservation (Speer et al.).

Lab work-related factors

Three lab work-related factors are target loci, DNA extraction protocol, and method of library preparation.

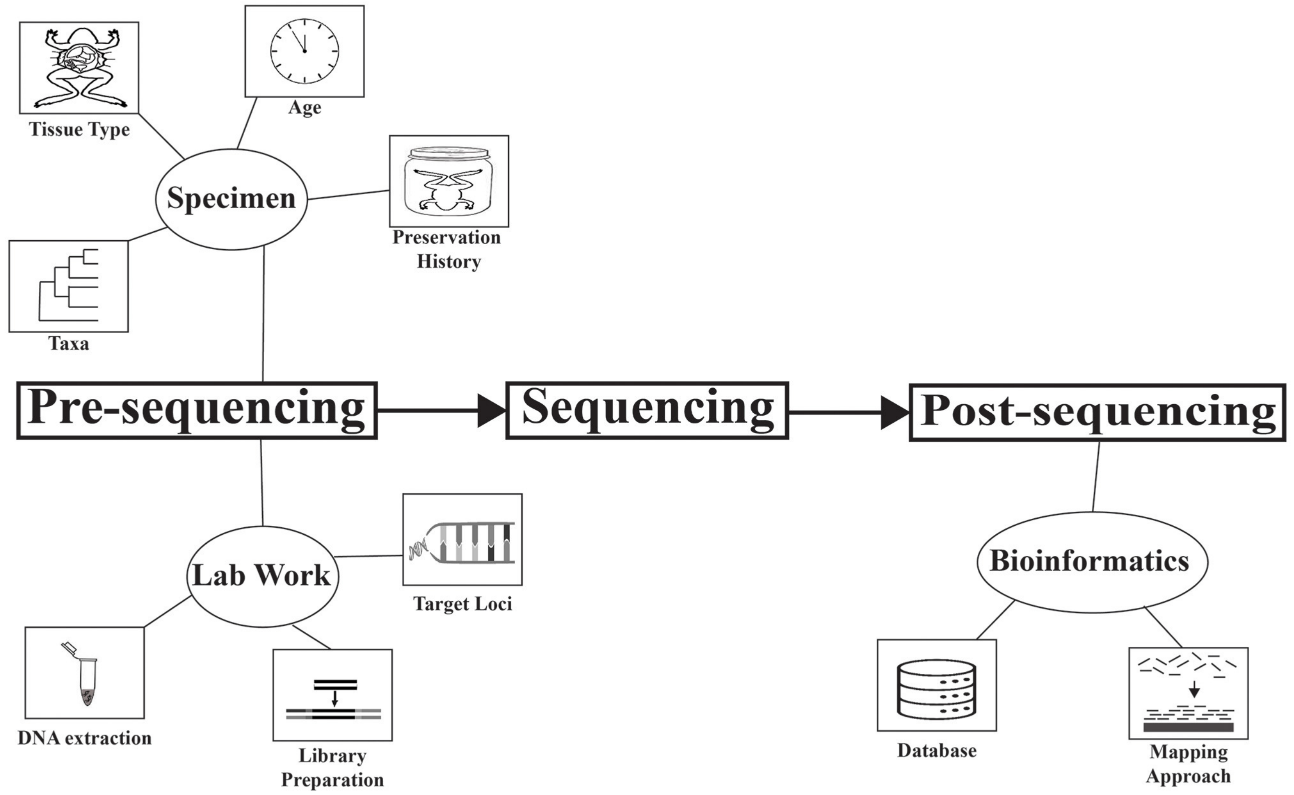


FIGURE 1
Factors that influence the data-quality and success of museomic studies, addressed in this Research Topic. Factors are organized temporally in the research process: pre-sequencing and post-sequencing.

For target loci, four major approaches were used—target capture, barcoding, shotgun sequencing, and cDNA sequencing. The approach used was largely determined by the objective of the study. One common theme is that the loci targeted are short in length, due to the tendency of DNA to fragment over time in historical and ancient tissues.

For DNA extraction, Hawkins et al. compared four methods (spin column, spin column with aDNA modifications, magnetic beads, and phenol chloroform) and found that the spin column and phenol chloroform methods outperformed magnetic beads. The spin column with aDNA modifications retained smaller fragments but took more time and was more expensive. Taking into consideration performance, cost, time, and toxicity, they recommended the spin column method.

For library preparation, Roycroft et al. compared the performance of single and dual barcoded library indexing strategies. They found that sequencing performance was better with dual barcoded libraries, having more reads and lower heterozygosity (=less cross contamination) compared to single barcoded libraries.

Post-sequencing

Post-sequencing factors addressed in these studies are related to the bioinformatic approaches.

Bioinformatic approaches

Two bioinformatic approaches were addressed in these studies: database and mapping approach.

Databases are important in genetic studies, especially when identifying an unknown sample or determining its evolutionary relationship with other taxa. Existing data in a database may affect the resolution of genetic analyses. Nakazato and Jinbo compared two commonly used DNA databases (GenBank and BOLD) and found that data for barcode loci are not the same in each database, despite each database importing from each other. This finding highlights the need of researchers to cross reference databases for relevant data.

To identify the genetic location of sequence reads and compare homologous loci, a mapping approach can be used. Erroneous read mapping can impact the results of a population genetics study, such as estimation of selection or genetic parameters. Roycroft et al. compared the effect of two different mapping approaches (sample-specific historical *de novo* assembly vs. high-quality "closest sister" *de novo* assembly) and found that data quality was better when mapping to a high-quality "closest sister" *de novo* assembly.

Other specimen-based research

Lastly, we note one study that in the strict sense may not qualify as "museomics", since it is not a genetic study. Balmaki et al.

studied plant-pollinator relationships by preparing pollen slides, taking photographs, and using an artificial neural network to help in identification. This approach, compared to metabarcoding, had greater resolution when identifying plant species. We include this study in the Research Topic because it exemplifies the spirit of developing novel research uses of specimens.

Conclusion

In the early 1900s, natural history museums were recognized as an "indispensable feature of modern civilization" due to the growing public interest in nature, their recognition of evolutionary trends in nature, and concerns regarding disappearing biodiversity (Farrington, 1915). Despite their popularity and importance (Allmon, 1994; Suarez and Tsutsui, 2004), NHCs are currently facing a survival crisis of their own due to shrinking budgets (Dalton, 2003; Gropp, 2004; Pennisi, 2020). To survive, NHCs need to find creative ways to publicize and acknowledge the usefulness of specimens and their data (Schindel and Cook, 2018; Miller et al., 2020; National Academies of Sciences Engineering and Medicine, 2020). Some ideas proposed are to develop an "extended specimen network" digitizing and linking all associated data to a specimen (Lendemer et al., 2020) and to recognize NHCs as coauthors on research articles (Rouhan et al., 2017). We are heartened to see museomics helping to expand interest in specimen-based research while showcasing the importance of natural history collections, and we look forward to seeing how newly developed technologies are used to study existing specimens.

Author contributions

All authors listed have made a substantial, direct, and intellectual contribution to the work and approved it for publication.

Acknowledgments

We would like to thank all the contributors to this Research Topic.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Allmon, W. D. (1994). The value of natural history collections. *Curator* 37, 83–89. doi: 10.1111/j.2151-6952.1994.tb01011.x

Cook, J. A., Arai, S., Armien, B., Bates, J., Bonilla, C. A. C., Cortez, M. B. D. S., et al. (2020). Integrating biodiversity infrastructure into pathogen discovery and mitigation of emerging infectious diseases. *Bioscience* 70, 531–534. doi: 10.1093/biosci/biaa064

Dalton, R. (2003). Natural history collections in crisis as funding is slashed. *Nature* 423, 575. doi: 10.1038/423575a

Ellwood, E. R., Sessa, J. A., Abraham, J. K., Budden, A. E., Douglas, N., Guralnick, R., et al. (2020). Biodiversity science and the twenty-first century workforce. *Bioscience* 70, 119–121. doi: 10.1093/biosci/biz147

Farrington, O. C. (1915). The rise of natural history museums. *Science* 42, 197–208. doi: 10.1126/science.42.1076.197

Freelance, C. B., Magrath, M. J. L., Elgar, M. A., and Wong, B. B. M. (2022). Long-term captivity is associated with changes to sensory organ morphology in a critically endangered insect. *J. Appl. Ecol.* 59, 504–513. doi: 10.1111/1365-2664.14069

Gilbert, M. T. P., Tomsho, L. P., Rendulic, S., Packard, M., Drautz, D. I., Sher, A., et al. (2008). Whole-genome shotgun sequencing of mitochondria from ancient hair shafts. *Science* 317, 1927–1930. doi: 10.1126/science.1146971

Gropp, R. E. (2004). Budget cuts affecting natural history. *Science* 306, 811. doi: 10.1126/science.306.5697.811b

Heberling, J. M., Prather, L. A., and Tonsor, S. J. (2019). The changing uses of herbarium data in an era of global change: an overview using automated content analysis. *Bioscience* 69, 812–822. doi: 10.1093/biosci/biz094

Hickey, J. J., and Anderson, D. W. (1968). Chlorinated hydrocarbons and eggshell changes in raptorial and fish-eating birds. *Science* 162, 271–273. doi: 10.1126/science.162.3850.271

Irestedt, M., Thorn, F., Muller, I. A., Jonsson, K. A., Ericson, E. G. P., and Blom, M. P. K. (2022). A guide to avian museomics: insights gained from resequencing hundreds of avian study skins. *Mol. Ecol. Resour.* 22, 2672–2684. doi: 10.1111/1755-0998.13660

Lendemer, J., Thiers, B., Monfils, A. K., Zaspel, J., Ellwood, E. R., Bentley, A., et al. (2020). The extended specimen network: a strategy to enhance US biodiversity collections, promote research and education. *Bioscience* 70, 23–30. doi: 10.1093/biosci/biz140

Miller, S. E., Barrow, L. N., Ehlman, S. M., Goodheart, J. A., Greiman, S. E., Lutz, H. L., et al. (2020). Building natural history collections for the twenty-first century and beyond. *Bioscience* 70, 674–687. doi: 10.1093/biosci/biaa069

Miller, W., Drautz, D. I., Janecka, J. E., Lesk, A. M., Ratan, A., Tomsho, L. P., et al. (2009). The mitochondrial genome sequence of the Tasmanian tiger (Thylacinus cynocephalus). *Genome Res.* 19, 213–220. doi: 10.1101/gr.082628.108

Moritz, C., Patton, J. L., Conroy, C. J., Parra, J. L., White, G. C., and Beissenger, S. R. (2008). Impact of a century of climate change on small-mammal communities in Yosemite National Park, U. S. A. *Science* 322, 261–264. doi: 10.1126/science.1163428

National Academies of Sciences Engineering and Medicine (2020). *Biological Collections: Ensuring Critical Research and Education for the 21st Century*. Washington, DC: The National Academies Press.

Pennisi, E. (2020). Shuttered natural history museums fight for survival. *Science* 368, 1042–1043. doi: 10.1126/science.368.6495.1042

Ratcliffe, D. A. (1967). Decrease in eggshell weight in certain birds of prey. *Nature* 215, 208–210. doi: 10.1038/215208a0

Raxworthy, C. J., and Smith, B. T. (2021). Mining museums for historical DNA: advances and challenges in museomics. *Trends Ecol. Evol.* 36, 1049–1060. doi: 10.1016/j.tree.2021.07.009

Rouhan, G., Dorr, L. J., Gauthier, L., Clerc, P., Muller, S., and Gaudeul, M. (2017). The time has come for natural history collections to claim co-authorship of research articles. *Taxon* 66, 1014–1016. doi: 10.12705/665.2

Schindel, D. E., and Cook, J. A. (2018). The next generation of natural history collections. *PLoS Biol.* 16, e2006125. doi: 10.1371/journal.pbio.2006125

Soberon, J. (1999). Linking biodiversity information sources. *TREE* 14, 291. doi: 10.1016/S0169-5347(99)01617-1

Suarez, A. V., and Tsutsui, N. D. (2004). The value of museum collections for research and society. *Bioscience* 54, 67–74. doi: 10.1641/0006-3568(2004)054(0066:TVOMCF)2.0.CO;2

BRIEF RESEARCH REPORT
published: 23 May 2022
doi: 10.3389/fevo.2022.907889

Poor hDNA-Derived NGS Data May Provide Sufficient Phylogenetic Information of Potentially Extinct Taxa

Catharina Clewing[1]‡, Christian Kehlmaier[2]‡, Björn Stelbrink[1]†, Christian Albrecht[1] and Thomas Wilke[1]*

[1] Department of Animal Ecology and Systematics, Justus Liebig University Giessen, Giessen, Germany, [2] Museum of Zoology, Senckenberg Natural History Collections Dresden, Dresden, Germany

OPEN ACCESS

Edited by:
Jonathan J. Fong,
Lingnan University, China

Reviewed by:
Parin Jirapatrasilp,
Chulalongkorn University, Thailand
Leila Belén Guzmán,
Instituto de Biología Subtropical (CONICET-UNaM), Argentina

***Correspondence:**
Catharina Clewing
catharina.clewing@allzool.bio.uni-giessen.de

†Present address:
Björn Stelbrink,
Museum für Naturkunde – Leibniz Institute for Evolution and Biodiversity Science, Berlin, Germany

‡These authors share first authorship

Specialty section:
This article was submitted to Phylogenetics, Phylogenomics, and Systematics, a section of the journal Frontiers in Ecology and Evolution

Received: *30 March 2022*
Accepted: *04 May 2022*
Published: *23 May 2022*

Citation:
Clewing C, Kehlmaier C, Stelbrink B, Albrecht C and Wilke T (2022) Poor hDNA-Derived NGS Data May Provide Sufficient Phylogenetic Information of Potentially Extinct Taxa. Front. Ecol. Evol. 10:907889. doi: 10.3389/fevo.2022.907889

Museum material is an important source of metadata for past and recent biological events. With current sequencing technologies, it is possible to obtain historical DNA (hDNA) from older material and/or endangered species to answer taxonomic, systematic, and biogeographical questions. However, hDNA from museum collections is often highly degraded, making it difficult to assess relationships at or above the species level. We therefore studied two probably extinct gastropod species of the genus *Laevicaspia*, which were collected ∼140 years ago in the Caspian Sea, to map "standard" mitochondrial and nuclear markers and assess both the sequencing depth and the proportion of ambiguous sites as an indicator for the phylogenetic quality of the NGS data. Our study resulted in the first phylogenetically informative mitochondrial and nuclear markers for *L. caspia*. Assessment of both sequencing depth (mean coverage) and proportion of ambiguous sites suggests that our assembled consensus sequences are reliable for this species. In contrast, no informative gastropod-specific DNA was obtained for *L. conus*, likely due to a high degree of tissue digestion and contamination with non-gastropod DNA. Nevertheless, our results show that hDNA may in principle yield high-quality sequences for species-level phylogenetic analyses, which underlines the importance of museum collections as valuable archives of the biological past.

Keywords: historical DNA, museomics, Gastropoda, Caspian Sea, mapping, mitochondrial makers, nuclear markers

INTRODUCTION

Biological collections in museums represent archives of the recent and remote past, providing a variety of metadata that allow to address a wide range of research questions (e.g., Bakker et al., 2020; Miralles et al., 2020). In recent years, advances in molecular technology have enabled access to valuable genetic and genomic resources from both comparatively old ethanol- and formalin-fixed or dry materials (Bi et al., 2013; Hykin et al., 2015; Ruane and Austin, 2017; Derkarabetian et al., 2019; Kehlmaier et al., 2020; Card et al., 2021; Ernst et al., 2021; Orlando et al., 2021; Raxworthy and Smith, 2021). DNA from museum materials is often highly degraded (i.e., represented as ultrashort fragments) and sometimes cross-linked with proteins or other DNA fragments and thus difficult to access (see e.g., Card et al., 2021; Orlando et al., 2021; Raxworthy and Smith, 2021). Moreover, the corresponding DNA sequences may contain a high number of read errors, which usually makes population-level analyses infeasible. However, even small amounts of genetic (and genomic) information can still be valuable when placing individual species in a phylogenetic context (e.g., Guschanski et al., 2013; Fabre et al., 2014). This is of particular importance when the taxon of interest has gone extinct in the wild and/or its habitat is no longer accessible.

A prime example is the endemic Pontocaspian molluscan fauna that evolved in the Caspian Sea, the Black Sea, and the Aral Sea region. It has suffered from major anthropogenic disturbances since the mid-twentieth century and is facing a severe biodiversity crisis (Wesselingh et al., 2019). A large share of the c. 55–99 endemic species (see Wesselingh et al., 2019; Gogaladze et al., 2021) declined in abundance or completely vanished in the course of human activities in the last century, and have been replaced by invasive species. This affected both relatively large and highly abundant species such as the Caspian bivalves *Dreissena caspia* and *D. elata*, but also microgastropod species with restricted ranges such as *Laevicaspia* spp. (Hydrobiidae, Pyrgulinae). The latter genus comprises a total of 12 species, of which 10 are endemic to the Caspian Sea and 2 to the Black Sea (Wesselingh et al., 2019). However, with the exception of *L. lincta* from the Black Sea (Wilke et al., 2007), none of these species have been found alive recently and are thus only known from the fossil record and older museum materials (Gogaladze et al., 2021).

The lack of comparative genetic data not only complicates taxonomic decisions. More importantly, it makes the reconstruction of biogeographic patterns and evolutionary processes—such as the timing and causes of faunal separation between the Black Sea and Caspian Sea taxa—very difficult. Given the lack of recent material for these tiny species from the Caspian Sea, the question arises whether degraded historical DNA (hDNA; Raxworthy and Smith, 2021) from old museum collections is of sufficient quality to assess relationships at or above the species level. Mollusks might be particularly problematic as their soft bodies are typically rich in mucopolysaccharides, which hamper DNA isolation (Jaksch et al., 2016; Adema, 2021).

In this study, we therefore subjected two ∼140-year-old museum specimens of *Laevicaspia* from the Caspian Sea, *L. caspia* (Eichwald, 1838) and *L. conus* (Eichwald, 1838), to next-generation sequencing (NGS) protocols, which were developed for ancient and heavily degraded DNA. Specifically, we aimed to (i) map "standard" mitochondrial and nuclear markers from quality-filtered reads that are frequently used for taxonomic assignments and (ii) evaluate whether the quality of the NGS data is sufficient to establish reliable DNA barcode references and thus to provide robust phylogenetic information of potentially extinct taxa.

MATERIALS AND METHODS

Materials

The ∼140-year-old specimens of *Laevicaspia caspia* and *L. conus* (Hydrobiidae, Pyrgulinae) were provided by the Zoological Institute of Russian Academy of Science (ZIN RAS), St. Petersburg, Russia (lot no. 4387/5 and 4614/4, respectively). *Laevicaspia caspia* was collected by O.A. Grimm in the Caspian Sea, ∼20 km off the eastern coast of Kazakhstan at a depth of ∼74 m (coordinates 43.28°N/51.05°E) on 9 July 1876. The individual of *L. conus* was collected by O.A. Grimm in the Caspian Sea, offshore near the city Baku at a depth of ∼11 m (geographical coordinates are not available) on 10 July 1874. In recent years, both specimens were stored in ethanol. However, it is not known in which fixative the individuals were originally preserved.

Genomic DNA was extracted from c. 3 mm^3 of soft tissue using the GEN-IAL All-tissue DNA-Kit (GEN-IAL GmbH, Troisdorf, Germany) basic protocol for forensic material. The final DNA pellet was dissolved in 50 μL TE buffer. DNA concentration and average fragment length were measured with a Qubit Fluorometer High Sensitivity assay kit (Invitrogen, Carlsbad, CA, United States) and a TapeStation High Sensitivity D1000 assay kit (Agilent, Santa Clara, CA, United States), respectively (**Supplementary Figures 1,2**). A final amount of 12.9 ng (*L. caspia*) and less than 0.2 ng (*L. conus*) of extracted DNA with average fragment lengths between 50 and 75 bp were converted into single-indexed, single-stranded Illumina sequencing libraries (see Gansauge and Meyer, 2013; Korlević et al., 2015), including the removal of uracil residues by uracil-DNA glycosylase (UDG) treatment. An Illumina MiSeq platform (Illumina, San Diego, CA, United States) housed at the Senckenberg Natural History Collections Dresden (Germany) was used for shotgun sequencing (75 bp paired-end reads), with each specimen being processed in its own private sequencing run.

Quality Control and Data Preparation

Raw reads were quality-checked and filtered using a previously established analytical pipeline (see Kehlmaier et al., 2017, 2019; Stelbrink et al., 2019). Adapters were trimmed with Skewer version 0.2.2 (Jiang et al., 2014), reads were merged (minimum length = 35 bp), filtered for quality (minimum Q-score = 20, corresponding to a base call accuracy of 99%), and duplicates were removed using BBMap version 37.24[1] (Bushnell, 2014). Per base sequence quality (i.e., base call accuracy) and read length distribution of trimmed (but unmerged) reads was analyzed and visualized using FastQC 0.11.9.[2]

Genomic Analysis

For the mitogenome assembly (see **Table 1**), the filtered reads (reduced readpool) were mapped against eight gastropod mitogenomes using Geneious Prime version 2021.1.1.[3] Because no mitogenome is publicly available for the family

[1] https://sourceforge.net/projects/bbmap

[2] http://www.bioinformatics.babraham.ac.uk/projects/fastqc

[3] https://www.geneious.com

TABLE 1 | Mitogenome mapping for *L. caspia* and *L. conus*.

Reference taxon (GenBank acc. no.)	Source	Reference length (bp)	Assembled reads *L. caspia* | *L. conus*	Coverage of reference sequence *L. caspia* | *L. conus*	Maximum coverage *L. caspia* | *L. conus*	Mean coverage *L. caspia* | *L. conus*
Bithynia leachii (MT410857)	Direct submission (DNAmark project)	15,682	2,918 | 95	39.6% | 3.1%	1,437 | 77	8.0 | 0.3
Caecum sp. (MT877093)	Sevigny et al., 2021	15,398	110 | 232	3.6% | 0.9%	80 | 223	0.3 | 0.6
Oncomelania h. hupensis (NC_012899)	Direct submission (NCBI genome project)	15,186	10,548 | 5	35.3% | 1.7%	5,109 | 2	157.4 | <0.1
Oncomelania h. robertsoni (NC_013187)	Direct submission (NCBI genome project)	15,191	913 | 10	38.2% | 2.2%	451 | 3	2.9 | <0.1
Potamopyrgus antipodarum (MG979468)	Sharbrough et al., 2018	15,149	3,126 | 5	39.3% | 1.4%	1,208 | 2	9.8 | <0.1
Potamopyrgus estuarinus (GQ996415)	Neiman et al., 2010	15,120	7,041 | 3	43.3% | 1.0%	5,175 | 2	21.0 | <0.1
Stenothyra glabra (MN548735)	Qi et al., 2020	15,830	5,341 | 301	41.6% | 2.9%	3,293 | 204	16.8 | 1.0
Tricula hortensis (NC_013833)	Direct submission (NCBI genome project)	15,179	519 | 10	45.1% | 2.6%	15 | 2	1.8 | <0.1

Hydrobiidae, we chose the following representatives of the superfamily Truncatelloidea: (1) *Bithynia leachii* (Bithyniidae; GenBank acc. no. MT410857; N/A = locality unknown), (2) *Caecum* sp. (Caecidae; MT877093; Belize), (3) *Oncomelania hupensis hupensis* (Pomatiopsidae; NC_012899; China), (4) *O. h. robertsoni* (Pomatiopsidae; NC_013187; China), (5) *Potamopyrgus antipodarum* (Tateidae; MG979468; New Zealand), (6) *Potamopyrgus estuarinus* (Tateidae; GQ996415; N/A), (7) *Stenothyra glabra* (Stenothyridae; MN548735; China), and (8) *Tricula hortensis* (Pomatiopsidae; NC_013833; China). Geneious Prime settings used for the mitogenome mapping were: sensitivity = medium-low sensitivity/fast; 5 iterations; annotation similarity = 25%. Finally, the consensus sequence was generated using the default settings (threshold for highest quality = 60%; call Sanger heterozygotes > 50%).

In addition, single-gene mapping was performed (see **Table 2**) against standard genetic markers used for phylogenetic analyses (see phylogenies of truncatelloids of Wilke et al., 2013; Delicado et al., 2019; Layton et al., 2019). Overall, we focused on the following three mitochondrial and five nuclear gene fragments: (1) mitochondrial cytochrome *c* oxidase subunit I (COI), (2) mitochondrial small subunit ribosomal RNA (SSU rRNA, 12S), mitochondrial large subunit ribosomal RNA (LSU rRNA, 16S), (4) nuclear small subunit ribosomal RNA (SSU rRNA, 18S), (5) nuclear large subunit ribosomal RNA (LSU rRNA, 28S), (6) nuclear internal transcribed spacer 1 (ITS1), (7) nuclear internal transcribed spacer 2 (ITS2), and (8) nuclear histone 3 (H3). For the selection of gene fragments, we chose those seed reference sequences that were as closely related as possible, depending on the availability in GenBank (e.g., for COI, *Laevicaspia lincta* from the Azov Sea in Russia was selected; see **Table 2**). Settings for the single-gene fragment mapping in Geneious Prime were as follows: sensitivity = medium-low sensitivity/fast; 5 iterations. The consensus sequences were generated using the following settings: threshold for highest quality = 60%; call "N" if coverage < 2; call Sanger heterozygotes > 50%. Ambiguous sites (i.e., "N") at the beginning and end of each sequence were removed afterward (see **Table 2** for trimmed sequence lengths).

Phylogenetic Analysis

In order to place these two species in a phylogenetic context, we compiled a reduced multigene dataset (COI, 16S, and 18S) from Wilke et al. (2007). The dataset included *Hydrobia acuta* (Hydrobiinae; France; GenBank acc. no.: AF278808, AY222659, AF367680) and *Pseudamnicola lucensis* (Pseudamnicolinae; Italy; AF367651, AF478394, AF367687) as outgroup and the following taxa belonging to the Pyrgulinae: *Dianella thiesseana* (Greece; AY676127, AY676121, AY676125), *Falsipyrgula pfeiferi* (Turkey; EF379296, EF379312, EF379283), *Laevicaspia lincta* (=*Euxinipyrgula milachevitchi*; Russia; EF379290, EF379306, EF379280), *Laevicaspia lincta* (=*Turricaspia* sp.; Ukraine; EF379294, EF379310, EF379282), *Laevicaspia lincta* (=*Micromelania lincta*; Romania; EF379292, EF379308, EF379281), *Ohridopyrgula macedonica* (North Macedonia; EF379287, EF379302, EF379278), *Pyrgula annulata* (Italy; AY341258, AY676122, AY676124), and *Xestopyrgula dybowskii* (North Macedonia; EF379289, EF379304, EF379279). The 16S and 18S partitions were aligned with the MAFFT web service (Katoh and Toh, 2008; Katoh and Standley, 2013) with default settings, and best-fit substitution models for each partition were selected using jModelTest 2.1.4 (Darriba et al., 2012). Bayesian inference (BI) was performed as implemented in MrBayes 3.2.6 (Ronquist et al., 2012), with two independent MCMC searches running for 1,000,000 generations and sampling each 500th tree. A burn-in of 50% was applied *a posteriori*.

RESULTS

Quality of Reads

A total of 37,339,378 (*L. caspia*) and 39,219,244 (*L. conus*) raw reads (read pairs) was generated in the two sequencing runs. The per base sequence quality was comparatively high for both untrimmed and trimmed reads. However, because the majority of

TABLE 2 | Overview of achieved gene fragments for *Laevicaspia caspia* and *L. conus* (for the latter, only the mapping results are shown).

				Mapping *L. caspia* | *L. conus*			**Consensus sequence *L. caspia***			
Gene fragment	**Gene code**	**Reference taxon (GenBank acc. no.)**	**Source**	**Assembled reads**	**Maximum coverage**	**Mean coverage**	**Achieved sequence length (bp)**	**Trimmed length (bp)**	**% N**	**GenBank accession #**
I. Mitochondrial gene fragments										
Cytochrome *c* oxidase subunit I	COI	*Laevicaspia lincta** (EF379290)	Wilke et al., 2007	44 | –	6 | –	3.0 | –	750	723	0.97%	ON365469
Small subunit ribosomal RNA (SSU rRNA)	12S	*Pyrgula annulata* (AF445350)	Hausdorf et al., 2003	51 | –	11 | –	4.8 | –	645	586	0.17%	ON362239
Large subunit ribosomal RNA (LSU rRNA)	16S	*Laevicaspia lincta** (EF379306)	Wilke et al., 2007	70 | –	9 | –	5.1 | –	766	736	0.14%	ON362224
II. Nuclear gene fragments										
Small subunit ribosomal RNA (SSU rRNA)	18S	*Laevicaspia lincta** (EF379280)	Wilke et al., 2007	1,293 | 92	105 | 24	57.4 | 8.8	1,070	1,021	0.10%	ON362237
Large subunit ribosomal RNA (LSU rRNA)	28S	*Hydrobia acuta* (KC110011)	Criscione and Ponder, 2013	1,948 | 115	94 | 18	45.3 | 4.2	1,873	1,840	0.00%	ON362238
Internal transcribed spacer 2	ITS2	*Pyrgula annulata* (MT594179)	Stelbrink et al., 2020	833 | –	95 | –	68.7 | –	717	697	0.00%	ON362234
Histone 3	H3	*Belgrandiella krupensis* (MG551341)	Osikowski et al., 2018	130 | –	22 | –	13.8 | –	455	328	0.00%	ON377370

**Note that this species was originally identified as Euxinipyrgula milachevitchi in Wilke et al. (2007), however, it has recently been synonymized with L. lincta (see Wesselingh et al., 2019).*

trimmed reads was very short, i.e., ≤ 35 bp (c. 69.4% for *L. caspia* and 52.2% for *L. conus*; **Figure 1**), only c. 20.0% (*L. caspia*) and 45.1% (*L. conus*) of the read pairs could be joined in BBMap. After quality filtering, a total number of 6,036,414 (*L. caspia*) and 5,282,339 (*L. conus*) reads and thus only c. 16.2% (*L. caspia*) and 13.5% (*L. conus*) of the total reads sequenced could be used for subsequent analyses.

Mitogenome Mapping

Eight truncatelloid mitogenomes were used to map the reduced readpool of *L. caspia* and *L. conus*. For *L. caspia*, the highest mean coverage (157.4) and second-highest maximum coverage (5,109) was obtained using the mitogenome data of *Oncomelania hupensis hupensis* (NC_012899; China) as seed reference (see **Figure 2** and **Table 1**). Thereby, 10,548 reads from the reduced readpool could be assembled, covering 35.3% of the reference sequence and parts of the following five genes: COI (cytochrome *c* oxidase subunit 1; 84% similarity), 12S (small subunit rRNA), 18S (large subunit rRNA), ND2 (NADH-ubiquinone oxidoreductase chain 2), and ATP8 (ATP synthase protein 8). Neither the number nor the coverage of mapped tRNAs were examined here, although they were also found by the mapping algorithm. A similar number of genes was obtained when the reduced readpool was mapped against *Tricula hortensis* from China (NC_013833; see **Table 1**). For the *Oncomelania hupensis hupensis* mapping, the high coverage was, however, mainly due to an overrepresentation of mapped reads against ND2 starting at position 15,016. When this 778 bp-long fragment was removed, maximum and mean coverages were considerably lower (32 and 1.5, respectively; **Figure 2**). The lowest mean coverage (0.3), as well as coverage of the reference sequence (3.6%), was obtained with the mitogenome data of *Caecum* sp. (MT877093; Belize). For all other selected reference mitogenomes, mean and maximum coverage ranged from 1.8–21.0 to 15–5,175, respectively. This was sometimes a result of overrepresented mapped genes such as 16S, cyt *b*, and ND2. The coverage of these reference sequences was between 38.2 and 45.1% (for details see **Table 1**). For *L. conus*, considerably fewer reads (5–301) were mapped against all mitogenomes selected (**Table 1**). We therefore did not analyze these results in detail.

Single-Gene Mapping

All selected “standard” genetic markers used for molecular phylogenies of truncatelloids could be successfully mapped using the reduced readpool of *L. caspia* (see **Table 2**). Mean and maximum coverage of the three mitochondrial markers (COI, 12S, and 16S) ranged from 3.0–5.1 to 6–11, respectively. Mean and maximum coverage of the four nuclear gene fragments (18S, 28S, ITS2, and H3) was considerably higher with values ranging from 13.8–68.7 to 22–105, respectively. The

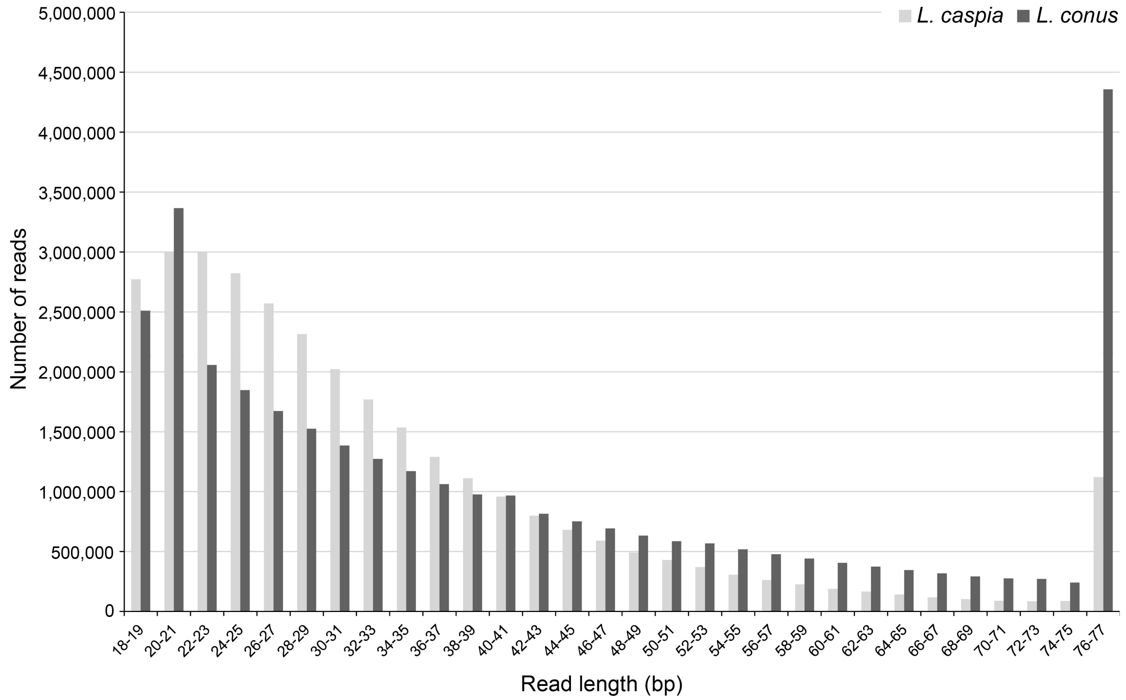


FIGURE 1 | Fragment length distribution (in bp) of trimmed and quality-filtered reads (reduced readpool) of both *Laevicaspia* species.

proportion of ambiguous sites ("N") in the trimmed consensus sequence was used as an additional quality measure of the respective gene fragment. Thereby, the mitochondrial markers showed a generally higher N-content (0.14–0.97%) compared to the nuclear markers (0.00–0.10%), with COI having the highest (0.97%) and 28S, ITS2, and H3 having the lowest values (0.00%).

In contrast, the single-gene mapping was not successful for *L. conus*, similar to the mitogenome mapping (see above). Accordingly, only 18S and 28S could be mapped, though with a very low number of assembled reads (92 and 115, respectively; see **Table 2**). We therefore did not analyze these mapping results further. However, we applied a megablast search (as implemented in Geneious Prime; settings: nr/nt, maximum hits = 1) to the reduced readpool for fragments >100 bp (N = 365,445). Accordingly, 28,143 hits were found, of which 10,283 had a query coverage of 100%, i.e., a fragment length of 100 bp. In total, 1,490 unique organisms were found that mainly belong to bacteria (**Supplementary Figure 3**).

Phylogenetic Analysis

Due to the different mapping success, only sequence information from *L. caspia* could be used in the phylogenetic analyses. Accordingly, *L. caspia* from the Caspian Sea represents a genetically distinct lineage and forms a highly supported (Bayesian posterior probability, BPP = 1.00) clade within the Pyrgulinae, together with *Falsipyrgula pfeiferi* from Lake Egirdir (Turkey) and three individuals of *Laevicaspia lincta* sampled from different localities in the Black Sea basin (see **Supplementary Figure 4**).

DISCUSSION

Leveraging genomic resources from historical museum material is a promising tool for addressing research topics related to the fields of biodiversity, conservation, taxonomy, and systematics, particularly for species that are rare or even extinct. Depending on age, tissue amount, and condition of the museum material, and the quality of generated sequences, complete mitogenomes and various nuclear loci of interest may, in principle, be assembled from raw sequencing data (e.g., Raxworthy and Smith, 2021). However, such analyses might be problematic in mollusks due to their high mucopolysaccharide content (Jaksch et al., 2016; Adema, 2021). Here, we used ∼140-year-old hydrobiid microgastropod specimens of *Laevicaspia caspia* and *L. conus* to map "standard" mitochondrial and nuclear markers for taxonomic assignments. We further assessed both the sequencing depth (mean coverage) as well as the proportion of ambiguous sites as an indicator of the phylogenetic quality of the NGS data.

The main problem in generating genomic information for both *Laevicaspia* species was probably not the DNA isolation and sequencing itself, but the preservation condition of the source tissue. Despite the overall high per base sequence quality, the reduced readpool was dominated by a large share of short DNA fragments and thus a low number of merged reads. Therefore, it was not possible to assemble a complete or near-complete mitogenome, although a high-quality mitogenome

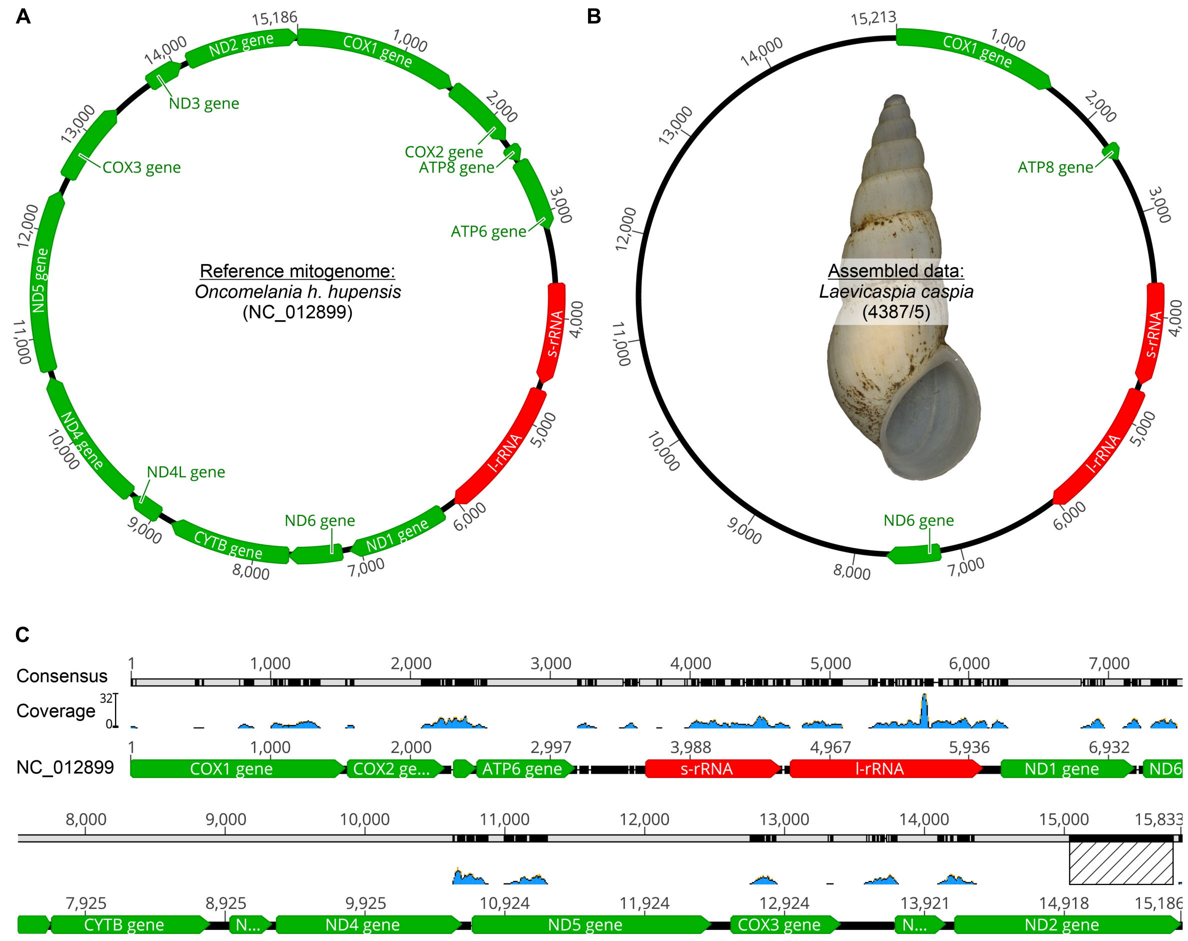


FIGURE 2 | Mitogenome mapping using Geneious Prime (version 2021.1.1). **(A)** Seed reference mitogenome (*O. h. hupensis*: GenBank acc. no. NC_012899) where the highest coverage could be achieved, **(B)** Assembled data for *Laevicaspia caspia* (10,548 out of 6,036,414 reads; maximum coverage = 5,109; mean coverage = 157.4, see also **Table 1**) including a shell image of the sequenced individual (shell height = c. 12.4 mm), **(C)** Illustration of mitogenome coverage. Note that the overrepresented 778 bp fragment of ND2 (marked with a hatched rectangle) has been removed (see text for details).

was previously generated for another ∼80-year-old freshwater gastropod specimen using the same laboratory pipeline in the same laboratory (see Stelbrink et al., 2019). However, applying our mitogenome and single-gene mapping approach, we were able to assemble taxonomically and phylogenetically informative mitochondrial and nuclear markers such as COI, 16S, and 18S (see e.g., Wilke et al., 2007), at least for *L. caspia* (**Supplementary Figure 4**). In contrast, virtually none of our mapping strategies were successful for *L. conus*. It is very likely that the specimen of this species was preserved under such poor conditions that the already small amount of tissue was too heavily digested and thus fragmented and further contaminated with non-gastropod DNA during decomposition (e.g., Raxworthy and Smith, 2021).

For assessing the reliability of the consensus sequences in *L. caspia*, we compared both the mean coverage as well as the proportion of ambiguous sites (see **Table 2**). The coverage of the mitochondrial target fragments was by an order of magnitude lower compared to the nuclear genes of interest and also considerably lower than the mean coverage of the previously published near-complete mitogenome of the paludomid gastropod *Pseudocleopatra dartevellei* (Stelbrink et al., 2019). We assume that this is related to the conditions under which the material was preserved. However, it would require a larger sequencing approach with several (fresh and old) samples to make a reliable statement. Similarly, given the higher mean coverage, the proportion of ambiguous sites was negligible for the nuclear fragments (<0.1%), whereas this ratio was higher, yet very low, for the mitochondrial markers (<1%). Overall, both factors—the moderate to high mean coverage together with the low amount of ambiguous sites—indicate that our assembled consensus sequences are reliable and can be used for taxonomic and phylogenetic purposes at the species level.

In summary, our pipeline using a set of single-gene and mitogenome seed reference sequences allowed us to map several phylogenetically relevant markers for *L. caspia*. These loci enabled us to provide the first DNA barcode sequences of this genus for the Caspian Sea. This will allow researchers to calculate genetic distances to other relatives, and to infer the phylogenetic position of this probably extinct species within the Pyrgulinae. Importantly, despite the relatively poor quality of our data, we here present information about an endangered ecosystem (e.g., Prange et al., 2020), whose endemic fauna is under increasing human pressure (e.g., Wesselingh et al., 2019).

DATA AVAILABILITY STATEMENT

The data presented in the study are deposited in the NCBI GenBank repository, accession numbers ON362224, ON362234, ON362237, ON362238, ON362239, ON365469, and ON377370.

AUTHOR CONTRIBUTIONS

CC analyzed the data, created the figures, and wrote the first draft of the manuscript. CK performed lab work and performed preliminary analyses. BS helped analyzing the data. CA and TW conceived the study. All authors contributed to drafting and reviewing the manuscript.

FUNDING

This study was funded by the German Research Foundation (DFG, Grant Nos. WI 1902-14 and WI 1902-17).

ACKNOWLEDGMENTS

We thank Vitaliy Anistratenko (I.I. Schmalhausen Institute of Zoology, Kiev, Ukraine), Dmitry M. Palatov (Zoological Institute of Russian Academy of Sciences, St. Petersburg, Russia), and Maxim V. Vinarski (Saint-Petersburg State University, St. Petersburg, Russia) for providing the *Laevicaspia* material. We thank two reviewers for their constructive comments on a previous version of the manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fevo.2022.907889/full#supplementary-material

REFERENCES

Adema, C. M. (2021). Sticky problems: extraction of nucleic acids from molluscs. *Philos. Trans. R. Soc. Lond. B* 376, 20200162. doi: 10.1098/rstb.2020.0162

Bakker, F. T., Antonelli, A., Clarke, J., Cook, J. A., Edwards, S. V., Faurby, S., et al. (2020). The Global Museum: natural history collections and the future of evolutionary biology and public education. *PeerJ* 8:e8225. doi: 10.7717/peerj.8225

Bi, K., Linderoth, T., Vanderpool, D., Good, J. M., Nielsen, R., and Moritz, C. (2013). Unlocking the vault: next-generation museum population genomics. *Mol. Ecol.* 22, 6018–6032. doi: 10.1111/mec.12516

Bushnell, B. (2014). *BBMap: A Fast, Accurate, Splice-Aware Aligner.* Berkeley, CA: Ernest Orlando Lawrence Berkeley National Laboratory.

Card, D. C., Shapiro, B., Giribet, G., Moritz, C., and Edwards, S. V. (2021). Museum genomics. *Annu. Rev. Genet.* 55, 633–659. doi: 10.1146/annurev-genet-071719-020506

Criscione, F., and Ponder, W. F. (2013). A phylogenetic analysis of rissooidean and cingulopsoidean families (Gastropoda: Caenogastropoda). *Mol. Phylogenet. Evol.* 66, 1075–1082. doi: 10.1016/j.ympev.2012.11.026

Darriba, D., Taboada, G. L., Doallo, R., and Posada, D. (2012). jModelTest 2: more models, new heuristics and parallel computing. *Nat. Methods* 9, 772–772. doi: 10.1038/nmeth.2109

Delicado, D., Arconada, B., Aguado, A., and Ramos, M. A. (2019). Multilocus phylogeny, species delimitation and biogeography of Iberian valvatiform springsnails (Caenogastropoda: Hydrobiidae), with the description of a new genus. *Zool. J. Linn. Soc.* 186, 892–914. doi: 10.1093/zoolinnean/zly093

Derkarabetian, S., Benavides, L. R., and Giribet, G. (2019). Sequence capture phylogenomics of historical ethanol-preserved museum specimens: unlocking the rest of the vault. *Mol. Ecol.* 19, 1531–1544. doi: 10.1111/1755-0998.13072

Eichwald, E. (1838). Faunae Caspii Maris primitiae. *Bull. Soc. Imp. Nat. Moscou* 11, 125–174.

Ernst, R., Kehlmaier, C., Baptista, N. L., Pinto, P. V., Branquima, M. F., Dewynter, M., et al. (2021). Filling the gaps: the mitogenomes of Afrotropical egg-guarding frogs based on historical type material and a re-assessment of the nomenclatural status of *Alexteroon* Perret, 1988 (Hyperoliidae). *Zool. Anz.* 293, 215–224. doi: 10.1016/j.jcz.2021.06.002

Fabre, P.-H., Vilstrup, J. T., Raghavan, M., Der Sarkissian, C., Willerslev, E., Douzery, E. J. P., et al. (2014). Rodents of the Caribbean: origin and diversification of hutias unravelled by next-generation museomics. *Biol. Lett.* 10:20140266. doi: 10.1098/rsbl.2014.0266

Gansauge, M. T., and Meyer, M. (2013). Single-stranded DNA library preparation for the sequencing of ancient or damaged DNA. *Nat. Protoc.* 8, 737–748. doi: 10.1038/nprot.2013.038

Gogaladze, A., Son, M. O., Lattuada, M., Anistratenko, V. V., Syomin, V. L., Pavel, A. B., et al. (2021). Decline of unique Pontocaspian biodiversity in the Black Sea Basin: a review. *Ecol. Evol.* 11, 12923–12947. doi: 10.1002/ece3.8022

Guschanski, K., Krause, J., Sawyer, S., Valente, L. M., Bailey, S., Finstermeier, K., et al. (2013). Next-generation museomics disentangles one of the largest primate radiations. *Syst. Biol.* 62, 539–554. doi: 10.1093/sysbio/syt018

Hausdorf, B., Röpstorf, P., and Riedel, F. (2003). Relationships and origin of endemic Lake Baikal gastropods (Caenogastropoda: Rissooidea) based on mitochondrial DNA sequences. *Mol. Phylogenet. Evol.* 26, 435–443. doi: 10.1016/s1055-7903(02)00365-2

Hykin, S. M., Bi, K., and McGuire, J. A. (2015). Fixing formalin: a method to recover genomic-scale DNA sequence data from formalin-fixed museum specimens using high-throughput sequencing. *PLoS One* 10:e0141579. doi: 10.1371/journal.pone.0141579

Jaksch, K., Eschner, A., von Rintelen, T., and Haring, E. (2016). DNA analysis of molluscs from a museum wet collection: a comparison of different extraction methods. *BMC Res. Notes* 9:348. doi: 10.1186/s13104-016-2147-7

Jiang, H., Lei, R., Ding, S. W., and Zhu, S. (2014). Skewer: a fast and accurate adapter trimmer for next-generation sequencing paired-end reads. *BMC Bioinformatics* 15:182. doi: 10.1186/1471-2105-15-182

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

Katoh, K., and Toh, H. (2008). Recent developments in the MAFFT multiple sequence alignment program. *Brief. Bioinform.* 9, 286–298. doi: 10.1093/bib/bbn013

Kehlmaier, C., Barlow, A., Hastings, A. K., Vamberger, M., Paijmans, J. L. A., Steadman, D. W., et al. (2017). Tropical ancient DNA reveals relationships of the extinct bahamian giant tortoise *Chelonoidis alburyorum*. *Proc. R. Soc. Lond. B* 284:20162235. doi: 10.1098/rspb.2016.2235

Kehlmaier, C., Graciá, E., Campbell, P. D., Hofmeyr, M. D., Schweiger, S., Martínez-Silvestre, A., et al. (2019). Ancient mitogenomics clarifies radiation of extinct Mascarene giant tortoises (*Cylindraspis* spp.). *Sci. Rep.* 9:17487. doi: 10.1038/s41598-019-54019-y

Kehlmaier, C., Zinenko, O., and Fritz, U. (2020). The enigmatic Crimean green lizard (*Lacerta viridis magnifica*) is extinct but not valid: mitogenomics of a 120-year-old museum specimen reveals historical introduction. *J. Zool. Syst. Evol. Res.* 58, 303–307. doi: 10.1111/jzs.12345

Korlević, P., Gerber, T., Gansauge, M., Hajdinjak, M., Nagel, S., Aximu-Petri, A., et al. (2015). Reducing microbial and human contamination in DNA extractions from ancient bones and teeth. *Biotechniques* 59, 87–93. doi: 10.2144/000114320

Layton, K. K. S., Middelfart, P. U., Tatarnic, N. J., and Wilson, N. G. (2019). Erecting a new family for *Spirostyliferina*, a truncatelloidean microgastropod, and further insights into truncatelloidean phylogeny. *Zool. Scr.* 48, 727–744. doi: 10.1111/zsc.12374

Miralles, A., Bruy, T., Wolcott, K., Scherz, M. D., Begerow, D., Beszteri, B., et al. (2020). Repositories for taxonomic data: where we are and what is missing. *Syst. Biol.* 69, 1231–1253. doi: 10.1093/sysbio/syaa026

Neiman, M., Hehman, G., Miller, J. T., Logsdon, J. M. Jr., and Taylor, D. R. (2010). Accelerated mutation accumulation in asexual lineages of a freshwater snail. *Mol. Biol. Evol.* 27, 954–963. doi: 10.1093/molbev/msp300

Orlando, L., Allaby, R., Skoglund, P., Der Sarkissian, C., Stockhammer, P. W., Ávila-Arcos, M. C., et al. (2021). Ancient DNA analysis. *Nat. Rev. Methods Prim.* 1:14. doi: 10.1038/s43586-020-00011-0

Osikowski, A., Hofman, S., Rysiewska, A., Sket, B., Prevorènik, S., and Falniowski, A. (2018). A case of biodiversity overestimation in the Balkan *Belgrandiella* A. J. Wagner, 1927 (Caenogastropoda: Hydrobiidae): molecular divergence not paralleled by high morphological variation. *J. Nat. Hist.* 52, 323–344. doi: 10.1080/00222933.2018.1424959

Prange, M., Wilke, T., and Wesselingh, F. P. (2020). The other side of sea level change. *Commun. Earth Environ.* 1:69. doi: 10.1038/s43247-020-00075-6

Qi, L., Kong, L., and Li, Q. (2020). Redescription of *Stenothyra glabra* A. Adam, 1861 (Truncatelloidea, Stenothyridae), with the first complete mitochondrial genome in the family Stenothyridae. *Zookeys* 991, 69–83. doi: 10.3897/zookeys.991.51408

Raxworthy, C. J., and Smith, B. T. (2021). Mining museums for historical DNA: advances and challenges in museomics. *Trends Ecol. Evol.* 36, 1049–1060. doi: 10.1016/j.tree.2021.07.009

Ronquist, F., Teslenko, M., van der Mark, P., Ayres, D. L., Darling, A., Höhna, S., et al. (2012). MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Syst. Biol.* 61, 539–542. doi: 10.1093/sysbio/sys029

Ruane, S., and Austin, C. C. (2017). Phylogenomics using formalin-fixed and 100+ year-old intractable natural history specimens. *Mol. Ecol. Resour.* 17, 1003–1008. doi: 10.1111/1755-0998.12655

Sevigny, J., Leasi, F., Simpson, S., Di Domenico, M., Jörger, K. M., Norenburg, J. L., et al. (2021). Target enrichment of metazoan mitochondrial DNA with hybridization capture probes. *Ecol. Indic.* 121:106973. doi: 10.1016/j.ecolind.2020.106973

Sharbrough, J., Luse, M., Boore, J. L., Logsdon, J. M. Jr., and Neiman, M. (2018). Radical amino acid mutations persist longer in the absence of sex. *Evolution* 72, 808–824. doi: 10.1111/evo.13465

Stelbrink, B., Kehlmaier, C., Wilke, T., and Albrecht, C. (2019). The near-complete mitogenome of the critically endangered *Pseudocleopatra dartevellei* (Caenogastropoda: Paludomidae) from the Congo River assembled from historical museum material. *Mitochondrial DNA B Resour.* 4, 3229–3231. doi: 10.1080/23802359.2019.1669081

Stelbrink, B., Wilke, T., and Albrecht, C. (2020). Ecological opportunity enabled invertebrate radiations in ancient Lake Ohrid. *J. Great Lakes Res.* 46, 1156–1161. doi: 10.1016/j.jglr.2020.06.012

Wesselingh, F. P., Neubauer, T. A., Anistratenko, V. V., Vinarski, M. V., Yanina, T., ter Poorten, J. J., et al. (2019). Mollusc species from the Pontocaspian region – an expert opinion list. *Zookeys* 827, 31–124. doi: 10.3897/zookeys.827.31365

Wilke, T., Albrecht, C., Anistratenko, V. V., Sahin, S. K., and Yildirim, Z. (2007). Testing biogeographical hypotheses in space and time: faunal relationships of the putative ancient Lake Egirdir in Asia Minor. *J. Biogeogr.* 34, 1807–1821. doi: 10.1111/j.1365-2699.2007.01727.x

Wilke, T., Haase, M., Hershler, R., Liu, H.-P., Misof, B., and Ponder, W. (2013). Pushing short DNA fragments to the limit: phylogenetic relationships of "hydrobioid" gastropods (Caenogastropoda: Rissooidea). *Mol. Phylogenet. Evol.* 66, 715–736. doi: 10.1016/j.ympev.2012.10.025

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

BRIEF RESEARCH REPORT
published: 17 June 2022
doi: 10.3389/fevo.2022.909846

Simultaneous Barcode Sequencing of Diverse Museum Collection Specimens Using a Mixed RNA Bait Set

Stefanie Agne[1]*, *Michaela Preick*[1], *Nicolas Straube*[2] *and Michael Hofreiter*[1]

[1] Evolutionary Adaptive Genomics, Institute for Biochemistry and Biology, Department of Mathematics and Natural Sciences, University of Potsdam, Potsdam, Germany, [2] Department of Natural History, University Museum of Bergen, University of Bergen, Bergen, Norway

A growing number of publications presenting results from sequencing natural history collection specimens reflect the importance of DNA sequence information from such samples. Ancient DNA extraction and library preparation methods in combination with target gene capture are a way of unlocking archival DNA, including from formalin-fixed wet-collection material. Here we report on an experiment, in which we used an RNA bait set containing baits from a wide taxonomic range of species for DNA hybridisation capture of nuclear and mitochondrial targets for analysing natural history collection specimens. The bait set used consists of 2,492 mitochondrial and 530 nuclear RNA baits and comprises specific barcode loci of diverse animal groups including both invertebrates and vertebrates. The baits allowed to capture DNA sequence information of target barcode loci from 84% of the 37 samples tested, with nuclear markers being captured more frequently and consensus sequences of these being more complete compared to mitochondrial markers. Samples from dry material had a higher rate of success than wet-collection specimens, although target sequence information could be captured from 50% of formalin-fixed samples. Our study illustrates how efforts to obtain barcode sequence information from natural history collection specimens may be combined and are a way of implementing barcoding inventories of scientific collection material.

Keywords: target capture, type specimens, molecular species identification, museum specimens, cross-species capture

OPEN ACCESS

Edited by:
Jonathan J. Fong,
Lingnan University, China

Reviewed by:
Daniel Janzen,
University of Pennsylvania, United States
Sachithanandam Veeraragavan,
National Centre for Sustainable Coastal Management, India
André Soares,
Uppsala University, Sweden

****Correspondence:***
Stefanie Agne
agne@uni-potsdam.de

Specialty section:
This article was submitted to Phylogenetics, Phylogenomics, and Systematics, a section of the journal Frontiers in Ecology and Evolution

Received: *31 March 2022*
Accepted: *17 May 2022*
Published: *17 June 2022*

Citation:
Agne S, Preick M, Straube N and Hofreiter M (2022) Simultaneous Barcode Sequencing of Diverse Museum Collection Specimens Using a Mixed RNA Bait Set. Front. Ecol. Evol. 10:909846. doi: 10.3389/fevo.2022.909846

INTRODUCTION

The growing interest in accessing DNA of natural history wet-collection specimens, which have long been recalcitrant regarding DNA analyses, is reflected in increasing numbers of publications reporting sequencing of this highly fragmented DNA (e.g., Lyra et al., 2020; Rancilhac et al., 2020; Scherz et al., 2020; Hahn et al., 2021; Straube et al., 2021a,b). Combining ancient DNA extraction methods, single stranded DNA library construction and short-read high throughput sequencing technology allows for obtaining DNA sequences of museum specimens at unprecedented scales (e.g., Hahn et al., 2021; Straube et al., 2021a). In taxonomy, unlocking DNA sequence information from rare and extinct species as well as type material is of particular interest. Numerous described species are only known from few, aged museum specimens and often re-collection efforts are hindered by several factors such as extensive sampling efforts, conservation concerns, politically instable situations in countries of origin or simply rareness of the species in question. However, rare species described from remote localities are of special concern in conservation, directing the attention to museum specimens as potential alternative DNA sources for taxonomic evaluation as basis for conservation efforts. Besides their undoubted importance for taxonomic research (e.g., Lyra et al., 2020; Rancilhac et al., 2020; Scherz et al., 2020; Straube et al., 2021b), type specimens may as well represent the only representatives of a rare or extinct species. Most of such specimens lack a phylogenetically close reference genome, but for taxonomy, barcode genes for species delimitation are generally sufficient as references for phylogenetic placement of species' haplotypes. In these circumstances, DNA sequences from type material can play a key role.

Ancient DNA methods have paved the way for accessing DNA sequence information from archival samples, including formalin-fixed wet-collection samples (Stiller et al., 2016; Gansauge et al., 2017; Straube et al., 2021a), even on the genome level (Hahn et al., 2021). These approaches are laborious and time consuming, however. As shown previously in Straube et al. (2021a), the level of target DNA in initial test-sequencing datasets may be low. Shotgun sequencing of such DNA libraries then becomes inefficient in terms of associated costs necessary to attain coverage levels allowing for reconstructing specific barcode loci. Target gene capture as alternative can be an additional costly and time-intensive step, especially when a second round of capture is performed which has been shown to increase sequencing success (e.g., Li et al., 2013, 2015; Templeton et al., 2013; Springer et al., 2015; Paijmans et al., 2016). In an effort to increase efficiency and decrease overall costs for target capture of sample specific barcode markers in museum specimens, we report here on the design and successful application of an RNA bait set targeting taxonomically useful barcode markers in a variety of natural history collection samples of different phyla. Undergoing this process, we also aim to detect factors that may have an impact on the capture success such as different target regions, tissue type, fixation history, and genetic distance between bait and target sequences.

MATERIALS AND METHODS

We obtained 37 samples including dried bone, teeth, and soft tissue samples as well as muscle and skin from wet-collection specimens. Representatives of the following classes were included: Demospongiae, Gastropoda, Polychaeta, Malacostraca, Insecta, Actinopterygii, Chondrichthyes, Amphibia, and Reptilia. The investigated samples range in age from 25 to 192 years (**Supplementary Table 1**). Along with the tissue samples, we obtained information on the samples using a standardised sample sheet (**Supplementary Table 2**). The requested information relates to the age, fixation, and preservation details as far as available, target barcode loci, bait sequences to capture specific barcode loci, reference genomes and taxonomic history of the sample. DNA was extracted from samples listed in **Supplementary Table 1** following the different DNA extraction treatments described in Straube et al. (2021a) based on the ancient DNA extraction protocol specified in Dabney et al. (2013) using a GuSCN based extraction buffer (Rohland et al., 2004). Subsequently, single stranded DNA libraries were prepared for each sample following the protocol by Gansauge et al. (2017). For obtaining information on the presence of target DNA, test-sequencing as described in Straube et al. (2021a) was performed. Independent of presence of endogenous DNA, target capture was subsequently performed for all samples to test if the limited information of the test-sequencing data may fail to detect endogenous DNA even though it is present in the DNA library.

For target capture of barcode loci, specific bait sequences and reference genomes provided partially by our collaborators, but mostly obtained from public resources (**Supplementary Table 3**) were sent to Arbor Biosciences® and split into a mitochondrial and a nuclear bait set. For both sets of sequences, 80 nt, 3x tiled baits were designed. While the mitochondrial baits were not further processed bioinformatically, the nuclear baits were filtered in two steps. First, baits were blasted to reference genomes from available most closely related species (**Supplementary Table 1**). Any bait that had blast hits to a region of the genome that was greater than 25% soft-masked for repeats was removed. The second filtering step was based on the number of bait hits and the predicted melting temperatures between the bait and those blast hits to detect the number of binding sites a bait may have, which ultimately resulted in the exclusion of 97 nuclear baits. A final set of 2,492 mitochondrial and 530 nuclear RNA baits was produced. Target capture was performed for each sample listed in **Supplementary Table 1** following the manufacturer's protocol for $N = 2$ samples. For the remaining 35 samples a target-gene enrichment protocol based on the Mybaits-manual-v3 was used, which is cost-reducing and requires less of RNA baits per sample compared to the recommended amount but maintaining the same level of target capture success (Huang et al., 2021). For both protocols, we used an in-solution hybridisation temperature of 65°C for 24 h. The capture was performed twice including a second amplification of libraries after the first round of target capture. Optimal number of amplification cycles was estimated for each library by performing a qPCR. DNA libraries were double-indexed during amplification and sequenced as described in Paijmans et al. (2017). Sequencing was performed on an Illumina Nextseq 500 sequencing platform, using 500/550 High Output v2.5 (75 cycles, Illumina 20024906) kits (75 bp single-end reads). All laboratory steps as well as sequencing was conducted in the molecular laboratories of the AG Hofreiter at the University of Potsdam. At least three million sequencing reads were targeted for each sample to gain sufficient coverage of target markers. Sequencing reads available after target capture underwent quality checking and trimming as in Straube et al. (2021a) and were subsequently used to reconstruct the target barcode loci using mapping and consensus sequence generation in BWA-ALN v.0.7.17 (Li and Durbin, 2009) and Bcftools v.1.9 (Li, 2011). We used either the bait sequences or phylogenetically closer reference sequences which became available after bait production (**Supplementary Table 4**). Afterwards, consensus sequences were analysed for phylogenetic position and classification.

We tested for correlation between the completeness of target genes after hybridisation capture and the phylogenetic distance of RNA bait sequences to target consensus sequences (p-distances). Therefore, each target consensus sequence was aligned to the appropriate bait sequence as listed in **Supplementary Table 1** using Mafft v.7.49 (Katoh et al., 2002) and resulting p-distances were calculated using MEGA v.11.0.10 (Kumar et al., 2016). If several bait sequences were available for aligning to a genetic locus of a species, the reference with the smallest p-distance to the consensus sequence was used. For correlation analysis, Pearson's correlation coefficient was calculated, and a *t*-test was performed. Specimens with too low endogenous DNA content to create a consensus sequence after target capture were not included in the analysis. We further tested for correlation between sequencing depth and completeness as described above.

RESULTS

After test-sequencing, we detected endogenous DNA in most of our samples (91.9%; **Supplementary Table 1**). For samples that showed no endogenous DNA after test sequencing, target capture attempts failed. Available sequencing data after target capture ranged from 391,964 to 12,195,369 raw reads and 91,101 to 10,611,372 reads after trimming. Trimmed reads including PCR duplicates that mapped to the reference sequences ranged between 0 and 69.23% (**Supplementary Table 4**). We were able to capture DNA sequence information of target barcode loci from 84% of our samples (**Figure 1**), 73.52% for mitochondrial and 94.28% for nuclear target genes, respectively. The completeness of all nuclear barcode loci is 85.15% and higher than that of the mitochondrial loci, the completeness of which is 72.25% (**Figure 1**). The best results in terms of consistency and sequence completeness were obtained from the crocodilian bone and dry skin samples with an average consensus sequence completeness of 98.31%. For wet-collection material we obtained sequence information for 86.2% of the target genes and an average sequence completeness of 71.74%. Similar differences are observed when comparing the different materials of the Demospongiae samples, with an average consensus sequence completeness of 55.02% for the wet collection tissues and 74.38% for the dried tissues, respectively. Three of the ten specimens for which formalin fixation is assumed resulted in target gene completeness above 75% (**Figure 1**). The Mollusca samples in particular showed a high target sequence completeness with an average of 96.12% in all five target loci tested.

The p-distance, defined as proportion of different nucleotides per total numbers of nucleotides compared, was on average 7.62% (range between 0 and 56.80%) and did not correlate with the target gene completeness (Pearson correlation coefficient: $r = -0.20$; $p = 0.0$). We found similar results when calculating the correlation coefficient for mitochondrial and nuclear data separately (Pearson correlation coefficient: $r = -0.40$; $p = 0.49$ for mitochondrial data; $r = -0.22$; $p = 0.21$ for nuclear data). A correlation between sequencing depth and target marker completeness was not detected ($r = 0.29$).

DISCUSSION

In this report, we present results from a target capture experiment using a mixed bait set covering specific taxa across several animal phyla (Porifera, Annelida, Mollusca, Arthropoda, and Chordata) set on a range of museum collection samples. We were able to obtain sequence information for 75% of all samples which is

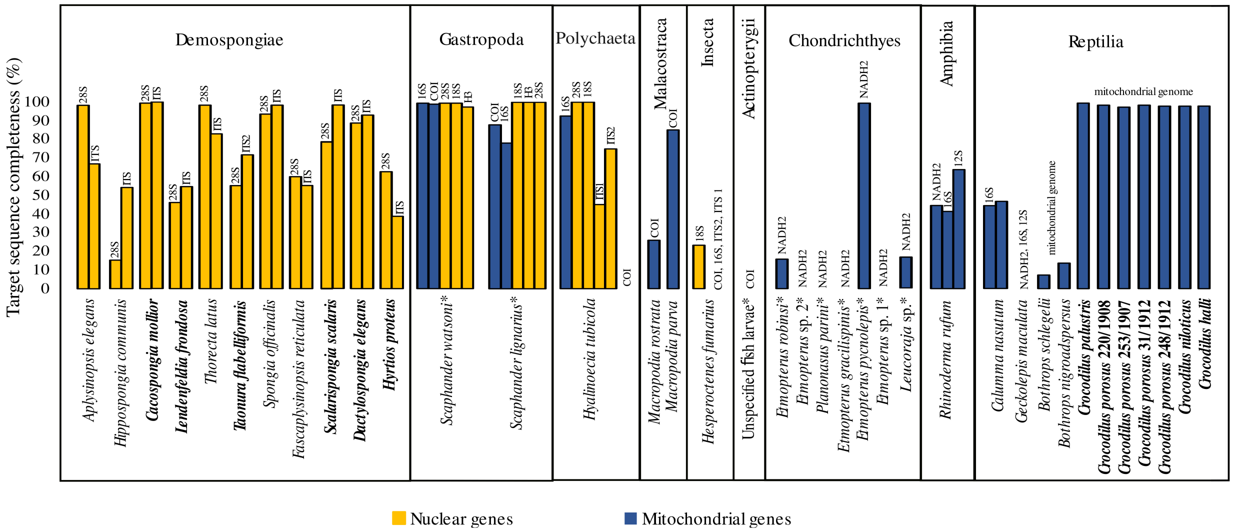


FIGURE 1 | Completeness of target genes after hybridisation capture. Dry material is indicated in bold, all other samples originate from wet-collection specimens. Assumed formalin-fixation before wet-collection preservation of specimens is indicated by an asterisk.

promising to be useful sequence information for phylogenetic placement of specimens. The obtained sequences will further be used for sample specific phylogenetic analyses. For the samples of the classes Demospongiae, Gastropoda, Polychaeta and Amphibia, we received consistently high sequence completeness (**Figure 1** and **Supplementary Table 4**). Above all, dry crocodilian material (tooth and bone) that are up to 100 years old (**Supplementary Table 1**) have shown to be a reliable source of DNA. Several samples of the classes Malacostraca, Insecta and Chondrichthyes targeted for mitochondrial and nuclear loci show low capture success. The single actinopterygian sample failed, which may have been due to long-term formalin preservation (N. Schnell pers. comm.). Although our results imply that target capture of nuclear markers outperforms capture of mitochondrial markers, the differences are likely introduced by samples with a generally low completeness of target sequences. In cases where both nuclear and mitochondrial markers were captured, similar results regarding the target sequence completeness were obtained (**Figure 1**). In general, wet-collection specimens showed poorer results compared to dry material. Water in ethanol solutions used for long-term storage intensifies hydrolysis (Lindahl, 1993) and may have contributed to our results.

To overcome potential disadvantages of large phylogenetic distances between bait and target sequences, a second round of target capture, as performed herein, can increase capture efficiency (e.g., Li et al., 2013; Paijmans et al., 2016). In this study, the p-distances between the bait sequences and the completeness of the consensus sequences are not correlated, which might be different if all consensus sequences were complete and should be investigated in further studies. Further experimental optimisation such as hybridisation temperature and time may allow for increasing capture efficiency in samples with low or no target gene completeness. However, our study also includes samples that should have small phylogenetic distances between bait and target sequences (e.g., *Etmopterus* spp., **Figure 1**). We were able to recover the complete mitochondrial marker sequence from only a single of these specimens (*E. pycnolepis*). As insufficient sequencing effort can be ruled out, reasons for the failure of the remaining samples could be related to fixation and preservation induced DNA damage. Details on the fixation history of most samples are poorly known (**Supplementary Table 1**), however, formalin has severe DNA damaging effects (Hoffman et al., 2015). Different ways of formalin fixation can also play a role in the success of DNA recovery (e.g., Paireder et al., 2013) ultimately influencing the amount and complexity of available target DNA for the target capture experiment. Besides these factors degradation and associated short DNA fragment size may have impeded the mapping attempts (Huson et al., 2007).

An alternative to commercially purchased RNA baits as used in this study are home-made DNA baits using PCR products of amplified target markers for DNA bait library production (González Fortes and Paijmans, 2019). In general, bait production for a small sample number targeting a single or few barcode markers of phylogenetically close taxonomic units is costly and inefficient. A combination of taxon-specific bait sequences for target capturing widely different taxa can overcome these limitations and enables the simultaneous sequencing of several phylogenetically distant taxa of interest. Our approach allows for cost-sharing between collection subsections and paves the way for implementing barcoding inventories in natural history collections, for example barcoding inventories of type specimens.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: 10.6084/m9.figshare.19619052.

ETHICS STATEMENT

Ethical review and approval was not required for the animal study because no living animals were collected or examined. DNA samples were taken solely from museum specimens.

AUTHOR CONTRIBUTIONS

NS and MH designed the study. SA and NS performed the laboratory work under supervision of MP. SA analysed the data under supervision of NS and MH. NS and SA wrote the manuscript with contributions from all authors. All authors contributed to the article and approved the submitted version.

FUNDING

This work was funded by the German Research Foundation (DFG; project number 351649567 to NS and MH within the DFG SPP 1991 "Taxon-Omics"). The publication of this article was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation; Project No. 491466077).

ACKNOWLEDGMENTS

We would like to sincerely thank our colleagues for entrusting us with their valuable samples and for their support with compiling the reference sequence set: Dirk Erpenbeck (LMU), Steffen Roth (UM), Justine Siegwald (UM), Nataliya Budaeva (UM), Frank Glaw (ZSM), Gavin Naylor (UF), Nalani Schnell (MNHN), Ralf Thiel (UH), Simon Weigmann (Elasmo-Lab), Michael Raupach (ZSM), Claudia Koch (ZFMK), Silke Schweiger (NHMW), and Axel Barlow (UWB). We would also like to thank the HPC team at the University of Potsdam for creating and maintaining the university's server on which we ran some of our analyses.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fevo.2022.909846/full#supplementary-material

REFERENCES

Dabney, J., Knapp, M., Glocke, I., Gansauge, M. T., Weihmann, A., Nickel, B., et al. (2013). Complete mitochondrial genome sequence of a Middle Pleistocene cave bear reconstructed from ultrashort DNA fragments. *Proc. Natl. Acad. Sci. U.S.A.* 110, 15758–15763. doi: 10.1073/pnas.1314445110

Gansauge, M. T., Gerber, T., Glocke, I., Korlević, P., Lippik, L., Nagel, S., et al. (2017). Single-stranded DNA library preparation from highly degraded DNA using T4 DNA ligase. *Nucleic Acids Res.* 45:e79. doi: 10.1093/nar/gkx033

González Fortes, G., and Paijmans, J. L. (2019). "Whole-genome capture of ancient DNA using homemade baits," in *Ancient DNA*. eds S. Beth, B. Axel, D. H. Peter, H. Michael, L. A. P. Johanna, E. R. S. André, (New York: Humana Press), 93–105. doi: 10.1007/978-1-4939-9176-1_11

Hahn, E. E., Alexander, M. R., Grealy, A., Stiller, J., Gardiner, D. M., and Holleley, C. E. (2021). Unlocking inaccessible historical genomes preserved in formalin. *Mol. Ecol. Resour.* [Epub ahead of print]. doi: 10.1111/1755-0998.13505

Hoffman, E. A., Frey, B. L., Smith, L. M., and Auble, D. T. (2015). Formaldehyde crosslinking: a tool for the study of chromatin complexes. *J. Biol. Chem.* 290, 26404–26411. doi: 10.1074/jbc.R115.651679

Huang, J. M., Yuan, H. and Li, C. H. (2021). Protocol for Cross-species Target-gene Enrichment. *Bio*. 101:e1010606. doi: 10.21769/BioProtoc.1010606

Huson, D. H., Auch, A. F., Qi, J., and Schuster, S. C. (2007). MEGAN analysis of metagenomic data. *Genom. Res.* 17, 377–386. doi: 10.1101/gr.5969107

Katoh, K., Misawa, K., Kuma, K. I., and Miyata, T. (2002). MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* 30, 3059–3066. doi: 10.1093/nar/gkf436

Kumar, S., Stecher, G., and Tamura, K. (2016). MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets. *Mol. Biol. Evol.* 33, 1870–1874. doi: 10.1093/molbev/msw054

Li, C., Corrigan, S., Yang, L., Straube, N., Harris, M., Hofreiter, M., et al. (2015). DNA capture reveals transoceanic gene flow in endangered river sharks. *Proc. Natl. Acad. Sci. U.S.A.* 112, 13302–13307. doi: 10.1073/pnas.1508735112

Li, C., Hofreiter, M., Straube, N., Corrigan, S., and Naylor, G. J. P. (2013). Capturing protein-coding genes across highly divergent species. *BioTechniques* 54, 321–326. doi: 10.2144/000114039

Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* 27, 2987–2993. doi: 10.1093/bioinformatics/btr509

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25, 1754–1760. doi: 10.1093/bioinformatics/btp324

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

Lyra, M. L., Carolina, A., Lourenço, C., Pinheiro, P. D. P., Pezzuti, T. L., Baêta, D., et al. (2020). High-throughput DNA sequencing of museum specimens sheds light on the long-missing species of the Bokermannohyla claresignata group (*Anura: Hylidae: Cophomantini*). *Zool. J. Linnean Soc.* 190, 1235–1255. doi: 10.1093/zoolinnean/zlaa033

Paijmans, J. L. A., Baleka, S., Henneberger, K., Taron, U. H., Trinks, A., Westbury, M. V., et al. (2017). Sequencing single-stranded libraries on the Illumina NextSeq 500 platform. *arXiv*. [preprint]. doi: 10.48550/arXiv.1711.11004

Paijmans, J. L. A., Fickel, J., Courtiol, A., Hofreiter, M., and Förster, D. W. (2016). Impact of enrichment conditions on cross-species capture of fresh and degraded DNA. *Mol. Ecol. Resour.* 16, 42–55. doi: 10.1111/1755-0998.12420

Paireder, S., Werner, B., Bailer, J., Werther, W., Schmid, E., Patzak, B., et al. (2013). Comparison of protocols for DNA extraction from long-term preserved formalin fixed tissues. *Anal. Biochem.* 439, 152–160. doi: 10.1016/j.ab.2013.04.006

Rancilhac, L., Bruy, T., Scherz, M. D., Pereira, E. A., Preick, M., Straube, N., et al. (2020). Target-enriched DNA sequencing from historical type material enables a partial revision of the Madagascar giant stream frogs (genus Mantidactylus). *J. Nat. Hist.* 54, 87–118. doi: 10.1080/00222933.2020.1748243

Rohland, N., Siedel, H., and Hofreiter, M. (2004). Nondestructive DNA extraction method for mitochondrial DNA analyses of museum specimens. *Biotechniques* 36, 814–821. doi: 10.2144/04365ST05

Scherz, M. D., Rasolonjatovo, S. M., Köhler, J., Rancilhac, L., Rakotoarison, A., Raselimanana, A. P., et al. (2020). 'Barcode fishing' for archival DNA from historical type material overcomes taxonomic hurdles, enabling the description of a new frog species. *Sci. Rep.* 10:19109. doi: 10.1038/s41598-020-75431-9

Springer, M. S., Signore, A. V., Paijmans, J. L. A., Vélez-Juarbe, J., Domning, D. P., Bauer, C. E., et al. (2015). Interordinal gene capture, the phylogenetic position of Steller's sea cow based on molecular and morphological data, and the macroevolutionary history of Sirenia. *Mol. Phylogenet. Evol.* 91, 178–193. doi: 10.1016/j.ympev.2015.05.022

Stiller, M., Sucker, A., Griewank, K., Aust, D., Baretton, G. B., Schadendorf, D., et al. (2016). Single-strand DNA library preparation improves sequencing of formalin-fixed and paraffin-embedded (FFPE) cancer DNA. *Oncotarget* 7:59115. doi: 10.18632/oncotarget.10827

Straube, N., Lyra, M. L., Paijmans, J. L. A., Preick, M., Basler, N., Penner, J., et al. (2021a). Successful application of ancient DNA extraction and library construction protocols to museum wet collection specimens. *Mol. Ecol. Resour.* 21, 2299–2315. doi: 10.1111/1755-0998.13433

Straube, N., Preick, M., Naylor, G. J. P., and Hofreiter, M. (2021b). Mitochondrial DNA sequencing of a wet-collection syntype demonstrates the importance of type material as genetic resource for lantern shark taxonomy (*Chondrichthyes: Etmopteridae*). *R. Soc. Open Sci.* 8:210474. doi: 10.1098/rsos.210474

Templeton, J. E., Brotherton, P. M., Llamas, B., Soubrier, J., Haak, W., Cooper, A., et al. (2013). DNA capture and next-generation sequencing can recover whole mitochondrial genomes from highly degraded samples for human identification. *Investig. Genet*. 4:26. doi: 10.1186/2041-2223-4-26

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

ORIGINAL RESEARCH
published: 30 June 2022
doi: 10.3389/fevo.2022.893088

Maximizing Molecular Data From Low-Quality Fluid-Preserved Specimens in Natural History Collections

Justin M. Bernstein[1]* *and Sara Ruane*[2]

[1] Department of Biological Sciences, Rutgers University – Newark, Newark, NJ, United States, [2] Life Sciences Section, Negaunee Integrative Research Center, Field Museum, Chicago, IL, United States

OPEN ACCESS

Edited by:
Jimmy McGuire,
University of California, Berkeley,
United States

Reviewed by:
Alex Pyron,
George Washington University,
United States
Jairo Arroyave,
National Autonomous University
of Mexico, Mexico

***Correspondence:**
Justin M. Bernstein
jmbernst223@gmail.com

Specialty section:
This article was submitted to
Phylogenetics, Phylogenomics,
and Systematics,
a section of the journal
Frontiers in Ecology and Evolution

Received: 10 March 2022
Accepted: 30 May 2022
Published: 30 June 2022

Citation:
Bernstein JM and Ruane S (2022)
Maximizing Molecular Data From
Low-Quality Fluid-Preserved
Specimens in Natural History
Collections.
Front. Ecol. Evol. 10:893088.
doi: 10.3389/fevo.2022.893088

Over the past decade, museum genomics studies have focused on obtaining DNA of sufficient quality and quantity for sequencing from fluid-preserved natural history specimens, primarily to be used in systematic studies. While these studies have opened windows to evolutionary and biodiversity knowledge of many species worldwide, published works often focus on the success of these DNA sequencing efforts, which is undoubtedly less common than obtaining minimal or sometimes no DNA or unusable sequence data from specimens in natural history collections. Here, we attempt to obtain and sequence DNA extracts from 115 fresh and 41 degraded samples of homalopsid snakes, as well as from two degraded samples of a poorly known snake, *Hydrablabes periops*. *Hydrablabes* has been suggested to belong to at least two different families (Natricidae and Homalopsidae) and with no fresh tissues known to be available, intractable museum specimens currently provide the only opportunity to determine this snake's taxonomic affinity. Although our aim was to generate a target-capture dataset for these samples, to be included in a broader phylogenetic study, results were less than ideal due to large amounts of missing data, especially using the same downstream methods as with standard, high-quality samples. However, rather than discount results entirely, we used mapping methods with references and pseudoreferences, along with phylogenetic analyses, to maximize any usable molecular data from our sequencing efforts, identify the taxonomic affinity of *H. periops*, and compare sequencing success between fresh and degraded tissue samples. This resulted in largely complete mitochondrial genomes for five specimens and hundreds to thousands of nuclear loci (ultra-conserved loci, anchored-hybrid enrichment loci, and a variety of loci frequently used in squamate phylogenetic studies) from fluid-preserved snakes, including a specimen of *H. periops* from the Field Museum of Natural History collection. We combined our *H. periops* data with previously published genomic and Sanger-sequenced datasets to confirm the familial designation of this taxon, reject previous taxonomic hypotheses, and make biogeographic inferences for *Hydrablabes*. A second *H. periops* specimen, despite being seemingly similar for initial raw sequencing results and after being put through the same protocols, resulted in little usable molecular data. We discuss the successes and failures of using different pipelines and methods to maximize the products from these data and provide expectations for others who are looking to use DNA sequencing efforts on specimens that likely have degraded DNA.

Life Science Identifier (*Hydrablabes periops*): zoobank.org:pub:F2AA44E2-D2EF-4747-972A-652C34C2C09D

Keywords: formalin, *Hydrablabes*, museum genomics, Natricidae, natural history collections, phylogenomics, snakes, systematics

INTRODUCTION

Advances in DNA sequencing technologies have allowed for the rapid accumulation of genomic or subgenomic datasets with thousands of loci. These datasets have provided opportunities to determine genomic correlates of phenotypic traits (Card et al., 2019; Stuckert et al., 2021), understand the links between recombination landscapes and genetic diversity (Schield et al., 2020), and reconstruct evolutionary histories in megadiverse groups (Hime et al., 2021). Research on the latter topic in particular, focusing on the discipline of systematics, has included continuously growing datasets to discover undescribed diversity in poorly studied taxa (Weinell and Brown, 2018), time-calibrate the diversification of extant groups (Álvarez-Carretero et al., 2021), and infer historical biogeography in comparative frameworks to better understand patterns of biodiversity (de Bruyn et al., 2014). However, these research findings are only possible due to the now-common practice of explicitly preserving fresh tissues upon collection of study organisms for subsequent DNA/RNA analyses. There remain large gaps of knowledge for thousands of organisms only known from museum specimens in natural history collections, often collected before practices of tissue preservation and DNA extraction. For centuries, vertebrates have been fixed using formalin or ethanol (Simmons, 2014), degrading the DNA by shearing, cross-linking, and deamination/depurination (Zimmermann et al., 2008; Campos and Gilbert, 2012; Do and Dobrovic, 2015), typically leading to DNA quality insufficient for sequencing, especially for systematic studies. The current era of genomics has been met with several protocols to obtain useable DNA from these intractable museum specimens (e.g., Rohland et al., 2004; Hykin et al., 2015; Ruane and Austin, 2017; O'Connell et al., 2021; reviewed in Ruane, 2021). As a result, the taxonomic identity and phylogenetic placement of poorly known snakes (Allentoft et al., 2018; Deepak et al., 2018), lizards (Hykin et al., 2015; McGuire et al., 2018), frogs (Rancilhac et al., 2020), salamanders (Pyron et al., 2022), crustaceans (France and Kocher, 1996), spiders (Wood et al., 2018), and birds (McCormack et al., 2016) have been successful, and with some studies on birds (Linck et al., 2017; Tsai et al., 2019) and mammals (Roycroft et al., 2021) obtaining levels of informativeness adequate to determine biogeographic histories and extinction patterns. Studies involving 'museum genomics' research involve materials from traditional museums and cryogenic collections, as well as the respective supporting infrastructure (Card et al., 2021). In this study, we use the term 'museum genomics' to refer to the more focused goal of obtaining useable DNA from often intractable, preserved museum specimens, which has undoubtedly created new directions for what is possible with DNA from voucher specimens and allowed us to leverage these data for biodiversity knowledge and evolutionary inference. However, many attempts at these endeavors still result in less than optimal (and frequently unusable) results, and studies often only report the successes that are obtained, leaving expectations of data quality, processing, and manipulation as a black box in such efforts.

While often viewed as a single task, the success of acquiring DNA from preserved museum specimens and obtaining DNA raw reads of sufficient quality for systematic studies each present separate difficulties. Hot alkali treatments (Campos and Gilbert, 2012; Hykin et al., 2015), heavy use of proteinase-K (Ruane and Austin, 2017), and development of digestion buffers (Allentoft et al., 2015) have all been used with varying success to break formalin cross-links and retrieve DNA from fixed specimens. However, different tissues (e.g., skin, liver, muscle, and bone) may yield varying DNA concentrations upon extraction (Appleyard et al., 2021; Zacho et al., 2021), and the lysis of soft tissue using enzymes like proteinase-K may be unsuccessful depending on the age, storage conditions, and preservation history of the source tissue. Even if DNA is extracted from intractable specimens, decreases in number and uneven distribution of mapped reads to reference genomes (Hykin et al., 2015; Allentoft et al., 2018), short fragment lengths, and low numbers of loci (Ruane and Austin, 2017) are commonly reported. Reduced-quality DNA from museum specimens is expected, but issues when using bioinformatic pipelines, the efficacy of using different types of loci and approaches for locus acquisition, and expectations of phylogenetic placements are seldom discussed. Additionally, when bioinformatic-related problems arise using published software, not all researchers have the expertise to edit, troubleshoot, or modify the source code. Predicting the analytical difficulties from museum genomics studies and understanding how data from degraded DNA can be processed will allow for higher success rates in understanding the biological histories of taxa only known from natural history collections. With increased global extinction rates (Pimm et al., 2014; De Vos et al., 2015) due to anthropogenic-related causes, it is important to elucidate the systematics and biodiversity of poorly-studied, yet ecologically important, groups or taxa that are rare or even possibly extinct.

Snakes are an excellent system for studying evolutionary processes (Esquerré et al., 2020; Schield et al., 2020; Westeen et al., 2020; Burbrink et al., 2021), and the utilization of preserved museum specimens has expanded our knowledge on both extant and extinct diversity (Ruane and Austin, 2017; Allentoft et al., 2018; Zacho et al., 2021). Southeast Asia in particular includes a diverse assemblage of snakes with multiple endemic lineages, many concentrated in biodiversity hotspots, which have been affected by the region's complex geological history (Hall, 2009; de Bruyn et al., 2014). Borneo, one of the largest islands in the world, harbors 160+ species of snakes, including multiple species that are only known from one to a few museum specimens and for which almost no natural history information is available (Stuebing et al., 2014; Das, 2018; Uetz et al., 2021). One such taxon is *Hydrablabes*, a genus consisting of two small-sized, aquatic snake species endemic to Borneo: *Hydrablabes periops* and *Hydrablabes praefrontalis*. Although the former species is more frequently encountered, *Hydrablabes* representation in natural history collections worldwide is lacking, with less than 10 and 0 specimens of each taxon in United States institutions, respectively. While these species are currently considered members of the family Natricidae, which contains hundreds of Old and New World aquatic species, they have also been hypothesized to belong to Homalopsidae (Murphy and Voris, 2014), a smaller family of mostly aquatic, mildly venomous snakes, also found across Southeast Asia. Much of Borneo's herpetofauna and its respective natural history is still in the midst of being fully described and understood (Quah et al., 2019; Das and Wong, 2021; Fukuyama et al., 2021). Indeed, Southeast Asia's undescribed diversity promises exciting discoveries, but it is equally worrisome that this diversity may disappear before ever being discovered (Sodhi et al., 2004; Strang and Rusli, 2021). Studying the systematics of rare snakes such as *Hydrablabes* can act as a first step in filling in the current knowledge gaps in the known biodiversity and evolutionary processes of Southeast Asia.

Here, we extract and sequence the DNA of homalopsid snakes from several natural history collections, and two specimens of *H. periops* from the Field Museum of Natural History (FMNH), as part of an ongoing study on homalopsids. We use a high-throughput target capture approach to sequence ultraconserved elements (UCEs; Faircloth et al., 2012), anchored hybrid enrichment loci (AHEs; Lemmon et al., 2012), and nuclear protein-coding genes (NPCGs) commonly used in squamate (lizards and snakes) phylogenetic studies *via* the SqCL v2 probe set from Singhal et al. (2017a,b). We use multiple pipelines and methods to isolate nuclear and mitochondrial loci to (i) maximize the utility of data obtained from museum specimens that would otherwise be considered 'failed' sequencing attempts, (ii) compare sequencing results between fresh and degraded tissue samples, (iii) place *H. periops* in a molecular phylogeny for the first time amongst all major extant snake lineages, and (iv) test competing taxonomic hypotheses for the familial designation of *Hydrablabes* (Natricidae vs. Homalopsidae). We focus on the failures/difficulties encountered pre- and post-sequencing, and make suggestions for future studies working with degraded DNA so as to increase expectations of error during project workflow and maximize the success of museum genomics for phylogenetic studies.

MATERIALS AND METHODS

Sample Collection and Morphological Identification

For the homalopsids, we obtained 115 fresh liver/muscle samples and 41 degraded (39 liver/muscle; 2 bone) tissue samples from several natural history museums (**Supplementary Table 1**). Additionally, we extracted liver tissue from two specimens of *H. periops* from the herpetology collection of the Field Museum of Natural History (specimens FMNH 158616, 251051; **Supplementary Table 1**). These *H. periops* specimens were collected by the late Mr. William Hosmer in 1964 (FMNH 158616) and Curator Emeritus at the FMNH, the late Dato Dr. Robert F. Inger, respective collaborators from the Field Research Team of Sabah Parks, Malaysia and Datin Tan Fui Lian in 1993 (FMNH 251051). Although not generated and used comparatively in this study, we note that computed tomography (CT) scans of FMNH 251051 are available on the MorphoSource repository (ark: /87602/m4/415178). To confirm the taxonomic identity of the specimens of *H. periops*, we conducted morphological examinations and compared those to species accounts in the literature (Mocquard, 1890; Stuebing et al., 2014). We looked at the following characters: color pattern; total length (TtL); tail length (TL), measured from the cloaca to the tip of the tail; snout-vent-length (SVL), measured from the tip of the rostral scale to the vent; TL:TtL ratio; dorsal scale rows (DSR) at 10 scales behind the head (anterior), midbody (half of the total length), and 5 scales anterior to the cloaca (posterior); number of subcaudal scales; number supraocular, preocular, subocular, and postocular scales; number of supralabial and infralabial scales; temporal scale formula; and the state of the prefrontal scales (divided vs. complete); morphological data can be found in **Supplementary Table 2**. While we only attempted molecular work from two of the FMNH specimens, we also obtained morphological data of a third *H. periops* specimen (FMNH 146230) and report it here. Our sampling also includes molecular data, as part of an in-progress study (Bernstein et al., unpublished data), from 115 fresh and 41 degraded samples of homalopsid snakes; we also included 3 viperids (*Bothrops moojeni* and *Bothrops pauloensis*), a colubrid (*Chironius exoletus*), a dipsadid (*Philodryas olfersii*), and an elapid (*Micrurus brasiliensis*) from Singhal et al. (2017b) as outgroups (**Supplementary Table 1**). While all of the tissues in our study have been stored in natural history museums, we use the terms 'museum specimens,' 'degraded,' and 'intractable' interchangeably to refer specifically to historic, fixed specimens with degraded DNA. We reference the homalopsids, viperids, colubrid, and dipsadid to draw quantitative and qualitative comparisons between degraded and fresh samples when attempting to recover nuclear and mitochondrial DNA, test the hypothesis of *H. periops* being a member of Homalopsidae, and help establish expectations for museum genomics studies. However, we limit our phylogenetic results and corresponding discussion on the *H. periops* samples.

Museum Specimen DNA Extraction and Sequencing

Total genomic DNA (gDNA) was extracted using published protocols for target capture sequencing of museum specimens (Ruane and Austin, 2017). This method uses a heated alkali buffer solution and a modified protocol of Qiagen® DNeasy Blood and Tissue kits to increase the gDNA yield from intractable specimens. Briefly, 100–200 mg liver tissue was cut into 15–25 mg pieces and washed in distilled water for 6 h to remove excess ethanol. Tissue was then pulverized or cut up to a mashed consistency. We then added ∼25–50 mg of the tissue to a 2-mL microcentrifuge tube with 300 μL of preheated (98°C) ATL buffer, and incubated the samples at 98°C for 15 min. The tubes were then cooled on ice for 2 min. Finally, we added 40 μL of proteinase-K to the samples and digested them for 48–72 h at 65°C, vortexing samples periodically and adding more proteinase-K if undigested tissue was visible. We then followed the post-digestion steps from the Qiagen® DNeasy Blood and Tissue kits protocol, except with two 100-μL final AE elution steps, rather than a single 200-μL elution. Different extraction attempts on the same sample were combined to increase the total gDNA per sample. Two of the samples were extracted from bone (122 and 14 mg), and we followed published protocols for obtaining DNA from hard tissue (Allentoft et al., 2015, 2018), with the exception that we used Qiagen® DNeasy Blood and Tissue kits after the proteinase-K digestion step. All DNA extractions were performed in an area isolated from fresh DNA work, on surfaces that were sterilized with bleach, and with UV-sterilized equipment and filter pipette tips. We used a Qubit 3 fluorometer (high sensitivity; Thermo Fisher Scientific: Invitrogen) to quantify the DNA yield of all extractions. Genomic DNA was sent to Daicel Arbor Biosciences (Ann Arbor, MI, United States) and optimized for target capture using the SqCL v2 probe set (Singhal et al., 2017b) for UCEs, AHEs, and NPCGs. To increase the likelihood of recovering reads from each sample, we had a small percentage of the non-captured libraries spiked into the sequencing pool to increase the number of bycatch molecules, thus increasing the chance of obtaining mitochondrial DNA (mtDNA) from our museum samples. The final sequencing pool for the degraded samples was prepared by combining the enriched (85%) and unenriched (15%) pools. Samples were sequenced on the Illumina NovaSeq 6000 platform on partial S4 PE150 lanes. Raw fastq files for the two specimens of *H. periops* have been deposited in the National Center for Biotechnology Information (NCBI) Sequence Read Archive (SRA) under the BioProject Accession Number PRJNA796637. The raw fastq files for the fresh and museum homalopsid specimens and outgroups are deposited under the SRA BioProject Accession Number PRJNA792597, PRJNA667001, and PRJNA382381 (see **Supplementary Table 1** for note on data availability for PRJNA792597).

Bioinformatics

We checked the quality of raw reads using FastQC (Andrews, 2010) and tested for contamination with FastQ Screen (Wingett and Andrews, 2018). To trim adapters and barcodes, we used illumiprocessor (Faircloth, 2011; Lohse et al., 2012; Del Fabbro et al., 2013) with default settings, and then assembled paired-end reads using SPAdes (Bankevich et al., 2012) in the *Phyluce* v1.7.1 (Faircloth, 2016) pipeline for processing UCEs. We also used Assembly By Short Sequences (ABySS; Simpson et al., 2009), separately, with k-mer values set to 30 and 60, to determine if more loci can be recovered. However, both assemblers achieved comparable run statistics, so we continued with using SPAdes, which averages over multiple *k*-mer values, as it is conveniently integrated into *Phyluce*. To test the hypothesis if *H. periops* is a homalopsid, we included *H. periops* into a dataset of homalopsid snakes and outgroup taxa (see **Supplementary Table 1**). We created DNA alignments for each locus following the workflow for *Phyluce*. Many of our museum samples had a high number of raw reads (>10–15 million); due to computational constraints, we used seqtk[1] to subsample the raw data from these samples to 3.5 million reads from each pair (7 million total). To align our *H. periops* samples with homalopsid samples, alignments of homologous nucleotide sites for each locus were edge-trimmed with Gblocks, and data matrices were created for each locus that contained at least 75% of the taxa in the dataset.

Because *Phyluce* yielded poor final phylogenetic results for the museum specimens (see Section "Concatenated and Species Tree Analyses"), we also extracted individual loci from the cleaned and trimmed raw data using Geneious v11.1.5. We tested a few approaches to compare their success of recovering targeted loci. Some of these approaches involved the use of a pseudoreference genome consisting of SqCL concatenated loci (with 10 ambiguous bases [Ns] between each locus) from one of the fresh homalopsids or the *Myanophis thanlyinensis* (Homalopsidae) reference genome from Köhler et al. (2021), hereafter 'pseudoreference' and 'reference genome,' respectively. Our approaches include: (i) mapping the trimmed and cleaned raw reads to the pseudoreference, (ii) BLASTing (Basic Local Alignment Search Tool) the trimmed and cleaned raw reads to the pseudoreference, (iii) mapping the unaligned.loci file of each individual (containing fastas of all individual loci) from *Phyluce* to the pseudoreference, (iv) BLASTing the unaligned.loci file from *Phyluce* to the pseudoreference, and (v) mapping the trimmed and cleaned raw reads to the reference genome. BLAST databases containing the pseudoreference and references were created using the 'Add sequence database' function in Geneious (**Figure 1**). All mapping and BLAST methods were performed under default parameters. For BLAST, Megablast (5 iterations) was used so that only matches with high similarity were returned. Our goal was to find the most efficient way to retrieve loci for the museum samples, thus we considered methods that took >24 h per sample to take too long (this is especially important if done

[1] https://github.com/lh3/seqtk

on computer clusters that have time constraints) and aborted the process. We test these approaches using two degraded samples as preliminary runs: (raw reads for *Calamophis ruuddelangi* RMNH.RENA 47517 and *Gyiophis maculosa* KU 92395 = ∼30 million reads each; unaligned.loci files for *C. ruuddelangi* and *G. maculosa* containing 1,527 and 2,016 loci, respectively), and then ran the rest of the samples using the approach that retrieved loci in the shortest amount of time and with the highest success. All approaches were run on a Digital Storm PC with a 64-bit operating system, x64-based processor, 16 cores, and 40 GB RAM allocated to Geneious for computation.

To increase the robustness of our phylogenetic results and determine if *H. periops* is a homalopsid, natricid, or a member of another family of snake, we used data from previously published studies to create additional DNA alignments based on our loci obtained (see Section "Read and Locus Acquisition"). We used the 50 longest UCEs of homalopsids (Bernstein et al., unpublished data) obtained from Geneious (see Section "Read and Locus Acquisition") to create a UCE-only alignment with *H. periops* and homalopsid snakes (plus outgroups). We also created alignments using AHEs from almost all lizard and snake families using the data from Burbrink et al. (2019, 2020) and a multilocus dataset from several genera of natricids (and one outgroup colubrid) from Deepak et al. (2021a) and other studies (Alfaro and Arnold, 2001; de Queiroz et al., 2002; Nagy et al., 2005; Guo et al., 2012, 2019; McVay and Carstens, 2013; Pyron et al., 2013; Kindler et al., 2014; McVay et al., 2015; Alencar et al., 2016; Ren et al., 2019; Lalronunga et al., 2020; Deepak et al., 2021b) consisting of two mitochondrial genes (cytochrome *b* [cyt-b], NADH-ubiquinone oxidoreductase chain 4 [ND4]), and one nuclear gene [Brain Derived Neurotrophic Factor (BDNF)]. We used the AHEs from Burbrink et al. (2020) as a pseudoreference to identify the same loci from this and our study. Specimen metadata from other published or in-progress studies can be found in **Supplementary Table 1**.

To obtain mtDNA from museum specimens, we tried three methods: (i) mitochondrial baiting and iterative mapping using MITObim (Hahn et al., 2013), (ii) mitogenomic data extraction using MitoFinder (Allio et al., 2020), with metaSPAdes (Nurk et al., 2017) used as the assembler, and (iii) mapping raw reads to a mitochondrial reference genome in Geneious, all under default parameters. Both MITObim and MitoFinder require mitochondrial reference genomes to extract loci, thus we used a *Hypsiscopus plumbea* (Homalopsidae) mitochondrial genome (Genbank accession: DQ343650; Yan et al., 2008). We used this reference for our mapping approach in Geneious as well. We used the 'Highest Quality' consensus option in Geneious, which only creates consensus sequences out of mapped reads using high-quality chromatograms.

All loci obtained through Geneious were manually incorporated into DNA alignments from *Phyluce* by using the 'Multiple Align' tool in Geneious. We left these alignments untrimmed, which has been found to achieve the best phylogenetic results in studies using UCEs (Portik and Wiens, 2021). We used the *lm* function in the *stats* package in R (R Core Team, 2021) to graph linear model relationships of (i) number of raw reads obtained and (ii) number of nuclear loci recovered, with the age of specimen (years since collected), DNA yield, and total gDNA used for sequencing. We used the same methods to determine trends between the number of raw reads and base pairs (in bp and percentage) of museum specimen reads that mapped to the mtDNA reference genome. We note that our linear models are heavily influenced by FMNH 251051 due to its high DNA yield and sequencing success compared to other specimens (see Section "Results"). We leave this specimen out of our analyses due to it being an outlier, but as it represents an important piece of information in regards to museum genomics success, linear regressions with and without this specimen can be found in **Supplementary Figures 1, 2**.

Phylogenetic Analyses

To determine the phylogenetic placement of *H. periops* using loci obtained from the *Phyluce* pipeline, we reconstructed a phylogeny by concatenating the alignments with the 'phyluce_align_concatenate_alignments' command and then used this as input for IQ-TREE v1.6.12 (Nguyen et al., 2015) under a GTR+G substitution model and with 1,000 bootstrap replicates. Due to high levels of missing data from formalin specimens and failed phylogenetic reconstruction with the loci obtained from *Phyluce* (see Section "Read and Locus Acquisition"), we ran species tree analyses, taking gene-tree-species-tree discordance into account, selectively choosing AHEs and UCEs with minimal missing data (see below).

For the species tree analysis using the multilocus dataset (cyt-b, ND4, and BDNF), we used Bayesian inference in StarBEAST2 (Ogilvie et al., 2017) under a birth-death evolutionary model, partitioning each alignment using the best partitioning scheme determined by PartitionFinder2 (Lanfear et al., 2017). We ran the model for 50 million generations, inferred and marginalized site models for our analysis using the 'bModelTest' plugin (Bouckaert and Drummond, 2017), and used an uncorrelated lognormal clock rate to allow branch rate heterogeneity. We assessed the convergence of our runs in Tracer v1.7 (Rambaut et al., 2018), discarding 25% of the run as burn-in, and considered effective sample sizes (ESS) > 200 to indicate sufficient sampling of parameter space. In addition to the StarBEAST2 tree, we reconstructed gene trees and a concatenated tree for cyt-b, ND4, and BDNF in IQ-TREE, searching for the best nucleotide model for each dataset with ModelFinder (Kalyaanamoorthy et al., 2017). Branch support was assessed by 1,000 ultrafast (UF) bootstrap iterations and SH-aLRT tests (Guindon et al., 2010; Hoang et al., 2018); relationships with UF bootstraps and SH-aLRT tests ≥95 and ≥80, respectively, were considered to be well-supported.

Because we had more loci in our AHE (n = 33) and UCE (n = 50) alignments compared to the multilocus dataset, and Bayesian approaches can be computationally demanding, we used polynomial time species tree reconstruction in ASTRAL-III (Zhang et al., 2018) for divergence date estimation. ASTRAL-III uses individual gene trees as input, so we built genealogies with the UCEs and AHEs. Gene trees were created using IQ-TREE with the same parameters used for the multilocus dataset loci. The individual AHE and UCE trees were used as input for

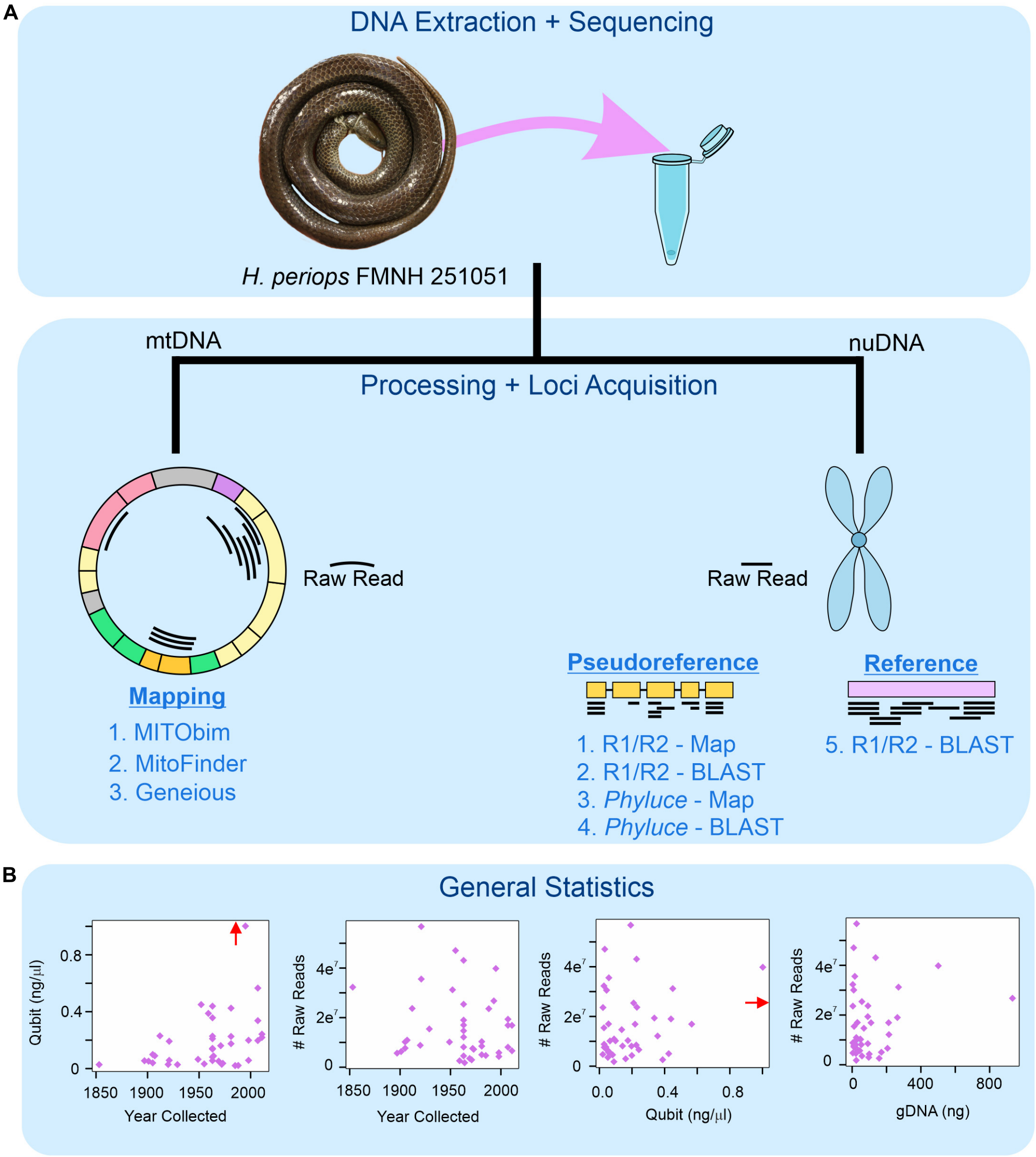


FIGURE 1 | Workflow for museum genomics in this project. **(A)** Three approaches for mitochondrial mapping to a reference genome (left) and five for nuclear locus acquisition using mapping and BLAST (right). *Phyluce* = unaligned.fasta file from *Phyluce* pipeline; R1/R2 = *.R1.fastq.gz and *.R2.fastq.gz files for raw reads. **(B)** Read statistics for DNA yield (Qubit) and total number of raw reads obtained by age of specimen (year collected), DNA yield, and total genomic DNA used for sequencing (gDNA). Red arrow shows data point for *Hydrablabes periops* FMNH 251051 (Qubit = ~6.0 ng/μl), arrow positioned along axis corresponding to respective *X*-axis and *Y*-axis values.

our species tree analysis in ASTRAL-III (Zhang et al., 2018), under default parameters. Relationships with Bayesian posterior probabilities (Bpp) ≥ 0.95 are considered strongly supported. As mentioned above, Bayesian methods can be a computationally difficult task with high numbers of loci, so we used treePL v1.0 (Smith and O'Meara, 2012) to estimate divergence times of our AHE dataset, which has the highest familial-level sampling. This software uses a semi-parametric penalized likelihood approach to estimate rates of gene evolution on branches of a concatenated input tree. We created a concatenated alignment of our AHEs and obtained a phylogeny using the parameters described above, with the exception that a GTR+G+I model was used. We used the 'thorough' and 'prime' commands to find the optimal parameters of our treePL analysis and to ensure the analysis ran until convergence. To identify the optimal smoothing parameter, which affects the rate variation penalty across the tree, we used the random subsample and replicate cross-validation (RSRCV) function. To calibrate the divergence times, we used squamate fossils that have been described and used in previous studies (Jones et al., 2013; Alencar et al., 2016; Zaher et al., 2018;

Burbrink et al., 2020; see **Supplementary Table 3**). Although we date the entire tree, we focus on the Natricidae given our phylogenetic results, discussed below.

RESULTS

Read and Locus Acquisition

Despite extremely low yields of DNA during extraction (see **Supplementary Table 1**), we were successful in sequencing many of the targeted museum specimens. Using *Phyluce*, we were able to recover 9–3,889 loci (median [M] = 403) from the museum (intractable) samples. However, these locus alignments were significantly smaller and contained fewer samples than loci obtained from fresh specimens, with DNA alignments from fresh samples and museum samples being 224–2,633 (average [$\bar{x}$] = 845.8) and 10–864 ($\bar{x}$ = 121.5), respectively. The alignments with degraded samples only contained an average of 6.7 (out of 43) formalin specimens across all alignments. Using alternative approaches, all mapping and BLASTing of raw reads to the pseudoreference and reference genomes took >24 h, and thus were terminated. When mapping the loci from these unaligned.loci files of *C. ruuddelangi* and *G. maculosa*, the analyses took, respectively, 17 and 10 seconds (s) to map 1,819 and 1,248 loci to the pseudoreference (remaining loci had no match). While these were faster, some loci that mapped to the pseudoreference spanned more than one gene, albeit rarely. When using the BLAST approach of the unaligned.loci files to the pseudoreference, run time took 60 s for *C. ruuddelangi* (2,016 recovered loci) and 45 s for *G. maculosa*, (1,800 recovered loci). Because BLASTing the *Phyluce* unaligned.loci file to the pseudoreference recovered more loci (and the DNA sequence of each locus is conveniently created in a separate file), we use this approach and considered it the most efficient. We note that despite two rounds of cleaning and trimming of adapter sequences, tens to hundreds of loci from museum specimens contained portions of adapters, which were trimmed off when we BLASTed sequences. No clear correlations were observed between DNA yield, specimen age, or raw reads obtained (**Supplementary Figure 1**).

BLASTing the unaligned.loci file to the pseudoreference retrieved a total of 25,126 UCEs, 1,657 AHEs, and 199 nuclear genes across all 43 museum specimens (**Supplementary Table 4**). Locus lengths (bp) ranged from 28–1,627 ($\bar{x}/M$ = 218.84/152) for UCEs, 28–1,988 ($\bar{x}/M$ = 156.37/94) for AHEs, and 34–1,225 ($\bar{x}/M$ = 233.47/120.5) for nuclear genes. The two specimens of *H. periops* we sequenced did not yield similar numbers of loci: we obtained 5 loci (44–75 bp) from the older specimen FMNH 158616, and 3,530 loci (33–1,532 bp; $\bar{x}/M$ = 292.5/218) from the more recently collected specimen FMNH 251051. This latter sample contained 275 loci that were ≥500 bp, and this specimen was used to determine *Hydrablabes* phylogenetic placement amongst other snake lineages. Taking all of our museum specimens into account, we recovered UCEs, AHEs, and NPCGs ≥ 250 bp for 25, 18, and 9 specimens, respectively (**Figure 2A**). We found positive relationships between DNA yield with the number of AHEs and NPCGs obtained, but not UCEs (**Figure 2B**). These patterns were also seen when comparing total gDNA used for sequencing with the number of loci obtained. Contrarily, there was no correlation between the number of loci obtained and the age of the specimen (**Figure 2B**). Graphs with lines, R^2, and p-values from linear models are in **Supplementary Figures 1, 2**.

All approaches for obtaining mitochondrial DNA from *H. periops* FMNH 158616 failed. However, attempts for FMNH 251051 were successful, depending on which approach was used to isolate mitochondrial bycatch. MITObim failed to extract any loci from the raw read data, while MitoFinder was successful in extracting 70 unique sequences of 15 mtDNA genes, across seven museum specimens (**Supplementary Table 5**). These sequences range from 162 to 1,785 bp ($\bar{x}/M$ = 898.3/914). Our most successful attempts to obtain mtDNA was using Geneious. Our mapping method of raw reads to the mitochondrial genome of *Hypsiscopus plumbea* resulted in a near-complete mitochondrial genome of the more recently collected *H. periops* specimen, mapping ∼1.17 million reads (15,649 non-ambiguous [A, C, T, G] bp), obtaining whole or partial coverage of every gene, control region, and tRNA (except tRNA-Phe). Out of the 43 museum specimens sequenced in this study, 27 specimens had at least one read mapped to the *H. plumbea* mitochondrial reference genome, with >1,000 reads mapped for 9 of these specimens (**Supplementary Table 6**). We recovered a range of 0.72–98.95% (126–17,215 bp) of the mitochondrial genome (reference = 17,397 bp). For five specimens, we obtained near-complete mitochondrial genomes, with >85% of the genome sequenced with almost all protein-coding genes, tRNAs, control regions, and the replication origin (**Figure 3A**). We observed no relationship between age or total gDNA with the number of non-ambiguous bp mapped to the mtDNA reference, number of raw reads mapped to the mtDNA reference, or percent of mtDNA genome sequenced (**Figure 3B** and **Supplementary Figure 1**). This was also seen when comparing DNA yield to the number of raw reads mapped to the mtDNA reference. However, DNA yield had a positive relationship with the number of bp mapped to the reference and percent of mtDNA genome sequenced (**Figure 3B**). Graphs with lines, R^2, and p-values from linear models are in **Supplementary Figures 1, 2**.

Concatenated and Species Tree Analyses

Our concatenated tree of homalopsids (fresh and degraded samples) + *Hydrablabes* (degraded samples) using all loci obtained from *Phyluce* (4,822 alignments concatenated to 2,346,038 bp) placed the two *H. periops* specimens within a group containing all museum sample homalopsids (**Supplementary Figure 4**). However, all of the museum samples randomly cluster (i.e., no sensible evolutionary relationships) together close to the outgroup taxa with long branches, and are placed outside of the fresh homalopsid specimens.

Our concatenated and genomic trees using the molecular data obtained from Geneious combined with published datasets supported the placement of *H. periops* in Natricidae. The multilocus dataset of natricids from Deepak et al. (2021a)

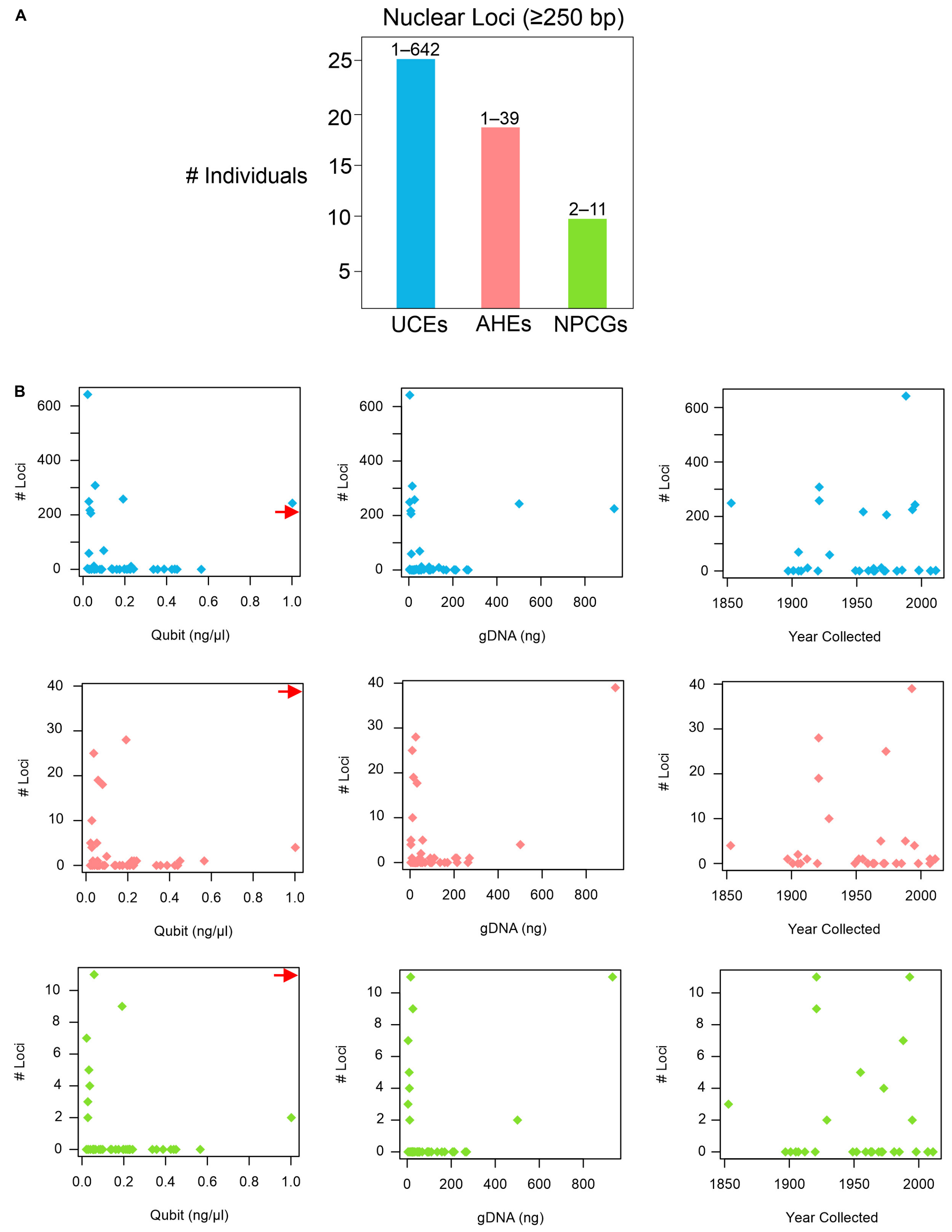


FIGURE 2 | Nuclear loci recovered from pseudoreference BLASTing. **(A)** Number of individuals that yielded UCEs, AHEs, and NPCGs ≥ 250 bp. Numbers above bars represent the range of number of loci recovered amongst all individuals. **(B)** Graphs showing the relationships of the number of UCEs (top), AHEs (middle), and NPCGs (bottom) and specimen age (year collected), DNA yield (Qubit), and total genomic DNA used for sequencing (gDNA). Red arrow shows data point for *H. periops* FMNH 251051 (Qubit = ~6.0 ng/μl), arrow positioned along axis corresponding to respective *Y*-axis values. Complete list of all nuclear loci and lengths for each specimen is in **Supplementary Table 4**.

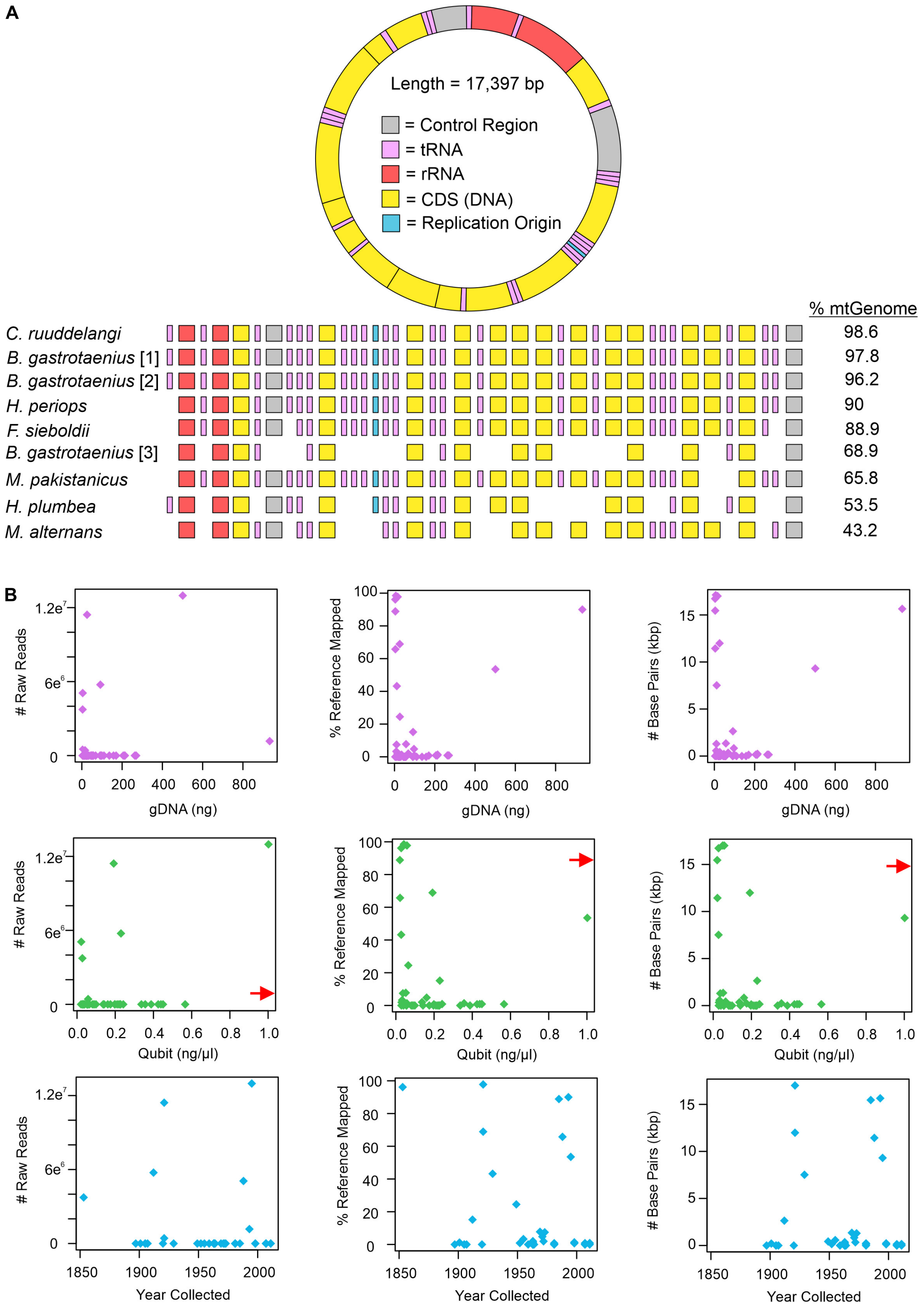


FIGURE 3 | Mitochondrial mapping results from Geneious. **(A)** Mapped loci from specimens genera in order include Calamophis, Brachyorrhos, Hydrablabes, Ferania, Mintonophis, Hypsiscopus, and Miralia with ≥25% of the whole mitochondrial reference genome (mtGenome). Blocks indicate successfully recovered regions (minimum ≥ 20% coverage). Gene regions colored in order as they appear on the reference genome. **(B)** Graphs showing the number of mitochondrial raw reads mapped to the mtGenome, percent of the mtGenome obtained, and the total number of base pairs obtained relative to total genomic DNA used for sequencing (gDNA; purple dots), DNA yield (Qubit score; green dots), and sample age (year collected; blue dots). Red arrow shows data point for *H. periops* FMNH 251051 (Qubit = ~6.0 ng/μl), arrow positioned along axis corresponding to respective *Y*-axis value. The complete list of mtDNA regions and respective coverage for each specimen is in **Supplementary Table 6**.

recovered a monophyletic Natricidae, with *Hydrablabes periops* as the sister taxon to *Trimerodytes praemaxillaris* (**Supplementary Figure 3**). The single gene trees resulted in multiple positions for *H. periops*, including an unresolved placement (BDNF), as sister to *T. praemaxillaris* (cyt-b), and the most closely related lineage to the sister pair *Smithophis atemporalis* + *Opisthotropis voquyi* (ND4) (**Supplementary Figure 3**). The species tree constructed from StarBEAST2 reached convergence and most ESS values were >200, with the exception of a few bModelTest and shape parameters, and one gene subset likelihood; the posterior, likelihood, prior, and species coalescent parameters all had ESS > 200. The species tree shows *H. periops* in the same phylogenetic position as the ND4 gene tree. While there is low support for placement with respect to the generic relationships, *H. periops* is strongly supported as a natricid (**Figure 4A**) in this multilocus tree.

The genomic trees using loci from published data also support that *H. periops* is in the family Natricidae, with its placement outside Homalopsidae. The UCE species tree strongly recovered Homalopsidae as a clade, and *H. periops* positioned with the outgroups *Micrurus*, *Chironius*, and *Philodryas* with strong support (**Figure 4B**). Specifically, *H. periops* is sister to *Chironius* (Colubridae) + *Philodryas* (Dipsadidae), although with poor support (**Figure 4B**). For the AHEs, we were able to obtain 33 loci that aligned to the AHEs of several snake families from Burbrink et al. (2020). Our species tree using 33 loci was broadly consistent with the full dataset from Burbrink et al. (2020), with most nodes strongly supported (**Figure 4C**). Higher-level relationships were identical between both trees, with the exception of the placement of Dibamia and Iguania, both of which have low support in our tree and in the species tree from Burbrink et al. (2020). Of the 37 snake families, our species tree shares the same relationships with ones seen in the full dataset tree, with the exception of the placements of Atractaspididae, Bolyeriidae, and Lamprophiidae, the latter poorly supported in both trees. Similar to the UCE tree, using AHEs recovered *H. periops* within Natricidae with strong support (**Figure 4C**); *H. periops* is sister to a clade containing Eurasian natricids (*Trimerodytes percarinatus* and *Natrix natrix*) and North American natricids (*Tropidoclonion lineatum*, *Storeria dekayi*, *Thamnophis marcianus*, *Liodytes pygaea*) (**Figure 4C**). Our AHE concatenated tree, used for divergence dating, is similar to the one obtained in Burbrink et al. (2020), with the exception that the ancestral Iguania lineage subtends Anguiformes (sister to Anguiformes in Burbrink et al., 2020). Our divergence dates of Natricidae are within 1–5 myr of those obtained from Burbrink et al. (2020) with the divergence of *H. periops* from its sister group ∼20.9 mya (**Figure 4C**).

DISCUSSION

Museum Genomics Is Successful With a Range of Specimen Ages and DNA Yields

Our results emphasize that the outcomes of several phases of museum genomics projects (e.g., DNA extraction, DNA sequencing, and bioinformatics) may not be optimal, yet valuable results can still be obtained. Our DNA extraction attempts on museum specimens yielded poor concentrations of DNA (often <0.1 ng/μl; see **Supplementary Table 1**). However, we were still able to extract a large number of nuclear and mitochondrial loci for several specimens, and even near-complete mitochondrial genomes for 5 individuals. We observe comparable findings with respect to other studies in that DNA extractions on museum specimens yielded extremely low levels of quantifiable DNA, often resulting in Qubit readings of 'Too Low' or <0.01 ng/μl (rarely ≥1 ng/μl), which are similar to reported quantifications in museum genomics studies (Hykin et al., 2015; Ruane and Austin, 2017). Though some studies achieve a wide range of DNA yields ("Too Low"–11.5, Ruane and Austin, 2017; "Too Low"–92.4, Zacho et al., 2021) from specimens when using identical methods within the respective study, it is likely that this is the result of different preservation treatment and environmental factors from time of collection to DNA extraction [e.g., exposure to UV light, time span from the death of an animal to preservation, preservation methods (ethanol vs. formalin), etc.]. While such information may not be (and are often not) recorded, the year of collection is typically known, providing an estimate of the age of the specimen post-mortem. Our study obtained significantly more nuclear and mitochondrial loci from the 28-year old (1993) specimen compared to the 57-year old (1964) specimen, the latter of which we only obtained five nuclear genes (44–75 bp; **Supplementary Table 4**). However, we cannot draw concrete patterns in relation to sequencing success and age (or even DNA yield), as some specimens from 1963 yielded DNA concentrations of 0.3–0.8 ng/μl (compared to <0.1 for other specimens) with failed sequencing results, while others for which we obtained near-complete mitochondrial genomes and hundreds of nuclear loci had concentrations <0.1 or even 'Too Low' and were collected in 1853 and 1921. These latter specimens are 41–109 years older than the 1964 *H. periops* specimen with failed sequencing attempts. Other studies have also found positive correlations between specimen age and DNA sequencing success in reptiles (Hykin et al., 2015) and birds (McCormack et al., 2016), but this is not ubiquitous amongst museum genomic studies with the same study organisms (e.g., Linck et al., 2017; Ruane and Austin, 2017). It is worth noting that obtaining samples that have had less time in fixatives will be ideal for recovering and sequencing DNA, as has been seen in recent studies using AHEs from salamanders that are ∼50 years old (Pyron et al., 2022). We note that while many of our linear regressions do not support significant relationships between DNA yield, age of specimens, and particular locus types retrieved (**Supplementary Figures 1**, **2**), future studies that include more specimens with an even sampling of specimen ages and tissue types may find better consistency with respect to specimen age and quality of results.

Museum Genomics Confirms *Hydrablabes* Is an Asian Natricid

The placement of *H. periops* supports the familial taxonomic status of *Hydrablabes* as a natricid, rejecting the hypothesis that this species is a homalopsid (Murphy and Voris, 2014). Divergence dates of natricids estimated here are slightly younger

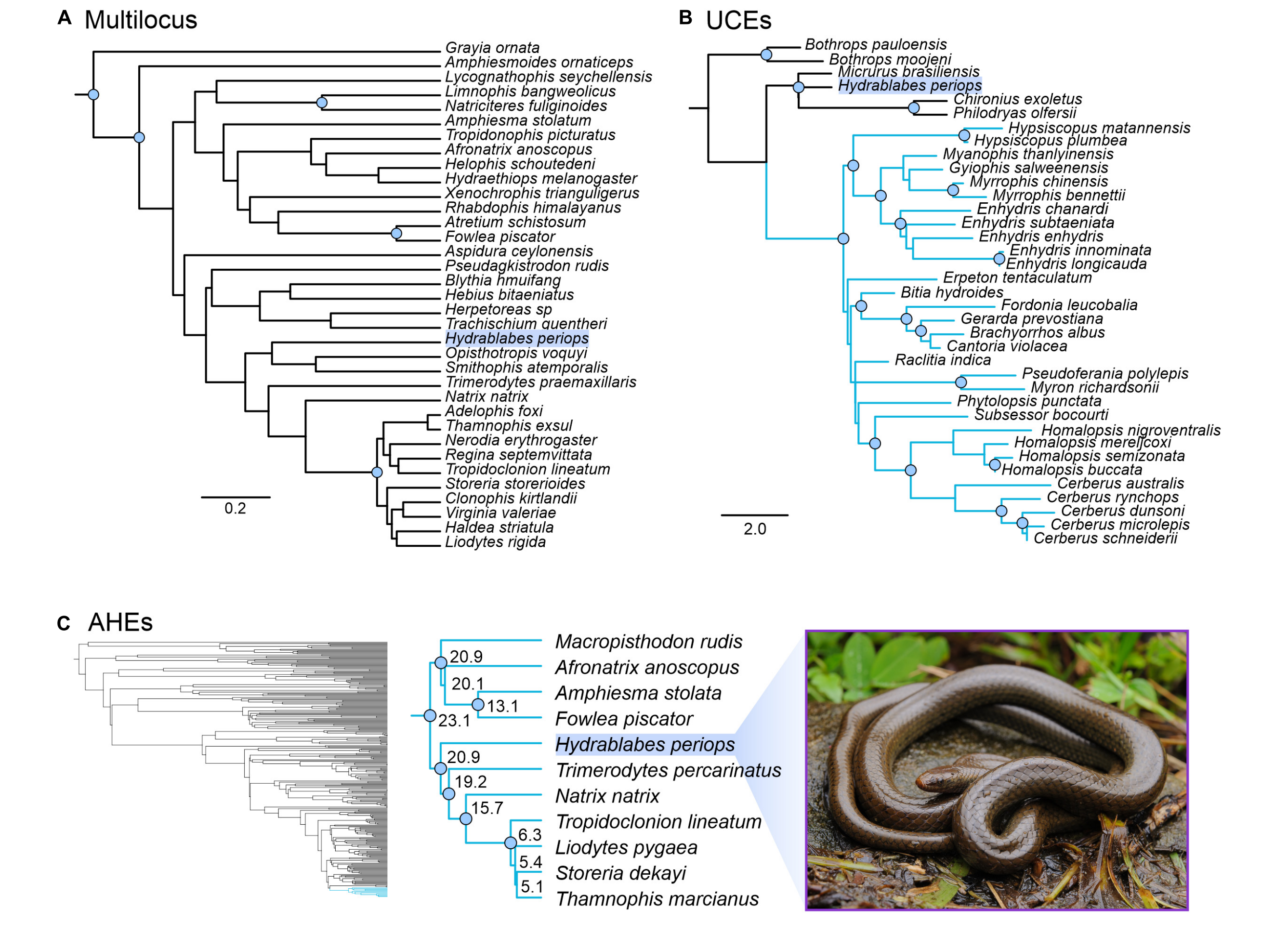


FIGURE 4 | Phylogenetic placement of *Hydrablabes periops* using three datasets from other studies. **(A)** Species tree using one nuclear and two mitochondrial genes (outgroup = *Grayia*); scale bar in nucleotide substitutions per site. **(B)** Homalopsidae species tree using UCEs; blue clade = Homalopsidae, scale bar in coalescent units. **(C)** Squamate species tree using AHEs; blue clade = Natricidae, numbers on enlarged Natricidae clade represent divergence dates in millions of years. Node circles in all trees indicate strongly supported relationships (Bpp ≥ 0.95; not shown for Squamata AHE tree). Photo credit of live *H. periops*: Chien C. Lee.

(**Figure 4C**) than those of Burbrink et al. (2020) (likely due to our reduced dataset and the inclusion of *Hydrablabes*), but are still broadly consistent with what is hypothesized about this family's diversification. *Hydrablabes* is a genus that is endemic to Borneo, a continental island that has only recently separated from mainland Southeast Asia (Hall, 2009). Cenozoic Sundaland was composed of the islands of Borneo, Java, and Sumatra, connected to the mainland by a land bridge. At ∼400,000 years ago (kya), Pleistocene sea-level fluctuations caused cycles of emergence and submergence of this land bridge during respective glacial and interglacial periods, up until the Last Glacial Maximum ∼20 kya (Voris, 2000; Sarr et al., 2019; Husson et al., 2020). Given its distribution, the close relation of *H. periops* to other Asian natricids is not unexpected, and our multilocus and AHE datasets provide different, yet valuable, information regarding the evolution of *Hydrablabes*. The multilocus tree supports *H. periops* as most closely related to Asian natricids in South and Southeast Asia, but outside of the clade with Indochinese *Trimerodytes*, Eurasian *Natrix natrix*, and North American natricids. Similarly, the AHE tree supports *H. periops* as sister to a group containing these taxa. These topologies are congruent with that of Deepak et al. (2021a), whose study also supported an Asian origin of Natricidae. Results from Deepak et al. (2021a) show a Mainland Asia + Japan origin over 20 mya for the clade that *Hydrablabes* is sister to. Biogeographic scenarios make sense in relation to our findings, as Borneo was still connected to the mainland prior to 400 kya (Hall, 2009; Husson et al., 2020). Our age of the clade containing *H. periops* (∼20.9 mya) may indicate population dispersal into Borneo from the mainland, with subsequent extinction events outside Borneo. Alternatively, the lineage ancestral to this clade may have dispersed into Borneo, followed by an *in situ* speciation event. The divergence dates of *H. periops* and *Trimerodytes* occur at interesting points in Indochina's geological record. Specifically, the rise of the Hengduan Mountains of the eastern Tibetan Plateau (Western

China) are considered to have been a diversification driver of *Trimerodytes* ~23.9 mya (Guo et al., 2020). While greater taxonomic sampling of *Hydrablabes* and other Borneo-endemics, as well as increased molecular sampling, will help to elucidate the evolutionary history of *Hydrablabes* amongst natricids, these initial results provide evolutionary and phylogeographic hypotheses for future testing.

Expectations and Suggestions for Museum Genomics

Museum genomics has advanced significantly in the last decade, undoubtably a product of newer, high-throughput sequencing technologies and the rapid accumulation of genome-scale datasets. A variety of biochemical protocols have been developed with varying degrees of success in different organismal systems. Some protocols (Campos and Gilbert, 2012; Hykin et al., 2015) rely on hot alkali treatments to increase the odds of extracting DNA by breaking formalin-induced crosslinking between DNA and protein. Other methods increase the amount of digestive enzymes (e.g., proteinase-K) and lengthen the digestion times to breakdown more total amount of tissue (Ruane and Austin, 2017; this study). Additional steps, such as 'pre-digestion' steps with buffers for removing surface contaminants, have also been used when working with bone (Allentoft et al., 2015). All of these biochemical differences in workflows will inevitably have differing success rates depending on tissue type and quality (discussed below), as well as the type of loci being targeted (mtDNA, UCEs, AHEs, NPCGs, introns, whole genomes, etc.). While we are starting to better understand the damaging effects of formalin- and ethanol-fixation on museum specimens that hampers their input into evolutionary studies (Card et al., 2019), there is still a paucity for expectations during various parts of the project workflow when dealing with intractable specimens. Below, we provide expectations and tips based on our experiences in dealing with museum specimen genomics, in hopes that other researchers can maximize the data obtained from seemingly failed sequencing attempts.

DNA Extraction

The success of extracting quantifiable DNA of sufficient quality is dependent on numerous variables, most of which are unknown (e.g., specimen storage conditions since collection, preservation technique, time span from death to preservation, etc.). If available, a variety of tissue types (liver, muscle, and bone) from multiple individuals of different ages, should be used. Some studies have found that higher DNA yields were extracted from soft tissues, such as liver and muscle, compared to hard bone tissue (Zacho et al., 2021). Though, bone tissue can yield surprisingly high amounts of DNA (Zacho et al., 2021) of sufficient quality for sequencing, as we found from our homalopsid specimens, potentially due to greater protection against chemicals inside dense tissue. Specifically, the two bone samples used here, with few nuclear loci, provided ~66 and ~89% of the mitochondrial genome. We also note that while bone digestion protocols (Allentoft et al., 2015, 2018) may be more tedious than using Qiagen or phenol-chloroform procedures, they should not be overlooked as a tissue source. Additionally, this may prevent destructive sampling of fluid specimens if skeletons are already available in natural history collections (especially in snakes, which have hundreds of ribs). We used minimal amounts of bone (122 and 14 mg) for the bone extractions here.

Our protocols used proteinase-K as a tissue digesting agent. We note that proteinase-K often failed to digest tissues completely or even partially (over 48–72 h). This was true even if tissue was pulverized, with or without the aid of liquid nitrogen, to increase surface area. Vortexing samples every 6 h could facilitate tissue lysis, as well as adding 25 μl of proteinase-K every 24 h. While relationships between proteinase-K concentration and DNA yield for museum samples are lacking, formalin-fixed, paraffin-embedded tissue sections have provided increased DNA amounts when subjected to more digestion enzyme (Frazer et al., 2020). The failure of tissues to digest completely may result in clogging of filter tubes in spin columns, thus using multiple tubes per tissue specimen is recommended during such scenarios. While museum genomics studies typically report low DNA yields during extraction, our results highlight that even quantifications of <0.1 ng/μl can lead to successful sequencing of nuclear and mitochondrial loci. Nonetheless, combining aliquots of DNA extractions from the same individuals will increase the total gDNA and increase the likelihood of sequencing success, especially with specimens that were preserved more recently. We found more mtDNA was caught as bycatch when more total gDNA was used for sequencing (**Supplementary Figure 2**). Though, linear regression results were influenced by the inclusion of FMNH 251051; without this specimen, total gDNA was positively correlated with the number of AHEs and UCEs, but not the amount of mtDNA obtained (**Supplementary Figure 1**). Our study included samples with 3.78–932.85 ng, with successful sequencing at the lowest and highest ends of this range (**Supplementary Table 1**).

Dataset Integration

One of the greatest potentials of museum genomics is to combine these fluid specimen data with already published datasets. While studies will differ in scientific disciplines, aims, and taxonomic groups, projects should preemptively focus on how potentially-obtained data from intractable specimens will or can be combined with available datasets (or future datasets for that matter). In this study, we use the SqCL v2 (Singhal et al., 2017b) probe set as it targets three different locus types: UCEs, AHEs, and NPCGs common to squamate studies. Here, we leveraged data from studies that used traditional nuclear and mitochondrial markers (Deepak et al., 2021b), AHEs (Burbrink et al., 2020), or UCEs (Bernstein et al., unpublished data), each providing information that ultimately allowed us to determine the taxonomic affinity and biogeographic hypotheses of *H. periops*. We note that we obtained significantly more UCEs than AHEs and NPCGs, but this may be due to the SqCL probe set targeting thousands more UCEs compared to the other nuclear loci. Additionally, spiking libraries with unenriched pools can increase the likelihood of mitochondrial bycatch. We found that mitochondrial protein coding genes (e.g., 16S, ATPase, COX1, and cyt-b) were more often recovered than tRNAs (**Supplementary Table 6**). Future studies that standardize the nuclear and mitochondrial loci targeted, as well as specimen sampling, might identify patterns that show specific types of loci are more likely to be sequenced than others. Regardless of the loci being targeted, we suggest the sampling of a conspecific or congener from fresh tissues to aid in downstream analyses (see below).

Bioinformatics and Locus Acquisition

The bioinformatics phase of museum genomics depends on the amount of DNA that was successfully sequenced, but pipelines may still run to completion, even with extremely short DNA fragments. We only use one pipeline here (*Phyluce*) and found that all steps ran to completion with no errors. However, a resulting concatenated phylogeny recovered all museum specimens in a clade near the outgroup, with extremely long branches and showing no reasonable evolutionary relationships (**Supplementary Figure 4**). While this is no reflection on the pipeline itself, a variety of assembly methods and parameters can be used to optimize results (e.g., ABySS with different *k*-values and SPAdes with *k*-value averaging). Averaging of *k*-values using SPAdes in *Phyluce* worked best for our data. We note that assembly took >72 h (a common computer cluster time limit if nodes are public), thus we had to subsample to 7 million reads from our paired-end raw read data when using computer nodes with 182 GB and 28 cores. This was often needed for samples that had >15 million raw reads (~42% of our museum samples; **Supplementary Table 1**).

While we found that a mapping method using individual loci obtained from *Phyluce* was best for creating DNA alignments, we emphasize that loci should be visually checked for remaining adapter contamination. We found that even with two rounds of trimming and cleaning, partial or whole adapter sequences were still appended to one or both ends of many loci of only the formalin specimens. Sequencing a fresh sample to create a pseudoreference of the target loci (or using a reference genome if one is available) can allow for BLASTing of loci to the pseudoreference and eliminating remaining adapter contamination. Finally, we find that mapping to mitochondrial genomes in Geneious was the most efficient method for obtaining mtDNA loci, with results from Geneious obtaining better coverage of the mitochondrial genome across more specimens when compared to MitoFinder.

Terminology of Museum Genomics Specimens

In our study, we refer to our sampling as 'museum' or 'intractable' specimens. As biochemical protocols improve to increase the success rate of extracting and sequencing DNA from voucher specimens, the terminology we use to describe these processes may change as well. Currently, 'museum,' 'intractable,' 'degraded,' 'fixed,' 'preserved,' and 'historic' samples/specimens have all been used. However, this terminology may not always be accurate, and can be misleading or confusing. For example, the use of the word 'specimen' denotes the physical voucher animal, but this voucher is not always in a degraded or poor condition; contrarily, voucher specimens are often in great physical condition and it is the DNA and other molecular compounds that are damaged. Additionally, the term 'museum samples' does not distinguish between tissues of high quality (i.e., 'fresh' tissue) versus those that are degraded, as both tissue types likely came from natural history specimens. Furthermore, 'preserved' or 'fixed' samples may not include dried out samples that were not chemically treated, and the word 'historic' is ambiguous in relation to time. Though, the word 'intractable' may be informative in regards to sample/specimen quality; even as new protocols are developed, chemically-fixed samples will be more difficult (intractable) to sequence DNA from than freshly-extracted tissues prior to preservation. We also suggest, when known, stating the method of preservative (e.g., 'formalin-fixed' and 'ethanol-fixed' tissues) when referring to samples from liquid fixatives.

CONCLUSION

Museum genomics is rapidly advancing as new protocols are developed and resulting datasets are then used for a range of evolutionary studies (Guschanski et al., 2013; Mikheyev et al., 2017; Ruane and Austin, 2017; Allentoft et al., 2018; Deepak et al., 2018). Museum collections have been viewed as a window into the past of natural history, which is vital for understanding evolutionary processes, ecological dynamics, and global change (Suarez and Tsutsui, 2004; Bradley et al., 2014; Meineke et al., 2019). Morphological inspection of voucher specimens is indeed important for studying diversity and the processes that generate it, but may be hampered by poor specimen quality or changes in commonly-measured traits (Maayan et al., 2022). Opportunities provided by natural history collections have now expanded, with genetic and genomic sequencing of museum specimens facilitating species (re)discovery (Rasmussen et al., 2012) and determining past evolutionary dynamics (Mikheyev et al., 2017; Tsai et al., 2019; Roycroft et al., 2021). Our approach and methodology were successful for incorporating both nuclear and mitochondrial data for phylogenetics, the latter of which is often used in museum genomics studies due to its often-easier acquisition over nuclear loci. In this study, we maximized both nuclear and mitochondrial data from a seemingly-failed attempt at producing useful sequence data from a preserved specimen of *H. periops*, a poorly known snake endemic to an island biodiversity hotspot in Southeast Asia. The expectations of outcomes when conducting museum genomics projects are important for planning such studies, increasing the likelihood of success, and maximizing data use and interpretation of results. As more studies discuss both the successes, 'failures,' and difficulties when sequencing DNA from voucher specimens, the field of museum genomics will advance even further, as well as our knowledge of rare and even extinct species.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in the supplementary material and in online repositories. The names of the repository/repositories and accession number(s) can be found below: https://www.ncbi.nlm.nih.gov/, PRJNA796637, https://www.ncbi.nlm.nih.gov/, PRJNA792597 and https://www.zoobank.org/, F2AA44E2-D2EF-4747-972A-652C34C2C09D. The specimen data can also be found in

Supplementary Table 1. Any further queries should be directed to the corresponding author.

ETHICS STATEMENT

Ethical review and approval was not required for the animal study because this research was performed on tissues of specimens that were already deceased and on already-published data that is available on public repositories.

AUTHOR CONTRIBUTIONS

JB contributed to the data collection, analyses, visualization, original writing of manuscript, and project conceptualization. SR contributed to the original writing of manuscript, project conceptualization, and funding for project. Both authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by funds from Rutgers University – Newark provided to the Ruane Lab/SR, and by funds from the Field Museum of Natural History (including the Grainger Foundation and Negaunee Integrative Research Center) provided to SR.

ACKNOWLEDGMENTS

We acknowledge the following individuals and associated institutions for providing tissues used in DNA extraction and sequencing for this and the in-progress homalopsid study: Ned Gilmore (Academy of Natural Sciences of Philadelphia); Frank Burbrink, David Kizirian, and Lauren Vonnahme (American Museum of Natural History); Molly Hagemann (Bernice P. Bishop Museum); Rayna Bell, Lauren Scheinberg, and Erica Ely (California Academy of Sciences); Jennifer Sheridan (Carnegie Museum of Natural History); José Rosado (Museum of Comparative Zoology); David Blackburn and Coleman Sheehy III (Florida Museum of Natural History); Alan Resetar and Joshua Mata (Field Museum of Natural History); Rafe Brown and Jeffrey Weinell (KU Biodiversity Institute & Natural History Museum); Lee Grismer (La Sierra University Herpetology Collection); Christopher Austin (Louisiana State University Museum of Natural Science); Jadranka Rota and Maria Mostadius (Lund Museum of Zoology); Kevin de Queiroz and Addison Wynn (National Museum of Natural History); Gavin Dally (Museum and Art Gallery of the Northern Territory); Jimmy McGuire and Carol Spencer (Museum of Vertebrate Zoology); Esther Dondorp and Wendy van Bohemen (Naturalis Biodiversity Center); Bryan Stuart (North Carolina Museum of Natural Sciences); Patrick Couper and Andrew Amey (Queensland Museum); Bob Murphy (Royal Ontario Museum); Sally South (South Australian Museum); Bo Delling, Sven Kullander, and Andrea Hennyey (Swedish Museum of Natural History); Paul Doughty and Rebecca Bray (Western Australian Museum); Gregory Watkins-Colwell (Yale Peabody Museum of Natural History), and Chan Kin Onn (Lee Kong Chian Natural History Museum). We thank David Kizirian and Lauren Vonnahme for help with tissue importation from the South Australian Museum and Western Australian Museum, Justin Davis and Juan Flores (Rutgers University – Newark) for helpful conservations on formalinized tissues, and Huri Mucahit and Edward Bonder (Rutgers University – Newark) for the use of facilities related to DNA extractions. We acknowledge Jackson Roberts, Edward A. Myers, and Marcelo Gehara for assistance with bioinformatic pipelines and R coding. Finally, we thank W. Hosmer, the late Robert F. Inger, Tan Fui Lian, and the Sabah Parks field research team and colleagues for the collection of the original *H. periops* specimens during their field work in Borneo.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fevo.2022.893088/full#supplementary-material

Supplementary Figure 1 | Linear regression models for nuclear and mitochondrial data, without the inclusion of *Hydrablabes periops* FMNH 251051. R^2, *F*-, and *p*-values are given (α = 0.05). Trend lines shown in red.

Supplementary Figure 2 | Linear regression models for nuclear and mitochondrial data, with *H. periops* FMNH 251051 included in linear models. R^2, *F*-, and *p*-values are given (α = 0.05). Trend lines shown in red.

Supplementary Figure 3 | Maximum likelihood trees of BDNF, ND4, cyt-b, and a concatenated tree of these loci, using sequence data from Deepak et al. (2021a). Values at nodes represent SH-aLRT (left) and UF bootstraps (right); scale bar in number of substitutions per site.

Supplementary Figure 4 | Phylogeny of Homalopsidae (fresh and degraded specimens) using a concatenated dataset of all loci obtained from the *Phyluce* pipeline. Degraded (Museum) specimens are shown in pink, with fresh (non-degraded specimens with high quality DNA) colored in black. Scale bar in number of substitutions per site.

Supplementary Table 1 | Specimen information, read data, and adapter and barcode sequences for all specimens used in this study. Raw read data, number of nuclear loci obtained from Geneious, and Qubit quantifications are given for museum samples only.

Supplementary Table 2 | Morphological data recorded for *Hydrablabes periops* specimens.

Supplementary Table 3 | Node calibrations used for treePL divergence dating for the AHE Squamata phylogeny. Fossil calibrations used, respective minimum and maximum fossil ages, and their uses in previous studies are obtained from Burbrink et al. (2020).

Supplementary Table 4 | Nuclear loci from the SqCL probe set that were extracted using BLAST in Geneious. Locus lengths are given in base pairs (bp).

Supplementary Table 5 | Successful MitoFinder results for museum specimens. Taxon, voucher ID, mitochondrial gene, and locus lengths (base pairs; bp) are given.

Supplementary Table 6 | Percentages and lengths (base pairs; bp) of mitochondrial reference genome and mitochondrial gene regions recovered for museum specimens using the mapping method in Geneious.

REFERENCES

Alencar, L. R. V., Quental, T. B., Grazziotin, F. G., Alfaro, M. L., Martins, M., Venzon, M., et al. (2016). Diversification in vipers: phylogenetic relationships, time of divergence and shifts in speciation rates. *Mol. Phylogenet. Evol.* 105, 50–62. doi: 10.1016/j.ympev.2016.07.029

Alfaro, M. E., and Arnold, S. J. (2001). Molecular systematics and evolution of regina and the thamnophiine snakes. *Mol. Phylogenet. Evol.* 21, 408–423. doi: 10.1006/mpev.2001.1024

Allentoft, M. E., Rasmussen, A. R., and Kristensen, H. V. (2018). Centuries-Old DNA from an extinct population of aesculapian snake (zamenis longissimus) offers new phylogeographic insight. *Diversity* 10, 1–14. doi: 10.3390/d10010014

Allentoft, M. E., Sikora, M., Sjögren, K.-G., Rasmussen, S., Rasmussen, M., Stenderup, J., et al. (2015). Population genomics of bronze age eurasia. *Nature* 522, 167–172. doi: 10.1038/nature14507

Allio, R., Schomaker-Bastos, A., Romiguier, J., Prosdocimi, F., Nabholz, B., and Delsuc, F. (2020). MitoFinder: efficient automated large-scale extraction of mitogenomic data in target enrichment phylogenomics. *Mol. Ecol. Resour.* 20, 892–905. doi: 10.1111/1755-0998.13160

Álvarez-Carretero, S., Tamuri, A. U., Battini, M., Nascimento, F. F., Carlisle, E., Asher, R. J., et al. (2021). A species-level timeline of mammal evolution integrating phylogenomic data. *Nature* 2021, 1–8. doi: 10.1038/s41586-021-04341-1

Andrews, S. (2010). *FastQC: A Quality Control Tool For High Throughput Sequence Data*. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/ (accessed January 1, 2022)

Appleyard, S. A., Maher, S., Pogonoski, J. J., Bent, S. J., Chua, X.-Y., and McGrath, A. (2021). Assessing DNA for fish identifications from reference collections: the good, bad and ugly shed light on formalin fixation and sequencing approaches. *J. Fish Biol.* 98, 1421–1432. doi: 10.1111/jfb.14687

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., et al. (2012). SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* 19, 455–477. doi: 10.1089/cmb.2012.0021

Bouckaert, R. R., and Drummond, A. J. (2017). bModelTest: bayesian phylogenetic site model averaging and model comparison. *BMC Evol. Biol.* 17:42. doi: 10.1186/s12862-017-0890-6

Bradley, R. D., Bradley, L. C., Garner, H. J., and Baker, R. J. (2014). Assessing the value of natural history collections and addressing issues regarding long-term growth and care. *BioScience* 64, 1150–1158. doi: 10.1093/biosci/biu166

Burbrink, F., Grazziotin, F., Pyron, R., Cundall, D., Donnellan, S., Irish, F., et al. (2019). Data from: interrogating genomic-scale data for squamata (lizards, snakes, and amphisbaenians) shows no support for key traditional morphological relationships. *Dryad Digital Rep.* 2019:sm6jb0. doi: 10.5061/dryad.sm6jb0p

Burbrink, F. T., Bernstein, J. M., Kuhn, A., Gehara, M., and Ruane, S. (2021). Ecological divergence and the history of gene flow in the nearctic milksnakes (*Lampropeltis triangulum* complex). *Syst. Biol.* 2021:syab093. doi: 10.1093/sysbio/syab093

Burbrink, F. T., Grazziotin, F. G., Pyron, R. A., Cundall, D., Donnellan, S., Irish, F., et al. (2020). Interrogating genomic-scale data for squamata (lizards, snakes, and amphisbaenians) shows no support for key traditional morphological relationships. *Syst. Biol.* 69, 502–520. doi: 10.1093/sysbio/syz062

Campos, P. F., and Gilbert, T. M. P. (2012). "DNA extraction from formalin-fixed material," in *Ancient DNA: Methods and Protocols Methods in Molecular Biology*, eds B. Shapiro and M. Hofreiter (Totowa, NJ: Humana Press), 81–85. doi: 10.1007/978-1-61779-516-9_11

Card, D. C., Adams, R. H., Schield, D. R., Perry, B. W., Corbin, A. B., Pasquesi, G. I. M., et al. (2019). Genomic basis of convergent island phenotypes in boa constrictors. *Genome Biol. Evol.* 11, 3123–3143. doi: 10.1093/gbe/evz226

Card, D. C., Shapiro, B., Giribet, G., Moritz, C., and Edwards, S. V. (2021). Museum genomics. *Annu. Rev. Genet.* 55, 633–659. doi: 10.1146/annurev-genet-071719-020506

Das, I. (2018). *A Naturalist's Guide to the Snakes of Southeast Asia*, 2nd Edn. Oxford, UK: John Beaufoy Publishing.

Das, I., and Wong, J. W. (2021). Predation on gonocephalus liogaster (agamidae) by ptyas carinata (colubridae) in sarawak. *Borneo. Herpetol. Notes* 13, 349–351.

de Bruyn, M., Stelbrink, B., Morley, R. J., Hall, R., Carvalho, G. R., Cannon, C. H., et al. (2014). Borneo and indochina are major evolutionary hotspots for southeast asian biodiversity. *Syst. Biol.* 63, 879–901. doi: 10.1093/sysbio/syu047

de Queiroz, A., Lawson, R., and Lemos-Espinal, J. A. (2002). Phylogenetic relationships of north American garter snakes (thamnophis) based on four mitochondrial genes: how much DNA sequence is enough? *Mol. Phylogenet. Evol.* 22, 315–329. doi: 10.1006/mpev.2001.1074

De Vos, J. M., Joppa, L. N., Gittleman, J. L., Stephens, P. R., and Pimm, S. L. (2015). Estimating the normal background rate of species extinction. *Conserv. Biol.* 29, 452–462. doi: 10.1111/cobi.12380

Deepak, V., Cooper, N., Poyarkov, N. A., Kraus, F., Burin, G., Das, A., et al. (2021a). Multilocus phylogeny, natural history traits and classification of natricine snakes (serpentes: natricinae). *Zool. J. Linn. Soc.* 2021:zlab099. doi: 10.1093/zoolinnean/zlab099

Deepak, V., Maddock, S. T., Williams, R., Nagy, Z. T., Conradie, W., Rocha, S., et al. (2021b). Molecular phylogenetics of sub-saharan African natricine snakes, and the biogeographic origins of the seychelles endemic lycognathophis seychellensis. *Mol. Phylogenet. Evol.* 161:107152. doi: 10.1016/j.ympev.2021.107152

Deepak, V., Ruane, S., and Gower, D. J. (2018). A new subfamily of fossorial colubroid snakes from the western ghats of peninsular India. *J. Nat. Hist.* 52, 2919–2934. doi: 10.1080/00222933.2018.1557756

Del Fabbro, C., Scalabrin, S., Morgante, M., and Giorgi, F. M. (2013). An extensive evaluation of read trimming effects on illumina NGS data analysis. *PLoS One* 8:e85024. doi: 10.1371/journal.pone.0085024

Do, H., and Dobrovic, A. (2015). Sequence artifacts in DNA from formalin-fixed tissues: causes and strategies for minimization. *Clin. Chem.* 61, 64–71. doi: 10.1373/clinchem.2014.223040

Esquerré, D., Donnellan, S., Brennan, I. G., Lemmon, A. R., Moriarty Lemmon, E., Zaher, H., et al. (2020). Phylogenomics, biogeography, and morphometrics reveal rapid phenotypic evolution in pythons after crossing wallace's line. *Syst. Biol.* 69, 1039–1051. doi: 10.1093/sysbio/syaa024

Faircloth, B. C. (2011). *Illumiprocessor - Software For Illumina Read Quality Filtering*. Availble online at: https://github.com/faircloth-lab/illumiprocessor (accessed November 1, 2021).

Faircloth, B. C. (2016). PHYLUCE is a software package for the analysis of conserved genomic loci. *Bioinformatics* 32, 786–788.

Faircloth, B. C., McCormack, J. E., Crawford, N. G., Harvey, M. G., Brumfield, R. T., and Glenn, T. C. (2012). Ultraconserved elements anchor thousands of genetic markers spanning multiple evolutionary timescales. *Syst. Biol.* 61, 717–726. doi: 10.1093/sysbio/sys004

France, S. C., and Kocher, T. D. (1996). DNA sequencing of formalin-fixed crustaceans from archival research collections. *Mol. Mar. Biol. Biotechnol.* 5, 304–313.

Frazer, Z., Yoo, C., Sroya, M., Bellora, C., DeWitt, B. L., Sanchez, I., et al. (2020). Effect of different proteinase k digest protocols and deparaffinization methods on yield and integrity of DNA extracted from formalin-fixed, paraffin-embedded tissue. *J. Histochem. Cytochem.* 68, 171–184. doi: 10.1369/0022155420906234

Fukuyama, R., Fukuyama, I., Kurita, T., Kojima, Y., Hossman, M. Y., Noda, A., et al. (2021). New herpetofaunal records from gunung mulu national park and its surrounding areas in borneo. *Herpetozoa* 2021, 89–97.

Guindon, S., Dufayard, J.-F., Lefort, V., Anisimova, M., Hordijk, W., and Gascuel, O. (2010). New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst. Biol.* 59, 307–321. doi: 10.1093/sysbio/syq010

Guo, P., Liu, Q., Xu, Y., Jiang, K., Hou, M., Ding, L., et al. (2012). Out of Asia: natricine snakes support the cenozoic beringian dispersal hypothesis. *Mol. Phylogenet. Evol.* 63, 825–833. doi: 10.1016/j.ympev.2012.02.021

Guo, P., Zhu, F., and Liu, Q. (2019). A new member of the genus sinonatrix (serpentes: colubridae) from western China. *Zootaxa* 2019:4623. doi: 10.11646/zootaxa.4623.3.5

Guo, P., Zhu, F., Liu, Q., Wang, P., Che, J., and Nguyen, T. Q. (2020). Out of the hengduan mountains: molecular phylogeny and historical biogeography of the asian water snake genus trimerodytes (squamata: colubridae). *Mol. Phylogenet. Evol.* 152:106927. doi: 10.1016/j.ympev.2020.106927

Guschanski, K., Krause, J., Sawyer, S., Valente, L. M., Bailey, S., Finstermeier, K., et al. (2013). Next-generation museomics disentangles one of the largest primate radiations. *Syst. Biol.* 62, 539–554. doi: 10.1093/sysbio/syt018

Hahn, C., Bachmann, L., and Chevreux, B. (2013). Reconstructing mitochondrial genomes directly from genomic next-generation sequencing reads—a baiting and iterative mapping approach. *Nucleic Acids Res.* 41:e129. doi: 10.1093/nar/gkt371

Hall, R. (2009). Southeast Asia's changing palaeogeography. *Blumea Biodivers Evol. Biogeogr. Plants* 54, 148–161. doi: 10.3767/000651909X475941

Hime, P. M., Lemmon, A. R., Lemmon, E. C. M., Prendini, E., Brown, J. M., Thomson, R. C., et al. (2021). Phylogenomics reveals ancient gene tree discordance in the amphibian tree of life. *Syst. Biol.* 70, 49–66. doi: 10.1093/sysbio/syaa034

Hoang, D. T., Chernomor, O., von Haeseler, A., Minh, B. Q., and Vinh, L. S. (2018). UFBoot2: improving the ultrafast bootstrap approximation. *Mol. Biol. Evol.* 35, 518–522. doi: 10.1093/molbev/msx281

Husson, L., Boucher, F. C., Sarr, A.-C., Sepulchre, P., and Cahyarini, S. Y. (2020). Evidence of sundaland's subsidence requires revisiting its biogeography. *J. Biogeogr.* 47, 843–853. doi: 10.1111/jbi.13762

Hykin, S. M., Bi, K., and McGuire, J. A. (2015). Fixing formalin: a method to recover genomic-scale DNA sequence data from formalin-fixed museum specimens using high-throughput sequencing. *PLoS One* 10:e0141579. doi: 10.1371/journal.pone.0141579

Jones, M. E., Anderson, C. L., Hipsley, C. A., Müller, J., Evans, S. E., and Schoch, R. R. (2013). Integration of molecules and new fossils supports a triassic origin for lepidosauria (lizards, snakes, and tuatara). *BMC Evol. Biol.* 13:208. doi: 10.1186/1471-2148-13-208

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

Kindler, C., Bringsøe, H., and Fritz, U. (2014). Phylogeography of grass snakes (natrix natrix) all around the baltic sea: implications for the holocene colonization of fennoscandia. *Amphib Reptil.* 35, 413–424. doi: 10.1163/15685381-00002962

Köhler, G., Khaing, K. P. P., Than, N. L., Baranski, D., Schell, T., Greve, C., et al. (2021). A new genus and species of mud snake from myanmar (reptilia, squamata, homalopsidae). *Zootaxa* 4915, 301–325. doi: 10.11646/zootaxa.4915.3.1

Lalronunga, S., Lalrinchhana, C., Vanramliana, V., Das, A., Gower, D. J., and Deepak, V. (2020). A multilocus molecular perspective on the systematics of the poorly known northeast indian colubrid snakes blythia reticulata (blyth, 1854), b. hmuifang vogel, lalremsanga amp; vanlalhrima, 2017, and hebius xenura (wall, 1907). *Zootaxa* 2020:4768. doi: 10.11646/zootaxa.4768.2.2

Lanfear, R., Frandsen, P. B., Wright, A. M., Senfeld, T., and Calcott, B. (2017). PartitionFinder 2: new methods for selecting partitioned models of evolution for molecular and morphological phylogenetic analyses. *Mol. Biol. Evol.* 34, 772–773. doi: 10.1093/molbev/msw260

Lemmon, A. R., Emme, S. A., and Lemmon, E. M. (2012). Anchored hybrid enrichment for massively high-throughput phylogenomics. *Syst. Biol.* 61, 727–744. doi: 10.1093/sysbio/sys049

Linck, E. B., Hanna, Z. R., Sellas, A., and Dumbacher, J. P. (2017). Evaluating hybridization capture with RAD probes as a tool for museum genomics with historical bird specimens. *Ecol. Evol.* 7, 4755–4767. doi: 10.1002/ece3.3065

Lohse, M., Bolger, A. M., Nagel, A., Fernie, A. R., Lunn, J. E., Stitt, M., et al. (2012). RobiNA: a user-friendly, integrated software solution for RNA-Seq-based transcriptomics. *Nucleic Acids Res.* 40, W622–W627. doi: 10.1093/nar/gks540

Maayan, I., Reynolds, R. G., Goodman, R. M., Hime, P. M., Bickel, R., Luck, E. A., et al. (2022). Fixation and preservation contribute to distortion invertebrate museum specimens: a 10-year study with the lizard anolis sagrei. *Biol. J. Linn. Soc.* 2022:blac040.

McCormack, J. E., Tsai, W. L. E., and Faircloth, B. C. (2016). Sequence capture of ultraconserved elements from bird museum specimens. *Mol. Ecol. Resour.* 16, 1189–1203. doi: 10.1111/1755-0998.12466

McGuire, J. A., Cotoras, D. D., O'Connell, B., Lawalata, S. Z. S., Wang-Claypool, C. Y., Stubbs, A., et al. (2018). Squeezing water from a stone: high-throughput sequencing from a 145-year old holotype resolves (barely) a cryptic species problem in flying lizards. *PeerJ* 6:e4470. doi: 10.7717/peerj.4470

McVay, J. D., and Carstens, B. (2013). Testing monophyly without well-supported gene trees: evidence from multi-locus nuclear data conflicts with existing taxonomy in the snake tribe thamnophiini. *Mol. Phylogenet. Evol.* 68, 425–431. doi: 10.1016/j.ympev.2013.04.028

McVay, J. D., Flores-Villela, O., and Carstens, B. (2015). Diversification of north american natricine snakes. *Biol. J. Linn. Soc.* 116, 1–12. doi: 10.1111/bij.12558

Meineke, E. K., Davies, T. J., Daru, B. H., and Davis, C. C. (2019). Biological collections for understanding biodiversity in the anthropocene. *Philos. Trans. R. Soc. B Biol. Sci.* 374:20170386. doi: 10.1098/rstb.2017.0386

Mikheyev, A. S., Zwick, A., Magrath, M. J. L., Grau, M. L., Qiu, L., Su, Y. N., et al. (2017). Museum genomics confirms that the lord howe island stick insect survived extinction. *Curr. Biol.* 27, 3157–3161.e4. doi: 10.1016/j.cub.2017.08.058

Mocquard, F. (1890). Recherches sur la faune herpétologique des îles de bornèo et de palawan. *Arch. Muséum Natl. Hist. Nat. Paris* 3, 115–168.

Murphy, J. C., and Voris, H. K. (2014). A checklist and key to the homalopsid snakes (reptilia, squamata, serpentes), with the description of new genera. *Fieldiana Life Earth Sci.* 2014, 1–43. doi: 10.3158/2158-5520-14.8.1

Nagy, Z. T., Vidal, N., Vences, M., Branch, W. R., Pauwels, O. S. G., Wink, M., et al. (2005). "Molecular systematics of african colubroidea (squamata: serpentes)," in *African Biodiversity*, eds B. A. Huber, B. J. Sinclair, and K.-H. Lampe (Boston, MA: Springer), 221–228. doi: 10.1007/0-387-24320-8_20

Nguyen, L.-T., Schmidt, H. A., von Haeseler, A., and Minh, B. Q. (2015). IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* 32, 268–274. doi: 10.1093/molbev/msu300

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

O'Connell, K. A., Mulder, K. P., Wynn, A., Queiroz, K., and Bell, R. C. (2021). Genomic library preparation and hybridization capture of formalin-fixed tissues and allozyme supernatant for population genomics and considerations for combining capture- and RADseq-based single nucleotide polymorphism data sets. *Mol. Ecol. Resour.* 2021:13481. doi: 10.1111/1755-0998.13481

Ogilvie, H. A., Bouckaert, R. R., and Drummond, A. J. (2017). StarBEAST2 brings faster species tree inference and accurate estimates of substitution rates. *Mol. Biol. Evol.* 34, 2101–2114. doi: 10.1093/molbev/msx126

Pimm, S. L., Jenkins, C. N., Abell, R., Brooks, T. M., Gittleman, J. L., Joppa, L. N., et al. (2014). The biodiversity of species and their rates of extinction, distribution, and protection. *Science* 344:1246752. doi: 10.1126/science.1246752

Portik, D. M., and Wiens, J. J. (2021). Do alignment and trimming methods matter for phylogenomic (UCE) analyses? *Syst. Biol.* 70, 440–462. doi: 10.1093/sysbio/syaa064

Pyron, R. A., Beamer, D. A., Holzheuser, C. R., Lemmon, E. M., Lemmon, A. R., Wynn, A. H., et al. (2022). Contextualizing enigmatic extinctions using genomic DNA from fluid-preserved museum specimens of *Desmognathus* salamanders. *Conserv. Genet.* 23, 375–386. doi: 10.1007/s10592-021-01424-4

Pyron, R. A., Kandambi, H. K. D., Hendry, C. R., Pushpamal, V., Burbrink, F. T., and Somaweera, R. (2013). Genus-level phylogeny of snakes reveals the origins of species richness in Sri Lanka. *Mol. Phylogenet. Evol.* 66, 969–978. doi: 10.1016/j.ympev.2012.12.004

Quah, E. S. H., Grismer, L. L., Lim, K. K. P., Anuar, M. S. S., and Imbun, A. Y. (2019). A taxonomic reappraisal of the smooth slug snake asthenodipsas laevis (boie, 1827) (squamata: pareidae) in borneo with the description of two new species. *Zootaxa* 2019:4646. doi: 10.11646/zootaxa.4646.3.4

R Core Team (2021). *R: A Language And Environment For Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Rambaut, A., Drummond, A. J., Xie, D., Baele, G., and Suchard, M. A. (2018). Posterior summarization in bayesian phylogenetics using Tracer 1.7. *Syst. Biol.* 67, 901–904. doi: 10.1093/sysbio/syy032

Rancilhac, L., Bruy, T., Scherz, M. D., Pereira, E. A., Preick, M., Straube, N., et al. (2020). Target-enriched DNA sequencing from historical type material enables a partial revision of the madagascar giant stream frogs (genus mantidactylus). *J. Nat. Hist.* 54, 87–118. doi: 10.1080/00222933.2020.1748243

Rasmussen, A. R., Elmberg, J., Sanders, K. L., and Gravlund, P. (2012). Rediscovery of the rare sea snake hydrophis parviceps smith 1935: identification and conservation status. *Copeia* 2012, 276–282. doi: 10.1643/CH-11-116

Ren, J., Wang, K., Guo, P., Wang, Y.-Y., Nguyen, T., and Li, J. (2019). On the generic taxonomy of opisthotropis balteata (cope, 1895) (squamata: colubridae: natricinae): taxonomic revision of two natricine genera. *Asian Herpetol. Res.* 10, 105–128. doi: 10.16373/j.cnki.ahr.180091

Rohland, N., Siedel, H., and Hofreiter, M. (2004). Nondestructive DNA extraction method for mitochondrial DNA analyses of museum specimens. *BioTechniques* 36, 814–821. doi: 10.2144/04365ST05

Roycroft, E., MacDonald, A. J., Moritz, C., Moussalli, A., Portela Miguez, R., and Rowe, K. C. (2021). Museum genomics reveals the rapid decline and extinction of Australian rodents since European settlement. *Proc. Natl. Acad. Sci. U.S.A.* 118:e2021390118. doi: 10.1073/pnas.2021390118

Ruane, S. (2021). New data from old specimens. *Ichthyol. Herpetol.* 109, 392–396. doi: 10.1643/t2019293

Ruane, S., and Austin, C. C. (2017). Phylogenomics using formalin-fixed and 100+ year-old intractable natural history specimens. *Mol. Ecol. Resour.* 17, 1003–1008. doi: 10.1111/1755-0998.12655

Sarr, A.-C., Husson, L., Sepulchre, P., Pastier, A.-M., Pedoja, K., Elliot, M., et al. (2019). Subsiding sundaland. *Geology* 47, 119–122. doi: 10.1130/G45629.1

Schield, D. R., Pasquesi, G. I. M., Perry, B. W., Adams, R. H., Nikolakis, Z. L., Westfall, A. K., et al. (2020). Snake recombination landscapes are concentrated in functional regions despite PRDM9. *Mol. Biol. Evol.* 37, 1272–1294. doi: 10.1093/molbev/msaa003

Simmons, J. E. (2014). *Fluid Preservation: A Comprehensive Reference*. Lanhan, MD: Rowman & Littlefield.

Simpson, J. T., Wong, K., Jackman, S. D., Schein, J. E., Jones, S. J. M., and Birol, I. (2009). ABySS: a parallel assembler for short read sequence data. *Genome Res.* 19, 1117–1123. doi: 10.1101/gr.089532.108

Singhal, S., Grundler, M., Colli, G., and Rabosky, D. L. (2017a). Data from: squamate conserved loci (SqCL): a unified set of conserved loci for phylogenomics and population genetics of squamate reptiles. *Dryad Digital Repository* 2017:r0q02. doi: 10.5061/dryad.r0q02

Singhal, S., Grundler, M., Colli, G., and Rabosky, D. L. (2017b). Squamate conserved loci (SqCL: a unified set of conserved loci for phylogenomics and population genetics of squamate reptiles. *Mol. Ecol. Resour.* 17, e12–e24. doi: 10.1111/1755-0998.12681

Smith, S. A., and O'Meara, B. C. (2012). treePL: divergence time estimation using penalized likelihood for large phylogenies. *Bioinformatics* 28, 2689–2690. doi: 10.1093/bioinformatics/bts492

Sodhi, N. S., Koh, L. P., Brook, B. W., and Ng, P. K. L. (2004). Southeast Asian biodiversity: an impending disaster. *Trends Ecol. Evol.* 19, 654–660. doi: 10.1016/j.tree.2004.09.006

Strang, K., and Rusli, N. (2021). "The challenges of conserving biodiversity: a spotlight on southeast asia," in *Wildlife Biodiversity Conservation: Multidisciplinary and Forensic Approaches*, eds S. C. Underkoffler and H. R. Adams (Cham: Springer International Publishing), 47–66. doi: 10.1007/978-3-030-64682-0_3

Stuckert, A. M. M., Chouteau, M., McClure, M., LaPolice, T. M., Linderoth, T., Nielsen, R., et al. (2021). The genomics of mimicry: gene expression throughout development provides insights into convergent and divergent phenotypes in a müllerian mimicry system. *Mol. Ecol.* 30, 4039–4061. doi: 10.1111/mec.16024

Stuebing, R. B., Inger, R. F., and Lardner, B. (2014). *A Field Guide to the Snakes of Borneo*. Kota Kinabalu: Natural History Publications.

Suarez, A. V., and Tsutsui, N. D. (2004). The value of museum collections for research and society. *BioScience* 54, 66–74.

Tsai, W. L. E., Mota-Vargas, C., Rojas-Soto, O., Bhowmik, R., Liang, E. Y., Maley, J. M., et al. (2019). Museum genomics reveals the speciation history of dendrortyx wood-partridges in the mesoamerican highlands. *Mol. Phylogenet. Evol.* 136, 29–34. doi: 10.1016/j.ympev.2019.03.017

Uetz, P., Freed, P., Aguilar, R., and Hošek, J. (2021). *The Reptile Database*. Available online at: http://www.reptile-database.org (accessed February 1, 2022)

Voris, H. K. (2000). Maps of pleistocene sea levels in southeast asia: shorelines, river systems and time durations. *J. Biogeogr.* 27, 1153–1167. doi: 10.1046/j.1365-2699.2000.00489.x

Weinell, J. L., and Brown, R. M. (2018). Discovery of an old, archipelago-wide, endemic radiation of Philippine snakes. *Mol. Phylogenet. Evol.* 119, 144–150. doi: 10.1016/j.ympev.2017.11.004

Westeen, E. P., Durso, A. M., Grundler, M. C., Rabosky, D. L., and Davis Rabosky, A. R. (2020). What makes a fang? Phylogenetic and ecological controls on tooth evolution in rear-fanged snakes. *BMC Evol. Biol.* 20:80. doi: 10.1186/s12862-020-01645-0

Wingett, S. W., and Andrews, S. (2018). FastQ screen: a tool for multi-genome mapping and quality control. *F1000Research* 7:1338. doi: 10.12688/f1000research.15931.2

Wood, H. M., González, V. L., Lloyd, M., Coddington, J., and Scharff, N. (2018). Next-generation museum genomics: phylogenetic relationships among palpimanoid spiders using sequence capture techniques (araneae: palpimanoidea). *Mol. Phylogenet. Evol.* 127, 907–918. doi: 10.1016/j.ympev.2018.06.038

Yan, J., Li, H., and Zhou, K. (2008). Evolution of the mitochondrial genome in snakes: gene rearrangements and phylogenetic relationships. *BMC Geno.* 9:569. doi: 10.1186/1471-2164-9-569

Zacho, C. M., Bager, M. A., Margaryan, A., Gravlund, P., Galatius, A., Rasmussen, A. R., et al. (2021). Uncovering the genomic and metagenomic research potential in old ethanol-preserved snakes. *PLoS One* 16:e0256353. doi: 10.1371/journal.pone.0256353

Zaher, H., Yánez-Muñoz, M. H., Rodrigues, M. T., Graboski, R., Machado, F. A., Altamirano-Benavides, M., et al. (2018). Origin and hidden diversity within the poorly known galápagos snake radiation (serpentes: dipsadidae). *Syst. Biodivers.* 16, 614–642. doi: 10.1080/14772000.2018.1478910

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

Zimmermann, J., Hajibabaei, M., Blackburn, D. C., Hanken, J., Cantin, E., Posfai, J., et al. (2008). DNA damage in preserved specimens and tissue samples: a molecular assessment. *Front. Zool.* 5:18. doi: 10.1186/1742-9994-5-18

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

ORIGINAL RESEARCH
published: 30 June 2022
doi: 10.3389/fevo.2022.910009

Taxonomic Identification of Two Poorly Known Lantern Shark Species Based on Mitochondrial DNA From Wet-Collection Paratypes

Stefanie Agne[1], Gavin J. P. Naylor[2], Michaela Preick[1], Lei Yang[2], Ralf Thiel[3,4], Simon Weigmann[3,5], Johanna L. A. Paijmans[6], Axel Barlow[7], Michael Hofreiter[1] and Nicolas Straube[8]*

[1] Evolutionary Adaptive Genomics, Department of Mathematics and Natural Sciences, Institute for Biochemistry and Biology, University of Potsdam, Potsdam, Germany, [2] Florida Museum of Natural History, University of Florida, Gainesville, FL, United States, [3] Centre for Taxonomy and Morphology, Zoological Museum, Leibniz Institute for the Analysis of Biodiversity Change (LIB), Hamburg, Germany, [4] Department of Biology, Biodiversity Research, University of Hamburg, Hamburg, Germany, [5] Elasmo-Lab, Elasmobranch Research Laboratory, Hamburg, Germany, [6] Department of Zoology, University of Cambridge, Cambridge, United Kingdom, [7] School of Natural Sciences, Bangor University, Bangor, United Kingdom, [8] Department of Natural History, University Museum of Bergen, Bergen, Norway

OPEN ACCESS

Edited by:
Anchalee Aowphol,
Kasetsart University, Thailand

Reviewed by:
Kieren James Mitchell,
University of Otago, New Zealand
Vladimir A. Trifonov,
Institute of Molecular and Cellular Biology (RAS), Russia

***Correspondence:**
Nicolas Straube
Nicolas.Straube@uib.no

Specialty section:
This article was submitted to Phylogenetics, Phylogenomics, and Systematics, a section of the journal Frontiers in Ecology and Evolution

Received: 31 March 2022
Accepted: 25 May 2022
Published: 30 June 2022

Citation:
Agne S, Naylor GJP, Preick M, Yang L, Thiel R, Weigmann S, Paijmans JLA, Barlow A, Hofreiter M and Straube N (2022) Taxonomic Identification of Two Poorly Known Lantern Shark Species Based on Mitochondrial DNA From Wet-Collection Paratypes. Front. Ecol. Evol. 10:910009. doi: 10.3389/fevo.2022.910009

Etmopteridae (lantern sharks) is the most species-rich family of sharks, comprising more than 50 species. Many species are described from few individuals, and re-collection of specimens is often hindered by the remoteness of their sampling sites. For taxonomic studies, comparative morphological analysis of type specimens housed in natural history collections has been the main source of evidence. In contrast, DNA sequence information has rarely been used. Most lantern shark collection specimens, including the types, were formalin fixed before long-term storage in ethanol solutions. The DNA damage caused by both fixation and preservation of specimens has excluded these specimens from DNA sequence-based phylogenetic analyses so far. However, recent advances in the field of ancient DNA have allowed recovery of wet-collection specimen DNA sequence data. Here we analyse archival mitochondrial DNA sequences, obtained using ancient DNA approaches, of two wet-collection lantern shark paratype specimens, namely *Etmopterus litvinovi* and *E. pycnolepis,* for which the type series represent the only known individuals. Target capture of mitochondrial markers from single-stranded DNA libraries allows for phylogenetic placement of both species. Our results suggest synonymy of *E. benchleyi* with *E. litvinovi* but support the species status of *E. pycnolepis*. This revised taxonomy is helpful for future conservation and management efforts, as our results indicate a larger distribution range of *E. litvinovi*. This study further demonstrates the importance of wet-collection type specimens as genetic resource for taxonomic research.

Keywords: type specimens, *Etmopterus litvinovi*, *Etmopterus pycnolepis*, deep-sea sharks, archival DNA

INTRODUCTION

Shark diversity is poorly represented in the scientific literature. Shark biologists have tended to focus on a few easy-to-access taxa that are assumed to be representative of the groups to which they belong. For example, though there are more than 40 different species of deep-sea lantern sharks (genus *Etmopterus*), nearly a quarter of the 2082 publications devoted to Lantern shark biology (Pollerspöck and Straube, 2021) has focussed on a single species (*Etmopterus spinax*). Thus, most of the diversity of this group remains relatively unexplored (**Figure 1**). To make matters worse, a substantial fraction of lantern shark diversity is known only from formalin preserved type material that was collected prior to the advent of DNA sequencing. Hence, tissue sampling, common practice today for performing DNA sequence-based analysis such as DNA barcoding (Hebert et al., 2003), was not conducted and fixation in formaldehyde and preservation in ethanol causes DNA damage (Gilbert et al., 2007; Hoffman et al., 2015; Hykin et al., 2015; Stiller et al., 2016; McGuire et al., 2018; Hahn et al., 2021). This means that, while we know that the group has diversified extensively (e.g., Straube et al., 2011a; Ebert et al., 2016, 2021; White et al., 2017; Dolganov and Balanov, 2018), it has been hard to decipher how the different species are related to one another and how different ecological pressures have contributed to their diversification. Recently developed tools allow us to obtain DNA sequence data from formalin preserved animals (Gansauge et al., 2017; Hahn et al., 2021; Straube et al., 2021a). In the current contribution we have applied these tools to type material for two species of *Etmopterus* and show how the data collected have implications, not only for understanding their taxonomy and evolution, but also their ranges, which has consequences for their conservation and management. The genus is subdivided into four clades supported by both DNA sequence data and morphological characters (Straube et al., 2010). Morphological characters therefore allow for tentative assignments of species lacking DNA sequence information to one of the four clades. Our first target species, *Etmopterus litvinovi* (Smalleye lantern shark, Parin and Kotlyar, 1990) has been assigned to the *E. spinax* clade (Straube et al., 2010, 2011a) comprising 11 species today (Ebert et al., 2021). The presence and shape of flank markings, dark patterns above the pelvic fins, is a key character allowing for species-to-clade assignments in many *Etmopterus* species. While the character is not present in all species and ontogenetic stages, every species of the *E. lucifer* clade shows distinct flank markings characterised by anterior and posterior branches. Species of the *E. lucifer* clade can further be subdivided into three subclades based on length comparisons of the anterior and posterior flank mark branches (Ebert et al., 2021). The three subclades are the *E. lucifer*, the *E. molleri* and the *E. burgessi* subclades. The *E. lucifer* subclade includes the four species *E. brosei*, *E. lailae*, *E. lucifer* and *E. sculptus*. *E. alphus*, *E. brachyurus*, *E. bullisi*, *E. decacuspidatus*, *E. dislineatus*, *E. molleri,* and *E. samadiae* are the seven species assigned to the *E. molleri* subclade. The *E. burgessi* subclade comprises four species, namely *E. burgessi*, *E. evansi*, *E. marshae*, and *E. pycnolepis* (Ebert et al., 2021). *Etmopterus pycnolepis* (Dense-scale lantern shark, Kotlyar, 1990) is our second target species. Both *E. litvinovi* and *E. pycnolepis* are known from their type specimens only and little is known regarding their biology as they were hitherto sampled only once each in the Salas y Gómez and Nazca submarine ridges in the Southeast Pacific (Kotlyar, 1990; Ebert et al., 2013).

MATERIALS AND METHODS

Etmopterus litvinovi (Smalleye Lantern Shark)

This species is known from 32 type specimens housed in three different museum collections, the Laboratory of Ichthyology at the Zoological Institute of the Russian Academy of Sciences (ZIN), St. Petersburg, Russia (holotype: ZIN 49228; six paratypes: ZIN 49229–32), the Zoological Museum (ZMMU), Biological Faculty, M. V. Lomonosov Moscow State University, Moscow, Russia (21 paratypes ZMMU: P-17989–91; two paratypes P-18222) and the ichthyological collection of the Zoological Museum (ZMH) of the LIB in Hamburg, Germany [paratype ZMH 24994 (ex ISH 6-1989); paratype ZMH 24993 (ex ISH 5-1989)]. We sampled muscle tissue from the paratype specimen ZMH 24994 (**Figure 2A**) at the caudal peduncle using a biopsy needle for minimally invasive sampling. The tissue was preserved in the original preservation fluid of the storage container. The specimen was captured at 25°21′S and 85°8′W at a depth of 720 m on 24.04.1987. It is a juvenile male of 445 mm total length (Thiel et al., 2009; Straube et al., 2011a). Although not explicitly mentioned in the original description, or tested by us, the overall condition of the specimen indicates a fixation in formaldehyde: both body and eyes do not show bleaching of exclusively ethanol preserved samples. Furthermore, the common procedure during research cruises at the time of sampling was a fixation of specimens in 4% formaldehyde and long-term preservation in 70% ethanol.

Laboratory steps and analysis of test-sequencing data of this specimen is described in detail in Straube et al. (2021a). The sample was incubated in a GuSCN-based buffer (Rohland et al., 2004) applying the protocol by Dabney et al. (2013) for DNA purification. A single-stranded DNA library was then constructed, and test-sequencing was performed to check for the ratio of target DNA and contamination. After detection of endogenous DNA in the test-sequencing dataset, target capture for mitochondrial DNA was performed using home-made baits. These were generated from long-range PCR products amplified from the DNA of *Etmopterus* cf. *molleri* tissue housed in the tissue sample collection of the Bavarian State Collection of Zoology (registration number: Ich-P-CH-0264). For the long-range PCR protocol and primers see Straube et al. (2021a). Hybridisation capture was then performed following the protocol of González Fortes and Paijmans (2019), where the single-stranded library is mixed with the denatured bait library after addition of blocking oligos. Hybridisation of target DNA to baits was carried out for 24 h at 65°C. The captured library was then amplified, and the capture procedure and amplification repeated. The resulting double captured library was then sequenced using

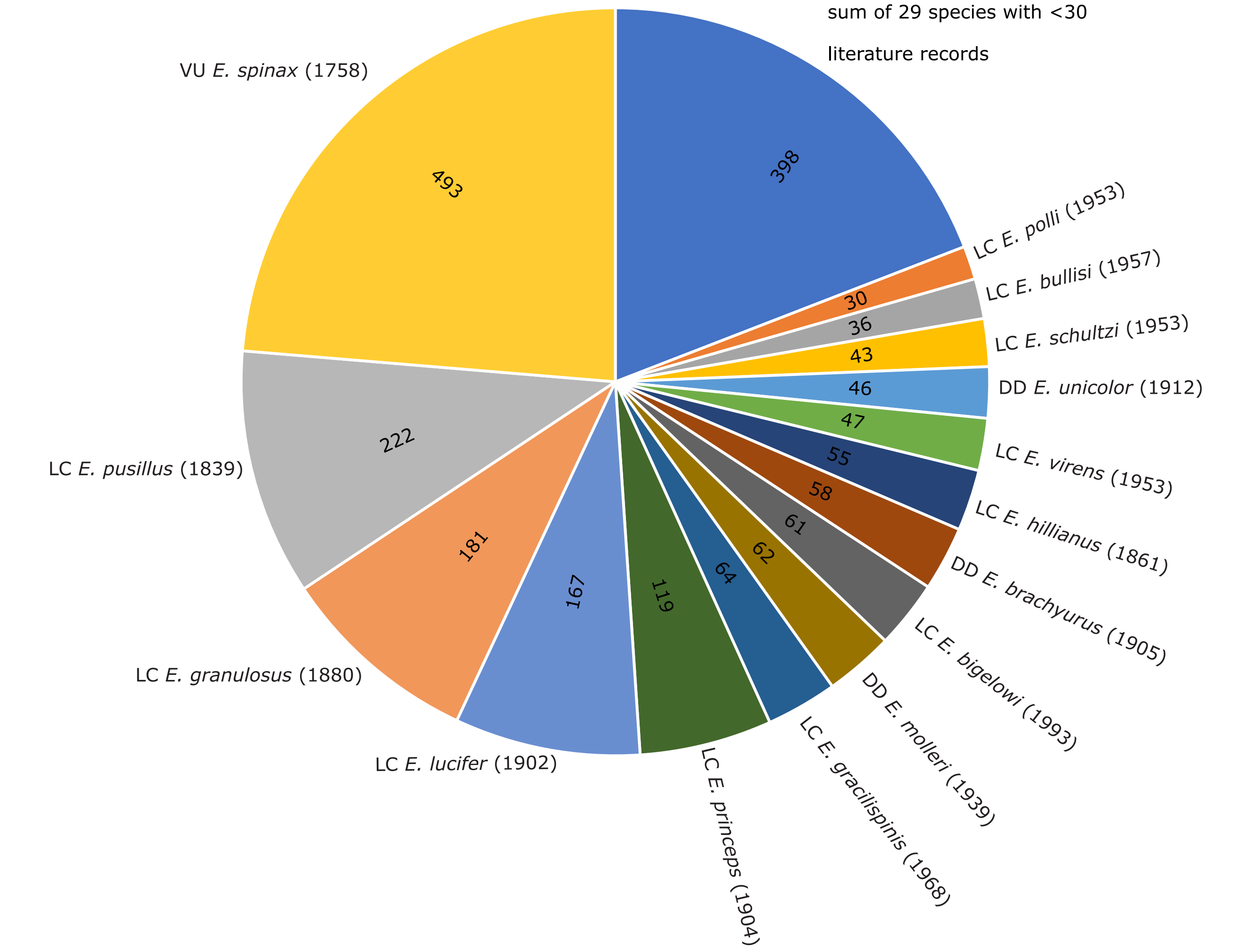


FIGURE 1 | Pie chart showing the number of scientific publications listed per *Etmopterus* species in the bibliographic database Shark References. Species with a total number of publications below 30 are summarised. Global IUCN Red List of Threatened Species status (VU = vulnerable; LC = least concern; DD = data deficient) is given in front of species names, year of description in brackets.

custom sequencing and index 2 read primers (Gansauge and Meyer, 2013; Paijmans et al., 2017) on an Illumina® MiniSeq instrument. We used a mid-output kit in a pool of double indexed samples.

Paired-end raw reads were quality and adapter trimmed with Cutadapt v.1.16 (Martin, 2011) using default settings. The iterative mapping algorithm MitoBim v. 1.9.1 (Hahn et al., 2013) was then used to reconstruct the mitochondrial genome sequence, using default settings and Genbank entry KU892588 (*Etmopterus pusillus*; Chen et al., 2016) as reference for initial baiting. Annotation was performed by aligning the paratype consensus sequence to KU892588 in Geneious® Prime 2021.1 (Biomatters Ltd. Auckland, New Zealand), and checked for internal stop codons. Protein coding genes could not be fully reconstructed. The tRNA-Phe and tRNA-Val transfer RNAs, and the 12S and 16s ribosomal RNAs could be completely reconstructed and were therefore extracted for phylogenetic analysis (2676 bp in total). Reads used in the last iteration of Mitobim were mapped back to the mitochondrial genome consensus sequence as well as to the tRNA and rRNA sequences using BWA aln v.0.7.17 (Li and Durbin, 2009), with default settings, to check if the reads could be unambiguously mapped. Further, BWA was used to align the trimmed and quality filtered reads excluding duplicates to the full mitochondrial genome sequence as well as the tRNA-Phe, the 12S ribosomal RNA, the tRNA-Val and the 16S ribosomal RNA of KU892588 to assess coverage. Obtained sequences were aligned to the sequences of specimens listed in **Supplementary Table 1**, covering nine of the eleven species of the *E. spinax* group (Straube et al., 2010; Ebert et al., 2021). Sequences used to determine the phylogenetic placement of *E. litvinovi* were obtained from the Chondrichthyan Tree of Life (2016) project[1] which are collected from vouchered and validated specimens, as described in White et al. (2018).

[1] https://sharkrays.org

A maximum likelihood tree was computed using RAxML v.8.2.4 (Stamatakis, 2014) under the general time reversible model. Heterogeneity of substitution rates among sites was modelled using a GAMMA distribution. To assess the statistical support for nodes, bootstrapping with 100 replicates was performed and plotted onto the maximum likelihood tree. A haplotype network was reconstructed with POPArt v. 1.7 (Leigh and Bryant, 2015) using the median joining network algorithm (Bandelt et al., 1999) under default settings. The RAxML tree served as a basis for calculating the p-distances between *E. litvinovi* and *E. spinax* clade species analysed herein using the Species Delimitation Plugin 1.4.5 (Masters et al., 2011) in Geneious®.

Etmopterus pycnolepis (Dense-Scale Lantern Shark)

This species is known from six specimens housed in three different museum collections, the ZIN (holotype: ZIN: 49226; two paratypes: ZIN: 49227); the ZMMU (paratype ZMMU: P-17992, paratype ZMMU P-17993) and the ZMH [paratype ZMH: 24995 (ex ISH 4-1989)]. We sampled tissue from the paratype specimen ZMH 24995 (**Figure 2B**) as described previously for the *E. litvinovi* paratype specimen. The specimen was captured at 25°56′ S and 88°33′ W at a depth of 580 m on 30.04.1987. It is an adult male of 426 mm total length (Thiel et al., 2009). As described for the *E. litvinovi* paratype, the overall condition and sampling date of the specimen suggests fixation with formaldehyde.

DNA extraction of the sample involved the same procedure as for *E. litvinovi*. Single stranded library preparation of *E. pycnolepis* DNA followed the protocol described in Gansauge et al. (2017). The *E. pycnolepis* sample underwent different laboratory procedures in comparison to the *E. litvinovi* sample, as the samples were processed with a considerable temporal gap, during which time the standard procedures in the historical laboratory at the University of Potsdam had been updated. Raw test-sequencing reads were analysed as in Straube et al. (2021a). FastQ Screen v0.14.0 (Wingett and Andrews, 2018) was used to check for unique hits to *Etmopterus* references and estimate contamination levels, before proceeding with target capture. After detection of target DNA in the test sequencing dataset, target capture was performed using an Arbor Bioscience myBaits® RNA bait kit. The baits were part of a multi-locus, multi species museum specimen barcoding approach described in Agne et al. (2022). NADH2 bait sequences were derived from representatives of all four *Etmopterus* clades (Straube et al., 2010) deposited in Genbank: *E. lucifer* (JQ518963), *E. gracilispinis* (JQ518960), *E. granulosus* (KF861686) and *E. bigelowi* (JQ518959). The four sequences were initially published in Naylor et al. (2012) and Straube et al. (2015). The single stranded DNA library was captured twice following the protocol described in Huang et al. (2021) using a hybridisation temperature of 65°C for 24 h. Sequencing of the double-captured, indexed library was performed on an Illumina NextSeq 500 System at the University of Potsdam as described in Paijmans et al. (2017). After quality filtering and adapter trimming using Cutadapt v. 2.10 (Martin, 2011) under default settings, reads were processed as described for *E. litvinovi* to reconstruct the NADH2 sequence of the paratype, using the NADH2 sequence of *E. lucifer* (JQ518963; Naylor et al., 2012) as reference.

The NADH2 consensus sequence (1044 bp) of the paratype was subsequently aligned with NADH2 sequences of other *Etmopterus* species with focus on the *E. lucifer* clade (**Supplementary Table 2**). Comparative sequences were obtained from the Chondrichthyan Tree of Life (2016) project (see text footnote 1). For details of NADH2 amplification and sequencing see Naylor et al. (2005, 2012). Forward and reverse sequences were aligned based on chromatograms and edited using Geneious® Pro v. 6.1.7 (Biomatters Ltd. Auckland, New Zealand). The consensus sequences were translated to amino acids and aligned with corresponding NADH2 sequences from representatives of closely related species using the MAFFT (Katoh et al., 2002, 2005) module in Geneious®. The aligned amino acid sequences were translated back in frame to their original nucleotide sequences, to yield a nucleotide alignment 1044 base pairs in length. Analysed samples are listed in **Supplementary Table 2** including 13 of the 15 *E. lucifer* clade species. Phylogenetic inference and species delimitation was performed as described for *E. litvinovi.*

RESULTS

Etmopterus litvinovi (Smalleye Lantern Shark)

A total of 3,734,481 trimmed and quality filtered reads were available after combining test-sequencing and target capture data, including duplicates. MitoBim ran for four iterations. 1,589,598 reads were used for baiting in the final iteration step. The consensus sequence shows 1.26% ambiguities scattered across the mitochondrial genome. Excluding duplicate sequences, 4985 reads map to the consensus sequence resulting from the Mitobim analysis providing an average coverage of 22 reads. The GC content is 40%. The mitochondrial tRNA and rRNA markers used for the phylogenetic analysis showed mapped read lengths mostly larger than 70 base pairs (**Supplementary Figure 1A**) and an average coverage of 58 reads, excluding duplicates (**Supplementary Figure 2A**). They did not show any ambiguous nucleotides.

The maximum likelihood phylogeny of the tRNA and rRNA sequences identifies lineages corresponding to species within the *E. spinax* clade. The relationships in the tree are mostly well-supported with many bootstrap values reaching 100% (**Figure 3A**). The *E. litvinovi* paratype sequence is sister to a sample identified as *E. benchleyi*. This clade also includes a specimen of *E. benchleyi* sampled in the Indian Ocean (GN4952). The clade as a whole is sister to the North Atlantic species *E. princeps* and *E. spinax,* which together form the sister clade to the Southern Hemisphere species *E. viator* (**Figure 3A**). The reconstructed haplotype network detected five haplotypes with 34 segregating sites and 16 parsimony-informative characters. **Figure 3B** shows that the haplotype sequence of the *E. litvinovi*

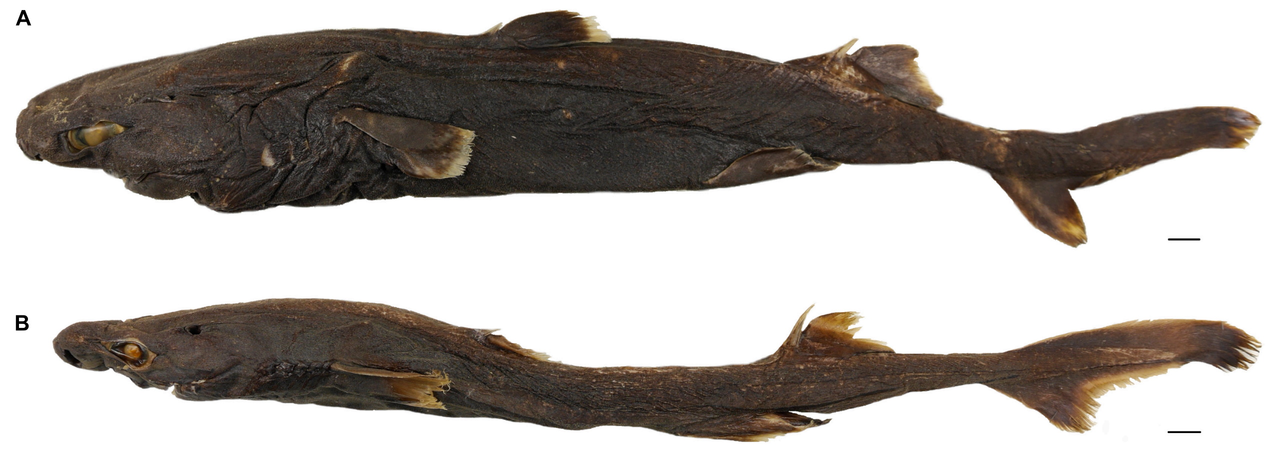


FIGURE 2 | Paratype images of **(A)** *Etmopterus litvinovi* (ZMH 24994) and **(B)** *Etmopterus pycnolepis* (ZMH 24995). Bars indicate 1 cm.

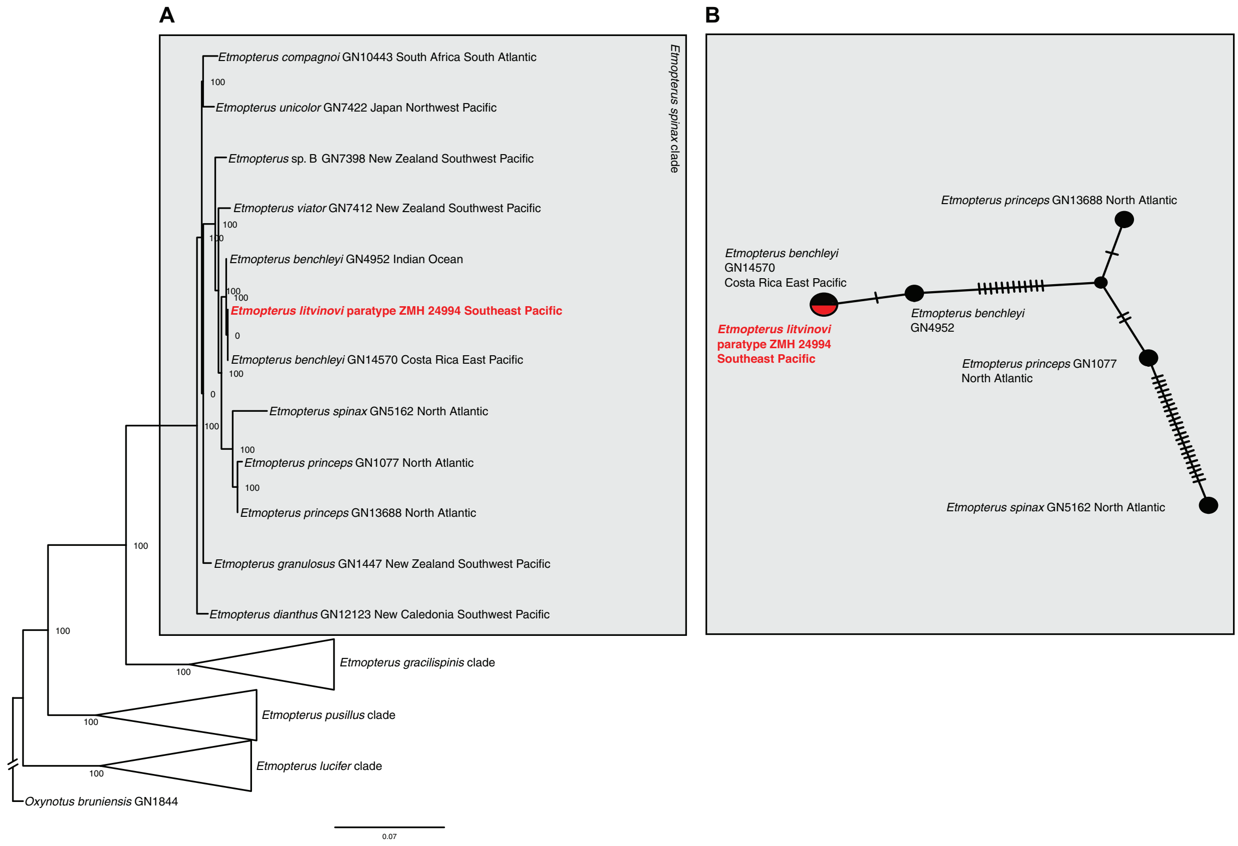

A
Etmopterus spinax clade
Etmopterus compagnoi GN10443 South Africa South Atlantic
100
Etmopterus unicolor GN7422 Japan Northwest Pacific
Etmopterus sp. B GN7398 New Zealand Southwest Pacific
Etmopterus viator GN7412 New Zealand Southwest Pacific
100
100
Etmopterus benchleyi GN4952 Indian Ocean
100
100
Etmopterus litvinovi paratype ZMH 24994 Southeast Pacific
0
Etmopterus benchleyi GN14570 Costa Rica East Pacific
100
0
Etmopterus spinax GN5162 North Atlantic
100
100
Etmopterus princeps GN1077 North Atlantic
100
Etmopterus princeps GN13688 North Atlantic
100
Etmopterus granulosus GN1447 New Zealand Southwest Pacific
Etmopterus dianthus GN12123 New Caledonia Southwest Pacific
100
Etmopterus gracilispinis clade
100
Etmopterus pusillus clade
100
Etmopterus lucifer clade
100
Oxynotus bruniensis GN1844
0.07
B
Etmopterus princeps GN13688 North Atlantic
Etmopterus benchleyi
GN14570
Costa Rica East Pacific
Etmopterus litvinovi
paratype ZMH 24994
Southeast Pacific
Etmopterus benchleyi
GN4952
Etmopterus princeps GN1077
North Atlantic
Etmopterus spinax GN5162 North Atlantic

FIGURE 3 | (A) Maximum likelihood analysis of *Etmopterus* 12s and 16s rRNA sequences. Numbers at nodes denote bootstrap support values. *Oxynotus bruniensis* was chosen as outgroup. **(B)** Haplotype network of 12s and 16s rRNA sequences including sequences from the *E. litvinovi* paratype, *E. benchleyi* and its sister clade comprising *E. princeps* and *E. spinax*. Crossbars indicate mutational steps. *E. litvinovi* and *E. benchleyi* GN14570 share a haplotype, GN4952 differs in a single mutational step.

paratype is identical to the *E. benchleyi* sample GN14570 and separated by a single mutational step from *E. benchleyi* sample GN4952. The species delimitation analysis shows that the interspecific K2P distance between two valid sister species within the *E. spinax* clade is on average 1.6% (**Supplementary Table 3A**), while the K2P distance between *E. litvinovi* and *E. benchleyi* is substantially smaller (K2P distance = 0.0518%; **Supplementary Table 3A**). Overall, our data does not support the validity of

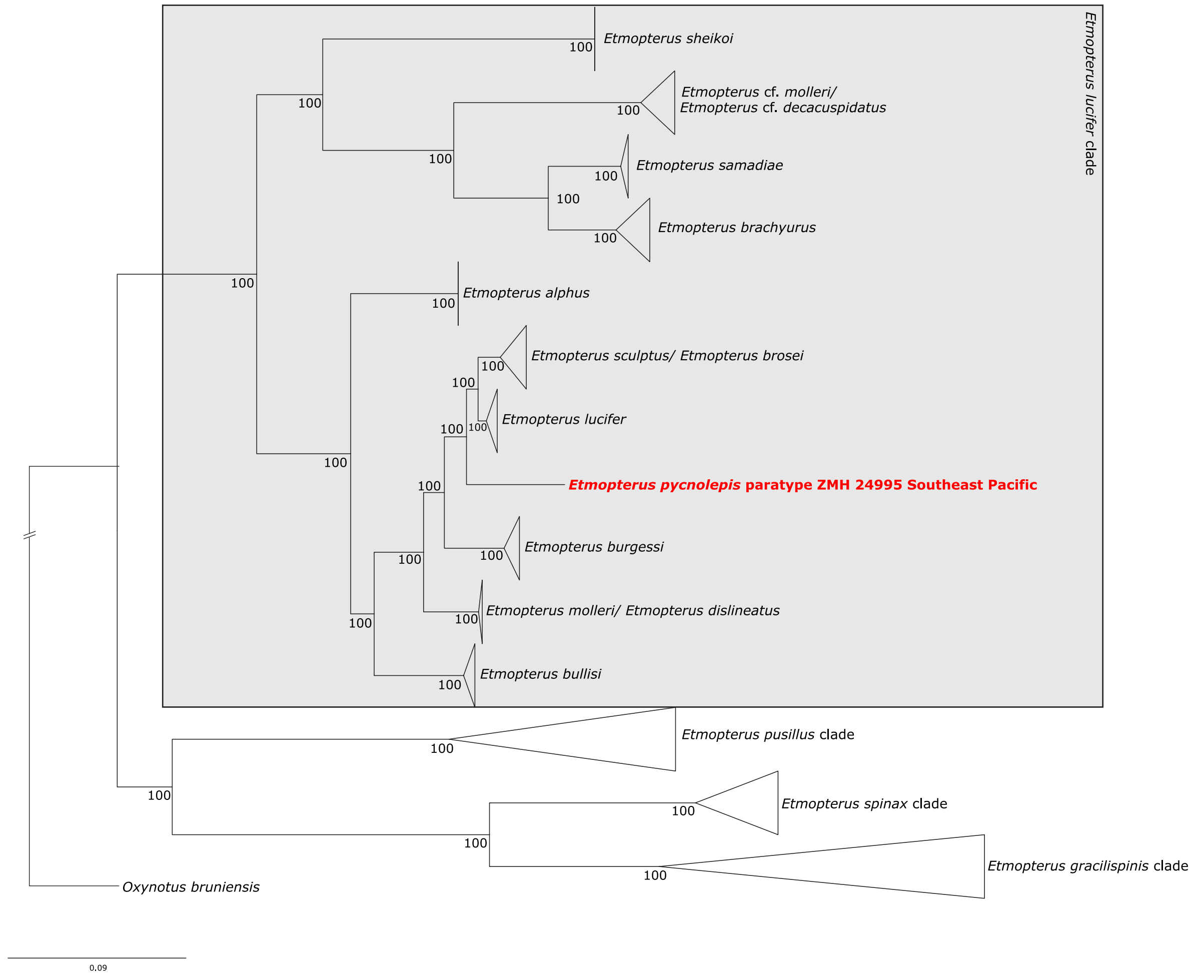


FIGURE 4 | Maximum likelihood analysis of *Etmopterus* NADH2 sequences. Numbers at branches denote bootstrap support values. *Oxynotus bruniensis* was chosen as outgroup.

both species due to the phylogenetic placement of the *E. litvinovi* paratype sequence in the *E. benchleyi* clade and a very small K2P distance between both species.

Etmopterus pycnolepis (Dense-Scale Lantern Shark)

The test-sequencing dataset of 398,752 trimmed reads detected the presence of *Etmopterus* DNA, as indicated by 5.62% unique hits to the *E. spinax* transcriptome used as reference in the FastQscreen analysis. 82.83% of reads were un-assigned to any of the provided references, and contamination with other samples processed simultaneously was not detected. 4,029,200 raw reads were produced by sequencing of the target captured library. Quality filtering and trimming reduced this to 2,680,159 reads. Of these, 139,219 reads mapped to the NADH2 gene of *E. lucifer* (JQ518963). The complete NADH2 sequence was reconstructed after three iterations in Mitobim, using 115,548 reads in the final iteration. The mapped read length distribution is shown in **Supplementary Figure 1B**. The modal fragment length is around 50 base pairs. Mapping those reads back to the reconstructed NADH2 consensus sequence, showed that 277 reads mapped with an average coverage of 13 reads, excluding duplicate sequences (**Supplementary Figure 2B**). The maximum likelihood NADH2 phylogeny shows well-supported clades within the *E. lucifer* group. Several clades do not correspond to species: *E. brosei* clusters with *E. sculptus, E.* cf. *molleri* clusters with *E.* cf. *decacuspidatus* and forms a distinct clade not including *E. molleri. Etmopterus molleri* and *E. dislineatus* do not form two distinct clades. The *E. pycnolepis* paratype specimen is sister to a clade containing the southern hemisphere samples of *E. lucifer* and *E. sculptus* (**Figure 4**). The species delimitation analysis shows that the interspecific K2P distance of the NADH2 gene between sequences of two valid species in the *E. lucifer* clade, excluding clades not corresponding to species, is on average 4% (**Supplementary Table 3B**). Comparing the K2P distance of *E. pycnolepis* to its closest sister taxon, *E. lucifer*, the K2P distance is 3.4%. Our data therefore supports the species status of *E. pycnolepis*.

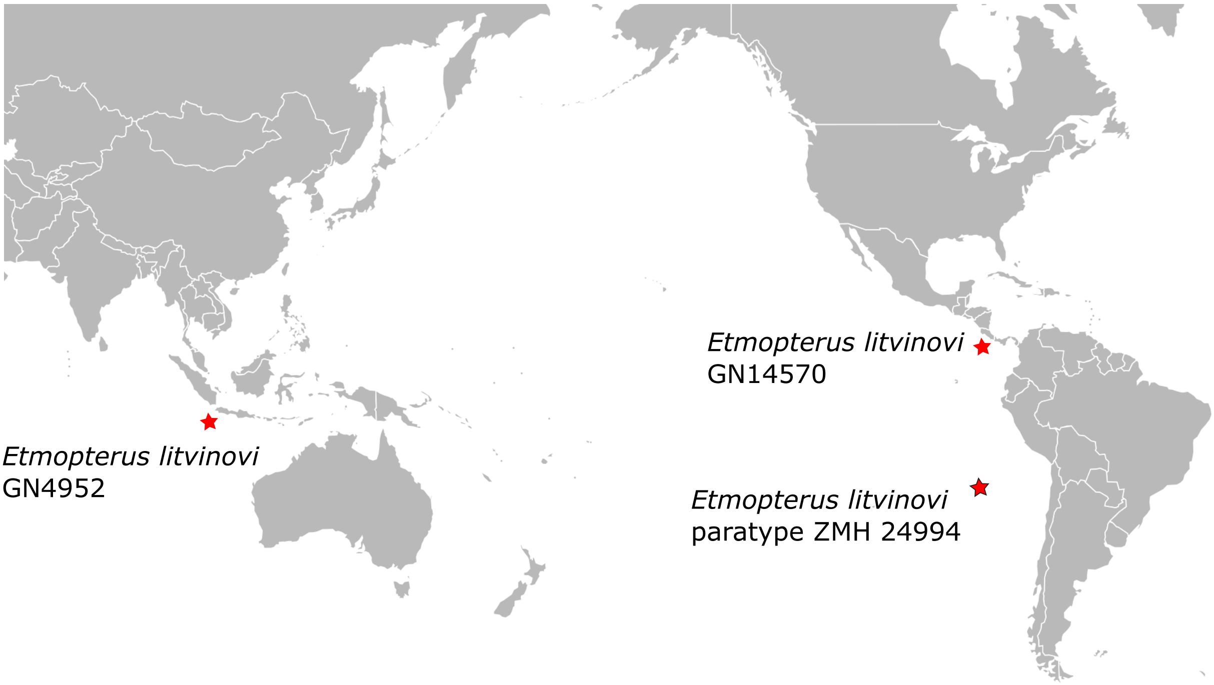


FIGURE 5 | Paratype locality of *Etmopterus litvinovi* ZMH 24994 (red star with black frame) and sampling locations of the two other *E. litvinovi* specimens (red stars) initially labeled as *E. benchleyi* (**Supplementary Table 1**).

DISCUSSION

Mitochondrial DNA Characteristics of the Two Paratype Specimens

The mitochondrial DNA we obtained from both paratype specimens, while fragmented (**Supplementary Figure 1**), was less degraded than other museum samples analysed in previous studies (Straube et al., 2021b). Fixation and preservation cause DNA damage (e.g., Stiller et al., 2016; Hahn et al., 2021; Straube et al., 2021a); however, the mitochondrial DNA of the two paratype samples analysed herein may be less affected due to the comparatively young age of 32 and 34 years, respectively, at the time of extraction. More comparative data is necessary to test if time is correlated with mitochondrial DNA fragmentation levels, and if fragmentation is ongoing under the current storage conditions.

Taxonomic Implications

Etmopterus litvinovi (Smalleye Lantern Shark)

The phylogenetic placement of the paratype sequence aligns with the morphology-based prediction that *E. litvinovi* is a member of the *E. spinax* species clade (Straube et al., 2010). A close relationship of *E. litvinovi* with morphological congeners, including cryptic species, was suggested in Straube et al. (2011a,b). This species complex was recently expanded with several species from which mitochondrial DNA sequence information is, however, available from only two species, *E. benchleyi* (Vásquez et al., 2015) and *E. viator* (Straube et al., 2011a). All analyses, i.e., the phylogenetic reconstruction, the haplotype network and the species delimitation analysis of the *E. litvinovi* paratype specimen suggest that *E. litvinovi* is conspecific with *E. benchleyi*, where *E. benchleyi* forms a junior synonym to *E. litvinovi* (**Figure 3** and **Supplementary Table 3A**). Notably, the sequenced paratype specimen's sampling locality is the Naszca Ridge in the Pacific Ocean, while another *E. litvinovi* haplotype (GN4952) is derived from a specimen sampled in the Indian Ocean (**Figure 5**). This suggests that *E. litvinovi* is widespread and occurs both in the Indian and Pacific oceans. Its overall distribution range may cover an even larger area of Southern Hemisphere oceans, and that its northern and southern distribution limits have yet to be identified. Similarly, wide distribution ranges are also documented for other closely related *Etmopterus* species in the *E. spinax* clade such as *E. granulosus* (Straube et al., 2011a,b, 2015) or *E. viator* (Straube et al., 2011a).

The new data presented herein will be helpful for future assessments of the species in the IUCN Red List of Threatened Species. As of today, the species is evaluated and listed under the "least concern" category justified due to limited fisheries in the area from which the species (i.e., the type material of *E. litvinovi*) was hitherto recorded (Ebert et al., 2020a). Based on our results, a notably larger distribution range should be considered in future evaluations taking different fishing pressure in other regions of occurrence into account. Our results do not confirm endemic occurrence (Kotlyar, 1990) but support its occurrence in the eastern Pacific as well as the Indian Ocean. The paratype sequence data analysed herein adds important alpha-level taxonomic information on the species, which will ease the collection of data on population size, as well as accurate geographic and depth distribution ranges, in the future. This forms the basis for conservation and management efforts for this poorly known deep-sea shark species.

Etmopterus pycnolepis (Dense-Scale Lantern Shark)

As already indicated by the distinct shape of its flank marking, our analysis further supports the assignment of *E. pycnolepis* as a distinct species (**Supplementary Table 3B**) within the *E. lucifer* clade (**Figure 4**; Straube et al., 2010). Its assignment to the *E. burgessi* subclade (Ebert et al., 2021) is not supported, however. The morphologically defined *E. lucifer* clade subclades described in Ebert et al. (2021) are generally not recovered in our molecular analysis (**Figure 4**). The phylogenetic inference displays to some extent geographic patterns of sampling locations instead. *E. lucifer*, *E. pycnolepis*, *E. brosei* and *E. sculptus* are represented by samples exclusively collected in Southern Hemisphere oceans (**Supplementary Table 2**), while *E. burgessi* samples stem from the Northwest Pacific. *E. brachyurus*, *E.* cf. *molleri*, and *E. samadiae* samples were also collected in the Northwest Pacific. *E. alphus* samples are from the Indian Ocean off Mauritius and *E. bullisi* was sampled in the Northwest Atlantic (**Supplementary Table 2**). Some species seem therefore confined to certain oceanic areas. The three different flank mark shapes characterising the *E. lucifer* clade subclades occur in three different ocean regions in parallel. In our study, *E. brachyurus E. samadiae E.* cf. *molleri*, *E.* cf. *decacuspidatus*, and *E. burgessi* represent the flank mark diversity of all three subclades in the Northwest Pacific; *E. lucifer*, *E. molleri*, *E. dislineatus*, *E. brosei*, *E. alphus*, *E. pycnolepis* and *E. sculptus* in the Indian and South Pacific oceans. In the Atlantic Ocean, only the *E. molleri* subclade type flank marking (Ebert et al., 2021) is represented by a single species (*E. bullisi*); however, *E. bullisi* is the only Atlantic species from the *E. lucifer* clade in general. A denser sampling is necessary to identify detailed species distribution boundaries and clarify indicated synonymies.

The IUCN Red List of Threatened species lists *E. pycnolepis* as least concern under the assumption that the area of origin of the six type specimens representing the species is not exposed to extensive fishing pressure as also in *E. litvinovi*. As mentioned in the evaluation justification, the species may be distributed in Chilean waters as well (Ebert et al., 2020b), which would amount to a large expansion of its distribution area. By providing the first DNA sequences for this species, newly collected samples available for NADH2 sequencing can be correctly assigned to the species and will therefore be useful for documenting its distribution range in the future.

CONCLUSION

Our results demonstrate the importance of archival DNA sequence information from type material for molecular based taxonomy. This is especially true for species which are known from few specimens only, and where re-sampling is hindered by remote sampling localities, as is the case for the two species analysed herein. Our results support the synonymy of *E. benchleyi* with *E. litvinovi,* and consequently suggest a notably larger distribution range than previously known, since the species was assumed to be endemic to the Salas y Gómez and Nazca Submarine Ridges (Kotlyar, 1990). The species status of *E. pycnolepis* is supported by our data, which is now available as reference for future molecular species-level identification of newly collected samples. This will help clarify the distribution of this species. Our results further show that genetic information from collection material can assist in the evaluation of species in a conservation and management context. While it is standard to evaluate morphological characters of wet-collection type material for descriptions of species new to science, the usage of wet-collection specimen DNA sequence information has only recently been established as such (Beermann et al., 2018; Lyra et al., 2020; Rancilhac et al., 2020; Scherz et al., 2020; Straube et al., 2021b) and our work is a further contribution to this.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: Figshare (doi: 10.6084/m9.figshare.19446992) and GenBank (accession numbers ON185623-ON185724).

ETHICS STATEMENT

Ethical review and approval was not required for the animal study because no living animals were collected or examined. Tissue samples for DNA sequencing were taken solely from museum specimens and combined with existing data for analysis.

AUTHOR CONTRIBUTIONS

NS, MH, JP, and AB designed the study. SA, MP, and NS conducted laboratory work. GN and LY provided comparative sequences for phylogenetics. SA, NS, LY, and GN analysed the data. RT and SW researched type specimen history, provided and helped sampling the type specimens. NS, SA, and MH wrote the manuscript with contributions from all authors. All authors contributed to the article and approved the submitted version.

FUNDING

This work was funded by the German Research Foundation (DFG; project number 351649567 to NS and MH within the DFG SPP 1991 "Taxon-Omics"). This project was further supported by the National Science Foundation (NSF), grant DEB-01132229 to GN.

ACKNOWLEDGMENTS

We would like to express our sincere thanks to Irina Eidus [Zoological Museum (ZMH) of the LIB] for her assistance during sampling of the paratype specimens. Jürgen Pollerspöck at shark references (www.shark-references.com) is thanked for literature search. We are also grateful to Mikhail Nazarkin (ZIN) and Matthias Stehmann (ICHTHYS) for information regarding handling of the specimens. Mikhail Nazarkin also kindly provided images of the holotypes of both species. Nataliya Budaeva (UM) is thanked for her help extracting collection information from the original description. We would also like to thank the HPC team at the University of Potsdam for creating and maintaining the university's server on which we ran some of our analyses.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fevo.2022.910009/full#supplementary-material

REFERENCES

Agne, S., Straube, N., Preick, M., and Hofreiter, M. (2022). Simultaneous barcode sequencing of diverse museum collection specimens using a mixed RNA bait set. *Front. Ecol. Evol.* doi: 10.3389/fevo.2022.909846

Bandelt, H., Forster, P., and Röhl, A. (1999). Median-joining networks for inferring intraspecific phylogenies. *Mol. Biol. Evol.* 16, 37–48. doi: 10.1093/oxfordjournals.molbev.a026036

Beermann, J., Westbury, M. V., Hofreiter, M., Hilgers, L., Deister, F., Neumann, H., et al. (2018). Cryptic species in a well-known habitat: applying taxonomics to the amphipod genus *Epimeria* (Crustacea, Peracarida). *Sci. Rep.* 8:6893. doi: 10.1038/s41598-018-25225-x

Chen, H., Chen, X., Gu, X., Wan, H., Chen, X., and Ai, W. (2016). The phylogenomic position of the Smooth lanternshark *Etmopterus pusillus* (Squaliformes: Etmopteridae) inferred from the mitochondrial genome. *Mitochondrial DNA B Resour.* 1, 341–342. doi: 10.1080/23802359.2016.1172274

Chondrichthyan Tree of Life (2016). *Chondrichthyan Tree of Life*. Available online at: https://sharksrays.org/(accessed March 07, 2022).

Dabney, J., Knapp, M., Glocke, I., Gansauge, M. T., Weihmann, A., Nickel, B., et al. (2013). Complete mitochondrial genome sequence of a Middle Pleistocene cave bear reconstructed from ultrashort DNA fragments. *Proc. Natl. Acad. Sci. U.S.A.* 110, 15758–15763. doi: 10.1073/pnas.1314445110

Dolganov, V. N., and Balanov, A. A. (2018). *Etmopterus parini* sp. n. (Squaliformes: Etmopteridae), a new shark species from the northwestern Pacific Ocean. *Biol. Morya* 44, 427–430.

Ebert, D. A., Concha, F., Herman, K., and Kyne, P. M. (2020a). *Etmopterus litvinovi. The IUCN Red List of Threatened Species 2020*. Gland: IUCN. doi: 10.2305/IUCN.UK.2020-3.RLTS.T63159A124463835.en

Ebert, D. A., Kyne, P. M., Concha, F., and Herman, K. (2020b). *Etmopterus pycnolepis. The IUCN Red List of Threatened Species 2020*. Gland: IUCN. doi: 10.2305/IUCN.UK.2020-3.RLTS.T63160A124463919.en

Ebert, D. A., Fowler, S. L., and Compagno, L. J. (2013). *Sharks of the World: a Fully Illustrated Guide*. Princeton, NJ: Wild Nature Press.

Ebert, D. A., Leslie, R. W., and Weigmann, S. (2021). *Etmopterus brosei* sp. nov.: a new lanternshark (Squaliformes: Etmopteridae) from the southeastern Atlantic and southwestern Indian oceans, with a revised key to the Etmopterus lucifer clade. *Mar. Biodivers.* 51, 1–17. doi: 10.1007/s12526-021-01173-0

Ebert, D. A., Straube, N., Leslie, R. W., and Weigmann, S. (2016). *Etmopterus alphus* n. sp.: a new lanternshark (Squaliformes: Etmopteridae) from the south-western Indian Ocean. *Afr. J. Mar. Sci.* 38, 329–340. doi: 10.2989/1814232X.2016.1198275

Gansauge, M. T., Gerber, T., Glocke, I., Korleviæ, P., Lippik, L., Nagel, S., et al. (2017). Single-stranded DNA library preparation from highly degraded DNA using T4 DNA ligase. *Nucleic Acids Res.* 45:e79. doi: 10.1093/nar/gkx033

Gansauge, M. T., and Meyer, M. (2013). Single-stranded DNA library preparation for the sequencing of ancient or damaged DNA. *Nat. Protocols* 8, 737–748. doi: 10.1038/nprot.2013.038

Gilbert, M. T. P., Haselkorn, T., Bunce, M., Sanchez, J. J., Lucas, S. B., Jewell, L. D., et al. (2007). The Isolation of Nucleic Acids from Fixed, Paraffin-Embedded Tissues-Which Methods Are Useful When? *PLoS One* 2:e537. doi: 10.1371/journal.pone.0000537

González Fortes, G., and Paijmans, J. L. (2019). *Ancient DNA*. Totowa, NJ: Human Press.

Hahn, C., Bachmann, L., and Chevreux, B. (2013). Reconstructing mitochondrial genomes directly from genomic next-generation sequencing reads - A baiting and iterative mapping approach. *Nucleic Acids Res.* 41:e129. doi: 10.1093/nar/gkt371

Hahn, E. E., Alexander, M. R., Grealy, A., Stiller, J., Gardiner, D. M., and Holleley, C. E. (2021). Unlocking inaccessible historical genomes preserved in formalin. *Mol. Ecol. Resour.* 1–18. doi: 10.1111/1755-0998.13505

Hebert, P. D., Cywinska, A., Ball, S. L., and DeWaard, J. R. (2003). Biological identifications through DNA barcodes. *Proc. R. Soc. Lond. Ser. B Biol. Sci.* 270, 313–321. doi: 10.1098/rspb.2002.2218

Hoffman, E. A., Frey, B. L., Smith, L. M., and Auble, D. T. (2015). Formaldehyde crosslinking: a tool for the study of chromatin complexes. *J. Biol. Chem.* 290, 26404–26411. doi: 10.1074/jbc.R115.651679

Huang, J. M., Yuan, H., and Li, C. H. (2021). Protocol for Cross-species Target-gene Enrichment. *Bio-protocol* 101:e1010606. doi: 10.21769/BioProtoc.1010606

Hykin, S. M., Bi, K., and McGuire, J. A. (2015). Fixing formalin: A method to recover genomic-scale DNA sequence data from formalin-fixed museum specimens using high-throughput sequencing. *PLoS One* 10:e0141579. doi: 10.1371/journal.pone.0141579

Katoh, K., Kuma, K. I., Toh, H., and Miyata, T. (2005). MAFFT version 5: Improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res.* 33, 511–518. doi: 10.1093/nar/gki198

Katoh, K., Misawa, K., Kuma, K. I., and Miyata, T. (2002). MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier Transform. *Nucleic Acids Res.* 30, 3059–3066. doi: 10.1093/nar/gkf436

Kotlyar, A. N. (1990). Dogfish sharks of the genus *Etmopterus* Rafinesque from the Nazca and Sala y Gómez submarine ridges. *Tr. Inst. Okeanol. AN USSR* 125, 127–147.

Leigh, J. W., and Bryant, D. (2015). PopART: Full-feature software for haplotype network construction. *Methods Ecol. Evol.* 6, 1110–1116. doi: 10.1111/2041-210X.12410

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25, 1754–1760. doi: 10.1093/bioinformatics/btp324

Lyra, M. L., Lourenço, A. C. C., Pinheiro, P. D., Pezzuti, T. L., Baêta, D., Barlow, A., et al. (2020). High-throughput DNA sequencing of museum specimens sheds light on the long-missing species of the *Bokermannohyla claresignata* group (Anura: Hylidae: Cophomantini). *Zool. J. Linn. Soc.* 190, 1235–1255. doi: 10.1093/zoolinnean/zlaa033

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *Tech. Notes* 7, 2803–2809. doi: 10.14806/ej.17.1.200

Masters, B. C., Fan, V., and Ross, H. A. (2011). Species delimitation-a geneious plugin for the exploration of species boundaries. *Mol. Ecol. Resour.* 11, 154–157. doi: 10.1111/j.1755-0998.2010.02896.x

McGuire, J. A., Cotoras, D. D., O'Connell, B., Lawalata, S. Z. S., Wang-Claypool, C. Y., Stubbs, A., et al. (2018). Squeezing water from a stone: High- throughput sequencing from a 145-year old holotype resolves (barely) a cryptic species problem in flying lizards. *PeerJ* 6:e4470. doi: 10.7717/peerj.4470

Naylor, G. J., Ryburn, J. A., Fedrigo, O., and Lopez, J. A. (2005). Phylogenetic relationships among the major lineages of modern elasmobranchs. *Reprod. Biol. Phylogeny* 3:25.

Naylor, G. J. P., Caira, J. N., Jensen, K., Rosana, K. A. M., White, W. T., and Last, P. R. (2012). A DNA sequence-based approach to the identification of shark and ray species and its implications for global elasmobranch diversity and parasitology. *Bull. Am. Mus. Nat. Hist.* 367, 1–262.

Paijmans, J. L. A., Baleka, S., Henneberger, K., Taron, U., Trinks, A., Westbury, M., et al. (2017). Sequencing single-stranded libraries on the Illumina NextSeq 500 platform. *arXiv* [Preprint]. doi: 10.48550/arXiv.1711.11004

Pollerspöck, J., and Straube, N. (2021). *Bibliography Database of Living/Fossil Sharks, Rays and Chimaeras (Chondrichtyes: Elasmobranchii, Holocephali) – List of Valid Extant Species; List of Described Extant Species; Statistic, World Wide Web Electronic Publication, Version 03/2021*. Available online at: www.shark-references.com

Rancilhac, L., Bruy, T., Scherz, M. D., Pereira, E. A., Preick, M., Straube, N., et al. (2020). Target-enriched DNA sequencing from historical type material enables a partial revision of the Madagascar giant stream frogs (genus *Mantidactylus*). *J. Nat. Hist.* 54, 87–118. doi: 10.1080/00222933.2020.1748243

Rohland, N., Siedel, H., and Hofreiter, M. (2004). Nondestructive DNAextraction method for mitochondrial DNA analyses of museum specimens. *BioTechniques* 36, 814–821. doi: 10.2144/04365ST05

Scherz, M. D., Rasolonjatovo, S. M., Köhler, J., Rancilhac, L., Rakotoarison, A., Raselimanana, A. P., et al. (2020). 'Barcode fishing'for archival DNA from historical type material overcomes taxonomic hurdles, enabling the description of a new frog species. *Sci. Rep.* 10:19109. doi: 10.1038/s41598-020-75431-9

Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30, 1312–1313. doi: 10.1093/bioinformatics/btu033

Stiller, M., Sucker, A., Griewank, K., Aust, D., Baretton, G. B., Schadendorf, D., et al. (2016). Single-strand DNA library preparation improves sequencing of formalin-fixed and paraffin-embedded (FFPE) cancer DNA. *Oncotarget* 7, 59115–59128. doi: 10.18632/oncotarget.10827

Straube, N., Duhamel, G., Gasco, N., Kriwet, J., and Schliewen, U. K. (2011a). "Description of a new deep-sea lantern shark *Etmopterus viator* sp. nov.(Squaliformes: Etmopteridae) from the Southern Hemisphere," in *The Kerguelen Plateau: Marine Ecosystem and Fisheries*, eds G. Duhamel & D.C. Welsford (Paris: Société Française d'Ichtyologie). doi: 10.13140/2.1.1107.4248

Straube, N., Kriwet, J., and Schliewen, U. K. (2011b). Cryptic diversity and species assignment of large lantern sharks of the *Etmopterus spinax* clade from the Southern Hemisphere (Squaliformes, Etmopteridae). *Zool. Scr.* 40, 61–75. doi: 10.1111/j.1463-6409.2010.00455.x

Straube, N., Iglésias, S. P., Sellos, D. Y., Kriwet, J., and Schliewen, U. K. (2010). Molecular phylogeny and node time estimation of bioluminescent Lantern Sharks (Elasmobranchii: Etmopteridae). *Mol. Phylogenet. Evol.* 56, 905–917. doi: 10.1016/j.ympev.2010.04.042

Straube, N., Leslie, R. W., Clerkin, P. J., Ebert, D. A., Rochel, E., Corrigan, S., et al. (2015). On the occurrence of the Southern Lanternshark, *Etmopterus granulosus*, off South Africa, with comments on the validity of *E. compagnoi*. *Deep Sea Res. II Top. Stud. Oceanogr.* 115, 11–17. doi: 10.1016/j.dsr2.2014.04.004

Straube, N., Lyra, M. L., Paijmans, J. L. A., Preick, M., Basler, N., Penner, J., et al. (2021a). Successful application of ancient DNA extraction and library construction protocols to museum wet collection specimens. *Mol. Ecol. Resour.* 21, 2299–2315. doi: 10.1111/1755-0998.13433

Straube, N., Preick, M., Naylor, G. J. P., and Hofreiter, M. (2021b). Mitochondrial DNA sequencing of a wet-collection syntype demonstrates the importance of type material as genetic resource for lantern shark taxonomy (Chondrichthyes: Etmopteridae). *R. Soc. Open Sci.* 8:210474. doi: 10.1098/rsos.210474

Thiel, R., Eidus, I., and Neumann, R. (2009). The Zoological Museum Hamburg (ZMH) fish collection as a global biodiversity archive for elasmobranchs and actinopterygians as well as other fish taxa. *J. Appl. Ichthyol.* 25(Suppl. 1), 9–32. doi: 10.1111/j.1439-0426.2009.01296.x

Vásquez, V. E., Ebert, D. A., and Long, D. J. (2015). *Etmopterus benchleyi* n. sp., a new lanternshark (Squaliformes: Etmopteridae) from the central eastern Pacific Ocean. *J. Ocean Sci. Found.* 17, 43–55.

White, W. T., Corrigan, S., Yang, L. E. I., Henderson, A. C., Bazinet, A. L., Swofford, D. L., et al. (2018). Phylogeny of the manta and devilrays (Chondrichthyes: Mobulidae), with an updated taxonomic arrangement for the family. *Zool. J. Linn. Soc.* 182, 50–75. doi: 10.1093/zoolinnean/zlx018

White, W. T., Ebert, D. A., Mana, R. R., and Corrigan, S. (2017). *Etmopterus samadiae* n. sp., a new lanternshark (Squaliformes: Etmopteridae) from Papua New Guinea. *Zootaxa* 4244, 339–354. doi: 10.11646/zootaxa.4244.3.3

Wingett, S. W., and Andrews, S. (2018). FastQ Screen: A tool for multi-genome mapping and quality control. *F1000Res.* 7:1338. doi: 10.12688/f1000research.15931.2

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

BRIEF RESEARCH REPORT
published: 14 July 2022
doi: 10.3389/fevo.2022.921106

DNA Barcoding Technology Used to Successfully Sub-Classify a Museum Whale Specimen as *Balaenoptera edeni edeni*

Xiaoying Ren[1†], Xiaolin Ma[2†], Edward Allen[1], Yuan Fang[3] and Shaoqing Wen[1,4]**

[1] Institute of Archaeological Science, Fudan University, Shanghai, China, [2] State Key Laboratory of Estuarine and Coastal Research, East China Normal University, Shanghai, China, [3] China (Hainan) Museum of the South China Sea, Qionghai, China, [4] Ministry of Education Key Laboratory of Contemporary Anthropology, Department of Anthropology and Human Genetics, School of Life Sciences, Fudan University, Shanghai, China

DNA barcoding technology is becoming an increasingly powerful tool in resolving issues of detailed species identification based on morphology, as commonly employed by museums. In the present study, we aimed to identify a stranded Bryde's whale on Hainan Island, China by extracting DNA from a vertebra pre-treated by physical and/or chemical processes. Based on morphological characteristics, this Bryde's whale was initially determined as *Balaenoptera edeni*. Then, DNA was efficiently extracted using ancient DNA techniques. The mitochondrial gene (COI) phylogenetic analysis further revealed that this museum whale specimen belonged to the sub-species *B. e. edeni*. This study provides a testable and rapid method for museum species verification, by using ancient DNA extraction methods to compensate the disadvantage of traditional DNA extraction methods that are difficult to extract valid DNA.

Keywords: museum specimen, Bryde's whale, morphological identification, DNA barcoding, DNA identification

OPEN ACCESS

Edited by:
Jonathan J. Fong,
Lingnan University, China

Reviewed by:
Utpal Smart,
University of North Texas Health Science Center, United States
Santiago Castroviejo-Fisher,
Seville University, Spain

****Correspondence:***
Yuan Fang
751938749@qq.com
Shaoqing Wen
wenshaoqing@fudan.edu.cn

†These authors have contributed equally to this work

Specialty section:
This article was submitted to Phylogenetics, Phylogenomics, and Systematics, a section of the journal Frontiers in Ecology and Evolution

Received: *15 April 2022*
Accepted: *13 June 2022*
Published: *14 July 2022*

Citation:
Ren X, Ma X, Allen E, Fang Y and Wen S (2022) DNA Barcoding Technology Used to Successfully Sub-Classify a Museum Whale Specimen as Balaenoptera edeni edeni. Front. Ecol. Evol. 10:921106. doi: 10.3389/fevo.2022.921106

INTRODUCTION

Identifying the species/sub-species of museum specimens has long been a major challenge. Traditional approaches to identification have been based on morphometric analysis and/or morphological criteria, often without the services of taxonomic specialists (Bacher, 2012). Genetic materials have recently emerged as a promising trend in the rapid resolution of species/sub-species identification for both fresh and ancient museum specimens (Barbanera et al., 2020; Pierson et al., 2020). These "non-invasive" approaches cost museums little to nothing with regard to the quantity and quality of specimens held. This development has been offset by the high degree of decomposition among much museum material, whose prior physical or chemical treatment can severely impede the process of effective DNA extraction. As such, increasingly effective means have been developed for the extraction of highly fragmented DNA in the presence of contaminants and inhibitors (Rohland et al., 2018). One specific new method, DNA barcoding, takes advantage of short standardized sequences in order to facilitate species identification (Hebert et al., 2003; Savolainen et al., 2005). In DNA barcoding, both intraspecific variation and interspecific divergence can be significant, with the mitochondrial cytochrome c oxidase 1 (COI) identified as the best gene based on its conserved amino acid sequence, and hence the key to distinguishing animal species and sub-species (Knowlton and Weigt, 1998; Hebert et al., 2003; Chapuis et al., 2016). The DNA barcoding approach has been used to identify a variety of museum species/sub-species ranging from insects to fishes to primates (Thomsen et al., 2009; Hawlitschek et al., 2017). Nevertheless, DNA barcoding of marine mammals—the subject of this study—remains in its relative infancy.

Stranded cetacean specimens are objects of public fascination when displayed in the exhibition halls of museums. Bryde's whale, or Bryde's whale complex, a baleen whale occupying warm-temperate waters on a year-round basis, can be recognized by the three distinct ridges on its rostrum (Penry et al., 2018). Bryde's whale is currently recognized as a single species (*Balaenoptera edeni* Anderson, 1879) with two recognized subspecies: a small coastal form (Eden's whale, *B. e. edeni*) and a large oceanic form (Bryde's whale, *B. e. brydei*) (Constantine et al., 2018; Penry et al., 2018; Liu et al., 2021; Committee on Taxonomy, 2022). Recently, sightings of Bryde's whale have been recorded from the coasts of East and Southeast Asia (Yamada et al., 2008; Chen et al., 2019; Liu et al., 2021), south west Indian Ocean (Penry et al., 2018), Southern Africa (Best, 2001; Penry et al., 2011) and Gulf of Mexico (Rosel and Wilcox, 2014). However, as the type specimen for *B. e. brydei* was not designated with the naming of the species, and genetic analysis of the type specimen of *B. e. edeni* was not completed, detailed taxonomy within the Bryde's whale group is unclear (Anderson, 1879; Constantine et al., 2015).

Tracking back at least four decades (1978–2016), about nine Bryde's-like whales were stranded along the coast of Hainan Province, China, whereas little information was available on age, gender and taxonomy (Zhang et al., 2015; Liu et al., 2019). In 2019, an adult Bryde's-like whale was discovered along the coast of Qiaotou Town, Chengmai County, Hainan Province, China. The specimen's corpse had already been buried by nearby villagers by the time the museum team arrived on site. Based on the extent of decomposition, the specimen was judged to have been deceased for approximately 2 weeks, and to have floated on the ocean for over a week prior to washing ashore. For better preservation, the carcass was subsequently transported to the museum. Whale skeleton, skin and residual tissue were repeatedly steamed at high temperatures, and finally soaked by chemical method using an anti-mold agent. Based on morphological characteristics (the presence of the diagnostic Bryde's whales triple head ridge consisting of a central ridge flanked by two lateral rostral ridges; Yamada, 2009; Constantine et al., 2018), the specimen was evaluated as Bryde's whale complex. However, no other features were available for further species/sub-species identification due to the high level of decomposition, leaving the sub-species undefined.

In the present study, we used ancient DNA methods and extracted DNA from the specimen's vertebra. Considering that museum specimens are usually treated by physical and/or chemical processes, we expected to find a more suitable DNA extraction method, and to achieve a precise species/sub-species identification through DNA analysis. We therefore sequenced one fragment of the mitochondrial gene (COI) and constructed a phylogenetic tree on this basis. Our expectation was to identify the sub-species of this stranded Bryde's whales by a DNA barcoding method outlined above.

MATERIALS AND METHODS

To verify the sub-species, a sample of the specimen was extracted and its DNA sequenced for further analyses. A section of vertebra was selected and rinsed with distilled water. After drilling off some surface bone with a sterile drill bit, bone powder was then collected using a new sterile bit. According to precautions established by previously published ancient human DNA (Knapp et al., 2012; Xiong et al., 2022), genomic DNA was extracted in a dedicated aDNA facility at Fudan University. In total, 200 mg of bone power was used for DNA extraction (no sample power was used as negative control) by rotating overnight with 0.25 mg/ml Proteinase K (Merck, Germany) and 0.5 M EDTA (pH 8.0) at 37°C. After centrifuging, the supernatant was added to binding buffer [5 M GuHCl, 40% Isopropanol, 25 mM sodium acetate, and 0.05% Tween-20 (PH 5.2)] and magnetic beads (Enlighten Biotech, China). Then, DNA was eluted by TET buffer (QIAGEN, Germany). Finally, DNA concentration was quantified using a Qubit 2.0 Fluorometer (Thermo Fisher Scientific, United States).

A∼700 bp segment of the mitochondrial COI gene was amplified to verify the specimen using primer pairs Balaenoptera-COI-F2 (ACACTAATCGGAGATGACCAAGTC) and Balaenoptera-COI-R2 (CTGATGTGAAATATGCTCGCG), designed by Primer Premier 5.0 (Lalitha, 2000). Polymerase chain reaction (PCR) was carried out in a total volume of 20 μL, consisting of 1 μL of genomic DNA, 1 μL of each primer, 7μL ddH$_2$O, and 10 μL Premix Taq (TaKaRa, Japan). The PCR temperature profile was as follows: incubation at 94°C for 3 min, 14 cycles of 30 s at 94°C, 30 s at 62°C (–0.5°C every cycle), and 30 s at 72°C; then 20 cycles of 30 s at 94°C, 30 s at 55°C, and 30 s at 72°C; and a final extension at 72°C for 10 min. PCR products were then purified and sequenced with forward primers on an ABI PRISM 3730 DNA capillary sequencer at MAP Tech (China). Newly obtained sequence with high quality was checked and submitted to GenBank under accession number: ON459534.

The COI nucleotide sequence was searched for its similarity using BLAST program from GenBank. Then, the relevant sequences were retrieved as reference sequences where available (mysticate families: Balaenopteridae, Eschrichtidea, Neobalaenidae, and Balaenide). The COI sequence was aligned with reference sequences using Clustal W (Thompson et al., 1994) in MEGA X (Kumar et al., 2018). The alignment was inspected visually and trimmed to the length of the shortest sequence. Phylogenetic analysis was performed on the alignment using the maximum-likelihood method in MEGA X with the bootstrap resampled 1,000 times. Here, a general time-reversible model with a gamma distribution (GTR + G) was gauged as the best-fit substitution model according to the corrected Akaike information criterion, using jModelTest v 2.1.3 (Darriba et al., 2012). Furthermore, the genetic distance (*p*-distance) of Balaenopteridae, Eschrichtiidae, Neobalaenidae and Balaenidae was calculated using MEGA X. To describe the intraspecific variation and relationship between newly obtained sequence and other related species (*Balaenoptera edeni edeni*, *Balaenoptera edeni brydei*, *Balaenoptera borealis,* and *Balaenoptera omurai*; these reference sequences were obtained from GenBank), a haplotype network was constructed by HAPLOVIEWER

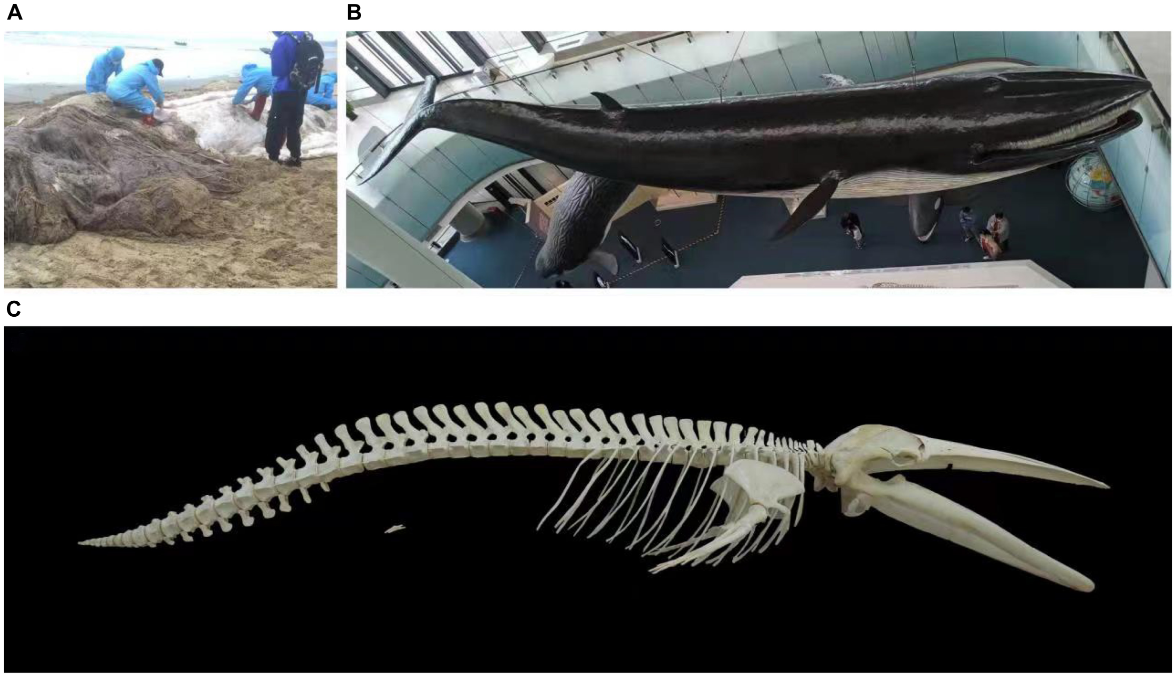


FIGURE 1 | Photographs of the museum whale specimen in this study. **(A)** Uncovering the stranded specimen; **(B)** whole body post-treatment, with characteristic three rostral cephalic ridges used as morphologic identification criterion; **(C)** complete skeleton on display in museum.

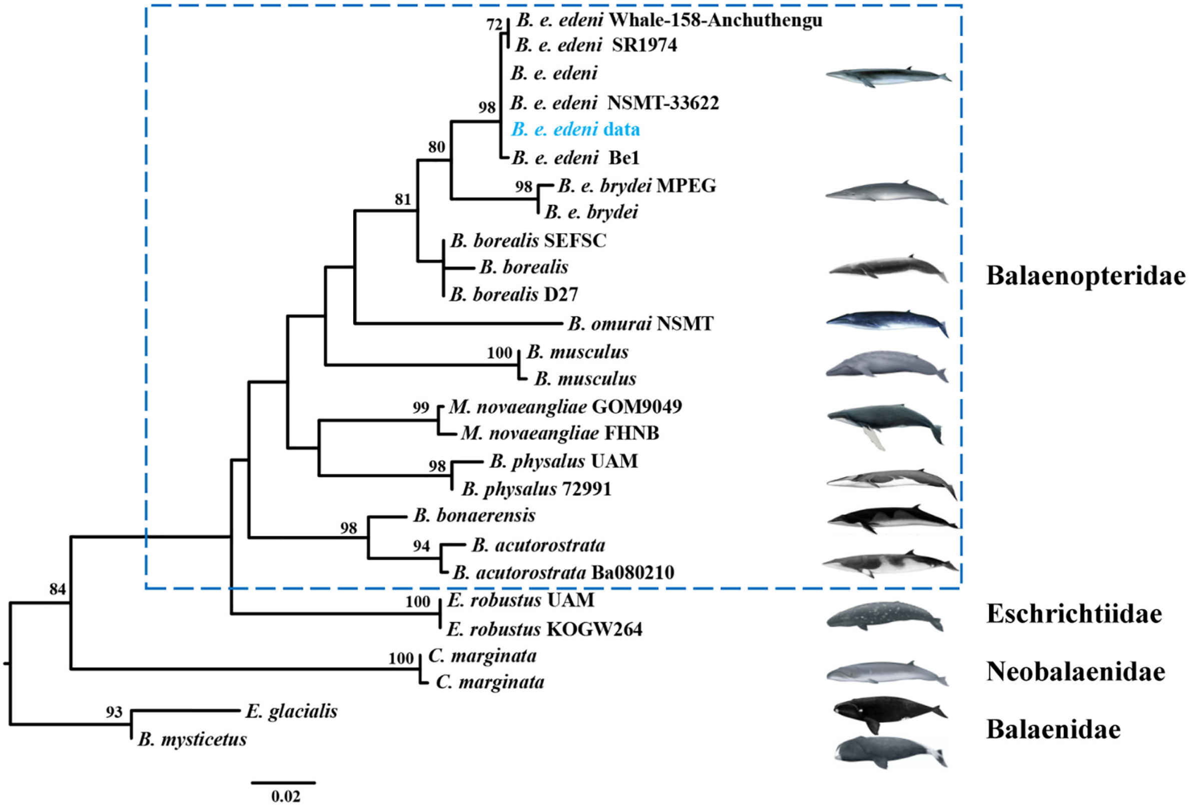


FIGURE 2 | Phylogenetic tree showing the relationship among Mysticeti. Bootstrap support values shown on each node. Data on references sequences provided in **Supplementary Table 1**. Bootstrap support values of under 70% are not displayed. Newly obtained sequence from present study highlighted in blue font.

(Salzburger et al., 2011). Unique COI haplotypes were identified in DnaSP 6 (Rozas et al., 2017).

RESULTS AND DISCUSSION

Other morphological features from the stranded whale specimen provided evidence for the presence of the Bryde's whale complex (**Figure 1A**). The key features of the Bryde's whale (three head ridges) were obvious macroscopically. The specimen's body size, estimated to have reached around 12.5 m in length (**Figure 1B**), was much larger than previous specimens found in the South China Sea (Liu et al., 2021). After physical and/or chemical treatments, the complete skeleton was presented in the museum (**Figure 1C**).

Previous research argued for similar morphological characteristics between *B. e. brydei* and *B. e. edeni* (Constantine et al., 2018; Penry et al., 2018; Castro et al., 2021). Generally, the body length of *B. e. brydei* may exceed *B. e. edeni* (Liu et al., 2021). Here, this stranded specimen is the largest individual of *B. edeni* recorded along the coast of Hainan Province and was initially considered as possibly *B. e. brydei*. Identifying the sub-species of the Bryde's whale specimen in this study was further problematized by serious specimen decomposition. In order to extract DNA successfully, we employed extraction approaches for ancient DNA from a section of vertebrae, avoiding issues such as high temperature, degreasing and EtOH or formalin fixation that are known causes of DNA extraction failures (Ruane and Austin, 2017; McGuire et al., 2018; Pierson et al., 2020). The DNA concentration was to 0.854 ng/μl and suitable for the subsequent analysis. Additional sub-species confirmation was possible after the mitochondrial COI gene (722 bp) of our specimen was successfully sequenced. The phylogenetic tree based on COI (resulting in a 424 bp alignment) of four families (Balaenopteridae, Eschrichtidea, Neobalaenidae and Balaenide; 26 reference sequences provided in **Supplementary Table 1**) within the Mysticeti (baleen whale) group was then reconstructed (**Figure 2**). This phylogenetic tree was congruent with relationships derived from previous combined parsimony analysis of 23 datasets (including morphology, transposon insertions, mitochondrial genomes, cetacean satellite sequences and so on; see Gatesy et al. (2013). Moreover, the phylogenetic relationship for Bryde's like, Sei, and Omura's whales also revealed the same pattern as found previously, split into four clades, corresponding to *B. e. edeni*, *B. e. brydei*, *B. borealis* and *B. omurai* (Rosel et al., 2021). Newly obtained sequence from China belonged to the *B. e. edeni* clade. Based on genetic distance analysis, it showed that the genetic relationship is close between this sequence and *B. e. edeni* (0.000–0.002; see **Supplementary Table 2**). The haplotype network of Bryde's like, Sei, and Omura's whales consisted of 12 haplotypes (1 newly obtained sequence and 30 reference sequences shown in **Figure 3**; reference sequences in **Supplementary Table 3**). Specifically, according to geographical origin, this newly obtained sequence from China, belonged to the lineage of *B. e. edeni*, and shared a COI haplotype (515 bp) with *B. e. edeni* from Japan (AB201258 and NC_007938) and India (JN190945 and GQ856370). In our study, genetic analysis by DNA barcoding (COI) indicated that

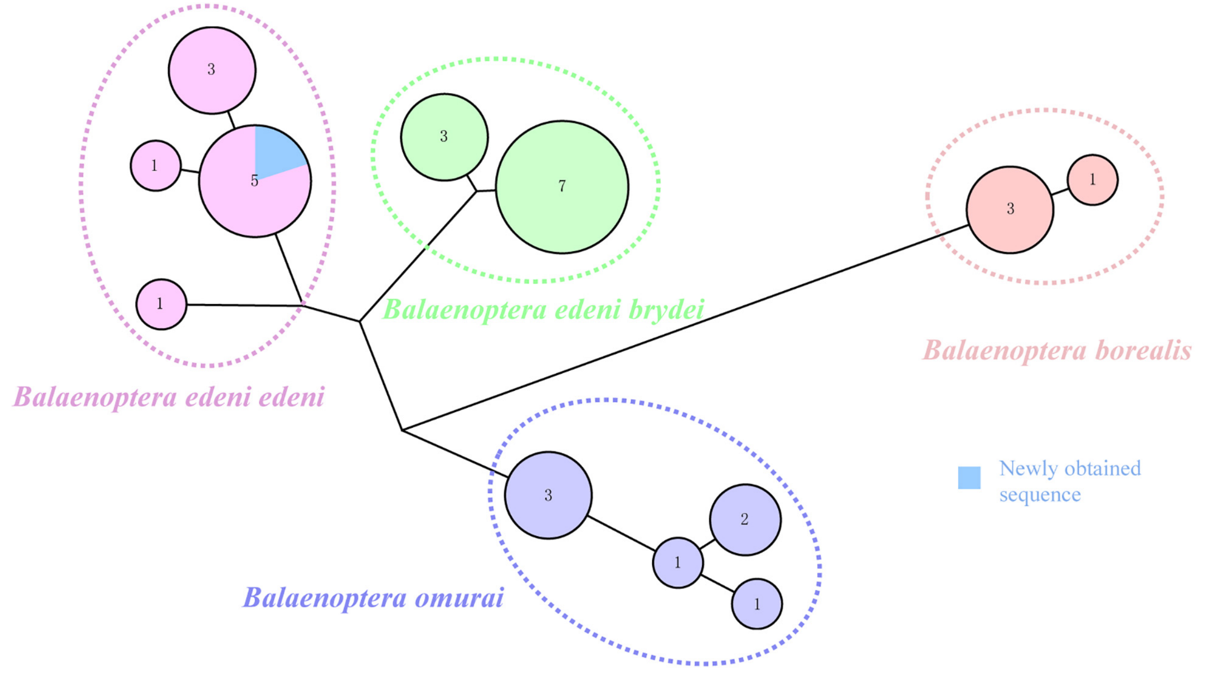


FIGURE 3 | Haplotype network of Bryde's like, Sei and Omura's whales based on the COI gene (515 bp). Each circle represents a unique haplotype and its size indicates the number of individuals carrying the haplotype. Color coding allow easy discrimination of species in the complex. Blue coloring represents a newly obtained sequence. Data for haplotypes from GenBank provided in **Supplementary Table 3**.

this Bryde's whale belonged to *B. e edeni*, a result that could not be conclusively confirmed based on morphology alone. We have proved the efficacy of genetic analysis for identifying cetacean museum specimens to the sub-species level, especially for specimens with non-obvious morphological characteristics, requiring minimal sample sizes without conferring visible damage (Gilbert et al., 2007; Rowley et al., 2007). In our analysis, phylogenetic analysis and haplotype networks provided ample confirmation of species/sub-species identity. This study shows that ancient DNA techniques and DNA barcoding technology can compensate for lack of morphological identification, making it amenable to questions of species/sub-species identification in the museum context.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: https://www.ncbi.nlm.nih.gov/nuccore/ON459534.

ETHICS STATEMENT

The animal study was reviewed and approved by Ethics Committee of Fudan University of Life Sciences.

AUTHOR CONTRIBUTIONS

YF and SW designed and supervised the study. YF provided materials and resources. XR performed genetic laboratory work. XM performed genetic data analysis. EA integrated the genetic data. XR, XM, EA, YF, and SW wrote and edited the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

This work was funded by the National Natural Science Foundation of China (32070576) to SW and the National Natural Science Foundation of China (31900308) to XM.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fevo.2022.921106/full#supplementary-material

Supplementary Table 1 | List of references sequences from Balaenopteridae, Eschrichtiidae, Neobalaenidae and Balaenidae in the COI phylogenetic analyses.

Supplementary Table 2 | The genetic distance (*p*-distance) of Balaenopteridae, Eschrichtiidae, Neobalaenidae and Balaenidae based on COI.

Supplementary Table 3 | List of reference sequences from GenBank used in the haplotype network analyses.

REFERENCES

Anderson, J. (1879). *Anatomical and Zoological Researches: Comprising an Account of the Zoological Results of the Two Expeditions to Western Yunnan in 1868 and 1875; and a Monograph of the Two Cetacean Genera, Platanista and Orcella*. London: B. Quaritch, doi: 10.5962/bhl.title.50434

Bacher, S. (2012). Still not enough taxonomists: reply to Joppa et al. *Trends Ecol. Evol.* 27, 65–66. doi: 10.1016/j.tree.2011.11.003

Barbanera, F., Moretti, B., Guerrini, M., Al-Sheikhly, O. F., and Forcina, G. (2020). Investigation of ancient DNA to enhance natural history museum collections: misidentification of smooth-coated otter (*Lutrogale perspicillata*) specimens across multiple museums. *Belg. J. Zool.* 146, 101–112. doi: 10.26496/bjz.2016.45

Best, P. B. (2001). Distribution and population separation of Bryde's whale *Balaenoptera* edeni off southern Africa. *Mar. Ecol. Prog. Ser.* 220, 277–289. doi: 10.3354/meps220277

Castro, J., Cid, A., and Laborde, M. I. (2021). Bryde's whale (Balaenoptera edeni) new record for mainland Portugal. *J. Cetacean Res. Manage.* 22, 75–80. doi: 10.47536/jcrm.v22i1.333

Chapuis, M.-P., Bazelet, C. S., Blondin, L., Foucart, A., Vitalis, R., and Samways, M. J. (2016). Subspecific taxonomy of the desert locust, Schistocerca gregaria (*Orthoptera*: *Acrididae*), based on molecular and morphological characters. *Syst. Entomol.* 41, 516–530. doi: 10.1111/syen.12171

Chen, B., Zhu, L., Jefferson, T. A., Zhou, K., and Yang, G. (2019). Coastal Bryde's Whales' (Balaenoptera edeni) Foraging Area Near Weizhou Island in the Beibu Gulf. *Aquat. Mamm.* 45, 274–280. doi: 10.1578/AM.45.3.2019.274

Committee on Taxonomy (2022). *List of Marine Mammal Species and Subspecies*. *Soc. Mar. Mammal*. Available online at: http://marinemammalscience.org. (accessed May 2022).

Constantine, R., Iwata, T., Nieukirk, S. L., and Penry, G. S. (2018). Future Directions in Research on Bryde's Whales. *Front. Mar. Sci.* 5:333. doi: 10.3389/fmars.2018.00333

Constantine, R., Johnson, M., Riekkola, L., Jervis, S., Kozmian-Ledward, L., Dennis, T., et al. (2015). Mitigation of vessel-strike mortality of endangered Bryde's whales in the Hauraki Gulf, New Zealand. *Biol. Conserv.* 186, 149–157. doi: 10.1016/j.biocon.2015.03.008

Darriba, D., Taboada, G. L., Doallo, R., and Posada, D. (2012). jModelTest 2: more models, new heuristics and high-performance computing. *Nat. Methods* 9:772. doi: 10.1038/nmeth.2109

Gatesy, J., Geisler, J. H., Chang, J., Buell, C., Berta, A., Meredith, R. W., et al. (2013). A phylogenetic blueprint for a modern whale. *Mol. Phylogenet. Evol.* 66, 479–506. doi: 10.1016/j.ympev.2012.10.012

Gilbert, M. T. P., Moore, W., Melchior, L., and Worobey, M. (2007). DNA extraction from dry museum beetles without conferring external morphological damage. *PLoS One* 2:e272. doi: 10.1371/journal.pone.0000272

Hawlitschek, O., Toussaint, E. F. A., Gehring, P.-S., Ratsoavina, F. M., Cole, N., Crottini, A., et al. (2017). Gecko phylogeography in the Western Indian Ocean region: the oldest clade of Ebenavia inunguis lives on the youngest island. *J. Biogeogr.* 44, 409–420. doi: 10.1111/jbi.12912

Hebert, P. D. N., Cywinska, A., Ball, S. L., and deWaard, J. R. (2003). Biological identifications through DNA barcodes. *Proc. Biol. Sci.* 270, 313–321. doi: 10.1098/rspb.2002.2218

Knapp, M., Clarke, A. C., Horsburgh, K. A., and Matisoo-Smith, E. A. (2012). Setting the stage - building and working in an ancient DNA laboratory. *Ann. Anat. Anat. Anz.* 194, 3–6. doi: 10.1016/j.aanat.2011.03.008

Knowlton, N., and Weigt, L. A. (1998). New dates and new rates for divergence across the Isthmus of Panama. *Proc. R. Soc. Lond. B Biol. Sci.* 265, 2257–2263. doi: 10.1098/rspb.1998.0568

Kumar, S., Stecher, G., Li, M., Knyaz, C., and Tamura, K. (2018). MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms. *Mol. Biol. Evol.* 35, 1547–1549. doi: 10.1093/molbev/msy096

Lalitha, S. (2000). Primer Premier 5. *Biotech. Softw. Internet Rep.* 1, 270–272. doi: 10.1089/152791600459894

Liu, M., Lin, M., Zhang, P., Xue, T., and Li, S. (2019). An overview of cetacean stranding around Hainan Island in the South China Sea, 1978–2016: implications for research, conservation and management. *Mar. Policy* 101, 147–153. doi: 10.1016/j.marpol.2018.04.029

Liu, M., Lin, W., Lin, M., Liu, B., Dong, L., Zhang, P., et al. (2021). The First Attempt of Satellite Tracking on Occurrence and Migration of Bryde's Whale (Balaenoptera edeni) in the Beibu Gulf. *J. Mar. Sci. Eng.* 9:796. doi: 10.3390/jmse9080796

McGuire, J. A., Cotoras, D. D., O'Connell, B., Lawalata, S. Z., Wang-Claypool, C. Y., Stubbs, A., et al. (2018). Squeezing water from a stone: high-throughput sequencing from a 145-year old holotype resolves (barely) a cryptic species problem in flying lizards. *PeerJ* 6:e4470. doi: 10.7717/peerj.4470

Penry, G. S., Hammond, P. S., Cockcroft, V. G., Best, P. B., Thornton, M., and Graves, J. A. (2018). Phylogenetic relationships in southern African Bryde's whales inferred from mitochondrial DNA: further support for subspecies delineation between the two allopatric populations. *Conserv. Genet.* 19, 1349–1365. doi: 10.1007/s10592-018-1105-4

Penry, G., Cockcroft, V., and Hammond, P. (2011). Seasonal fluctuations in occurrence of inshore Bryde's whales in Plettenberg Bay, South Africa, with notes on feeding and multispecies associations. *Afr. J. Mar. Sci.* 33, 403–414. doi: 10.2989/1814232x.2011.637617

Pierson, T. W., Kieran, T. J., Clause, A. G., and Castleberry, N. L. (2020). Preservation-Induced Morphological Change in Salamanders and Failed DNA Extraction from a Decades-Old Museum Specimen: Implications for Plethodon ainsworthi. *J. Herpetol.* 54:137. doi: 10.1670/19-012

Rohland, N., Glocke, I., Aximu-Petri, A., and Meyer, M. (2018). Extraction of highly degraded DNA from ancient bones, teeth and sediments for high-throughput sequencing. *Nat. Protoc.* 13, 2447–2461. doi: 10.1038/s41596-018-0050-5

Rosel, P. E., and Wilcox, L. A. (2014). Genetic evidence reveals a unique lineage of Bryde's whales in the northern Gulf of Mexico. *Endanger. Species Res.* 25, 19–34. doi: 10.3354/esr00606

Rosel, P. E., Wilcox, L. A., Yamada, T. K., and Mullin, K. D. (2021). A new species of baleen whale (*Balaenoptera*) from the Gulf of Mexico, with a review of its geographic distribution. *Mar. Mamm. Sci.* 37, 577–610. doi: 10.1111/mms.12776

Rowley, D. L., Coddington, J. A., Gates, M. W., Norrbom, A. L., Ochoa, R. A., Vandenberg, N. J., et al. (2007). Vouchering DNA-barcoded specimens: test of a nondestructive extraction protocol for terrestrial arthropods. *Mol. Ecol. Notes* 7, 915–924. doi: 10.1111/j.1471-8286.2007.01905.x

Rozas, J., Ferrer-Mata, A., Sánchez-DelBarrio, J. C., Guirao-Rico, S., Librado, P., Ramos-Onsins, S. E., et al. (2017). DnaSP 6: DNA Sequence Polymorphism Analysis of Large Data Sets. *Mol. Biol. Evol.* 34, 3299–3302. doi: 10.1093/molbev/msx248

Ruane, S., and Austin, C. C. (2017). Phylogenomics using formalin-fixed and 100+ year-old intractable natural history specimens. *Mol. Ecol. Resour.* 17, 1003–1008. doi: 10.1111/1755-0998.12655

Salzburger, W., Ewing, G. B., and Von Haeseler, A. (2011). The performance of phylogenetic algorithms in estimating haplotype genealogies with migration. *Mol. Ecol.* 20, 1952–1963. doi: 10.1111/j.1365-294X.2011.05066.x

Savolainen, V., Cowan, R. S., Vogler, A. P., Roderick, G. K., and Lane, R. (2005). Towards writing the encyclopaedia of life: an introduction to DNA barcoding. *Philos. Trans. R. Soc. B Biol. Sci.* 360, 1805–1811. doi: 10.1098/rstb.2005.1730

Thompson, J. D., Higgins, D. G., and Gibson, T. J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* 22, 4673–4680. doi: 10.1093/nar/22.22.4673

Thomsen, P. F., Elias, S., Gilbert, M. T. P., Haile, J., Munch, K., Kuzmina, S., et al. (2009). Non-Destructive Sampling of Ancient Insect DNA. *PLoS One* 4:e5048. doi: 10.1371/journal.pone.0005048

Xiong, J., Du, P., Chen, G., Tao, Y., Zhou, B., Yang, Y., et al. (2022). Sex-Biased Population Admixture Mediated Subsistence Strategy Transition of Heishuiguo People in Han Dynasty Hexi Corridor. *Front. Genet.* 13:827277. doi: 10.3389/fgene.2022.827277

Yamada, T. K. (2009). "Omura's Whale: *Balaenoptera* omurai," in *Encyclopedia of Marine Mammals (Second Edition)*, eds W. F. Perrin, B. Würsig, and J. G. M. Thewissen (London: Academic Press), 799–801. doi: 10.1016/B978-0-12-373553-9.00187-5

Yamada, T. K., Kakuda, T., and Tajima, Y. (2008). Middle sized balaenopterid whale specimens in the Philippines and Indonesia. *Mem. Natl. Sci. Mus. Tokyo* 45, 75–83.

Zhang, P., Li, S., Lin, M., and Xing, L. (2015). *Database of Cetacean Stranding Records around Hainan Island Science Data Bank*. Available Online at: http://doi.org/10.11922/sciencedb.37 (accessed on Jun 25, 2022).

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

ORIGINAL RESEARCH
published: 22 July 2022
doi: 10.3389/fevo.2022.931644

Sequence Capture From Historical Museum Specimens: Maximizing Value for Population and Phylogenomic Studies

Emily Roycroft[1,2]*, *Craig Moritz*[1], *Kevin C. Rowe*[2], *Adnan Moussalli*[2], *Mark D. B. Eldridge*[3], *Roberto Portela Miguez*[4], *Maxine P. Piggott*[1] *and Sally Potter*[1,3]

[1] *Division of Ecology and Evolution, Research School of Biology, The Australian National University, Acton, ACT, Australia,* [2] *Sciences Department, Museums Victoria, Melbourne, VIC, Australia,* [3] *Australian Museum Research Institute, Australian Museum, Sydney, NSW, Australia,* [4] *Department of Life Sciences, Natural History Museum, London, United Kingdom*

The application of high-throughput, short-read sequencing to degraded DNA has greatly increased the feasibility of generating genomic data from historical museum specimens. While many published studies report successful sequencing results from historical specimens; in reality, success and quality of sequence data can be highly variable. To examine predictors of sequencing quality, and methodological approaches to improving data accuracy, we generated and analyzed genomic sequence data from 115 historically collected museum specimens up to 180 years old. Data span both population genomic and phylogenomic scales, including historically collected specimens from 34 specimens of four species of Australian rock-wallabies (genus *Petrogale*) and 92 samples from 79 specimens of Australo-Papuan murine rodents (subfamily Murinae). For historical rodent specimens, where the focus was sampling for phylogenomics, we found that regardless of specimen age, DNA sequence libraries prepared from toe pad or bone subsamples performed significantly better than those taken from the skin (in terms of proportion of reads on target, number of loci captured, and data accuracy). In total, 93% of DNA libraries from toe pad or bone subsamples resulted in reliable data for phylogenetic inference, compared to 63% of skin subsamples. For skin subsamples, proportion of reads on target weakly correlated with collection year. Then using population genomic data from rock-wallaby skins as a test case, we found substantial improvement in final data quality by mapping to a high-quality "closest sister" *de novo* assembly from fresh tissues, compared to mapping to a sample-specific historical *de novo* assembly. Choice of mapping approach also affected final estimates of the number of segregating sites and Watterson's θ, both important parameters for population genomic inference. The incorporation of accurate and reliable sequence data from historical specimens has important outcomes for evolutionary studies at both population and phylogenomic scales. By assessing the outcomes of different approaches to specimen subsampling, library preparation and bioinformatic processing, our results provide a framework for increasing sequencing success for irreplaceable historical specimens.

Keywords: bioinformatics, collections, exon capture, genomics, historical DNA, *Petrogale*, phylogenomics, Murinae

OPEN ACCESS

Edited by:
Jonathan J. Fong,
Lingnan University, China

Reviewed by:
Nicolás Mongiardino Koch,
University of California, San Diego, United States
Niklas Wahlberg,
Lund University, Sweden

****Correspondence:***
Emily Roycroft
emily.roycroft@anu.edu.au

Specialty section:
This article was submitted to Phylogenetics, Phylogenomics, and Systematics,
a section of the journal
Frontiers in Ecology and Evolution

Received: *29 April 2022*
Accepted: *24 June 2022*
Published: *22 July 2022*

Citation:
Roycroft E, Moritz C, Rowe KC, Moussalli A, Eldridge MDB, Portela Miguez R, Piggott MP and Potter S (2022) Sequence Capture From Historical Museum Specimens: Maximizing Value for Population and Phylogenomic Studies.
Front. Ecol. Evol. 10:931644.
doi: 10.3389/fevo.2022.931644

INTRODUCTION

The application of high-throughput, short-read sequencing to historical museum specimens has accelerated the pace of collections-based genomics. Historical museum specimens that were not sampled with the intention to preserve genetic material (e.g., skins, skeletons and fluid-preserved specimens) are now the only viable source of genomic data for many rare, elusive or extinct species, or extirpated populations. Such specimens have proven critical in reconstructing evolutionary history across the Tree of Life (Mason et al., 2011; Guschanski et al., 2013; Besnard et al., 2016; McCormack et al., 2016; Roycroft et al., 2021b), and in understanding genetic responses of species to recent environmental change and anthropogenic impact (Bi et al., 2013, 2019; Roycroft et al., 2021b). Genomic data from museum specimens can fill crucial sample gaps for studies of evolutionary processes across scales of divergence, from population-level to macroevolutionary analyses.

For studies at the population level, historical museum specimens can provide more comprehensive geographic sampling, especially where species are threatened or extirpated (e.g., Ewart et al., 2019; Roycroft et al., 2021b), reducing the effects of sample-bias on population genomic inference (e.g., Battey et al., 2020). Sampling that spans the entire historical range of species allows more accurate estimation of population structure and divergence vs. isolation-by-distance, thereby enabling robust delimitation of species boundaries (Perez et al., 2018). The inclusion of historical specimens may also decrease the impact of "ghost" populations on inference, where failure to sample a population can misrepresent estimates of gene flow and our understanding of introgression (Beerli, 2004; Slatkin, 2005; Hey et al., 2018; Linck et al., 2019). Further, historical genomic data from across space and time increases scope for studies of adaptive evolution and selection (Alves et al., 2019), responses to environmental change (Bi et al., 2013, 2019; Schmitt et al., 2019) and genomic erosion during population decline (Hung et al., 2014; Irestedt et al., 2019; van der Valk et al., 2019; Gauthier et al., 2020; Roycroft et al., 2021b).

Historical museum specimens are also the only source of genetic data for type specimens, and for most rare, elusive or extirpated taxa that are otherwise missing from studies at a phylogenomic or macroevolutionary scale (Ruane and Austin, 2017; McGuire et al., 2018; Wood et al., 2018; Lyra et al., 2020). The inclusion of these specimens mitigates the impact of missing taxa on phylogenetic inference (Streicher et al., 2016), the estimation of speciation and extinction rates (Höhna et al., 2011; Höhna, 2014; Craig et al., 2022) and molecular dating (Linder et al., 2005). Recent studies have also demonstrated how genomic data from extinct taxa can provide unprecedented capacity to resolve long-standing taxonomic uncertainty and reconstruct recent population decline (Grewe et al., 2021; Roycroft et al., 2021b; Pyron et al., 2022). The ability to place extinct or elusive taxa in a phylogenetic and genomic context provides an opportunity to obtain a high-resolution evolutionary reconstruction of all recently extant species, with important implications for conservation biology of persisting species.

While many published studies report successful sequencing results from historical specimens across evolutionary scales, sequencing attempts that result in poor quality or unusable data are typically not reported in scientific literature. Predictors of sequencing success from museum specimens are therefore difficult to assess. Previous studies have suggested that DNA is preserved longer in certain tissue types, e.g., hard tissue like teeth and bone (Adler et al., 2011; Rowe et al., 2011; Burrell et al., 2015; Damgaard et al., 2015; Dabney and Meyer, 2019) and avian toe pads (Tsai et al., 2020) compared to soft tissues like skin. As well as specimen tissue type, decisions during library preparation and bioinformatic processing may also impact final data quality from historical specimens. The consequences of sequencing quality and accuracy on evolutionary inference depend on the research question, and differ between population and phylogenetic studies. For example, erroneous read mapping, variant calling, or missing data may have the most significant impact on the estimation of positive selection in studies of molecular evolution (e.g., Roycroft et al., 2021a), or on fine-scale population genomic parameters. In these cases, studies may focus on ensuring only high-quality and gap-free data are included. In contrast, phylogenomic or macroevolutionary studies may substantially benefit by the inclusion of rare or enigmatic taxa, while tolerating higher levels of missing data. In the latter case, there may be greater emphasis placed on minimizing specimen damage but optimizing sequencing success.

To optimize sequence success and quality at different evolutionary scales, we assessed (1) a phylogenomic dataset of 92 samples from 79 historical museum specimens of Australo-Papuan rodents (family Muridae, tribes Hydromyini and Rattini), and (2) a population genomic dataset from 34 historical skins of four species in the Australian rock-wallaby genus *Petrogale* (Macropodidae: Marsupialia). Using the rodent data, we assess the effect of tissue subsample type, specimen age and library indexing strategy on sequencing success. Using the rock-wallaby data, we test the impact of bioinformatic processing on data accuracy and estimation of population genomic parameters. Specimen collection years range from 1841 to 1997 and were sourced from six different museums spanning three continents. By integrating results across population and phylogenomic datasets, we highlight how steps from specimen subsampling, library preparation, to post-sequencing bioinformatics can be optimized to increase the usability and accuracy of genome sequence data obtained from historical museum specimens.

MATERIALS AND METHODS

Sampling

Rodents

We sequenced 92 samples from 79 specimens (63 species) of Australo-Papuan endemic rodents from the subfamily Murinae, including samples from the tribes Hydromyini and Rattini. Most of these species are known only from museum specimens, including seven extinct species, emphasizing the need to use historical museum specimens to ensure comprehensive sampling. Samples were obtained from museum collections in Australia (Museums Victoria, Australian

Museum, Western Australian Museum, Australian National Wildlife Collection), Europe (Natural History Museum in London), and America (American Museum of Natural History; **Supplementary Table 1**). Specimens were collected between 1841 and 1997 and preserved as dry preparations (also known as "study skins"). We sampled either skin ($n = 49$), toe pad ($n = 34$) or bone ($n = 9$) from each specimen. For nine specimens, we collected multiple samples comprising different tissue types. For skins, we sampled ∼25 mm^2 (5 × 5 mm) from the exposed area of the underbelly, where the preparatory incision had previously been made. For toe pads, we removed ∼1 mm^2 from a single digit. This subsample size difference was intended to maximize the amount of respective DNA obtained from each sample, as preliminary results indicated toe pad yielded more DNA than skin. The DNA quantity for each sample was later normalized during library preparation. Bone was sampled opportunistically, where the specimen had experienced previous damage resulting in broken/exposed bone that could be sampled without additional consequence to the specimen.

Rock-Wallabies

We sampled 56 museum skins from four species of rock-wallaby from the *brachyotis* group of the genus *Petrogale*, and eight reference samples from modern tissues (one from each known lineage; Potter et al., 2014). Historical specimens sampled included *P. brachyotis* ($n = 18$); *P. burbidgei* ($n = 3$); *P. concinna* ($n = 16$) and *P. wilkinsi* ($n = 19$). Samples were obtained from Australian museum collections (Australian National Wildlife Collection, Museums Victoria and the Western Australian Museum, **Supplementary Table 2**). Specimen collection years ranged from 1912 to 1977, and specimens were all preserved as dry study skins. To minimize invasive sampling, we took ∼ 5 mm x 5 mm pieces of skin from the ear, or dried skin still attached to skulls.

DNA Extraction

For rodent samples, DNA was extracted following a modified version of a standard phenol-chloroform-isoamyl DNA extraction protocol (Roycroft et al., 2021b, and provided in the Supplementary Material), in the Museums Victoria Ancient DNA facility. For rock-wallaby samples, DNA was extracted using the DNeasy Blood and Tissue Kit (Qiagen GmbH, Hilden, Germany) using aerosol barrier pipette tips, with working surfaces and equipment wiped down with Lookout DNA Erase (Sigma-Aldrich) before each use. Extractions were undertaken in a dedicated trace DNA laboratory at the Australian National University.

Library Preparation, Hybridisation, and Sequencing

Both the rodent and rock-wallaby datasets were obtained through exon capture target enrichment. All sample libraries were prepared using (Meyer and Kircher, 2010) protocol, including modifications made by Bi et al. (2013). For rodent samples, we used a murine-specific custom exon capture design (SeqCap EZ Developer Library; Roche NimbleGen), targeting 1.27 Mb of genomic DNA (1417 exons, see Roycroft et al., 2020). Rodent samples were either indexed with a single unique barcode, or with a dual-indexing approach, and pooled across multiple captures with up to 92 samples at equimolar ratios (1.2 µg total). Dual-indexed samples were barcoded with a combination of one of 96 unique p5 index sequences, and one of 24 unique p7 index sequences. For rock-wallaby samples, we used a *Petrogale*-specific custom exon capture approach (SeqCap EZ Developer Library; Roche NimbleGen), which targets 1.83 Mb of genomic DNA (3960 exons), designed using transcriptome data from a yellow-footed rock-wallaby (*Petrogale xanthopus*) (see Bragg et al., 2016; Potter et al., 2017, 2022). Rock-wallaby samples were indexed with a single unique barcode, and all 56 samples were pooled at equimolar ratios (1.2 µg total).

For both datasets, pooled libraries were then hybridized for ∼72 h, with 5 µg of mouse Cot-1 DNA (Life Technologies Corporation), barcode specific blocking oligos (1000 pmol) and target probes following the SeqCap EZ Developer Library protocol. Post incubation, the hybridization reaction was amplified in two independent enrichment PCRs and then cleaned up using the QIAquick PCR purification kit (Qiagen). Quality control checks were made using the DyNAmo Flash SYBR green qPCR kit (Thermo Fisher Scientific Inc.; see Bi et al., 2012) to assess global enrichment of the target exons by comparing pre-capture pooled genomic libraries to the post-capture cleaned hybridization reaction and specifically designed to hit targets of the hybridization probes. After passing these quality control checks, the enriched hybridization samples were run on a BioAnalyzer (2100; Agilent Technologies, Inc.) to check the quality and quantity of the libraries prior to sequencing. Each pooled library was then sequenced on a single lane of an Illumina HiSeq 2500 (100 bp paired-end run) at the ACRF Biomolecular Resource Facility.

Sample Processing and Bioinformatics

We processed raw sequencing data from all specimens using *Exon Capture Pipeline for Phylogenetics* (ECPP, https://github.com/Victaphanta/ECPP), following the protocol described in Roycroft et al. (2020). For a subset of rodent samples, we ran mapDamage2 (Jónsson et al., 2013) to assess the extent of DNA misincorporation. For rock-wallaby samples, reflecting population genetic sampling, we compared the effect of mapping to a sample-specific reference versus mapping to the highest-quality assembly from the closest non-historical sister sample. Initially, we implemented the sample-specific reference approach which creates a *de novo* assembly for each historical sample (the "historical *de novo*" dataset). This is the default approach in ECPP, and in other commonly used target capture assembly pipelines (e.g., Bragg et al., 2015; Faircloth, 2016; Singhal et al., 2017). These historical *de novo* assemblies were used to create sample-specific references, and raw reads were then mapped back to each reference using BBmap (version 35.82, sourceforge.net/projects/bbmap/) with a minid threshold of 0.95. As a comparison, we also used a high-quality "closest sister" reference approach to map reads (see Roycroft et al., 2021b). To do this, we generated a reference set of high-quality *de novo* assemblies from fresh tissue samples of various *Petrogale* sub-species. Using the same mapping approach as above, we mapped the raw reads from historical samples to the closest sister sample (i.e., lowest evolutionary distance from each historical population) with a high-quality assembly (the "high-quality *de novo*" dataset). The sample with the lowest evolution distance was determined based on divergence between the sample and the reference in substitutions per site, calculated in IQ-TREE 1.6.9 (Nguyen et al., 2015). In all cases, reads from historical samples were mapped to a high-quality *de novo* assembly of the same sub-species. For all rodent specimens, we only applied this "high-quality *de novo*" mapping approach, as preliminary results showed this had superior performance over the default method. Final alignments for all data were filtered at a threshold of 3% heterozygosity per locus, and processed with BMGE (Criscuolo and Gribaldo, 2010) to remove poorly represented regions.

Summary Statistics and Branch Length Estimation

To estimate population genomic summary statistics, we filtered the rock-wallaby dataset to 3742 loci that were >90% sample-complete and split samples into seven populations (*Petrogale brachyotis brachyotis*; BB, *Petrogale brachyotis victoriae*; BV, *Petrogale concinna canescens*; CC, *Petrogale concinna monastria*; CM, *Petrogale wilkinsi* core population; W, *Petrogale wilkinsi* Gulf of Carpentaria population; wGU, and *Petrogale wilkinsi* Groote Eylandt population; wGR). A total of 22 originally sequenced rock-wallaby samples were excluded due to insufficient coverage and poor data quality (see **Supplementary Table 2**). For each population, we calculated the number of segregating sites and proportion of segregating sites to valid sites (to account for missing data) in PopGenome in R (Pfeifer et al., 2014). We also estimated Watterson's theta (θ, Watterson, 1975) and Tajima's *D* (Tajima, 1989) in PopGenome, as these are common metrics used to assess genetic diversity and population dynamics. We repeated all calculations for both the "historical *de novo*" and "high-quality *de novo*" datasets, across all loci. We also performed all calculations using only exons which matched between the two *Petrogale* datasets, to directly compare the effect of mapping strategy on parameter estimation. We tested for significant differences across datasets using Welch's two-sample *t*-test. As a further comparison, we used IQ-TREE 1.6.9 (Nguyen et al., 2015) with codon partitions to infer terminal branch lengths (in substitutions per site) for both the "historical *de novo*" and "high-quality *de novo*" datasets. Accurate estimation of tip branch lengths are important, as they are increasingly used as metrics for speciation rates (e.g., ClaDS, Maliet et al., 2019) and in analyses of variation in rates of molecular evolution (e.g., Ivan et al., 2022).

Using the final processed rodent phylogenomic data, we calculated the proportion of reads on target (i.e., the total proportion of deduplicated sequenced reads that mapped to the target region) and the proportion of total target loci successfully captured (>40% of target region) for all specimens. For each sample, we also calculated the average heterozygosity across all loci as a measure of sequence quality and accuracy, where outliers with high values are assumed to contain a higher rate of error. We then compared these metrics across sampled tissue type (toe pad or bone vs. skin) and library indexing strategy (single vs. dual-indexed). Toe pad and bone samples were grouped, due to the comparatively high success rate among these two tissue types and the low overall number of samples from bone. Using the *stats* package in R, we applied generalized linear models (GLM) to model two categorical predictors (indexing strategy and tissue type) and a continuous predictor (specimen age) on four continuous response variables; proportion of reads on target, loci captured, average coverage and heterozygosity. We also used a two-sample *t*-test to test each of these variables for significant differences in response to indexing strategy and tissue type.

RESULTS

Predictors of Capture Efficacy and Sequence Quality in Phylogenomic Data

Across all rodent genomic libraries sequenced, 91% (31 out of 34) toe pad and 89% (8 out of 9) bone subsamples resulted in useable sequence data, compared to 63% (31 out of 49) of skin samples (**Supplementary Table 1**). Unusable samples were those that either returned no sequence data after processing in ECPP, or where data was returned, were primarily sequencing contaminants. We took a conservative approach to screening for contaminant samples, by excluding all samples that showed a terminal branch length at least ∼20% greater than close relatives sequenced using high-quality DNA. Results from mapDamage2 suggested that the effect of DNA damage was relatively minor, but was more evident in subsamples taken from the skin, compared to bone or toe pad of the sample specimen (see **Supplementary Figure 1** for an example).

The proportion of reads on target (**Figure 1A**) was significantly lower for single-indexed samples than for dual-indexed samples, while the difference between the number of loci captured (**Figure 1B**) or average coverage (**Figure 1C**) was not significant (**Table 1**). Generalized linear models (GLM) found that collection year was a significant predictor ($p < 0.05$) for the proportion of reads on target (**Supplementary Table 3**). Proportion of reads on target tended to be higher for samples that were collected more recently, especially for skin subsamples (**Figure 1E**). Our GLMs also indicated indexing approach was a significant predictor ($p < 0.01$) of heterozygosity (**Supplementary Table 3**), with average heterozygosity across loci (**Figure 1D**) significantly higher for single-indexed samples than for dual-indexed samples (**Table 1**). Interactions between collection year, tissue type, and indexing strategy also had significant effect on heterozygosity (**Supplementary Table 3**).

When dual-indexed samples were grouped by source tissue type, the average coverage, reads on target and loci captured were all significantly higher in toe pad/bone subsamples compared to skin subsamples (**Table 2**). There was no significant difference in average heterozygosity when comparing tissue types. There was a weak relationship between specimen age and reads on target (**Figure 1E**) for skin subsamples ($r = 0.29$, $p < 0.05$), and no relationship for toe pad/bone subsamples ($r = 0.037$, $p = 0.84$). There was no relationship between specimen age and the

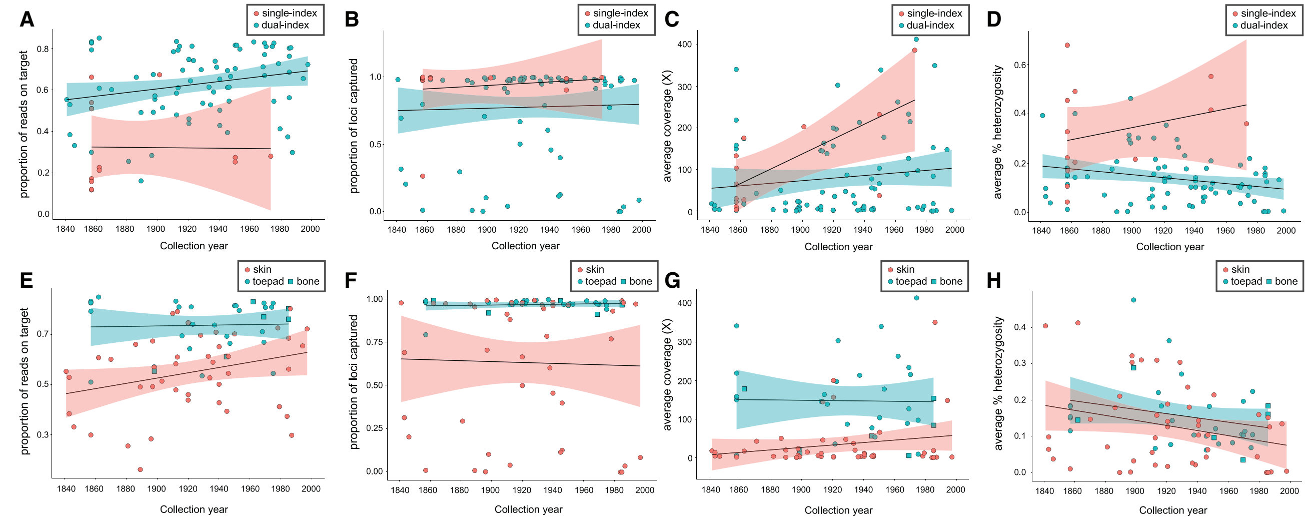


FIGURE 1 | Relationship between year of specimen collection and indexing strategy for **(A)** proportion of reads on target, **(B)** proportion of targeted loci captured, **(C)** average coverage, **(D)** average percent heterozygosity, and for dual-indexed samples only; between year of specimen collection and tissue type for **(E)** proportion of reads on target, **(F)** proportion of targeted loci captured, **(G)** average coverage, **(H)** average percent heterozygosity.

TABLE 1 | Sequencing success of single- vs. dual-indexed rodent samples.

	Single (mean)	Dual (mean)	Difference *p*-value (*t*-test)
Average coverage	110.41	81.41	n.s.
Prop. reads on target	0.32	0.63	<0.001
Prop. loci captured	0.93	0.78	n.s.
Average heterozygosity	0.33	0.14	<0.001

TABLE 2 | Sequencing success of different tissue types for dual-indexed rodent samples.

	Toe pad/bone (mean)	Skin (mean)	Difference *p*-value (*t*-test)
Average coverage	147.65	33.90	<0.001
Prop. reads on target	0.74	0.55	<0.001
Prop. loci captured	0.97	0.64	<0.001
Average heterozygosity	0.16	0.13	n.s.

proportion of loci captured (**Figure 1F**) for either skin or toe pad/bone subsamples.

The Effect of Mapping Strategy on Population Genomic Parameters

To assess the potential impact of mapping strategy on downstream population genomic inference, we compared inferred population parameters between the "historical *de novo*" and "high-quality *de novo*" datasets for *Petrogale* rock-wallabies. A total of 34 out of 56 rock-wallaby skins (61%) yielded sufficient data for use in population genomic analyses. The proportion of segregating sites to valid sites was always higher for the historical *de novo* approach, suggesting a higher error rate (**Table 3**). In five out of seven populations, the "historical *de novo*" mapping approach resulted in a greater absolute number of apparent variable sites (up to 10% more per population) than the "high-quality *de novo*" approach. In both cases where the "high-quality *de novo*" dataset had a greater absolute number of variable sites, this was explained by a 38% (CM) and 77% (wGU) increase, respectively, in total number of recovered sites when using the "high-quality *de novo*" mapping approach compared to the "historical *de novo*" approach. In five out of seven populations, the "high-quality *de novo*" mapping approach also resulted in fewer missing sites than the "historical *de novo*" approach.

In all cases except for the wGU population, the inferred number of segregating sites and the estimated θ values were consistently higher in the "historical *de novo*" dataset compared to the "high-quality *de novo*" dataset (**Figure 2A** and **Table 3**). The impact of mapping strategy resulted in a significant difference ($p < 0.05$) in the number of segregating sites for the CC, CM, BB and wGU populations using a two-sample *t*-test (**Figure 2A**). However, the differences were not significant when only overlapping loci were considered (**Supplementary Table 4**). We also recovered a significant difference in θ estimates for CC, CM, BB and wGU (**Figure 2B** and **Table 3**), and for CC, CM and BB in overlapping loci (**Supplementary Table 4**). The proportion of segregating sites to valid sites was smaller for the "high-quality *de novo*" dataset compared to the "historical *de novo*" dataset across populations, with a significant difference in the CC, CM, BB, WGR populations (**Table 3**), and only for CM in overlapping

TABLE 3 | Population genomic summary statistics across rock-wallaby populations for both "historical *de novo*" and "high-quality *de novo*" mapping approaches.

Population	Mapping approach	# loci	# valid sites (vs)	# unknown sites	# segregating sites (ss)	Proportion of ss to vs	Mean Tajima's *D* ± SD	Mean θ ± SD
CC	historical *de novo*	340	229,038	1,062	737	0.0032	−0.401 ± 0.715	1.18 ± 2.52
	high-quality *de novo*	316	220,405	2,507	457	0.0021	−0.405 ± 0.697	0.79 ± 0.59
CM	historical *de novo*	1,329	546,333	163,714	3,423	0.0063	−0.926 ± 0.638	0.99 ± 2.89
	high-quality *de novo*	1,385	755,282	108,429	2,587	0.0034	−0.938 ± 0.606	0.72 ± 0.97
BV	historical *de novo*	222	137,951	3,816	781	0.0057	−0.257 ± 0.963	1.69 ± 4.97
	high-quality *de novo*	203	139,645	848	515	0.0037	−0.281 ± 0.942	1.23 ± 4.04
BB	historical *de novo*	512	214,860	70,883	1,740	0.0081	−1.039 ± 0.433	1.39 ± 4.94
	high-quality *de novo*	464	314,115	10,080	784	0.0025	−1.054 ± 0.358	0.69 ± 0.54
W	historical *de novo*	583	366,996	19,419	1,339	0.0037	−0.867 ± 0.743	0.81 ± 2.69
	high-quality *de novo*	558	364,575	14,739	1,090	0.0030	−0.882 ± 0.694	0.69 ± 2.62
wGR	historical *de novo*	85	53,588	785	371	0.0069	NA	2.91 ± 11.08
	high-quality *de novo*	79	49,156	734	99	0.0020	NA	0.84 ± 0.47
wGU	historical *de novo*	755	279,258	128,192	1,638	0.0059	−0.434 ± 0.846	1.04 ± 2.50
	high-quality *de novo*	979	494,550	217,020	2,639	0.0053	−0.520 ± 0.818	1.29 ± 1.79

For Tajima's D and Watterson's theta (θ) the mean and the standard deviation (SD) are reported.

loci (**Supplementary Table 4**). While estimated mean Tajima's *D* values were consistently higher for the "high-quality *de novo*" dataset compared to the "historical *de novo*" dataset, the difference was only significant in the wGU population (**Table 3**).

The difference in terminal branch length for *Petrogale* samples in the "historical *de novo*" and "high-quality *de novo*" datasets was variable, but with all large differences having over-inflated branch length in the "historical *de novo*" dataset. This was especially evident for individuals within the CM and BB populations (**Figure 2C**). The impact of these individuals is also reflected in population-level significant differences in summary statistics (**Table 3** and **Figure 2**). Overall, samples with comparatively lower quality (i.e., lower coverage, higher heterozygosity in historical *de novo* dataset) tended to show the most reduction in terminal branch length when using the "high-quality *de novo*" reference compared to the "historical *de novo*" reference. Higher quality (i.e., higher coverage, fewer errors in historical *de novo* dataset) samples tend to show slightly longer terminal branch length using the "high-quality *de novo*" reference compared to the "historical *de novo*" reference.

DISCUSSION

Using population and phylogenomic data generated from historical museum specimens, we demonstrate that choices prior to DNA extraction (i.e., type of tissue subsampled), during library preparation (i.e., indexing) and post-sequencing bioinformatic processing (i.e., mapping) have significant impacts on the success, usability, and quality of genomic sequence data and inference. Further, we show how the use of a high-quality reference assembly for mapping reads from historical specimens can result in significant differences in the amount of final data recovered, inferred population genomic summary statistics and phylogenetic tip lengths compared to a *de novo* sample-specific approach. This demonstrates the importance of the availability of high-quality reference assemblies from closely-related taxa, especially for studies including sequence data from historical specimens where maximum data recovery and accurate variant calling are crucial.

In synthesizing our results, we provide a framework for optimizing pre- and post-sequencing protocols for irreplaceable historical dried mammal specimens at both population and phylogenomic scales.

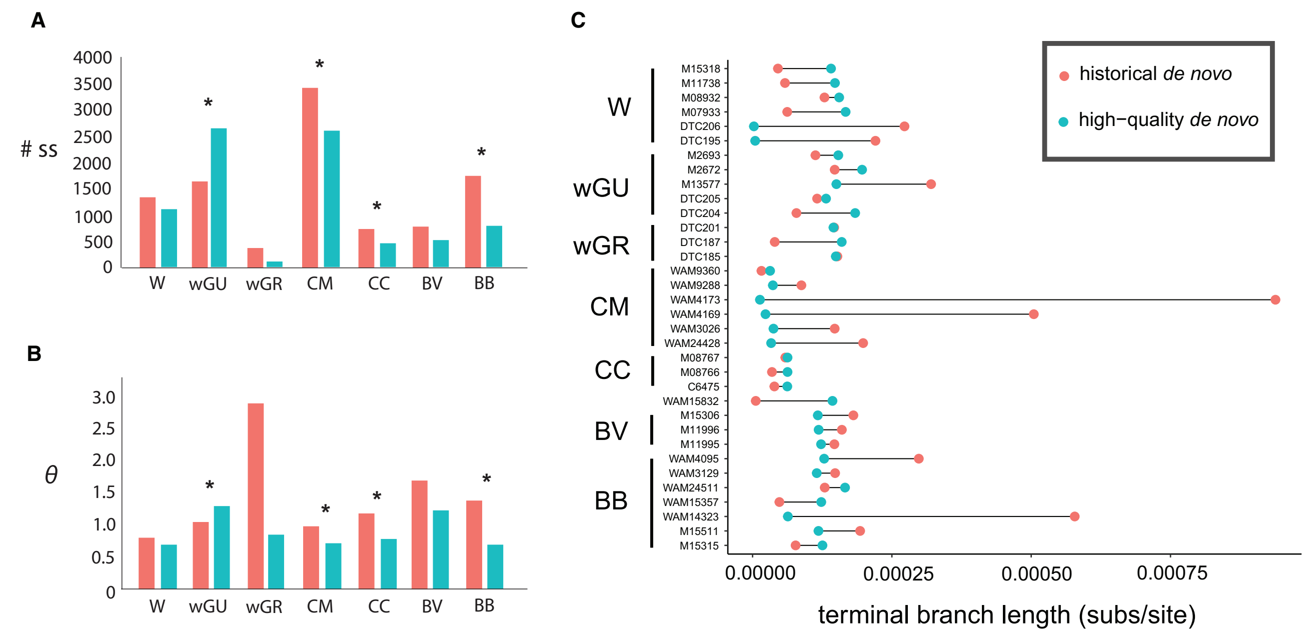


FIGURE 2 | Comparison between the "historical *de novo*" (pink) and "high-quality *de novo*" (blue) datasets for **(A)** total number of segregating sites, and **(B)** average Watterson's theta (θ) estimates, and **(C)** terminal branch length (substitutions per site). Samples are identified by the voucher number and grouped by population; W, *wilkinsi*; wGU, *wilkinsi* Gulf of Carpentaria; wGR, *wilkinsi* Groote Eylandt; CM, *concinna monastria*; CC, *concinna canescens*; BV, *brachyotis victoriae*; BB, *brachyotis brachyotis*. Significant differences between the "historical *de novo*" and "high-quality *de novo*" datasets for each population are designated by * in **(A,B)**.

Pre-sequencing Predictors of Data Quality From Historical Specimens

Across rodent specimens, DNA extracted from subsamples of toe pad or bone consistently performed better than skins in terms of sequencing success, capture specificity (proportion of sequence reads on target), data quality and accuracy (heterozygosity), and completeness (final number of loci captured). This is consistent with previous studies of avian toe pads and bone compared to skin (Tsai et al., 2020), and from ancient DNA studies that have found harder tissues like bones and teeth to preserve DNA for longer than soft tissues (Adler et al., 2011; Burrell et al., 2015; Damgaard et al., 2015; Dabney and Meyer, 2019). Notably however, we found no effect of specimen age on the proportion of reads on target or number of loci captured for toe pad or bone subsamples, indicating a high level of protection from post-mortem DNA damage and degradation in these tissue types. This is in contrast with results from McCormack et al. (2016), who found a decrease in total assembled sequence data for avian specimens with age, but appears to be consistent with results from Sawyer et al. (2012), who found minimal effect of DNA fragmentation across time in specimens up to 60,000 years. In our data, rodent toe pads and bone also yield high-quality endogenous DNA with no observable relationship to specimen age, demonstrating the feasibility of obtaining reliably high-quality genomic sequence data from specimens spanning the last three centuries.

In contrast, DNA sequence libraries prepared from skin subsamples had a significantly lower rate of sequencing success (63% for skins, compared to 93% in toe pad/bone), and a weak but significant relationship ($r = 0.29$, $p < 0.05$) between specimen age and sequenced reads on target (**Figure 1E**). Where DNA is more fragmented and has a greater degree of post-mortem damage (e.g., **Supplementary Figure 1**), overall capture efficiency is likely to be lower. This would explain the difference in loci captured for skin subsamples compared to toe pad and bone (**Figure 1F**). If capture efficiency is lower in these samples, then the relative amplification of off-target DNA in the post-capture PCR is likely to be greater. In turn, this may explain the effect of tissue type and specimen age we observed for skin subsamples for proportion of reads on target (**Figure 1E**). Previous studies (Pääbo et al., 2004) have suggested a relationship between specimen age and DNA quality, as well as a decrease in endogenous DNA via degradation and an increase in exogenous DNA via contamination over time. However, recent studies suggest that the specimen preservation and storage may be crucial factors for collections-age material (McCormack et al., 2016; McDonough et al., 2018). In our data, DNA degradation with specimen age was only evident for skin subsamples, and not for toe pad and bone. As the skins of prepared museum specimens are thinner and more exposed to the environment than toe pad or bone, DNA content and quality in these tissues is likely dependent on the conditions of specimen storage, superficial treatment of the skin with chemicals (e.g., arsenic), and handling of the specimen.

We also found a significant difference between proportion of reads on target and heterozygosity in single- vs. dual-indexed samples, with reads on target being lower and heterozygosity higher in single-indexed samples. Higher average heterozygosity in single-indexed samples may be explained by a combination of cross-index contamination during genomic library preparation, and background cross-indexing during sequencing. Both types of contamination can be reduced by using unique (or partially unique) dual indexes (Kircher et al., 2012). Interestingly, we also report overall fewer reads on target for single-indexed samples. This may be explained by contaminant libraries using the same index from the laboratory environment, resulting in an apparent lower overall capture efficiency. In this case, dual-indexing also reduces the likelihood that any cross-library contamination contains a matching pair of indexes, especially where effort is made to alternate between combinations of indexes across experiments. However, we note that the overall number of single-indexed libraries in our data was low compared to the number dual-indexed libraries, and so it is possible that the patterns we observe are an artifact of sample library variability.

Practical Guidelines for Specimen Selection and Subsampling

While historically preserved DNA has the potential to be highly valuable, this is in addition to the existing intrinsic taxonomic, morphological and historical significance of specimens. Destructive sampling may interfere with potential diagnostic characters, which are often taxonomic group specific (e.g., ear length and shape, nose leaf morphology, toe pad morphology and number, etc.). Further, museum specimens are finite and irreplaceable sources of genomic material, especially for specimens of rare or extinct taxa. As such, it is critical to follow minimally invasive procedures when subsampling material from historical specimens, as well as ensuring optimal genomic library preparation and bioinformatic post-processing decisions to maximize data accuracy and utility. For dry museum skins of small mammals like rodents, subsamples of skin from around the preparatory incision may be the least invasive, however our results suggest that DNA quality and sequencing success from such subsamples is variable, and as such there is an increased chance that DNA extraction and sequencing from these subsamples will fail. Sampling from harder tissue types like toe pad or bone is therefore more likely to result in high-quality genomic data. Where practical considerations warrant subsampling from skin in the first instance, our results show that library preparation using a dual-indexing, rather than single-indexing, may minimize contamination and maximize the chance of obtaining useable data. Recent advances in sequencing genomic DNA from formalin-fixed specimens (e.g., Hykin et al., 2015; Ruane and Austin, 2017), and historical ethanol-preserved specimens (e.g., Derkarabetian et al., 2019) may also present viable options for sampling as an alternative to skins, although with variable success.

Post-sequencing Optimisation of Historical Sequence Data

Using population genomic data from *Petrogale* rock-wallabies, we demonstrate that increased reference quality can have substantial impact on population genomic parameters and terminal branch length estimation. Previous studies have also demonstrated the impact of reference choice, for example Shafer et al. (2017) found that a reference-based approach recovered lower inbreeding coefficient (F_{IS}) values than a *de novo* approach for RAD-seq data. In our case, we hypothesize that for historical samples, mapping to a sample-specific *de novo* assembly can reinforce error that is present at low levels in the historical sequence data (e.g., due to DNA damage or sequencing error). Our results show that the use of a high-quality *de novo* reference can both reduce error and increase data completeness.

At an individual level, we found that samples with the overall lowest quality tended to show the most significant reduction in terminal branch length when using the "high-quality *de novo*" reference compared to the "historical *de novo*" reference. For cases where historical samples were comparatively high quality from the outset, terminal branch length tended to be slightly longer using the "high-quality *de novo*" reference compared to the "historical *de novo*" reference. This was due to an increase in legitimate variable sites when using a more complete and contiguous reference for mapping historical reads. While the "high-quality *de novo*" mapping approach is likely to have the most impact on samples of lower initial sequence quality, total population sample size may also be a contributing factor. For example, although the overall inferred terminal branch length differences were relatively small for samples within the CC population (**Figure 2C**), the alternative mapping approaches resulted in significant differences in population genomic summary statistics (**Table 3**). For populations with lower sample sizes, small changes in allele frequencies may have greater relative effect on estimated summary statistics (e.g., Fumagalli, 2013).

When summary statistics were inferred at a population-level, we saw a significant impact on the inferred number of segregating sites and Watterson's θ estimates in four of the seven populations. This was despite most populations containing individuals with higher-quality sequence data (see **Figure 2C**), which may be expected to mask the impact of low-frequency errors. In the CM and BB populations, pronounced differences in terminal branch length of individuals correspond to significant differences in inferred summary statistics. However, populations with significant differences in inferred summary statistics at a population level did not always show obvious differences in terminal branch length (e.g., the CC population). In cases where sequence quality is reduced pervasively across individuals in a population, errors introduced by DNA damage, sequencing error or bioinformatic processing are likely to have greater consequences. This may then impact the accuracy of downstream inference of genetic diversity, population structure, population size and demographic processes.

The Importance of Data Accuracy for Population Genomic Inference From Historical Specimens

The inclusion of historical museum specimens in population genomics provides the opportunity to sample extirpated populations, potentially contributing to the delimitation of species boundaries and conservation units, assessment of extinction risk and studies of population decline (e.g., Mondol et al., 2013; Nakahama and Isagi, 2018; Nakahama, 2021). Optimizing sequence quality from historical specimens is crucial in empirical systems like *Petrogale*, where complex patterns of mito-nuclear discordance (Potter et al., 2012, 2014), introgression (Potter et al., 2015, 2017, 2022), and incomplete lineage sorting across the landscape can only be resolved with comprehensive geographic sampling. In addition, data quality and completeness are especially important in studies using targeted exon capture approaches for population genomics (e.g., Bi et al., 2012; Belkadi et al., 2016; Potter et al., 2016), where there are often limited segregating sites within exonic loci. In such cases, a decrease in data completeness can reduce power to detect genuine population level variation, but equally, the impact of erroneous variant calling can be more severe.

It has long been recognized that variation in data quality, accuracy and completeness can have considerable impact on inference and conclusions in population genomic studies. The allele frequency spectrum, a summary of the distribution of derived allele frequencies, is commonly used in population genomic inference. Estimated allele frequencies can be highly sensitive to bioinformatic approaches, potentially impacting estimates of demographic expansion and isolation-with-migration models (Shafer et al., 2017). Many analytical approaches use allele frequency estimates to determine population structure (e.g., STRUCTURE, Pritchard et al., 2000), gene flow (e.g., DILS, Fraïsse et al., 2021; ABBA-BABA tests, Durand et al., 2011; TreeMix, Pickrell and Pritchard, 2012), and demographic history (e.g., δaδi, Gutenkunst et al., 2009; range expansion tests, Peter and Slatkin, 2013, 2015). Inflation of the number of variable sites, as reported in our results, could have profound influence if skewed to increase the number of minor alleles in a population, influencing patterns of demographic expansion, and evaluation of selection and adaptation, common population genomic analyses where museum specimens have been incorporated (e.g., Bi et al., 2013; Ewart et al., 2019; Dussex et al., 2021). Low frequency variants, or minor alleles, can significantly influence population structure (Linck and Battey, 2019) and estimates of demographic history (e.g., Shafer et al., 2017).

Our results showing the effect of *de novo* assembly quality on population genomic summary statistics demonstrate the importance of maximizing the quality and contiguity of the mapping reference and highlight the complexities in bioinformatic processing and analyzing data from historical museum specimens. This is especially true in contexts where accuracy is crucial. While sample-specific *de novo* assemblies have been routinely used in many target capture bioinformatic pipelines (e.g., Bragg et al., 2015; Faircloth, 2016; Singhal et al., 2017) to mitigate against reference bias (Sousa and Hey, 2013), we caution against a true "sample-specific" approach for historical specimens. While some historical specimens can provide high-quality *de novo* assemblies, these are typically not as contiguous as *de novo* assembly obtained from fresh tissue. Where fresh tissues are available from the same or closely related species, studies should endeavor to generate "high-quality *de novo*" assemblies from close relatives as a reference prior to sampling historical specimens. For population level studies, bias may be further reduced by selecting loci at random from multiple fresh specimens per lineage (e.g., Potter et al., 2016), or data recovery increased by mapping individuals to a common and highly complete reference for each population (e.g., Potter et al., 2018). The application of iterative mapping approaches (e.g., "pseudoreferencing," Sarver et al., 2017) may also serve to further reduce bias where raw data is mapped to a divergent reference. It is likely that the consequence of "reference bias," even at moderate evolutionary divergences (e.g., above population level to 10 million years), is less than the consequence of potential error and loss of data introduced using a *de novo* assembly generated from historical sequencing reads, however further studies are needed to quantify the impact of evolutionary divergence. Reference genomes for diverse taxa are also now being generated by the research community at a rapid rate, providing an additional source for mapping reads from historical specimens in future work.

DATA AVAILABILITY STATEMENT

The data presented in this study are deposited in the NCBI sequence read archive (SRA) at BioProject Accession PRJNA846960.

AUTHOR CONTRIBUTIONS

ER, SP, and CM conceived and designed the study. ER, SP, ME, RPM, and KR contributed to specimen sampling. ER, SP, KR, and MP performed laboratory work. AM provided bioinformatic tools. ER and SP analyzed the data and wrote the manuscript. All authors edited and approved the final manuscript.

FUNDING

This project was supported by funding from Bioplatforms Australia through the Australian Government National Collaborative Research Infrastructure Strategy, as part of the Oz Mammals Genomics Initiative. This project was also supported by funding (to ER) from the Holsworth Wildlife Research Endowment (Equity Trustees Charitable Foundation and the Ecological Society of Australia), Dame Margaret Blackwood Soroptimist Scholarship, the Museums Victoria 1854 Scholarship, and to SP and MP from a Center for Biodiversity Analysis Ignition Grant. MP was a recipient of a Discovery Early Career Research Award from the Australian Research Council (DE130100777).

ACKNOWLEDGMENTS

The authors wish to acknowledge the Traditional Owners and Custodians of the lands on which we work. We recognize the past atrocities against First Peoples of Australia, who have never ceded sovereignty. We acknowledge the contribution of the Oz Mammals Genomics Initiative consortium in the generation of data used in this publication (https://ozmammalsgenomics.com/consortium/). We thank Niccy Aitken and Anna MacDonald for advice with laboratory protocols. We are also indebted to collections staff and to many individuals who contributed material at the Australian Museum, Museums Victoria, Western Australian Museum, Australian National Wildlife Collection (https://ror.org/059mabc80), American Museum of Natural History and Natural History Museum in London. We thank Pierre-Henri Fabre for assistance in sampling specimens from London.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fevo.2022.931644/full#supplementary-material

REFERENCES

Adler, C. J., Haak, W., Donlon, D., and Cooper, A. (2011). Survival and recovery of DNA from ancient teeth and bones. *J. Archaeol. Sci.* 38, 956–964. doi: 10.1016/j.jas.2010.11.010

Alves, J. M., Carneiro, M., Cheng, J. Y., de Matos, A. L., Rahman, M. M., Loog, L., et al. (2019). Parallel adaptation of rabbit populations to myxoma virus. *Science* 363, 1319–1326. doi: 10.1126/science.aau7285

Battey, C. J., Ralph, P. L., and Kern, A. D. (2020). Space is the place: Effects of continuous spatial structure on analysis of population genetic data. *Genetics* 215, 193–214. doi: 10.1534/genetics.120.303143

Beerli, P. (2004). Effect of unsampled populations on the estimation of population sizes and migration rates between sampled populations. *Mol. Ecol.* 13, 827–836. doi: 10.1111/j.1365-294X.2004.02101.x

Belkadi, A., Pedergnana, V., Cobat, A., Itan, Y., Vincent, Q. B., Abhyankar, A., et al. (2016). Whole-exome sequencing to analyze population structure, parental inbreeding, and familial linkage. *Proc. Natl. Acad. Sci. U.S.A.* 113, 6713–6718. doi: 10.1073/pnas.1606460113

Besnard, G., Bertrand, J. A. M., Delahaie, B., Bourgeois, Y. X. C., Lhuillier, E., and Thébaud, C. (2016). Valuing museum specimens: High-throughput DNA sequencing on historical collections of New Guinea crowned pigeons (Goura). *Biol. J. Linn. Soc.* 117, 71–82. doi: 10.1111/bij.12494

Bi, K., Linderoth, T., Singhal, S., Vanderpool, D., Patton, J. L., Nielsen, R., et al. (2019). Temporal genomic contrasts reveal heterogeneous evolutionary responses within and among montane chipmunk species during recent climate change. *PLoS Genet.* 15, e1008119. doi: 10.1371/journal.pgen.1008119

Bi, K., Linderoth, T., Vanderpool, D., Good, J. M., Nielsen, R., and Moritz, C. (2013). Unlocking the vault: Next-generation museum population genomics. *Mol. Ecol.* 22, 6018–6032. doi: 10.1111/mec.12516

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

Bragg, J. G., Potter, S., Bi, K., Catullo, R., Donnellan, S. C., Eldridge, M. D. B., et al. (2016). Resources for phylogenomic analyses of Australian terrestrial vertebrates. *Mol. Ecol. Resour.* 17, 869–876. doi: 10.1111/1755-0998.12633

Bragg, J. G., Potter, S., Bi, K., and Moritz, C. (2015). Exon capture phylogenomics: efficacy across scales of divergence. *Mol. Ecol. Resour.* 16, 1059–1068. doi: 10.1111/1755-0998.12449

Burrell, A. S., Disotell, T. R., and Bergey, C. M. (2015). The use of museum specimens with high-throughput DNA sequencers. *J. Hum. Evol.* 79, 35–44. doi: 10.1016/j.jhevol.2014.10.015

Craig, J. M., Kumar, S., and Hedges, S. B. (2022). Limitations of phylogenomic data can drive inferred speciation rate shifts. *Mol. Biol. Evol.* 39, 1–11. doi: 10.1093/molbev/msac038

Criscuolo, A., and Gribaldo, S. (2010). BMGE (Block Mapping and Gathering with Entropy): A new software for selection of phylogenetic informative regions from multiple sequence alignments. *BMC Evol. Biol.* 10:210. doi: 10.1186/1471-2148-10-210

Dabney, J., and Meyer, M. (2019). Extraction of highly degraded DNA from ancient bones and teeth. *Methods Mol. Biol.* 1963, 25–29. doi: 10.1007/978-1-4939-9176-1_4

Damgaard, P. B., Margaryan, A., Schroeder, H., Orlando, L., Willerslev, E., and Allentoft, M. E. (2015). Improving access to endogenous DNA in ancient bones and teeth. *Sci. Rep.* 5, 1–12. doi: 10.1038/srep11184

Derkarabetian, S., Benavides, L. R., and Giribet, G. (2019). Sequence capture phylogenomics of historical ethanol-preserved museum specimens: Unlocking the rest of the vault. *Mol. Ecol. Resour.* 19, 1531–1544. doi: 10.1111/1755-0998.13072

Durand, E. Y., Patterson, N., Reich, D., and Slatkin, M. (2011). Testing for ancient admixture between closely related populations. *Mol. Biol. Evol.* 28, 2239–2252. doi: 10.1093/molbev/msr048

Dussex, N., van der Valk, T., Morales, H. E., Wheat, C. W., Díez-del-Molino, D., von Seth, J., et al. (2021). Population genomics of the critically endangered kākāpo. *Cell Genomics* 1:100002. doi: 10.1016/j.xgen.2021.100002

Ewart, K. M., Johnson, R. N., Ogden, R., Joseph, L., Frankham, G. J., and Lo, N. (2019). Museum specimens provide reliable SNP data for population genomic analysis of a widely distributed but threatened cockatoo species. *Mol. Ecol. Resour.* 19, 1578–1592. doi: 10.1111/1755-0998.13082

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

Fraïsse, C., Popovic, I., Mazoyer, C., Spataro, B., Delmotte, S., Romiguier, J., et al. (2021). DILS: Demographic inferences with linked selection by using ABC. *Mol. Ecol. Resour.* 21, 2629–2644. doi: 10.1111/1755-0998.13323

Fumagalli, M. (2013). Assessing the effect of sequencing depth and sample size in population genetics inferences. *PLoS ONE* 8, 14–17. doi: 10.1371/journal.pone.0079667

Gauthier, J., Pajkovic, M., Neuenschwander, S., Kaila, L., Schmid, S., Orlando, L., et al. (2020). Museomics identifies genetic erosion in two butterfly species across the 20th century in Finland. *Mol. Ecol. Resour.* 20, 1191–1205. doi: 10.1111/1755-0998.13167

Grewe, F., Kronforst, M. R., Pierce, N. E., and Moreau, C. S. (2021). Museum genomics reveals the Xerces blue butterfly (Glaucopsyche xerces) was a distinct species driven to extinction. *Biol. Lett.* 17:e0123. doi: 10.1098/rsbl.2021.0123

Guschanski, K., Krause, J., Sawyer, S., Valente, L. M., Bailey, S., Finstermeier, K., et al. (2013). Next-generation museomics disentangles one of the largest primate radiations. *Syst. Biol.* 62, 539–554. doi: 10.1093/sysbio/syt018

Gutenkunst, R. N., Hernandez, R. D., Williamson, S. H., and Bustamante, C. D. (2009). Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS Genet.* 5, e100695. doi: 10.1371/journal.pgen.1000695

Hey, J., Chung, Y., Sethuraman, A., Lachance, J., Tishkoff, S., Sousa, V. C., et al. (2018). Phylogeny estimation by integration over isolation with migration models. *Mol. Biol. Evol.* 35, 2805–2818. doi: 10.1093/molbev/msy162

Höhna, S. (2014). Likelihood inference of non-constant diversification rates with incomplete taxon sampling. *PLoS ONE* 9, e84184. doi: 10.1371/journal.pone.0084184

Höhna, S., Stadler, T., Ronquist, F., and Britton, T. (2011). Inferring speciation and extinction rates under different sampling schemes. *Mol. Biol. Evol.* 28, 2577–2589. doi: 10.1093/molbev/msr095

Hung, C. M., Shaner, P. J. L., Zink, R. M., Liu, W. C., Chu, T. C., Huang, W. S., et al. (2014). Drastic population fluctuations explain the rapid extinction of the passenger pigeon. *Proc. Natl. Acad. Sci. U.S.A.* 111, 10636–10641. doi: 10.1073/pnas.1401526111

Hykin, S. M., Bi, K., and McGuire, J. A. (2015). Fixing formalin: A method to recover genomic-scale DNA sequence data from formalin-fixed museum specimens using high-throughput sequencing. *PLoS ONE* 10, e141579. doi: 10.1371/journal.pone.0141579

Irestedt, M., Ericson, P. G. P., Johansson, U. S., Oliver, P., Joseph, L., and Blom, M. P. K. (2019). No signs of genetic erosion in a 19th century genome of the extinct Paradise Parrot (Psephotellus pulcherrimus). *Diversity* 11:40058. doi: 10.3390/d11040058

Ivan, J., Moritz, C., Potter, S., Bragg, J., Turakulov, R., and Hua, X. (2022). Temperature predicts the rate of molecular evolution in Australian Eugongylinae skinks. *Evolution* 76, 252–261. doi: 10.1111/evo.14342

Jónsson, H., Ginolhac, A., Schubert, M., Johnson, P. L. F., and Orlando, L. (2013). MapDamage2.0: Fast approximate Bayesian estimates of ancient DNA damage parameters. *Bioinformatics* 29, 1682–1684. doi: 10.1093/bioinformatics/btt193

Kircher, M., Sawyer, S., and Meyer, M. (2012). Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform. *Nucleic Acids Res.* 40, 1–8. doi: 10.1093/nar/gkr771

Linck, E., and Battey, C. J. (2019). Minor allele frequency thresholds strongly affect population structure inference with genomic data sets. *Mol. Ecol. Resour.* 19, 639–647. doi: 10.1111/1755-0998.12995

Linck, E., Epperly, K., Van Els, P., Spellman, G. M., Bryson, R. W., McCormack, J. E., et al. (2019). Dense geographic and genomic sampling reveals paraphyly and a cryptic lineage in a classic sibling species complex. *Syst. Biol.* 68, 956–966. doi: 10.1093/sysbio/syz027

Linder, H. P., Hardy, C. R., and Rutschmann, F. (2005). Taxon sampling effects in molecular clock dating: An example from the African Restionaceae. *Mol. Phylogenet. Evol.* 35, 569–582. doi: 10.1016/j.ympev.2004.12.006

Lyra, M. L., Lourenço, A. C. C., Pinheiro, P. D. P., Pezzuti, T. L., Baêta, D., Barlow, A., et al. (2020). High-throughput DNA sequencing of museum specimens sheds light on the long-missing species of the Bokermannohyla claresignata group (Anura: Hylidae: Cophomantini). *Zool. J. Linn. Soc.* 190, 1235–1255. doi: 10.1093/zoolinnean/zlaa033

Maliet, O., Hartig, F., and Morlon, H. (2019). A model with many small shifts for estimating species-specific diversification rates. *Nat. Ecol. Evol.* 3, 1086–1092. doi: 10.1038/s41559-019-0908-0

Mason, V. C., Li, G., Helgen, K. M., and Murphy, W. J. (2011). Efficient cross-species capture hybridization and next-generation sequencing of mitochondrial genomes from noninvasively sampled museum specimens. *Genome Res.* 21, 1695–1704. doi: 10.1101/gr.120196.111

McCormack, J. E., Tsai, W. L. E., and Faircloth, B. C. (2016). Sequence capture of ultraconserved elements from bird museum specimens. *Mol. Ecol. Resour.* 16, 1189–1203. doi: 10.1111/1755-0998.12466

McDonough, M. M., Parker, L. D., McInerney, N. R., Campana, M. G., and Maldonado, J. E. (2018). Performance of commonly requested destructive museum samples for mammalian genomic studies. *J. Mammal.* 99, 789–802. doi: 10.1093/jmammal/gyy080

McGuire, J. A., Cotoras, D. D., O'Connell, B., Lawalata, S. Z. S., Wang-Claypool, C. Y., Stubbs, A., et al. (2018). Squeezing water from a stone: High- throughput sequencing from a 145-year old holotype resolves (barely) a cryptic species problem in flying lizards. *PeerJ* 2018, 1–16. doi: 10.7717/peerj.4470

Meyer, M., and Kircher, M. (2010). Illumina sequencing library preparation for highly multiplexed target capture and sequencing. *Cold Spring Harb. Protoc.* 2010, pdb.prot5448. doi: 10.1101/pdb.prot5448

Mondol, S., Bruford, M. W., and Ramakrishnan, U. (2013). Demographic loss, genetic structure and the conservation implications for indian tigers. *Proc. R. Soc. B Biol. Sci.* 280:e0496. doi: 10.1098/rspb.2013.0496

Nakahama, N. (2021). Museum specimens: An overlooked and valuable material for conservation genetics. *Ecol. Res.* 36, 13–23. doi: 10.1111/1440-1703.12181

Nakahama, N., and Isagi, Y. (2018). Recent transitions in genetic diversity and structure in the endangered semi-natural grassland butterfly, Melitaea protomedia, in Japan. *Insect Conserv. Divers.* 11, 330–340. doi: 10.1111/icad.12280

Nguyen, L. T., Schmidt, H. A., Von Haeseler, A., and Minh, B. Q. (2015). IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* 32, 268–274. doi: 10.1093/molbev/msu300

Pääbo, S., Poinar, H., Serre, D., Jaenicke-Després, V., Hebler, J., Rohland, N., et al. (2004). Genetic analyses from ancient DNA. *Annu. Rev. Genet.* 38, 645–679. doi: 10.1146/annurev.genet.37.110801.143214

Perez, M. F., Franco, F. F., Bombonato, J. R., Bonatelli, I. A. S., Khan, G., Romeiro-Brito, M., et al. (2018). Assessing population structure in the face of isolation by distance: Are we neglecting the problem? *Divers. Distrib.* 24, 1883–1889. doi: 10.1111/ddi.12816

Peter, B. M., and Slatkin, M. (2013). Detecting range expansions from genetic data. *Evolution* 67, 3274–3289. doi: 10.1111/evo.12202

Peter, B. M., and Slatkin, M. (2015). The effective founder effect in a spatially expanding population. *Evolution* 69, 721–734. doi: 10.1111/evo.12609

Pfeifer, B., Wittelsbürger, U., Ramos-Onsins, S. E., and Lercher, M. J. (2014). PopGenome: An efficient swiss army knife for population genomic analyses in R. *Mol. Biol. Evol.* 31, 1929–1936. doi: 10.1093/molbev/msu136

Pickrell, J. K., and Pritchard, J. K. (2012). Inference of population splits and mixtures from genome-wide allele frequency data. *PLoS Genet.* 8, e1002967. doi: 10.1371/journal.pgen.1002967

Potter, S., Bragg, J. G., Blom, M. P., Deakin, J. E., Kirkpatrick, M., Eldridge, M. D., et al. (2017). Chromosomal speciation in the genomics era: disentangling phylogenetic evolution of rock-wallabies. *Front. Genet.* 8, e00010. doi: 10.3389/fgene.2017.00010

Potter, S., Bragg, J. G., Peter, B. M., Bi, K., and Moritz, C. (2016). Phylogenomics at the tips: inferring lineages and their demographic history in a tropical lizard, Carlia amax. *Mol. Ecol.* 25, 1367–1380. doi: 10.1111/mec.13546

Potter, S., Bragg, J. G., Turakulov, R., Eldridge, M. D. B., Deakin, J., Kirkpatrick, M., et al. (2022). Limited Introgression between Rock-Wallabies with Extensive Chromosomal Rearrangements. *Mol. Biol. Evol.* 39:msab333. doi: 10.1093/molbev/msab333

Potter, S., Close, R. L., Taggart, D. A., Cooper, S. J. B., and Eldridge, M. D. B. (2014). Taxonomy of rock-wallabies, Petrogale (Marsupialia: Macropodidae). IV. Multifaceted study of the brachyotis group identifies additional taxa. *Aust. J. Zool.* 62, 401–414. doi: 10.1071/ZO13095

Potter, S., Eldridge, M. D. B., Taggart, D. A., and Cooper, S. J. B. (2012). Multiple biogeographical barriers identified across the monsoon tropics of northern Australia: phylogeographic analysis of the brachyotis group of rock-wallabies. *Mole. Ecol.* 21, 2254–2269. doi: 10.1111/j.1365-294X.2012.05523.x

Potter, S., Moritz, C., and Eldridge, M. D. B. (2015). Gene flow despite complex Robertsonian fusions among rock-wallaby (Petrogale) species. *Biol. Lett.* 11, 20150731. doi: 10.1098/rsbl.2015.0731

Potter, S., Xue, A. T., Bragg, J. G., Rosauer, D. F., Roycroft, E. J., and Craig, M. (2018). Pleistocene climatic changes drive diversification across a tropical savanna. *Mol. Ecol.* 27, 520–532. doi: 10.1111/mec.14441

Pritchard, J. K., Stephens, M., and Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics* 155, 945–959. doi: 10.1093/genetics/155.2.945

Pyron, R. A., Beamer, D. A., Holzheuser, C. R., Lemmon, E. M., Lemmon, A. R., Wynn, A. H., et al. (2022). Contextualizing enigmatic extinctions using genomic DNA from fluid-preserved museum specimens of Desmognathus salamanders. *Conserv. Genet.* 23, 375–386. doi: 10.1007/s10592-021-01424-4

Rowe, K. C., Singhal, S., Macmanes, M. D., Ayroles, J. F., Morelli, T. L., Rubidge, E. M., et al. (2011). Museum genomics: Low-cost and high-accuracy genetic data from historical specimens. *Mol. Ecol. Resour.* 11, 1082–1092. doi: 10.1111/j.1755-0998.2011.03052.x

Roycroft, E., Achmadi, A., Callahan, C. M., Esselstyn, J. A., Good, J. M., Moussalli, A., et al. (2021a). Molecular evolution of ecological specialisation: genomic insights from the diversification of murine rodents. *Genome Biol. Evol.* 13, 1–16. doi: 10.1093/gbe/evab103

Roycroft, E., MacDonald, A. J., Moritz, C., Moussalli, A., Miguez, R. P., and Rowe, K. C. (2021b). Museum genomics reveals the rapid decline and extinction of Australian rodents since European settlement. *Proc. Natl. Acad. Sci. U.S.A.* 118:e2021390118. doi: 10.1073/pnas.2021390118

Roycroft, E. J., Moussalli, A., and Rowe, K. C. (2020). Phylogenomics uncovers confidence and conflict in the rapid radiation of australo-papuan rodents. *Syst. Biol.* 69, 431–444. doi: 10.1093/sysbio/syz044

Ruane, S., and Austin, C. C. (2017). Phylogenomics using formalin-fixed and 100+ year-old intractable natural history specimens. *Mol. Ecol. Resour.* 17, 1003–1008. doi: 10.1111/1755-0998.12655

Sarver, B., Keeble, S., Cosart, T., Tucker, P., Dean, M., and Good, J. (2017). Phylogenomic insights into mouse evolution using a pseudoreference approach. *Genome Biol. Evol.* 9, 726–739. doi: 10.1093/gbe/evx034

Sawyer, S., Krause, J., Guschanski, K., Savolainen, V., and Pääbo, S. (2012). Temporal patterns of nucleotide misincorporations and DNA fragmentation in ancient DNA. *PLoS ONE* 7, e0034131. doi: 10.1371/journal.pone.0034131

Schmitt, C. J., Cook, J. A., Zamudio, K. R., and Edwards, S. V. (2019). Museum specimens of terrestrial vertebrates are sensitive indicators of environmental change in the Anthropocene. *Philos. Trans. R. Soc. B Biol. Sci.* 374:387. doi: 10.1098/rstb.2017.0387

Shafer, A. B. A., Peart, C. R., Tusso, S., Maayan, I., Brelsford, A., Wheat, C. W., et al. (2017). Bioinformatic processing of RAD-seq data dramatically impacts downstream population genetic inference. *Methods Ecol. Evol.* 8, 907–917. doi: 10.1111/2041-210X.12700

Singhal, S., Grundler, M., Colli, G., and Rabosky, D. L. (2017). Squamate Conserved Loci (SqCL): A unified set of conserved loci for phylogenomics and population genetics of squamate reptiles. *Mol. Ecol. Resour.* 17, e12–e24. doi: 10.1111/1755-0998.12681

Slatkin, M. (2005). Seeing ghosts: The effect of unsampled populations on migration rates estimated for sampled populations. *Mol. Ecol.* 14, 67–73. doi: 10.1111/j.1365-294X.2004.02393.x

Sousa, V., and Hey, J. (2013). Understanding the origin of species with genome-scale data: modelling gene flow. *Nat. Rev. Genet.* 14, 404–414. doi: 10.1038/nrg3446

Streicher, J. W., Schulte, J. A., and Wiens, J. J. (2016). How should genes and taxa be sampled for phylogenomic analyses with missing data? An Empirical Study in Iguanian Lizards. *Syst. Biol.* 65:58. doi: 10.1093/sysbio/syv058

Tajima, F. (1989). Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* 123, 585–595. doi: 10.1093/genetics/123.3.585

Tsai, W. L. E., Schedl, M. E., Maley, J. M., and McCormack, J. E. (2020). More than skin and bones: Comparing extraction methods and alternative sources of DNA from avian museum specimens. *Mol. Ecol. Resour.* 20, 1220–1227. doi: 10.1111/1755-0998.13077

van der Valk, T., Díez-del-Molino, D., Marques-Bonet, T., Guschanski, K., and Dalén, L. (2019). Historical genomes reveal the genomic consequences of recent population decline in eastern gorillas. *Curr. Biol.* 29, 165–170.e6. doi: 10.1016/j.cub.2018.11.055

Watterson, G. A. (1975). On the number of segregating sites in genetical models without recombination. *Theor. Popul. Biol.* 7, 256–276. doi: 10.1016/0040-5809(75)90020-9

Wood, H. M., González, V. L., Lloyd, M., Coddington, J., and Scharff, N. (2018). Next-generation museum genomics: Phylogenetic relationships among palpimanoid spiders using sequence capture techniques (Araneae: Palpimanoidea). *Mol. Phylogenet. Evol.* 127, 907–918. doi: 10.1016/j.ympev.2018.06.038

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers | Frontiers in Ecology and Evolution

TYPE Original Research
PUBLISHED 28 July 2022
DOI 10.3389/fevo.2022.910250

Insights into phylogenetic divergence of *Dalbergia* (Leguminosae: Dalbergiae) from Mexico and Central America

Solange Sotuyo[1]*†, Euler Pedraza-Ortega[1]†, Esteban Martínez-Salas[1], José Linares[2] and Lidia Cabrera[1]

[1]Departamento de Botánica, Instituto de Biología, Universidad Nacional Autónoma de México, Mexico City, Mexico, [2]Centro Universitario Regional del Litoral Atlántico (CURLA), La Ceiba, Honduras

Check for updates

OPEN ACCESS

EDITED BY
Jonathan J. Fong,
Lingnan University, China

REVIEWED BY
Harald Schneider,
Xishuangbanna Tropical Botanical Garden (CAS), China
Maria M. Romeiras,
University of Lisbon, Portugal

*CORRESPONDENCE
Solange Sotuyo
jssotuyo@ib.unam.mx

†These authors have contributed equally to this work and share first authorship

SPECIALTY SECTION
This article was submitted to Phylogenetics, Phylogenomics, and Systematics, a section of the journal Frontiers in Ecology and Evolution

RECEIVED 01 April 2022
ACCEPTED 28 June 2022
PUBLISHED 28 July 2022

CITATION
Sotuyo S, Pedraza-Ortega E, Martínez-Salas E, Linares J and Cabrera L (2022) Insights into phylogenetic divergence of *Dalbergia* (Leguminosae: Dalbergiae) from Mexico and Central America. *Front. Ecol. Evol.* 10:910250. doi: 10.3389/fevo.2022.910250

The pantropical genus *Dalbergia* includes more than 250 species. Phylogenetic studies of the group are scarce and have only included two or three species distributed in Mexico. We obtained herbarium samples of Mexican, Central American, and South American species (sourced from MEXU). In addition, sequences of GenBank accessions were used to complement the study. Using internal transcribed spacer (*ITS*), the *matK* and *rbcL* sequences from 384 accessions comprising species from America, Asia, and Africa were sampled to evaluate phylogenetic relationships of Mexican species and infrageneric classifications based on morphological data. Phylogenetic analyses suggest that the genus *Dalbergia* is monophyletic and originated in South America. The species distributed in Mexico are not a monophyletic clade but are divided into four clades with affinities to South American and Asian species clades. There is no correlation between geography and large-scale phylogeny. The estimated ages of the Mexican and Central American clades ranged from 11.32 Ma (*Dalbergia granadillo* clade) to 1.88 Ma (*Dalbergia ecastaphyllum* clade). Multiple long-distance dispersal events should be used to explain the current genus distribution.

KEYWORDS

barcode, *Dalbergia*, diversification, Miocene, Mexico

Introduction

The subfamily Papilionoideae includes an important clade, the Dalbergiodeae group. The "Dalbergioides" represent a monophyletic group comprising all genera referred to as the tribes Adesmieae and Aeschynomeneae, the subtribe Bryinae of Desmodieae and Dalbergieae except the genera *Andira*, *Hymenolobium*, *Vatairea*, and *Vataireopsis* (Lavin et al., 2001). This group consists of the subclades *Adesmia*, *Dalbergia*, and *Pterocarpus*, supported and identified mainly on a molecular data basis (chloroplast sequences; the trnK/matK spacer and the trnL intron, Lavin et al., 2001).

The *Dalbergia* genus is a pantropical group with around 250 species and centers of diversity in Central and South America, Africa, Madagascar, and Asia (Klitgård and Lavin, 2005). In Mexico, *Dalbergi*a comprises 20 species, of which six are endemic (Sousa et al., 2001; Linares and Sousa, 2007; Ricker et al., 2013). *Dalbergia,* or rosewoods as are generally known, distinguishes because their heartwood is considered of high economic value, owing to its beauty, durability, and excellent physical, mechanical, and acoustic properties (Pittier, 1922). They also produce metabolites, used as antimicrobial (Rutiaga-Quiñones et al., 2010), antifungal (Rutiaga-Quiñones et al., 1995; Barragán-Huerta et al., 2004), antibiotic, antioxidants, and cytotoxic agents (Hamburger et al., 1987; Lianhe et al., 2011; Pérez-Gutiérrez and García-Baez, 2013). In addition, it has been reported that *Dalbergia* species establish symbiotic relationships with rhizobia for nitrogen fixation. This plays an important role in ecosystems since it improves soil fertility (Rasolomampianina et al., 2005). The species populations are at risk because of its intensive use, habitat loss, and fragmentation as well as slow recruitment rate and growth. Several species of the genus are used as timber species, and they are intensively exploited and subject to international traffic. Conservation of timber species threatened by illegal logging and deforestation is essential. Barcodes of the species could help to monitor species of *Dalbergia* subject to international traffic and reconstruct a phylogenetic hypothesis of the genus.

Phylogenetic analyses of the tribe Dalbergieae are based on molecular and morphological data (Lavin et al., 2001), placing *Dalbergia* in the Dalbergia clade as sister to the genus *Machaerium* Pers. and *Aeschynomene* L. subgen. *Ochopodium.* Later on, Ribeiro et al. (2007) concluded that *Aeschynomene* subgen. Ochopodium Vogel is more closely related than *Machaerium* to *Dalbergia*. Vatanparast et al. (2013) made other attempts to resolve the generic relationships between *Aeschynomene* and other segregate genera (*Bryaspis*, *Geissaspis* Wight & Arn. and *Kotschya* Endl.), but they are weakly resolved still. Cardoso et al. (2020), studying the phylogenetic relationships of *Aeschynomene* subgen. Ochopodium, find that both, A*eschynomene* and *Machaerium,* are sister taxa of *Dalbergia*. Ochopodium section was newly circumscribed as *Ctenodon*, and the genus is particularly diverse in Mexico and Brazil and has a few endemic species in the Andes.

In Mexico, 20 species have been described, 15 of which are potentially threatened by illegal logging (Linares and Sousa, 2007). Due to the characteristics of its wood, they are over-exploited, placing them in danger of extinction (NOM-059 SEMARNAT, 2010). According to the red list of the IUCN (International Union for Conservation of Nature), the species of the most concern are *Dalbergia granadillo* and *D. retusa*, as their natural populations are decreasing considerably and are therefore considered Critically Endangered (IUCN). Mexican species inhabit the west and center of the country and the Yucatan peninsula. However, most species grown in the southeast are associated with tropical forests, cloud forests, tropical deciduous and sub-deciduous forests, and pine-oak forests (Standley, 1922) (**Table 1**). Only three Mexican species were used in previous phylogenetic studies.

Dalbergia species are morphologically variable and possess a wide range of habitat preferences which made it difficult to classify the New and Old World species into natural groups (Bentham, 1860; Prain, 1904; Carvalho, 1989). It is necessary to use several specimens for each species in a broadscale distribution to get a more clear idea of which are the taxonomic circumscription within *Dalbergia* species. We employed a relatively wide taxonomic sampling using several *Dalbergia* accessions from species occurring in Mexico and included species from its centers of diversity in America, Africa, and Asia. The objectives of this study were to (1) provide a phylogenetic framework for Mexican *Dalbergia* species, (2) test up barcode molecular markers in Mexican species, and (3) provide an age of divergence for the Mexican species.

Materials and methods

Taxa sampling and deoxyribonucleic acid sequencing

To obtain an in-depth view of the phylogenetic relationships within the genus, we increased the previous sampling by the addition of Mexican, Central America, and South American species of *Dalbergia*. We included a total of 287 *Dalbergia* accessions. Outgroup selection was based on previous phylogenetic studies ensuring that accession sequences from *Ctenodon*, *Machaerium,* and *Pictetia* close relative genera were represented (**Supplementary Table 1**). A summary of accessions used for species from Mexico, Central America, and the Caribbean is listed in **Table 2**.

The sample tissue material for DNA extraction was obtained from specimens in the MEXU herbarium. Total genomic DNA was extracted from leaves, flowers, or fruit samples using a modified DNeasy Plant Mini Kit (Qiagen). The target DNA regions, *rbcL* and *matK,* were amplified with universal barcoding primers (CBOL Plant Working Group, 2009). In the case of internal transcribed spacer (ITS), AB101 and AB102 primers (Sun et al., 1994) were used. PCR amplification of *rbcL*, *matK,* and *ITS* was carried out on a Gene Amp 2700 (Applied Biosystems, United States) with a Thermo PCR Master Mix kit (Thermo Fisher), using the manufacturer's instructions. PCR conditions for *matK* and *rbcL* were as follows: 2 min initial denaturation at 94°C, 35 cycles (94°C 1 min, 52°C 1 min, and 72°C 1 min), and 10 min of final extension at 72°C. PCR conditions for *ITS* were as follows: 2 min initial denaturation at 94°C, 35 cycles (94°C 1 min, 53°C 1 min, and 72°C 1 min), and 7 min of final extension at 72°C. Amplified PCR products were checked on 1% agarose gel

TABLE 1 Ecological and morphological information of *Dalbergia* species distributed in Mexico, Central America, and the Caribbean.

Species	IUCN	Habit	Habitat	Altitude (m)	Leaflet number	Flower size (mm)	Ovary indumentum	Fruit shape	Fruit texture	Fruit dispersion
D. agudeloi	NT	tree	oak forest, seasonal dry forest	750-200	(11—)13 (—15)	4-4.5	villose	unknown	unknown	anemocory?
D. brownei	LC	Scandent shrub or liana	coastal scrub, mangroves, flooded forests	0-20	1	7-11	grabrous	oblong-lunate	woody	Hydrocoric
D. calderonii	CR	tree	tropical deciduous forest, medium deciduous forests	400-1200	5-6	4-5	velutine	oblong	woody	Anemocory
D. calycina	VU	tree	Quercus forest, cloud forest	(800—) 1000-1900	(5-)9(-11)	17-20	glabrous?	oblong	chartaceous	Anemocory
D. chontalensis	VU	shrub	seasonal dry forest, riparian vegetation, coastal vegetation.	0-1000	11-15	10-12	glabrous?	elliptic	subchartaceous	anemocory
D. congestiflora	EN	tree	tropical deciduous forest	0-600	7-13	3-4	pubescent	oblong	papyraceous	Anemocory
D. cubilquitzensis	LC	tree	tropical evergreen forest	0-900	(11—) 13-15	5-6	pubescent	elliptic-oblong	papyraceous	Anemocory
D. ecastaphyllum	LC	Scandent shrub or liana	coastal scrub, mangroves, flooded forests	0-20	1	8-9	glabrous?	suborbicular	woody	hydrocoric
D. glabra	LC	Scandent shrub or liana	seasonal dry forest, riparian vegetation	0-800	(7—)9	7-11.5	glabrous/pubescent	elliptic to oblong	chartaceous	hydrocoric?
D. glomerata	CR	tree	tropical evergreen forest, tropical oak forest	0-900	(5—) 9-11 (—12)	4.7-5.5	glabrous?	elliptic-oblong	chartaceous	anemocory
D. granadillo	CR	tree	tropical deciduous forest, oak forest, rain forest	0-100	(13—) 11 (—15)	20	?	elliptic-oblong	chartaceous	anemocory
D. longepedunculata	EN	tree	deciduous forest, evergreen forest	600-1100	7 (8—)	6	pubescent adaxially	oblong	chartaceous	anemocory
D. luteola	CR	tree	seasonal dry forest, riparian vegetation in oak forest	800-1100	11-13	3-3.6	glabrous?	unknown	unknown	anemocory
D. melanocardium	EN	tree	montane rain forest	1300-1600	7-11 (—13)	5-6	villose	oblong	woody	anemocory
D. monetaria	LC	Scandent shrub or liana	humid forests, mangroves	0-30	3–5 (–6)	5–6	glabrous?	orbicular	woody	hydrocoric
D. palo-escrito	EN	tree	cloud forest	1000-2000	9-13	3-5.5	puberulus	oblong	papyraceous	anemocory
D. retusa	CR	tree	dry seasonal forest, rain forest?, gallery forest	20–1000	7–15 (—17)	(8–) 15–18 (–20)	glabrous?	elliptic to oblong	woody	anemocory
D. stevensonii	CR	tree	low deciduous forest	0-200	5	5-6	villose	oblong	woody	anemocory
D. tabascana	?	Scandent shrub or liana	swamps, mangroves, lagoons, savannahs and coastal vegetation	0-100	5-7	10-11.5	glabrous?	oblong-lunate	woody	hydrocoric
D. tucurensis	EN	tree	cloud forests, pine and pine-oak forest	1400-2500	(11—) 13-15	4.5-6	densely villose	oblong	papyraceous	anemocory
D. tilarana	EN	tree	pine-oak forest, medium forests	600-1450	5-9	4-11	densely strigose	elliptic to oblong	woody	anemocory

TABLE 2 The phylogenetic clades recognized in the present study for Mexican species of *Dalbergia*.

Clade	Species	N° accessions sampled
Dalbergia ecastaphyllum	*D. monetaria*	0
	D. ecastaphyllum	3
Dalbergia glabra	*D. brownei*	2
	D. chontalensis	2
	D. glabra	4
	D. tabascana	2
Dalbergia glomerata	*D. agudeloi*	2
	D. calderoni	2
	D. congestiflora	1
	D. cubilquitzensis	7
	D. glomerata	2
	D. longepedunculata	2
	D. luteola	3
	D. melanocardium	4
	D. palo-escrito	1
	D. stevensonii	2
	D. tucurensis	2
Dalbergia granadillo	*D. calycina*	3
	D. granadillo	3
	D. retusa	5

electrophoresis. Both strands of the clean PCR products were directly sequenced using BigDye Terminator v.3.1 (Thermo Fisher, Foster City, CA, United States) cycle sequencing kit and visualized on an ABI 3730 (Applied Biosystems) at Laboratorio de Secuenciación Genomica de la Biodiversidad y la Salud, Instituto de Biología, using the same primers as for amplification.

Distribution maps

We constructed distribution maps with collection information accessed from Global Biodiversity Information Facility (GBIF.org, 2022).[1] We downloaded 11,941 herbaria records for Mexican, Central American, and Caribbean species of *Dalbergia* sampled in the molecular phylogeny. Data cleaning involved, first, standardizing data, deleting duplicate specimens, deleting records without any geographical coordinates, and any georeference erroneously georeferenced. After that, we used the R package "CoordinateCleaner" (Zizka et al., 2019) for further cleaning about coordinates at sea, country and province centroids, country capitals, urban areas, and around biodiversity institutions, which often come from cultivated individuals or with incorrect data. From the records downloaded, 5840 records were georeferenced, and after filtering and cleaning, 4,014 records were suitable to be used to generate distribution maps by species and phylogenetic clade.

Data analysis

Sequences were edited and assembled using SeqTrace software (Stucky, 2012). All sequences generated in this study were deposited in GenBank (**Supplementary Table 1**). Edited sequences for each gene region were aligned separately with MAFFT (Katoh et al., 2009). After an initial alignment, the alignments were manually adjusted using AliView (Larsson, 2014) if needed, following the principles described in Kelchner and Clark (1997). In addition, we compiled all *ITS, matK,* and *rbcL* sequences publicly available in GenBank for *Dalbergia* and added our newly generated three loci sequences to that dataset to produce a phylogenetic tree with a denser sampling across *Dalbergia*. Sequences generated from the same voucher from at least two loci have been used in the combined dataset to reduce the missing data. The combined dataset has 194 accessions and 336 accessions for the unique ITS dataset. A total of 384 accessions were analyzed.

Phylogenetic reconstruction of all the taxa sampled was undertaken using Bayesian inference (BI). We used three datasets: (1) the individual ITS dataset (unique ITS), (2) the plastid data set, and (3) the concatenated dataset. A Bayesian analysis without a molecular clock for the concatenated matrix was inferred with MrBayes. Gene trees for calibration were inferred with BEAST2 (Bouckaert et al., 2019). The GTR + Γ was selected as the best fit model based on the Akaike information criterion (Akaike, 1974) using the software jModelTest 2 (Darriba et al., 2012). The combined analysis for the three markers was run in 20 × 106 generations, sampling every 1,000. For the ITS dataset, 40 × 106 generations were run. Trees were sampled for 1,000 generations, and 20% of them were discarded as burn-in. The convergence of MCMC chain trees was visualized with the Bestiary software (Wirth and Duchene, 2021). Calibrated time trees were estimated using BEAST2 (Bouckaert et al., 2019) with a Yule tree prior model, lognormal relaxed molecular clock, and the node *Machaerium-Dalbergia* data according to Lavin et al. (2005). The trees were visualized with ggtree for R (Yu, 2020). Alignments in FASTA format can be seen in the **Supplementary Material** (S2).

Results

Phylogenetic relationships, combined tree

Phylogenetic trees show that *Dalbergia* is monophyletic (1.0 PP) with a basal clade formed by South American Neotropical species (*Dalbergia miscolobium, Dalbergia spruceana,* and *Dalbergia villosa,* sect Dalbergia sensu Carvalho, 1997) resolved sister to the remaining species (**Figures 1, 2**). The second clade of Asian species, containing two subclades, is then sister to the remaining species. Subclade II-A of Mexican climbing or woody vine species part of Ecastaphyllum sensu Carvalho

[1] www.gbif.org

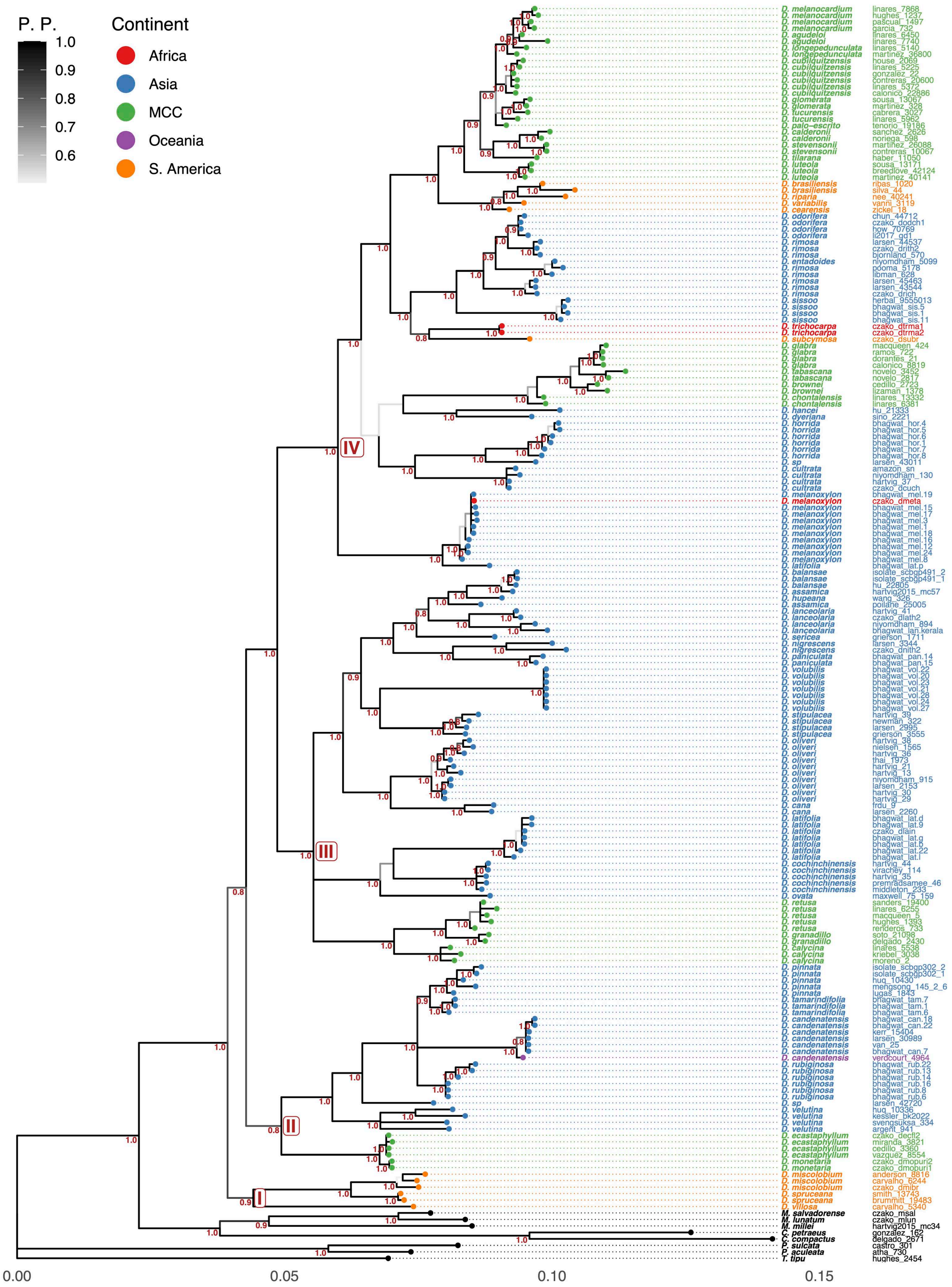


FIGURE 1

Bayesian combined phylogram of *Dalbergia*. Under the branches, posterior probabilities (pp) are in red font.

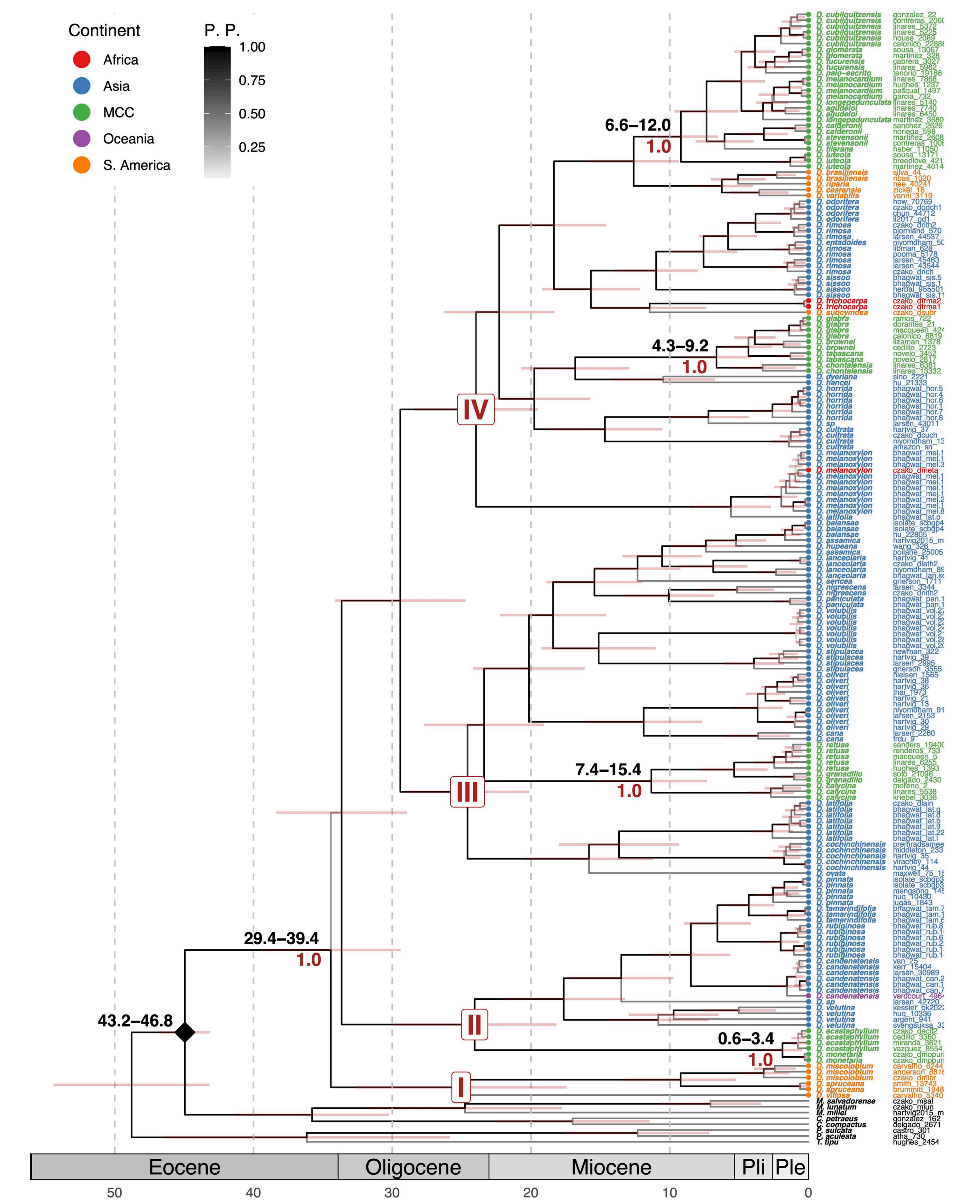


FIGURE 2

Combined Bayesian calibrated tree of *Dalbergia*. Above the branches, the estimated age range of the clades is in black font. Local posterior probabilities are shown under branches in red font. Shading bottom bars represent geological epochs. The diamond mark represents the calibration node. The bars on the branches represent the range of the 95% confidence interval in the Bayesian tree.

(1997) (*Dalbergia ecastaphyllum, Dalbergia monetaria*) is sister to subclade II-B of woody vine Asian species: *Dalbergia velutina, Dalbergia pinnata* (synonym of *Dalbergia tamarindifolia*), and *Dalbergia rubiginosa* (series Polyphyllae, Rubiginosae, Sericeae, and Velutina sensu Prain, 1904).

Clade III contains three subclades (III-A, III-B, and III-C). Subclade III-A contains Asian species *Dalbergia ovata* (Ovatae), *Dalbergia cochinchinensis,* and *Dalbergia latifolia* (serie Latifoliae), both parts of section Miscolobium. Subclade III-B contains Mexican and Central American species: *Dalbergia calycina, D. granadillo, D. retusa,* and *Dalbergia cuscatlanica.* Subclade III-C contains Asian species, and the species are part of subgenus Amerimmnon sensu Prain (1904) section Dalbergaria; *Dalbergia cana* (Canae) is the sister species of *Dalbergia oliveri* (Lanceolarieae); then, they are the sister group of *Dalbergia stipulacea* (Stipulaceae)-*Dalbergia volubilis* (Volubilis), and the group formed by seven species part of Lanceolarieae (*Dalbergia lanceolaria* subsp. *lanceolaria, Dalbergia balansae, Dalbergia huapeana, Dalbergia assamica, Dalbergia nigrescens, and Dalbergia paniculata)* and Sericeae (*Dalbergia sericea*).

Clade IV is formed by five subclades (IV-A, IV-B, IV-C, IV-D, and IV-E). Subclade IV-A is formed by the tree species of *Dalbergia latifolia* (Latifoliae) and *Dalbergia melanoxylon* (Phyllanthoides). Subclade IV-B is formed by *Dalbergia dyeriana, Dalbergia hancei* (Foliaceae), *Dalbergia cultrata* (Cultratae), and *Dalbergia horrida*. Subclade IV-C is formed with four climbing Mexican–Central American species (*Dalbergia brownei, Dalbergia chontalensis, Dalbergia glabra,* and *Dalbergia tabascana*). Subclade IV-D includes a divergent climbing species *Dalbergia subcymosa* (Ecastaphyllum sensu Carvalho, 1997) from South America as a sister to the following Asian species: *Dalbergia trichocarpa* (Madagascar-African tree, unknown sect.), the tree *Dalbergia sissoo* (serie Sisso), the woody climbers *Dalbergia rimosa* (serie Rimosae), and *Dalbergia entadoides* (unknown sect.-serie), and the tree *Dalbergia odorifera* (unknown sect.-serie). Subclade IV-E resolves a group of South American tree species of Triptolemea sensu Carvalho, 1997 (*Dalbergia variabilis = Dalbergia frutescens, Dalbergia cearensis, Dalbergia riparia,* and *Dalbergia brasiliensis*) as the sister group of 10 Mexican–Central American species of trees (*Dalbergia agudeloi, Dalbergia melanocardium, Dalbergia palo-escrito, Dalbergia tucurensis, Dalbergia calderonii, Dalbergia stevensonii, Dalbergia cubilquitzensis, Dalbergia glomerata, Dalbergia longepedunculata,* and *D. calycina*).

Phylogenetic relationships, internal transcribed spacer tree

In this tree, we included a larger number of species from Africa (**Figure 3**). *Dalbergia* is monophyletic with a basal clade formed by South American Neotropical species (*D. miscolobium, D. spruceana, Dalbergia foliolosa, Dalbergia cuiabensis, D. villosa, Dalbergia acuta, Dalbergia revoluta, Dalbergia inundata,* and *Dalbergia lateriflora*) resolved sister to *Dalbergia afzeliana* from Africa. The second clade is formed by two subclades: one of Mexican climbing or woody vine species (*D. ecastaphyllum* and *D. monetaria*) as sister to a subclade of woody vine Asian species (*D. pinnata, D. tamarindifolia, D. rubiginosa, Dalbergia candenatensis, Dalbergia rostrata, D. stipulacea,* and *D. velutina*) and an African bush species (*Dalbergia microphylla*).

Clade III contains three subclades. The first contains Asiatic species (*Dalbergia ovata*, *D. cochinchinensis, Dalbergia sissoides,* and *D. latifolia)* and African tree species (*Dalbergia maritima, Dalbergia capuronii,* and *Dalbergia boehmi*). The second one is with Mexican and Central American species (*D. calycina, D. granadillo,* and *D. retusa*). The last subclade has a group of African species (*Dalbergia lactea, Dalbergia aurea,* and *Dalbergia bignonae)* as sister of a clade with Asian species grouped into three clades: the first one grouping *D. cana*, *D. oliveri, D. hancei,* and *Dalbergia lakhonensis;* second one grouping *D. stipulacea*, *Dalbergia yunnanensis*, *D. volubilis, D. paniculata,* and *D. nigrescens;* and the third one formed by D. sericea, *D. lanceolaria, Dalbergia godefroyi, D. stipulacea, Dalbergia huepeana, D. balansae,* and *D. assamica*.

Clade IV is formed by five subclades. The first subclade is formed by a climber African species *Dalbergia hostilis* and two Asian species (*Dalbergia sandakanensis* and *Dalbergia bintuluensis*). In the second subclade, three climbing and one small tree species from Mexico-Central America (*D. brownei, D. chontalensis, D. glabra,* and *D. tabascana*) are nested with *D. nigra* from Brazil; these species are the sisters of an Asian group formed by *D. dyeriana, D. hancei, D. cultrata, Dalbergia thorelii, Dalbergia lunghuhnii,* and *D. horrida*. The third subclade has African species *Dalbergia bracteolata* and *D. boehmii* as the sister group of a clade with the tree species *Dalbergia latifolia* and *D. melanoxylon*. The fourth subclade has *Dalbergia canescens* and *Dalbergia benthamii* from Asia, South American species (*D. cearensis, Dalbergia decipularis*, *D. variabilis = D. frutescens, D. brasiliensis, D. riparia,* and *Dalbergia frutenscens* var. *tomentosa*), and 12 Mexican–Central American species of trees (*D. agudeloi, D. calderonii, Dalbergia congestiflora, D. cubilquitzensis, D. glomerata, D. longepedunculata, Dalbergia luteola, D. melanocardium, D. palo-escrito, Dalbergia tilarana, D. tucurensis,* and *D. stevensonii*). The fifth subclade includes an African group of trees (*D. trichocarpa, Dalbergia greveana, Dalbergia abrahamii, Dalbergia humbertii, Dalbergia bojeri,* and *Dalbergia baronii*) with two divergent climbing species (*D. subcymosa* from South America and *Dalbergia martii* from Africa) as sister to Asian species. This group of Asiatic species consists of a divergent tree species (*D. cultrata*) sister to a subclade formed by woody climbers (*D. rimosa*, *Dalbergia cf. kingiana*, and *Dalbergia dialoides*), plus a tree species (*D. sissoo*), and a mix of trees,

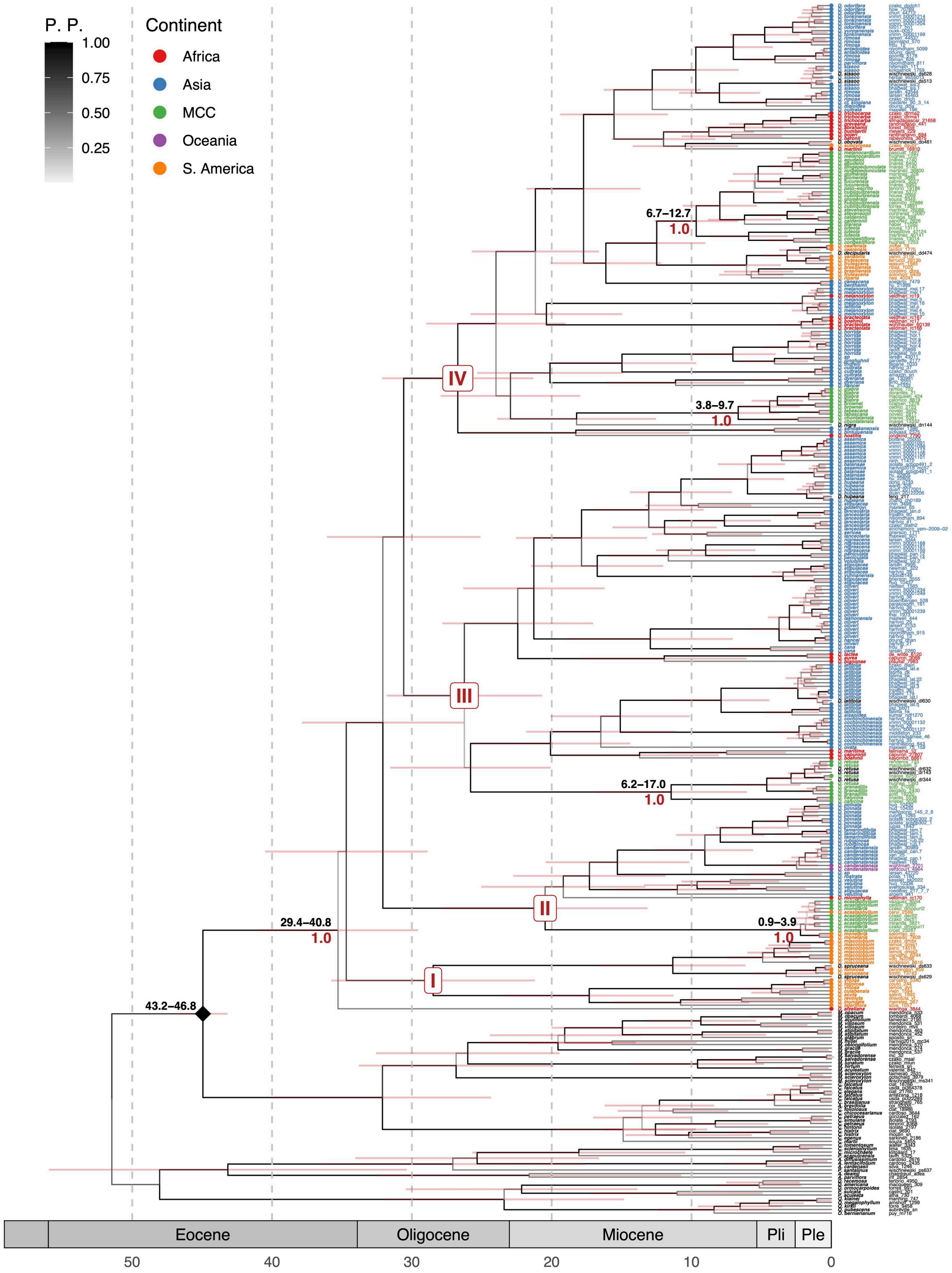


FIGURE 3

ITS calibrated tree of *Dalbergia*. Above the branches, the estimated age range of the clades is in black font. Local posterior probabilities are shown under branches in red font. Shading bottom bars represent geological epochs. The diamond mark represents the calibration node. The bars on the branches represent the range of the 95% confidence interval in the Bayesian tree.

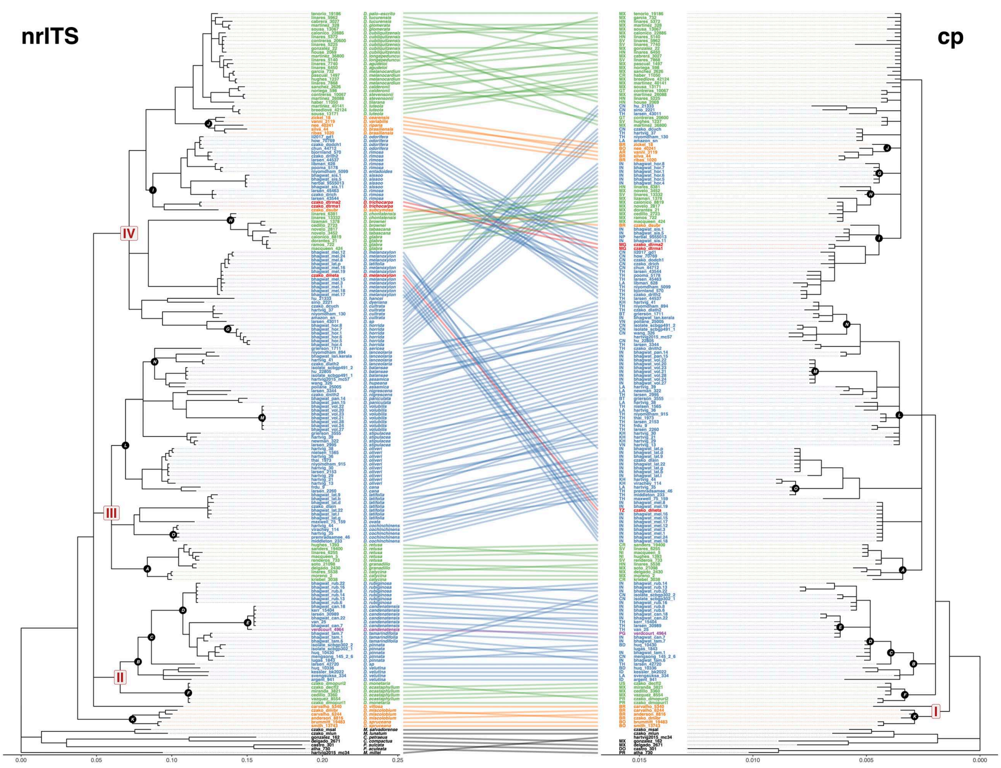


FIGURE 4
Tanglegram illustrating the discordance between the ITS gene tree (**nrITS**) and the combined plastid gene tree (**cp**) for *Dalbergia*. Links connect identical tips, with nodes rotated to minimize link overlap. Links are colored by geographical distribution. Clades that are similar between the two trees are indicated by black circles with white font. *Tipuana tipu* was eliminated because only the ITS1 sequence was available.

lianas, and woody climbers (*D. odorifera, Dalbergia tonkinensis, D. yunnanensis, Dalbergia rimosa var. foliacea, D. entadoides,* and *Dalbergia parviflora*).

There are not many obvious relationships conflicting between the nuclear loci tree and the two loci concatenated plastid tree (**Figure 4**). The backbone of the plastid loci tree is nearly identical to the nuclear ITS tree, with all major nodes and monophyly receiving strong support. In the plastid tree, there are many polytomies but a geographical structuring of the species is observed. The major clades disappear hierarchically but still form a group. The major conflict in the tanglegram is *Dalbergia melanoxylum*, in ITS tree is a sister species to *D. glabra* clade, and in plastid tree is part of a polytomy.

Geographical distribution of Mexican, Central American, and Caribbean species of *Dalbergia*

Most tree species of *Dalbergia* are restricted in distribution with the exception of *D. congestiflora*. Populations of *D. granadillo* clade are mostly distributed along the Pacific coast of Mexico from Colima to Panama. Climbing species have more widespread distributions, like *D. ecastaphyllum* and *D. monetaria* whose distribution reaches South America and Africa. *D. brownei* is a climbing species that has managed to spread as far as the coast of Florida. Distribution maps by clade can be found in **Figure 5**. Distribution maps by species can be found in **Figures 1–5** of the **Supplementary Material**.

Divergence time estimates

Divergence time estimation provided a robust time-calibrated tree of *Dalbergia* (**Figure 2**). The *Dalbergia* group arose 34.42 Ma during the Oligocene and diversification of the present day occurred during the Miocene to Pleistocene from 34.42 to 1.88 Ma. *Dalbergia* diverged from their sister genera 44.95 Ma and diversified during the Miocene (24.73–5.23 Ma). The divergence ages for Mexican *Dalbergia* species are between Quaternary (Pleistocene) and Tertiary (Neogene). The oldest Mexican clade is *Dalbergia granadillo* with 11.32 Ma (Miocene), *D. congestiflora* clade with 9.2 Ma (Miocene, Tortonian), *D. glabra* with 6.63 Ma (Miocene, Zancleane), and *D. ecastaphyllum* with 1.88 Ma (Pleistocene, Calabrian). Divergence estimations from the combined tree to the ITS tree do not vary considerably (**Figure 3**). The estimation age for the *Dalbergia granadillo* clade was 11.5 Ma, for *D. glomerata*

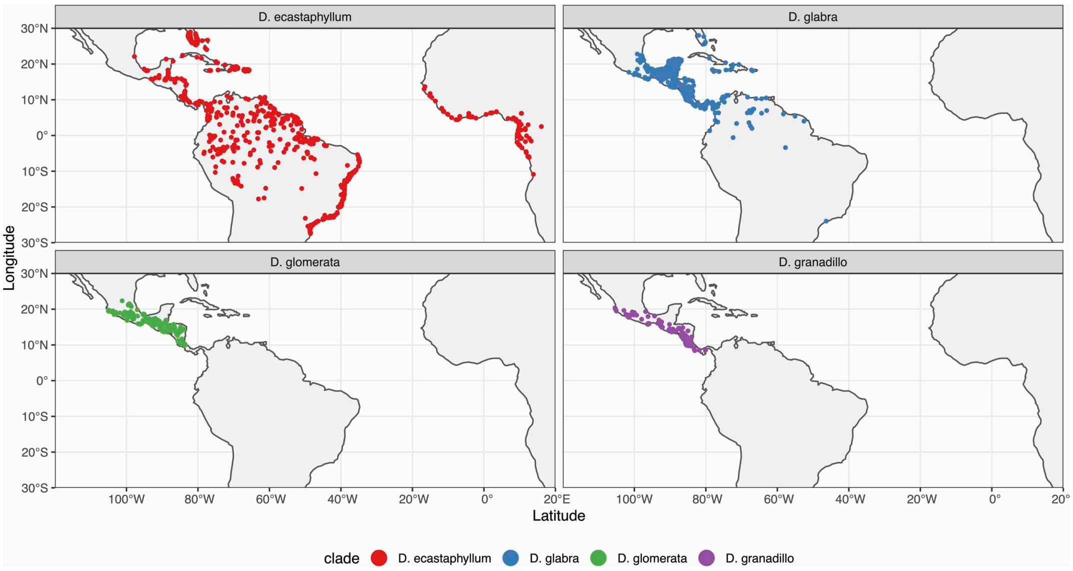


FIGURE 5
Distribution maps by clade for *Dalbergia* species from Mexico, Central America, and the Caribbean.

clade was 9.7 Ma, for *D. glabra* clade was 6.7 Ma, and for *D. ecastaphyllum* clade was 2.2 Ma.

Discussion

Taxonomy

Bentham (1860) divided the 64 species of *Dalbergia* known into six series (Triptolemea Americanae, Triptolemea, Sissoae Americanae, Sissoae Gerontogee, Dalbergariae, and Selenolobium). von Taubert (1894) divided the species into four sections (Triptolemaea, Sissoa, Dalbergaria, and Selenolobium). In the Neotropics, the 44 Brazilian species of *Dalbergia* were divided into five sections by Carvalho (1989, 1997) based on inflorescence and fruit types (Dalbergia, Ecastaphyllum, Pseudoecastaphyllum, Selenobium, and Triptolemea). In Asia, Prain (1904) classified the 86 South-East Asian species of *Dalbergia* into two subgenera (Amerimnon and Sissoa), five sections (Dalbergaria, Endespermum, Miscolobium, Podiopetalum, and Triptolemea) and 24 series. Finally, Thothathri (1987) categorized the 46 *Dalbergia* species, present in the Indian subcontinent, into four sections and seven series based on androecium and fruit types.

The sect. Triptolemea, with cymose inflorescences and samaroid legume, and sect. Ecastaphyllum, with racemose or paniculate inflorescences and orbicular to reniform legume sensu Carvalho (1997), are monophyletic. These sister relationships between species have also been found by Ribeiro et al. (2007); Vatanparast et al. (2013), and Hartvig et al. (2015). We also found relationships between *D. candenatensis, D. pinnata,* and *D. velutina* as other authors do (Vatanparast et al., 2013; Hartvig et al., 2015) but no as sister species. They are part of the same clade with *D. tamarindifolia* (sister to *D. pinnata*), *D. rubiginosa*, and *D. sericea*. Niyomdham et al. (1997) recognized that *D. pinnata, D. candenatensis,* and *D. velutina* have morphological affinities in the lower calyx tooth as long as or slightly longer than the laterals, standard equal to at least 3/4 of the blade, sometimes exceeding it. Niyomdham et al. (1997) treated *D. tamarindifolia* as a synonym of *D. pinnata,* specimens occurring together into the clade in two groups; these results might be indicative of taxonomic differences.

We found that sect. Dalbergia sensu Carvalho (1997) is also monophyletic. Species sampled from section Dalbergiaria sensu Prain (1904) are monophyletic too (series Lanceolarieae, Stipulaceae, and Volubilis). Vatanparast et al. (2013) treated *Dalbergia balansae* and *D. assamica* as separate species, but Hartvig et al. (2015) treated both species as a synonym. In this study, we treated the species separately and we included *Dalbergia hupeana*. Specimens in the phylogenetic analyses occur together; *D. hupeana* and *D. balansae* are sister species of *D. assamica,* as well as *D. sericea, D. nigrescens,* and *D. paniculata*. These results might be indicative that *D. balansae* and *D. assamica* are different species.

The Latifoliae series (*D. latifolia, D. cochinchinensis,* and *D. ovata*) from section Miscolobium (Prain, 1904) is monophyletic, and this group was also found by Vatanparast et al. (2013) and Hartvig et al. (2015). Morphological characters between *D. cochinchinensis* and *D. ovata* are lower calyx teeth as long or slightly longer than the lateral ones; standard longer than wide; leaves with (5—) 7-9 leaflets; leaflets acute to acuminate, apiculate, rarely obtuse, or rounded; flowers white to whitish, 5.5–6 mm long; fruits thin, papyraceous, glabrous, light brown (Niyomdham et al., 1997).

Accessions of *Dalbergia rimosa* are in three different groups in the same subclade. The first group is the sister species to *D. odorifera*, the second one is sister to *D. entadoides*, and the third one is basal *D. rimos*a accessions. The species is distributed in India, Myanmar, South of China, Thailand, Laos, and Vietnam from 200 to 1,300 m in mixed deciduous forests and scrub forests. When we see herbarium specimens from the different country distributions, it is clear that a morphological taxonomic revision must be carried out (e.g., Hooker J.D. sn. from Myanmar, Berhaman et al. SAN 134566 from Malaysia).

Mexican, Central American, and Caribbean species of *Dalbergia*

Although there is a taxonomic treatment for Flora Mesoamericana (Linares, in press), since Pittier (1922) there have been no attempts at subgeneric-level classifications of Mexican *Dalbergia*. Richter et al. (1996), citing personal comments by Richter et al. (1996), suggested four groups for the Mexican species. The first consists of *D. retusa*, *Dalbergia hypoleuca*, *D. granadillo*, and *Dalbergia lineata*, probably *D. cuscatlanica* and *Dalbergia pacifica*, all of which are similar in wood structure and metabolites.

The second group comprises Central American and Mexican species, *D. tucurensis* (including *D. cubilquitzensis*), *D. palo-escrito*, *D. melanocardium*, *D. glomerata, D. congestiflora, D. calderonii* (including *Dalbergia funera*), and probably *D. stevensoni*. No differentiation was detected in the wood of these species, although Richter et al. (1996) underline that *D. stevensonii* may be different from the rest. The third group, with the species *D. calycina* and *Dalbergia intibucana* (nowadays synonyms), were not sampled for Richter's study because of their lack of commercial value at that time. Finally, the fourth group with *D. brownei* whose wood parenchyma banding is similar to that found in *D. congestiflora* and *D. funera* but different in the uniseriate rays.

Three of the groups hypothesized by Rudd are phylogenetically valid. The first is referred to here as the *Dalbergia granadillo* clade (because it is the most traded species in the clade). Currently, only the following species are recognized by Linares (in press), *Dalbergia granadillo*, *D. retusa* var. *retusa*, and *D. retusa* var. *cuscatlanica*. To the same clade belongs *D. calycina* which Rudd recognized as a separate group and which is within the clade as the most divergent species. The species are distributed in the seasonally dry forests of the Pacific Coast of Mexico from Jalisco to Oaxaca (*D. granadillo*), in the seasonally dry forests of Southeastern Honduras, Nicaragua, and Costa Rica (*D. retusa* var. *retusa*), and in humid environments of Honduras and El Salvador (*D. retusa* var. *cuscatlanica*).

The second group is what we called the *Dalbergia glomerata* clade. The *D. glomerata* clade is a group of 12 species distributed in the Pacific Coast of Mexico from Jalisco to Costa Rica, in the Rio Balsas Depression, in the Gulf Coastal plain from Veracruz to North of Chiapas, and in Guatemala and Belize. Species can be found in cloud forests, seasonally dry tropical forests, tropical rainforests, or secondary vegetation. Sister species clade is from South America (*D. brasiliensis, D. riparia, D. cearensis*, and *D. variabilis* (*D. frutescens)).* They are part of Triptolemea and characterized by inflorescence cymose in terminal racemes; fruit oblong to elliptical, samaroid, with reticulate venation more prominent over the seed cavity. Species occur mainly in central and eastern Brazil, with the exception of *D. riparia* that inhabits the central Amazon Basin and less frequently on the lower Amazon.

Finally, the third group, referred to here as the *Dalbergia glabra* clade (fourth group for Rudd), includes the species *D. brownei*, *D. chontalensis*, *D. glabra,* and *D. tabascana*. Rudd does not include the last three species, although, in 1995, she described varieties within *D. glabra*. The *D. glabra* clade species are distributed from Veracruz, Mexico, to Honduras on the Atlantic and from Oaxaca, Mexico, to El Salvador in the Pacific Coast. Furthermore, Rudd did not say anything about the species *D. ecastaphyllum* and *D. monetaria* (*D. ecastaphyllum* clade). The *D. ecastaphyllum* clade is distributed from Florida, United States, to Brazil, passing through Mexico, and in Caribbean islands. Plant distribution records exist in the Western part of Africa. Species inhabits riparian vegetation, coastal dunes, mangrove forests, and mangrove-associated forests.

Time of diversification in *Dalbergia*

The origin of *Dalbergia* is probably South America, as the South American species are the earliest divergent. Later, the genus must have migrated to North America (possibly when Central America did not yet exist) and diversified into the four lineages we recognize today. In **Figure 6** (**Supplementary Material**), we can see that all haplotypes are central (in green). The Asian species evolved from the North American ones. There are different lineages between them, probably during the boreotropic (the only issue is that the geological data date this stage in the Eocene, implying that the genus is possibly older), which is in agreement with the fossil record found in America and Europe. The South American lineage of *Dalbergia frutescens* is a more recent arrival and derives from a southern migration of the *Dalbergia glomerata* clade (**Figure 1** and **Supplementary Material**).

The *Dalbergia granadillo* clade must have had an ancestor in mountain areas, tolerant of metamorphic rocky soils and dry conditions, and equivalent to mixed pine-oak forest. The earliest diverged species is *D. calycina*, a species found in montane areas such as Bochil or in the Cañon del Sumidero, both in Chiapas.

The *Dalbergia glomerata* clade can be divided into two groups: (1) species related to *D. cubilquitzensis* (species complex) that inhabit humid environments and are tolerant of limestone and metamorphic soils, and (2) species related to *D. congestiflora* that inhabit areas with marked seasonality and dryness. *Dalbergia stevensonii*, which is the most recently diversified species in the clade, has a morphology that resembles species from the seasonal dry forest and not from the humid and flooded area where it is currently distributed.

The *Dalbergia glabra* clade mostly consists of climbing species. The earliest divergent species, *D. chontalensis*, is a shrub distributed in floodplains or near low-lying streams. *D. brownei* is a shrubby, climbing species distributed on coastal dunes and has dispersed as far as Florida. *Dalbergia tabascana* is another lianoid species that has "specialized" to grow in freshwater swamp areas. The area where it is currently distributed was once a wetland area (San Lorenzo Tenochtitlán). The most recently diversified species (Pleistocene) is D. glabra, the only species in the clade that succeeded in diversifying from seasonal dry forest environments associated with water bodies to rain forests. Populations of this species can be found in the interior of the country but are always associated with water bodies.

The *Dalbergia ecastaphyllum* clade is the most recent in origin, and the species that comprise it are two lianoid species that inhabit mostly coastal regions. *D. monetaria* is the most recent species, is tolerant of freshwater bodies, and can therefore be found in different areas of the Amazon and in the African Congo. The fruit is floating and woody and has a "spongy" endocarp.

Most of the ages obtained here are younger than previous estimates (Lavin et al., 2005). Lavin et al. (2005) using a *matK* phylogenetic reconstruction estimated the age of divergence between *Dalbergia sisso* and *Tipuana tipu* in 49.1 ± 0.8 Ma (47.1–51.4 Ma). In the same study, they estimated the age of divergence between *D. sissoo* and *Ormocarpum* in 45.6 ± 0.8 Ma (43.9–47.3 Ma). However, in the study of Lavin et al. (2004), the reported *Dalbergia* estimation age from stem and crown clades is 40.4–43.3 Ma, and they give divergence estimates ranging from 12.7-3.8 to 7-12.2 Ma. Later on, Hung et al. (2020) with transcriptomes data (256 single-copy orthologs, 479,064 bp) established that the *Dalbergia miscolobium* clade is basal with an age of ± 14.78 Ma (Miocene-Langhian). The divergence ages found for *D. cochinchinensis* and *D. oliveri* in Indochina were estimated to be 11.68 Ma (Lower Miocene), which corresponds with the separation of the Thai–Malay Peninsula from Borneo ± 15 Ma ago (Vatanparast et al., 2013). The fossils of *Dalbergia* found in Europe are from the Miocene such as *Dalbergia nostratum* (15.97–23.03 Ma; lower Miocene), *Dalbergia lucida* (5.33–11.61 Ma; late Miocene), or *Dalbergia phleboptera* (27.82–23.2 Ma; Oligocene–Miocene). For Mexico, fossils of wood from Puebla are dated from the Oligocene (32 Ma, Sainz-Reséndiz, 2011).

Miocene diversification of *Dalbergia* reflects patterns shown in other tropical genera (Choo et al., 2020; Schley et al., 2022) in accordance with the climatic and ecological changes that occurred in the Tropics during the Miocene.

The combined marker phylogenetic reconstruction indicated that in Mexico, several ancestral independent lineages within *Dalbergia* might began their diversification consecutively during the Miocene. The ancestors of Mexican *Dalbergia* clades came from South America and Asia. How species were exchanged from South America to Mexico can be explained by migrations through Central America via the narrow Isthmus of Panama, which existed above sea level from the late Eocene to the Oligocene (38–28 Mya, Montes et al., 2012), and through which the exchanges of flora and fauna may have taken place (Cody et al., 2010). In the warm periods from the Eocene to Miocene through the transport of seeds and after the consolidation of the Isthmus, the contributions of flora must have increased.

For the *Dalbergia glomerata* clade, the sister group of taxa *D. brasiliensis, D. cearensis*, and *D. variabilis* (*D. frutescens)* have a fruit that could be wind-dispersed, while *D. riparia* has a fruit that is dispersed by water. The ancestor of the *D. glomerata* clade could have been wind-dispersed through Panama Isthmus. Physiographic conditions in Mexico at that time must have facilitated the introduction of coastal and low-elevation species through efficient mechanisms of long-distance seed dispersal (e.g., ancestors of *Dalbergia glomerata* and *D. granadillo*). These species then evolved in the Sierra Madre del Sur, which was still active during the early Miocene, and, later on, in some areas during the Pleistocene (Ferrari et al., 2005; Moran-Zenteno et al., 2007). Likewise, the complex Trans-Mexican Volcanic Belt generated hundreds of scenarios in Central Mexico from the Miocene to the present (Ferrari et al., 2012) promoting population divergence, and thus speciation.

Ancestors from Asia and Africa must have arrived in America due to long-distance dispersal. How did they arrive could have been by different routes. One of these ways could have been by ocean currents. The tropical Atlantic belts where surface currents and winds are simultaneously favorable for East-West crossing are found between the Congo delta and the Maranhão in Brazil and just North of the Senegal river delta and Northern Brazil and the Guianas. Both streams originate in river deltas in Africa. Parrish (1993) has suggested "rafting" transport of organisms between South America and Africa during the Tertiary and was probably predominantly from East to West rather than the other way around (Renner, 2004). Although these currents may have been different during the warm climates of the Eocene–Miocene, there is no evidence that they were different from those of today. The only current that may have been different is the one in the vicinity of the Isthmus of Panama, before it closed. There are also data that the Rio Grande rise (Southeastern of the coast in Brazil) and the western end of the Walvis Ridge (Southwest African Coast, Cape Town) may have been above water until the Oligocene (Parrish, 1993; Morley, 2000), reducing the distance between continental coasts. Another form of long-distance dispersal using ocean currents may have been by fruit flotation. Currently, we have a clear example with *Dalbergia ecastaphyllum* and *D. monetaria* found in America but also in the western portion of Africa and whose fruits are frequent buoyants in marshes and rivers and can remain floating up to nine months (Gunn et al., 1976). There are *Dalbergia* species reported in the literature as *Dalbergia monosperma* that are waterborne (Ridley, 1990). Some of the Asian species of *Dalbergia* have fruits with similar characteristics to *D. ecastaphyllum* and *D. monetaria* to be transported, because they have coriaceous to woody fruits with a single seed which would form an air chamber inside, allowing them to float (e.g., *Dalbergia albertesii, Dalbergia beccarii, D. horrida,* and *D. tamarindifolia*).

Other options for long-distance transport are migratory birds, but except for some Psittacidae that consume *Dalbergia* seeds, there are no migratory species that could transport them from Asia to America or vice versa. While only ocean currents are heard as consistent for *Dalbergia* to be dispersed over long distances, all of the above mechanisms together could shape the diversity encountered in the genus today. Species distribution must also be related to soil type and microbiome. Rasolomampianina et al. (2005) found 68 strains of fixative bacteria in eight endemic *Dalbergia* species from Madagascar. Some of these strains such as *Bradhirhizobium* are common in tropical legumes, but the others are specific.

Concluding remarks

The reconstructed evolutionary history of *Dalbergia* from Mexico and Central America provides insights on how the number of species present in the area may have originated.

Regarding genetic barcodes, the most commonly used for *Dalbergia* have been ITS, matK, and rbcL, either alone or in different combinations (Bhagwat et al., 2015; Hassold, 2015; Li et al., 2017). Li et al. (2017) recommend the combination *ITS + matK + rbcL* to identify *Dalbergia* species. Our results show that, for species from Mexico, Central America, and the Caribbean, the ITS region is acceptable to distinguish at the species level, and in combination with chloroplast markers, we can know the area of provenance. Hassold (2015), in her study with chloroplast markers, indicates that plastid sequences reflect the geographical range and shared haplotypes between species. The data obtained in this study demonstrate that the whole piece of ITS alone can help us to differentiate between *Dalbergia* species. If the area of provenance is also required, it will be necessary to use chloroplast sequences.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/**Supplementary Material**.

Author contributions

SS designed the study, collected the materials, conducted the experiments, drafted the manuscript, and secured funding for the project. EP-O conceived and conducted the bioinformatic analyses, reviewed the manuscript, and assisted in the discussions. EM-S assisted in the discussions and reviewed the manuscript. JL provided taxonomic advice, reviewed the manuscript, and assisted in the discussions. LC trained students in the laboratory and performed a first draft assembly of sequences generated by this study. All authors contributed to the article and approved the submitted version.

Funding

This study was supported by a grant from PAPIIT project IA2015 to SS and operative funding to SS from Instituto de Biología.

Acknowledgments

We thank Berenice Mendoza, Mayra Castillo, and José Antonio Contreras-Chijate for their help in the laboratory. Our special gratitude goes to Laura Márquez and Nelly López-Ortiz for sequencing.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fevo.2022.910250/full#supplementary-material

SUPPLEMENTARY FIGURES 1–5
Distribution maps by species.

SUPPLEMENTARY FIGURE 6
Haplotype network for plastid markers of *Dalbergia*.

SUPPLEMENTARY TABLE 1
List of *Dalbergia* species and allies used in this study.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Trans. Automat. Control* 19, 716–723.

Barragán-Huerta, B. E., Peralta-Cruz, J., González-Laredo, R. F., and Karchesy, J. (2004). Neocandenatone, an isoflavan-cinnamylphenol quinone methide pigment from *Dalbergia congestiflora*. *Phytochemistry* 65, 925–928. doi: 10.1016/j.phytochem.2003.11.011

Bentham, G. (1860). Synopsis of Dalbergieæ, a tribe of leguminosæ. *J. Proc. Linn. Soc. Lond. Bot.* 4, 1–128. doi: 10.1111/j.1095-8339.1860.tb02464.x

Bhagwat, R. M., Dholakia, B. B., Kadoo, N. Y., and Balasundaran, M. (2015). Two new potential barcodes to discriminate *Dalbergia* species. *PLoS One* 10:e0142965. doi: 10.1371/journal.pone.0142965

Bouckaert, R., Vaughan, T. G., Barido-Sottani, J., Duchêne, S., Fourment, M., Gavryushkina, A., et al. (2019). BEAST 2.5: an advanced software platform for Bayesian evolutionary analysis. *PLoS Comput. Biol.* 15:e1006650. doi: 10.1371/journal.pcbi.1006650

Cardoso, D. B., Mattos, C. M., Filardi, F. L., Delgado-Salinas, A., Lavin, M., Moraes, P. L., et al. (2020). A molecular phylogeny of the pantropical papilionoid legume *Aeschynomene* supports reinstating the ecologically and morphologically coherent genus *Ctenodon*. *Neodiversity* 13, 1–38. doi: 10.13102/neod.131.1

Carvalho, A. M. (1989). *Systematic Studies of the Genus Dalbergia L.f. in Brazil*. Reading: University of Reading.

Carvalho, A. M. (1997). A synopsis of the genus *Dalbergia* (*Fabaceae*: *Dalbergieae*) in Brazil. *Brittonia* 49, 87–109. doi: 10.2307/2807701

CBOL Plant Working Group (2009). A DNA barcode for land plants. *Proc. Natl. Acad. Sci. U.S.A.* 106, 12794–12797. doi: 10.1073/pnas.0905845106

Choo, L. M., Forest, F., Wieringa, J. J., Bruneau, A., and de la Estrella, M. (2020). Phylogeny and biogeography of the *Daniellia* clade (*Leguminosae*: *Detarioideae*), a tropical tree lineage largely threatened in Africa and Madagascar. *Mol. Phylogenet. Evol.* 146:106752. doi: 10.1016/j.ympev.2020.106752

Cody, S., Richardson, J. E., Rull, V., Ellis, C., and Pennington, T. (2010). The Great American Biotic Interchange revisited. *Ecography* 33, 326–332. doi: 10.1111/j.1600-0587.2010.06327.x

Darriba, D., Taboada, G. L., Doallo, R., and Posada, D. (2012). jModelTest 2: more models, new heuristics and high-performance computing. *Nat. Methods* 9:772. doi: 10.1038/nmeth.2109

Ferrari, L., Orozco-Esquivel, T., Manea, V., and Manea, M. (2012). The dynamic history of the Trans-Mexican Volcanic Belt and the Mexico subduction zone. *Tectonophysics* 522–523, 122–149. doi: 10.1016/j.tecto.2011.09.018

Ferrari, L., Valencia-Moreno, M., and Bryan, S. (2005). Magmatismo y tectónica en la Sierra Madre Occidental y su relación con la evolución de la margen occidental de Norteamérica. *Boletín de la Soc. Geol. Mexicana* 57, 343–378. doi: 10.18268/bsgm2005v57n3a5

GBIF.org (2022). *GBIF Occurrence Download*. Copenhagen: GBIF. doi: 10.15468/dl.hw5mmh

Gunn, C. R., Dennis, J. V., and Paradine, J. (1976). *Gunn World Guide To Tropical Drift Seeds and Fruits*. New York, NY: Demeter Press.

Hamburger, M. O., Cordell, G. A., Tantivatana, P., and Ruangrungsi, N. (1987). Traditional Medicinal Plants of Thailand, VIII. Isoflavonoids of *Dalbergia candenatensis*. *J. Nat. Prod.* 50, 696–699. doi: 10.1021/np50052a020

Hartvig, I., Czako, M., Kjær, E. D., Nielsen, L. R., and Theilade, I. (2015). The Use of DNA Barcoding in Identification and Conservation of Rosewood (*Dalbergia* spp.). *PLoS One* 10:e0138231. doi: 10.1371/journal.pone.0138231

Hassold, S. (2015). *Molecular Identification of Malagasy Dalbergia species (rosewoods) for Biodiversity Conservation*. Ph.D. thesis. Zürich: ETH Zurich, doi: 10.3929/ethz-a-010782581

Hung, T. H., So, T., Sreng, S., Thammavong, B., Boounithiphonh, C., Boshier, D. H., et al. (2020). Reference transcriptomes and comparative analyses of six species in the threatened rosewood genus *Dalbergia*. *Sci. Rep.* 10:17749. doi: 10.1038/s41598-020-74814-2

Katoh, K., Asimenos, G., and Toh, H. (2009). Multiple alignment of DNA sequences with MAFFT. *Methods Mol. Biol.* 537, 39–64. doi: 10.1007/978-1-59745-251-9_3

Kelchner, S. A., and Clark, L. G. (1997). Molecular evolution and phylogenetic utility of the chloroplast rpl16 intron in *Chusquea* and the Bambusoideae (Poaceae). *Mol. Phylogenet. Evol.* 8, 385–397. doi: 10.1006/mpev.1997.0432

Klitgård, B., and Lavin, M. (2005). "*Dalbergieae*," in *Legumes of the World*, eds G. P. Lewis, B. Schrire, B. MacKinder, and M. Lock (Kew, UK: Royal Botanic Gardens), 307–335.

Larsson, A. (2014). AliView: a fast and lightweight alignment viewer and editor for large datasets. *Bioinformatics* 30, 3276–3278. doi: 10.1093/bioinformatics/btu531

Lavin, M., Herendeen, P. S., and Wojciechowski, M. (2005). Evolutionary rates analysis of *leguminosae* implicates a rapid diversification of lineages during the tertiary. *Syst. Biol.* 54, 575–594. doi: 10.1080/10635150590947131

Lavin, M., Schrire, B. P., Lewis, G., Pennington, R. T., Delgado-Salinas, A., et al. (2004). Metacommunity Process Rather than Continental Tectonic History Better Explains Geographically Structured Phytogenies in Legumes. *Philos. Trans. Biol. Sci.* 359, 1509–1522. doi: 10.1098/rstb.2004.1536

Lavin, M., Wojciechowski, M. F., Richman, A. D., Rotella, J. J., Sanderson, M. J., and Beyra-Matos, A. (2001). Identifying tertiary radiations of fabaceae in the greater antilles: alternatives to cladistic vicariance analysis. *Int. J. Plant Sci.* 162, S53–S76.

Li, Q., Wu, J., Wang, Y., Lian, X., Wu, F., Zhou, L., et al. (2017). The phylogenetic analysis of *Dalbergia* (*Fabaceae*: *Papilionaceae*) based on different DNA barcodes. *Holzforschung* 71, 939–949. doi: 10.1515/hf-2017-0052

Lianhe, Z., Li, W., Xing, H., and Zhengxing, C. (2011). Antioxidant activities of seed extracts from *Dalbergia odorifera* T. Chen. *Afr. J. Biotechnol.* 10, 11658–11667.

Linares, J., and Sousa, M. (2007). Nuevas especies de *Dalbergia* (*Leguminosae*: *Papilionoideae*: *Dalbergieae*) en México y Centroamérica. *Ceiba* 48, 61–82.

Linares, J. L. (in press). *Dalbergia. Flora Mesoamericana: Fabaceae a Begoniaceae*, Vol. 3. St. Louis, MO: Missouri Botanical Garden.

Montes, C., Cardona, A., McFadden, R., Morón, S. E., Silva, C. A., Restrepo-Moreno, S., et al. (2012). Evidence for middle Eocene and younger land emergence in central Panama: implications for Isthmus closure. *GSA Bull.* 124, 780–799. doi: 10.1130/B30528.1

Moran-Zenteno, D., Cerca, M., and Keppie, J. D. (2007). "The Cenozoic tectonic and magmatic evolution of southwestern Mexico: advances and problems of interpretation," in *Geology of Mexico: Celebrating the Centenary of the Geological Society of Mexico*, eds S. A. Alaniz-Alvarez and A. F. Nieto-Samaniego (Boulder, CO: Geological Society of America), 71–90. doi: 10.1130/2007.2422(03)

Morley, R. J. (2000). *Origin and Evolution of Tropical Rain Forests*. Hoboken: John Wiley & Sons.

Niyomdham, C., Hô, P. H., Dy Phon, P., and Vidal, J. E. (eds) (1997). *Flore du Cambodge, du Laos et du Viêtnam. 29 Légumineuses-Papilionoïdées-Dalbergiées*. Paris: Muséum National d'Histoire Naturelle.

Parrish, J. T. (1993). Climate of the supercontinent Pangea. *J. Geol.* 101, 215–233.

Pérez-Gutiérrez, R. M., and García-Baez, E. (2013). Cytotoxic activity of isoflavan-cinnamylphenols from *Dalbergia congestiflora* on HeLa cells. *J. Med. Plants Res.* 7, 2992–2998.

Pittier, H. (1922). On the Species of *Dalbergia* of Mexico and Central America. *J. Wash. Acad. Sci.* 12, 54–64.

Prain, D. (1904). The species of *Dalbergia* of southeastern Asia. *Ann. R. Bot. Gard.* 10, 1–114.

Rasolomampianina, R., Bailly, X., Fetiarison, F., Rabevohitra, R., Béna, G., Ramaroson, L., et al. (2005). Nitrogen-fixing nodules from rosewood legume trees (*Dalbergia* spp.) endemic to Madagascar host seven different genera belonging to α-and β-*Proteobacteria*. *Mol. Ecol.* 14, 4135–4146. doi: 10.1111/j.1365-294X.2005.02730.x

Renner, S. S. (2004). Plant dispersal across the tropical Atlantic by wind and sea currents. *Int. J. Plant Sci.* 165, S23–S33.

Ribeiro, R. A., Lavin, M., Lemos-Filho, J. P., Mendonça-Filho, C. V., Rodrigues dos Santos, F., and Lovato, M. B. (2007). The genus *Machaerium* (*Leguminosae*) is more closely related to *Aeschynomene* sect. *Ochopodium* than to *Dalbergia*: inferences from combined sequence data. *Syst. Bot.* 32, 762–771. doi: 10.1600/036364407783390700

Richter, H. G., Krause, V. J., and Muche, C. (1996). *Dalbergia congestiflora* Standl.: wood structure and physico-chemical properties compared with other Central American species of *Dalbergia*. *IAWA J.* 17, 327–341. doi: 10.1163/22941932-90001583

Ricker, M., Hernández, H. M., Sousa, M., and Ochoterena, H. (2013). Tree and tree-like species of Mexico: *Asteraceae*, *Leguminosae*, and *Rubiaceae*. *Rev. Mex. Biodivers.* 84, 439–470. doi: 10.7550/rmb.32013

Ridley, H. N. (1990). *Dispersal of Plants Throughout the World*. Ashford: L. Reeve & Co., Ltd.

Rutiaga-Quiñones, J. G., Pedraza-Bucio, F. E., and López-Albarrán, P. (2010). Componentes químicos principales de la madera de *Dalbergia granadillo* Pittier y de *Platymiscium lasiocarpum* Sandw. *Rev. Chapingo Ser. Cienc. For. Ambient.* 16, 179–186. doi: 10.5154/R.RCHSCFA.2010.04.023

Rutiaga-Quiñones, J. G., Windeisen, E., and Schumacher, P. (1995). Anti fungal activity of heartwood extracts from *Dalbergia granadillo* and *Enterolobium cyclocarpum*. *Holz als Rohund Werkstoff* 53, 308–308.

Sainz-Reséndiz, B. A. (2011). *Descripción e identificación de maderas del Paleógeno de San Juan Atzingo, Puebla, México, Facultad de Estudios Superiores, Iztacala*. Ph.D. thesis. México: Universidad Nacional Autónoma de México.

Schley, R. J., Qin, M., Vatanparast, M., Malakasi, P., de la Estrella, M., Lewis, G. P., et al. (2022). Pantropical diversification of padauk trees and relatives was influenced by biome-switching and long-distance dispersal. *J. Biogeogr.* 49, 391–404. doi: 10.1111/jbi.14310

SEMARNAT (2010). *Norma Oficial Mexicana NOM-059-SEMARNAT-2010, Protección ambiental – Especies nativas de México de flora y fauna silvestres – Categorías de riesgo y especificaciones para su inclusión, exclusión o cambio – Lista de especies en riesgo. Diario Oficial de la Federación 30 de diciembre de 2010.* Mexico: SEMARNAT.

Sousa, S. M., Ricker, M., and Hernández, H. M. (2001). Tree Species of the Family *Leguminosae* in Mexico. *Harv. Pap. Bot.* 6, 339–365.

Standley, P. C. (1922). Trees and shrubs of Mexico. *Contr. U. S. Nat. Herb.* 23, 1–1312.

Stucky, B. J. (2012). SeqTrace: a graphical tool for rapidly processing DNA sequencing chromatograms. *J. Biomol. Tech.* 23, 90–93. doi: 10.7171/jbt.12-2303-004

Sun, Y., Skinner, D. J., Liang, G. H., and Hulbert, S. H. (1994). Phylogenetic analysis of Sorghum and related taxa using internal transcribed spacers of nuclear ribosomal DNA. *Theor. Appl. Gen.* 89, 26–32.

Thothathri, K. (1987). *Taxonomic Revision of the Tribe Dalbergieae in the Indian Subcontinent*. Kolkata: Botanical Survey of India.

Vatanparast, M., Klitgård, B., Frits, A. C., Adema, R., Pennington, T., Yahara, T., et al. (2013). First molecular phylogeny of the pantropical genus *Dalbergia*: implications for infrageneric circumscription and biogeography. *S. Afr. J. Bot.* 89, 143–149. doi: 10.1016/j.sajb.2013.07.001

von Taubert, P. (1894). III sb. Papilionatae-Dalbergieae-Lonchocarpinae. *Nat. Pflanzenfamilien* 3, 341–348.

Wirth, W., and Duchene, S. (2021). Real-Time and Remote MCMC Trace Inspection with Beastiary. *bioRxiv* [Preprint]. doi: 10.1101/2021.11.21.469478v1

Yu, G. (2020). Using ggtree to visualize data on tree-like structures. *Curr. Prot. Bioinformatics* 69:e96. doi: 10.1002/cpbi.96

Zizka, A., Silvestro, D., Andermann, T., Azevedo, J., Ritter, C. D., Edler, D., et al. (2019). CoordinateCleaner: standardized cleaning of occurrence records from biological collection databases. *Methods Ecol. Evol.* 10, 744–751. doi: 10.1111/2041-210X.13152

TYPE Brief Research Report
PUBLISHED 02 August 2022
DOI 10.3389/fevo.2022.953131

A comparative study of RNA yields from museum specimens, including an optimized protocol for extracting RNA from formalin-fixed specimens

Kelly A. Speer[1,2]*, Melissa T. R. Hawkins[3], Mary Faith C. Flores[3], Michael R. McGowen[3], Robert C. Fleischer[1], Jesús E. Maldonado[1], Michael G. Campana[1] and Carly R. Muletz-Wolz[1]

[1]Center for Conservation Genomics, Smithsonian's National Zoo and Conservation Biology Institute, Washington, DC, United States, [2]Department of Invertebrate Zoology, National Museum of Natural History, Washington, DC, United States, [3]Department of Vertebrate Zoology, National Museum of Natural History, Washington, DC, United States

Animal specimens in natural history collections are invaluable resources in examining the historical context of pathogen dynamics in wildlife and spillovers to humans. For example, natural history specimens may reveal new associations between bat species and coronaviruses. However, RNA viruses are difficult to study in historical specimens because protocols for extracting RNA from these specimens have not been optimized. Advances have been made in our ability to recover nucleic acids from formalin-fixed paraffin-embedded samples (FFPE) commonly used in human clinical studies, yet other types of formalin preserved samples have received less attention. Here, we optimize the recovery of RNA from formalin-fixed ethanol-preserved museum specimens in order to improve the usability of these specimens in surveys for zoonotic diseases. We provide RNA quality and quantity measures for replicate tissues subsamples of 22 bat specimens from five bat genera (*Rhinolophus*, *Hipposideros*, *Megareops*, *Cynopterus,* and *Nyctalus*) collected in China and Myanmar from 1886 to 2003. As tissues from a single bat specimen were preserved in a variety of ways, including formalin-fixed (8 bats), ethanol-preserved and frozen (13 bats), and flash frozen (2 bats), we were able to compare RNA quality and yield across different preservation methods. RNA extracted from historical museum specimens is highly fragmented, but usable for short-read sequencing and targeted amplification. Incubation of formalin-fixed samples with Proteinase-K following thorough homogenization improves RNA yield. This optimized protocol extends the types of data that can be derived from existing museum specimens and facilitates future examinations of host and pathogen RNA from specimens.

KEYWORDS

Coronaviridae, Chiroptera (bats), natural history collection, historical specimens, RNA

OPEN ACCESS

EDITED BY
Jonathan J. Fong,
Lingnan University, China

REVIEWED BY
Bryan Stuart,
North Carolina Museum of Natural Sciences, United States
Joon-Yong Chung,
National Cancer Institute (NIH), United States

*CORRESPONDENCE
Kelly A. Speer
speerk@si.edu

SPECIALTY SECTION
This article was submitted to Evolutionary and Population Genetics, a section of the journal Frontiers in Ecology and Evolution

RECEIVED 25 May 2022
ACCEPTED 13 July 2022
PUBLISHED 02 August 2022

CITATION
Speer KA, Hawkins MTR, Flores MFC, McGowen MR, Fleischer RC, Maldonado JE, Campana MG and Muletz-Wolz CR (2022) A comparative study of RNA yields from museum specimens, including an optimized protocol for extracting RNA from formalin-fixed specimens.
Front. Ecol. Evol. 10:953131.
doi: 10.3389/fevo.2022.953131

Introduction

Natural history collections are an essential and underused resource for emerging infectious disease research (Talley et al., 2015; Schmitt et al., 2018; Colella et al., 2021; Thompson et al., 2021). These collections preserve snapshots of animal and plant populations and their associated parasites and pathogens through time. This quality has been used to track the spread of invasive parasites and pathogens in wildlife (Kleindorfer and Sulloway, 2016) and emergence of human pathogens (Childs et al., 1994; Yates et al., 2002). Natural history collections also maintain voucher specimens that can be revisited and compared between projects and institutions, a feature that can make pathogen surveillance more effective and reproducible (Colella et al., 2021; Thompson et al., 2021). Lastly, these collections maintain multiple specimen types that can be analyzed in new ways as new technology is developed, enabling novel data to be derived from existing resources. For example, DNA sequencing revolutionized our understanding of the information stored within a natural history specimen. Now, with the development of more sensitive and accurate sequencing and imaging technologies, we can also detect the community of pathogens associated with a specimen.

The spillover of SARS-CoV-2 from wildlife to humans has led to increased screening for coronaviruses, a highly diverse family (Coronaviridae) of positive-sense single-stranded RNA viruses, in bats globally (Valitutto et al., 2020; Becker et al., 2022). There are also efforts to revisit bat specimens housed in natural history collections to examine the evolution and host associations of coronaviruses. However, few protocols are available for extracting RNA from museum specimens (although see Fanning et al., 2002), limiting the use of museum specimens in viral screening efforts. RNA is a rapidly deteriorating molecule that requires specialized stabilization (Camacho-Sanchez et al., 2013), and many museum specimens are not preserved with RNA in mind. While RNA is less stable than DNA, RNA can persist even in ancient plant and animal tissues (Fordyce et al., 2013; Shaw et al., 2019; Smith et al., 2019) and ancient RNA methods have been used to examine viruses (Castello et al., 1999; Fanning et al., 2002; Smith et al., 2014; Düx et al., 2020). The persistence of RNA in historical and ancient tissues supports the value of natural history specimens in examining viral pathogens through time and other downstream uses, including host gene expression profiles.

Here, we examine the quality and quantity of RNA that can be extracted from bat specimens ranging in age (19–136 years old) and varying in preservation method (i.e., formalin-fixed ethanol-preserved at room temperature, ethanol-preserved and stored at room temperature, ethanol-preserved and frozen, and flash-frozen without buffer; **Figure 1**). We present an optimized protocol for extracting RNA from formalin-fixed specimens that is refined from existing protocols developed for extracting RNA from formalin-fixed paraffin-embedded (FFPE) tissues (Krafft et al., 1997; Fanning et al., 2002; Sharma et al., 2012). We used a suite of tools to confirm the success of RNA extractions and examine the downstream usability of these extractions, including Qubit, Bioanalyzer, qPCR, and RNA-seq. This research builds on the growing body of evidence that natural history specimens capture an extended suite of data that can be used beyond the original intent for which that specimen was vouchered, reinforcing the value of natural history collections.

Methods

Specimen subsampling

We sampled bat specimens and tissues (n = 22 unique bats) housed in the Smithsonian National Museum of Natural History (NMNH) that represented species from five genera (*Rhinolophus*, *Hipposideros*, *Megareops*, *Cynopterus* and *Nyctalus*) collected in 1886, 1888, and 2002–2003 in Myanmar, and in 1989 in China (**Supplementary Table 1**). Whole voucher specimens were either preserved in ethanol at room temperature (1886, 1888: ethanol-RT) or were fixed in formalin and transferred to ethanol for long-term storage (2002–2003: formalin-fixed). From some of these whole specimens (n = 4), organ and/or muscle tissue was sampled in the field and preserved in ethanol and then frozen for long-term storage at –20°C (ethanol-F). For other bats, only organ and muscle tissues were available. Some tissues sampled in the field were flash frozen using liquid nitrogen and stored in vapor phase liquid nitrogen freezers (1989: flash-frozen). As there are multiple sample types taken from the same bat individual, we use paired tissue subsamples to examine the impact of preservation method on RNA (**Supplementary Table 1** and **Figure 1**).

Fluid vouchers

We collected lung and small intestine tissue samples from formalin-fixed bat vouchers (n = 8 bats; **Figure 1**). Specimens were removed from their jars and blotted dry to remove excess 70% ethanol. Next, thoracic and abdominal cavities were dissected using sterilized instruments (forceps, scissors, hemostats and scalpels) treated with RNase AWAY^{TM} (Thermo Fisher Scientific, Waltham, MA, United States). Approximately 50 mg subsamples of lung and small intestine were weighed and then placed in a 1.5 mL tube containing PBS buffer (to remove remaining ethanol), shaken for approximately 5 s, moved to a 1.5 mL tube containing ddH_20, again shaken for 5 s, and finally transferred to a 1.5 mL tube containing Trizol^{TM} buffer (Invitrogen, Waltham, MA, United States). After dissections were completed, all tubes

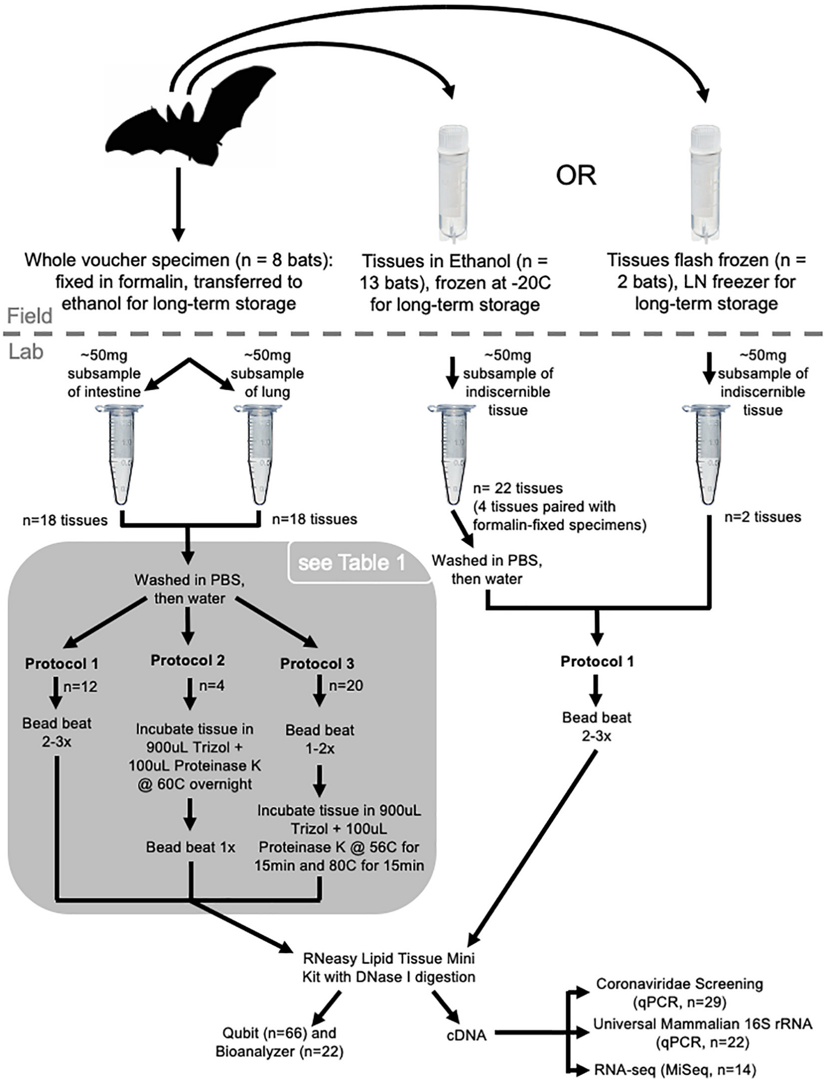

OR
Whole voucher specimen (n = 8 bats): fixed in formalin, transferred to ethanol for long-term storage
Tissues in Ethanol (n = 13 bats), frozen at -20C for long-term storage
Tissues flash frozen (n = 2 bats), LN freezer for long-term storage
Field
Lab
~50mg subsample of intestine
~50mg subsample of lung
~50mg subsample of indiscernible tissue
~50mg subsample of indiscernible tissue
n=18 tissues
n=18 tissues
n= 22 tissues (4 tissues paired with formalin-fixed specimens)
n=2 tissues
see Table 1
Washed in PBS, then water
Washed in PBS, then water
Protocol 1
n=12
Bead beat 2-3x
Protocol 2
n=4
Incubate tissue in 900uL Trizol + 100uL Proteinase K @ 60C overnight
Bead beat 1x
Protocol 3
n=20
Bead beat 1-2x
Incubate tissue in 900uL Trizol + 100uL Proteinase K @ 56C for 15min and 80C for 15min
Protocol 1
Bead beat 2-3x
RNeasy Lipid Tissue Mini Kit with DNase I digestion
Qubit (n=66) and Bioanalyzer (n=22)
cDNA
Coronaviridae Screening (qPCR, n=29)
Universal Mammalian 16S rRNA (qPCR, n=22)
RNA-seq (MiSeq, n=14)

FIGURE 1
Sampling and processing design. Whole voucher specimens collected in 1886 and 1888 (*n* = 3 bats, *n* = 6 tissues; ethanol-RT) are not shown in the figure.

were transferred to a –20°C freezer for temporary storage and then transferred to a –80°C freezer until extraction occurred. Ethanol-RT bat vouchers pre-dated the use of formalin in museums (collected in 1886 and 1888). Tissues from these specimens were sampled in the same way as formalin-fixed vouchers.

Tissue samples

A subsample of frozen tissue samples preserved in ethanol or flash frozen were loaned from the NMNH Biorepository and stored at –80°C until extraction. During subsampling, all instruments were treated with RNase AWAY^{TM}. Frozen tissues were moved from –80°C to a –20°C freezer for approximately 1 h prior to subsampling. Flash-frozen tissues were then stored at 4°C and processed individually. Tubes containing ethanol-F tissues were removed from the –20°C freezer one at a time and stored on ice during subsampling. It was not possible to discern the tissue type of these subsamples as multiple organ types (usually heart, liver, lung, kidney, spleen) and muscle are frequently sampled in the field and put in the same tube for long-term storage. These tissues do not always maintain diagnostic morphology during long-term storage and become indiscernible from each other. Prior to extractions tissues were weighed to confirm they did not exceed ∼50 mg and washed in 1× nuclease-free PBS and nuclease-free water (all except flash-frozen samples) as described for tissues sampled from fluid vouchers. Following this washing step, samples were transferred to Trizol^{TM} and extracted immediately.

RNA extraction and quality assessment

RNA extraction protocols

Following the PBS and water washes for samples in ethanol (all except flash-frozen samples), all tissues were transferred to Trizol^{TM} and either processed immediately or stored frozen until RNA extraction. We tested three protocols for extracting RNA from the tissues (**Figure 1**). Six negative controls used during extraction yielded no measurable RNA.

For protocol 1, tissues were homogenized manually or by bead-beating and then RNA was extracted using the RNeasy Lipid Tissue Mini Kit (Qiagen, Hilden, Germany). Ethanol-RT samples were collected in 1886 and 1888, and therefore were processed in the Smithsonian National Zoo and Conservation Biology Institute's ancient DNA laboratory. These samples were homogenized in 40μL Trizol using BioMashers II^{TM} (Kimble, Rockwood, TX, United States) until no chunks of tissue were visible. Then 960μL of Trizol was added, the pestle was removed, and the tube was centrifuged for 1 min. Supernatant was transferred from the BioMasher tube to a screw cap tube before proceeding with the RNeasy Lipid Tissue Mini Kit extraction protocol, beginning after the homogenization steps in the Kit handbook (start at step 12, the first step under Preparation of Total RNA, handbook v. 07/2018). Ethanol-F samples were homogenized using a Mini-BeadBeater-96 (BioSpec, Bartlesville, OK, United States) and one 3 mm chrome steel bead. Prior to use, we soaked beads in RNase AWAY^{TM} for 5 min and washed them twice with RNase-free water. We poured off the water washes and irradiated the beads with UV light for 5 min (UV Clave, Benchmark Scientific, Sayreville, NJ, United States) before transferring one bead to a screw-cap tube containing tissue sample and 1 mL Trizol buffer. Each sample was bead beat two times at maximum speed (40 oscillations/second) for 30 s and incubated at –20°C for 2 min following each bout of bead beating. If large chunks of tissue were visible, we repeated bead beating and incubation once more. Supernatant was transferred to a new tube and RNA extraction proceeded using the RNeasy Lipid Tissue Mini Kit (step 12 as above). Kit extraction followed the manufacturer's protocol and included DNase I digestion (RNase-free DNase Set, Qiagen). RNA was eluted in 40 μL of RNase-free water and the elution was repeated using the original eluate to re-wet the filter as recommended to increase RNA concentration. Extractions were split into two aliquots to reduce freeze-thaw cycles and stored at –80°C.

Protocols 2 and 3 were optimized from Protocol 1 and Sharma et al. (2012) to improve RNA yield from formalin-fixed samples. For protocol 2, tissues in 900 μL Trizol and 100 μL Proteinase K (Qiagen) were incubated overnight at 60°C with agitation and then bead beat once as described above. Following bead beating, the supernatant was moved to a new tube and RNA extraction proceeded using the RNeasy Lipid Tissue Mini Kit as described in protocol 1. To improve access of Proteinase K to tissues, we switched the order of the homogenization and digestion steps in protocol 3. For protocol 3, tissues in 900μL Trizol and 100 μL Proteinase K were bead beat 1–2 times as described in protocol 1, followed by incubation with agitation at 56°C for 15 min, then 80°C for 15 min. Following incubation, the supernatant was moved to a new tube for extraction with the RNeasy Lipid Tissue Mini Kit as described in protocol 1.

Qubit and bioanalyzer

To examine the quality and downstream use of RNA derived from museum specimens, we estimated RNA yield for all samples using Qubit (n = 66; Invitrogen, RNA HS or BR assay; **Supplementary Table 1**), the RNA Integrity Number (RIN) and DV200 (proportion of RNA fragments > 200 nucleotides in length) for 22 representative samples, and 260/280 and 260/230 ratios of RNA purity using NanoDrop, Thermo Fisher Scientific, Waltham, MA, United States for 56 representative samples. RIN and DV200 are a measures of RNA degradation and were quantified using the Bioanalyzer RNA 6000 Pico (Agilent, Santa Clara, CA, United States) Eukaryote Total RNA analysis following the manufacturer's instructions (**Supplementary Figures 1**, **2**). We compared RNA purity (i.e., 260/280 and 260/230 ratios) across preservation methodologies and, within formalin-fixed samples, across extraction protocols using one-way ANOVA. We used Tukey's HSD to compare all groups to each other if a significant difference was detected. Friedman's test was used to compare RNA purity between intestine samples used in optimization of the RNA extraction protocol from formalin-fixed samples.

Screening for mammalian and viral RNA using qPCR

For qPCR, we synthesized cDNA from RNA extractions using the ProtoScript II First Strand cDNA Synthesis Kit (New England Biolabs, Ipswich, MA, united States) using the Randomized Primer Mix and following the manufacturer's instructions. We confirmed the presence of mammalian RNA in 22 representative samples by targeting a 100 bp region of the 16S rRNA gene using universal mammalian primers (Tillmar et al., 2013) using the SsoAdvanced SYBR Green Supermix (BioRad, Hercules, CA, United States), following the manufacturer's instructions for a 20 μL final reaction volume. Reactions were incubated at 95°C for 30 s followed by 40 cycles of 95°C for 15 s and 58°C for 30 s. We included negative and positive controls (*Leontopithecus rosalia*, *Callithrix geoffroyi*, *Choloepus didactylus,* and *Desmodus rotundus*) with each assay.

We screened 29 samples (derived from 15 unique bats) for viruses in the subgenus *Sarbecoronavirus* by targeting the N gene region (HKU-N; primers and probe from Chu et al., 2020) and more broadly for alpha- and betacoronaviruses by targeting the RdRp gene region (RdRP; primers and probe I and probe III

TABLE 1 Optimization of RNA extraction from formalin-fixed tissues.

USNM	Duplicate	Tissue	Proteinase K	No. rounds bead beating	Conc. RNA extraction (ng/μL)	RIN	DV200
583864	a	Intestine	N	2	Too low	2.5	<30%
583864	b	Lung	N	2	Too low		
583866	a	Intestine	N	2	Too low	2.6	<30%
583866	b	Lung	N	2	Too low		
583873	a	Intestine	N	2	Too low		
583873	b	Lung	N	2	Too low		
583877	a	Intestine	N	2	Too low		
583877	b	Lung	N	2	Too low		
Negtive1		NA	N	2	Too low		
Negtive2		NA	N	2	Too low		
583864	c	Intestine	Y	2	5.2	2.5	<30%
583864	d	Lung	Y	2	Too low		
583866	c	Intestine	Y	2	27.6	2.5	<30%
583866	d	Lung	Y	2	Too low		
583873	c	Intestine	Y	2	2.12		
583873	d	Lung	Y	2	Too low		
583877	c	Intestine	Y	2	1.61	2.6	<30%
583877	d	Lung	Y	2	Too low		
Negtive3		NA	Y	2	Too low		
Negtive4		NA	Y	2	Too low		
583864	e	Intestine	Y	1	2.45	2.5	<30%
583864	f	Lung	Y	1	Too low		
583866	e	Intestine	Y	1	3.67	2.6	<30%
583866	f	Lung	Y	1	Too low		
583873	e	Intestine	Y	1	Too low		
583873	f	Lung	Y	1	Too low		
583877	e	Intestine	Y	1	Too low		
583877	f	Lung	Y	1	Too low		
Negtive5		NA	Y	1	Too low		
Negtive6		NA	Y	1	Too low		

Qubit concentrations and RIN quality estimates of RNA extracted from replicate tissue and lung subsamples taken from four formalin-fixed bat vouchers, corresponding to Protocols 1 and 3 in Figure 1.

originally developed by Muradrasoli et al., 2009 and modified by Joffrin et al., 2020). Our sample size is smaller than what has previously been used for screening bats for coronaviruses (Joffrin et al., 2020). For the HKU-N assay, each 25 μL reaction contained 12.5 μL KlearKall Hot Start 2× Master Mix (LGC, Biosearch Technologies, Hoddeston, United Kingdom), 0.5 μM of each forward and reverse primer, 0.2 μM of Cy5-labeled probe, 20 μg BSA, and 2.5 μL cDNA. Reactions were incubated at 95°C for 15 min per manufacturer's instructions, followed by 50 cycles of 95°C for 15 s then 58°C for 45 s. For the RdRp assay, 20 μL were used with each reaction containing 10 μL Luna Universal Probe qPCR 2x Master Mix (New England Biolabs), 0.4 μM of each forward and reverse primer, 0.2 μM FAM-labeled probe I, 0.2 μM HEX-labeled probe III, 20 μg BSA, and 2.5 μL cDNA. Thermal conditions followed Joffrin et al. (2020), with an initial incubation at 95°C for 1 min, followed by 2 cycles of 95°C for 15 s and 56°C for 30 s, 2 cycles of 95°C for 15 s and 54°C for 30 s, 2 cycles of 95°C for 15 s and 52°C for 30 s, and 50 imaged cycles of 95°C for 15 s and 50°C for 30 s. All virus-screening assays were performed in duplicate and included negative and positive controls (IDT Gblocks) with each assay.

RNA sequencing

Library preparation

We sequenced a subset of cDNA libraries to evaluate the composition of extracted products. Library preparation was performed following Hawkins et al. (2016) with a KAPA

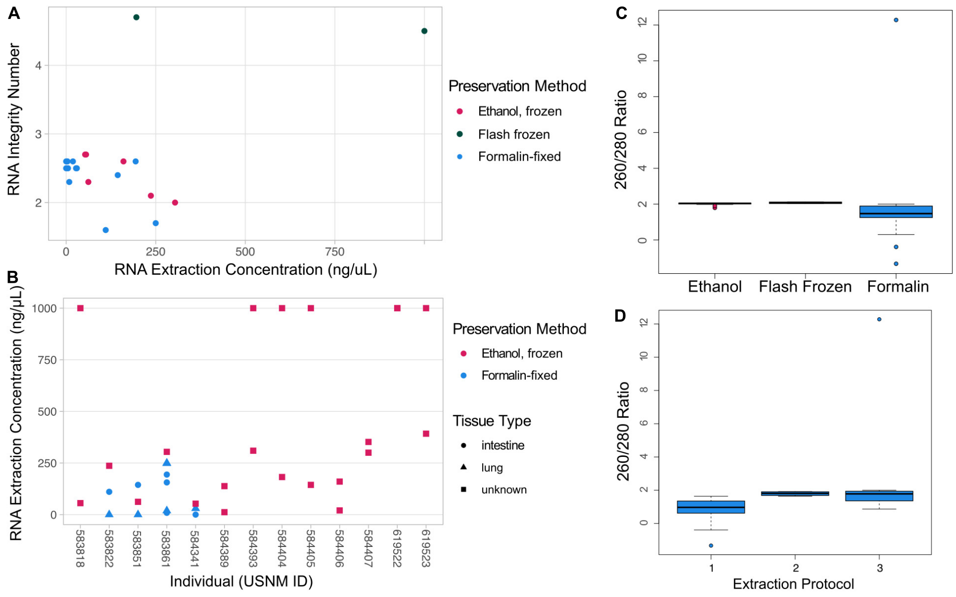


FIGURE 2
RNA Quality and Quantity. **(A)** RNA extraction concentration by RIN for different sample types with preservation method indicated by color. **(B)** Plot of the RNA extraction concentrations for individual bats (NMNH identification number, USNM ID) that had duplicate tissue samples taken. In this plot preservation is mapped as color and tissue type is indicated by shape. In both **(A,B)** plots, samples that have an indicated concentration of 1,000 ng/μL were above the detection threshold of the Qubit HS Kit and should be interpreted as having at least 1,000 ng/μL concentrations. **(C)** Estimate of RNA purity using Nanodrop 260/280 ratio compared to preservation of tissues used for extractions. **(D)** Comparison of Nanodrop 260/280 ratio across protocols used to extract RNA from formalin-fixed tissues. A 260/280 ratio of ∼2 is considered pure RNA.

528 Biosystems LTP Library Preparation kit Roche, Basel, Switzerland, and with UGA iTru style dual indices (Glenn et al., 2016). Due to the low input amount ∼10 μL and low concentration, libraries were amplified for 30 cycles instead of 14 as described in Hawkins et al. (2016). Following amplification, a 1 × SPRI purification (Rohland and Reich, 2012) was performed to remove primer and adapter dimer. Qubit fluorometry and TapeStation, Agilent, Santa Clara, CA, United States traces were completed for each sample to recover both the concentration and size distribution of each library. An Illumina MiSeq 2 × 150 PE v2 run was performed on 14 samples and two controls (**Supplementary Table 1**). Due to the insert length, the run was limited to 75 cycles.

Analysis of RNA sequencing

Samples were demultiplexed by MiSeq Reporter software and adapters were trimmed using cutadapt v.2.4 (Martin, 2011). Sequence quality was assessed before and after trimming using fastqc v.0.11.8 (Andrews, 2010). Trimmed reads were mapped to GenBank reference genomes using STAR v.2.7.10a (Dobin et al., 2013) and Bowtie2 v.2.3.5 (Langmead and Salzberg, 2012) using default parameters. Reads were also mapped to reference transcriptomes when available using Bowtie2. For samples in the family Pteropodidae, reads were mapped to the *Cynopterus brachyotis* genome (GCA_009793145.1; Chattopadhyay et al., 2020) and the *Rousettus aegyptiacus* genome and transcriptome (GCF_014176215.1; Jebb et al., 2020). For samples in the genus *Hipposideros*, reads were mapped to the reference genome and transcriptome of *Hipposideros armiger* (GCF_001890085.1; Dong et al., 2017). For samples in the genus *Rhinolophus,* reads were mapped to the reference genome and transcriptome of *Rhinolophus ferrumequinum* (GCF_004115265.1; Jebb et al., 2020). In instances where multiple libraries were prepared for the sample bat individual (i.e., from replicate tissues), reads were concatenated for mapping. The function *featureCounts* in the Subread package was used to examine the genes to which reads mapped (Liao et al., 2013, 2014).

Metagenomic analysis was performed on sequenced reads to evaluate content using the software MEGAN6 Community Edition (Huson et al., 2007; Bağcı et al., 2021). Prior to taxonomic assignment from MEGAN, the DIAMOND (Buchfink et al., 2015) protein BLAST method was performed on the Smithsonian High Performance Computing Cluster using the Genbank NR database to compare all sequenced reads. Following DIAMOND, the .daa files were imported to MEGAN using the February 2022 database "MEGAN map." All individual files were imported from DIAMOND input and "MEGANIZED" to make RMA6 files. Comparisons were performed between samples where replicates were sequenced as well as across individuals. Sample preservation, tissue type, and specimen were all used in comparisons.

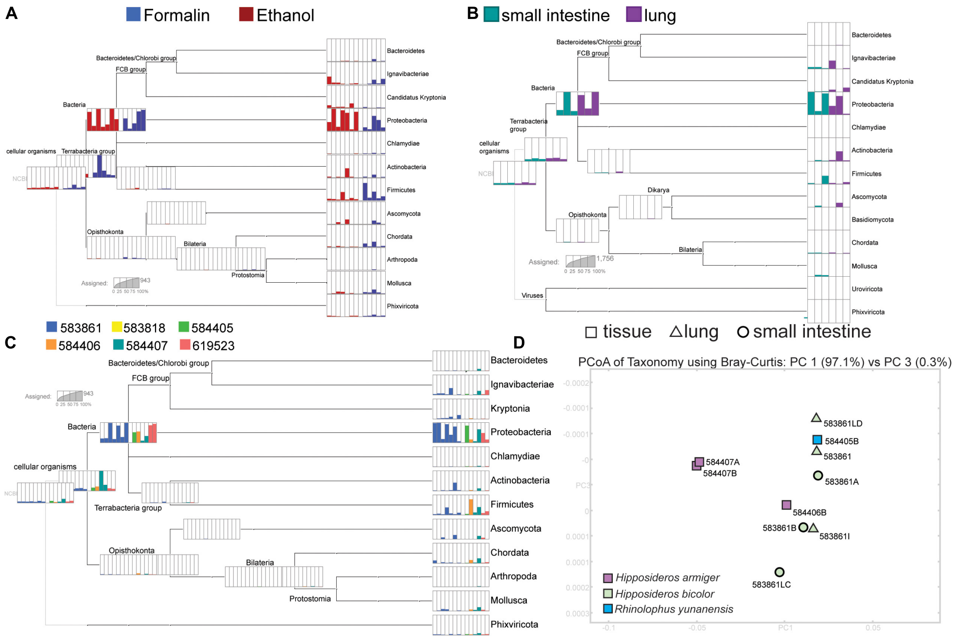


FIGURE 3
Results from the MEGAN analysis. Parts A-C each have separate scales inset in each section to indicate read counts. **(A)** The relative composition of phyla detected from of each sequenced library separated by preservation type. **(B)** Variation in phyla recovered across lung and small intestine replicates from specimen USNM 583861 (*Hipposideros bicolor*). **(C)** The relative proportions of reads identified to phylum from each sequenced library, colored by individual bat (USNM ID). **(D)** A PCoA of all sequenced insectivorous bats, with PC's 1 and 3 shown. Tissue type is indicated by the shape (corresponding to **Figure 2**), and color indicates the species as shown in the bottom left corner of the PCoA.

Results

Optimization of RNA extraction from formalin-fixed specimens

We optimized our RNA extraction protocol using replicate lung and intestine tissues sampled from four formalin-fixed bat individuals (**Table 1**). Extractions from formalin-fixed tissues that were homogenized by bead beating twice prior to Proteinase K digestion more reliably yielded measurable RNA via Qubit quantification than samples homogenized with only one round of bead beating or those extracted without Proteinase K (**Table 1**). No tissue subsamples yielded measurable RNA when we extracted following Protocol 1, and two tissues that yielded measurable RNA when bead beat twice did not yield measurable RNA when bead beat once. In all cases, lung tissues did not yield measurable RNA. This is likely due to variation in how quickly formalin was able to penetrate these tissues compared to the intestine when the bat was originally preserved. We detected RNA from other formalin-fixed lung tissue (i.e., 583861c,f,g; **Supplementary Table 1**), suggesting that the specific preservation protocol used in the field may have substantial impact on resulting RNA preservation. We found that Proteinase K incubation and one additional bead-beating step did not impact RIN quality estimates.

Quality and quantity of RNA from museum specimens

The quality and quantity of RNA extracted from museum specimens varied and was related to preservation method (**Figure 2**). While there was no significant difference in the 260/280 ratio between preservation methods [one-way ANOVA: $F(2, 47) = (0.226)$, $p= 0.799$], formalin-fixed samples typically had a 260/280 ratio lower than the target of 2, likely indicating protein contamination (**Figure 2C**). The 260/280 ratio got closer to the target of 2 when Proteinase K was used in the extraction [i.e., protocols 2 and 3; **Figure 2D**; one-way ANOVA: $F(1, 31) = (3.659)$, $p = 0.0654$]. Estimates of RNA purity using 260/230 ratios are largely consistent with evidence from 260/280 ratios, except that there is a significant difference between the mean 260/230 ratios observed in formalin-fixed and ethanol-F samples [**Supplementary Figure 3**; one-way ANOVA: $F(2, 47) = (8.037)$, $p = 0.001$; Tukey's HSD: $p = 0.0014$, 95% C.I. = (−2.05, −0.44)]. There was no significant difference in RNA purity measured from repeated extractions of intestine tissue from the same formalin-fixed bat individuals [Friedman's test: $\chi^2(2) = 3.4545$, $p = 0.178$]. Flash-frozen samples yielded the highest RIN values (4.5, 4.7) and high RNA quantity, while ethanol-F and formalin-fixed samples typically yielded lower RIN values (1.6–2.7) and low RNA quantity (**Figure 2A**). However, six ethanol-F samples yielded high RNA quantity (> 1 μg/μL). All ethanol-F samples yielded detectable RNA, while many formalin-fixed samples did not. No ethanol-RT samples yielded measurable RNA; these individuals were collected in 1886 and 1888 and are much older than the rest of our samples. RNA quantity varied by individual bat, again suggesting a strong impact of the specific field preservation protocol on RNA persistence (**Figure 2B**). Tissue type did not influence RNA yield (Wilcoxon signed rank test, $p = 0.078$). Six negative controls yielded no detectable RNA, suggesting lab precautions were sufficient to protect even poorly preserved tissues from contamination during extraction.

Downstream usability of RNA from museum specimens

Targeted amplification with qPCR

All samples screened using mammalian universal 16S rRNA primers showed successful amplification with Cq values comparable to those of positive controls (i.e., modern mammal DNA). There was no impact of preservation on qPCR amplification. We did not detect any coronaviruses using our targeted qPCR assays ($n = 15$ bats; $n = 29$ tissues).

RNA sequencing

A small proportion of the RNA-seq data was mappable to bat genomic/transcriptomic references, as is expected for highly degraded libraries. A total of 34,484,466 reads passed sequencing quality filters. Of the reads passing filters 79.3% were demultiplexed (20.7% undetermined reads); the high proportion of undetermined reads is likely from excess sequencing adapters forming dimers. Following adapter trimming, the number of reads was reduced (ranging from 4,867 to 36,790 remaining per sample). Endogenous RNA content, estimated by mapping reads to annotated genomes, ranged from 1.1 to 8.71% using splice-aware mapping (i.e., STAR). Estimates of endogenous content were slightly higher when reads were mapped to reference genomes using Bowtie2, ranging from 3.1 to 18.86%. Overall alignment rate varied less when reads were mapped to transcriptomes, but was less successful (0.73–4.75%). Of the uniquely mapped reads for *Hipposideros armiger* and *H. bicolor*, the most well-represented species in our data, most aligned to the MAT1A gene (**Supplementary Table 2**). Reads were mapped to other genes, but coverage was shallow across the board.

Metagenomic analysis revealed high representation of bacteria in sequenced reads, with some reads mapping to mammals and viruses (**Figure 3**). There was large variation in the representation of different bacterial taxa between replicates, which did not correspond to preservation (**Figure 3A**) or tissue type (**Figure 3B**). Ordination of metagenome communities indicated differentiation between bat species (**Figure 3D**). A low proportion of reads mapped to taxa not likely represented in our sample, possibly as a consequence of the short read length or biases from the GenBank NR database.

Discussion

Museum specimens are an underutilized resource in building foundational knowledge of zoonotic viruses (Colella et al., 2021; Thompson et al., 2021), other emerging infectious diseases (Talley et al., 2015; Byrne et al., 2019), and host gene expression responses to environmental change. These specimens can be used to screen a broad range of host species for pathogens that would be difficult to sample in the wild, track the host and geographic occurrence of pathogens through time, and gain historical snapshots of host and pathogen evolution (Colella et al., 2021; Thompson et al., 2021). However, methods have not been optimized for deriving RNA from natural history specimens, limiting the use of these specimens in RNA virus screening. Here, we present an optimized protocol for extracting RNA from formalin-fixed specimens and explore the quality and quantity of RNA that can be derived from museum specimens, extending the value and possible uses of these specimens.

The RNA extracted from museum specimens is highly degraded, but usable for downstream applications, including qPCR and sequencing (**Figure 3** and **Supplementary Table 2**). Following recommendations from FFPE protocols (Krafft et al., 1997; Sharma et al., 2012), we found that incubation of formalin-fixed samples with Proteinase K following thorough homogenization improves RNA yield (**Table 1**). RNA from formalin-fixed and ethanol-F samples is highly degraded, but may include persistent mRNA as shown by our recovery of bat gene transcripts in RNA-seq data. We did not detect viral RNA in RNA-seq data or targeted qPCR methods, possibly due to the small number of bat individuals screened for viruses in our study. We found that a high proportion of RNA is of bacterial origin, which is likely due to persistence of rRNA in these degraded samples and may also be reflective of database bias, as bacterial rRNA is well-represented on GenBank. We suggest that highly degraded samples may be better suited for targeted approaches, like RT-PCR and qPCR (Castello et al., 1999; Fanning et al., 2002; Worobey et al., 2016).

Age and field preservation are the most important factors influencing the quantity and quality of RNA derived from museum specimens (**Figure 2**). Flash-frozen samples had high RIN values compared to ethanol-F and formalin-fixed samples. While these RIN values are lower than typically targeted for contemporary tissues (RIN > 7), these samples are likely still valuable for RNA-seq applications or other more targeted approaches. Within the ethanol-F and formalin-fixed samples, there was variation in RNA quantity between individual bats, which may indicate lasting impact of field-based preservation protocol. For example, the amount of time between when a bat was euthanized and when its tissues were sampled and preserved has a large impact on quality and quantity of RNA remaining in those tissues (Camacho-Sanchez et al., 2013). The details of in-the-field preservation method matter for the quantity of RNA derived from these samples and should be viewed as valuable specimen metadata. However, this type of metadata is not often recorded in enough detail to tease apart the preservation, storage, and handling of a specimen.

The practice of storing multiple tissue types within one sample tube, a common practice in field mammalogy, is not ideal for viral zoonotic disease screening and gene expression studies. In many instances, viruses are known to aggregate differentially across tissue types and gene expression studies often seek to compare expression profiles between tissues. Long-term storage of different tissue types in the same vial can make it difficult to separate and differentiate them as tissues do not maintain distinct morphology through long periods of storage, even at cryogenic temperatures. While separating tissues into individual tubes has its own limitations (i.e., space, sample tracking), we suggest that, when possible, storing each tissue type in an individual tube may improve the value of these tissues for pathogen screening and gene expression studies.

We find that museum specimens are a valuable source of RNA, even in cases where tissues have not been preserved with RNA in mind. This finding broadens the use of historical specimens in pathogen detection to include viruses. Further work is needed to examine the persistence of mRNA compared to rRNA in these specimens. However, findings from aRNA research provide evidence that mRNA may be maintained under specific conditions through thousands of years (Schmitt et al., 2018). Through efforts to derive new information from existing specimens, we continue to reaffirm the value of natural history collections and the necessity of expanding and maintaining these critical scientific resources.

Data availability statement

The original contributions presented in the study are included in the article/**Supplementary material**, further inquiries can be directed to the corresponding author/s. Raw sequence data has been uploaded to NCBI SRA database. Available at: https://www.ncbi.nlm.nih.gov/sra/PRJNA838638.

Author contributions

KS, MH, CM-W, MM, RF, JM, and MC designed the study. MH, CM-W, MM, RF, JM, and MC secured funding. KS, MH, MF, and CM-W collected and analyzed the data. All authors contributed to writing and revising the manuscript.

Funding

KS was funded by the Smithsonian Institution Biodiversity Genomics Postdoctoral Fellowship and George E. Burch Postdoctoral Fellowship in Theoretical Medicine. CM-W was funded by the Robert and Arlene Kogod Secretarial Scholar's donation.

Acknowledgments

We thank Suzan Murray for input early in the project and Smithsonian Channel's Mission Critical program for featuring this research in the film “Virus Hunting: Caves to COVID. We also thank the Smithsonian Institution's Office of the Undersecretary of Science and Research and the Smithsonian Channel's Mission Critical program for research funding. Gratitude is extended to Nancy McInerney (Smithsonian Conservation Biology Institute, Center for Conservation Genomics), Katie Murphy (NMNH Laboratory of Analytical Biology), Chris Huddleston and Daniel DiMichele (NMNH Biorepository), and Darrin Lunde (NMNH Division of Mammals) for assistance in tissue acquisition and data generation. Some of the computations in this manuscript were conducted on the Smithsonian High Performance Cluster (SI/HPC; https://doi.org/10.25572/SIHPC).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fevo.2022.953131/full#supplementary-material

References

Andrews, S. (2010). *FastQC: A Quality Control Tool for High Throughput Sequence Data*. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc (accessed February 1, 2022).

Bağcı, C., Patz, S., and Huson, D. H. (2021). DIAMOND+MEGAN: fast and easy taxonomic and functional analysis of short and long microbiome sequences. *Curr. Protoc.* 1:e59. doi: 10.1002/cpz1.59

Becker, D. J., Albery, G. F., Sjodin, A. R., Poisot, T., Bergner, L. M., Chen, B., et al. (2022). Optimising predictive models to prioritise viral discovery in zoonotic reservoirs. *Lancet Microbe* [Epub ahead of print]. doi: 10.1016/S2666-5247(21)00245-7

Buchfink, B., Xie, C., and Huson, D. H. (2015). Fast and sensitive protein alignment using Diamond. *Nat. Methods* 12, 59–60.

Byrne, A. Q., Vredenburg, V. T., Martel, A., Pasmans, F., Bell, R. C., Blackburn, D. C., et al. (2019). Cryptic diversity of a widespread global pathogen reveals expanded threats to amphibian conservation. *Proc. Natl. Acad. Sci. U. S. A.* 116, 20382–20387. doi: 10.1073/pnas.1908289116

Camacho-Sanchez, M., Burraco, P., Gomez-Mestre, I., and Leonard, J. A. (2013). Preservation of RNA and DNA from mammal samples under field conditions. *Mol. Ecol. Resour.* 13, 663–673.

Castello, J. D., Rogers, S. O., Starmer, W. T., Catranis, C. M., Ma, L., Bachand, G. D., et al. (1999). Detection of tomato mosaic tobamovirus RNA in ancient glacial ice. *Polar Biol.* 22, 207–212.

Chattopadhyay, B., Garg, K. M., Ray, R., Mendenhall, I. H., and Rheindt, F. E. (2020). Novel de Novo Genome of *Cynopterus brachyotis* Reveals Evolutionarily Abrupt Shifts in Gene Family Composition across Fruit Bats. *Genome Biol. Evol.* 12, 259–272. doi: 10.1093/gbe/evaa030

Childs, J. E., Ksiazek, T. G., Spiropoulou, C. F., Krebs, J. W., Morzunov, S., Maupin, G. O., et al. (1994). Serologic and genetic identification of *Peromyscus maniculatus* as the primary rodent reservoir for a new hantavirus in the southwestern United States. *J. Infect. Dis.* 169, 1271–1280. doi: 10.1093/infdis/169.6.1271

Chu, D. K. W., Pan, Y., Cheng, S. M. S., Hui, K. P. Y., Krishnan, P., Liu, Y., et al. (2020). Molecular Diagnosis of a Novel Coronavirus (2019-nCoV) Causing an Outbreak of Pneumonia. *Clin. Chem.* 66, 549–555.

Colella, J. P., Bates, J., Burneo, S. F., Camacho, M. A., Carrion Bonilla, C., Constable, I., et al. (2021). Leveraging natural history biorepositories as a global, decentralized, pathogen surveillance network. *PLoS Pathog.* 17:e1009583. doi: 10.1371/journal.ppat.1009583

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., et al. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 29, 15–21. doi: 10.1093/bioinformatics/bts635

Dong, D., Lei, M., Hua, P., Pan, Y.-H., Mu, S., Zheng, G., et al. (2017). The Genomes of Two Bat Species with Long Constant Frequency Echolocation Calls. *Mol. Biol. Evol.* 34, 20–34. doi: 10.1093/molbev/msw231

Düx, A., Lequime, S., Patrono, L. V., Vrancken, B., Boral, S., Gogarten, J. F., et al. (2020). Measles virus and rinderpest virus divergence dated to the sixth century BCE. *Science* 368, 1367–1370. doi: 10.1126/science.aba9411

Fanning, T. G., Slemons, R. D., Reid, A. H., Janczewski, T. A., Dean, J., and Taubenberger, J. K. (2002). 1917 avian influenza virus sequences suggest that the 1918 pandemic virus did not acquire its hemagglutinin directly from birds. *J. Virol.* 76, 7860–7862. doi: 10.1128/jvi.76.15.7860-7862.2002

Fordyce, S. L., Ávila-Arcos, M. C., Rasmussen, M., Cappellini, E., Romero-Navarro, J. A., Wales, N., et al. (2013). Deep sequencing of RNA from ancient maize kernels. *PLoS One* 8:e50961. doi: 10.1371/journal.pone.0050961

Glenn, T. C., Nilsen, R. A., Kieran, T. J., Sanders, J. G., Bayona-Vásquez, N. J., Finger, J. W. Jr., et al. (2016). Adapterama I: universal stubs and primers for 384 unique dual-indexed or 147,456 combinatorially-indexed Illumina libraries (iTru & iNext). *bioRxiv* [Preprint]. doi: 10.1101/049114

Hawkins, M. T. R., Hofman, C. A., Callicrate, T., McDonough, M. M., Tsuchiya, M. T. N., Gutiérrez, E. E., et al. (2016). In-solution hybridization for mammalian mitogenome enrichment: pros, cons and challenges associated with multiplexing degraded DNA. *Mol. Ecol. Resour.* 16, 1173–1188. doi: 10.1111/1755-0998.12448

Huson, D. H., Auch, A. F., Qi, J., and Schuster, S. C. (2007). MEGAN analysis of metagenomic data. *Genome Res.* 17, 377–386.

Jebb, D., Huang, Z., Pippel, M., Hughes, G. M., Lavrichenko, K., Devanna, P., et al. (2020). Six reference-quality genomes reveal evolution of bat adaptations. *Nature* 583, 578–584. doi: 10.1038/s41586-020-2486-3

Joffrin, L., Goodman, S. M., Wilkinson, D. A., Ramasindrazana, B., Lagadec, E., Gomard, Y., et al. (2020). Bat coronavirus phylogeography in the Western Indian Ocean. *Sci. Rep.* 10:6873. doi: 10.1038/s41598-020-63799-7

Kleindorfer, S., and Sulloway, F. J. (2016). Naris deformation in Darwin's finches: experimental and historical evidence for a post-1960s arrival of the parasite *Philornis downsi*. *Glob. Ecol. Conserv.* 7, 122–131.

Krafft, A. E., Duncan, B. W., Bijwaard, K. E., Taubenberger, J. K., and Lichy, J. H. (1997). Optimization of the isolation and amplification of RNA from formalin-fixed, paraffin-embedded tissue: the armed forces institute of pathology experience and literature review. *Mol. Diagn.* 2, 217–230. doi: 10.1054/MODI00200217

Langmead, B., and Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nat. Methods* 9, 357–359. doi: 10.1038/nmeth.1923

Liao, Y., Smyth, G. K., and Shi, W. (2013). The subread aligner: Fast, accurate and scalable read mapping by seed-and-vote. *Nucleic Acids Res.* 41:e108. doi: 10.1093/nar/gkt214

Liao, Y., Smyth, G. K., and Shi, W. (2014). featureCounts: An efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* 30, 923–930. doi: 10.1093/bioinformatics/btt656

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet J.* 17, 10–12. doi: 10.1089/cmb.2017.0096

Muradrasoli, S., Mohamed, N., Hornyák, A., Fohlman, J., Olsen, B., Belák, S., et al. (2009). Broadly targeted multiprobe QPCR for detection of coronaviruses: coronavirus is common among mallard ducks (*Anas platyrhynchos*). *J. Virol. Methods* 159, 277–287. doi: 10.1016/j.jviromet.2009.04.022

Rohland, N., and Reich, D. (2012). Cost-effective, high-throughput DNA sequencing libraries for multiplexed target capture. *Genome Res.* 22, 939–946.

Schmitt, C. J., Cook, J. A., Zamudio, K. R., and Edwards, S. V. (2018). Museum specimens of terrestrial vertebrates are sensitive indicators of environmental change in the Anthropocene. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 374:20170387. doi: 10.1098/rstb.2017.0387

Sharma, M., Mishra, B., Vandana Saikia, U. N., Bahl, A., Ratho, R. K., et al. (2012). Ribonucleic acid extraction from archival formalin fixed paraffin embedded myocardial tissues for gene expression and pathogen detection. *J. Clin. Lab. Anal.* 26, 279–285. doi: 10.1002/jcla.21518

Shaw, B., Burrell, C. L., Green, D., Navarro-Martinez, A., Scott, D., Daroszewska, A., et al. (2019). Molecular insights into an ancient form of Paget's disease of bone. *Proc. Natl. Acad. Sci. U. S. A.* 116, 10463–10472. doi: 10.1073/pnas.1820556116

Smith, O., Clapham, A., Rose, P., Liu, Y., Wang, J., and Allaby, R. G. (2014). A complete ancient RNA genome: identification, reconstruction and evolutionary history of archaeological Barley Stripe Mosaic Virus. *Sci. Rep.* 4:4003. doi: 10.1038/srep04003

Smith, O., Dunshea, G., Sinding, M.-H. S., Fedorov, S., Germonpre, M., Bocherens, H., et al. (2019). Ancient RNA from Late Pleistocene permafrost and historical canids shows tissue-specific transcriptome survival. *PLoS Biol.* 17:e3000166. doi: 10.1371/journal.pbio.3000166

Talley, B. L., Muletz, C. R., Vredenburg, V. T., Fleischer, R. C., and Lips, K. R. (2015). A century of *Batrachochytrium dendrobatidis* in Illinois amphibians (1888–1989). *Biol. Conserv.* 182, 254–261.

Thompson, C. W., Phelps, K. L., Allard, M. W., Cook, J. A., Dunnum, J. L., Ferguson, A. W., et al. (2021). Preserve a Voucher Specimen! The Critical Need for Integrating Natural History Collections in Infectious Disease Studies. *MBio* 12, e2698–e2620. doi: 10.1128/mBio.02698-20

Tillmar, A. O., Dell'Amico, B., Welander, J., and Holmlund, G. (2013). A universal method for species identification of mammals utilizing next generation sequencing for the analysis of DNA mixtures. *PLoS One* 8:e83761. doi: 10.1371/journal.pone.0083761

Valitutto, M. T., Aung, O., Tun, K. Y. N., Vodzak, M. E., Zimmerman, D., Yu, J. H., et al. (2020). Detection of novel coronaviruses in bats in Myanmar. *PLoS One* 15:e0230802. doi: 10.1371/journal.pone.0230802

Worobey, M., Watts, T. D., McKay, R. A., Suchard, M. A., Granade, T., Teuwen, D. E., et al. (2016). 1970s and "Patient 0" HIV-1 genomes illuminate early HIV/AIDS history in North America. *Nature* 539, 98–101. doi: 10.1038/nature19827

Yates, T. L., Mills, J. N., Parmenter, C. A., Ksiazek, T. G., Parmenter, R. R., Vande Castle, J. R., et al. (2002). The Ecology and Evolutionary History of an Emergent Disease: hantavirus Pulmonary Syndrome. *Bioscience* 52, 989–998.

Frontiers in Ecology and Evolution

TYPE Brief Research Report
PUBLISHED 17 August 2022
DOI 10.3389/fevo.2022.984056

A comparative analysis of extraction protocol performance on degraded mammalian museum specimens

Melissa T. R. Hawkins[1]*, Mary Faith C. Flores[1], Michael McGowen[1] and Arlo Hinckley[1,2]

[1]Department of Vertebrate Zoology, National Museum of Natural History, Smithsonian Institution, Washington, DC, United States, [2]Departamento de Zoología, Universidad de Sevilla, Seville, Spain

Check for updates

OPEN ACCESS

EDITED BY
Jonathan J. Fong,
Lingnan University, China

REVIEWED BY
Paúl M. Velazco,
American Museum of Natural History,
United States
Guinevere O. U. Wogan,
Oklahoma State University,
United States

*CORRESPONDENCE
Melissa T. R. Hawkins
hawkinsmt@si.edu

SPECIALTY SECTION
This article was submitted to Phylogenetics, Phylogenomics, and Systematics, a section of the journal Frontiers in Ecology and Evolution

RECEIVED 01 July 2022
ACCEPTED 01 August 2022
PUBLISHED 17 August 2022

CITATION
Hawkins MTR, Flores MFC, McGowen M and Hinckley A (2022) A comparative analysis of extraction protocol performance on degraded mammalian museum specimens. *Front. Ecol. Evol.* 10:984056. doi: 10.3389/fevo.2022.984056

The extraction of nucleic acids is one of the most routine procedures used in molecular biology laboratories, yet kit performance may influence the downstream processing of samples, particularly for samples which are degraded, and in low concentrations. Here we tested several commercial kits for specific use on commonly sampled mammalian museum specimens to evaluate the yield, size distribution, and endogenous content. Samples were weighed and had approximately equal input material for each extraction. These sample types are typical of natural history repositories ranged from 53 to 130 years old. The tested protocols spanned spin-column based extractions, magnetic bead purification, phenol/chloroform isolation, and specific modifications for ancient DNA. Diverse types of mammalian specimens were tested including adherent osteological material, bone and teeth, skin, and baleen. The concentration of DNA was quantified via fluorometry, and the size distributions of extracts visualized on an Agilent TapeStation. Overall, when DNA isolation was successful, all methods had quantifiable concentrations, albeit with variation across extracts. The length distributions varied based on the extraction protocol used. Shotgun sequencing was performed to evaluate if the extraction methods influenced the amount of endogenous versus exogenous content. The DNA content was similar across extraction methods indicating no obvious biases for DNA derived from different sources. Qiagen kits and phenol/chloroform isolation outperformed the Zymo magnetic bead isolations in these types of samples. Statistical analyses revealed that extraction method only explained 5% of the observed variation, and that specimen age explained variation (29%) more effectively.

KEYWORDS

museomics, degraded DNA, high throughput sequencing, bone, skin, baleen, osteological tissue

Introduction

High throughput sequencing (HTS) has revolutionized the ability to recover genomic DNA from many unconventional sources. Most ancient DNA (aDNA) studies have been published since the first high throughput sequencer was available in 2008 (Knapp and Hofreiter, 2010). As such, it became more tangible to obtain nucleic acid sequences from samples which had historically performed poorly with standard Sanger sequencing and polymerase chain reaction (PCR) (Paabo et al., 2004). Degraded samples which did not yield high molecular weight DNA (fragments <1,000 bp), particularly benefited from this technology various starting sources can be considered degraded DNA (feces, eDNA, etc.); however, the focus of this work is dry mammalian museum collections.

Since the exponential decrease in sequencing cost, museum collections have become invaluable sources of degraded samples for genetic and genomic analyses. Natural history repositories house millions of specimens around the world and contain both temporally and geographically wide-ranging specimens for inclusion in genetic studies (Rowe et al., 2011; Bi et al., 2013; Holmes et al., 2016; Lopez et al., 2020; Buckner et al., 2021; Card et al., 2021; Colella et al., 2021). Museum specimens also allow for endangered, extinct, or elusive species to be represented when fresh tissues are not available (Ho and Gilbert, 2010; Fabre et al., 2014; Brüniche-Olsen et al., 2018; White et al., 2018). Additionally, by optimizing methodology for museum specimens, genomic signatures can be generated from type specimens, the individual specimen (or series) from which species descriptions are generated. This is important for the study of taxonomy as well as conservation and biodiversity (Guschanski et al., 2013; Chomicki and Renner, 2015; Zedane et al., 2016; Raxworthy and Smith, 2021).

Despite being beneficial for using such material to recover genetic signatures, the resulting DNA molecules are in low copy number and concentration as well as highly fragmentary (Burrell et al., 2015). Here we test several types of DNA extraction including phenol/chloroform, silica membrane, and magnetic bead isolation to determine if one method is superior for recovering DNA from degraded mammalian museum specimens. In addition to standard quality metrices (DNA concentration, size distribution, etc.) shotgun sequencing was performed to evaluate if any extraction methods appeared to bias the amount of endogenous versus exogenous DNA in each sample as assessed by metagenomic analyses.

Methods

Samples

A set of mammalian museum specimens were selected to represent common sources of nucleic acids from non-tissue-based museum holdings. A total of 17 samples were included in various comparisons of extraction protocols. First, 12 samples were extracted across three different extraction kits/protocols with approximately the same input mass per sample per extraction (see Table 1). In order to make extractions as comparable as possible all samples were weighed on the same scale, digested overnight, and manually processed in the same way across all three treatments. Depending on the type of sample (e.g., bone, dried tissue/osteocrusts, skin, baleen, and teeth) the amount of physical processing varied. For example, the baleen was shaved via a Dremel tool from a 2″ × 2″ square of baleen from the growth plate, and the fine powder was collected and divided into three replicates. Teeth were ground into a fine powder using a mortar and pestle. In contrast, the adherent muscle tissue and bone fragments required less manipulation and were weighed and divided in thirds for each replicate. This included cutting skin with scissors, or breaking osteocrusts and bone with a blade to allow each replicate as similar sample as possible. None of our sample types were subjected to a prewash as most (osteocrusts and bone fragments) are very fragile and the risk of losing sample outweighed the potential benefits of a prewash. Once weighed and placed in a 2.0 ml tube, the samples were broken down with forceps against the wall of the tube. After overnight digestion the samples were vortexed and evaluated for complete lysis of tissue. If large pieces remained, additional Proteinase K was added, and more physical manipulation of the tissues was performed with sterile instruments (particularly cutting up the skin into smaller fragments). After adding more Proteinase K the samples were vortexed and placed back into the shaker/incubator for another 1–2 h. Specific details for each kit are provided below.

Extraction kits

The first protocol included using minor modifications (detailed below) to a Qiagen QIAamp DNA extraction kit. First, a Qiagen QIAamp DNA Mini Kit (#51306) was used to extract DNA following the manufacturer's protocol (180 μl ATL plus 20 μl Proteinase K) with an overnight digestion at 56°C in a shaking incubator. The final elution step was done twice, with 50 μl of AE buffer added to the membrane, incubated, then centrifuged for a total elution volume of 100 μl.

The second kit used in this comparison was the Zymo DNA/RNA Viral MagBead Kit (#R2140). This kit was selected due to the increased recovery of low concentration, short insert size nucleic acids. The digestion was modified to have 10 μl of 10 mg/ml Proteinase K added (optional for viral studies of extracellular molecules, but part of the manufacturer's protocol for tissue usage) with an overnight digestion at 56°C. The standard elution volume of 30 μl was retained. While performing this extraction it became clear that undigested tissues could potentially interfere with the magnetic bead steps, so upon magnetic separation any remaining undigested particles

TABLE 1 A summary of the 12 samples extracted across the QIAamp®, Zymo®, and phenol/chloroform extractions.

	Species	Catalog #	Sample type	Year collected	Weight per replicate (if possible)				Qubit concentration (ng/µl)					
					1	2	3	Total sample weight	Replicate:1 Qiagen (ng/µl)	Total DNA	2-Zymo (ng/µl)	Total DNA	3-Phenol/ Chloroform (ng/µl)	Total DNA
1	*Propithecus diadema*	USNM 063348	Skin clip	1895	0.015	0.015	0.017	0.048	2.74	274	0.534	16.02	3.82	382
2	*Propithecus diadema*	USNM 063348	Bone fragments	1895	0.002	0.002	0.003	0.0087	too low	N/A	Too low	N/A	Too low	N/A
3	*Propithecus diadema*	USNM 063349	Osteocrust	1895	0.0095	0.015	0.011	0.0312	0.26	26	0.312	9.36	2.44	**244**
4	*Propithecus diadema*	USNM 063349	Skin clip	1895	0.005	0.005	0.0035	0.0103	0.658	65.8	0.826	24.78	1.99	**199**
5	*Callosciurus nigrovittatus*	USNM 154902	Bone fragments	1909	0.003	0.003	0.003	0.009	too low	N/A	0.122	3.66	0.764	**76.4**
6	*Callosciurus notatus*	USNM 101686	Osteocrust	1900	0.004	0.004	0.004	0.013	0.212	21.2	0.538	16.14	0.736	**73.6**
7	*Callosciurus notatus*	USNM 196712	Osteocrust	1913	0.006	0.009	0.007	0.022	0.454	45.4	2.4	72	4.08	**408**
8	*Callosciurus notatus*	USNM 145405	Osteocrust	1911	0.006	0.006	0.006	0.021	0.874	87.4	26.4	792	10.5	**1050**
9	*Balaeonptera physalus*	USNM 617703	Baleen	1948	0.059	0.0102	0.0153	0.085	19.7	**1970**	43.2	1296	16.4	1640
10	*Balaeonptera physalus*	USNM 617538	Baleen	1948	0.06	0.0202	0.08	0.168	9.54	954	2.22	66.6	46	**4600**
11	*Orcaella brevirostris*	FMNH 99613	Tooth	1966	0.069	0.089	0.097	0.216	Too low	N/A	0.1	3	Too low	N/A
12	*Orcaella brevirostris*	MCZ 21929	Tooth	1892	0.0993	0.1074	0.1005	0.3709	Too low	N/A	Too low	N/A	Too low	N/A
13	*extraction negative*	NA	NA	NA	NA	NA	NA	NA	Too low	N/A	Too low	N/A	Too low	N/A
								Mean:	1.28	128.45	1.56	46.82	4.11	411.33
								stdev	6.52	651.8	14.62	438.65	13.75	1375.06

Each specimen contains details about the species, museum catalog number, type of sample (skin clip, bone, osteocrust, baleen, or tooth), the year collected, weight across each replicate and total sample weight, and the recovered DNA concentration across extraction replicated and total recovered DNA. Total DNA was calculated as the protocols had variable elution volumes. Summary statistics are shown below the extraction types, with the mean, standard deviation, and minimum calculated. Samples in bold represent extraction replicate with the highest recovered DNA concentration.

TABLE 2 Sample details for the ancient DNA extraction method (Hagan et al., 2020) and the modified QIAamp protocol which included DTT.

	Catalog #	Species	Sample type	Weight 1	Weight 2	Total sample weight	Ancient DNA protocol	Total DNA (ng/μl)	Qiamp with DTT added (ng/μl)	Total DNA
1	USNM 114629	*Callosciurus saturatus*	Skin	10	10.8	21.3	0.464	**27.84**	0.234	23.4
2	USNM 063347	*Propithecus diadema*	Skin	12.2	13.2	24.5	0.124	7.44	0.396	**39.6**
3	USNM 256833	*Ratufa bicolor*	Osteocrusts	27.2	27.2	54.4	1.28	76.8	6.26	**626**
4	USNM 257721	*Ratufa bicolor*	Osteocrusts	26.8	26.7	53.5	3.8	228	2.82	**282**
5	USNM 488162	*Ratufa bicolor*	Osteocrusts/nasal turbinates	22	22	44	49.8	2,988	70	**7,000**
6	Extraction blank			NA	NA	NA	Too low	N/A	Too low	N/A
						Mean	1.69	101.62	2.58	258.09
						min	0.12	7.44	0.23	23.4
						max	49.8	2,988	70	7,000
						stdev	19.4	1,163.76	27.12	2,711.68

Species, catalog number, sample type, weight of samples (miligrams) and the recovered Qubit concentrations (in ng/μl), and total DNA are shown. Summary statistics of the mean, minimum, maximum, and standard deviation are shown. Samples in bold represent extraction replicate with the highest recovered DNA concentration.

were removed via pipette tips. It is noteworthy that some museum specimen sample types (osteocrusts and bone) are difficult to fully lyse, but in spin column and phenol/chloroform extractions the particles do not interfere with subsequent steps.

The third extraction protocol followed a standard phenol/chloroform isolation as described in Hawkins et al. (2016), and originally detailed in Leonard et al. (2000). Briefly, an extraction buffer was prepared containing Tris + EDTA (100×), EDTA (0.5 M), NaCl (5 M), and water plus 10% SDS, DTT (400 mg/ml), and 20 μl of 10 mg/ml Proteinase K. The extraction buffer composition can be found in the **Supplementary material**. After samples were placed in the 1× extraction buffer they were incubated in a shaker/incubator overnight at 56°C. The next day two washes of phenol were used to separate proteins from nucleic acids followed by a chloroform wash. Top aqueous layers were removed and placed in clean tubes at each step and the final product was washed with 2 ml of water (1 ml washes performed twice) via an Amicon Ultra-4 centrifugal column and centrifuged at 3,300 RPM for 9 min. After the final spin, the volume was evaluated in each Amicon filter and an additional 8–12 min of centrifugation was performed to yield approximately 100 μl of purified DNA.

Ancient DNA MinElute modified protocol versus Qiamp DNA extraction kit

Ancient DNA laboratories have published various modifications to Qiagen spin column-based DNA extractions (Dabney and Meyer, 2019; Hagan et al., 2020; Xavier et al., 2021; Dehasque et al., 2022). Unfortunately we were not able to compare the 12 samples across four extraction protocols without resulting in extremely limited input for each replicate. However, as the aDNA protocol uses similar chemistry, reagents, and procedures as the QIAamp DNA extraction kit, we extracted five additional samples with both the aDNA protocol and the standard QIAamp protocol described above, with one minor modification (the addition of 20 μl of DTT 400 mg/ml to the extraction buffer) since the aDNA protocol also includes the usage of DTT. Samples spanned skin, adherent muscle tissue and nasal turbinates to determine if DNA concentration or size distribution of recovered molecules varied between protocols. The samples used for this comparison are provided in **Table 2**.

The aDNA protocol used here adhered closely to that described in Hagan et al. (2020), specifically "Method B," which uses a Zymo reservoir attached to a MinElute Spin Column to allow for a larger volume of Qiagen Buffer PB (binding buffer) to be mixed with sample lysate following overnight digestion. From the published protocol we made a few modifications for better comparison to our QIAamp extractions. We added 1 ml of 0.5 M EDTA to the weighed sample, then added 100 μl proteinase K, and placed in a shaker incubator overnight at 37°C (recommended for aDNA). After the overnight digest samples were checked, 50 µl of additional proteinase K was added, as well as 20 µl of DTT (Sigma). Samples were vortexed and placed back in the shaker/incubator for two additional hours. We did not add a second milliliter of EDTA, and as such we did not use the full 13 ml of Buffer PB detailed in Hagan et al. (2020), and instead used 7 ml Buffer PB. The 50 ml conical tubes had the Zymo reservoir added, with the MinElute spin column snugly attached. Then a 5 ml tube had the lid removed and was placed inside the Falcon tube to help hold the spin column in place. Finally, the 5 ml tube had a hole drilled in the side with a Dremel to allow the buffer to flow out during centrifugation and prevent contaminating the MinElute column with flowthrough. The remaining spin and wash steps were the same as detailed in Hagan et al. (2020). Two washes of Buffer EB were done with 30 µl each wash for a final elution volume of 60 µl.

Quantification and visualization of extracts

After extractions were completed, each sample was quantified via a Qubit (Invitrogen) dsDNA High Sensitivity Kit (# Q33230). From the Qubit concentrations the total DNA yield from each extraction was calculated. This was necessary as the different methods resulted in varying volumes of DNA. We also generated electropherograms of each extract to visualize if the protocols recovered varying size distributions based on the protocol for DNA isolation. An Agilent TapeStation 4200 was used with a high-sensitivity kit to evaluate the size distribution of the extracted DNA.

Sequencing and analysis

Dual indexed sequencing libraries were generated for all replicates and all extracts in this study with an Illumina Library Preparation—Kapa Biosystems Kit (Catalog # KK8232). Qubit values were used to pool samples in equimolar ratios with all replicates across this study. Once this was completed qPCR was performed on this pool of all samples with replication and dilution and the average size fragments of 250 bp (from a final TapeStation electropherogram) on an ABI ViiA7 using KAPA library quantification kit (#KK4824).

To evaluate recovered DNA content, shotgun sequencing was performed on an Illumina HiSeq X with an insert length of 150 bp PE. Sequencing was performed at Admera Health Biopharma Services, NJ, United States. Reads were demultiplexed via the BaseSpace Hub. Standard sequence quality filtering was performed with BBDUK (Bushnell, 2014) version 38.84 via Geneious Prime (Biomatters). All Illumina adapters were removed, and low-quality reads were removed from both ends (quality score minimum 20), and reads under 10 bp in length were discarded.

Metagenomic analysis of sample contents was performed to evaluate any biases between the extraction methods. Using the program MEGAN (Huson et al., 2016), we evaluated the major composition of identifiable taxa from each extract and used these results to identify any patterns across samples. DIAMOND (Bağcı et al., 2021) BLAST was performed prior to importing results to MEGAN, and the DIAMOND "*.daa" were "meganized" and transformed into "*.RMA6" files to make comparisons across the replicates for each individual. Samples were compared using "MeganMap" from February 2022, and trees were made at the Phyla level. Comparisons were made across each sample by the different extraction method.

Statistical analyses

We used paired t-tests (two tailed $p = 0.05$) to determine if the quantified differences were statistically significant across extract method. We also performed linear regressions across the sample type, extraction type, age of sample, against recovered DNA concentration to determine which relationships explained our results better.

Results

DNA recovery across protocols

After comparing the three major kit types [silica-membrane (Qiagen), phenol/chloroform and magnetic beads (Zymo)] we found that most extractions recovered quantifiable DNA across the museum specimens. Samples that recovered detectable concentrations via Qubit recovered quantifications across all three extractions, and the samples which were too low to measure via Qubit were generally not quantified in any extraction method. When quantifiable, the minimum DNA concentrations ranged from 3 ng (Zymo) to 76.3 ng (phenol/chloroform) across replicates. The maximum across methods ranged from 1,296 to 4,600 ng, again with Zymo recovering the least and phenol/chloroform the most. The average yields were: 128.45 ng (Qiagen), 46.82 ng (Zymo), and 411.33 ng (phenol/chloroform) with the standard deviation high across all extraction methods. All quantification results are shown in **Table 1**. Despite the wide range of DNA concentrations, the yields were not statistically significant in any t-tests, details are provided in the **Supplementary material**. Linear regressions were performed on total DNA yield versus extraction method, as well as starting sample type. DNA extraction method only explained about 5% of the observed variation in DNA concentrations (**Supplementary**

Table 1, $p = 0.25$). Alternatively, the starting template explained about 16% of the variation (Supplementary Table 2, $p = 0.04$). When a regression was performed using sample age and recovered DNA concentration, a total of 29.4% of the variation was explained by age (Supplementary Table 3, $p = 0.003$).

Ancient DNA versus Qiagen kit

The recovered DNA concentrations between the aDNA and the QIAamp extraction kits were similar. For these samples recovery did not increase with use of the aDNA protocol. In fact, all concentrations were higher with the modified Qiagen kit. The average yield from the aDNA protocol was 101.62 ng versus 258.09 ng with the modified Qiagen extraction. The minimum for each was 7.44 ng (aDNA) and 23.4 ng (Qiagen), and maximum yield was 2,988 and 7,000 ng for aDNA and Qiagen kits respectively. The highest yield for both kits in this comparison was from the same sample (USNM 488162). Neither a one nor two tailed *t*-test was significant for this comparison. Details of all these comparisons can be found in Table 2. Regressions showed that extraction method was not significant, and explained 5% of the variation (Supplementary Table 4, $p = 0.4$), sample type was also not significant, and explained 18% (Supplementary Table 5, $p = 0.22$). Sample age explained 61% of the variation and was the only significant comparison for these samples (Supplementary Table 6, $p = 0.007$).

Sequencing and length variation

Despite samples being pooled in equimolar ratios, the number of reads varied across replicates. Quality filtering results recovered a general trend in which each sample had the same percentage of reads removed. However, the Zymo extracts appeared to have more reads removed than either Qiagen or phenol/chloroform. The MEGAN analysis showed that the proportion of endogenous and exogenous sequences was fairly consistent across replicates (Figure 1). Due to the variable components in each extract some amount of stochasticity between replicates was expected. Plots of all samples can be found in the Supplementary material.

The second set of extraction comparisons was between an aDNA protocol (Hagan et al., 2020) and modifications to a Qiagen QIAamp extraction protocol. These samples recovered quite different size distributions as evaluated on the TapeStation (summarized in **Supplementary Table 7**). The aDNA protocol recovered TapeStation traces for all five samples, ranging in size from 50 to >850 bp. The modified Qiagen extraction only recovered traces for three samples, one of which was too large to determine the average size (obscured the upper marker). The sequencing results were also different between extraction methods, with the aDNA protocol retaining a much higher percentage of starting reads (average of 79%) than the modified Qiagen protocol (55%). Details of quality filtering are shown in Table 3. Subsequent research evaluating biases in length recovery across kits is warranted. It is also worth noting that despite many replicates lacking a visible peak, all samples yielded usable sequence.

Discussion

Variation in extraction methods

The Qiagen and phenol/chloroform protocols performed better than the Zymo kit in all metrics evaluated here. The Zymo kit was the most cost effective but is also marketed toward intracellular viruses and may not be geared for optimization of degraded vertebrate DNA. We modified the protocol as detailed by the manufacturer to lyse tissues but it appears to do so at the cost of losing the smallest fragments. The magnetic bead-based protocol is a quick and less toxic method for nucleic acid isolation, and is scalable for large tissue samples; however, it is not an effective protocol for the museum specimens tested here. Our results mirror those of McDonough (McDonough et al., 2018) supporting a similar DNA yield, fragment size and percentage of starting reads for Qiagen and phenol/chloroform.

One caveat of this study is that it is impossible to know if subsamples are truly representative across replicates. For example, one weighed subsample may contain bone of different density, and thus provide a better template for extraction than another. Similarly, the ratio of endogenous/contaminant DNA may vary among different osteocrust or skin clip subsamples of the same specimen. To our knowledge there is no way to use real-world specimens and account for this variable.

Concentration does not necessarily imply target DNA (Straube et al., 2021). Sample USNM 488162 had the highest DNA yield but also the lowest percentage of raw reads. Previous studies found that the highest concentration was associated with the highest amount of contamination in some samples (Campana et al., 2012; McDonough et al., 2018). The two included baleen samples also had a large amount of exogenous DNA (particularly bacteria). Specimen age has the largest impact on DNA yield. This is in contrast with other studies which showed no correlation between these variables (McDonough et al., 2018; Straube et al., 2021), but in corroboration with other studies (Yuan et al., 2021).

The average DNA yield and size from the ancient protocol was smaller than the modified Qiagen extraction, but the average percentage of reads after trimming was higher for the former than the latter, particularly for the two oldest samples (USNM 114629 and USNM 063347) which had a shorter fragment size (50–60 bp). The aDNA protocol does appear to do exactly as intended by retaining the smallest fragments; however, depending on the quantity and value of individual samples, the modified Qiagen kit may perform nearly as well while removing the smallest fragments. The aDNA protocol is more time consuming and expensive making the Qiagen kit a nearly equivalent option. Individual projects should perform their own cost-benefit analyses to determine which methods to employ.

Sequencing

Shotgun sequencing of samples provides valuable data to understand DNA content within a degraded sample. Studies of aDNA have found that the percent of endogenous DNA varies tremendously based on the preservation, age, and handling conditions following excavation from substrate (Dabney and Meyer, 2019). The most common phyla represented across taxa were Proteobacteria and Chordata. However, our results indicate that extraction methods had less of an effect on DNA recovery than either pre-extraction preservation or contamination. Other considerations for selection of an extraction protocol may be more important than sample contents as it does not appear that a bias was recovered across the extraction methods tested here. From this study we show that the Qiagen extractions (both standard and modified) recovered nearly the same profile of endogenous DNA as the more expensive phenol/chloroform, and more labor intensive aDNA protocol. Samples of the same specimen derived from different input (Figure 1; skin versus osteocrust) also show variation in the amount of endogenous DNA, with osteocrust out performing skin in this sample.

Cost difference across protocols

A comparison of kit and reagent costs is an important factor when budgeting for a grant and planning a project. Cost is important, as is efficiency, especially when using limited starting material which can be difficult or impossible to replace or resample. The cost of the kits tested here varied substantially and are detailed in the Supplementary material. Extractions ranged from $3.05 per sample to $11.54 per sample. Overall, the cheapest per sample cost was the Zymo kit, which ultimately had the poorest recovery in terms of concentration and appeared to lose more short fragments based on the TapeStation electropherograms.

Phenol/chloroform extractions were the most expensive ($8.39–11.54), especially when the Amicon Ultra-4 spin columns were used to wash the sample. It is possible to use different more cost-effective centrifugal columns (such as a Qiagen MinElute column), which reduces the cost to approximately $8.40 per sample. In any case, future studies should assess the efficiency of this and other modified phenol/chloroform protocols with alternative cost-effective washing steps following chloroform precipitation.

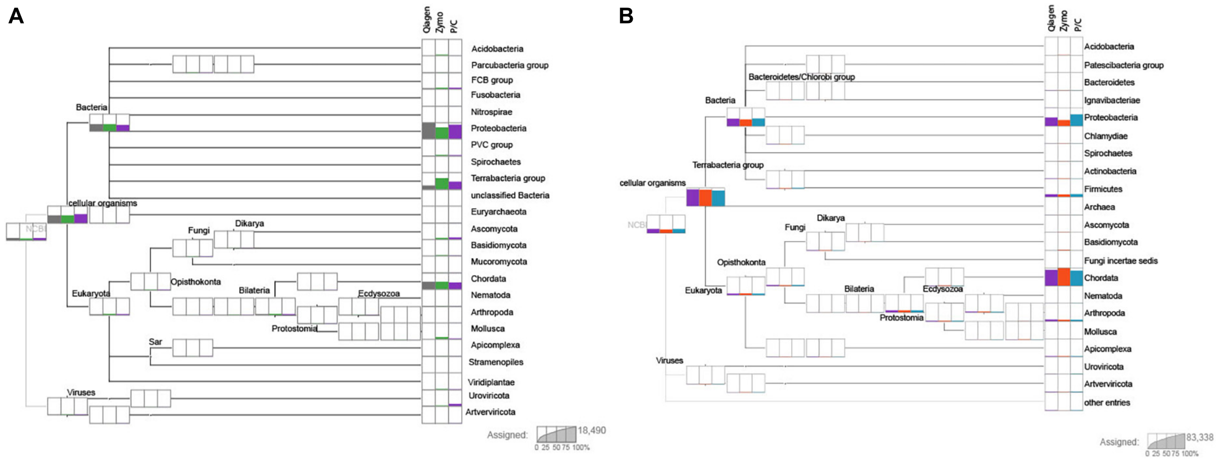


FIGURE 1

Two individuals after performing DIAMOND (Bağcı et al., 2021) BLAST and importing through MEGAN (Huson et al., 2016). Sample 1 (USNM 063348, *Propithecus diadema*, skin clip) and sample 3 (USNM 063349, *P. diadema*, osteocrusts) are shown in panels **(A,B)**, respectively. The columns at each terminal represent the extraction method, with Qiagen shown first, then Zymo and finally P/C for phenol/chloroform. The number of reads of each group are proportionally represented on each plot. Resolution is at the level of the phylum. Note the stochasticity between extraction method, but generally similar proportions of each phylum are represented across methods. Most samples recovered a high amount of Chordata and Proteobacteria with other phyla having more variable proportions. Individual plots for each sample can be found in the **Supplementary material**.

TABLE 3 Quality filtering results following BBDUK (Dehasque et al., 2022).

Species	Catalog #	Sample type	Year collected	Raw reads								
				Qiagen	Reads following trimming	% after trim	Zymo		Reads following trimming	P/C		Reads following trimming
1 *Propithecus diadema*	USNM 063348	Skin clip	1895	1,789,744	566,702	32	4,601,708	193,946	4%	2,696,330	858,070	32%
2 *Propithecus diadema*	USNM 063348	Bone fragments	1895	2,104,764	643,908	31	3,391,274	1,628,704	48%	3,427,410	791,212	23%
3 *Propithecus diadema*	USNM 063349	Osteocrusts	1895	2,454,632	2,070,534	84	2,295,604	2,147,874	94%	4,302,074	3,918,248	91%
4 *Propithecus diadema*	USNM 063349	Skin clip	1895	3,175,338	1,384,718	44	3,569,950	583,680	16%	3,647,742	968,074	27%
5 *Callosciurus nigrovittatus*	USNM 154902	Bone fragments	1909	4,362,226	242,968	6	1,352,868	360,320	27%	3,554,152	1,569,516	44%
6 *Callosciurus notatus*	USNM 101686	Osteocrusts	1900	3,509,020	1,893,724	54	3,390,104	346,568	10%	691,686	306,438	44%
7 *Callosciurus notatus*	USNM 196712	Osteocrusts	1913	4,071,290	3,583,970	88	4,282,886	3,578,628	84%	3,623,668	2,415,302	67%
8 *Callosciurus notatus*	USNM 145405	Osteocrusts	1911	4,568,240	4,266,038	93	5,992	3,904	65%	447,326	390,658	87%
9 *Balaeonptera physalus*	USNM 617703	Baleen	1948	2,246,600	1,992,748	89	3,086,536	2,724,516	88%	3,107,516	2,704,140	87%
10 *Balaeonptera physalus*	USNM 617538	Baleen	1948	2,217,512	2,109,102	95	3,714,458	3,160,742	85%	2,413,430	1,740,710	72%
11 *Orcaella brevirostris*	FMNH 99613	Tooth	1966	3,052,618	209,266	7	485,582	370,162	76%	3,462,662	479,232	14%
12 *Orcaella brevirostris*	MCZ 21929	Tooth	1892	3,384,956	561,364	17	3,774,084	166,242	4%	4,076,568	115,646	3%
					Average:	53		Average:	50%		Average:	49%

	Species	Catalog #	Sample type	Year collected	Ancient protocol			Qiagen (modified)		
					Raw reads	After trimming		Raw reads	After trimming	
1	*Callosciurus saturatus*	USNM 114629	Skin	1902	4,509,928	3,022,340	67%	4,373,798	194,016	4%
2	*Propithecus diadema*	USNM 063347	Skin	1895	2,659,176	1,295,370	49%	4,166,987	441,518	11%
3	*Ratufa bicolor*	USNM 256833	Osteocrusts	1931	2,067,418	1,944,496	94%	2,325,696	2,213,562	95%
4	*Ratufa bicolor*	USNM 257721	Osteocrusts	1931	1,615,948	1,552,336	96%	382,810	309,308	81%
5	*Ratufa bicolor*	USNM 488162	Osteocrusts/nasal turbinates	1969	106,960	97,006	91%	50,152	41,704	83%
						Average:	79%		Average:	55%

Raw reads were paired and trimmed with the specified quality filtering parameters. The number of reads remaining after filtering are shown for each extraction method, as well as the percentage of remaining reads. P/C represent phenol/chloroform extractions.

Based on the cost and endogenous DNA recovery, we show here that QIAamp kits perform well on mammalian museum specimens. With a price point at $3.76 per sample it is difficult to beat the savings, and the kits are well vetted and can be processed in a higher throughput on a QIAcube robot if throughput is a concern (although historical materials should always be performed in small batches with negative controls).

Future prospects

Single-tube and single-strand library preparation methods have been shown to yield better results than other approaches when working with highly degraded DNA (Gansauge et al., 2017; Carøe et al., 2018). Future research should evaluate the performance of combinations of DNA extractions and library preparation methods. The most expensive yet efficient phenol/chloroform extraction might yield better results in combination with a single-tube library preparation, since it has less bead cleaning steps than the KAPA library preparation protocol and will potentially lose fewer short fragments that are retained by the Amicon column during the extraction. However, if funding is limited and savings on DNA extraction are desirable, the QIAamp DNA extraction was fairly comparable to phenol/chloroform, at a much lower price point (under $4 an extraction) versus ∼$11.50 when using the Amicon filters. The phenol/chloroform protocol with a Qiagen spin column clean up saves approximately $3 per sample. Finally, the aDNA protocol did retain the smallest fragments, but it does not appear overly important for samples derived from museum specimens, as the fragmentary sequences are much more difficult to reconstruct, and with a price point of over $8.50/sample. The aDNA protocol did recover higher proportions of Chordata sequences in four of the five tested samples, so individual decisions should be made when determining the best methods to use for each project weighing the extraction cost, and availability of samples.

Data availability statement

The datasets presented in this study can be found in online repositories. The data can be found at: https://figshare.com/s/260c150dcfa53a6f405b.

Ethics statement

Ethical review and approval was not required for the animal study because all samples from this study involved long preserved museum specimens.

Author contributions

MH conceived the study and analyzed the data. MF, MH, and AH performed the laboratory work. All authors contributed to the article and approved the submitted version.

Funding

MH funded the study with discretionary startup funds provided by the Smithsonian Institution. AH was supported by a "Margarita Salas" postdoctoral fellowship funded by the Ministerio de Universidades de España and the European Union "NextGenerationEU."

Acknowledgments

We thank Darrin Lunde, Ingrid Rochon from USNM, Adam Ferguson from FMNH, and Mark Omura from MCZ. Jeff Hunt, Katie Murphy, and Carrie Craig are thanked for assistance with laboratory logistics from the Laboratory of Analytical Biology. Some of the computations in this manuscript were conducted on the Smithsonian High Performance Cluster (SI/HPC), Smithsonian Institution (https://doi.org/10.25572/SIHPC).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fevo.2022.984056/full#supplementary-material

References

Bağcı, C., Patz, S., and Huson, D. H. (2021). DIAMOND+MEGAN: Fast and easy taxonomic and functional analysis of short and long Microbiome sequences. *Curr. Protoc.* 1:e59. doi: 10.1002/cpz1.59

Bi, K., Linderoth, T., Vanderpool, D., Good, J. M., Nielsen, R., and Moritz, C. (2013). Unlocking the vault: Next-generation museum population genomics. *Mol. Ecol.* 22, 6018–6032. doi: 10.1111/mec.12516

Brüniche-Olsen, A., Jones, M. E., Burridge, C. P., Murchison, E. P., Holland, B. R., and Austin, J. J. (2018). Ancient DNA tracks the mainland extinction and island survival of the Tasmanian devil. *J. Biogeogr.* 45, 963–976. doi: 10.1111/jbi.13214

Buckner, J. C., Sanders, R. C., Faircloth, B. C., and Chakrabarty, P. (2021). The critical importance of vouchers in genomics. *eLife* 10:e68264. doi: 10.7554/eLife.68264

Burrell, A. S., Disotell, T. R., and Bergey, C. M. (2015). The use of museum specimens with high-throughput DNA sequencers. *J. Hum. Evol.* 79, 35–44. doi: 10.1016/j.jhevol.2014.10.015

Bushnell, B. (2014). *BBMap: A fast, accurate, splice-aware aligner*. Available online at: https://www.osti.gov/biblio/1241166 (accessed June 27, 2022).

Campana, M. G., Lister, D. L., Whitten, C. M., Edwards, C. J., Stock, F., Barker, G., et al. (2012). Complex relationships between mitochondrial and nuclear DNA preservation in historical DNA extracts. *Archaeometry* 54, 193–202. doi: 10.1111/j.1475-4754.2011.00606.x

Card, D. C., Shapiro, B., Giribet, G., Moritz, C., and Edwards, S. V. (2021). Museum genomics. *Annu. Rev. Genet.* 55, 633–659. doi: 10.1146/annurev-genet-071719-020506

Carøe, C., Gopalakrishnan, S., Vinner, L., Mak, S. S. T., Sinding, M.-H. S., Samaniego, J. A., et al. (2018). Single-tube library preparation for degraded DNA. *Methods Ecol. Evol.* 9, 410–419. doi: 10.1111/2041-210X.12871

Chomicki, G., and Renner, S. S. (2015). Watermelon origin solved with molecular phylogenetics including Linnaean material: Another example of museomics. *New Phytol.* 205, 526–532. doi: 10.1111/nph.13163

Colella, J. P., Bates, J., Burneo, S. F., Camacho, M. A., Carrion Bonilla, C., Constable, I., et al. (2021). Leveraging natural history biorepositories as a global, decentralized, pathogen surveillance network. *PLoS Pathog.* 17:e1009583. doi: 10.1371/journal.ppat.1009583

Dabney, J., and Meyer, M. (2019). Extraction of Highly Degraded DNA from Ancient Bones and Teeth. *Methods Protoc.* 1963, 25–29. doi: 10.1007/978-1-4939-9176-1_4

Dehasque, M., Pečnerová, P., Kempe Lagerholm, V., Ersmark, E., Danilov, G. K., Mortensen, P., et al. (2022). Development and Optimization of a Silica Column-Based Extraction Protocol for Ancient DNA. *Genes* 13:687. doi: 10.3390/genes13040687

Fabre, P.-H., Vilstrup, J. T., Raghavan, M., Sarkissian, C. D., Willerslev, E., Douzery, E. J., et al. (2014). Rodents of the Caribbean: Origin and diversification of hutias unravelled by next-generation museomics. *Biol. Lett.* 10:20140266. doi: 10.1098/rsbl.2014.0266

Gansauge, M.-T., Gerber, T., Glocke, I., Korlevic, P., Lippik, L., Nagel, S., et al. (2017). Single-stranded DNA library preparation from highly degraded DNA using T4 DNA ligase. *Nucleic Acids Res.* 45:e79. doi: 10.1093/nar/gkx033

Guschanski, K., Krause, J., Sawyer, S., Valente, L. M., Bailey, S., Finstermeier, K., et al. (2013). Next-generation museomics disentangles one of the largest primate radiations. *Syst. Biol.* 62, 539–554. doi: 10.1093/sysbio/syt018

Hagan, R. W., Hofman, C. A., Hübner, A., Reinhard, K., Schnorr, S., Lewis, C. M., et al. (2020). Comparison of extraction methods for recovering ancient microbial DNA from paleofeces. *Am. J. Phys. Anthropol.* 171, 275–284. doi: 10.1002/ajpa.23978

Hawkins, M. T. R., Leonard, J. A., Helgen, K. M., McDonough, M. M., Rockwood, L. L., and Maldonado, J. E. (2016). Evolutionary history of endemic Sulawesi squirrels constructed from UCEs and mitogenomes sequenced from museum specimens. *BMC Evol. Biol.* 16:80. doi: 10.1186/s12862-016-0650-z

Ho, S. Y., and Gilbert, M. T. P. (2010). Ancient mitogenomics. *Mitochondrion* 10, 1–11. doi: 10.1016/j.mito.2009.09.005

Holmes, M. W., Hammond, T. T., Wogan, G. O., Walsh, R. E., LaBarbera, K., Wommack, E. A., et al. (2016). Natural history collections as windows on evolutionary processes. *Mol. Ecol.* 25, 864–881. doi: 10.1111/mec.13529

Huson, D. H., Beier, S., Flade, I., Górska, A., El-Hadidi, M., Mitra, S., et al. (2016). MEGAN Community Edition - Interactive Exploration and Analysis of Large-Scale Microbiome Sequencing Data. *PLoS Comput. Biol.* 12:e1004957. doi: 10.1371/journal.pcbi.1004957

Knapp, M., and Hofreiter, M. (2010). Next Generation Sequencing of Ancient DNA: Requirements, strategies and perspectives. *Genes* 1, 227–243. doi: 10.3390/genes1020227

Leonard, J. A., Wayne, R. K., and Cooper, A. (2000). Population genetics of ice age brown bears. *Proc. Natl. Acad. Sci. U.S.A.* 97, 1651–1654. doi: 10.1073/pnas.040453097

Lopez, L., Turner, K. G., Bellis, E. S., and Lasky, J. R. (2020). Genomics of natural history collections for understanding evolution in the wild. *Mol. Ecol. Resour.* 20, 1153–1160. doi: 10.1111/1755-0998.13245

McDonough, M. M., Parker, L. D., Rotzel McInerney, N., Campana, M. G., and Maldonado, J. E. (2018). Performance of commonly requested destructive museum samples for mammalian genomic studies. *J. Mammal.* 99, 789–802. doi: 10.1093/jmammal/gyy080

Paabo, S., Poinar, H., Serre, D., Jaenicke-Despres, V., Hebler, J., Rohland, N., et al. (2004). Genetic Analyses from Ancient DNA. *Annu. Rev. Genet.* 38, 645–679. doi: 10.1146/annurev.genet.37.110801.143214

Raxworthy, C. J., and Smith, B. T. (2021). Mining museums for historical DNA: Advances and challenges in museomics. *Trends Ecol. Evol.* 36, 1049–1060. doi: 10.1016/j.tree.2021.07.009

Rowe, K. C., Singhal, S., Macmanes, M. D., Ayroles, J. F., Morelli, T. L., Rubidge, E. M., et al. (2011). Museum genomics: Low-cost and high-accuracy genetic data from historical specimens. *Mol. Ecol. Resour.* 11, 1082–1092. doi: 10.1111/j.1755-0998.2011.03052.x

Straube, N., Lyra, M. L., Paijmans, J. L. A., Preick, M., Basler, N., Penner, J., et al. (2021). Successful application of ancient DNA extraction and library construction protocols to museum wet collection specimens. *Mol. Ecol. Resour.* 21, 2299–2315. doi: 10.1111/1755-0998.13433

White, L. C., Mitchell, K. J., and Austin, J. J. (2018). Ancient mitochondrial genomes reveal the demographic history and phylogeography of the extinct, enigmatic thylacine (Thylacinus cynocephalus). *J. Biogeogr.* 45, 1–13. doi: 10.1111/jbi.13101

Xavier, C., Eduardoff, M., Bertoglio, B., Amory, C., Berger, C., Casas-Vargas, A., et al. (2021). Evaluation of DNA extraction methods developed for forensic and ancient dna applications using bone samples of different age. *Genes* 12:146. doi: 10.3390/genes12020146

Yuan, S. C., Malekos, E., and Hawkins, M. T. R. (2021). Assessing genotyping errors in mammalian museum study skins using high-throughput genotyping-by-sequencing. *Conserv. Genet. Resour.* 13, 303–317. doi: 10.1007/s12686-021-01213-8

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

frontiers | Frontiers in Ecology and Evolution

TYPE Methods
PUBLISHED 30 August 2022
DOI 10.3389/fevo.2022.924941

OPEN ACCESS

EDITED BY
Pamela Soltis,
University of Florida, United States

REVIEWED BY
Hugo Fort,
Universidad de la República, Uruguay
Kazuhiro Takemoto,
Kyushu Institute of Technology, Japan

*CORRESPONDENCE
Behnaz Balmaki
bbalmaki@unr.edu

SPECIALTY SECTION
This article was submitted to Phylogenetics, Phylogenomics, and Systematics, a section of the journal Frontiers in Ecology and Evolution

RECEIVED 20 April 2022
ACCEPTED 29 July 2022
PUBLISHED 30 August 2022

CITATION
Balmaki B, Rostami MA, Christensen T, Leger EA, Allen JM, Feldman CR, Forister ML and Dyer LA (2022) Modern approaches for leveraging biodiversity collections to understand change in plant-insect interactions. *Front. Ecol. Evol.* 10:924941. doi: 10.3389/fevo.2022.924941

Modern approaches for leveraging biodiversity collections to understand change in plant-insect interactions

Behnaz Balmaki[1,2]*, Masoud A. Rostami[1,2], Tara Christensen[1,2], Elizabeth A. Leger[1,2], Julie M. Allen[1,2], Chris R. Feldman[1,2], Matthew L. Forister[1,2] and Lee A. Dyer[1,2]

[1]Department of Biology, University of Nevada, Reno, Reno, NV, United States, [2]Museum of Natural History, University of Nevada, Reno, Reno, NV, United States

Research on plant-pollinator interactions requires a diversity of perspectives and approaches, and documenting changing pollinator-plant interactions due to declining insect diversity and climate change is especially challenging. Natural history collections are increasingly important for such research and can provide ecological information across broad spatial and temporal scales. Here, we describe novel approaches that integrate museum specimens from insect and plant collections with field observations to quantify pollen networks over large spatial and temporal gradients. We present methodological strategies for evaluating insect-pollen network parameters based on pollen collected from museum insect specimens. These methods provide insight into spatial and temporal variation in pollen-insect interactions and complement other approaches to studying pollination, such as pollinator observation networks and flower enclosure experiments. We present example data from butterfly pollen networks over the past century in the Great Basin Desert and Sierra Nevada Mountains, United States. Complementary to these approaches, we describe rapid pollen identification methods that can increase speed and accuracy of taxonomic determinations, using pollen grains collected from herbarium specimens. As an example, we describe a convolutional neural network (CNN) to automate identification of pollen. We extracted images of pollen grains from 21 common species from herbarium specimens at the University of Nevada Reno (RENO). The CNN model achieved exceptional accuracy of identification, with a correct classification rate of 98.8%. These and similar approaches can transform the way we estimate pollination network parameters and greatly change inferences from existing networks, which have exploded over the past few decades. These techniques also allow us to address critical ecological questions related to mutualistic networks, community ecology, and conservation biology. Museum collections remain a bountiful source of data for biodiversity science and understanding global change.

KEYWORDS

plant-pollinator, interaction network, pollen analysis, museum collection, convolutional neural network

Introduction

Global change is one of the most pressing issues for modern ecologists, and increases in habitat loss, fragmentation, climate change, invasive species, and pollutants are leading to unprecedented losses of biological diversity and less reticulate ecological networks (Alarcon et al., 2008; Ferrarini et al., 2017; Harrison et al., 2020; Salcido et al., 2020; Wagner et al., 2021). Pollination is one of the essential ecosystem services impacted by global change, but it is difficult to document these impacts without thorough natural history observations of plant-pollinator associations and estimates of network relationships (Seltmann et al., 2017; Balmaki et al., 2022).

Entomopalynology, the study of pollen grains associated with insects, is a relatively new approach developed to track pollination ecology through time and space (Jones and Jones, 2001). This approach has recently received greater attention and has provided more demand for museum specimens because insects collected across different temporal or spatial gradients provide invaluable data for reconstructing networks of insect-pollen interactions. A limited number of studies have used this method to estimate parameters related to bee pollination biology (Silberbauer et al., 2004; Wood et al., 2019). Expanding this approach to other insects that are important pollinators, such as Lepidoptera, can reveal unique aspects of pollen-insect interaction networks, and their sensitivity or resilience to change (Balmaki et al., 2022).

Pollen grains are the common currency of pollination ecology. Insects may consume, passively carry, or actively transport pollen (to a stigma or other plant parts), and pollen grains can cover an insect's body, either passively through the air column, or actively while an insect is feeding on nectar or pollen (Jones, 2012a,b, 2014). Analysis of pollen grains on the body of a pollinator can reveal dietary associations and patterns of floral visitation. Examining pollen grains on pollinators approximates a measure of pollen availability, and with repeated sampling can illustrate changes in plant-pollinator interactions over time. Tracking these changes is key to understanding the effects of environmental change on pollination ecology. Precise and quantitative descriptions of plant-pollinator interactions are required to make inferences about changing interaction networks, and pollination ecosystem services through time, and analysis of pollen on insect specimens is a powerful approach to address this need (Burkle et al., 2013).

Traditional palynology, the study of pollen grains and spores, depends on morphological characters of pollen grains to identify pollen taxa. Typical morphological traits used to distinguish pollen include general shape, polarity, symmetry, apertures, size, and ornamentation. Nevertheless, the morphological similarities of pollen grains make it difficult to effectively use these features to identify pollen species quickly and accurately. In addition, identifying pollen grains under the microscope is time-consuming and expensive, and the results are typically dependent on partly-subjective criteria for identifications that are associated with a relatively high error rate (Gonçalves et al., 2016; Sevillano et al., 2020). Alternatively, pollen metabarcoding is a high-throughput approach that can characterize multiple taxa in a mixed sample, but is frequently unable to resolve lower taxonomic levels, and is not an effective method for estimating abundance (Bell et al., 2017). While some studies using pollen identification only warrant a coarse level of taxonomic resolution (family), most approaches to insect-pollen networks benefit from finer taxonomic resolution, at the level of species. An effective method for pollen identification should be efficient, precise, and accurate, and machine learning approaches are well suited for this goal. Here, we provide an example using convolutional neural networks (CNN) which is a deep learning algorithm that can be part of an integrated approach to collections-based research. The approach should be especially useful for museums with large herbaria and entomological collections, because pollen can be collected from herbarium specimens as well as insects (Daood et al., 2016; Carranza-Rojas et al., 2017; Romero et al., 2020; Polling et al., 2021).

We analyzed a plant-pollinator interaction network using museum specimens collected in the Great Basin Desert and Sierra Nevada Mountains and stored at the University of Nevada, and our goal here is to present these methods and analytical tools to encourage adoption in other collections. The main objectives of these methodological innovations are to quantify historic and contemporary pollen-butterfly interaction networks, and to use this information for hypothesis testing about changes in pollination networks in response to extreme weather events and other commonly measured parameters of global change. This approach will transform the way we quantify pollinator networks and present an efficient alternative to pollen identification that provides reliable species-level accuracy.

With this integrative approach to studying plant-pollinator interactions using museum specimens, it is possible to address important questions in ecology and conservation biology, such as: How have plant-pollinator interaction networks changed over time? Is climate change associated with changes in interaction networks? How do habitat loss, fragmentation, biological invasions, and other disturbances affect these networks? Can we improve accuracy and decrease the time-consuming methods of pollen grain identifications using deep learning?

General methodological approach

Data collection and pollen analysis

The best methods for documenting plant-pollinator species interactions are likely to combine quantitative approaches with well-informed natural history descriptions. Historically, these approaches include flower bagging experiments, observations of floral visitation, and pollen identification from insect specimens as described above. These approaches are rarely combined, and pollination studies are dominated by observational methods and quantitative literature reviews, typically with a focus on flower visitation observations for estimating network parameters (Yamaji and Ohsawa, 2016; Colom et al., 2021; Mendes et al., 2022). Visitation network studies typically consist of observation periods in which the researcher observes and records the visitors to a particular plant in an allotted time period. On its own, this approach falls shorts because it disregards the effectiveness of particular pollinators and treats all floral visitors as pollinators, when some are not (Ballantyne et al., 2015). Additionally, many observation hours over relatively long temporal scales may be required to accurately and adequately characterize these interaction networks (Kaiser-Bunbury et al., 2009). Flower bagging experiments involve isolating inflorescences with bags to assess the effects of pollinator exclusions, and pollination events can be closely monitored upon removal of the bag (Yamaji and Ohsawa, 2016; Aslan et al., 2019). This method is valuable for assessing the effectiveness of individual pollinators, but can be time-consuming, and may be inefficient and impractical for community-level studies.

In recent decades, ecologists have used pollen analysis to study the effects of habitat loss and alteration on pollinators and plants (Silberbauer et al., 2004; Bosch et al., 2009; Jones, 2014; Wood et al., 2019; Balmaki et al., 2022). Pollen collections have typically focused on pollen from sediment or soil cores, but collecting pollen grains directly from the bodies of pollinators is a more recent approach to estimating changes in plant-pollinator interactions (Bosch et al., 2009). In addition, collecting historical ecological data associated with museum specimens can increase the accuracy of pollinator-plant interactions and expand our knowledge of pollination networks through space and time (Kleijn and Raemakers, 2008; Colla et al., 2012; Bartomeus et al., 2013; Balmaki et al., 2022). Natural history museums are underutilized repositories of historical interaction diversity and rapidly declining biodiversity (Johnson et al., 2011; Castillo-Figueroa, 2018; Jones and Daeler, 2018). Data from pollen associated with pollinators stored in museums can be used for the estimation of interaction networks between plants and flower visitors through time and space.

Collecting data from historical museum specimens, especially butterflies, presents a unique set of challenges, particularly with older specimens. Using museum samples precludes us from using the acetolysis technique, in which organic materials, in this case insect tissue, are dissolved to recover pollen from insects and reveal diagnostic characters of pollen grains (Jones, 2014). In order to preserve museum specimens, we use entomological pins under a binocular microscope to manually collect pollen grains from the external surface of pollinators, which can be exacting and delicate work. On Lepidoptera, pollen grains typically aggregate on the proboscis, legs, and compound eyes (**Figure 1**). Pollen grains can be mounted on glass slides by adding two drops of 2000 cs silicone oil volume. Suspension in silicon oil allows for the rotation of pollen grains under a microscope to examine the dimensions and shape of pollen in different orientations (Cushing, 2011). The next step is sealing the slide with a cover slip and nail polish to protect the slides from damage. This method is prevalent among quaternary researchers who make pollen slides from sediment samples in cores for palynology purposes (Cushing, 2011; Balmaki et al., 2019; Riding, 2021). Once pollen slides are prepared, they can serve as reference slides for identification of pollen grains to the genus or species level. Having pollen reference slides from all plant taxa in our study region increases the accuracy of pollen grain identification. A high-resolution light microscope and camera can create detailed images for pollen morphology, which can illustrate the number of apertures, exine sculpture, and internal texture, to analyze and identify pollen grains. In addition, electron microscopes (SEM) can examine the surface structures for pollen identification. **Figure 2** indicates the summary of the procedure, from collecting pollen to analyzing the data.

Network analysis and parameters

It is useful to quantify species interaction networks because of the importance of biotic interactions for ecosystem functions, from primary productivity to community stability, especially in the context of environmental change (Tylianakis et al., 2010;

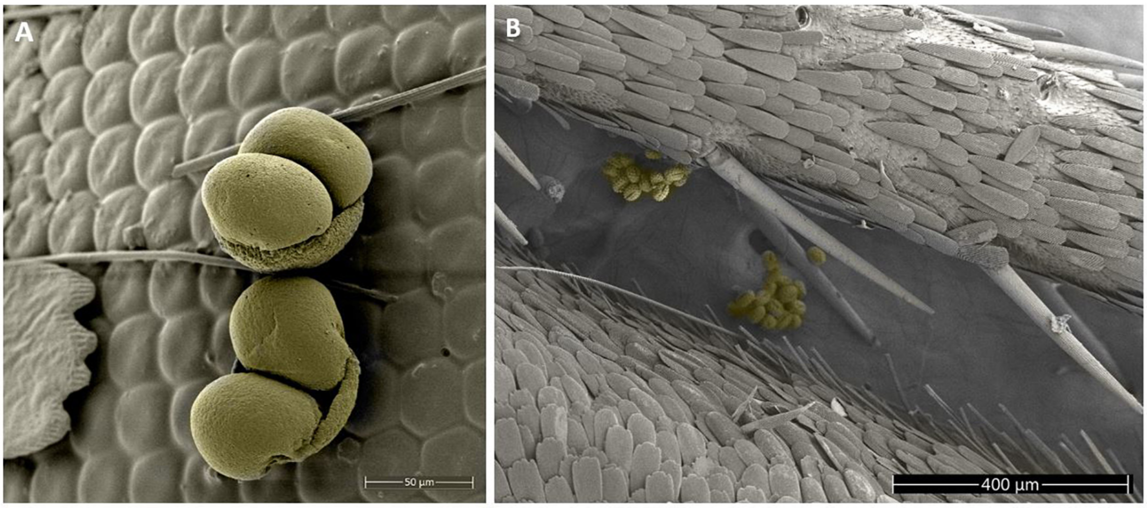


FIGURE 1
Scanning electron microscope images of pollen grains on the legs and eyes of a skipper (*Hesperopsis libya*, Hesperiidae) from the entomological collections of the University of Nevada Reno Museum of Natural History (UNRMNH). **(A)** Pinaceae pollen grains adhered to the butterfly's eye. **(B)** Asteraceae pollen grains on the butterfly leg.

Losapio et al., 2018; Aslan et al., 2019). The documented relationships between interaction diversity and stability of ecological communities are partly a consequence of the number of network links, their relative strength, nestedness, and degree of specialization (Pawar, 2014; Metelmann et al., 2020). Large disturbances, extreme weather events, and continued global change can decrease the number of potential and realized interactions in mutualistic networks (Balmaki et al., 2022). Extending analyses to examine interaction diversity at multiple scales may provide mechanistic insights into the community and ecosystem-level consequences of climate change. Including interaction diversity and network approaches should contribute to predicting how species interactions will change over time in response to global change as well as across different environmental and disturbance gradients, especially if they are used to construct and validate predictive or forecasting models (Strydom et al., 2021).

Typically, plant-pollinator interaction networks are considered as bipartite, or two-sided networks, in which the nodes indicate plant and pollinator taxa, and the edges represent their interactions. Commonly, the width of the edges represents the frequency of interactions, with wider edges representing higher frequencies of interaction. Dozens of network parameters can be used to summarize bipartite and more complex networks; for example, some useful network metrics for community ecology are connectance, nestedness, and network specialization (H2) (Dormann et al., 2009). Connectance represents the number of links between nodes, and it summarizes the number of realized possible connections (Martinez, 1992). Nestedness describes the degree of subsetting that occurs compared to a random network; in other words, nestedness describes the extent to which more specialized interactions form subsets within more generalized interactions (Bascompte et al., 2003; Pawar, 2014). H2 is an index that quantifies the degree of specialization and is useful for comparisons across multiple networks (Blüthgen et al., 2006).

Automation of pollen grain classification

Deep learning as a subset of artificial intelligence is not a new approach, but it has become more popular in the past decade with the advance of technology, including computational power and the availability of large datasets (Wäldchen and Mader, 2018). Deep learning algorithms are computationally expensive, but for researchers who do not have access to appropriate computational resources and high-speed internet to handle large datasets with many parameters, there are platforms such as Colaboratory by Google (Google Colab), which is a Jupyter notebook-based runtime environment that allows running code entirely on the cloud, that can help train large-scale deep learning models using a standard computer. The main purpose of neural networks in deep learning is to receive a set of inputs, perform complex linear and non-linear calculations on them, and provide output to aid classification or provide classic regression parameter estimates. Deep learning is a technique that enables us to train huge and complex datasets, and applies to many fields, including crop or weed detection (Buddha et al., 2019; Afonso et al., 2020), leaf detection (Younis et al., 2020), detection and classification of plant diseases (Geetharamani and Pandian, 2019; Albattah et al., 2022), species identification

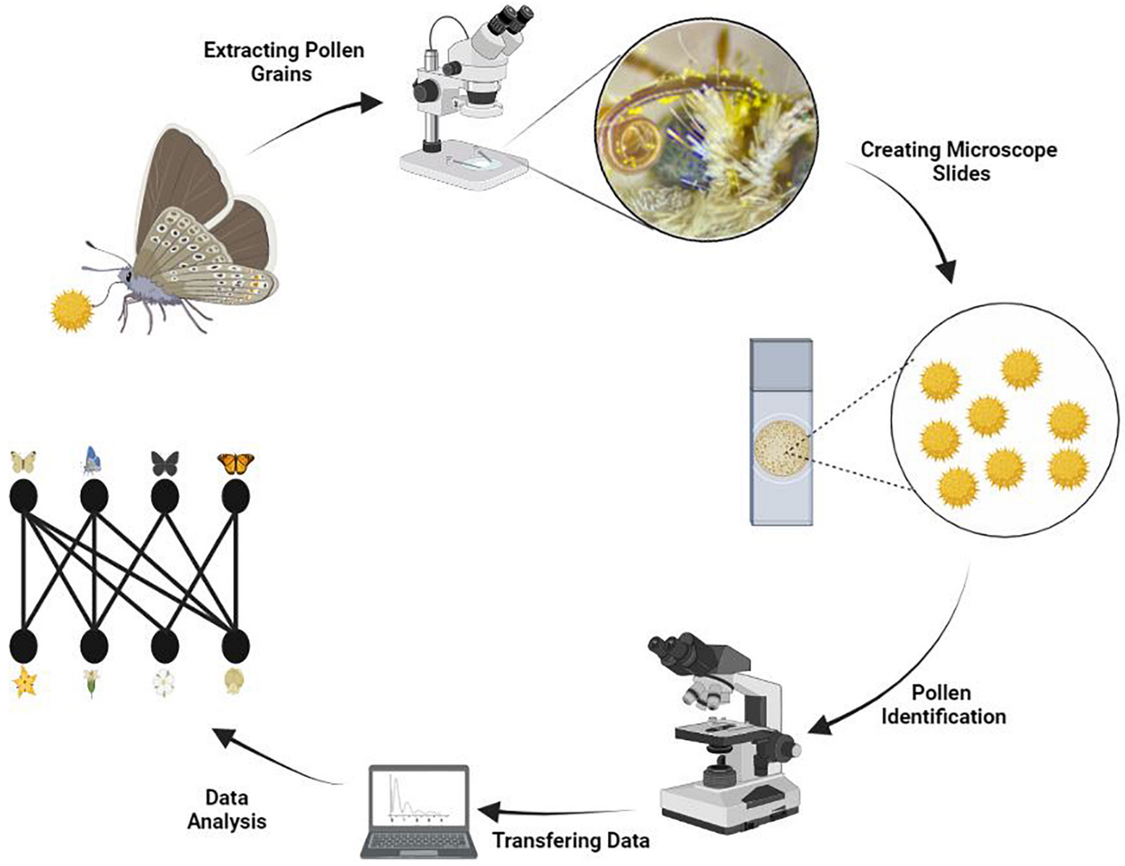


FIGURE 2
Summary of pollen analysis method for plant-pollinator studies. Pollen grains are manually extracted from the insect specimen using an entomology pin under a microscope. The grains are then oriented and slide mounted for pollen identification *via* machine learning methods. The direct associations between insect and pollen are then combined with similar data from several specimens or several species collected at various spatial or temporal scales for examination *via* network analysis (or other downstream analyses).

(Galanty et al., 2021), and animal counts using camera traps (Norouzzadeh et al., 2018, 2021; Wäldchen and Mader, 2018).

Convolutional neural networks (CNN) are utilized for deep learning (e.g., Norouzzadeh et al., 2018; Astolfi et al., 2020; Polling et al., 2021). A CNN model contains multiple layers, including convolutional layers, pooling, and fully connected (FC) layers (**Figure 3**). For example, utilizing a pollen image as input, the first layer would include dimensions such as height, width, and color channels (Red, Green, Blue). The neuron in the first convolutional layer transforms this information into a three-dimensional output, yielding non-linear combinations of the input layer or feature extraction. These learned features are utilized as inputs for the next layer, allowing for pooling and data reduction, and at each step, the next node reclassifies the previous node. Learned features become inputs for statistical models, taking advantage of the hierarchical nature of the input data, and summarizing complex patterns using nested patterns that are smaller and simpler. These approaches have rarely been used for pollen identification (Daood et al., 2016; Khanzhina et al., 2018; Sevillano and Aznarte, 2018; Gallardo-Caballero et al., 2019; Astolfi et al., 2020; Romero et al., 2020; Sevillano et al., 2020; Polling et al., 2021), whether the goal is for identifying allergens in the air column or monitoring change in pollinator-plant interactions through time. Whatever the goal, CNN models are ideal for image classification and will be useful for species-level determinations.

Convolutional neural networks models often achieve prediction capabilities not seen by any other modeling approach (Flagel et al., 2019; Sevillano et al., 2020; Polling et al., 2021). This is because CNN models contain many filters and neural network layers that can extract low and high-level features from images or data matrices. In fact, the CNN method develops algorithms that automatically extract discriminant features from images without human involvement, in contrast to standard statistical approaches, such as ordination (PCA, NMDS) and Support Vector Machine (SVM) analyses, with extraction and preprocessing steps that require user iterations and are time-consuming (O'Mahony et al., 2019; Alzubaidi et al., 2021).

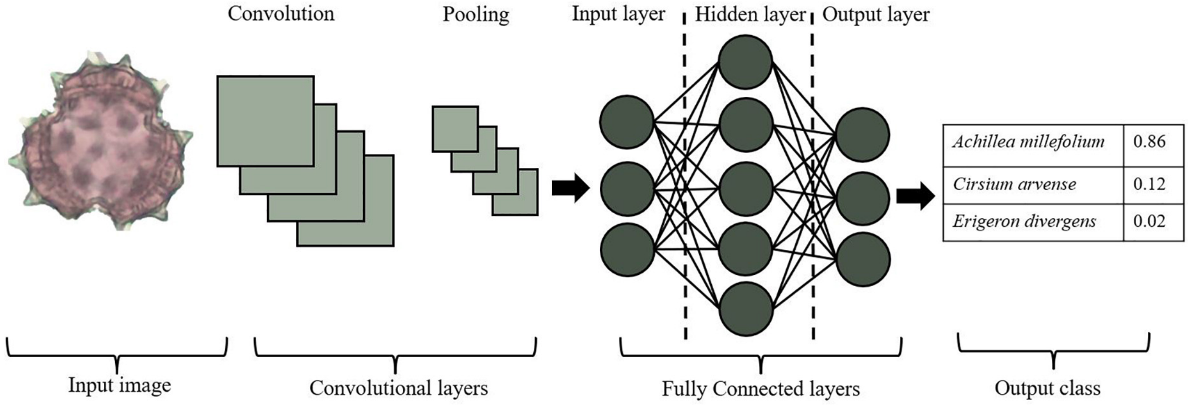


FIGURE 3
Basic convolutional neural network (CNN) architecture, including an input image, convolutional layers (convolution and pooling), fully connected layers, and output classes.

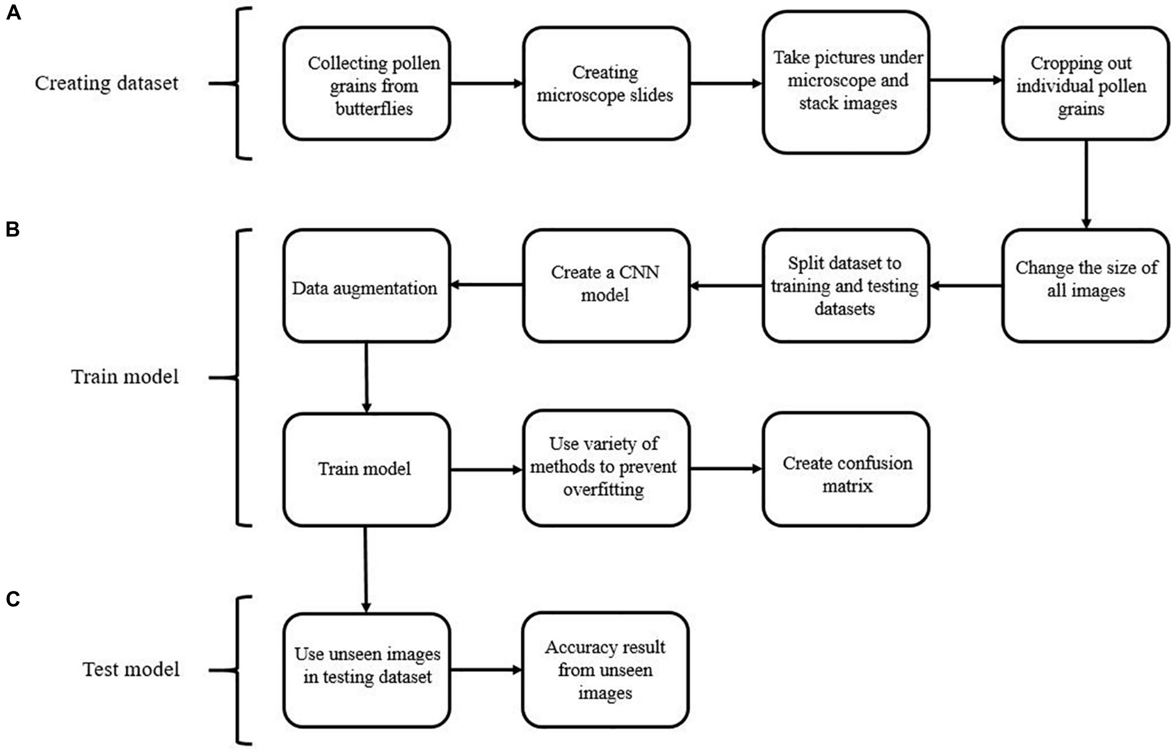

A
Creating dataset
Collecting pollen grains from butterflies
Creating microscope slides
Take pictures under microscope and stack images
Cropping out individual pollen grains
B
Train model
Change the size of all images
Split dataset to training and testing datasets
Create a CNN model
Data augmentation
Train model
Use variety of methods to prevent overfitting
Create confusion matrix
C
Test model
Use unseen images in testing dataset
Accuracy result from unseen images

FIGURE 4
Flowchart showing the pollen image classification process across several steps, including: **(A)** creating the image dataset; **(B)** training the model; **(C)** testing the model.

There are several advantages of CNN compared to traditional supervised machine learning methods. The CNN method often achieves a higher accuracy score in tasks such as image classification and object detection (Viertel and Konig, 2022). The CNN can be re-trained which allows us to utilize it in different custom datasets (O'Mahony et al., 2019).

In the example presented here for identification of pollen from the Great Basin Desert and Sierra Nevada Mountains, two popular transfer learning (pertained models) approaches have been used, including AlexNet and VGG19, to create and train our models and extract the critical features automatically from the pollen images (Krizhevsky et al., 2012;

Simonyan and Zisserman, 2014). AlexNet was initially created to classify millions of images in 1000 categories in ImageNet datasets (Krizhevsky et al., 2012). It takes input images by size 224 × 224 RGB. This method includes five convolutional layers and three fully connected (FC) layers with around 60 million parameters. Through different layers of the CNN network, the first layer extracts the basic features such as color and edges; then, in the deeper layers, the model learns more convoluted features such as spines and pores in pollen grains. After the convolutional layers and extracting the features, AlexNet has three FC layers with 1000 neurons for each category. The output layer in the AlexNet model is interpreted as the probability of an image belonging to each pollen species category. The VGG (Visual Geometry Group) model takes input images with the size of 224 × 224 RGB. This model has five convolutional blocks with a filter size of 3 × 3, a fixed stride size of 1, and each of these convolutional blocks followed by max-pooling with size 2 × 2 with a stride of 2. Also, the VGG has three FC layers, including Rectified Linear Unit (ReLU) and softmax function in the final layer. The main VGG transfer-learning models are VGG16 and VGG19, and the critical difference between them is the number of convolutional layers which are 16 and 19, respectively (Simonyan and Zisserman, 2014). Here, we used VGG19 for our case study.

The pollen image datasets are divided into training and validation sets to evaluate the training error and prevent overfitting, compromising 80% training set and 20% validation sets. There are several regularization approaches to avoid overfitting, including early stopping, batch normalization, dropout, L1 and L2 regularization, increasing the number of training datasets, and data augmentation. For our approach, we used dropout, increasing the number of training datasets, and data augmentation. Dropout is a regularization strategy that involves randomly excluding some number of layer outputs during the training of the CNN model. It helps to force nodes within a layer to probabilistically take on more or less responsibility for the inputs, decreasing the complexity of the model. The data augmentation method was also used on the training dataset after separating the dataset into two training and validation datasets to prevent overfitting and increase the accuracy of the model. The deep learning models need enormous datasets, and it is one of the most significant challenges that researchers face in the case of collecting a large number of samples (e.g., Najafabadi et al., 2015; Polling et al., 2021).

Data augmentation is an approach commonly used in computer vision to increase the amount of training data by adding slightly modified copies of already existing data, only using information from the training data (Perez and Wang, 2017). This method can act as a regularization strategy, and the model is not able to overfit all the image samples, which allows for greater model generalizations (Perez and Wang, 2017). For the data augmentation, we used several transformation methods, such as resizing the images (all of which were the same size), rotating the images across multiple angles, and horizontal flips. All these transformations generate new images from the original. This approach balances the sample sizes for images of different species, it delivers a wider variety of features found in images of the pollen grains, and it increases the number of images in the training datasets (**Figure 4**).

To evaluate our CNN algorithm, we used the accuracy metric. The accuracy metric equation includes the terms TP (True Positive), TN (True Negative), FP (False Positive), and FN (False Negative) and provides an estimate of how the model performs through all the classes. It calculates the ratio between the number of correct predictions and the total number of predictions.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

Case study: Pollen analysis of historic Lepidoptera in Great Basin Desert and Sierra Nevada Mountains

Great Basin Desert and Sierra Nevada Mountains pollen-butterfly networks

Our pollen analysis included pollen grains collected from lepidopteran specimens from the UNR Museum of Natural History (UNRMNH) from historic collections dating back to 1910 in the Great Basin and Sierra Nevada Mountains near Reno, NV. Beginning in 2020, we started regular collections of butterflies from three sites to supplement and expand the UNR collections and to improve the resolution of plant-pollinator networks from museum specimens. We selected 266 specimens, including 20 locally abundant native butterfly species from five families in the Great Basin and the Sierra Nevada, for pollen analysis for this study (**Supplementary Table 1**). Part of the dataset was published recently to reconstruct the butterfly-pollen interaction network in the Great Basin and Sierra Nevada Mountains over the past century (Balmaki et al., 2022); that study used the methods described here, and more specific methods for data collection and statistical analyses are described in that paper. While Balmaki et al. (2022) focused on characterizing changes in pollen-butterfly networks over the past century and comparing these networks to contemporary visitation networks, the current paper focuses more generally on pollinator network methodology with an expanded pollen-butterfly network from the UNRMNH collections.

We prepared more than 400 pollen reference slides from native flowers found in the Reno herbarium (RENO) for cross-validation of pollen identifications. We used a ZEISS, Axiolab 5 light microscope, and Axiocam 208 color microscope camera

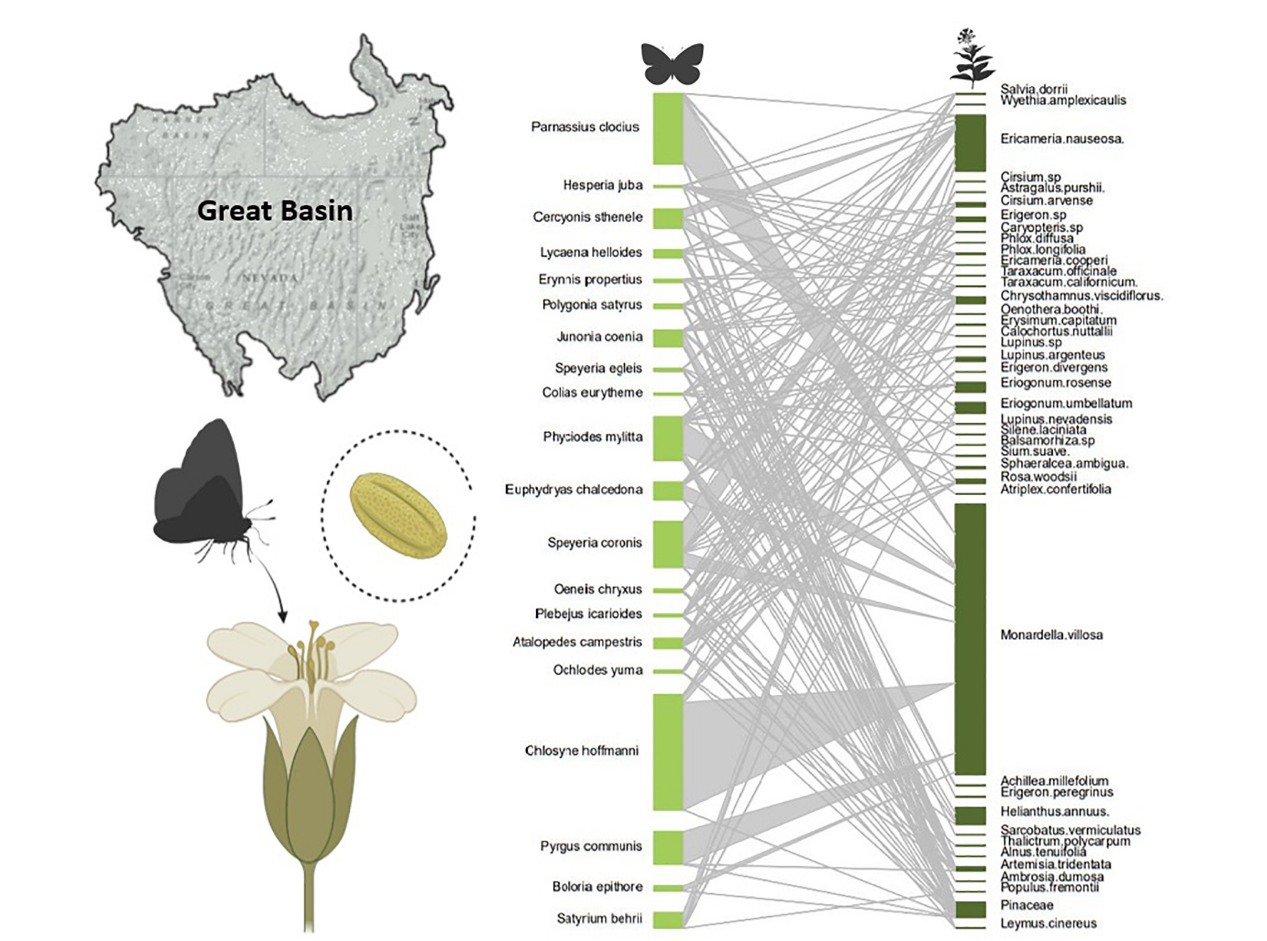
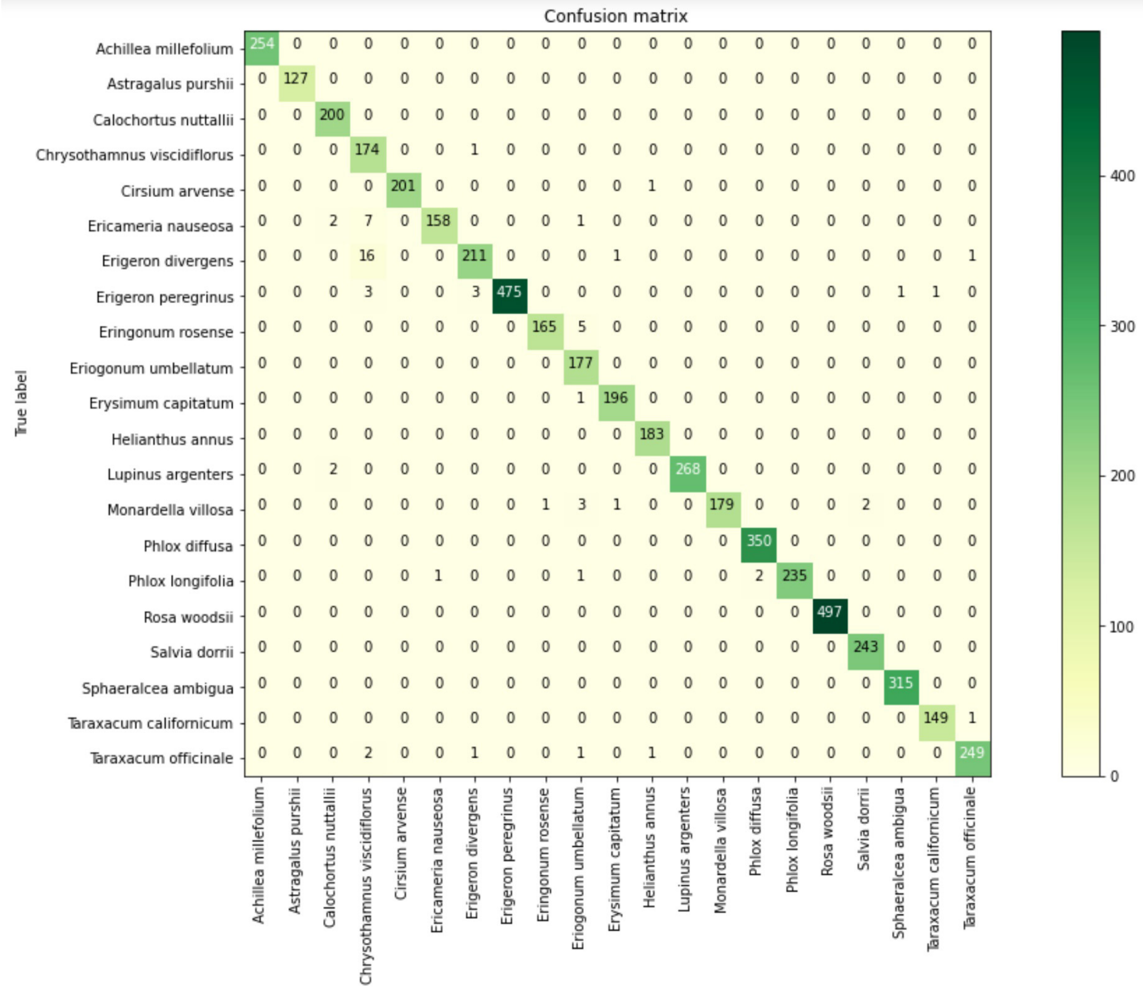


FIGURE 5
Bipartite pollen-butterfly networks of 20 butterfly species from museum collections of butterflies in the Great Basin Desert and Sierra Nevada Mountains (United States). Light green nodes are butterfly species, dark green nodes pollen species, and the size of the nodes indicate the frequency of those species in the dataset, while the edge thickness (gray) indicates the frequency of interactions (or strength of the association) between the insect and plant species.

for pollen identification and photography of pollen grains, and the images were captured using 40× objective lenses and 10× ocular lenses. Z-stack images show the vertical details of pollen grains at various focus levels. To train the model for automating the identification of pollen grains, we cropped all the images using Adobe Photoshop (CS6, 13.0.1.3). We removed images with high levels of noise due to debris, air bubbles, and aggregated pollen.

We then estimated the richness and frequency of butterfly-plant interactions over time and space by bipartite interaction networks, and estimated network parameters using network methods outlined by Dormann et al. (2009). This network provided a summary of butterfly-plant interactions over the last century in the Great Basin Desert and Sierra Nevada Mountains (**Figure 5**). Using temporal subsets of these networks from 1910 to 2021, Balmaki et al. (2022) demonstrated that there have been shifts in plant species associated with butterflies, with strong shifts in network structure when comparing pre- and post-drought time intervals. For that analysis, pollen species known to be from wind-pollinated plants were excluded. Insect-pollinated plants have spikey, sticky pollen grains that easily attach to butterflies' bodies when they are foraging for nectar. Wind-pollinated species in the Great Basin Desert and Sierra Nevada Mountains butterfly-pollen network shown here included species in the families of Pinaceae and Poaceae, and insect-pollinated plants are in the Asteraceae, Lamiaceae, Fabaceae, Polemoniaceae, Malvaceae, and Rosaceae. We found pollen grains of these wind-pollinated families were attached to the legs and wings of butterfly specimens, which means they likely were picked up incidentally from the environment (e.g., as butterflies visit or perch on these plants).

Results from Balmaki et al. (2022) indicated that the plant community associated with butterflies is shifting and that this shift is temporally associated with periods of extreme drought



	Achillea millefolium	Astragalus purshii	Calochortus nuttallii	Chrysothamnus viscidiflorus	Cirsium arvense	Ericameria nauseosa	Erigeron divergens	Erigeron peregrinus	Eringonum rosense	Eriogonum umbellatum	Erysimum capitatum	Helianthus annus	Lupinus argenters	Monardella villosa	Phlox diffusa	Phlox longifolia	Rosa woodsii	Salvia dorrii	Sphaeralcea ambigua	Taraxacum californicum	Taraxacum officinale
Achillea millefolium	254	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Astragalus purshii	0	127	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Calochortus nuttallii	0	0	200	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Chrysothamnus viscidiflorus	0	0	0	174	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Cirsium arvense	0	0	0	0	201	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
Ericameria nauseosa	0	0	2	7	0	158	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
Erigeron divergens	0	0	0	16	0	0	211	0	0	0	1	0	0	0	0	0	0	0	0	0	1
Erigeron peregrinus	0	0	0	3	0	0	3	475	0	0	0	0	0	0	0	0	0	0	1	1	0
Eringonum rosense	0	0	0	0	0	0	0	0	165	5	0	0	0	0	0	0	0	0	0	0	0
Eriogonum umbellatum	0	0	0	0	0	0	0	0	0	177	0	0	0	0	0	0	0	0	0	0	0
Erysimum capitatum	0	0	0	0	0	0	0	0	0	1	196	0	0	0	0	0	0	0	0	0	0
Helianthus annus	0	0	0	0	0	0	0	0	0	0	0	183	0	0	0	0	0	0	0	0	0
Lupinus argenters	0	0	2	0	0	0	0	0	0	0	0	0	268	0	0	0	0	0	0	0	0
Monardella villosa	0	0	0	0	0	0	0	0	1	3	1	0	0	179	0	0	0	2	0	0	0
Phlox diffusa	0	0	0	0	0	0	0	0	0	0	0	0	0	0	350	0	0	0	0	0	0
Phlox longifolia	0	0	0	0	0	1	0	0	0	1	0	0	0	0	2	235	0	0	0	0	0
Rosa woodsii	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	497	0	0	0	0
Salvia dorrii	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	243	0	0	0
Sphaeralcea ambigua	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	315	0	0
Taraxacum californicum	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	149	1
Taraxacum officinale	0	0	0	2	0	0	1	0	0	1	0	1	0	0	0	0	0	0	0	0	249

FIGURE 6
VGG19 confusion matrix for the 21 pollen species used for the training dataset pollen images from the Great Basin. Rows are species identities and columns are convolutional neural network (CNN) species assignments. The color bar indicates frequency, with dark green being most frequent. The diagonal elements are frequency of correctly classified outcomes, while misclassified outcomes are on the off diagonals.

in the Western United States. This study also showed that pollen richness associated with butterflies has declined over the past 100 years, which can be a consequence of lower local plant diversity or fewer floral resources (Balmaki et al., 2022). Fewer floral resources could potentially lead to the decline of pollinator species, especially specialized butterflies that may depend on nectar or pollen from a limited number of plant species (Schowalter, 2006). These temporal changes in plant-pollinator interaction networks are an example of how anthropogenic change may be influencing biodiversity.

Anthropogenic climate change has been characterized by increased drought frequency and intensity, and extreme temperatures in the Western United States, and has in some cases been linked to phenological mismatches between pollinators and their food plants (Stemkovski et al., 2020). Museum specimens are one of the best options for examining predicted changes in plant-pollinator interactions over time due to specific global change parameters.

Convolutional neural network models for the Great Basin Desert and Sierra Nevada Mountains pollen identification

We used two pretrained CNN models (AlexNet and VGG19) to classify the 21 most common pollen species in the Great Basin, including *Achillea millefolium*, *Cirsium arvense*, *Erigeron divergens*, *Erigeron peregrinus*, *Helianthus annus*, *Taraxacum officinale*, *Taraxacum californicum*, *Ericameria nauseosa*, *Chrysothamnus viscidiflorus* (Asteraceae); *Erysimum capitatum* (Brassicaceae); *Astragalus purshii*, *Lupinus argenters* (Fabaceae); *Monardella villosa*, *Salvia dorrii* (Lamiaceae);

Calochortus nuttallii (Liliaceae); *Sphaeralcea ambigua* (Malvaceae); *Phlox diffusa, Phlox longifolia* (Polemoniaceae); *Eriogonum umbellatum, Eriogonum rosense* (Polygonaceae); *Rosa woodsii* (Rosaceae). Our pollen image datasets included 5709 images from 21 different pollen species. The number of images per species ranges between 200 and 650, and the majority of the images belong to these four species (*E. peregrinus*, *S. ambigua*, *P. diffusa*, and *R. woodsii*).

To evaluate the accuracy of our model, we used the validation set, which was composed of unseen images by the model during the training process. These images did not go through the data augmentation, which let us get the realistic accuracy of our model when encountering a new observation. Our AlexNet model achieved the training and validation accuracy of 96.5 and 92.1%, respectively. On the other hand, we acquired higher training and validation accuracy using VGG19, including 98.8 and 93.1%, respectively. This is likely because our VGG19 model architecture, compared to the AlexNet, has a higher number of parameters (VGG19: 102,850,581, AlexNet: 9,459,733) and deeper layers (VGG19: 19, AlexNet:8 layers) which let the VGG19 model better differentiate features within images. The accuracy obtained by the validation dataset was similar to the accuracy obtained by training datasets in the VGG19 model. The low deviation between training and validation accuracy indicates that our model is robust and rules out the possibility of overfitting, which occurs when a model is too complex.

In addition, to see how our VGG19 model acts in different pollen species, we created a confusion matrix that shows just a few mislabeled species (**Figure 6**). Finally, we believe this accuracy in VGG19 is high enough to build a web and phone application to create an automatic classification system for pollen grains at the species level.

Conclusion

Decades of research have focused on coevolution between plants and insects; these coevolutionary interactions have generated broad-scale geographic patterns of interactions that can be summarized with network parameters (Olesen et al., 2007; Tylianakis et al., 2010; Pellissier et al., 2018). This plant-insect interaction research is often limited by poor natural history data, for which museum collections can serve as an untapped and unparalleled resource. Current challenges include incorrectly inferring relationships from brief visits (i.e., a butterfly landing on a flower implies pollination), assuming interactions are present throughout the geographic range of a species and inferring interactions from literature sources. Consequently, inferences used for ecological networks, for understanding of plant-pollinator coevolution, and for pollinator conservation efforts are formed using incomplete data (Dyer, 2018). Despite the abundance of lepidopterans in collections, their importance as pollinators still lacks rigorous quantification for many taxa. Mining pollinator interaction data from museum specimens can help to fill this critical knowledge gap.

In this time of well-documented declines in pollinators, there is a clear need for innovative methods for studying plant-pollinator interaction networks using museum collections (Potts et al., 2010; Burkle et al., 2013). Species interactions, and their impact on community structure, and ultimately, ecosystem functioning, can be explored through better-informed network methods, which can help us to describe spatial and temporal changes in these dynamics (Burkle and Alarcón, 2011; Campos-Moreno et al., 2021). For example, many specialist pollinators are more susceptible to declines as their more restricted niches provide less redundancy in resource availability (Weiner et al., 2014). It is also likely that the occupancy of specialists across the landscape is low compared to generalists (Sudta et al., 2022) and that more specialized pollinators are less abundant overall (Fort et al., 2016). In either case, there is an expectation of a strong positive correlation between generalization and abundance at some scale, which has conservation implications for threatened specialized plant-pollinator interactions and overall network complexity. It is difficult to assess such network responses without careful networks that are backed by natural history observations and that take into account changes across spatial and temporal gradients. In particular, because museum collections can provide multiple observations over space and time, they can be a more powerful tool for differentiating specialist and generalist pollinators than more limited field observations. Analyzing pollen grains on butterflies from museum collections adds valuable natural history data to specimens and is an efficient and accurate method for documenting the frequency and richness of interactions with plants. These methods should be used to explore how networks have changed over time and may help us predict further network change. Lastly, this approach can help us identify relationships that are most at risk to environmental perturbations and those that are robust to perturbations associated with global change.

Data availability statement

The original contributions presented in this study are included in the article/**Supplementary material**, further inquiries can be directed to the corresponding author. **Supplementary data** associated with CNN models can be found at this link: https://github.com/masoudrostami/Pollen-Great-Basin.

Author contributions

BB and MR: methodology, conceptualization, laboratory analysis, data analysis, visualization, original draft preparation, and writing – reviewing and editing. TC: methodology, conceptualization, field work, original draft preparation, and writing – reviewing and editing. EL, JA, CF, and MF: methodology, conceptualization, and writing – reviewing and editing. LD: methodology, conceptualization, data analysis, visualization, original draft preparation, and writing – reviewing and editing. All authors contributed to the article and approved the submitted version.

Funding

This work was supported by NSF grant DEB 2114942 to LD, EL, and JA and NSF DEB-2114793 to MF and LD.

Acknowledgments

We appreciate the contributions of Jerry Tiehm, curator of the herbarium (RENO) at the University of Nevada, Reno, Museum of Natural History (UNRMNH) for maintaining this valuable collection.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fevo.2022.924941/full#supplementary-material

References

Afonso, M., Fonteijn, H., Fiorentin, F. S., Lensink, D., Mooij, M., Faber, N., et al. (2020). Tomato fruit detection and counting in greenhouses using deep learning. *Front. Plant Sci.* 11:571299. doi: 10.3389/fpls.2020.571299

Alarcon, R., Waser, N., and Ollerton, J. (2008). Year-to-year variation in the topology of a plant-Pollinator interaction network. *Oikos* 117, 1796–1807.

Albattah, W., Nawaz, M., Javed, A., Masood, M., and Albahli, S. (2022). Novel deep learning method for detection and classification of plant diseases. *Complex Intellig. Syst.* 8, 507–524.

Alzubaidi, L., Zhang, J., Humaidi, A. J., Al-Dujaili, A., Duan, Y., Al-Shamma, O., et al. (2021). Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. *J. Big Data* 8:53. doi: 10.1186/s40537-021-00444-8

Aslan, C. E., Shiels, A. B., Haines, W., and Liang, C. T. (2019). Non-native insects dominate daytime pollination in a high-elevation Hawaiian dryland ecosystem. *Am. J. Bot.* 106, 313–324.

Astolfi, G., Gonçalves, A. B., Menezes, G. V., Borges, F. S. B., Astolfi, A. C. M. N., Matsubara, E. T., et al. (2020). POLLEN73S: an image dataset for pollen grains classification. *Ecol. Inform.* 60:101165.

Ballantyne, G., Baldock, K. C. R., and Willmer, P. G. (2015). Constructing more informative plant-pollinator networks: visitation and pollen deposition networks in a heathland plant community. *Proc. R. Soc. B* 282:1130. doi: 10.1098/rspb.2015.1130

Balmaki, B., Christensen, T., and Dyer, L. A. (2022). Reconstructing butterfly-pollen interaction networks through periods of anthropogenic drought in the Great Basin (USA) over the past century. *Anthropocene* 37:100325.

Balmaki, B., Wigand, P., Frontalini, F., Shaw, A. T., Avnaim-Katav, S., and Asgharian Rostami, M. (2019). Late holocene paleoenvironmental changes in the seal beach wetland (California, USA): a micropaleontological perspective. *Quatern. Int.* 530-531, 14–24.

Bartomeus, I., Ascher, J. S., Gibbs, J., Danforth, B. N., Wagner, D. L., Hedtke, S. M., et al. (2013). Historical changes in northeastern US bee pollinators related to shared ecological traits. *Proc. Natl. Acad. Sci. U.S.A.* 110, 4656–4660. doi: 10.1073/pnas.1218503110

Bascompte, J., Jordano, P., Meliá, N. C. J., and Olesen, J. M. (2003). The nested assembly of plant-animal mutualistic networks. *Proc. Natl. Acad. Sci. U.S.A.* 100, 9383–9387. doi: 10.1073/pnas.1633576100

Bell, K. L., Fowler, J., Burgess, K. S., Dobbs, E. K., Gruenewald, D., Lawley, B., et al. (2017). Applying pollen dna metabarcoding to the study of plant–pollinator interactions. *Appl. Plant Sci.* 5:1600124. doi: 10.3732/apps.1600124

Blüthgen, N., Menzel, F., and Blüthgen, N. (2006). Measuring specialization in species interaction networks. *BMC Ecol.* 6:9. doi: 10.1186/1472-6785-6-9

Bosch, J., Martín Gonzá Lez, A. M., Rodrigo, A., and Navarro, D. (2009). Plant-pollinator networks: adding the pollinator's perspective. *Ecol. Lett.* 12, 409–419. doi: 10.1111/j.1461-0248.2009.01296.x

Buddha, K., Nelson, H., Zermas, D., and Papanikolopoulos, N. (2019). "Weed detection 401 and classification in high altitude aerial images for robot-based precision 402 agriculture," in *Proceedings of the 2019 27th Mediterranean Conference on Control and Automation 403 (MED)*, (Akko: IEEE), 280–285.

Burkle, L. A., and Alarcón, R. (2011). The future of plant-pollinator diversity: understanding interaction networks across time, space, and global change. *Am. J. Bot.* 98, 528–538. doi: 10.3732/ajb.1000391

Burkle, L. A., Marlin, J. C., and Knight, T. M. (2013). Plant-pollinator interactions over 120 years: loss of species, co-occurrence, and function. *Science* 340, 1611–1615. doi: 10.1126/science.1232728

Campos-Moreno, D. F., Dyer, L. A., Salcido, D., Massad, T. J., Pérez-Lachaud, G., Tepe, E. J., et al. (2021). Importance of interaction rewiring in determining spatial and temporal turnover of tritrophic (Piper-caterpillar-parasitoid) metanetworks in the Yucatán Península, México. *Biotropica* 53, 1071–1081.

Carranza-Rojas, J., Goeau, H., Bonnet, P., Mata-Montero, E., and Joly, A. (2017). Going deeper in the automated identification of *Herbarium* specimens. *BMC Evol. Biol.* 17:181. doi: 10.1186/s12862-017-1014-z

Castillo-Figueroa, D. (2018). Beyond specimens: linking biological collections, functional ecology and biodiversity conservation. *Rev. Peruana Biol.* 25, 343–348.

Colla, S. R., Gadallah, F., Richardson, L., Wagner, D., and Gall, L. (2012). Assessing declines of North American bumble bees (*Bombus* spp.) using museum specimens. *Biodivers. Conserv.* 21, 3585–3595.

Colom, P., Traveset, A., and Stefanescu, C. (2021). Long-term effects of abandonment and restoration of Mediterranean meadows on butterfly-plant interactions. *J. Insect Conserv.* 25, 383–393.

Cushing, E. (2011). Longevity of reference slides of pollen mounted in silicone oil. *Rev. Palaeobot. Palynol.* 164, 121–131.

Daood, A., Ribeiro, E., and Bush, M. (2016). "Pollen grain recognition using deep learning," in *Advances in Visual Computing. Lecture Notes in Computer Science*, eds G. Bebis, R. Boyle, B. Parvin, D. Koracin, I. Pavlidis, R. Feris, et al. (Cham: Springer).

Dormann, C. F., Fründ, J., Blüthgen, N., and Gruber, B. (2009). Indices, graphs and null models: analyzing bipartite ecological networks. *Open Ecol. J.* 2, 7–24.

Dyer, L. A. (2018). Multidimensional diversity associated with plants: a view from a plant-insect interaction ecologist. *Am. J. Bot.* 105, 1439–1442. doi: 10.1002/ajb2.1147

Ferrarini, A., Alatalo, J. M., Gervasoni, D., and Foggi, B. (2017). Exploring the compass of potential changes induced by climate warming in plant communities. *Ecol. Complex* 29, 1–9.

Flagel, L., Brandvain, Y., and Schrider, D. R. (2019). The unreasonable effectiveness of convolutional neural networks in population genetic inference. *Mol. Biol. Evol.* 36, 220–238. doi: 10.1093/molbev/msy224

Fort, H., Vázquez, D. P., and Lan, B. L. (2016). Abundance and generalisation in mutualisticnetworks: solving the chicken-and-egg dilemma. *Ecol. Lett.* 19, 4–11. doi: 10.1111/ele.12535

Galanty, A., Danel, T., Wegrzyn, M., Podolak, I., and Podolak, I. (2021). Deep convolutional neural network for preliminary in-field classification of lichen species. *Biosyst. Eng.* 204, 15–25.

Gallardo-Caballero, R., García-Orellana, C. J., García-Manso, A., González-Velasco, H. M., Tormo-Molina, R., and Macías-Macías, M. (2019). Precise pollen grain detection in bright field microscopy using deep learning techniques. *Sensors* 19:3583. doi: 10.3390/s19163583

Geetharamani, G., and Pandian, A. (2019). Identification of plant leaf diseases using a nine-layer deep convolutional neural network. *Comput. Electr. Engineer* 76, 323–338.

Gonçalves, A. B., Souza, J. S., Silva, G. G. D., Cereda, M. P., Pott, A., Naka, M. H., et al. (2016). Feature extraction and machine learning for the classification of brazilian savannah pollen grains. *PLoS One* 11:e0157044. doi: 10.1371/journal.pone.0157044

Harrison, S., Spasojevic, M. J., and Li, D. (2020). Climate and plant community diversity in space and time. *Proc. Natl. Acad. Sci. U.S.A.* 117, 4464–4470.

Johnson, K. G., Brooks, S. J., Fenberg, P. B., Glover, A. G., James, K. E., Lister, A. M., et al. (2011). Climate change and biosphere response: unlocking the collections vault. *BioScience* 61, 147–153.

Jones, C. A., and Daeler, C. C. (2018). Herbarium specimens can reveal impacts of climate change on plant phenology; a review of methods and applications. *PeerJ* 6:e4576. doi: 10.7717/peerj.4576

Jones, G. D. (2012a). Pollen analyses for pollination research, unacetolyzed pollen. *J. Pollinat. Ecol.* 9, 96–107.

Jones, G. D. (2012b). Pollen extraction from insects. *Palynology* 36, 86–109.

Jones, G. D. (2014). Pollen analyses for pollination research, acetolysis. *J. Pollinat. Ecol.* 13, 203–217. doi: 10.1590/s1519-566x2009000200005

Jones, G. D., and Jones, S. D. (2001). The uses of pollen and its implication for entomology. *Neotrop. Entomol.* 30, 341–350.

Kaiser-Bunbury, C. N., Memmott, J., and Müller, C. B. (2009). Community structure of pollination webs of Mauritian heathland habitats. *Perspect. Plant Ecol. Evol. Syst.* 11, 241–254.

Khanzhina, N., Putin, E., Filchenkov, A., and Zamyatina, E. (2018). "Pollen grain recognition using convolutional neural network," in *Proceedings of the ESANN 2018-Proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, (Piscataway, NJ: IEEE), 409–414.

Kleijn, D., and Raemakers, I. (2008). A retrospective analysis of pollen host plant use by stable and declining bumble bee species. *Ecology* 89, 1811–1823. doi: 10.1890/07-1275.1

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. *NIPS* 1, 1–4.

Losapio, G., de la Cruz, M., Escudero, A., Schmid, B., and Schöb, C. (2018). The assembly of a plant network in alpine vegetation. *J. Veg. Sci.* 29, 999–1006.

Martinez, N. (1992). Constant connectance in community food webs. *Am. Nat.* 139, 1208–1218.

Mendes, S. B., Timóteo, S., Loureiro, J., and Castro, S. (2022). The impact of habitat loss on pollination services for a threatened dune endemic plant. *Oecologia* 198, 279–293. doi: 10.1007/s00442-021-05070-y

Metelmann, S., Sakai, S., Kondoh, M., and Telschow, A. (2020). Evolutionary stability of plant–pollinator networks: efficient communities and a pollination dilemma. *Ecol. Lett.* 23, 1747–1755. doi: 10.1111/ele.13588

Najafabadi, M. M., Villanustre, F., Khoshgoftaar, T. M., Seliya, N., Wald, R., and Muharemagic, E. (2015). Deep learning applications and challenges in big data analytics. *J. Big Data* 2, 1–21.

Norouzzadeh, M. S., Morris, D., Beery, S., Joshi, N., Jojic, N., and Clune, J. (2021). A deep active learning system for species identification and counting in camera trap images. *Methods Ecol. Evol.* 12, 150–161.

Norouzzadeh, M. S., Nguyen, A., Kosmala, M., Swanson, A., Palmer, M. S., Packer, C., et al. (2018). Automatically identifying, counting, and describing wild animals in camera-trap images with deep learning. *Proc. Natl. Acad. Sci. U.S.A.* 1150, 5716–5725. doi: 10.1073/pnas.1719367115

Olesen, J. M., Bascompte, J., Dupont, Y. L., and Jordano, P. (2007). The modularity of pollination networks. *Proc. Natl. Acad. Sci. U.S.A.* 104, 19891–19896.

O'Mahony, N., Campbell, S., Carvalho, A., Harapanahalli, S., Velasco-Hernandez, G., Krpalkova, L., et al. (2019). "Deep learning vs. traditional computer vision," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, (Piscataway, NJ: IEEE).

Pawar, S. (2014). Why are plant-pollinator networks nested? Mutualistic communities maximize their structural stability. *Science* 345:282.

Pellissier, L., Albouy, C., Bascompte, J., Farwig, N., Graham, C., Loreau, M., et al. (2018). Comparing species interaction networks along environmental gradients. *Biol. Rev.* 93, 785–800.

Perez, L., and Wang, J. (2017). The effectiveness of data augmentation in image classification using deep learning. *arXiv* [preprint]. Available online at: https://arxiv.org/abs/1712.04621 (accessed November, 2021).

Polling, M., Li, C., Cao, L., Verbeek, F., de Weger, L. A., Belmonte, J., et al. (2021). Neural networks for increased accuracy of allergenic pollen monitoring. *Sci. Rep.* 11, 11357–11367.

Potts, S. G., Biesmeijer, J. C., Kremen, C., Neumann, P., Schweiger, O., and Kunin, W. E. (2010). Global pollinator declines: trends, impacts and drivers. *Trends Ecol. Evol.* 25, 345–353.

Riding, J. B. (2021). A guide to preparation protocols in palynology. *Palynology* 45, 1–110.

Romero, I. C., Kong, S., Fowlkes, C. C., Jaramillo, C., Urban, M. A., Oboh-Ikuenobe, F., et al. (2020). Improving the taxonomy of fossil pollen using convolutional neural networks and superresolution microscopy. *Proc. Natl. Acad. Sci. U.S.A.* 117, 28496–28505. doi: 10.1073/pnas.2007324117

Salcido, D. M., Forister, M., Lopez, H. G., and Dyer, L. A. (2020). Loss of dominant caterpillar genera in a protected tropical forest. *Sci. Rep.* 10:422. doi: 10.1038/s41598-019-57226-9

Schowalter, T. D. (2006). *Insect Ecology: An Ecosystem Approach.* Cambridge, MA: Academic Press.

Seltmann, K. C., Cobb, N. S., Gall, L. F., Bartlett, C. R., Basham, M. A., Betancourt, I., et al. (2017). LepNet: the lepidoptera of north america network. *Zootaxa* 4247, 73–77. doi: 10.11646/zootaxa.4247.1.10

Sevillano, V., and Aznarte, J. L. (2018). Improving classification of pollen grain images of the polen23e dataset through three different applications of deep learning convolutional neural networks. *PLoS One* 13:e0201807. doi: 10.1371/journal.pone.0201807

Sevillano, V., Holt, K., and Aznarte, J. L. (2020). Precise automatic classification of 46 different pollen types with convolutional neural networks. *PLoS One* 15:e0229751. doi: 10.1371/journal.pone.0229751

Silberbauer, L., Yee, M., Socorro, A. D., Weatten, S., Gregg, P., and Bowie, M. (2004). Pollen grains as markers to track the movements of generalist predatory insects in agroecosystems. *Int. J. Pest Manag.* 50, 165–171.

Simonyan, K., and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv* [Preprint]. Available online at: https://arxiv.org/abs/1409.1556 (accessed November, 2021).

Stemkovski, M., Pearse, W. D., Griffin, S. R., Pardee, G. L., Gibbs, J., Griswold, T., et al. (2020). Bee phenology is predicted by climatic variation and functional traits. *Ecol. Lett.* 23, 1589–1598.

Strydom, T., Catchen, M. D., Banville, F., Caron, D., Dansereau, G., Desjardins-Proulx, P., et al. (2021). A roadmap towards pre-dicting species interaction networks (across space and time). *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 376:20210063. doi: 10.1098/rstb.2021.0063

Sudta, C., Salcido, D. M., Forister, M. L., Walla, T. R., Villamarín-Cortez, S., and Dyer, L. A. (2022). Jack-of-all-trades paradigm meets long-term data: generalist herbivores are more widespread and locally less abundant. *Ecol. Lett.* 25, 948–957. doi: 10.1111/ele.13972

Tylianakis, J. M., Laliberte, E., Nielsen, A., and Bascompte, J. (2010). Conservation of species interaction networks. *Biol. Conserv.* 143, 2270–2279.

Viertel, P., and Konig, M. (2022). Pattern recognition methodologies for pollen grain image classification: a survey. *Mach. Vis. Appl.* 33:18.

Wagner, D., Fox, R., Salcido, D. M., and Dyer, L. A. (2021). A window to the world of global insect declines: moth biodiversity trends are complex and heterogeneous. *Proc. Natl. Acad. Sci. U.S.A.* 118:e2002549117. doi: 10.1073/pnas.2002549117

Wäldchen, J., and Mader, P. (2018). Machine learning for image based species identification. *Methods Ecol. Evol.* 9, 2216–2225.

Weiner, C. N., Werner, M., Linsenmair, K. E., and Blüthgen, N. (2014). Land-use impacts on plant-pollinator networks: interaction strength and specialization predict pollinator declines. *Ecology* 95, 466–474. doi: 10.1890/13-0436.1

Wood, T. J., Gibbs, J., Graham, K. K., and Isaacs, R. (2019). Narrow pollen diets are associated with declining Midwestern bumble bee species. *Ecology* 100:e0193822. doi: 10.1002/ecy.2697

Yamaji, F., and Ohsawa, T. A. (2016). Field experiments of pollination ecology: the case of *Lycoris sanguinea* var. sanguinea. *J. Visual. Exp.* 2016:54728. doi: 10.3791/54728

Younis, S., Schmidt, M., Weiland, C., Dressler, S., Seeger, B., and Hickler, T. (2020). Detection and annotation of plant organs from digitized herbarium scans using deep learning. *Biodivers. Data J.* 8:e57090. doi: 10.3897/BDJ.8.e57090

frontiers | Frontiers in Ecology and Evolution

TYPE Original Research
PUBLISHED 08 September 2022
DOI 10.3389/fevo.2022.970249

Assessing the performance of historical skins and bones for museomics using wolf specimens as a case study

Carolina Pacheco[1,2,3†], Diana Lobo[1,2,3†], Pedro Silva[1,3], Francisco Álvares[1,3], Emilio J. García[4], Diana Castro[1,3], Jorge F. Layna[5], José Vicente López-Bao[4] and Raquel Godinho[1,2,3,6*]

[1]CIBIO, Centro de Investigação em Biodiversidade e Recursos Genéticos, *InBIO* Laboratório Associado, Universidade do Porto, Vairão, Portugal, [2]Departamento de Biologia, Faculdade de Ciências, Universidade do Porto, Porto, Portugal, [3]BIOPOLIS, Program in Genomics, Biodiversity and Land Planning, CIBIO, Vairão, Portugal, [4]Biodiversity Research Institute (CSIC), Oviedo University, Mieres, Spain, [5]Consultores en Iniciativas Ambientales, S.L., Madrid, Spain, [6]Department of Zoology, Centre for Ecological Genomics and Wildlife Conservation, University of Johannesburg, Johannesburg, South Africa

OPEN ACCESS

EDITED BY
Jimmy McGuire,
University of California, Berkeley,
United States

REVIEWED BY
Shaoqing Wen,
Fudan University, China
Nicolas Dussex,
Stockholm University, Sweden

*CORRESPONDENCE
Raquel Godinho
rgodinho@cibio.up.pt

†These authors have contributed equally to this work and share first authorship

SPECIALTY SECTION
This article was submitted to Evolutionary and Population Genetics, a section of the journal Frontiers in Ecology and Evolution

RECEIVED 15 June 2022
ACCEPTED 11 August 2022
PUBLISHED 08 September 2022

CITATION
Pacheco C, Lobo D, Silva P, Álvares F, García EJ, Castro D, Layna JF, López-Bao JV and Godinho R (2022) Assessing the performance of historical skins and bones for museomics using wolf specimens as a case study. *Front. Ecol. Evol.* 10:970249. doi: 10.3389/fevo.2022.970249

Advances in the field of museomics have promoted a high sampling demand for natural history collections (NHCs), eventually resulting in damage to invaluable resources to understand historical biodiversity. It is thus essential to achieve a consensus about which historical tissues present the best sources of DNA. In this study, we evaluated the performance of different historical tissues from Iberian wolf NHCs in genome-wide assessments. We targeted three tissues—bone (jaw and femur), maxilloturbinal bone, and skin—that have been favored by traditional taxidermy practices for mammalian carnivores. Specifically, we performed shotgun sequencing and target capture enrichment for 100,000 single nucleotide polymorphisms (SNPs) selected from the commercial Canine HD BeadChip across 103 specimens from 1912 to 2005. The performance of the different tissues was assessed using metrics based on endogenous DNA content, uniquely high-quality mapped reads after capture, and enrichment proportions. All samples succeeded as DNA sources, regardless of their collection year or sample type. Skin samples yielded significantly higher amounts of endogenous DNA compared to both bone types, which yielded equivalent amounts. There was no evidence for a direct effect of tissue type on capture efficiency; however, the number of genotyped SNPs was strictly associated with the starting amount of endogenous DNA. Evaluation of genotyping accuracy for distinct minimum read depths across tissue types showed a consistent overall low genotyping error rate (<7%), even at low (3x) coverage. We recommend the use of skins as reliable and minimally destructive sources of endogenous DNA for whole-genome and target enrichment approaches in mammalian carnivores. In addition, we provide a new 100,000 SNP capture array validated for historical DNA (hDNA) compatible to the Canine HD BeadChip for high-quality DNA. The increasing demand for NHCs as DNA sources should encourage the generation of genomic datasets comparable among studies.

KEYWORDS

endogenous DNA, historical DNA, museomics, natural history collections, SNP genotyping errors, target enrichment

Introduction

Natural history collections (NHCs) have been gathered since the seventeenth century, motivated by human curiosity about our planet's biodiversity and the breakthrough in preserving perishable material (Farrington, 1915). Currently, national, regional, and private collections worldwide own irreplaceable natural resources, offering wide perspectives across distinct temporal and spatial scales that inspire research in many scientific areas (Casas-Marce et al., 2012; Tsangaras and Greenwood, 2012; Lopez et al., 2020; Pearson et al., 2020). Such collections provide unique overviews of historical biodiversity, from both extinct and extant species, and are also essential resources for addressing questions about species that have sampling limitations due to financial, bureaucratic, or conservation constraints (Burrell et al., 2015). Advances in molecular genetics and sequencing technology have promoted the use of the naturally fragmented DNA of historical specimens, transforming NHCs into invaluable sources of material for investigating genetics-related questions among different fields, including phylogenetics, biogeography, and conservation (Suarez and Tsutsui, 2004; Holmes et al., 2016; Bi et al., 2019).

Whereas a growing body of studies using historical DNA (hDNA) illustrate the potential of NHCs in genetic research, their regular use still poses different challenges. Sampling of genetic material is often destructive, eventually compromising the integrity of specimens and jeopardizing their future use. Therefore, because sampling techniques that minimize damages are prioritized (Pálsdóttir et al., 2019), the amount of genetic material collected is often limited (Horváth et al., 2005). Additionally, traditional taxidermy practices, which commonly use hazardous chemicals, and general carelessness in protecting specimens from environmental damage, do not favor DNA preservation. Thus, historical samples often yield limited and highly degraded DNA (Raxworthy and Smith, 2021) that presents major challenges during laboratory and analytical procedures (Allentoft et al., 2012; Dabney et al., 2013). Moreover, hDNA extracts can contain a non-negligible proportion of exogenous DNA from pre- or post-mortem sources, frequently in overwhelming ratios (Weiß et al., 2016; McDonough et al., 2018; Eisenhofer et al., 2019). These factors may explain why genetic studies using hDNA have often relied on the amplification of short nuclear or mitochondrial fragments (e.g., Schwartz et al., 2007; Maebe et al., 2016; Lonsinger et al., 2019). However, the use of hDNA is nowadays facilitated by high-throughput sequencing and by recent developments in molecular methods (e.g., Rowe et al., 2011; Staats et al., 2013; Hung et al., 2014). Methods like sequence capture of target loci, which limits the representation of the genome to specific loci, are among the most used in genome-wide studies of low-quality DNA to achieve large and cost-effective datasets (Jones and Good, 2016; McCormack et al., 2016; Derkarabetian et al., 2019). Furthermore, despite requiring *a priori* availability of the target genome to design specific baits, target enrichment has been shown to be successful using bait designs based on closely related species (Vallender, 2011).

Mammals have been traditionally preserved in NHCs by archiving skins, bones, teeth, or mounted specimens (Rowe et al., 2011). It is thus not surprising that most of the available genetic studies using hDNA rely on these tissues (Raxworthy and Smith, 2021). However, DNA yields may vary greatly among different tissues and be dependent on curation history (Burrell et al., 2015). Hard tissues, such as teeth and bones, were at first thought to provide higher-quality DNA (Wandeler et al., 2007; Casas-Marce et al., 2010), encouraging proposals to use hard tissues assumed to minimize sampling damage, such as maxilloturbinal bone (Wisely et al., 2004). Yet, recent studies have shown conflicting results (Rowe et al., 2011; Lonsinger et al., 2019; Tsai et al., 2020), revealing soft tissues to be good sources of hDNA when preserved appropriately (Burrell et al., 2015). To date, few studies have implemented genomic resources to assess differences in the performance of distinct tissues, and those tackling this question are often based on very low sample sizes that hamper reliable statistical comparisons (e.g., Rowe et al., 2011; McDonough et al., 2018). Thus, the tissue of choice for increasing the quality of genomic data, while sampling mammal NHCs with minimal damaging, remains an open question (Raxworthy and Smith, 2021).

In this work, we sought to evaluate the performance of different mammalian carnivore tissues generally available at NHCs in genome-wide assessments. Using the Iberian wolf (*Canis lupus signatus*) as a case study, we first collected bones (jaw and femur), maxilloturbinal bones, and skins from

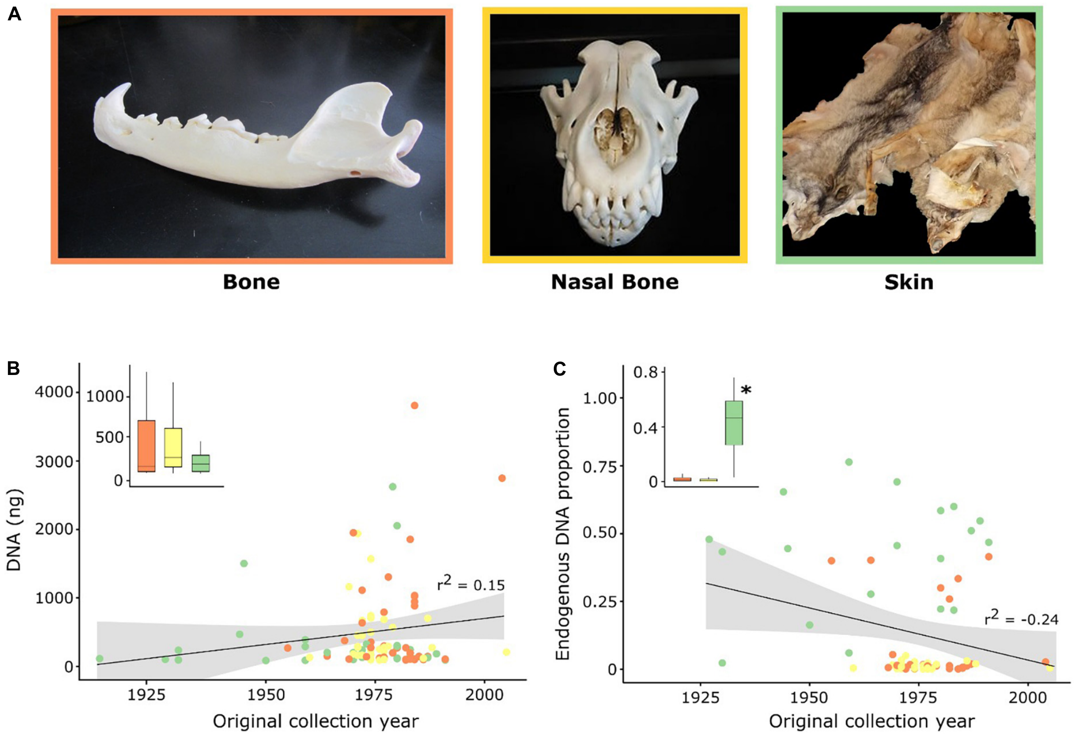


FIGURE 1
Characterization of samples and hDNA used in this work. **(A)** Photographs illustrating the specimens used for the collection of each tissue type (Image credits Raquel Godinho). **(B,C)** Correlation (r^2) between DNA yield (ng) following extraction and the proportion of endogenous DNA content with the original collection date of the specimens. Each dot represents a historical sample and inset boxplots display the median (central line) and distribution per tissue type. Significant differences ($p < 0.05$) between tissues are identified with an asterisk. Colors depict the three different tissue types used in this study: orange—bones (jaw and femur); yellow—maxilloturbinal (nasal) bones; and, green—skins.

103 historical specimens (Figure 1A). Then, we performed a shotgun sequencing of these samples to characterize endogenous DNA content. Third, we developed and tested a capture array of 100,000 regions overlapping single nucleotide polymorphisms (SNPs) contained in the commercial Illumina Canine HD BeadChip, ensuring compatibility between datasets generated with both low- and high-quality DNA. We intended to answer three main questions: Does endogenous DNA content differ across historical tissues? How is capture efficiency affected by historical tissue type? What is the effect of read depth on SNP genotyping error rates from hDNA?

Materials and methods

Sampling, DNA extraction, and shotgun sequencing

We collected 103 samples from Iberian wolf specimens housed at the three largest museum collections in the Iberian Peninsula and at 15 private NHCs, consisting of 43 bones (jaw and femur), 31 maxilloturbinal bones (hereafter nasal bones), and 29 skins. The original collection years for these specimens ranged from 1912 to 1990, except for two samples from 2004 and 2005 (Figure 1, Supplementary Table 1). All our samples conform to the definition of historical samples by Raxworthy and Smith (2021), which restricts hDNA to that fortuitously obtained from traditional museum specimens not intended to serve as sources of DNA. Jaw and femur bones were sampled by drilling ca. 1 g of bone powder with a Dremel tool (Dremel, WI, United States; Supplementary Figure 1). Femurs were sampled in the patellar surface region, whereas jaws were sampled in the posterior lower region of the mandibula (see Supplementary Figure 1 for examples of the drilling location in each bone type). Nasal bones were collected following Wisely et al. (2004), by inserting sterilized forceps into the nasal cavity of the skull to extract the bones. Nasal bone material was posteriorly crushed into small fragments. Bones were not bleached prior to DNA extraction. Skin samples were collected from pelts or mounted specimens by extracting a patch of approximately 2 cm^2. Collecting tools, including drill bits, were cleaned with bleach, and flamed with 96% ethanol between samples to minimize cross contamination.

We prepared DNA extracts using 50 mg of sample following Dabney and Meyer (2019). For skins, we favored the inner layer for DNA extraction and discarded hairs to minimize contamination with external DNA sources. DNA concentration was measured using the Qubit fluorometer dsDNA HS Assay Kit (Thermo Fisher Scientific, MA, United States). We used 100–300 ng of DNA to prepare blunt-end dual-indexed DNA sequencing libraries using a full-uracil-DNA-glycosylase treatment following Meyer and Kircher's (2010) protocol with the modifications described in Kircher et al. (2012). To limit DNA contamination, DNA extractions and library preparations were conducted in dedicated rooms under sterile conditions and positive air pressure, and negative controls were used alongside the procedures. DNA libraries were diluted based on concentration measurements obtained with the Qubit fluorometer dsDNA HS Assay Kit, and library size ranges were characterized using a Bioanalyzer 2100 with High Sensitivity DNA kits (Agilent Technologies, CA, United States). To characterize the endogenous DNA content of each sample, we performed a shotgun sequencing run using one lane of an Illumina HiSeq X instrument (Illumina, CA, United States) in PE 150 bp mode. For this, we selected libraries with concentrations > 15 ng/μl ($N = 79$; $N_{\text{bone}} = 29$; $N_{\text{nasal bone}} = 30$; $N_{\text{skin}} = 20$) to ensure that enough library material remained for the following steps (based on the required starting amount for target capture enrichment of 14–72 ng/μl).

We also generated genomic information from two contemporary wolf muscle samples to be used as positive controls to validate the implemented approach. No animals were killed or injured for this study. We isolated DNA from these samples using the DNeasy Blood and Tissue Kit (Qiagen, Hilden, Germany). DNA quantification and library preparation were performed following the same procedures as described above for historical samples but with two modifications during library preparation: (i) a shearing step using a Bioruptor Pico sonication device (Diagenode, NJ, United States) was performed to obtain fragments of ∼250 bp, and (ii) the USER enzyme treatment was not performed. These two samples were genotyped using two different approaches: the Illumina CanineHD BeadChip (Illumina, CA, United States; ∼170,000 SNPs) and the capture array of 100,000 regions developed in this work (see next sections).

Bait design

We targeted a set of autosomal genome-wide SNPs whose positions were defined based on the coordinates available on the CanineHD BeadChip. With this experimental design, we ensured the compatibility of historical datasets with other datasets generated using the same SNP chip. From the ∼170,000 SNPs in the CanineHD BeadChip, only uniquely mapped autosomal SNPs were considered for probe design. Based on this list of putative SNPs, we custom designed a sequence capture panel containing a final set of 100,000 SNPs distributed across the wolf genome using MYbaits Target Capture technology (Arbor Biosciences, Ann Arbor, MI, United States). The implemented methods for probe design followed the strategy described by Haak et al. (2015) and Cruz-Dávalos et al. (2017), in which four probes of 60 bp each were designed to target fragments of 120 bp per SNP: one upstream of the SNP, one downstream, and one for each possible allele, with the allele positioned at the center of the probe. RNA probes were designed and synthesized by Arbor Biosciences (product code #302016). The selection of the final set of MYbaits RNA probes was based on their capture efficiency (stringent criteria $\leq 25\%$ repeat masked). A FASTA file providing the sequences of all 400,000 synthesized probes is available online.

Capture enrichment

We performed target enrichment using the MYbaits Custom Target Capture Kit following the manufacturer protocol (v.3.02). DNA libraries with similar endogenous DNA content were pooled in equimolar sets of 8–10 samples per capture reaction (600 ng in the final pool). Hybridization between RNA probes and the DNA library occurred at 65°C for 40 h. Real-time PCRs were performed to determine the number of amplification cycles required to obtain sufficient molarity for sequencing. Post-enrichment libraries were amplified using KAPA HiFi Hotstart Ready Mix (KAPA Biosystems, MA, United States), following the manufacturer recommendations, with an annealing temperature of 60°C. The amplified enriched DNA libraries were purified in 20 μl of EB buffer using the MinElute PCR Purification kit (Qiagen, Hilden, Germany) and quantified by the Qubit dsDNA HS Assay Kit. Library size ranges were characterized using a Bioanalyzer 2100 with High Sensitivity DNA kits and pooled in equimolar ratios. Pooled libraries were then sequenced using 2 lanes of an Illumina HiSeq X instrument in PE 150 bp mode.

Processing of sequencing reads and single nucleotide polymorphism calling

Raw sequences from shotgun sequencing and target enrichment were processed similarly in different time periods. First, sequence reads were demultiplexed and quality assessments were done using FastQC (Andrews, 2018) and MultiQC (Ewels et al., 2016). Sequence reads were then processed and aligned using the PALEOMIX v.1.2.13.2 BAM pipeline (Schubert et al., 2014). Briefly, the pipeline consisted of the following steps: (i) adapter sequences were removed, low-quality and N bases were trimmed, and overlapping read pairs were collapsed using AdapterRemoval v.2.2.2 (Schubert et al., 2016) with the default parameters (on average, 96 and 95%

of reads were collapsed for shotgun and capture sequencing, respectively); (ii) all sequence reads were then aligned against the CanFam3.1 dog reference genome (Lindblad-Toh et al., 2005) using BWA-MEM v.0.7.17 (Heng, 2013); and (iii) PCR duplicates for each library were marked by the "paleomix rmdup_collapsed" tool but were not used for the following steps. The final alignment file was subjected to local realignment around indels using the GATK v.3.8 IndelRealigner tool (DePristo et al., 2011). The endogenous content was determined by the ratio of unique reads (no duplicates) that mapped to the dog reference genome with mapping quality above 20 (MQ > 20) to the total number of available reads (also referred as library complexity by some authors; Dehasque et al., 2022). For the target enrichment experiment, the number of reads mapping on target was defined as the total number of unique high-quality (MQ > 20) reads overlapping at least 1 base of the 120 bp target region. Enrichment success was determined by the ratio of reads on target in relation to (i) the total number of available reads and (ii) the total number of uniquely and high-quality mapped reads. These metrics were retrieved from the PALEOMIX summary report and, additionally, with the help of SAMtools v.1.9 (Li et al., 2009). Fold enrichment was determined by the ratio of the number of on-target reads to the total number of uniquely and high-quality mapped reads, divided by the expected representation of the target regions without enrichment (i.e., the ratio of genome length, 2.4 Gb, to target length, 100,000 × 120 bp, corresponding to 0.5%).

Following target enrichment, genotypes were called using BCFtools v.1.10.2 (Li, 2011) mpileup/call -m tools, with minimum Phred-scaled thresholds of 20 for base quality and read mapping quality. At the end, to evaluate the effect of sequencing depth on genotyping quality rates, we considered genotypes supported by at least three (DP $\geq$ 3), four (DP $\geq$ 4), or more reads (DP $\geq$ 5) using the custom python script gtvalues2plink.py, which was also used to convert the final VCF file to plink format. The distribution and density of SNPs in our dataset (using DP $\geq$ 4) was visualized using the R/Bioconductor package karyoploteR (Gel and Serra, 2017) in R (R Development Core Team, 2017).

Genotypes were ultimately validated by estimating their concordance rates with the genotypes obtained from the control samples using the Canine HD BeadChip. In this last approach, genotype calling was performed using GenomeStudio software (Illumina), following Illumina's recommendations. Sex-chromosome-related SNPs and non-uniquely mapped SNPs (SNPs with multiple positions attributed) were removed from the dataset using PLINK v.1.9 (Purcell et al., 2007), resulting in a final dataset of ∼ 121,000 SNPs. Concordance rates were calculated for the entire set of genome-wide SNPs obtained across the three depth thresholds for each control sample, using a second custom script, SNP_concordance.py. We also calculated concordance rates for SNPs called exclusively with 3x and 4x coverage in each control sample and estimated the associated error rates. Additionally, to assess the genotype quality across all historical samples, we estimated the potential genotyping error rate associated with low read depth using the ErrorCount.sh script from the dDocent pipeline (Puritz et al., 2014). Briefly, this script reports a low range based on a 50% binomial probability of observing the second allele in a heterozygote and a high range based on a 25% probability. All genotyped SNPs were considered for this analysis without any filter for missing data.

Statistical analysis

To determine whether the amount of endogenous DNA and capture efficiency were influenced by sample type (bone, nasal bone, and skin), we implemented a set of generalized linear models (GLMs). We ran four different GLMs with the following dependent variables: (i) proportion of endogenous DNA, i.e., number of reads that mapped uniquely with MQ > 20 in relation to the total number of available reads, following shotgun sequencing; (ii) mapped reads after capture, i.e., proportion of the number of reads that mapped uniquely and with MQ > 20 to the total number of reads available after target enrichment; (iii) reads on target (all), i.e., number of unique and high-quality (MQ > 20) reads that mapped on target regions in relation to the total number of reads available; and (iv) reads on target (mapped), i.e., the same as above, but in relation to the total number of mapped reads. All GLMs were fitted with a binomial error distribution and a logit link. The fit of each model was further assessed using the Pearson's χ^2 residuals, which test whether any significant patterns remain in the residuals. Given the unavailability of shotgun sequencing data for all the samples, we tested levels of correlation between the proportion of endogenous DNA and the proportion of reads mapping after capture to understand if the latter could be interpreted as a proxy for endogenous content. We also tested correlation coefficients using the following variables: original collection year, DNA yield (ng) following extraction, endogenous DNA proportion, fragment length (average length of filtered reads from shotgun sequencing), and mapping length (average length of mapped and unique reads, with MQ > 20). The effect of sample type on DNA concentration was also evaluated using a Kruskal-Wallis test. All the previously mentioned tests were performed in R, and all the plots were constructed using the R package ggpubr (Kassambara, 2020).

Results

Endogenous DNA content

We successfully obtained DNA extracts for all 103 historical samples, with an average DNA yield of 469.91 ± 63.18 (s.e.)

ng; (range 84–3,812 ng). DNA yield was not correlated with the original specimen collection year (r^2 = 0.15) nor with the sample type (Kruskal-Wallis test, p = 0.256; Figure 1B). Nevertheless, even using the same quantity of starting material for DNA extraction, we cannot rule out greater effects due to histological differences between hard and soft tissues. The initial sample characterization by shotgun sequencing resulted in an average fragment length of 97 bp (range 42–145 bp). Endogenous DNA content across all samples varied from 0.05 to 76.35% (mean: 14.48 ± 2.47%; Figure 1C) and was not correlated with the fragment length (r^2 = −0.42; Supplementary Figure 2A) nor with the original collection year (r^2 = −0.24; Figure 1C). The average mapping length across all samples was 75 bp (range 40–116 bp) and was not correlated with specimen original collection year (r^2 = 0.15; Supplementary Figure 2B). Among sample types, skin samples retrieved the highest proportion of endogenous DNA (43.49 ± 4.76%, p = 0.008; Figure 1C and Supplementary Tables 2, 3) in relation to bones (8.14 ± 2.66%) and nasal bones (1.28 ± 0.22%).

After capture enrichment, the proportion of reads mapping to the reference genome ranged from 0.13 to 78.26% (mean: 19.89 ± 2.59%; Supplementary Table 2). Consistently, skin samples presented a significantly higher proportion of mapped reads (49.01 ± 4.69%, p = 0.001; Figure 2 and Supplementary Tables 2, 4) relative to bones (12.57 ± 2.99%) and nasal bones (2.79 ± 0.64%). Across contemporary control samples, the average proportion of reads mapping after capture was 32.85 ± 0.56% (Supplementary Table 5), with 65.4% of filtered reads being duplicates. Endogenous DNA content and the proportion of reads mapping to the reference genome following capture were highly correlated (r^2 = 0.93; Supplementary Figure 3A).

Capture efficiency across sample types

Following capture enrichment, the proportion of reads mapping on target regions ranged from 0.01 to 27.79% (mean: 4.75 ± 0.73%; Supplementary Table 2) in relation to all available reads and from 1 to 37% (mean: 19.82 ± 0.84%; Supplementary Table 2) in relation to all mapped reads. The average fold enrichment was 39.64x, with 91% of all samples presenting an enrichment > 10x. Among sample types, skin samples showed the highest proportion of reads mapping on target regardless of the metric used (12.35 ± 1.58% of all reads; 22.49 ± 1.65% of mapped reads; Figure 2). Bones (2.68 ± 0.78% of all reads; 18.13 ± 1.41% of mapped reads) and nasal bones (0.50 ± 0.11% of all reads; 19.68 ± 1.18% of mapped reads) worked less successfully; however, differences were not significant among the three sample types (Figure 2 and Supplementary Tables 6, 7). For the two contemporary control samples, the proportion of reads mapping on target regions was 17.54 ± 0.01% of all available reads and 53.42 ± 0.69% of mapped reads (Supplementary Table 5).

Genotyping errors and coverage effect

We were able to generate genotypes for a panel of 99,982 genome-wide SNPs (Supplementary Figure 4) using the target enrichment approach, with distinct levels of missing data across samples. For contemporary control samples, concordance rates between genotypes obtained from the capture array developed in this study and those from the Canine HD BeadChip were above 99%. Error rates for the two control samples (i.e., rates of genotype discordance) were almost negligible but increased with decreasing coverage at the target sites (0.64 and 0.91%

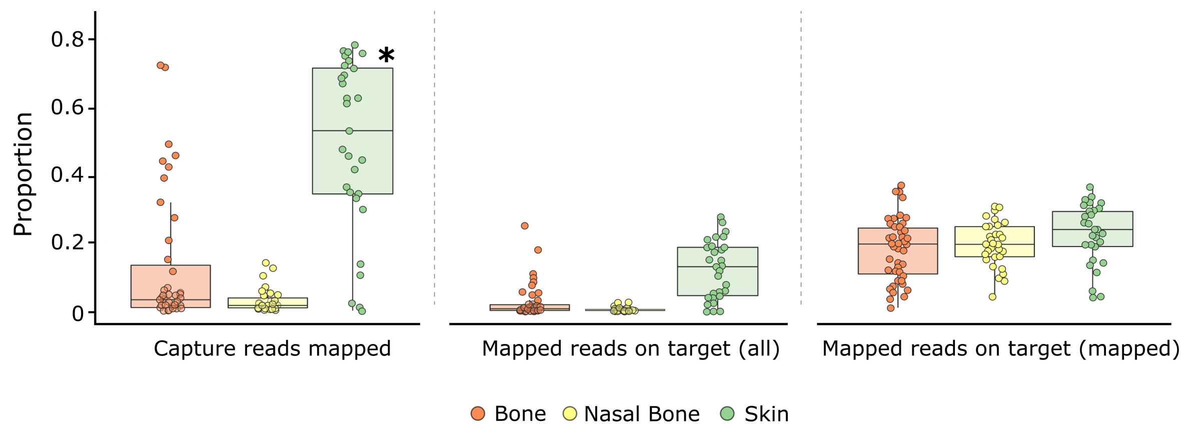


FIGURE 2
Genomic metrics assessed from capture enrichment for each tissue type: bones (orange), maxilloturbinal (nasal) bones (yellow), and skins (green). Boxplots display the proportion of reads mapping after capture enrichment and reads mapping on target region in relation to all or mapped reads for each tissue. Within boxplots, dots represent historical samples, and the central line indicates the median value. Significant differences (p < 0.05) between tissues for each metric are identified with an asterisk.

error rate, at coverage ≥ 5x and ≥ 3x, respectively; Figure 3A and Supplementary Table 8). When calculating the genotype concordance rates using SNPs called exclusively with 3x and 4x, we found a decrease to 94.1 and 96.4%, respectively (Supplementary Table 8). The most common error found in genotypes called with the lowest coverage (3x) was the dropout of a second allele (miscalled homozygous in relation to the SNP chip), corresponding to 79.5% of the observed discordances. The number of loci obtained in contemporary control samples increased as the coverage decreased; for example, the average number of SNPs obtained across both samples declined from 93,948 to 88,173 at coverage ≥ 3x and ≥ 5x, respectively (Figure 3B and Supplementary Table 9).

We found the same pattern across historical samples, with increased estimates of potential genotyping error rates associated with SNPs called with decreasing coverage (Figure 3A). Nevertheless, the highest error rate estimate for coverage ≥ 3x did not reach 7%. Concordantly, we also found an increase in the average number of SNPs across all samples for lower read depths (24,663 and 18,523 SNPs at coverage ≥ 3x and ≥ 5x; Figure 3B and Supplementary Table 9). Overall, the average number of SNPs obtained across historical samples was substantially lower than in contemporary control samples. Still, the highest genotyping success rates were found among three historical samples, which presented very similar rates (>95%) to those found in the contemporary samples (Figure 3B and Supplementary Table 9). Skin samples presented the highest average numbers of SNPs genotyped (Figure 3B). Genotyping success rate (number of genotyped SNPs) was positively correlated with the proportion of reads mapping to the genome after capture (r^2 = 0.85; Supplementary Figure 3B).

Discussion

In this study, we evaluated the performance of different historical tissues—bones, nasal bones, and skins—across

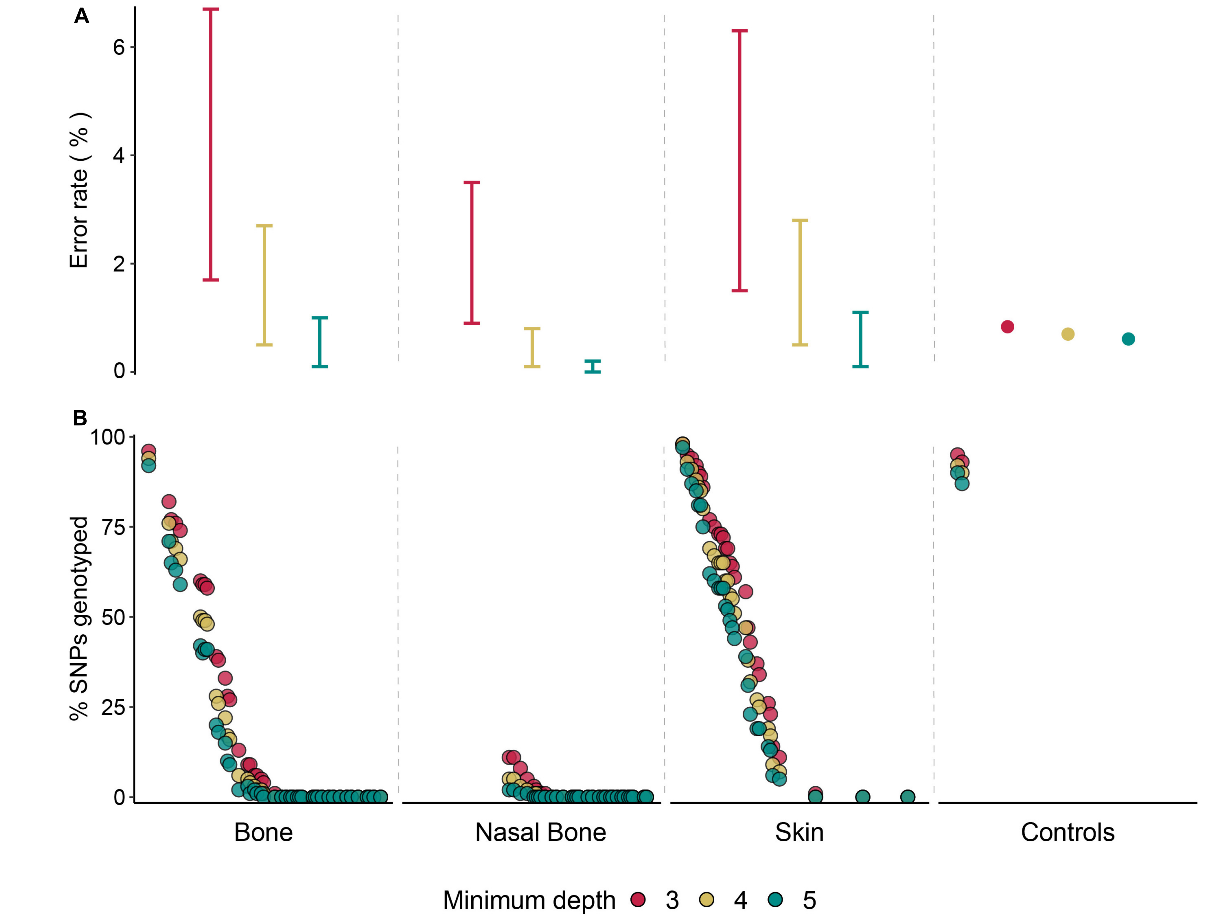


FIGURE 3
Genotyping performance of hDNA from three sample types. **(A)** Interval of potential genotyping error rate estimates (%) for the three hDNA sample types, and average empirical genotyping error rate for control samples, considering the three minimum depth levels. **(B)** Percentage of genotyped SNPs (*N* = 99,982) considering a minimum depth of 3x, 4x, or 5x per SNP across 103 samples. Each dot represents a sample.

a population sampling of Iberian wolf NHCs, in whole-genome sequencing and target enrichment. Specifically, we analyzed metrics based on endogenous DNA content, uniquely high-quality mapped reads after capture, and enrichment proportions.

Bones, nasal bones, and skins all succeeded as DNA sources but differed in endogenous DNA content. Based on equal starting amounts of DNA for shotgun sequencing, skins yielded significantly higher amounts of endogenous DNA than did the other sample types. This concords with other studies that have previously shown that soft tissues of mammals (van der Valk et al., 2017) and birds (Tsai et al., 2020) provide higher endogenous DNA content than hard tissues, although these are based on less comprehensive sample sizes than the one used here. Although nasal bones were initially recommended for presenting higher genotyping success rates than other bones (Wisely et al., 2004), a recent study showed conflicting results (Lonsinger et al., 2019). Here, we assessed for the first time the proportion of endogenous DNA retrieved from nasal bones, showing that this bone type provides lower content than other commonly used bones. Despite the observed significant effect of tissue type on endogenous DNA, we cannot disregard the impact of distinct collection histories in this result (Raxworthy and Smith, 2021). We did not observe significant associations between endogenous DNA content, or its mapping length, and the original collection year, suggesting no substantial DNA degradation across the period under evaluation.

There was no direct effect of tissue type on capture efficiency, with most samples presenting an enrichment > 10x. This result emphasizes the efficiency of target capture as a powerful method for building genomic datasets from hDNA and is in line with other studies using capture approaches (Carpenter et al., 2013; van der Valk et al., 2017). An *a priori* understanding of the importance of endogenous DNA content in the success of a capture experiment (Hernandez-Rodriguez et al., 2018) has driven our decision to perform shotgun sequencing prior to capture enrichment. Using this approach, we were able to demonstrate a perfect association between the number of resulting SNPs and the initial amount of endogenous DNA, emphasizing that this is a practical and cost-effective way to select samples prior to capture experiments. Furthermore, knowing the endogenous DNA content across samples also allows to minimize variation within each sequencing experiment, ensuring low rates of index hopping and reducing possible bias in downstream analysis (van der Valk et al., 2020). Our SNP genotyping results demonstrate that historical samples with high amounts of endogenous DNA can behave similarly to, or even better than, contemporary samples in capture procedures, further supporting the use of capture enrichment to genotype thousands of genome-wide SNPs in hDNA samples (Bi et al., 2013; Smith et al., 2014; Harvey et al., 2016; Lim and Braun, 2016).

The final number of SNPs in a dataset can be a trade-off between coverage and genotyping error rates: relaxing the minimum read depth to increase the number of SNPs is accompanied by higher uncertainty in genotype calling. Still, we were able to generate an average of ca. 25,000 genome-wide SNPs per sample, called with $\geq$ 3x coverage and with low potential error rates (<7%). Reducing coverage led to an increase in the number of SNPs, mostly across samples with intermediate genotyping success, i.e., those where SNPs were captured but read depth was generally low. In such cases, decreasing the coverage from $\geq$ 4x to $\geq$ 3x represented an increase of ~11,000 SNPs. Regardless of the threshold used for minimum coverage, genotypes obtained for the contemporary control samples were confirmed and validated against the SNP chip genotypes. The average error rates found in the controls were within, or very close to, the potential error rate intervals estimated for the historical dataset. This overlap suggests that the use of such an analytical approach (Puritz et al., 2014) can be a reliable alternative to estimate potential error rates when no control samples are available. Observed error rates in our work for $\geq$ 5x coverage, which is a widely accepted threshold for SNP calling (Yi et al., 2010), were much lower than those reported by Fountain et al. (2016) at the same coverage for fresh tissue samples.

The most common error found among control samples was the dropout of a second allele after target enrichment. Given the overall low error rates, we predict minimal impact associated with allelic dropouts in downstream analysis. However, even low rates of genotyping error tend to overestimate genetic variation and can affect population genetic studies in different ways, or, to a greater extent linkage and association studies (Pompanon et al., 2005; Gautier et al., 2013). Since most studies do not have control samples available to calculate empirical concordance rates and validate genotypes (Fountain et al., 2016), predicting the effects of errors might be difficult, and in this case higher minimal depth thresholds (>5x) may be considered. Choosing between increasing the number of loci or having high reliability in the genotypes should be considered on a case-by-case basis to ensure compatibility with downstream applications. Alternatively, historical datasets can be analyzed based on genotype likelihoods instead of genotypes, in order to take the inherent uncertainty of the genotypes into account (Nielsen et al., 2011).

Reducing sample damage without compromising data recovery

Skin samples were revealed to be a good source of endogenous DNA while still being minimally destructive to specimens, as small patches can be easily sampled from non-unique morphological features of hides or skin mounts. Bone sampling is generally more destructive and does not necessarily translate into higher endogenous DNA content. This applies in particular to nasal bones, which yield the lowest endogenous DNA content while relying on a quite destructive sampling process in which a large part of the structure is removed. Although we acknowledge that the consumptive sampling of nasal bones does not compromise the utility of a specimen for morphometric or character studies (Wisely et al., 2004), our results discourage their sampling for genomic studies of mammalian carnivores.

As the demand for NHCs in molecular studies increases, conscious sampling of specimens should now become routine to not compromise their future use. The drilling procedure we used for bone sampling left a small hole in the bone tissue (**Figure 1A** and **Supplementary Figure 1**) but did not hamper future morphometric studies. This sampling approach is less damaging than cutting bone fragments (**Supplementary Figure 1**) and still ensures that enough material is collected for genomic analysis without further manipulation. Sampling the petrous bone, a recognized excellent source of endogenous DNA (Pinhasi et al., 2015), would only be recommendable for NHC specimens if entailing the use of damaged skulls, as its sampling is highly destructive to the skull (Charlton et al., 2019). The starting amount of endogenous DNA can also be maximized through wet lab procedures by performing several rounds of DNA extraction, creating multiple and differentially indexed DNA libraries to increase complexity levels, and/or captures per sample (Hernandez-Rodriguez et al., 2018; White et al., 2019; Fontsere et al., 2021; von Seth et al., 2021). Understanding that a reduced fragment of the appropriate tissue is enough to successfully recover molecular data without compromising the reusability of the specimen is of remarkable importance for managing NHCs.

Limitations and opportunities for further development

In contrast to current practices for preserving biological material in controlled environments and using sophisticated resources that prevent DNA damage, older traditional methods did not prioritize DNA integrity (Hall et al., 1997; Burrell et al., 2015; Card et al., 2021). Thus, these preservation techniques and storage conditions can greatly impact the quality and quantity of DNA. In this work, we used wolf specimens from three museums and 15 private collections distributed across a 90-year period, where most samples (∼60%) were collected between 1960 and 1980. These samples likely have different collection histories, particularly those in private collections, where less-standardized preservation methods can be expected. Such heterogeneity may explain the observed levels of variability in endogenous DNA content, although our sample size hampers a statistical evaluation of these metrics across sampling origins. We acknowledge that it would be relevant to assess the effects of different preservation techniques on DNA degradation; unfortunately, information about the method applied to each specimen used for this study was not available, as is often the case in NHCs.

One of the major challenges the museum community currently faces is improving the documentation of collection histories for individual specimens and to make it accessible to the scientific community. Global initiatives have been promoting the digitization of specimens' metadata and digital images to make them available in electronic databases that can be easily accessed (e.g., The Global Information Facility; Robertson et al., 2014). As this process evolves, it is essential that researchers working in the field of museomics synergize genomic data with NHCs metadata to enhance the scientific impact and traceability of their studies by, for example, always providing the catalog numbers of specimens used (Card et al., 2021).

Concluding remarks

To our knowledge, this study uses the most comprehensive dataset to date—in terms of sample size and genome representation—to test the performance of three wolf tissues as sources of hDNA. Based on our findings, we recommend the use of skins for sampling mammalian carnivore specimens, as these are reliable and minimally destructive sources of endogenous DNA suitable for whole-genome and target enrichment approaches. This study should also encourage future research with the same aims but targeting different vertebrate and invertebrate groups. In addition, we provide a validated genome-wide SNP tool (i.e., probe design) that allows for direct comparison between historical and contemporary data. Although the enrichment approach presented here was based on canid genomes, its conceptual design can be implemented in any species for which SNP chips are available. We believe that the increasing demand for NHCs as DNA sources, and the requirements for minimal damage to the specimens, should encourage the generation of genomic datasets comparable among studies.

Data availability statement

Individual SNP genotypes and the sequence of the 400,000 RNA probes are available from the OSF repository: https://osf.io/j3r7x/?view_only=a747acf5297c49309bbb89f2e7414104. Raw sequence reads from whole-genome resequencing are available on NCBI (accession SRA PRJNA860381): https://www.ncbi.nlm.nih.gov/sra/PRJNA860381. Custom python scripts (gtvalues2plink.py and SNP_concordance.py) used in this study are available at https://github.com/pdroslva84/SNPcap.

Author contributions

RG coordinated the project. RG, CP, DL, and PS designed the study. RG, JL-B, and FÁ coordinated historical sample collection efforts. RG, JL-B, EG, FÁ, JL, and DL collected samples from historical specimens. CP, DL, and DC conducted laboratory work. CP and DL performed data analysis under the guidance of RG and PS. PS did the bioinformatic scripting. CP, DL, and RG wrote the manuscript with input from all the other authors. All authors contributed to the article and approved the submitted version.

Funding

This work was co-funded by the Portuguese Foundation for Science and Technology (FCT) under the project PTDC/BIA-EVF/2460/2014 and the project NORTE-01-0246-FEDER-000063, the Norte Portugal Regional Operational Programme (NORTE2020), under the PORTUGAL 2020 Partnership Agreement, through the European Regional Development Fund (ERDF). DL, CP, RG, and PS were supported by the FCT (PD/BD/132403/2017, PD/BD/135026/2017, 2021/00647/CEECIND, and PTDC/BIA-EVL/31902/2017, respectively). JL-B was supported by the Spanish Ministry of Economy, Industry and Competitiveness (RYC-2015-18932), and the GRUPIN research grant from the Regional Government of Asturias (IDI/2021/000075).

Acknowledgments

We acknowledge the Museu Nacional de História Natural e da Ciência (Lisbon, Portugal), Museo Nacional de Ciencias Naturales (Madrid, Spain), Estación Biológica de Doñana (Seville, Spain) and all owners of private collections who shared historical wolf samples. The two contemporary tissue samples were provided by the Portuguese Institute for Nature Conservation and Forestry (ICNF) and the Consejería de Medio Ambiente de la Junta de Castilla y León. We are grateful to R. Pedra and S. Roque for helping collecting samples, and to S. Mourão and M. Magalhães for lab assistance.

Conflict of interest

JL was employed by Consultores en Iniciativas Ambientales, S.L.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fevo.2022.970249/full#supplementary-material

References

Allentoft, M. E., Collins, M., Harker, D., Haile, J., Oskam, C. L., Hale, M. L., et al. (2012). The half-life of DNA in bone: Measuring decay kinetics in 158 dated fossils. *Proc. R. Soc. B Biol. Sci.* 279, 4724–4733. doi: 10.1098/rspb.2012.1745

Andrews, S. (2018). *FastQC: A quality control tool for high throughput sequence data (v. 0.11.7).* Berlin: ScienceOpen, Inc.

Bi, K., Linderoth, T., Singhal, S., Vanderpool, D., Patton, J. L., Nielsen, R., et al. (2019). Temporal genomic contrasts reveal rapid evolutionary responses in an alpine mammal during recent climate change. *PLoS Genet.* 15:e1008119. doi: 10.1371/journal.pgen.1008119

Bi, K., Linderoth, T., Vanderpool, D., Good, J. M., Nielsen, R., and Moritz, C. (2013). Unlocking the vault: Next-generation museum population genomics. *Mol. Ecol.* 22, 6018–6032. doi: 10.1111/mec.12516

Burrell, A. S., Disotell, T. R., and Bergey, C. M. (2015). The use of museum specimens with high-throughput DNA sequencers. *J. Hum. Evol.* 79, 35–44. doi: 10.1016/j.jhevol.2014.10.015

Card, D. C., Shapiro, B., Giribet, G., Moritz, C., and Edwards, S. V. (2021). Museum Genomics. *Annu. Rev. Genet.* 55, 633–659. doi: 10.1146/annurev-genet-071719-020506

Carpenter, M. L., Buenrostro, J. D., Valdiosera, C., Schroeder, H., Allentoft, M. E., Sikora, M., et al. (2013). Pulling out the 1%: Whole-Genome capture for the targeted enrichment of ancient dna sequencing libraries. *Am. J. Hum. Genet.* 93, 852–864. doi: 10.1016/j.ajhg.2013.10.002

Casas-Marce, M., Revilla, E., and Godoy, J. A. (2010). Searching for DNA in museum specimens: A comparison of sources in a mammal species. *Mol. Ecol. Resour.* 10, 502–507. doi: 10.1111/j.1755-0998.2009.02784.x

Casas-Marce, M., Revilla, E., Fernandes, M., Rodriguez, A., Delibes, M., and Godoy, J. A. (2012). The value of hidden scientific resources: Preserved animal specimens from private collections and small museums. *Bioscience* 62, 1077–1082. doi: 10.1525/bio.2012.62.12.9

Charlton, S., Booth, T., and Barnes, I. (2019). The problem with petrous? A consideration of the potential biases in the utilization of pars petrosa for ancient DNA analysis. *World Archaeol.* 51, 574–585. doi: 10.1080/00438243.2019.1694062

Cruz-Dávalos, D. I., Llamas, B., Gaunitz, C., Fages, A., Gamba, C., Soubrier, J., et al. (2017). Experimental conditions improving in-solution target enrichment for ancient DNA. *Mol. Ecol. Resour.* 17, 508–522. doi: 10.1111/1755-0998.12595

Dabney, J., and Meyer, M. (2019). Extraction of highly degraded DNA from ancient bones and teeth. *Methods Mol. Biol.* 1963, 25–29. doi: 10.1007/978-1-4939-9176-1_4

Dabney, J., Knapp, M., Glocke, I., Gansauge, M. T., Weihmann, A., Nickel, B., et al. (2013). Complete mitochondrial genome sequence of a Middle Pleistocene cave bear reconstructed from ultrashort DNA fragments. *Proc. Natl. Acad. Sci. U. S. A.* 110, 15758–15763. doi: 10.1073/pnas.1314445110

Dehasque, M., Pečnerová, P., Kempe Lagerholm, V., Ersmark, E., Danilov, G. K., Mortensen, P., et al. (2022). Development and Optimization of a Silica Column-Based Extraction Protocol for Ancient DNA. *Genes* 13:687. doi: 10.3390/genes13040687

DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., et al. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* 43, 491–498. doi: 10.1038/ng.806

Derkarabetian, S., Benavides, L. R., and Giribet, G. (2019). Sequence capture phylogenomics of historical ethanol-preserved museum specimens: Unlocking the rest of the vault. *Mol. Ecol. Resour.* 19, 1531–1544. doi: 10.1111/1755-0998.13072

Eisenhofer, R., Minich, J. J., Marotz, C., Cooper, A., Knight, R., and Weyrich, L. S. (2019). Contamination in Low Microbial Biomass Microbiome Studies: Issues and Recommendations. *Trends Microbiol.* 27, 105–117. doi: 10.1016/j.tim.2018.11.003

Ewels, P., Magnusson, M., Lundin, S., and Käller, M. (2016). MultiQC: Summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* 32, 3047–3048. doi: 10.1093/bioinformatics/btw354

Farrington, O. C. (1915). The rise of Natural History Museums. *Science* 42, 197–208. doi: 10.1126/science.42.1076.197

Fontsere, C., Alvarez-Estape, M., Lester, J., Arandjelovic, M., Kuhlwilm, M., Dieguez, P., et al. (2021). Maximizing the acquisition of unique reads in noninvasive capture sequencing experiments. *Mol. Ecol. Resour.* 21, 745–761. doi: 10.1111/1755-0998.13300

Fountain, E. D., Pauli, J. N., Reid, B. N., Palsbøll, P. J., and Peery, M. Z. (2016). Finding the right coverage: The impact of coverage and sequence quality on single nucleotide polymorphism genotyping error rates. *Mol. Ecol. Resour.* 16, 966–978. doi: 10.1111/1755-0998.12519

Gautier, M., Gharbi, K., Cezard, T., Foucaud, J., Kerdelhué, C., Pudlo, P., et al. (2013). The effect of RAD allele dropout on the estimation of genetic variation within and between populations. *Mol. Ecol.* 22, 3165–3178. doi: 10.1111/mec.12089

Gel, B., and Serra, E. (2017). KaryoploteR: An R/Bioconductor package to plot customizable genomes displaying arbitrary data. *Bioinformatics* 33, 3088–3090. doi: 10.1093/bioinformatics/btx346

Haak, W., Lazaridis, I., Patterson, N., Rohland, N., Mallick, S., Llamas, B., et al. (2015). Massive migration from the steppe was a source for Indo-European languages in Europe. *Nature* 522, 207–211. doi: 10.1038/nature14317

Hall, L. M., Willcox, M. S., and Jones, D. S. (1997). Association of enzyme inhibition with methods of museum skin preparation. *Biotechniques* 22, 928–934. doi: 10.2144/97225st07

Harvey, M. G., Smith, B. T., Glenn, T. C., Faircloth, B. C., and Brumfield, R. T. (2016). Sequence Capture versus Restriction Site Associated DNA Sequencing for Shallow Systematics. *Syst. Biol.* 65, 910–924. doi: 10.1093/sysbio/syw036

Heng, L. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM*. *arXiv* [Preprint]. doi: 10.48550/arXiv.1303.3997

Hernandez-Rodriguez, J., Arandjelovic, M., Lester, J., de Filippo, C., Weihmann, A., Meyer, M., et al. (2018). The impact of endogenous content, replicates and pooling on genome capture from faecal samples. *Mol. Ecol. Resour.* 18, 319–333. doi: 10.1111/1755-0998.12728

Holmes, M. W., Hammond, T. T., Wogan, G. O. U., Walsh, R. E., Labarbera, K., Wommack, E. A., et al. (2016). Natural history collections as windows on evolutionary processes. *Mol. Ecol.* 25, 864–881. doi: 10.1111/mec.13529

Horváth, M. B., Martínez-Cruz, B., Negro, J. J., Kalmár, L., and Godoy, J. A. (2005). An overlooked DNA source for non-invasive genetic analysis in birds. *J. Avian Biol.* 36, 84–88. doi: 10.1111/j.0908-8857.2005.03370.x

Hung, C. M., Shaner, P. J. L., Zink, R. M., Liu, W. C., Chu, T. C., Huang, W. S., et al. (2014). Drastic population fluctuations explain the rapid extinction of the passenger pigeon. *Proc. Natl. Acad. Sci. U. S. A.* 111, 10636–10641. doi: 10.1073/pnas.1401526111

Jones, M. R., and Good, J. M. (2016). Targeted capture in evolutionary and ecological genomics. *Mol. Ecol.* 25, 185–202. doi: 10.1111/mec.13304

Kassambara, A. (2020). *ggpubr R Package: Ggplot2-Based Publication Ready Plots*. Available online at: https://rpkgs.datanovia.com/ggpubr/ (accessed July, 2021).

Kircher, M., Sawyer, S., and Meyer, M. (2012). Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform. *Nucleic Acids Res.* 40:e3. doi: 10.1093/nar/gkr771

Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* 27, 2987–2993. doi: 10.1093/bioinformatics/btr509

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25, 2078–2079. doi: 10.1093/bioinformatics/btp352

Lim, H. C., and Braun, M. J. (2016). High-throughput SNP genotyping of historical and modern samples of five bird species *via* sequence capture of ultraconserved elements. *Mol. Ecol. Resour.* 16, 1204–1223. doi: 10.1111/1755-0998.12568

Lindblad-Toh, K., Wade, C. M., Mikkelsen, T. S., Karlsson, E. K., Jaffe, D. B., Kamal, M., et al. (2005). Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature* 438, 803–819. doi: 10.1038/nature04338

Lonsinger, R. C., Daniel, D., Adams, J. R., and Waits, L. P. (2019). Consideration of sample source for establishing reliable genetic microsatellite data from mammalian carnivore specimens held in natural history collections. *J. Mammal.* 100, 1678–1689. doi: 10.1093/jmammal/gyz112

Lopez, L., Turner, K. G., Bellis, E. S., and Lasky, J. R. (2020). Genomics of natural history collections for understanding evolution in the wild. *Mol. Ecol. Resour.* 20, 1153–1160. doi: 10.1111/1755-0998.13245

Maebe, K., Meeus, I., Vray, S., Claeys, T., Dekoninck, W., Boevé, J. L., et al. (2016). A century of temporal stability of genetic diversity in wild bumblebees. *Sci. Rep.* 6:38289. doi: 10.1038/srep38289

McCormack, J. E., Tsai, W. L. E., and Faircloth, B. C. (2016). Sequence capture of ultraconserved elements from bird museum specimens. *Mol. Ecol. Resour.* 16, 1189–1203. doi: 10.1111/1755-0998.12466

McDonough, M. M., Parker, L. D., McInerney, N. R., Campana, M. G., and Maldonado, J. E. (2018). Performance of commonly requested destructive museum samples for mammalian genomic studies. *J. Mammal.* 99, 789–802. doi: 10.1093/jmammal/gyy080

Meyer, M., and Kircher, M. (2010). Illumina sequencing library preparation for highly multiplexed target capture and sequencing. *Cold Spring Harb. Protoc.* 2010:db.rot5448. doi: 10.1101/pdb.prot5448

Nielsen, R., Paul, J. S., Albrechtsen, A., and Song, Y. S. (2011). Genotype and SNP calling from next-generation sequencing data. *Nat. Rev. Genet.* 12, 443–451. doi: 10.1038/nrg2986

Pálsdóttir, A. H., Bläuer, A., Rannamäe, E., Boessenkool, S., and Hallsson, J. H. (2019). Not a limitless resource: Ethics and guidelines for destructive sampling of archaeofaunal remains. *R. Soc. Open Sci.* 6:191059. doi: 10.1098/rsos.191059

Pearson, K. D., Nelson, G., Aronson, M. F. J., Bonnet, P., Brenskelle, L., Davis, C. C., et al. (2020). Machine learning using digitized herbarium specimens to advance phenological research. *Bioscience* 70, 610–620. doi: 10.1093/biosci/biaa044

Pinhasi, R., Fernandes, D., Sirak, K., Novak, M., Connell, S., Alpaslan-Roodenberg, S., et al. (2015). Optimal ancient DNA yields from the inner ear part of the human petrous bone. *PLoS One* 10:e0129102. doi: 10.1371/journal.pone.0129102

Pompanon, F., Bonin, A., Bellemain, E., and Taberlet, P. (2005). Genotyping errors: Causes, consequences and solutions. *Nat. Rev. Genet.* 6, 847–859. doi: 10.1038/nrg1707

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., et al. (2007). PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *Am. J. Hum. Genet.* 81, 559–575. doi: 10.1086/519795

Puritz, J. B., Hollenbeck, C. M., and Gold, J. R. (2014). dDocent: A RADseq, variant-calling pipeline designed for population genomics of non-model organisms. *PeerJ* 2014:e431. doi: 10.7717/peerj.431

R Development Core Team (2017). *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing.

Raxworthy, C. J., and Smith, B. T. (2021). Mining museums for historical DNA: Advances and challenges in museomics. *Trends Ecol. Evol.* 36, 1049–1060. doi: 10.1016/j.tree.2021.07.009

Robertson, T., Döring, M., Guralnick, R., Bloom, D., Wieczorek, J., Braak, K., et al. (2014). The GBIF Integrated Publishing Toolkit: Facilitating the Efficient

Publishing of Biodiversity Data on the Internet. *PLoS One* 9:e102623. doi: 10.1371/journal.pone.0102623

Rowe, K. C., Singhal, S., Macmanes, M. D., Ayroles, J. F., Morelli, T. L., Rubidge, E. M., et al. (2011). Museum genomics: Low-cost and high-accuracy genetic data from historical specimens. *Mol. Ecol. Resour.* 11, 1082–1092. doi: 10.1111/j.1755-0998.2011.03052.x

Schubert, M., Ermini, L., Der Sarkissian, C., Jónsson, H., Ginolhac, A., Schaefer, R., et al. (2014). Characterization of ancient and modern genomes by SNP detection and phylogenomic and metagenomic analysis using PALEOMIX. *Nat. Protoc.* 9, 1056–1082. doi: 10.1038/nprot.2014.063

Schubert, M., Lindgreen, S., and Orlando, L. (2016). AdapterRemoval v2: Rapid adapter trimming, identification, and read merging. *BMC Res. Notes* 9:88. doi: 10.1186/s13104-016-1900-2

Schwartz, M. K., Luikart, G., and Waples, R. S. (2007). Genetic monitoring as a promising tool for conservation and management. *Trends Ecol. Evol.* 22, 25–33. doi: 10.1016/j.tree.2006.08.009

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

Staats, M., Erkens, R. H. J., Vossenberg, B., van de Wieringa, J. J., Kraaijeveld, K., Stielow, B., et al. (2013). Genomic Treasure Troves: Complete Genome Sequencing of Herbarium and Insect Museum Specimens. *PLoS One* 8:e69189. doi: 10.1371/JOURNAL.PONE.0069189

Suarez, A., and Tsutsui, N. (2004). The Value of Museum Collections for Research and Society. *Bioscience* 54, 66–74. doi: 10.1641/0006-35682004054

Tsai, W. L. E., Schedl, M. E., Maley, J. M., and McCormack, J. E. (2020). More than skin and bones: Comparing extraction methods and alternative sources of DNA from avian museum specimens. *Mol. Ecol. Resour.* 20, 1220–1227. doi: 10.1111/1755-0998.13077

Tsangaras, K., and Greenwood, A. D. (2012). Museums and disease: Using tissue archive and museum samples to study pathogens. *Ann. Anat.* 194, 58–73. doi: 10.1016/j.aanat.2011.04.003

Vallender, E. J. (2011). Expanding whole exome resequencing into non-human primates. *Genome Biol.* 12:R87. doi: 10.1186/GB-2011-12-9-R87

van der Valk, T., Lona Durazo, F., Dalén, L., and Guschanski, K. (2017). Whole mitochondrial genome capture from faecal samples and museum-preserved specimens. *Mol. Ecol. Resour.* 17, e111–e121. doi: 10.1111/1755-0998.12699

van der Valk, T., Vezzi, F., Ormestad, M., Dalén, L., and Guschanski, K. (2020). Index hopping on the Illumina HiseqX platform and its consequences for ancient DNA studies. *Mol. Ecol. Resour.* 20, 1171–1181. doi: 10.1111/1755-0998.13009

von Seth, J., Dussex, N., Díez-del-Molino, D., van der Valk, T., Kutschera, V. E., Kierczak, M., et al. (2021). Genomic insights into the conservation status of the world's last remaining Sumatran rhinoceros populations. *Nat. Commun.* 12:2393. doi: 10.1038/s41467-021-22386-8

Wandeler, P., Hoeck, P. E. A., and Keller, L. F. (2007). Back to the future: Museum specimens in population genetics. *Trends Ecol. Evol.* 22, 634–642. doi: 10.1016/j.tree.2007.08.017

Weiß, C. L., Schuenemann, V. J., Devos, J., Shirsekar, G., Reiter, E., Gould, B. A., et al. (2016). Temporal patterns of damage and decay kinetics of dna retrieved from plant herbarium specimens. *R. Soc. Open Sci.* 3:160239. doi: 10.1098/rsos.160239

White, L. C., Fontsere, C., Lizano, E., Hughes, D. A., Angedakin, S., Arandjelovic, M., et al. (2019). A roadmap for high-throughput sequencing studies of wild animal populations using noninvasive samples and hybridization capture. *Mol. Ecol. Resour.* 19, 609–622. doi: 10.1111/1755-0998.12993

Wisely, S. M., Maldonado, J. E., and Fleischer, R. C. (2004). A technique for sampling ancient DNA that minimizes damage to museum specimens. *Conserv. Genet.* 5, 105–107. doi: 10.1023/B:COGE.0000014061.04963.da

Yi, X., Liang, Y., Huerta-Sanchez, E., Jin, X., Cuo, Z. X. P., Pool, J. E., et al. (2010). Sequencing of 50 human exomes reveals adaptation to high altitude. *Science* 329, 75–78. doi: 10.1126/science.1190371

TYPE Original Research
PUBLISHED 03 October 2022
DOI 10.3389/fevo.2022.930356

Museomics and the holotype of a critically endangered cricetid rodent provide key evidence of an undescribed genus

Susette Castañeda-Rico[1,2,3]*, Cody W. Edwards[1,3], Melissa T. R. Hawkins[3,4] and Jesús E. Maldonado[1,2,3]

[1]Smithsonian-Mason School of Conservation, Front Royal, VA, United States, [2]Center for Conservation Genomics, Smithsonian National Zoo and Conservation Biology Institute, Washington, DC, United States, [3]Department of Biology, George Mason University, Fairfax, VA, United States, [4]Division of Mammals, Department of Vertebrate Zoology, National Museum of Natural History, Washington, DC, United States

OPEN ACCESS

EDITED BY
Jonathan J. Fong,
Lingnan University, China

REVIEWED BY
Andreanna J. Welch,
Durham University, United Kingdom
Mariana Lyra,
São Paulo State University, Campus Rio Claro, Brazil

*CORRESPONDENCE
Susette Castañeda-Rico
susetteazul@gmail.com;
castanedaricos@si.edu

SPECIALTY SECTION
This article was submitted to Phylogenetics, Phylogenomics, and Systematics, a section of the journal Frontiers in Ecology and Evolution

RECEIVED 27 April 2022
ACCEPTED 09 September 2022
PUBLISHED 03 October 2022

CITATION
Castañeda-Rico S, Edwards CW, Hawkins MTR and Maldonado JE (2022) Museomics and the holotype of a critically endangered cricetid rodent provide key evidence of an undescribed genus. *Front. Ecol. Evol.* 10:930356. doi: 10.3389/fevo.2022.930356

Historical DNA obtained from voucher specimens housed in natural history museums worldwide have allowed the study of elusive, rare or even extinct species that in many cases are solely represented by museum holdings. This has resulted in the increase of taxonomic representation of many taxa, has led to the discovery of new species, and has yielded stunning novel insights into the evolutionary history of cryptic or even undescribed species. *Peromyscus mekisturus*, is a critically endangered cricetid rodent endemic to Mexico and is only known from two museum specimens collected in 1898 and 1947. Intensive field work efforts to attempt to determine if viable populations still exist have failed, suggesting that this species is extinct or is nearing extinction. In addition, a recent study using mitogenomes demonstrated that *P. mekisturus* forms a well-supported clade outside the genus *Peromyscus* and hypothesized that this taxon is the sister group of the genus *Reithrodontomys*. Here, we used target enrichment and high-throughput sequencing of several thousand nuclear ultraconserved elements and mitogenomes to reconstruct dated phylogenies to test the previous phylogenetic hypothesis. We analyzed the holotype and the only other known specimen of *P. mekisturus* and museum samples from other peromyscine rodents to test the phylogenetic position of the species. Our results confirm that the only two specimens known to science of *P. mekisturus* belong to the same species and support the hypothesis that this species belongs to an undescribed genus of cricetid rodents that is sister to the genus *Reithrodontomys*. We dated the origin of *P. mekisturus* together with other speciation events in peromyscines during the late Pliocene – early Pleistocene and related these events with the Pleistocene climatic cycles. In light of our results, we recommend a taxonomic re-evaluation of this enigmatic species to properly recognize its taxonomic status as a new genus. We also acknowledge the relevance of generating genomic data from type specimens and highlight the need and importance of continuing to build the scientific heritage of the collections to study and better understand past, present, and future biodiversity.

KEYWORDS

historical DNA, genomics, mitogenomes, *Peromyscus mekisturus,* type specimen, ultraconserved elements

Introduction

Museomics is a booming field that leverages the potential of natural history museums as a source of DNA (ancient DNA – aDNA – naturally preserved, heavily degraded trace amounts with both low quality and quantity yields, and usually between thousands to a million years old; historical DNA – hDNA – fortuitously preserved in voucher specimens almost always collected during the last 200 years, highly degraded with both low quality and quantity yields; and modern DNA – mDNA – tissues stored frozen or in preservatives, usually of high DNA quality and quantity, but in some cases, they can be affected by the mode of preservation regardless of time) coupled with genomic methods and techniques (Schmitt et al., 2018; Raxworthy and Smith, 2021). It has transformed the field of collection-based research, extending research possibilities for paleontological and natural history specimens (Buerki and Baker, 2016; Rubi et al., 2020). Museomics-based research has yielded new insights into the evolutionary history of organisms and has greatly impacted our knowledge regarding the tree of life, filling gaps in the majority of its branches and revealing unknown or controversial phylogenetic positions (Buerki and Baker, 2016; Kehlmaier et al., 2019; Cong et al., 2021).

This innovative tool has been applied to discover and delimit species (Abreu-Jr et al., 2020; Lyra et al., 2020; McDonough et al., 2022), to sample extinct species (Roycroft et al., 2021) and extirped populations (Shepherd and Lambert, 2008), to clarified taxonomic classifications with type specimens (Prosser et al., 2016; Kehlmaier et al., 2019). It has also been used in population genetic studies (Yuan et al., 2022), to document changes in genetic diversity through time (Schmitt et al., 2018; Bi et al., 2019) and the species' response to environmental change and genetic erosion (Bi et al., 2013; Dussex et al., 2019). Museomics has even been used to track the origins and spread of infectious diseases (Schmitt et al., 2018; Karwacki et al., 2021), and to investigate epigenetic effects (Rubi et al., 2020).

In the current biodiversity crisis, the discovery and documentation of biodiversity on earth should be a priority (Campana et al., 2021). It is of great concern that many species could be lost before they or their ecological roles have been described, without even being aware of what is being lost (Kehlmaier et al., 2019). Although accurate species identification should be the backbone of biodiversity research it is not sufficient to just identify and count these species, but we also need to better understand their evolutionary and environmental history. In this sense, the use of type specimens, within a taxonomic and phylogenetic framework, is essential to ensure the accurate identification of specimens (Buerki and Baker, 2016; Kehlmaier et al., 2019). Type specimens (or simply referred to as types) are the exemplar specimens that are representative of the species description, and as such, determine the correct application of nomenclature and represent the link between a name and a taxonomic unit (Buerki and Baker, 2016; Cong et al., 2021). Within types, a holotype is a single specimen designated, in the original publication, as the name-bearing exemplar of a species (International Commission on Zoological Nomenclature[1]). Despite their great contribution to science, the representation of holotypes in genetic studies is scarce. This is due to the impact that "destructive sampling" can have on these invaluable and irreplaceable specimens because it will likely involve damaging or destroying a portion of the specimen to obtain the genomic data. These specimens are, in general, very old – between 10 and 200 years old, and therefore they are understandably highly protected and valued by the curators and collection managers of their museum collections. Since specimens represent finite resources, most museums have strict policies governing destructive sampling, limiting the availability of samples (Holmes et al., 2016). However, recent phylogenetic studies have successfully demonstrated the importance of including type specimens, and as such museum curators are carefully evaluating the proper use and sampling of these unique specimens (e.g., Prosser et al., 2016; McGuire et al., 2018; Kehlmaier et al., 2019; Cong et al., 2021; Reyes-Velasco et al., 2021; Roos et al., 2021; Roycroft et al., 2021).

The Puebla deer mouse, *Peromyscus mekisturus*, is a critically endangered cricetid rodent endemic to Mexico and is only known from two museum specimens. The holotype (Smithsonian Institution's National Museum of Natural History – USNM64108) collected by Merriam (1898) in Chalchicomula (= Ciudad Serdán) and a second individual captured by Hooper (1947) in Tehuacán, both in the state

[1] https://www.iczn.org/the-code/the-code-online/

of Puebla (University of Michigan Museum of Zoology – UMMZ88967). Unfortunately, multiple expeditions after 1947 targeting this species have failed to find more specimens. This suggests that the Puebla deer mouse may have already become extinct or is close to extinction.

Peromyscus mekisturus, based on morphology (Osgood, 1909; Carleton, 1989; Musser and Carleton, 1993, 2005) and on a few mitochondrial genes (Castañeda-Rico et al., 2014), had been traditionally placed within the *Peromyscus melanophrys* group, together with *P. melanophrys* and *P. perfulvus*. However, Castañeda-Rico et al. (2020) using mitogenomes and ultraconserved elements (UCE) obtained from the *P. mekisturus* specimen collected in 1947 [University of Michigan Museum of Zoology – UMMZ88967– the same specimen analyzed by Castañeda-Rico et al. (2014)], found that this species was not part of the *Peromyscus melanophrys* group, as previously suggested. In addition, with a denser sampling of mitogenomes including more cricetid species, they also uncovered that *P. mekisturus* was more closely related to the genera *Reithrodontomys* and *Isthmomys* than to any other member of the genus *Peromyscus*. However, they suggested that the latest results needed to be confirmed with a denser taxon sampling of the nuclear genome. Castañeda-Rico et al. (2020) also found that the mitochondrial sequence obtained by Castañeda-Rico et al. (2014) was incorrect due to (i) cross contamination with other *Peromyscus* samples processed in the same lab during extraction and/or PCR steps, (ii) a chimera sequence product of jumping PCR, and/or (iii) contamination from the environment caused by not performing the extractions in a dedicated facility for ancient DNA analysis.

In this study, we show how museomics has revolutionized phylogenetic studies, improving our understanding of the biodiversity of our planet. Importantly, we demonstrate that holotype specimen data is crucial for confirming the accurate identification of poorly studied species, especially, when it concerns rare, extinct or under-collected species such as *P. mekisturus*. Here, we improved on the previous study by Castañeda-Rico et al. (2020) by obtaining genome-wide data, specifically mitogenomes and thousands of UCE loci from a larger number of representative species of the genus *Peromyscus* and some of their outgroups obtained from specimens in museum collections. We tested the phylogenetic hypothesis that *P. mekisturus* is more closely related to the genera *Reithrodontomys* and *Isthmomys* than it is to members of the genus *Peromyscus*. We analyzed the holotype of *P. mekisturus* and compared it to the previously sequenced museum specimen from Tehuacán, Puebla, in order to confirm its correct identification and the phylogenetic position of the species. Finally, we conducted molecular dating to estimate the timing of the divergence events of *P. mekisturus*. Our results conclusively support the genetic uniqueness of *P. mekisturus* and have important implications for taxonomy and the impact of biodiversity loss.

Materials and methods

Sample collection and laboratory methods

We obtained 12 samples (ca. 2 mm^2 of frozen tissue – internal organ– or dry skin) from specimens deposited at the Smithsonian Institution's National Museum of Natural History and the Museum of Texas Tech University (**Supplementary Table S1**). Sampling comprised of one sample per each species (*Peromyscus attwateri, P. aztecus, P. megalops, P. polionotus, P. crinitus, Neotomodon alstoni, Podomys floridanus, Onychomys leucogaster, Reithrodontomys mexicanus, Isthmomys pirrensis,* and *Neotoma mexicana*), including the holotype specimen of *P. mekisturus* (collected in 1898) see **Supplementary Figure S1**. We selected these species so that we could incorporate all of the species used in mitochondrial phylogeny obtained by Castañeda-Rico et al. (2020) and test their hypothesis using nuclear genome-wide data. We followed strict protocols to avoid cross-contamination during sampling, as described in McDonough et al. (2018) and Castañeda-Rico et al. (2020).

We performed all laboratory work at the Center for Conservation Genomics (CCG), Smithsonian National Zoo and Conservation Biology Institute, Washington, DC. DNA was extracted from frozen-preserved internal organs (i.e., liver or muscle, hereafter modern samples), in the modern lab at the CCG, using a DNeasy Blood and Tissue Kit (Qiagen Inc., Valencia, CA, USA) following the manufacturer's protocol. We conducted all pre-PCR steps for the historical samples in a laboratory specifically dedicated to processing of historical and ancient DNA at the CCG. We extracted DNA from historical samples (i.e., dry skin), using the silica column extraction protocol (McDonough et al., 2018). We quantified DNA samples with a Qubit 4 fluorometer (Thermo Fisher, Waltham, MA, USA) using a 1x dsDNA HS assay and visualized DNA with a TapeStation 4200 System (Agilent Technologies, Santa Clara, CA, USA) using High Sensitivity D1000 reagents. We sheared modern DNA to an average length of 250 base pairs (bp) using a Bioruptor® Pico sonicator (Diagenode Inc., Denville, NJ, USA) with a pulse of 30 s on/30 s off for 90 cycles. We did not shear DNA from historical samples due to its inherent degradation and fragmentation.

We prepared dual-indexed libraries using the Kapa HyperPrep kit (Roche Sequencing) with 1/2 reactions, following the manufacturer's protocol. To library prep the holotype specimen, we used the SRSLY PicoPlus NGS library prep kit (Claret Bioscience, LLC), according to the manufacturer's protocol. We performed dual indexing PCR with TruSeq-style indices (Meyer and Kircher, 2010) using Kapa HiFi HotStart Uracil + (Roche Sequencing) for historical samples and Kapa HiFi HotStart Ready Mix (Roche Sequencing) for modern samples, following the manufacturer's protocol. Libraries were amplified with 8–13 cycles of PCR. We cleaned the indexed libraries using 1.6x solid-phased reversible immobilization (SPRI) magnetic beads (Rohland and Reich, 2012), quantified concentration using a Qubit 4 fluorometer, and inspected size-ranges and quality with a TapeStation 4200 System (conditions as mentioned above). Each capture reaction contained pooled libraries, which consisted of equimolar pools of two individuals for historical samples and three individuals for modern samples. The holotype specimen was not pooled with any other sample and captured alone. We performed target enrichment using the myBaits® UCE Tetrapods 5Kv1 kit (Faircloth et al., 2012) produced by Daicel Arbor Biosciences following the myBaits protocol v3, and the myBaits® Mito kit (Daicel Arbor Biosciences) for the house mouse *Mus musculus* panel, following the myBaits protocol v4 to capture ultraconserved elements (UCE) and mitogenomes, respectively. We amplified post-enrichment UCE and mitogenomes libraries with 14–18 cycles of PCR using Kapa HiFi HotStart Ready Mix (Roche Sequencing), following the manufacturer's protocol. A 1.6x SPRI magnetic bead clean-up was performed subsequently. We quantified and visualized the enriched libraries pool using a Qubit 4 fluorometer and a TapeStation 4200 System, respectively (conditions as mentioned above). Finally, we pooled captured libraries equimolarly and sequenced on a NovaSeq 6000 SP PE 2 × 150 bp (Illumina, Inc., San Diego, CA, USA) at the Oklahoma Medical Research Foundation, Oklahoma City (combined with samples from unrelated projects). We used two lanes of NovaSeq, one for historical samples and another for modern samples, to avoid biased sequencing.

We also reanalyzed UCE and mitogenomes published by Castañeda-Rico et al. (2020), and mitogenomes from Bi (2017) and Sullivan et al. (2017). We should note that we detected a misidentification labeling error in a museum specimen that was previously designated as *P. eremicus* in Castañeda-Rico et al. (2020) (GenBank accession number MT078819). It has now been correctly identified as *P. pectoralis* based on a BLAST analysis of the *cytochrome b* gene in GenBank[2], and corroborated with the voucher specimen deposited at the Museo de Zoología, Facultad de Ciencias, Universidad Nacional Autónoma de México. A list of all samples used in this study is found in **Supplementary Table S1**.

Data processing and phylogenetic analyses of ultraconserved elements

We processed raw data, provided by the sequencing core, following the PHYLUCE v1.6.7 pipeline (Faircloth, 2016[3]). We used Illumiprocessor 2.10 (Faircloth, 2013) and Trim Galore 0.6.5[4] to trim adapters, barcode regions and low-quality bases. The PHYLUCE script *phyluce_assembly_get_fastq_lengths.py* was used to check average fragment size after trimming. Reads were assembled into contigs using Trinity 2.8.5 (Grabherr et al., 2011), and identified contigs matching UCE loci in the 5K UCE locus set[5]. We generated two "taxon sets": (1) containing all of our samples to query the database obtained during UCE contig identification and created a list of UCE loci by sample, and (2) without the holotype specimen to test if including a sample which recovered fewer loci and higher amounts of missing data could affect phylogenetic relationships. We produced a monolithic FASTA file to extract sequences from each sample. We aligned FASTA sequences using MAFFT 7.4 (Katoh and Standley, 2013; Nakamura et al., 2018) and performed edge trimming. We also tested the internal trimming using Gblocks 0.91b (Castresana, 2000; Talavera and Castresana, 2007), but we found that this approach increased branch lengths on samples with a high percentage of missing data. However, the phylogenetic relationships remained the same with both trimming methods (data not shown for the internal trimming). We filtered the resulting alignments to test them for various degrees of missing data (matrix completeness): 65% matrix for which 65% of the taxa were present for each UCE locus, 75% matrix (25% of taxa missing), 85% matrix (15% of taxa missing), and 95% matrix (5% of taxa missing), where the number of missing taxa is directly proportional to the number of UCE loci and missing data on the final matrices. We quantified informative sites with the PHYLUCE script *phyluce_align_get_informative_sites.py*. All of these analyses were performed on the Smithsonian Institution High Performance Computing Cluster (Smithsonian Institution[6]). The final UCE dataset included data generated in this study and in Castañeda-Rico et al. (2020).

We performed two independent phylogenetic analyses using: (1) a concatenated dataset including all of our samples ($N = 18$), and (2) a concatenated dataset without the holotype specimen of *P. mekisturus* ($N = 17$) due to high amounts of missing data. We tested the aforementioned levels of matrix completeness (65, 75, 85, and 95%) for both datasets.

First, we conducted a Maximum Likelihood (ML) analysis, for both datasets and all levels of matrix completeness, using RAxML 8.12 (Stamatakis, 2014) with a GTRGAMMA site rate substitution model and 20 ML searches for the phylogenetic tree that best fit each set of data. We generated non-parametric bootstrap replicates using the -N autoMRE option which runs until convergence is reached. We reconciled the best fitting ML tree with the bootstrap replicate to obtain the final phylogenetic tree with support values using the -f b command.

[2] https://blast.ncbi.nlm.nih.gov/Blast.cgi

[3] https://github.com/faircloth-lab/phyluce

[4] https://github.com/FelixKrueger/TrimGalore

[5] https://github.com/faircloth-lab/uce-probe-sets

[6] https://doi.org/10.25572/SIHPC

We performed a Bayesian Inference (BI) analysis, with all levels of matrix completeness and both datasets, using MrBayes 3.2.6 (Huelsenbeck and Ronquist, 2001; Ronquist and Huelsenbeck, 2003). First, we estimated the best evolutionary model of nucleotide substitution in jModelTest 2.1.1 (Guindon and Gascuel, 2003; Darriba et al., 2012) using the Akaike Information Criterion (AIC). The GTR + G model was selected as the best fitting model for both datasets ($N = 18$ and $N = 17$) with the following parameters: base frequencies A = 0.3041, C = 0.1960, G = 0.2039, T = 0.2960; nst = 6; and gamma shape = 0.1220; and base frequencies A = 0.2995, C = 0.2006, G = 0.2012, T = 0.2988; nst = 6; and gamma shape = 0.1270, respectively. The BI analyses were run using two independent runs with 50 million generations for the 95% matrix and 20 million generations for the 65, 75, and 85% matrices due to the high number of loci, sampling trees and parameters every 1,000 generations with four Markov-chains Monte Carlo (MCMC), three heated and one cold. Heating temperature was set at 0.2 to facilitate greater movement between the four MCMC chains. We visualized output parameters using Tracer v1.7.1 (Rambaut et al., 2018) to check for convergence between runs and we discarded the first 25% of the trees as burn-in.

Maximum Likelihood and Bayesian Inference analyses were performed without partitions (as mentioned above) and with partitions only on the 95% matrix of both datasets to test if there was any difference due to partitioning and to account for heterogeneity in rates and patterns of molecular evolution within each UCE loci. First, the Sliding-Window Site Characteristics (SWSC) partitioning method based on sites entropies (Tagliacollo and Lanfear, 2018) was used to generate partitions that account for within-UCE heterogeneity. We followed the code implemented by Tagliacollo and Lanfear (2018) in the SWSC-EN method[7]. Then, we used PartitionFinder 2.1.1 (Lanfear et al., 2016) to optimize the partition scheme, by joining together similar subsets, obtained with the SWSC-EN method. After the final partition scheme was obtained, we performed the ML and BI analyses as mentioned above.

Finally, we used the dataset without the holotype of *P. mekisturus* –high number of missing data– ($N = 17$) with all levels of matrix completeness, to conduct a species tree analysis under the multispecies coalescent (MSC) model with ASTRAL-III v.5.7.8 (Zhang et al., 2018). We used the uce2speciestree pipeline script (Campana, 2019[8]) to generate input files for ASTRAL. This script uses RAxML to infer individual gene trees under the GTRGAMMA substitution model, and 100 bootstrap replicates. The local posterior probability – LPP – (Sayyari and Mirarab, 2016) was used as branching support, where an LPP ≥ 0.95 is considered as strong support (Erixon et al., 2003).

Data processing and phylogenetic analyses of mitogenomes

We analyzed read quality of the FASTQ format files using FastQC v0.11.5 (Andrews, 2010[9]). We removed adapter sequences and low-quality reads using the default parameters (Phred:20, mean min-len:20) in Trim Galore 0.6.5 (see text footnote 4). We removed exact duplicates (-derep1,4) using Prinseq-lite v0.20.4 (Schmieder and Edwards, 2011). We mapped the resulting high quality reads to a reference genome according to a species-specific reference (see GenBank accession numbers for each reference genome in **Supplementary Table S1**), using the Geneious algorithm in Geneious Prime® 2021.2.2[10] with default parameters (Medium-Low sensitivity, Maximum mismatches = 20%, Maximum gaps = 10%). We generated consensus sequences with Geneious Prime® 2021.2.2 (see footnote 11), using 5X as the lowest coverage to call a base, a Highest Quality control, and the remaining default parameters, and aligned them using MAFFT 7.45 plug-in (Katoh and Standley, 2013). We transferred annotations from each species-specific reference (**Supplementary Table S1**) to rule out the presence of nuclear copies of mitochondrial genes (NUMTs), and translated all protein-coding genes to check for frame shifts or stop codons.

We aligned sequences with MAFFT 7.45 plug-in (Katoh and Standley, 2013) in Geneious Prime® 2021.2.2 (see footnote 11). We used samples generated in this study and data previously published by Bi (2017), Sullivan et al. (2017), and Castañeda-Rico et al. (2020) (**Supplementary Table S1**). For most of the species with a mitogenome previously published we generated a new mitogenome sequence from a different sample but same species (except for *Peromyscus melanophrys, P. perfulvus, P. mexicanus,* and *Habromys ixtlani*). We used the mitogenome alignment to infer the phylogenetic relationships of *P. mekisturus* in relation to other neotomine rodents. We performed a ML analysis using the concatenated dataset (without partitions) in RAxML 8.12 (Stamatakis, 2014) with a GTRGAMMA site rate substitution model. Clade support was assessed by bootstrapping with the -N autoMRE option for a bootstrap convergence criterion. We used the -f b option to reconcile the best fitting ML tree with the bootstrap replicate to obtain the final phylogenetic tree (as mentioned above).

We conducted a BI analysis, on a partitioned dataset, using MrBayes 3.2.6 (Huelsenbeck and Ronquist, 2001; Ronquist and Huelsenbeck, 2003). The best model and partition scheme were estimated using PartitionFinder 2.1.1 (Lanfear et al., 2016). Our search was limited to the models available in MrBayes, with linked, corrected Akaike Information Criterion (AICc) and greedy parameters. The data block was defined by codon

[7] https://github.com/Tagliacollo/PFinderUCE-SWSC-EN

[8] https://github.com/campanam/uce2speciestree

[9] www.bioinformatics.babraham.ac.uk/projects/fastqc

[10] https://www.geneious.com

position, tRNA, rRNA and D-loop selection, and the result was incorporated in the phylogenetic reconstruction. We used two independent runs with 50 million generations, sampling trees and parameters every 1,000 generations with four MCMC, and a heating temperature set at 0.2, as mentioned above, to perform the BI analysis. We checked convergence between runs using Tracer v1.7.1 (Rambaut et al., 2018), and we discarded the first 25% of the trees as burn-in.

DNA damage patterns were evaluated for the historical samples with mapDamage2.0 (Joìnsson et al., 2013). We analyzed the reads obtained from the mitogenome enrichment and mapped to the reference genome. We used the –rescale, -y 0.1, –plot-only commands.

Divergence time estimates

We estimated molecular dates of divergence using Bayesian MCMC searches implemented in BEAST2 v2.6.6 (Bouckaert et al., 2019) using the concatenated 95% matrix of the UCE data (N = 17) without partitions. The holotype specimen of *P. mekisturus* was not included in the matrix due to a high number of missing data. The analysis was performed under an uncorrelated lognormal relaxed molecular clock model. The calibrated Yule speciation processes model (Heled and Drummond, 2012) with a randomly generated starting tree were set up as priors. We used three calibration points with a lognormal distribution. Calibrations were based on fossil records (million years ago [mya]) of (1) *Reithrodontomys* (mean = 1.8, stdev = 1.076, offset = 1.63), as used by Steppan and Schenk (2017); (2) *Onychomys* (mean = 4.9, stdev = 1.169, offset = 4.753), as used by Steppan and Schenk (2017); and (3) the most recent common ancestor of *P. attwateri* (mean = 2.7, stdev = 0.9, offset = 2.4 [Dalquest, 1962; Karow et al., 1996; Wright et al., 2020]) (**Supplementary Tables S2, S3**). Two separated runs of 50 million iterations each were sampled every 1,000 iterations. We checked convergence statistics for Effective Sample Sizes (ESS) using Tracer v1.7.1 (Rambaut et al., 2018) and a 25% of burn-in was performed on each run. We used LogCombiner v2.6.6 to combine trees and TreeAnnotator v2.6.2 to get a consensus tree with node height distribution (both packages available in BEAST).

We also estimated the divergence times on the complete mitogenomes dataset. First, we obtained the best model and partition scheme in PartitionFinder 2.1.1 (Lanfear et al., 2016). Our search was limited to the models available in BEAST, with linked, AICc, and greedy parameters. The data block was defined by codon position, tRNA, rRNA and D-loop selection, and the result was incorporated in the dating analysis. The analysis was performed under the same conditions and priors set up for the UCE data (mentioned above). We used the same three calibrations points set up for the UCE analysis. Two separated runs of 50 million iterations each were sampled every 1,000 iterations, with a burn-in of 25% on each run. We evaluated convergence with Tracer v1.7.1 (Rambaut et al., 2018), and LogCombiner v2.6.6 was used to combine trees. Finally, we obtained a consensus tree with node height distribution in TreeAnnotator v2.6.2.

We visualized all phylogenetic and dated trees from the UCE and mitogenomes datasets in FigTree 1.4.4[11]. Phylogenetic, dating and DNA damage analyses were performed on the Smithsonian Institution High Performance Computing Cluster (Smithsonian Institution, see footnote 6).

Results

We successfully sequenced UCE's (raw data is available in GenBank under BioProject PRJNA838631), and mitogenomes (GenBank accession numbers ON528108 – ON528119), from all samples processed, three historical and nine modern samples. The average number of paired-end reads were 11,632,614 (ranging from 8,998,310 to 13,693,826) and 12,272,671 (ranging from 9,480,276 to 21,093,430) for historical and modern samples, respectively. The average fragment size after trimming ranged from 59 to 123 bp and from 134 to 144 bp for historical and modern samples, respectively.

Multilocus nuclear phylogenies

Trinity assemblies yielded an average of 24,543 contigs per sample (min = 2,056; max = 87,428) for historical samples and 197,503 contigs (min = 43,081; max = 450,450) for modern samples. We recovered 4,406 UCE loci in the incomplete matrix (N = 18; average = 2,537 min = 306 max = 3,575 for historical samples, and average = 3,305 min = 1,375 max = 3,859 for modern samples). We obtained 303 UCE loci for the holotype specimen of *P. mekisturus*.

We tested topologies with different levels of missing data for: (a) complete dataset (N = 18), and (b) dataset without the holotype specimen (N = 17). For the complete dataset the 65% matrix contained 3,659 UCE loci (NL) with an average of 13.9 informative sites per locus (AIS), the 75% matrix (NL = 2,899, AIS = 14.4), the 85% matrix (NL = 1,334, AIS = 14.4), and the 95% matrix (NL = 85, AIS = 14.1). For the dataset without the holotype specimen the 65% matrix contained 3,649 UCE loci with an average of 13.9 informative sites per locus, the 75% matrix (NL = 3,361, AIS = 14.1), the 85% matrix (NL = 2,155, AIS = 14.3), and the 95% matrix (NL = 417, AIS = 14.9).

Maximum Likelihood and Bayesian Inference analyses for both datasets (N = 18 and N = 17) with all levels of matrix completeness yielded the same topology with high support

[11] http://tree.bio.ed.ac.uk/software/figtree/

values for all branches (**Figure 1**, phylogenetic trees obtained from the 65, 75, and 85% matrices are shown in **Supplementary Figures S2, S3**). The different levels of missing data reflected with the percentage matrices showed, at least for these datasets, that the inclusion of more or less samples per locus did not affect the phylogenetic inferences nor the support values. Both phylogenetic trees (**Figure 1**) placed *Peromyscus mekisturus* as the sister species of *Reithrodontomys mexicanus* with high bootstrap support values (bootstrap > 92, pp = 1), and it is more closely related to *Isthmomys pirrensis* than to any other member of the genus *Peromyscus* (bootstrap > 98, pp = 1). We also confirmed that the holotype specimen, despite its amount of missing data, was confidently placed within the clade that include the only other known specimen of this species. The ML and BI trees, using the 95% matrix, with and without partitions (**Figure 1**), supported the same topology with high bootstrap and posterior probability values (bootstrap > 92, pp = 1) for all branches.

The species tree analysis, with all levels of matrix completeness and the dataset without the holotype specimen, estimated the same topology from all matrices (**Figure 2**) with high support values (local posterior probability – LPP > 0.95). The species tree was concordant with the ML and BI analyses, supporting the placement of *P. mekisturus* outside the genus *Peromyscus*, and as the sister species of *R. mexicanus* (LPP = 1). It also supported the relationship of *P. mekisturus* + *R. mexicanus* as the sister group of *I. pirrensis* (LPP > 0.95). The only difference between the concatenated analyses (ML and BI), and the species tree analysis was the phylogenetic relationship between *P. mexicanus* and *P. megalops*. The first analysis placed *P. mexicanus* and *P. megalops* as sister species, while the second analysis placed *P. mexicanus* as sister of *P. melanophrys* + *P. perfulvus*, and *P. megalops* as sister of *P. mexicanus* + (*P. melanophrys* + *P. perfulvus*).

Mitochondrial phylogenies

We recovered near-complete mitogenome sequences for all samples, including the holotype specimen of *P. mekisturus* (>95% of the reference mitogenome covered).

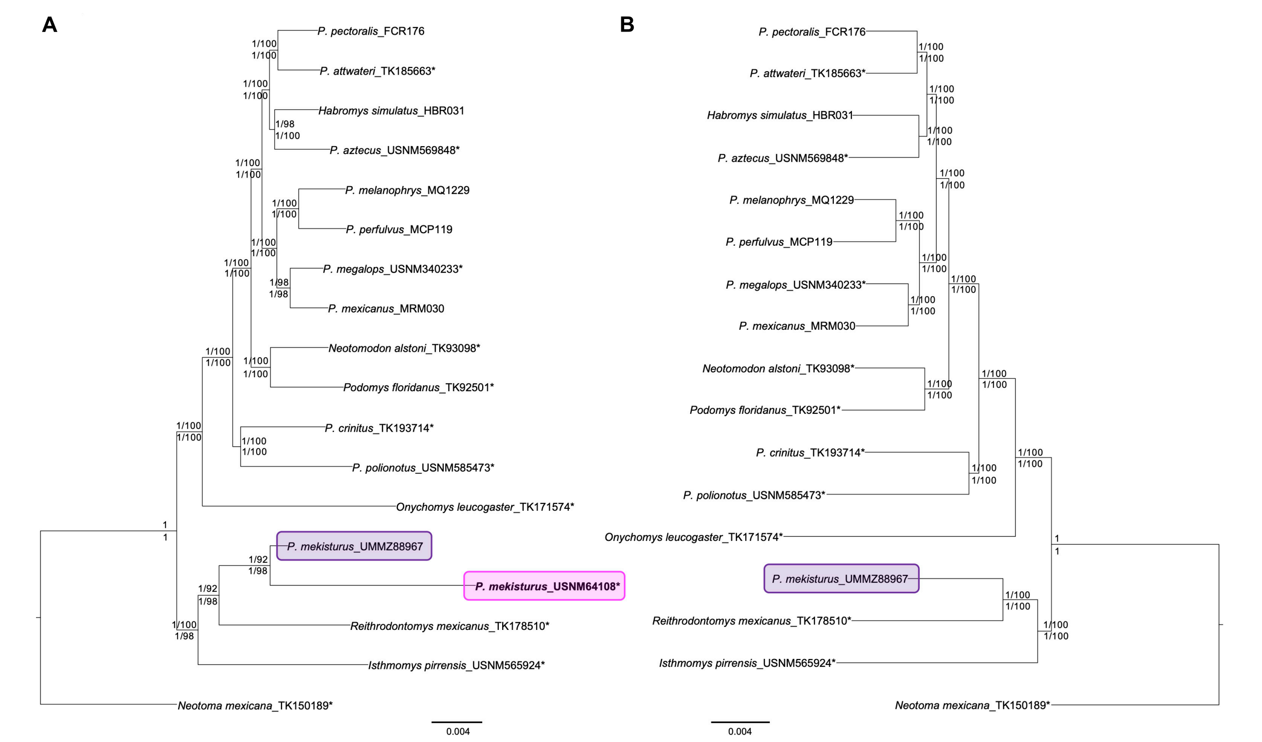


FIGURE 1

Ultraconserved elements (UCE) phylogenetic trees constructed using Bayesian Inference and Maximum Likelihood with and without partitions. Trees from all analyses yielded identical topologies. Nodal support is denoted with posterior probability/bootstrap values (numbers above the branches indicate results without partitions, those below with partitions). **(A)** Phylogenetic tree using a complete dataset (*N* = 18) based on 85 UCE loci (95% matrix) showing the phylogenetic position of the two *Peromyscus mekisturus* specimens. The pink block highlights the phylogenetic position of the *P. mekisturus* holotype collected in 1898, and the purple block shows the position of the *P. mekisturus* specimen collected in 1947; **(B)** phylogenetic tree based on 417 UCE loci (95% matrix, *N* = 17). Note that the removal of the holotype due to missing data (306 loci) does not change the tree topology but increases the nodal support between *P. mekisturus* and *R. mexicanus*. Asterisks* denote specimens that were sequenced from museum specimens for this study.

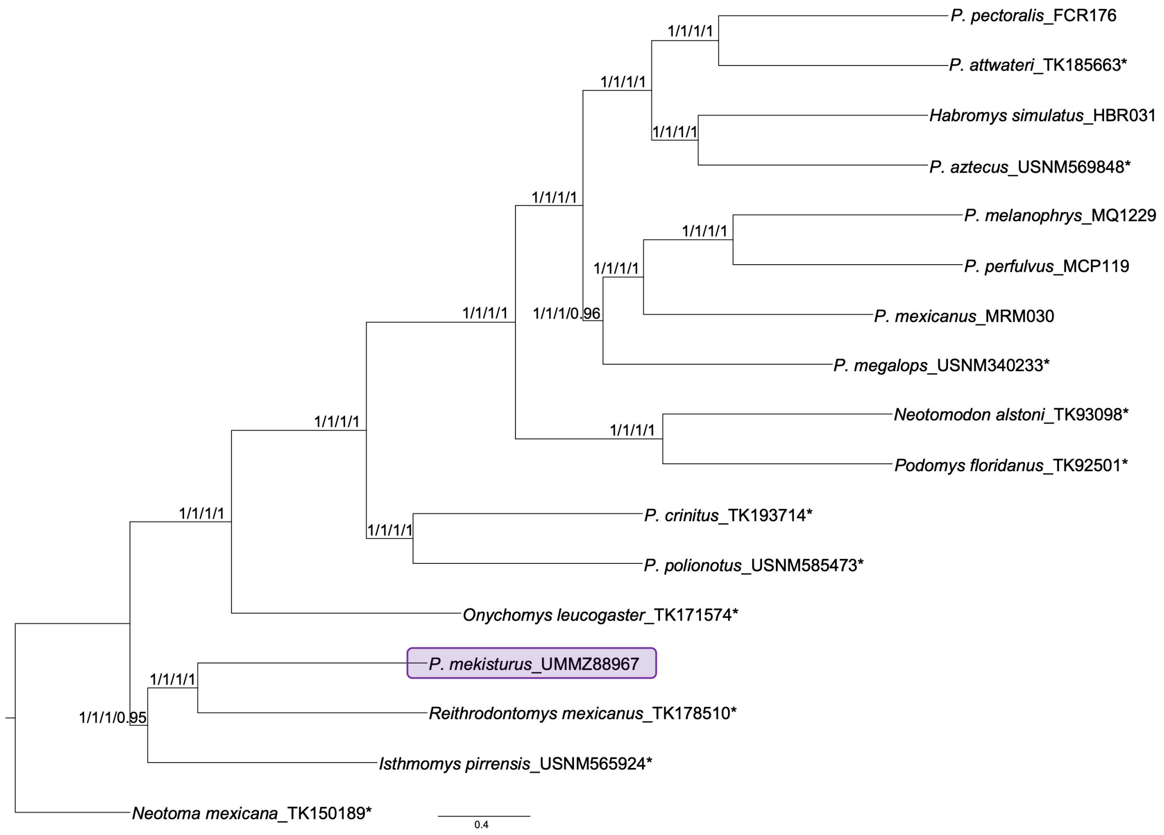

P. pectoralis_FCR176
P. attwateri_TK185663*
Habromys simulatus_HBR031
P. aztecus_USNM569848*
P. melanophrys_MQ1229
P. perfulvus_MCP119
P. mexicanus_MRM030
P. megalops_USNM340233*
Neotomodon alstoni_TK93098*
Podomys floridanus_TK92501*
P. crinitus_TK193714*
P. polionotus_USNM585473*
Onychomys leucogaster_TK171574*
P. mekisturus_UMMZ88967
Reithrodontomys mexicanus_TK178510*
Isthmomys pirrensis_USNM565924*
Neotoma mexicana_TK150189*
1/1/1/1
1/1/1/0.96
1/1/1/0.95
0.4

FIGURE 2
ASTRAL species tree estimation based on different levels of matrix completeness (65% –3,649 UCE loci–, 75% –3,361 UCE loci–, 85% –2,155 UCE loci–, and 95% –417 UCE loci–) and the *N* = 17 dataset. Nodal support is provided with local posterior probability in the same order as the matrices were mentioned. Note that the phylogenetic position of the *P. mekisturus* specimen collected in 1947 (purple block) also shows strong support for its close relationship to *R. mexicanus*. Asterisks* denote specimens that were sequenced from museum specimens for this study.

The mitochondrial sequences contain the standard features present in a mammalian genome as similar size, structure and gene arrangement. The final alignment was 16,228 bp length and included 30 individuals. The BI (with six partitions) and ML analyses (**Figure 3**) supported the placement of *P. mekisturus* outside the genus *Peromyscus*, and more closely related to the genus *Reithrodontomys* and *Isthmomys*. The closer phylogenetic relationship of *P. mekisturus* to *R. mexicanus*, as sister species, is strongly supported (pp = 1, bootstrap = 98). We also confirmed, with high support values (pp = 1, bootstrap = 100) that the two samples of *P. mekisturus* were closely related and placed in the same clade. The only difference between the BI and ML trees is the phylogenetic relationship of *Neotomodon alstoni* + *Podomys floridanus* with other peromyscine rodents. The BI analysis placed this clade as sister of *P. attwateri, P. pectoralis, P. aztecus, P. megalops, P. mexicanus, P. melanophrys, P. perfulvus,* and *H. ixtlani*, while the ML analysis place it as sister to *P. attwateri, P. pectoralis, P. aztecus,* and *H. ixtlani.*

In addition, all of the species which included both a mitogenome generated in this study and one obtained from GenBank were very similar and clustered together in our phylogenetic analysis. This allowed us to corroborate the taxonomic identity of the samples by using voucher specimens deposited in scientific collections. Finally, the results of mapDamage2.0 analysis showed a weak signal of DNA damage typical of historical DNA (**Supplementary Figure S4**). The weak damage signal is expected since the oldest sample was collected in 1898 and was well-preserved.

Divergence time estimates of *Peromyscus mekisturus* and its close relatives

For the UCE dataset, the analysis estimating the time to the most recent ancestor (TMRA) recovered that the divergence between *I. pirrensis* + *R. mexicanus* + *P. mekisturus* from the genus *Peromyscus* + *O. leucogaster* occurred ca. 8.60 mya (95% Highest Posterior Density [HPD]: 6.00 – 11.51 mya). While the split of *P. mekisturus* + *R. mexicanus* versus *I. pirrensis* is dated ca. 6.74 mya (95% HPD: 5.51 – 8.19 mya). Finally, the divergence

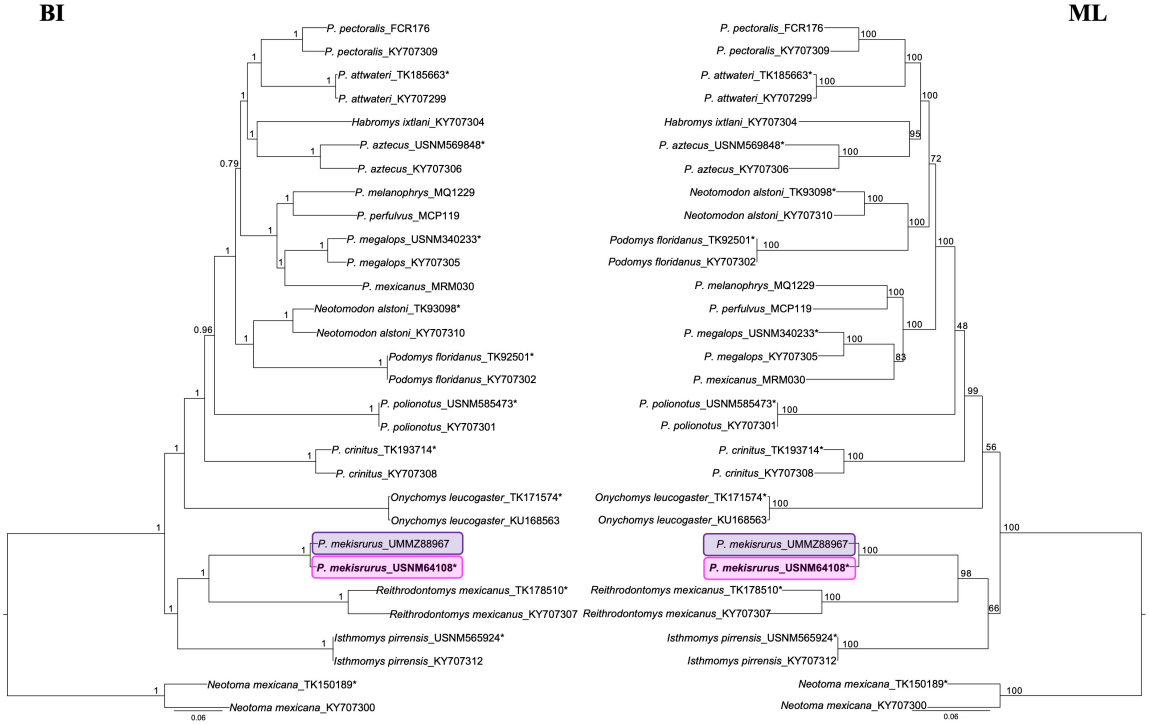

BI
P. pectoralis_FCR176
P. pectoralis_KY707309
P. attwateri_TK185663*
P. attwateri_KY707299
Habromys ixtlani_KY707304
P. aztecus_USNM569848*
P. aztecus_KY707306
P. melanophrys_MQ1229
P. perfulvus_MCP119
P. megalops_USNM340233*
P. megalops_KY707305
P. mexicanus_MRM030
Neotomodon alstoni_TK93098*
Neotomodon alstoni_KY707310
Podomys floridanus_TK92501*
Podomys floridanus_KY707302
P. polionotus_USNM585473*
P. polionotus_KY707301
P. crinitus_TK193714*
P. crinitus_KY707308
Onychomys leucogaster_TK171574*
Onychomys leucogaster_KU168563
P. mekisrurus_UMMZ88967
P. mekisrurus_USNM64108*
Reithrodontomys mexicanus_TK178510*
Reithrodontomys mexicanus_KY707307
Isthmomys pirrensis_USNM565924*
Isthmomys pirrensis_KY707312
Neotoma mexicana_TK150189*
Neotoma mexicana_KY707300
1
0.79
0.96
0.06
ML
P. pectoralis_FCR176
P. pectoralis_KY707309
P. attwateri_TK185663*
P. attwateri_KY707299
Habromys ixtlani_KY707304
P. aztecus_USNM569848*
P. aztecus_KY707306
Neotomodon alstoni_TK93098*
Neotomodon alstoni_KY707310
Podomys floridanus_TK92501*
Podomys floridanus_KY707302
P. melanophrys_MQ1229
P. perfulvus_MCP119
P. megalops_USNM340233*
P. megalops_KY707305
P. mexicanus_MRM030
P. polionotus_USNM585473*
P. polionotus_KY707301
P. crinitus_TK193714*
P. crinitus_KY707308
Onychomys leucogaster_TK171574*
Onychomys leucogaster_KU168563
P. mekisrurus_UMMZ88967
P. mekisrurus_USNM64108*
Reithrodontomys mexicanus_TK178510*
Reithrodontomys mexicanus_KY707307
Isthmomys pirrensis_USNM565924*
Isthmomys pirrensis_KY707312
Neotoma mexicana_TK150189*
Neotoma mexicana_KY707300
100
95
72
83
48
99
56
98
66
0.06

FIGURE 3
Mitogenome phylogenies based on Bayesian Inference (BI) and Maximum Likelihood (ML). Nodal support is provided with posterior probability and bootstrap values, respectively. The pink block highlights the phylogenetic position of the *Peromyscus mekisturus* holotype collected in 1898, and the purple block shows the position of the second specimen collected in 1947. Asterisks* denote specimens that were sequenced from museum specimens for this study and compared with previous GenBank accessioned mitogenome sequences.

between *P. mekisturus* and *R. mexicanus* occurred ca. 3.80 mya (95% HPD: 1.67 – 6.27 mya) (**Figure 4**).

For the mitogenome dataset with six partitions, we estimated the split between *I. pirrensis* + *R. mexicanus* + *P. mekisturus* versus the genus *Peromyscus* + *O. leucogaster* dated ca. 5.85 mya (95% HPD: 5.19 – 6.64 mya). While the divergence between *P. mekisturus* + *R. mexicanus* versus *I. pirrensis* occurred ca. 5.38 mya (95% HPD: 4.65 – 6.23 mya), followed by the split between *P. mekisturus* and *R. mexicanus* dated ca. 4.42 mya (95% HPD: 3.61 – 5.30 mya). Finally, we dated the diversification within *P. mekisturus* ca. 0.26 mya (95% HPD: 0.15 – 0.37 mya) (**Figure 5**).

Discussion

Phylogenetic relationships of *Peromyscus mekisturus* and its relatives

All of our ML, BI, and species tree analyses, with both mitochondrial and nuclear datasets, strongly supported that the Puebla deer mouse, *P. mekisturus*, is the sister species of the genus *Reithrodontomys*, and it is more closely related to the genus *Isthmomys* than to any other member of the genus *Peromyscus*. Therefore, our nuclear data results support previous mitochondrial hypothesis proposed by Castañeda-Rico et al. (2020). In addition, the successful sequencing of UCE loci and mitogenome from the holotype of *P. mekisturus* allowed us to confirm the identification of the only two known specimens of this species.

To better understand the phylogenetic position of *P. mekisturus*, it is important to outline some of the previous taxonomic problems that have emerged for the genus *Peromyscus* and its relationship with the genera *Isthmomys* and *Reithrodontomys*. *Peromyscus* is a very large and diverse group in which new species are still being described (Bradley et al., 2007, 2014) making taxonomic sampling challenging for this genus. It has also been demonstrated that this genus has a high and rapid diversification rate complicating the reconstruction of its phylogenetic relationships (Platt et al., 2015). Thus, *Peromyscus* has presented a great challenge to systematists and after over 100 years – since Osgood's (1909) monograph – its evolutionary boundaries remain unresolved (Carleton, 1980, 1989; Bradley et al., 2007; Miller and Engstrom, 2008; Platt et al., 2015). An additional conflict is the taxonomic status of *Habromys,*

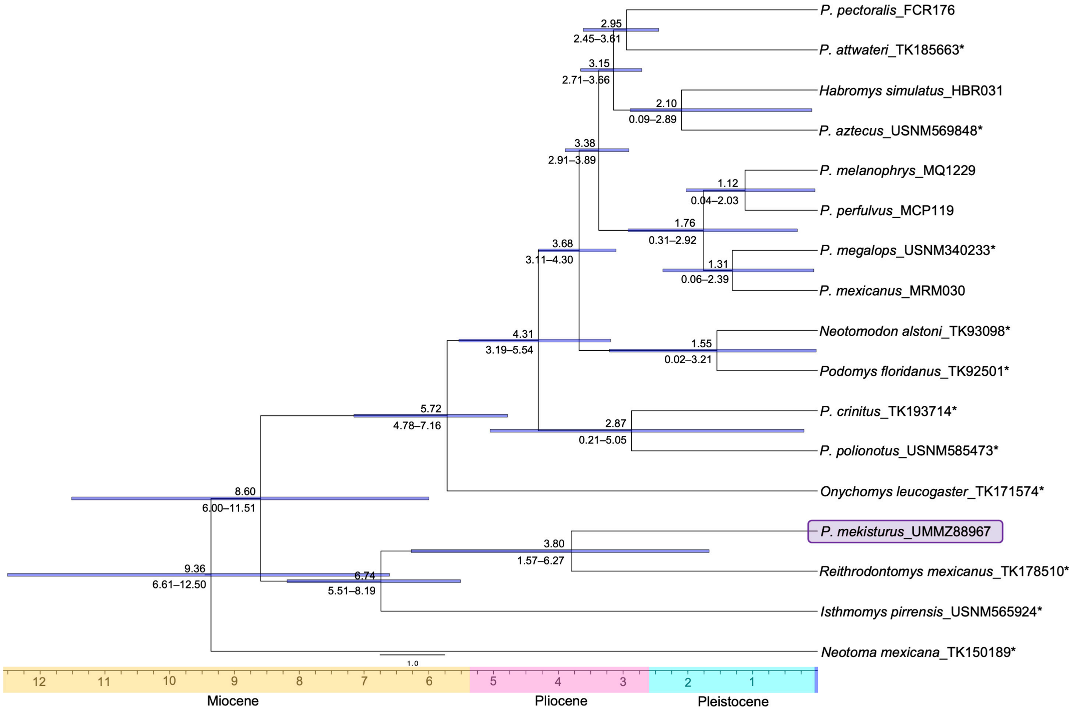


FIGURE 4
Divergence dated nuclear phylogeny based on 417 UCE loci (95% matrix, N = 17). Dates above the branches are provided in millions of years. Blue horizontal bars and numbers below the branches show the 95% confidence intervals. The purple block shows the phylogenetic position of the *Peromyscus mekisturus* specimen collected in 1947. Asterisks* denote specimens that were sequenced from museum specimens for this study.

Megadontomys, Neotomodon, Osgoodomys, Podomys, and *Isthmomys*, recognized at the generic (*Peromyscus – sensu lato –*) or subgeneric (*Peromyscus – sensu stricto –*) level. To date, no single classification fits perfectly into one category, and not a single study has been able to offer unambiguous taxonomic recommendations for *Peromyscus* and its close relatives (Platt et al., 2015).

Isthmomys was first suggested as a subgenus of *Peromyscus* (Hooper and Musser, 1964) but later it was elevated to a separate genus (Bradley et al., 2007; Miller and Engstrom, 2008; Platt et al., 2015; Sullivan et al., 2017). In addition, several studies have placed *Isthmomys* as the sister taxon of *Reithrodontomys,* and these two genera are the nearest taxa to *Onychomys* + *Peromyscus* (Bradley et al., 2007; Miller and Engstrom, 2008; Platt et al., 2015; Sullivan et al., 2017; Castañeda-Rico et al., 2020). In our study, we found that the divergence between *Peromyscus* versus *Isthmomys* + (*Reithrodontomys* + *P. mekisturus*) was strongly supported by the UCE's BI tree and the mitogenomes analyses (**Figures 1**, **3**), even though the ML and species tree analyses did not yield strong node support for this node (**Figures 1**, **2**). Furthermore, we confirmed that *Isthmomys* is the sister genus of *Reithrodontomys* and *P. mekisturus* with high support values for all analyses and datasets (**Figures 1–3**).

To date, no phylogenetic hypothesis has ever suggested that the genus *Reithrodontomys* should be nested within *Peromyscus* (Sullivan et al., 2017). Additionally, no morphological similarities have been found between *Reithrodontomys* and its close relative *Isthmomys* (Miller and Engstrom, 2008). Harvest mice belonging to the genus *Reithrodontomys*, are small-bodied rodents with long tails and are distinguished from other peromyscine rodents by possessing grooved or sulcate upper incisors – a key synapomorphy defining this genus – (Le Conte, 1853; Musser and Carleton, 1993; Arellano et al., 2005). *Peromyscus mekisturus* is also a small-bodied rodent with a very long tail, equaling three-fourths of the total length that is associated with its arboreal habits. This same character was used to place it as sister of *P. melanophrys* within the *Peromyscus melanophrys* group (Osgood, 1909; Carleton, 1989). In addition, *P. mekisturus* does not have grooved or sulcate incisors, but only a greater development of an incisor capsule on the dentary compared with other peromyscine rodents (Carleton, 1989). To date, no morphological character has shown similarities or has suggested a close phylogenetic relationship between *P. mekisturus* and *Reithrodontomys.*

A recent study by Castañeda-Rico et al. (2020), suggested the placement of *P. mekisturus* and *Isthmomys* at the same taxonomic level, i.e., still considered part of *Peromyscus* (*sensu*

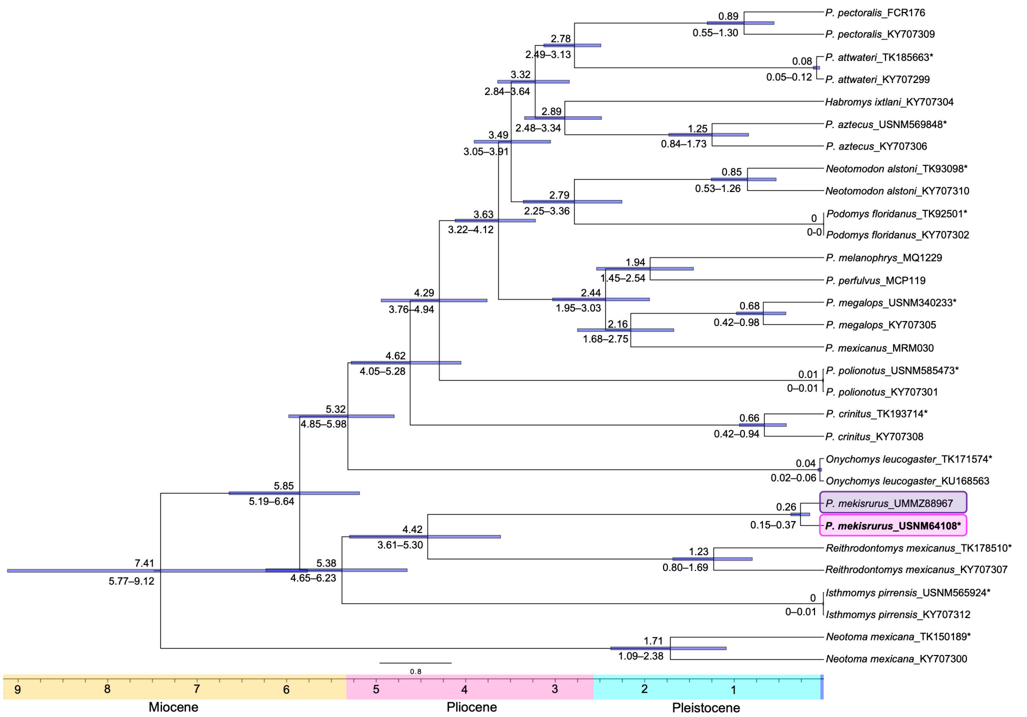


FIGURE 5

Divergence dated whole mitochondrial genome phylogeny. Dates above the branches are provided in millions of years. Blue horizontal bars and numbers below the branches show the 95% confidence intervals. The pink block highlights the phylogenetic position of the *Peromyscus mekisturus* holotype collected in 1898, and the purple block shows the position of the second specimen collected in 1947. Asterisks* denote specimens that were sequenced from museum specimens for this study and compared with previous GenBank accessioned mitogenome sequences.

lato or *sensu stricto*) but supporting the paraphyly of the genus as has been suggested (Bradley et al., 2007; Miller and Engstrom, 2008; Platt et al., 2015; Sullivan et al., 2017). However, based on the fact that *Isthmomys* is currently accepted as a separate genus (Sullivan et al., 2017) and coupled with our genomics results, we recommend that *P. mekisturus* be recognized at the generic level. In support of our recommendation is the phylogenetic position, and genetic uniqueness and distinctiveness of *P. mekisturus* alongside with its close relationship with *Reithrodontomys* but lacking the synapomorphy (i.e., grooved or sulcate upper incisors) that defines this genus. We also propose that a taxonomic revision of *P. mekisturus* should be undertaken to incorporate a morphological re-evaluation to formally recognize it as a new genus. If our results lead to a taxonomic re-evaluation and rearrangement of this group into a monotypic genus, this would have a great impact on their conservation management as it would likely represent the description of a nearly, or recently extinct unique lineage of rodents (Castañeda-Rico et al., 2020).

Even though the objectives of this study were not to further investigate the phylogenetic relationships within the genus *Peromyscus*, our sampling and novel results using mitogenomes and UCE loci of representative museum specimens within this genus allowed us to make some interesting inferences. First, all of our phylogenetic analyses (**Figures 1–3**) continue to support the paraphyly for the genus *Peromyscus*, including representatives of the genera *Habromys, Podomys,* and *Neotomodon*, as was previously suggested (Bradley et al., 2007; Miller and Engstrom, 2008; Platt et al., 2015; Sullivan et al., 2017; Castañeda-Rico et al., 2020). All of the analyses based on the UCE dataset showed a well-supported clade for *P. crinitus* and *P. polionotus* (**Figures 1, 2**). Sullivan et al. (2017) and Castañeda-Rico et al. (2020), both using mitogenomes, identified the same clade with high support values for the BI analysis but lacking support or low support values for the ML tree. In sharp contrast, our mitogenome trees did not support this clade, instead, *P. crinitus* was the most divergent species within *Peromyscus*, followed by the split of *P. polionotus* (**Figure 3**). This mito-nuclear discordance, commonly seen in mammals (Hawkins et al., 2016), requires further investigation. We suggest that future studies increase taxon sampling.

The relationship between *P. mexicanus* and *P. megalops* also recovered some discrepancies. The species tree analysis placed *P. mexicanus* as the sister of *P. melanophrys* + *P. perfulvus*, with *P. megalops* being the most closely related species to a clade containing all three, with high LPP support values (**Figure 2**). However, ML and BI nuclear UCE trees as well as all mitogenome trees showed a well-supported clade including

P. mexicanus and *P. megalops* being the sister to a clade containing *P. melanophrys* + *P. perfulvus* (**Figures 1**, **3**). The same phylogenetic relationships were also supported by Castañeda-Rico et al. (2020) but only the clade of *P. mexicanus* and *P. megalops* was identified by Sullivan et al. (2017) due to the inclusion of less taxa.

Neotomodon alstoni and *Podomys floridanus* constitute a well-supported clade across all the phylogenetic analyses, however, the placement of this clade is in conflict. All nuclear trees and the BI mitogenome tree with high support values (**Figures 1–3**) placed *N. alstoni* + *P. floridanus* as the sister clade of *P. megalops, P. mexicanus, P. perfulvus, P. melanophrys, P. aztecus, P. attwateri*, *P. pectoralis,* and *H. simulatus*/H. *ixtlani*. In contrast, the ML mitogenome tree and the time tree (**Figures 3**, **5**) placed *N. alstoni* + *P. floridanus* as the sister clade of *P. aztecus, P. attwateri*, *P. pectoralis,* and *H. ixtlani* but with a lower support value (bootstrap = 72). Sullivan et al. (2017) and Castañeda-Rico et al. (2020) also supported the same phylogenetic relationship with high support values but only using mitogenomes. Finally, our genome-wide analyses confirm the sister genera relationship of *Onychomys* and *Peromyscus* which had been previously suggested using single genes (Platt et al., 2015), and that the *Peromyscus melanophrys* group (*P. melanophrys* + *P. perfulvus*) is sister to *P. mexicanus* + *P. megalops* previously suggested using only mitogenomes (Castañeda-Rico et al., 2020).

In general, our nuclear and mitochondrial phylogenetic trees largely mirror the mitogenome trees of Sullivan et al. (2017) and Castañeda-Rico et al. (2020) save a few exceptions. However, here we present the first nuclear and mitogenome-wide phylogeny with the most complete taxon dataset of peromyscine rodents to date. Given our results, we consider that the next step to unraveling the phylogenic relationships within the genus *Peromyscus* and its close relatives is to increase taxon sampling. However, this study has demonstrated that using museum specimens to increase taxa using UCE and mitogenomes is suitable to address complex phylogenetic studies, particularly when some taxa are only known from museum specimens.

Divergence time estimation indicates a late Pliocene – early Pleistocene origin of *Peromyscus mekisturus*

Our divergence time estimates (based on separate UCE and mitogenome datasets) resulted in similar dates (**Figures 4**, **5**). Although mitochondrial divergence dates were slightly older than those obtained with nuclear data, with the exception of the three oldest splits [*Neotoma*, *Isthmomys* + (*Reithrodontomys* + *P. mekisturus*), and *Isthmomys*]. Nuclear and mitochondrial estimates indicated that the main speciation events started in the late Miocene and Pliocene up to the Pleistocene, when the diversification started within each species. A majority of the divergences appear to correspond with the timing of the Quaternary climatic fluctuations, mostly during the Pleistocene glacial/interglacial cycles.

We dated three late Miocene – Pliocene events: the divergence between *Isthmomys* + *Reithrodontomys* + *P. mekisturus* versus *Onychomys* + *Peromyscus* ca. 8.6–5.85 mya, the split of *Isthmomys* from *Reithrodontomys* + *P. mekisturus* ca. 6.74–5.38 mya, and the divergence between *Onychomys* and *Peromyscus* ca. 5.72–5.32 mya. The order of these divergence events coincides with those proposed by Platt et al. (2015) using a combined dataset of one mitochondrial and three nuclear genes, however, their dates are slightly older but still place these events during the Miocene and Pliocene (i.e., ca. 7.93, 7.30, and 7.20 mya, respectively). The divergence between *P. mekisturus* and *Reithrodontomys* dated ca. 3.80 – 4.42 mya coincides with the beginning of diversification within *Peromyscus* ca. 4.31 – 4.62 mya, both events during the late Pliocene and early Pleistocene. Platt et al. (2015) estimated an older origin for *Peromyscus* that began at approximately 8 mya, but its diversification appears to have been focused at ca. 5.71 mya (95% HPD: 3.37 – 9.08). Our estimates are placed within the range reported by them and with smaller 95% HPD values (**Figures 4**, **5**). Finally, based on the mitogenome calibrated tree including the holotype specimen, we dated the diversification within *P. mekisturus* at ca. 0.26 mya at the end of the Pleistocene.

Similar divergence times have been found in other studies of *Peromyscus* (e.g., Castañeda-Rico et al., 2014; Cornejo-Latorre et al., 2017; Bradley et al., 2019) but they also analyzed single genes. Here, we present the first dated phylogeny obtained from genome-wide data for these groups of rodents. We expect that future genomic studies will continue to investigate and provide new insights into the divergence times in neotomines and other groups of rodents.

The complexity of elucidating the evolutionary history of *P. mekisturus*, with only two specimens known to science, can be decreased by making inferences about its closest relatives. For example, among peromyscines, the genus *Peromyscus* ranks first in species richness, followed by *Reithrodontomys* [ca. 70 and 24 species, respectively] (Miller and Engstrom, 2008; Platt et al., 2015; Martínez-Borrego et al., 2022). Both genera are found in most habitats distributed in North and Central America but only *Reithrodontomys* is found in South America. However, Mesoamerica, specifically Mexico, has been recognized as the center of biodiversity and diversification for both genera due to the unique physiographic characteristics that have promoted the isolation and differentiation of taxa in this region (Hooper, 1952; Hall, 1981; Eisenberg, 1989; Sullivan et al., 2000; Bradley et al., 2004; Arellano et al., 2005; Dawson, 2005; Miller and Engstrom, 2008). Speciation and diversification processes for these peromyscines have also been driven by the Pleistocene climatic cycles that expanded North American taxa southward during glacial advances, and retracted them northward during interglacial warming, giving rise to numerous vicariant and dispersal events (Dawson, 2005; Castañeda-Rico et al., 2014; Platt et al., 2015; Martínez-Borrego et al., 2022). Future phylogenetic studies should also include a denser taxon sampling of members of the genus *Reithrodontomys* and incorporate *P. mekisturus* as its closest outgroup to validate the timing and process of diversification of this group.

Information on environmental fluctuations and the existence of corridors at that time that favored movement across the landscape followed by post-glacial isolation strongly support the role of Pleistocene climate changes in the diversification process of many taxa (Martin and Klein, 1984; Ceballos et al., 2010; Ferrusquía-Villafranca et al., 2010). Therefore, we propose that both *P. mekisturus* and *Reithrodontomys* were also greatly impacted by the climatic fluctuation events that occurred during the Pleistocene, in agreement with our molecular dating. These taxa generated evolutionary novelties after repeated cycles of expansion and isolation that gave rise to unique lineages at the generic level. However, a surprisingly interesting revelation of our study regarding the phylogenetic placement and evolutionary history of *P. mekisturus* is that despite their close relatives (*Peromyscus* and *Reithrodontomys*) show high diversification rates, *P. mekisturus* did not and remained isolated in a restricted geographic area in central Mexico. We can only speculate that the distribution of this unique lineage was once more widespread with larger population sizes and that the subsequent biotic and/or abiotic conditions in the Anthropocene drastically decreased its population sizes putting it on a trajectory toward extinction.

The impact of museomics on present and future research

The case of *P. mekisturus* is particularly interesting as it demonstrates the positive impact of museomics, highlighting the importance of the inclusion of holotypes in phylogenetic studies, but it also provides evidence of the biodiversity loss that we are currently facing due to the ongoing mass extinction caused during the Anthropocene (Ceballos et al., 2020). We also demonstrate that it is possible to carefully design a protocol for destructive sampling that requires a very small amount of skin sample and that causes minimal damage to the voucher holotype specimen, ensuring all diagnostic characters remain intact. We also show that hDNA from museum specimens coupled with high throughput capture hybridization technologies are capable of yielding powerful genome-scale data. From a small piece of dry skin from the holotype specimen of *P. mekisturus*, we recovered a near-complete mitogenome sequence and 306 UCE loci that were enough to obtain well-resolved dated phylogenies. Therefore, we confirmed that the removal of tiny amounts of material from museum specimens by best practices of destructive sampling may add enormous value to the content of collections and will allow them, together with hDNA, to meet their full and incredible research potential (Bailey et al., 2016; Schmitt et al., 2018; Raxworthy and Smith, 2021). We expect that this example, confirming the ability of even very old specimens to yield genomic data, will motivate researchers to utilize type specimens and give confidence to curatorial staff who are tasked with ensuring the proper use of these valuable specimens.

Throughout this manuscript, we have continuously mentioned the value and importance of natural history museums and the specimens that are currently housed in their collections to conduct a wide range of cutting-edge research as well as continue with more traditional studies. However, we also need to highlight and advocate for the need to continue collecting specimens and to continue building the scientific heritage of the collections in the forthcoming years to keep a record of the historical biodiversity on the planet for the future generations of researchers and society in general. From a general perspective, Schmitt et al. (2018) argued that creative and novel uses of museum specimens have provided diverse applications of value to society, among them, is the research on biodiversity and global sustainability. Continued support of museums by funding agencies and dedication to collect specimens by museums are urgently needed to build and maintain this critical scientific resource moving forward and this topic should be a global priority. Yet collecting new specimens is still criticized and overlooked as an invaluable investment in the future (Minteer et al., 2014). However, this criticism is often due to misconceptions about the perceived negative impact of museum collecting on wildlife populations (Remsen, 1995; Hope et al., 2018). We expect that our case study of *P. mekisturus* can be used as justification for the need to continue to direct efforts and funding support for the collection and preservation of specimens across time, space, and taxonomic diversity with sufficient sample sizes, metadata, and breadth to ensure maximum impact across multiple disciplines (Brooks et al., 2011; Ward et al., 2015; Schmitt et al., 2018). There is still much work to be done in the field of museomics and the forthcoming years will surely offer astonishing results, new applications and uses, and even more improvements in methods and technologies.

Data availability statement

The raw data generated for this study can be found in the GenBank under BioProject: PRJNA838631 and under GenBank accession numbers given in **Supplementary Table S1**.

Ethics statement

Ethical review and approval were not required for the animal study because we exclusively used museum specimens deposited in scientific collections. All destructive sampling requests of the museum specimens used in this study were approved by the destructive sampling committee of those museums. We also used publicly available data on GenBank.

Author contributions

SC-R, CE, and JM secured funding and designed the study. SC-R and MH performed the specimen sampling. SC-R conducted the lab experiments, analyzed and archived the data, produced the figures, and wrote the manuscript with contributions from all co-authors. All authors read and approved the submitted version.

Funding

SC-R received a fellowship from Smithsonian-Mason School of Conservation and George Mason University. Research funding was provided by the Center for Conservation Genomics, Smithsonian Conservation Biology Institute, Smithsonian-Mason School of Conservation, and George Mason University. This article was funded in part by the George Mason University Libraries Open Access Publishing Fund.

Acknowledgments

We specially thank the specimen collectors, collection managers, curators and all museum staff of the Museum of Texas Tech University (TTU) and the Smithsonian Institution's National Museum of Natural History (NMNH) that granted the destructive sampling of museum specimens and provided tissue sample loans.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fevo.2022.930356/full#supplementary-material

References

Abreu-Jr, E. F., Pavan, S. E., Tsuchiya, M. T. N., Wilson, D. E., Percequillo, A. R., and Maldonado, J. E. (2020). Museomics of Neotropical tree squirrels: A dense taxon sampling of mitogenomes shakes the squirrel tree and suggests deep changes on their taxonomy. *BMC Evol. Biol.* 20:77 doi: 10.1186/s12862-020-01639-y

Andrews, S. (2010). *FastQC: A quality control tool for high throughput sequence data*. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc

Arellano, E., González-Cozátl, F. X., and Rogers, D. S. (2005). Molecular systematics of Middle American harvest mice *Reithrodontomys* (Muridae), estimated from mitochondrial cytochrome b gene sequences. *Mol. Phylogenet. Evol.* 37, 529–540. doi: 10.1016/j.ympev.2005.07.021

Bailey, S. E., Mao, X., Struebig, M., Tsagkogeorga, G., Csorba, G., Heaney, L. R., et al. (2016). The use of museum samples for large-scale sequence capture: A study of congeneric horseshoe bats (family Rhinolophidae). *Biol. J. Linn. Soc.* 117, 58–70. doi: 10.1111/bij.12620

Bi, G. (2017). The complete mitochondrial genome of northern grasshopper mouse (*Onychomys leucogaster*). *Mitochondrial DNA B Resour.* 2, 393–394. doi: 10.1080/23802359.2017.1347905

Bi, K., Linderoth, T., Singhal, S., Vanderpool, D., Patton, J. L., Nielsen, R., et al. (2019). Temporal genomic contrasts reveal rapid evolutionary responses in an alpine mammal during recent climate change. *PLoS Genet.* 15:e1008119. doi: 10.1371/journal.pgen.1008119

Bi, K., Linderoth, T., Vanderpool, D., Good, J. M., Nielsen, R., and Moritz, C. (2013). Unlocking the vault: Next-generation museum population genomics. *Mol. Ecol.* 22, 6018–6032. doi: 10.1111/mec.12516

Bouckaert, R., Vaughan, T. G., Barido-Sottani, J., Duchêne, S., Fourment, M., Gavryushkina, A., et al. (2019). BEAST 2.5: An advanced software platform for Bayesian evolutionary analysis. *PLoS Comput. Biol.* 15:e1006650. doi: 10.1371/journal.pcbi.1006650

Bradley, R. D., Bradley, L. C., Garner, H. J., and Baker, R. J. (2014). Assessing the value of natural history collections and addressing issues regarding long-term growth and care. *BioScience* 64, 1150–1158. doi: 10.1093/biosci/biu166

Bradley, R. D., Durish, N., Rogers, D., Millar, J., Engstrom, M., and Kilpatrick, W. (2007). Toward a molecular phylogeny for *Peromyscus*: Evidence from mitochondrial cytochrome-b sequences. *J. Mammal.* 88, 1146–1159. doi: 10.1644/06-mamm-a-342r.1

Bradley, R. D., Francis, J. Q., Platt, R. N. I. I., Soniat, T. J., Alvarez, D., and Lindsey, L. (2019). *Mitochondrial DNA sequence data indicate evidence for multiple species within Peromyscus maniculatus. Special publications, Museum of Texas*, Vol. 70. Lubbock, TX: Museum of Texas Tech University, 1–68.

Bradley, R. D., Mendez-Harclerode, F., Hamilton, M. J., and Ceballos, G. (2004). A new species of *Reithrodontomys* from Guerrero, Mexico. *Occas. Pap. Mus. Texas Tech Univ.* 231, 1–12.

Brooks, S. J., Fenberg, P. B., Glover, A. G., James, G. E., Johnson, K. G., Lister, K. G., et al. (2011). Natural history collections as sources of long-term datasets. *Trends Ecol. Evol.* 26, 153–154. doi: 10.1016/j.tree.2010.12.009

Buerki, S., and Baker, W. J. (2016). Collections-based research in the genomic era. *Biol. J. Linn. Soc.* 117, 5–10. doi: 10.1111/bij.12721

Campana, M. G. (2019). *uce2speciestree*. Available online at: https://github.com/campanam/uce2speciestree

Campana, M. G., Hawkins, M. T. R., and Caballero, S. (2021). Editorial: Assessing biodiversity in the phylogenomic era. *Front. Ecol. Evol.* 9:803188. doi: 10.3389/fevo.2021.803188

Carleton, M. D. (1980). Phylogenetic relationships in neotominae-peromyscine rodents (Muroidea) and a reappraisal of the dichotomy with New World Cricetinae. *Misc. Publ. Mus. Zool. Univ. Mich.* 146, 1–43.

Carleton, M. D. (1989). "Systematics and evolution," in *Advances in the study of Peromyscus (Rodentia)*, eds G. L. Kirkland and J. Layne (Lubbock, TX: Texas Tech University Press), 7–141.

Castañeda-Rico, S., León-Paniagua, L., Edwards, C. W., and Maldonado, J. E. (2020). Ancient DNA from museum specimens and next generation sequencing help resolve the controversial evolutionary history of the critically endangered puebla deer mouse. *Front. Ecol. Evol.* 8:94. doi: 10.3389/fevo.2020.00094

Castañeda-Rico, S., León-Paniagua, L., Vázquez-Domínguez, E., and Navarro-Sigüenza, A. G. (2014). Evolutionary diversification and speciation in rodents of the Mexican lowlands: The *Peromyscus melanophrys* species group. *Mol. Phylogenet. Evol.* 70, 454–463. doi: 10.1016/j.ympev.2013.10.004

Castresana, J. (2000). Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol. Biol. Evol.* 17, 540–552. doi: 10.1093/oxfordjournals.molbev.a026334

Ceballos, G., Arroyo-Cabrales, J., and Ponce, E. (2010). Effects of Pleistocene environmental changes on the distribution and community structure of the mammalian fauna of Mexico. *Quat. Res.* 73, 464–473. doi: 10.1016/j.yqres.2010.02.006

Ceballos, G., Ehrlich, P. R., and Raven, P. H. (2020). Vertebrates on the brink as indicators of biological annihilation and the sixth mass extinction. *Proc. Natl. Acad. Sci. U.S.A.* 117, 13596–13602. doi: 10.1073/pnas.1922686117

Cong, Q., Shen, J., Zhang, J., Li, W., Kinch, L. N., Calhoun, J. V., et al. (2021). Genomics reveals the origins of historical specimens. *Mol. Biol. Evol.* 38, 2166–2176. doi: 10.1093/molbev/msab013

Cornejo-Latorre, C., Cortés-Calva, P., and Álvarez-Castañeda, S. T. (2017). The evolutionary history of the subgenus *Haplomylomys* (Cricetidae: *Peromyscus*). *J. Mammal.* 98, 1627–1640. doi: 10.1093/jmammal/gyx107

Dalquest, W. W. (1962). The good creek formation, Pleistocene of Texas, and its fauna. *J. Paleontol.* 36, 568–582.

Darriba, D., Taboada, G. L., Doallo, R., and Posada, D. (2012). jModelTest 2: More models, new heuristics and parallel computing. *Nat. Methods* 9:772. doi: 10.1038/nmeth.2109

Dawson, W. (2005). "Peromyscine biogeography, Mexican topography and pleistocene climatology," in *Contribuciones Mastozoológicas en Homenaje a Bernardo Villa*, eds V. Sánchez-Cordero and R. Medellín (México: UNAM-CONABIO), 145–156.

Dussex, N., von Seth, K., Knapp, M., Kardailsky, O., Robertson, B. C., and Dalén, L. (2019). Complete genomes of two extinct New Zealand passerines show responses to climate fluctuations but no evidence for genomic erosion prior to extinction. *Biol. Lett.* 15:20190491. doi: 10.1098/rsbl.2019.0491

Eisenberg, J. F. (1989). *Mammals of the Neotropics: The Northern Neotropics*, Vol. 1. Chicago, IL: University of Chicago Press.

Erixon, P., Svennblad, B., Britton, T. and Oxelman, B. (2003). Reliability of Bayesian posterior probabilities and bootstrap frequencies in phylogenetics. *Syst. Biol.* 52, 665–673.

Faircloth, B. C. (2013). Illumiprocessor: A Trimmomatic wrapper for parallel adapter and quality trimming. doi: 10.6079/J9ILL

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

Faircloth, B. C., McCormack, J. E., Crawford, N. G., Harvey, M. G., Brumfield, R. T., and Glenn, T. C. (2012). Ultraconserved elements anchor thousands of genetic markers spanning multiple evolutionary timescales. *Syst. Biol.* 61, 717–726. doi: 10.1093/sysbio/sys004

Ferrusquía-Villafranca, I., Arroyo-Cabrales, J., Martínez-Hernández, E., Gama-Castro, J., Ruíz-González, J., Polaco, O., et al. (2010). Pleistocene mammals of Mexico: A critical review of regional chronofaunas, climate change response and biogeographic provinciality. *Quat. Res.* 217, 53–104. doi: 10.1016/j.quaint.2009.11.036

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., et al. (2011). Full-length transcriptome assembly from RNA-seq data without a reference genome. *Nat. Biotechnol.* 29, 644–652. doi: 10.1038/nbt.1883

Guindon, S., and Gascuel, O. (2003). A simple, fast and accurate method to estimate large phylogenies by maximum-likelihood. *Syst. Biol.* 52, 696–704. doi: 10.1080/10635150390235520

Hall, E. R. (1981). *Mammals of North America*, 2nd Edn. New York, NY: John Wiley and Sons.

Hawkins, M. T. R., Leonard, J., Helgen, K. M., McDonough, M. M., Rockwood, L. L., and Maldonado, J. E. (2016). Evolutionary history of endemic Sulawesi squirrels constructed from UCEs and mitogenomes sequenced from museum specimens. *BMC Evol. Biol.* 16:80. doi: 10.1186/s12862-016-0650-z

Heled, J., and Drummond, A. J. (2012). Calibrated tree priors for relaxed phylogenetics and divergence time estimation. *Syst. Biol.* 61, 138–149. doi: 10.1093/sysbio/syr087

Holmes, M. W., Hammond, T. T., Wogan, G. O. U., Walsh, R. E., Labarbera, K., Wommack, E. A., et al. (2016). Natural history collections as windows on evolutionary processes. *Mol. Ecol.* 25, 864–881. doi: 10.1111/mec.13529

Hooper, E. T. (1947). Notes on Mexican mammals. *J. Mammal.* 28, 40–57.

Hooper, E. T. (1952). A systematic review of the harvest mice (genus *Reithrodontomys*) of Latin America. *Misc. Publ. Mus. Zool. Univ. Mich.* 77, 1–255.

Hooper, E. T., and Musser, G. G. (1964). Notes on classification of the rodent genus *Peromyscus*. *Occas. Pap. Mus. Zool. Univ. Mich.* 635, 1–13.

Hope, A. G., Sandercock, B. K., and Malaney, J. L. (2018). Collection of scientific specimens: Benefits for biodiversity sciences and limited impacts on communities of small mammals. *BioScience* 68, 35–42. doi: 10.1093/biosci/bix141

Huelsenbeck, J. P., and Ronquist, F. (2001). MRBAYES: Bayesian inference of phylogeny. *Bioinformatics* 17, 754–755. doi: 10.1093/bioinformatics/17.8.754

Joìnsson, H., Ginolhac, A., Schubert, M., Johnson, P., and Orlando, L. (2013). mapDamage2.0: Fast approximate Bayesian estimates of ancient DNA damage parameters. *Bioinformatics* 13, 1682–1684. doi: 10.1093/bioinformatics/btt193

Karow, P. F., Morgan, G. S., Portell, R. W., Simmons, E., and Auffenberg, K. (1996). "Middle Pleistocene (early Rancholabrean) vertebrates and associated marine and non-marine invertebrates from Oldsmar, Pinellas County, Florida," in *Palaeoecology and palaeoenvironments of Late Cenozoic mammals: Tributes to the career of C. S. (Rufus) Churcher*, eds K. Stewart and K. Seymour (Toronto, ON: University of Toronto Press), 97–133.

Karwacki, E. E., Martin, K. R., and Savage, A. E. (2021). One hundred years of infection with three global pathogens in frog populations of Florida, USA. *Biol. Conserv.* 257:109088. doi: 10.1016/j.biocon.2021.109088

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

Kehlmaier, C., Zhang, X., Georges, A., Campbell, P. D., Thomson, S., and Fritz, U. (2019). Mitogenomics of historical type specimens of Australasian turtles: Clarification of taxonomic confusion and old mitochondrial introgression. *Sci. Rep.* 9:5841. doi: 10.1038/s41598-019-42310-x

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

Le Conte, J. (1853). Descriptions of three new species of American arvicolae, with remarks upon some other American rodents. *Proc. Acad. Nat. Sci. Philos.* 5, 404–420.

Lyra, M. L., Lourenço, A. C. C., Pinheiro, P. D. P., Pezzuti, T. L., Baêta, D., Barlow, A., et al. (2020). High-throughput DNA sequencing of museum specimens sheds light on the long-missing species of the *Bokermannohyla claresignata* group (Anura: Hylidae: Cophomantini). *Zool. J. Linnean Soc.* 190, 1235–1255. doi: 10.1093/zoolinnean/zlaa033

Martin, P. S., and Klein, R. G. (eds) (1984). *Quaternary extinctions*. Tucson, AZ: University of Arizona Press.

Martínez-Borrego, D., Arellano, E., González-Cózatl, F. X., Castro-Arellano, I., León-Paniagua, L., and Rogers, D. S. (2022). Molecular systematics of the *Reithrodontomys tenuirostris* group (Rodentia: Cricetidae) highlighting the *Reithrodontomys microdon* species complex. *J. Mammal.* 103, 29–44. doi: 10.1093/jmammal/gyab133

McDonough, M. M., Ferguson, A. W., Dowler, R. C., Gompper, M. E., and Maldonado, J. E. (2022). Phylogenomic systematics of the spotted skunks

(Carnivora, Mephitidae, Spilogale): Additional species diversity and Pleistocene climate change as a major driver of diversification. *Mol. Phylogenet. Evol.* 167:107266. doi: 10.1016/j.ympev.2021.107266

McDonough, M. M., Parker, L. D., Rotzel, N., Campana, M. G., and Maldonado, J. E. (2018). Performance of commonly requested destructive museum samples for mammalian genomic studies. *J. Mammal.* 99, 789–802. doi: 10.1093/jmammal/gyy080

McGuire, J. A., Cotoras, D. D., O'Conell, B., Lawalata, S. Z. S., Wang-Claypool, C. Y., Stubbs, A., et al. (2018). Squeezing water from a stone: High-throughput sequencing from a 145-year old holotype resolves (barely) a cryptic species problem in flying lizards. *PeerJ* 6:e4470. doi: 10.7717/peerj.4470

Merriam, C. H. (1898). Descriptions of twenty new species and a subgenus of *Peromyscus* from Mexico and Guatemala. *Proc. Biol. Soc. Washington* 12, 115–125.

Meyer, M., and Kircher, M. (2010). Illumina sequencing library preparation for highly multiplexed target capture and sequencing. *Cold Spring Harb. Protoc.* 2010:pdb.rot5448. doi: 10.1101/pdb.prot5448

Miller, J. R., and Engstrom, M. D. (2008). The relationships of major lineages within peromyscine rodents: A molecular phylogenetic hypothesis and systematic reappraisal. *J. Mammal.* 89, 1279–1295. doi: 10.1644/07-mamm-a-195.1

Minteer, B. A., Collins, J. P., Love, K. E., and Puschendorf, R. (2014). Avoiding (Re)extinction. *Science* 344, 260–261. doi: 10.1126/science.1250953

Musser, G., and Carleton, M. D. (1993). "Family Muridae," in *Mammal species of the world: A taxonomic and geographic reference*, eds D. E. Wilson and M. Reeder (Washington DC: Smithsonian Institution Press), 501–755.

Musser, G., and Carleton, M. D. (2005). "Superfamily Muridae," in *Mammal species of the world: A taxonomic and geographic reference*, eds D. E. Wilson and M. Reeder (Baltimore, MD: Johns Hopkins University Press), 894–1531.

Nakamura, T., Yamada, K. D., Tomii, K., and Katoh, K. (2018). Parallelization of MAFFT for large-scale multiple sequence alignments. *Bioinformatics* 34, 2490–2492. doi: 10.1093/bioinformatics/bty121

Osgood, W. (1909). Revision of the mice of the American genus *Peromyscus*. *North Am. Fauna* 28, 1–285. doi: 10.3996/nafa.28.0001

Platt, R. N. II, Amman, A. M., Keith, M. S., Thompson, C. W., and Bradley, R. D. (2015). What is Peromyscus? Evidence from nuclear and mitochondrial DNA sequences suggests the need for a new classification. *J. Mammal.* 96, 708–719. doi: 10.1093/jmammal/gyv067

Prosser, S. W., Dewaard, J. R., Miller, S. E., and Hebert, P. D. N. (2016). DNA barcodes from century-old type specimens using next-generation sequencing. *Mol. Ecol. Resour.* 16, 487–497. doi: 10.1111/1755-0998.12474

Rambaut, A., Drummond, A. J., Xie, D., Baele, G., and Suchard, M. A. (2018). Posterior summarization in Bayesian phylogenetics using Tracer 1.7. *Syst. Biol.* 67, 901–904. doi: 10.1093/sysbio/syy032

Raxworthy, C. J., and Smith, B. T. (2021). Mining museums for historical DNA: Advances and challenges in museomics. *Trends Ecol. Evol.* 36, 1049–1060. doi: 10.1016/j.tree.2021.07.009

Remsen, J. V. (1995). The importance of continued collecting of bird specimens to ornithology and bird conservation. *Bird Conserv. Int.* 5, 145–180. doi: 10.1017/S095927090000099X

Reyes-Velasco, J., Goutte, S., Freilich, X., and Boissinot, S. (2021). Mitogenomics of historical type specimens clarifies the taxonomy of Ethiopian *Ptychadena* Boulenger, 1917 (Anura, Ptychadenidae). *ZooKeys* 1070, 135–149. doi: 10.3897/zookeys.1070.66598

Rohland, N., and Reich, D. (2012). Cost-effective, high-throughput DNA sequencing libraries for multiplexed target capture. *Genome Res.* 22, 939–946. doi: 10.1101/gr.128124.111

Ronquist, F., and Huelsenbeck, J. P. (2003). MRBAYES 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* 19, 1572–1574. doi: 10.1093/bioinformatics/btg180

Roos, C., Portela Miguez, R., Sabin, R., Louis, E. E. Jr., Hofreiter, M., and Zinner, D. (2021). Mitogenomes of historical type specimens unravel the taxonomy of sportive lemurs (*Lepilemur* spp.) in Northwest Madagascar. *Zool. Res.* 42, 428–432. doi: 10.24272/j.issn.2095-8137.2021.157

Roycroft, E., MacDonald, A. J., Moritz, C., and Rowe, K. C. (2021). Museum genomics reveals the rapid decline and extinction of Australian rodents since European settlement. *Proc. Natl. Acad. Sci. U.S.A.* 118:e2021390118. doi: 10.1073/pnas.2021390118

Rubi, T. L., Knowles, L. L., and Dantzer, B. (2020). Museum epigenomics: Characterizing cytosine methylation in historic museum specimens. *Mol. Ecol. Resour.* 20, 1161–1170. doi: 10.1111/1755-0998.13115

Sayyari, E., and Mirarab, S. (2016). Fast coalescent-based computation of local branch support from quartet frequencies. *Mol. Biol. Evol.* 33, 1654–1668. doi: 10.1093/molbev/msw079

Schmieder, R., and Edwards, R. (2011). Quality control and preprocessing of metagenomic datasets. *Bioinformatics* 27, 863–864. doi: 10.1093/bioinformatics/btr026

Schmitt, C. J., Cook, J. A., Zamudio, K. R., and Edwards, S. V. (2018). Museum specimens of terrestrial vertebrates are sensitive indicators of environmental change in the Anthropocene. *Philoss. Trans. R. Soc. B* 374:20170387. doi: 10.1098/rstb.2017.0387

Shepherd, L. D., and Lambert, D. M. (2008). Ancient DNA and conservation: Lessons from the endangered kiwi of New Zealand. *Mol. Ecol.* 17, 2174–2184. doi: 10.1111/j.1365-294X.2008.03749.x

Stamatakis, A. (2014). RAxML Version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30, 1312–1313. doi: 10.1093/bioinformatics/btu033

Steppan, S., and Schenk, J. J. (2017). Muroid rodent phylogenetics: 900-species tree reveals increasing diversification rates. *PLoS One* 12:e0183070. doi: 10.1371/journal.pone.0183070

Sullivan, J., Arellano, E., and Rogers, D. S. (2000). Comparative phylogeography of Mesoamerican highland rodents: Concerted versus independent response to past climatic fluctuations. *Am. Nat.* 155, 755–786. doi: 10.1086/303362

Sullivan, K. A. M., Platt, R. N. I. I., Bradley, R. D., and Ray, D. A. (2017). Whole mitochondrial genomes provide increased resolution and indicate paraphyly in deer mice. *BMC Zool.* 2:11. doi: 10.1186/s40850-017-0020-3

Tagliacollo, V. A., and Lanfear, R. (2018). Estimating improved partitioning schemes for ultraconserved elements. *Mol. Biol. Evol.* 35, 1798–1811. doi: 10.1093/molbev/msy069

Talavera, G., and Castresana, J. (2007). Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Syst. Biol.* 56, 564–577. doi: 10.1080/10635150701472164

Ward, D. F., Leschen, R. A. B., and Buckley, T. R. (2015). More from ecologists to support natural history museums. *Trends Ecol. Evol.* 30, 373–374. doi: 10.1016/j.tree.2015.04.015

Wright, E. A., Roberts, E. K., Evans, C. L., Schmidly, D. J., and Bradley, R. D. (2020). Evidence from mitochondrial DNA sequences suggest a recent origin for *Peromyscus truei comanche*. *Occas. Pap. Tex. Tech Univ. Mus.* 367, 1–19.

Yuan, S. C., Malekos, E., Cuellar-Gempeler, C., and Hawkins, M. T. R. (2022). Population genetic analysis of the Humboldt's flying squirrel using high-throughput sequencing. *J. Mammal.* 103, 287–302. doi: 10.1093/jmammal/gyac002

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

frontiers | Frontiers in Ecology and Evolution

TYPE Original Research
PUBLISHED 11 October 2022
DOI 10.3389/fevo.2022.910084

Life-history traits drive spatial genetic structuring in Dinaric cave spiders

Martina Pavlek[1,2,3*†], Jérémy Gauthier[4†], Vanina Tonzo[3], Julia Bilat[4], Miquel A. Arnedo[3] and Nadir Alvarez[4,5,6]

[1]Ruđer Bošković Institute, Zagreb, Croatia, [2]Croatian Biospeleological Society, Zagreb, Croatia, [3]Department of Evolutionary Biology, Ecology, and Environmental Sciences, Biodiversity Research Institute (IRBio), Universitat de Barcelona, Barcelona, Spain, [4]Geneva Natural History Museum, Geneva, Switzerland, [5]Department of Genetics and Evolution, University of Geneva, Geneva, Switzerland, [6]Natural Sciences Museum, Lausanne, Switzerland

OPEN ACCESS

EDITED BY
Mozes Blom,
Museum of Natural History Berlin (MfN), Germany

REVIEWED BY
Darija Josic,
Museum of Natural History Berlin (MfN), Germany
Shahan Derkarabetian,
Harvard University, United States
Luisa Teasdale,
Max Planck Institute for Developmental Biology, Germany

*CORRESPONDENCE
Martina Pavlek
martina.pavlek@irb.hr

†These authors share first authorship

SPECIALTY SECTION
This article was submitted to Phylogenetics, Phylogenomics, and Systematics, a section of the journal Frontiers in Ecology and Evolution

RECEIVED 31 March 2022
ACCEPTED 02 September 2022
PUBLISHED 11 October 2022

CITATION
Pavlek M, Gauthier J, Tonzo V, Bilat J, Arnedo MA and Alvarez N (2022) Life-history traits drive spatial genetic structuring in Dinaric cave spiders. *Front. Ecol. Evol.* 10:910084. doi: 10.3389/fevo.2022.910084

The subterranean ecosystem exerts strong selection pressures on the organisms that thrive in it. In response, obligate cave-dwellers have developed a series of morphological, physiological, and behavioral adaptations, such as eye reduction, appendage elongation, low metabolic rates or intermittent activity patterns, collectively referred to as troglomorphism. Traditionally, studies on cave organisms have been hampered by the difficulty of sampling (i.e., small population sizes, temporal heterogeneity in specimen occurrence, challenges imposed by the difficult-to-access nature of caves). Here, we circumvent this limitation by implementing a museomics approach. Specifically, we aim at comparing the genetic population structures of five cave spider species demonstrating contrasting life histories and levels of troglomorphism across different caves in the northern Dinarides (Balkans, Europe). We applied a genome-wide hybridization-capture approach (i.e., HyRAD) to capture DNA from 117 historical samples. By comparing the population genetic structures among five species and by studying isolation by distance, we identified deeper population structuring and more pronounced patterns of isolation by distance in the highly troglomorphic *Parastalita stygia* and *Stalita pretneri* ground dwellers, while the three web-building *Troglohyphantes* species, two of which can occasionally be found in surface habitats, showed less structured populations compatible with higher dispersal ability. The spatial distribution of genetic groups revealed common phylogeographic breaks among lineages across the studied species, which hint at the importance of environmental features in driving dispersal potential and shaping underground diversity.

KEYWORDS

cave-dwelling spiders, Dinarides, subterranean dispersal, subterranean gene flow, HyRAD, population genomics

Introduction

The underground habitat, with its conspicuous features such as complete darkness and low availability of food, is a very different environment compared to the surface. The absence of light and thus of primary producers, as well as stable climatic conditions over time, have imposed strong adaptive constraints on organisms that thrive therein, the troglobionts. As a response to the extreme environmental conditions, cave-dwelling organisms have evolved a series of morphological, physiological, and behavioral adaptations, such as elongated appendages or reduced visual system, collectively referred to as troglomorphisms (Christiansen, 2012). The two prevailing theories of the colonization and speciation of cave animals are the climate-relict and the adaptive-shift hypotheses (Barr, 1968; Howarth, 1987), which presume that isolation of cave populations happens either by extinction of the surface population (due to climatic changes) or by divergent natural selection resulting in reduced gene flow, respectively. Once diverged from their closest surface relatives, cave species, which have become highly adapted and dependent on stable underground conditions, are thought to be unable to use surface habitats for dispersal, which could still occur in a restricted manner through a single aquifer or a fissured and permeable geotectonic unit (Trontelj, 2018). The dispersal potential of cave species depends on the continuity and the size of the limestone outcrops in which caves develop (Barr, 1967; Bregović and Zagmajster, 2016), and on the potential of a given region for long cave passages, which is directly linked to the total length of subterranean voids (Curl, 1986; Culver et al., 2004). Most cave species have restricted distributions when compared to surface relatives, and some are known from one or from a limited number of nearby caves (Ribera et al., 2018). Still, there are some examples of species with large distributions (Lefébure et al., 2006; Eme et al., 2013).

Despite the possibility of dispersing through aquifers or fissures in the rock, connectivity among karstic areas is generally limited due to the impermeable landscape separating them, and thus organisms strictly associated with caves might show population dynamics similar to those found in island species (Barr and Holsinger, 1985), such as counter-selection for traits related to dispersal (Zimmerman, 1949; Borregaard et al., 2017; Salces-Castellano et al., 2020), and strong adaptive traits to their local environment. Since isolation and local adaptation are crucial mechanisms governing patterns of gene flow among populations, it is expected that the level of troglomorphism of a given species relates to the pattern of population structuring (Caccone, 1985; Sbordoni et al., 2000; Trontelj, 2018). Several studies on cave arthropods revealed lower levels of gene flow in troglobiont species compared to troglophiles (cave species able to survive and disperse through surface habitats) (Caccone, 1985; Sbordoni et al., 2000). However, reduced gene flow and highly structured populations were for instance found in North American *Nesticus* spiders regardless of the level of cave adaptation (Hedin, 1997). Still, in the absence of physical barriers to dispersal, the levels of gene flow among co-occurring cave arthropod populations should be better explained by the intrinsic characteristics of the organisms Caccone (1985).

The Dinarides, a mountain chain in the western Balkans (south-eastern Europe), is a global hotspot of cave biodiversity with more than 1,000, mostly endemic, obligate cave species (Sket et al., 2004; Culver et al., 2006; Sket, 2012; Jalžić et al., 2013). The spiders, with 101 species, rank second among the most species-rich terrestrial groups, only after beetles (Pavlek and Mammola, 2021). Around 1,000 cave spider species have been recorded worldwide (Mammola and Isaia, 2017), 10% of which are found in the Dinarides. This exceptional diversity could be explained by the abundance of suitable habitat composed of more than 20,000 karstic caves (Zupan Hajna, 2019), by habitat heterogeneity (Bregović and Zagmajster, 2016), and by the long-term climatic stability and high productivity of the region (Culver et al., 2006). During the Pleistocene, the Dinarides remained mostly ice-free (Mihevc et al., 2010), giving lineages the opportunity to survive, disperse and colonize new caves.

The two most speciose groups of cave spiders from the Dinarides belong to the families Dysderidae and Linyphiidae, in the second case mostly restricted to the genus *Troglohyphantes* (Sket et al., 2004). Dysderidae and *Troglohyphantes* species are characterized by contrasting lifestyles. Dysderidae do not build webs, but wander through the cave passages actively hunting their prey, while *Troglohyphantes* species build sheet webs near the substrate, from which they hang upside down (**Figure 1**). The majority of Dinaric Dysderidae species display extreme troglomorphic traits, while *Troglohyphantes* species exhibit levels of troglomorphisms ranging from shallow to extreme. Based on their different life-history traits, we predict that the highly troglomorphic Dysderidae species would show deeper population structure and steeper isolation by distance, compared to the less troglomorphic *Troglohyphantes* species, assuming the latter demonstrate better dispersal abilities. To test this hypothesis, we selected samples from a region in the north-western part of the Dinarides (**Figure 2**) where two Dysderidae species, *Stalita pretneri* (Deeleman-Reinhold, 1971) and *Parastalita stygia* (Joseph, 1882), and three *Troglohyphantes* species, *Troglohyphantes excavatus* (Fage, 1919), *T. croaticus* (Chyzer, 1894) and *Troglohyphantes kordunlikanus* (Deeleman-Reinhold, 1978), co-exist. All species are endemic to the northern Dinarides (**Figure 2**), except *T. excavatus* that reaches the southern Austrian Alps (Deeleman-Reinhold, 1978; Thaler, 1986; Pavlek and Mammola, 2021). The two Dysderidae species are eyeless, highly depigmented (**Figures 1A,B**), and have never been collected outside caves (Deeleman-Reinhold, 1971; Pavlek and Mammola, 2021). They both belong to a clade of highly cave-adapted species (Pavlek and Mammola, 2021). On the other hand, the three *Troglohyphantes* species have eyes of variable size, show variable levels of depigmentation, and two of them (*T. excavatus* and *T. kordunlikanus*) are occasionally found in dark and humid places outside caves (Deeleman-Reinhold, 1978). Although no molecular data is available for these species, morphological characters suggest that the three species belong to two distantly related lineages—*T. croaticus* and *T. excavatus* belong to the *croaticus* group, while *T. kordunlikanus* belongs to the *polyophthalmus* group (Deeleman-Reinhold, 1978; Isaia et al., 2017).

Cave fauna are generally difficult to sample due to the technical obstacles in entering and exploring the caves and pits (Zagmajster et al., 2010). In addition, population densities seem to be low and show a strong temporal variability of their presence. To overcome these sampling limitations, we took advantage of the sampling carried out over the past decades, stored in natural history collections. Despite not having been preserved in DNA-compliant conditions, these samples have the advantage of being accessible. Recent developments in molecular biology, such as hybridization capture methods, e.g., HyRAD (Suchan et al., 2016), facilitates recovery of genetic information from museum samples for which the DNA is often degraded and in low quantities (Toussaint et al., 2021). By applying museomics techniques, we retrieved DNA polymorphisms and inferred population genetic structures in each of the five species described above. Specifically, we applied the HyRAD technique and the popHyRAD pipeline (Gauthier et al., 2020) to specimens from the natural history collection of the Croatian Biospeleological Society (CBSS) from Zagreb, which holds one of the largest collections of cave spiders in the world. We then performed a comparative phylogeographic approach involving the five above mentioned co-distributed cave-dwelling spider species to test (1) if the degree of troglomorphism relates to the patterns of isolation by distance, and (2) whether the studied species underwent similar phylogeographic processes and thus share the same breaks in the spatial patterning of their genetic variation, or, alternatively, if spatial population structure is distinctive to each species.

Materials and methods

Sampling

Our sampling performed at the Croatian Biospeleological Society (CBSS Collection, Zagreb, Croatia), ZCSL (Zoological collection of SubBioLab, University of Ljubljana, Ljubljana, Slovenia), and ROC (Roman Ozimec collection, Zagreb, Croatia) natural history collections encompasses 117 specimens from the five studied species (i.e., 34 specimens of *P. stygia*, 26 of *S. pretneri*, 12 of *T. kordunlikanus*, 24 of *T. excavatus*, and 21 of *T. croaticus*) (**Figure 2** and **Supplementary Table 1A**). The oldest sample was collected in 1974, 5 samples were collected before 2000, 49 samples were collected between 2000 and 2010, and 61 were collected between 2010 and 2018. Their preservation conditions varied in different concentrations of ethanol, ranging from absolute to 40% ethanol, and all samples were kept at room temperature (**Supplementary Table 1**).

Since the information on the level of troglomorphy as suggested by eye reduction and depigmentation for the three *Troglohyphantes* species was only reported in a few relatively old papers (Fage, 1919; Deeleman-Reinhold, 1978), and include very limited geographic sampling, we checked the level of depigmentation and eye reduction in all available specimens from the CBSS collection. Observed troglomorphic patterns are reported in **Supplementary Table 1B**. We used the QGIS software in order to represent the distribution of specimens on maps.

DNA extraction and HyRAD protocol

To retrieve DNA from old and poorly preserved samples, the HyRAD protocol (Suchan et al., 2016) was applied following Toussaint et al. (2021) with some modifications specified below. To produce the probes used in the capture process, we first extracted DNA from four fresh specimens corresponding to the studied species: sample psty_5115_1 for *P. stygia*, sample spre_5118 for *S. pretneri*, sample tcro_4274 for *T. croaticus* and *T. excavatus* (as the two species are closely related), and sample tkor_5227_1 for *T. kordunlikanus* (**Supplementary Table 1A**). Genomic DNA was extracted using Speedtools Tissue DNA Extraction Kit (Biotools), DNeasy Blood, and Tissue Kit (Qiagen) and QIAamp DNA Micro Kit (Qiagen), depending on the sample condition. The preparation of the ddRAD library comprised a digestion with the restriction enzymes *Mse*I and *Pst*I-HF (New England Biolabs, Ipswich, MA, USA), ligation of adaptors and individual barcodes, size selection with Blue Pippin (2% dye-free Agarose Gel Cassette marker V1, Sage Science) with a range of 190–240 bp and a final amplification by PCR for 20 cycles using NEBNext Hi-Fi 2X PCR Master Mix (New England Biolabs, Ipswich, MA, USA). An aliquot of the final library was sequenced on one lane of an Illumina MiSeq 150 bp paired-end at the Lausanne Genomic Technology Facility (LGTF) in order to obtain a sequence catalog of the loci represented in the ddRAD probes, and the rest of the library was transcribed into RNA probes and biotinylated using HiScribe T7 High Yield RNA Synthesis Kit (New England Biolabs, Ipswich, MA, USA).

Historical DNA from collection samples was extracted using the same DNA extraction kits as for the probes design, and in some cases only one or few legs were used through non-destructive extraction consisting in putting the material in a buffer with proteinase K overnight, and returning it back to the collection specimen. The purified DNA was quantified and the quality was assessed using a Fragment

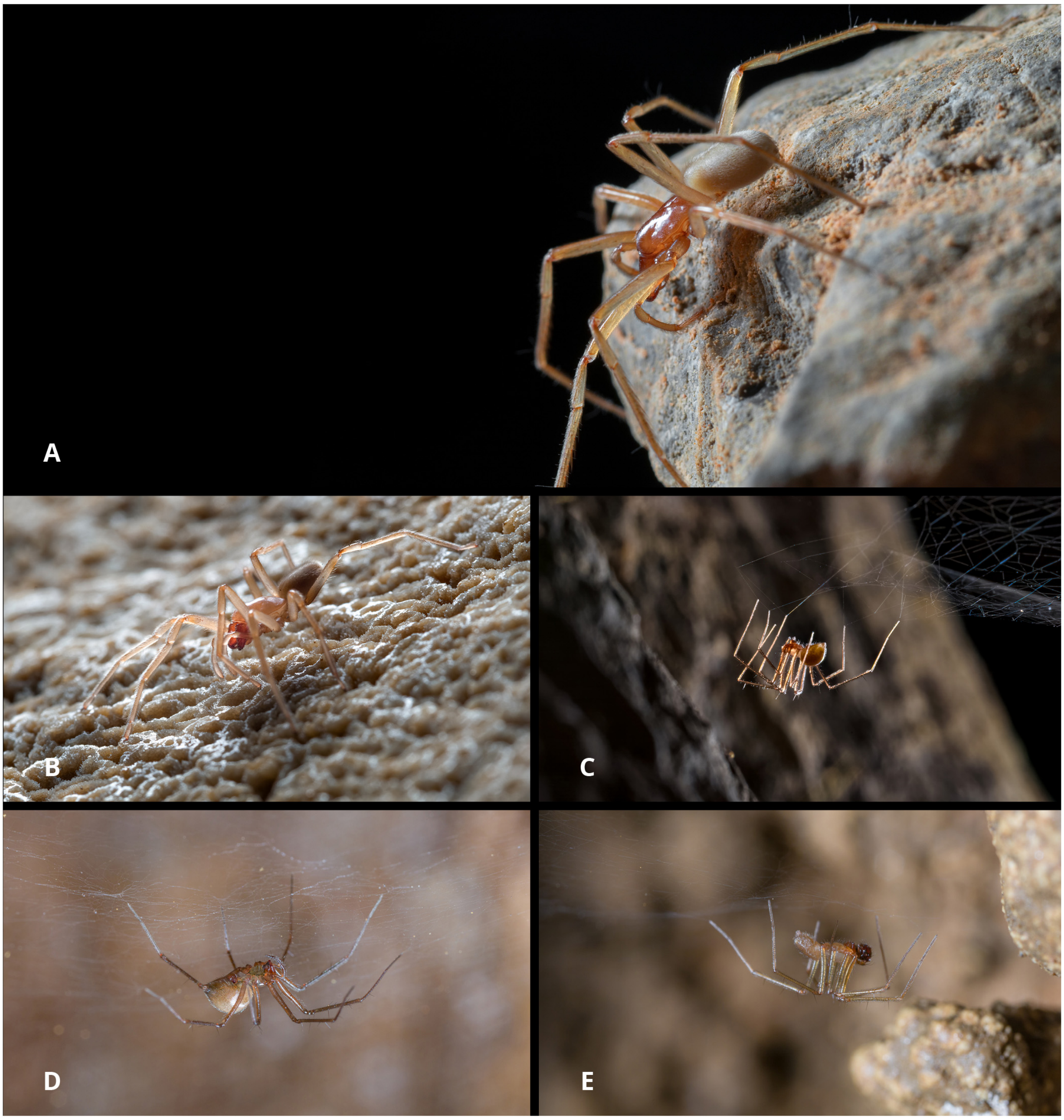


FIGURE 1
Photographs of the five studied species. **(A)** *Parastalita stygia*, **(B)** *Stalita pretneri*, **(C)** *Troglohyphantes croaticus*, **(D)** *Troglohyphantes excavatus*, **(E)** *Troglohyphantes kordunlikanus*. Photos by: **(A,B)** Tin Rožman; **(C)** Jana Bedek; **(D,E)** MP.

Analyzer. For specimens with large DNA fragment sizes (>1 kb) a shearing step was performed using the NEBNext® dsDNA Fragmentase® (New England Biolabs, Ipswich, MA, USA) protocol, except that only 1 μl of enzyme was used. A shotgun library preparation was applied to each sample following Suchan et al. (2016), comprising phosphorylation with T4 Polynucleotide Kinase, heat-denaturation, G-tailing with Terminal Transferase, second strand DNA synthesis with Klenow Fragment (3'–>5' exo-) using a poly-C oligonucleotide, blunt-end reaction with T4 DNA Polymerase, barcoded adapters ligation to the phosphorylated end with T4 DNA ligase, and PCR amplification using Phusion U Hot Start DNA Polymerase (Thermo Scientific). Libraries were purified and quantified using Quant-iT PicoGreen® dsDNA reagent (Invitrogen) on a Hidex Sense Microplate reader and pooled equimolarly based on their concentration. Hybridization capture was performed with a

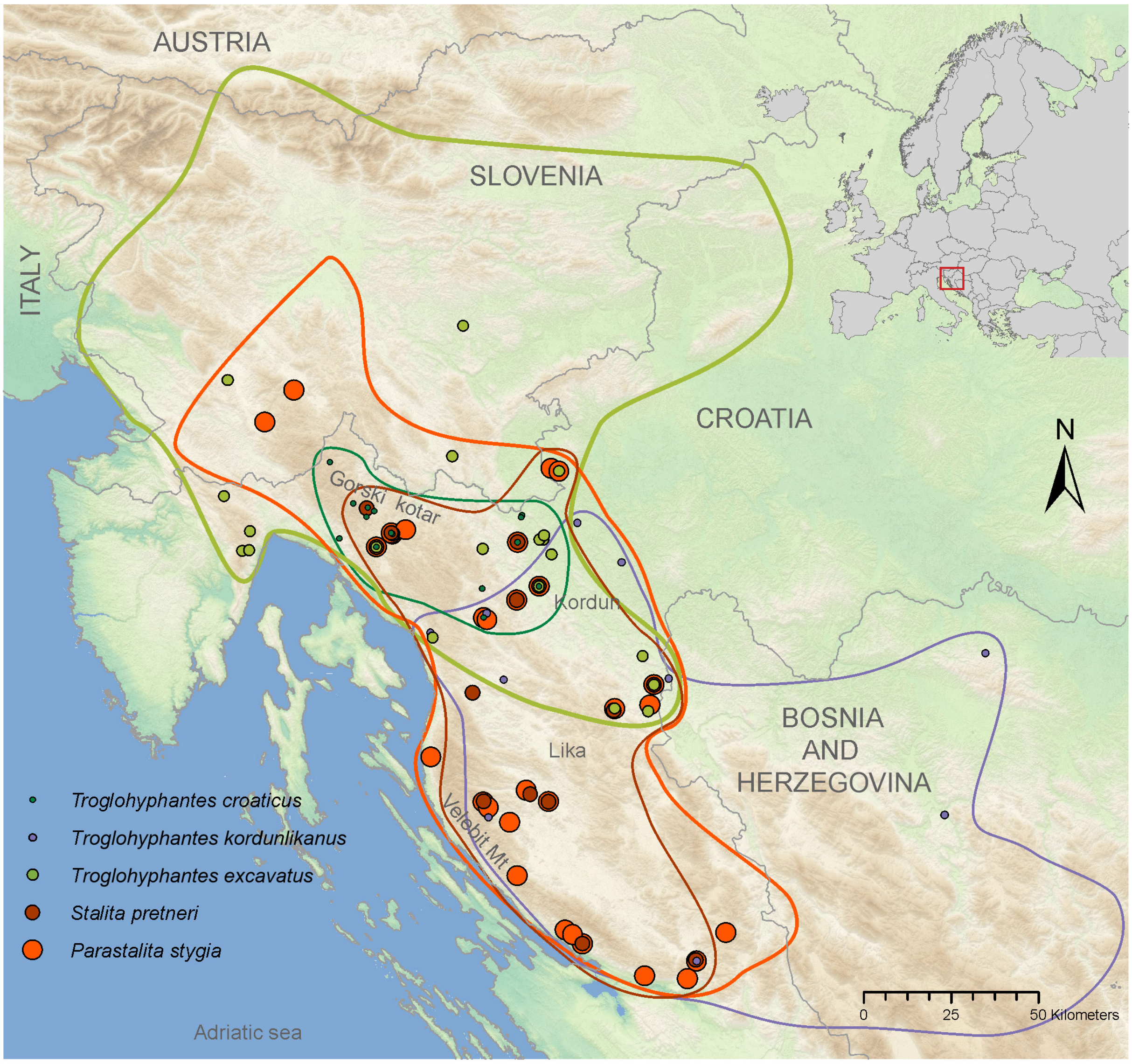


FIGURE 2
Map showing the distributions of each of the five studied species. Dots represent the localities from specimens used in this study, and lines in matching colors encircle the species' distributions. Overlapping dots represent caves with more than one species. Changes in dot sizes are used as a means of showing overlapping species presences.

two-step capture at two temperatures, i.e., 55 and 65°C to improve the stringency of the reaction, as suggested by Li et al. (2013) as well as Suchan et al. (2022). Enriched libraries were sequenced on two lanes of an Illumina HiSeq2500 using a paired-end 100 bp protocol at the LGTF.

Demultiplexing and Single Nucleotide Polymorphisms identification

The first part of the bioinformatic pipeline consisted in the identification of the ddRAD loci present in the probes specimens used for the capture, in order to build a reference catalog. The probe reads generated from the ddRAD libraries were demultiplexed according to barcodes and cleaned using Cutadapt v1.18 (Martin, 2011) to remove adaptors, bases with a quality lower than 20 and reads smaller than 30 bp. Read quality was checked using FastQC v0.11.8 (Babraham Institute, Babraham, England). Loci construction was performed for each species individually, except for the pair of sister species *T. croaticus* and *T. excavatus*. Ipyrad v0.7.30 (Eaton and Overcast, 2020) was used with a minimum depth of six and a clustering threshold of 0.80 (selected threshold after testing values of 0.70, 0.80, and 0.90). Shared loci were retained to build the reference catalogs for each species. In practice, a catalog of 23,031 loci was used for the mapping of *T. croaticus* and *T. excavatus*, a catalog of 20,691 loci for *T. kordunlikanus*, a catalog of 22,150 loci for *P. stygia*, and a catalog of 23,022 loci for *S. pretneri*. Reads from historical samples were cleaned using Cutadapt v1.18 (Martin, 2011) to remove barcodes, adaptors, terminal poly-Cs, and bases with a quality lower than 20 and reads smaller than 30 bp. Read quality was checked using FastQC v0.11.8 (Babraham Institute, Babraham, England). Clean reads were individually mapped on the corresponding probe catalogs using BWA-ALN v0.6 (Li and Durbin, 2009). Indels were realigned using the GATK IndelRealigner v3.8 (McKenna et al., 2010), PCR duplicates were removed using MarkDuplicates from the Picard toolkit v2.20.2[1], and base quality was rescaled using MapDamage v2.0 (Jónsson et al., 2013) to take into account post-mortem DNA deamination.

To verify the species status of each sample, the phyloHyRAD pipeline (Toussaint et al., 2021) was applied to reconstruct the ddRAD sequence alignment for each locus and perform phylogenetic inferences within the two groups. Consensus sequences were generated from the previous mapping files using samtools mpileup v1.4, bcftools v1.4 and vcfutils, keeping the main bases and a minimum coverage of three. Resulting consensus loci were combined from the shared loci resulting from the abovementioned iPyRAD analysis, i.e., at the family level for Dysderidae and at the genus level for *Troglohyphantes*. They were further aligned using MAFFT v7.407 (Katoh et al., 2002). Alignments were cleaned to keep loci shared by at least one third of the samples and checked manually. We finally obtained 1,915 loci for *P. stygia*, 989 loci for *S. pretneri* and 2,276 loci for the three *Troglohyphantes* species. Loci were concatenated using AMAS v1.02 (Borowiec, 2016). Best partitioning schemes were estimated using PartitionFinder2 (Lanfear et al., 2017), and corresponding models of nucleotide substitution were determined using ModelFinder (Kalyaanamoorthy et al., 2017). Phylogenetic inferences were performed on the concatenated alignment using IQ-TREE v 1.6.11 (Minh et al., 2020), and branch support was estimated using 1,000 ultrafast bootstraps along with 1,000 SH-aLRT tests.

To identify genetic variations in each species confirmed by the phylogenetic approach, the PopHyRAD pipeline was applied (Gauthier et al., 2020). A variant calling was performed between each sample from the same species using Freebayes v1.3.1 (Garrison and Marth, 2012). Single Nucleotide Polymorphisms (SNPs) were filtered to keep only bi-allelic SNPs with a calling quality above 100 and shared by at least 60% of the samples in each species using vcftools (Danecek et al., 2011). Finally, samples with more than 90% of missing data, i.e., 11 samples in total, were removed.

Population genetic structure analyses

In order to investigate population genetic structure in the five species, we extracted a subset of unlinked SNPs by randomly selecting one SNP per locus. To infer intraspecific genetic structures, we used principal component analysis (PCA) as implemented in the adegenet R package v2.1.5 (Jombart and Ahmed, 2011). PCA plots were performed using direct PCA values (**Figure 3**), and integrating eigenvalues to take into account the weight of the PC axis (**Supplementary Figure 1**). Secondly, a Bayesian admixture analysis as implemented in STRUCTURE v 2.3.4 (Pritchard et al., 2000) was performed. We ran analyses with *K* ranging from 1 to 10, assuming correlated allele frequencies and admixture, and performed 10 independent replicates for each *K* with 200,000 Markov chain Monte Carlo including a burn-in step of 10,000 iterations. We evaluated the number of genetic clusters that best describes our data according to log likelihoods of the data (LnPr (*X*| *K*) for each value of *K* (Pritchard et al., 2000) and the Δ*K* method (Evanno et al., 2005) with Structure Harvester (Earl and VonHoldt, 2012). We used *CLUMPP* and the Greedy algorithm to align multiple runs of STRUCTURE for the same *K* value (Jakobsson and Rosenberg, 2007), and distruct (Rosenberg, 2003) to plot the individual's cluster membership probabilities. Additionally, to overcome the STRUCTURE constraints related to discrete population inference and to investigate putative continuous patterns of population structure, we also analyzed data under a model-based method that simultaneously infers continuous and discrete patterns integrating geographic distances, as implemented in conStruct (Bradburd et al., 2018). To test the best fit to the data between discrete versus continuous clusters, we used the cross-validation procedure implemented in the conStruct package for each species with *K* ranging from 1 to 10, with three repetitions for each *K* value, 15,000 iterations per repetition, and a training proportion of 0.9. The best *K* was identified based on comparing predictive accuracy for each *K* and for each model, i.e., with and without geographic distances. Considering that the larger the number of K tested, the larger the number of parameters implemented (potentially inducing an overparameterization and artificially increase the accuracy; Bradburd et al., 2018), the best K can be considered as the lowest *K* for which the accuracy has reached a plateau. Results were visualized using plots generated by conStruct.

For each population, descriptive statistics including the observed heterozygosities (Ho), mean gene diversities within population (Hs), allelic richness (Ar), and inbreeding coefficient (FIS) were estimated using the hierfstat package v0.5-7 (Goudet, 2005). Genetic differentiation (FST) between pairs of populations was estimated using vcftools (Danecek et al., 2011).

Isolation by distance (IBD) was calculated as follows: first, we performed a simple Mantel test as implemented in the R package vegan (Oksanen et al., 2020) to compare genetic and

[1] http://broadinstitute.github.io/picard

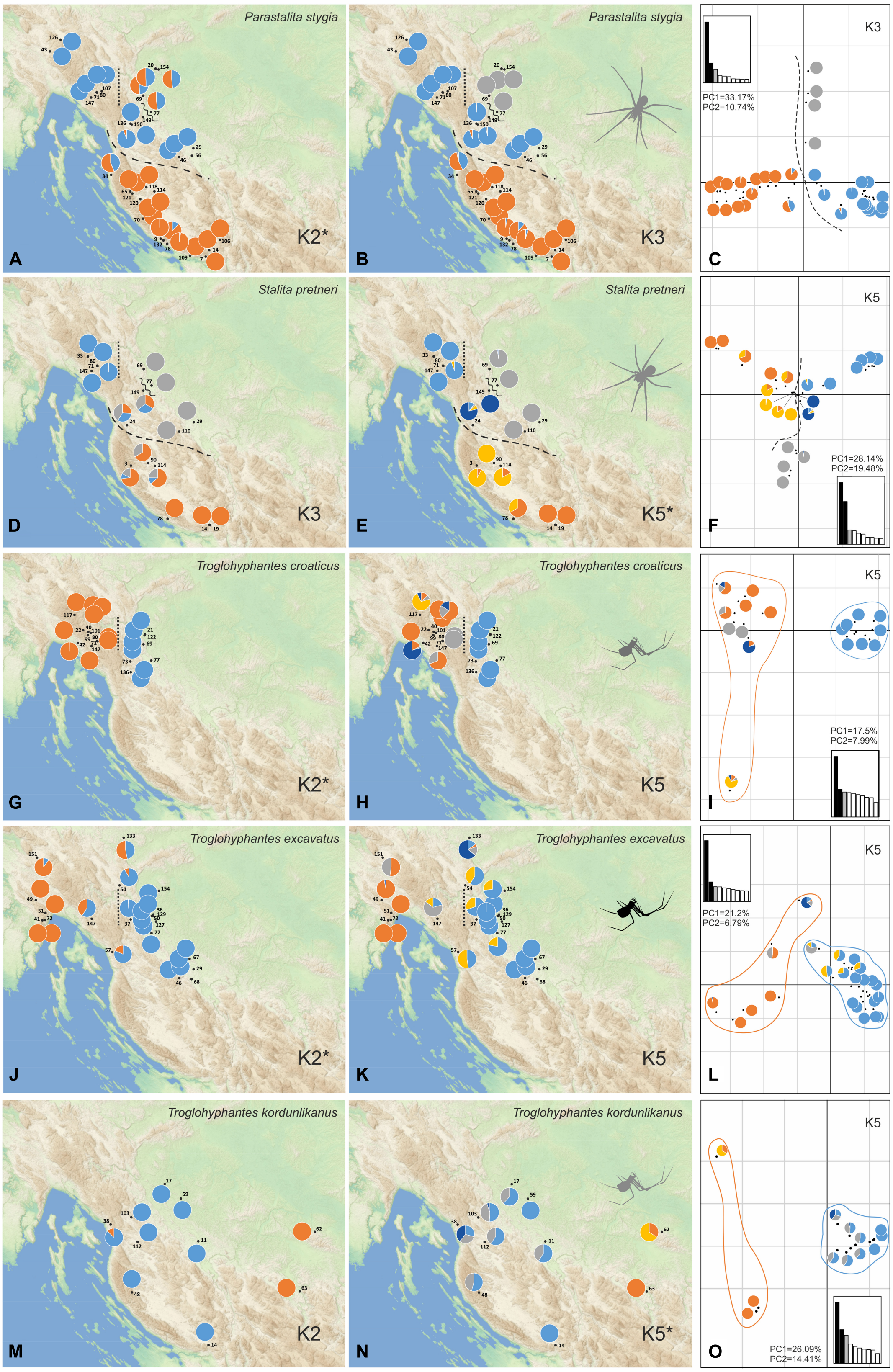


FIGURE 3
(Continued)

FIGURE 3
Population structure of the five studied species as inferred by STRUCTURE and principal component analyses (PCA) analyses. Colors on the pie charts correspond to the proportion of each specimen assigned to each of the genetic clusters inferred for a given *K* value. For each species, both the optimal *K* value and an additional one showing a relevant geographical pattern with good statistical support, according to log probabilities of the data (LnPr (*X*| *K*)), are presented. **(A)** Scatterplot of the first two PCA components is given for every species with a pie chart taken from one of the best *K* values associated with each locality. Main population breaks are marked with dashed and dotted lines. *Parastalita stygia* maps with K2 **(A)**, K3 **(B),** and PCA **(C)**; *Stalita pretneri* maps with K3 **(D)**, K5 **(E)** and PCA **(F)**; *Troglohyphantes croaticus* maps with K2 **(G)**, K5 **(H)** and PCA **(I)**; *Troglohyphantes excavatus* maps with K2 **(J)**, K5 **(K),** and PCA **(L)**; *Troglohyphantes kordunlikanus* maps with K2 **(M)**, K5 **(N),** and PCA **(O)**. In *Troglohyphantes* species PCAs **(I,L,O)**, pies denote clustering with *K*5, and lines encircle *K*2 grouping. Optimal *K* for each species is marked with an asterisk.

geographic distances for all sample pairs, and the significance was assessed with 999 permutations. Genetic distances were calculated as pairwise Nei's distance values between samples using the *dist.genpop* function in the R package adegenet (Jombart and Ahmed, 2011), and geographic coordinates of caves were used to calculate straight-line geographic distances between sampling sites using the *pointDist* function in the R package enmSdm (Morelli et al., 2020). Secondly, a Spearman's correlation test was applied for testing the correlation between genetic and geographic distances, and a linear model including an estimate of goodness-of-fit (*R*-squared), slope, and intercept, was set. These analyses were performed for each species. In addition, supplementary analyses were made for some species after excluding population pairs that demonstrated a pattern of very strong genetic isolation.

We used Stairway Plot 2 (Liu and Fu, 2020) to infer the demographic history of the two main genetic groups established by $K = 2$ in STRUCTURE for each species (**Supplementary Table 2**). Site frequency spectrums were estimated using easySFS.[2] We assumed a mutation rate per site per generation of 2.8×10^{-9} as estimated for *Drosophila melanogaster* (Keightley et al., 2014), and a generation time of 2 years according to our knowledge on the species biology.

Results

Museomics and HyRAD efficiency

The HyRAD approach allowed the recovery of a large number of reads and loci from the historical samples regardless of their age and storage conditions (**Supplementary Tables 1, 2**). The capture and sequencing method resulted in a total of 470 million reads distributed across the 106 samples for which HyRAD was successfully applied (of a total of 117 samples) with a mean of 4.44 million reads per sample (sd = 1.63 millions). The cleaning of reads resulted in a loss of 2.3% of reads (details in **Supplementary Table 2**). The oldest sample in the analysis was a sample of *T. croaticus* (ARCBSS 3422) collected in 1974 and despite being nearly 50 years old, the analysis yielded 6.14 million reads. Mapping of historical clean reads to the corresponding probe catalog resulted in mean mapping percentage of 55.6% (sd 7.2). After another stringent cleaning step, this proportion decreased to 25.4% (sd 7.3), corresponding to a mean of 14,367 bi-allelic SNPs (sd 4657) shared by at least 60% of the samples in each species and a mean read coverage of 43.85 (**Table 1** and **Supplementary Table 2**). Finally, in order to perform genetic structure analyses, we kept one single SNP per locus, resulting in a mean of 6847 unlinked SNP by sample (sd 1967). The summarized data on the number of samples, SNPs, and the overall percentage of missing data is given in **Table 1** (see also **Supplementary Table 2**).

Population genetic structure

Phylogenetic inferences performed on *P. stygia*, *S. pretneri*, and the three *Troglohyphantes* species, confirmed the morphological identifications made on the historical samples and their species-level status, at least on the basis of our sampling (**Supplementary Figure 2**). Population genetic analyses were then carried out within each species. The STRUCTURE and conStruct analyses showed different patterns of genetic clustering for each of the five studied species (STRUCTURE statistics and conStruc results are shown in **Supplementary Table 3** and **Supplementary Figure 3**, respectively). For each species, a map was made with pie charts corresponding to the STRUCTURE Bayesian assignment probabilities per sampling location exemplified for the best run of each selected *K* values (**Figure 3** and **Supplementary Figures 3, 4**). **Supplementary Figures 3, 4** show the summed results of all *K* runs for each species. Overall, the conStruct results showed that the spatial model presents, for each species, a better accuracy than the non-spatial model (**Supplementary Figure 3**), highlighting the fact that the clustering recovered in the STRUCTURE analysis is not the consequence of isolation by distance alone. Thus, additional phylogeographic breaks or ecological traits potentially driving lineage divergence might be at work in the case of all five species studied here.

For *P. stygia*, STRUCTURE analyses yielded an optimal clustering value for $K = 2$ according to the ΔK criterion,

[2] https://github.com/isaacovercast/easySFS

TABLE 1 Number of samples and single nucleotide polymorphisms (SNPs) per species after filtering steps.

Species	# Sample	# All SNP	# Unlinked SNP	% Missing
Parastalita stygia	32	17521	9212	24.69
Stalita pretneri	21	14911	8356	33.63
Troglohyphantes croaticus	18	27206	11883	26.82
Troglohyphantes excavatus	23	22163	9656	30.13
Troglohyphantes kordunlikanus	12	13621	7418	14.98

while the log likelihood of the data [L (K) (mean ± SD)] (see **Supplementary Table 3**) increased steadily up to K = 6. ConStruct results also confirm the structuring of the samples in clear genetic groups; we observe a plateau in the accuracy values after K = 3, a result that might be interpreted as three clusters being the optimal way to partition the data, as well as a sustained contribution of these three groups to the total covariation between the clusters (**Supplementary Figure 3A**). The genetic clusters inferred by STRUCTURE at K = 2 follow a north-south geographical distribution pattern with the break being located in the northern Lika region (dashed line in **Figures 3A–C**). Several hybrid specimens were identified, one near the break-line, and a group of four samples in the central-north area. This main genetic differentiation is also denoted in the PCA analyses, with axis 1 describing the two main genetic groups (**Figure 3C**). The second best STRUCTURE clustering scenario according to the ΔK criterion is K = 3, which is also considered as the best K value in conStruct. It revealed the presence of a third cluster, distributed across the central and northern sampling area (gray in **Figures 3B,C**) and separated by two break lines from the central (blue) group—one in the eastern part of Gorski kotar region (dotted line in **Figures 3A,B**), and a second in the area of the Ogulin-Plaški valley (wavy line in **Figures 3A,B**). On the PCA, this separation is supported by axis two (**Figure 3C**). Analyses with a higher number of clusters, e.g., K = 6, revealed additional structuring within the large southern group, south to the northern Lika break (**Supplementary Figure 4A**).

For *S. pretneri*, STRUCTURE analyses showed that log probabilities of the data [L (K) (mean ± SD)] reached a plateau at K = 5. Also, the ΔK criterion indicated K = 5 as the best clustering solution, and K = 3 as the second best (**Supplementary Table 3**). In the conStruct analysis, accuracy reached a plateau after K = 5, similarly as revealed by STRUCTURE results, but additional spatial layers beyond K = 4 only contributed marginally to the total covariance among clusters. The same breaks as in *P. stygia* were recognized in *S. pretneri*—one in the northern Lika (dashed line in **Figures 3D–F**), one in the eastern part of the Gorski kotar region (dotted line in **Figures 3D,E**), and one in Ogulin-Plaški valley (wavy line in **Figures 3D,E**). The PCA yielded analogous results to those obtained with STRUCTURE, separating northern from southern populations along axis 1, and grouping populations according to their geographical location.

For *T. croaticus* the optimal clustering solution identified with STRUCTURE was K = 2 based on both the ΔK criterion and the log probabilities of the data (**Supplementary Table 3**)—there is a clear genetic and geographic separation between eastern (blue in **Figure 3G**) and western (orange in **Figure 3G**) individuals, and no hybrids were detected. In the conStruct results, the predictive accuracy reached a plateau at K = 2, and the layer contribution is clearly more important for the two main clusters even when the number of K tested is larger (**Supplementary Figure 3C**). In addition, the conStruct analysis revealed some level of admixture between them. The position of a genetic break at the eastern part of the Gorski kotar region (dotted line in **Figures 3G,H**) is the same as in the two Dysderidae species (**Figures 3A,B,D,E**). Although less supported by the ΔK criterion, the K = 5 in STRUCTURE, revealed substructuring only in the western group, with the locality Sniježna jama (cave 117, yellow in **Figure 3H**) genetically distant (**Figure 3I**) from the other samples, even those that were geographically very close. The other localities of the western group were more admixed (**Figure 3H**).

For *T. excavatus* the best K value according to the ΔK criterion in STRUCTURE was also K = 2, with K = 5 being the second best (**Supplementary Table 3**). In conStruct, the layer contribution analysis showed that one of the clusters had a dominant contribution, and that after K = 2 the contribution of other clusters was very limited suggesting also an optimal clustering at K = 2. Overall, we observed a similar situation at K = 2 as in *T. croaticus* with a break at the eastern part of the Gorski kotar region (dotted line in **Figures 3J,K**), separating eastern and western genetic groups (blue and orange in **Figure 3J**, respectively), but in addition, this species demonstrated a hybrid zone in the middle of its geographic range. At K = 5 the STRUCTURE analysis showed that western- and eastern-most specimens were fully assigned to one of the two groups already found at K = 2, while the samples at the geographic boundary showed some levels of admixture (**Figure 3K**). The PCA plot showed the same pattern with a clear separation between the two main clusters according to axis 1, which is the axis carrying the main proportion of the genetic variation (**Figure 3L**).

For *T. kordunlikanus* the best K values according to the ΔK criterion in STRUCTURE were K = 2 and K = 5

TABLE 2 Descriptive statistics estimated in each population identified by the best ΔK criterion, including the observed heterozygosities (Ho), mean gene diversities within population (Hs), allelic richness (Ar), inbreeding coefficient (FIS) and genetic differentiation (FST).

Species	Population	# Sample	FST	Ho	Hs	Ar	FIS
Parastalita stygia	pop1	14	0.032	0.204	0.212	1.211	0.070
(17,521 SNPs)	pop2	14		0.205	0.286	1.281	0.291
Stalita pretneri	pop1	6	0.030	0.240	0.202	1.206	-0.190
(14,911 SNPs)	pop2	9		0.184	0.265	1.252	0.239
Troglohyphantes croaticus	pop1	8	0.033	0.104	0.152	1.147	0.270
(27,206 SNPs)	pop2	10		0.098	0.188	1.178	0.418
Troglohyphantes excavatus	pop1	16	0.070	0.124	0.167	1.350	0.261
(22,163 SNPs)	pop2	5		0.119	0.171	1.151	0.166
Troglohyphantes kordunlikanus	pop1	9	0.062	0.340	0.303	1.621	-0.095
(13,621 SNPs)	pop2	3		0.290	0.345	1.304	-0.060

TABLE 3 Correlations among pairwise genetic and geographic distances based on Mantel tests and Spearman's rho correlation coefficient. Adjusted *R*-squared (R^2) of the linear model are included and linear equations are indicated in slope-intercept form.

	Mantel statistic *r*	Significance	Spearman *P*-value	Adjusted R^2	Slope	Intercept
Parastalita stygia	0.547	0.001	$< 2.2e^{-16}$	0.2978	$1.53e^{-06}$	$1.77e^{-01}$
Stalita pretneri	0.547	0.001	$7.41e^{-10}$	0.2935	$1.44e^{-06}$	$1.49e^{-01}$
Troglohyphantes croaticus	0.743	0.001	$< 2.2e^{-16}$	0.5476	$9.84e^{-07}$	$1.03e^{-01}$
Troglohyphantes excavatus	0.711	0.001	$<2.2e^{-16}$	0.5023	$7.55e^{-07}$	$8.01e^{-02}$
Troglohyphantes kordunlikanus	0.639	0.015	$1.27e^{-05}$	0.3946	$1.33e^{-06}$	$1.08e^{-01}$

(**Supplementary Table 3**). In conStruct, the analysis of the layer contribution and the distribution of the clusters confirmed the separation into two clusters: the eastern group (orange on **Figure 3M**) was restricted to north-west Bosnia and Herzegovina, and the western group (blue in **Figure 3N**) was distributed across the whole species range in Croatia. This split into two clusters is also clear from the PCA along axis 1 (**Figure 3O**). At $K = 5$, although the statistical support in STRUCTURE was strong (**Supplementary Table 3**), none of the analyses did present any clear spatial pattern—the eastern group remains separated, while the whole western group is composed of admixed individuals from several genetic clusters, without any clear barrier to gene flow.

Our comparative analysis has shown distinctive spatial genetic structuring patterns for each species. At the same time several similarities were revealed: one genetic break at the eastern part of Gorski kotar region is common to four species (both Dysderidae, and *T. croaticus* and *T. excavatus*), and two breaks are shared by the two Dysderidae species, one in northern Lika, and the other in the area of the Ogulin-Plaški valley. In contrast, *T. kordunlikanus* did not share any pattern in genetic structure with the other four species, partly since its distribution only overlaps that of the others across a small area. *Troglohyphantes kordunlikanus* aside, the observed breaks among species have been revealed whatever method used (including the two conStruct variants, with or without embedding a spatial model). These results confirm that the discrete distribution of genetic variation reported in all species is not merely due to isolation by distance processes coupled with non-continuous sampling.

Despite a clear genetic structuring revealed by the population structure analyses, the genetic differentiation (FST) observed among the main populations remained limited (<0.07). Analysis of the different genetic diversity statistics showed a variability across populations and species, e.g., the species *T. kordunlikanus* seems to demonstrate a higher genetic diversity than the other species, as well as a lower inbreeding coefficient (FIS) (**Table 2**). However, these results should be considered with caution because of the small number of samples encompassed in our study. The demographic inference analysis showed a similar pattern among species and populations, i.e., an increase in population size 100,000–200,000 years ago, and then a progressive reduction (**Supplementary Figure 5**).

Isolation by distance

Significant isolation by distance was found in all five species (**Table 3**). Although Dysderidae species (*S. pretneri* and *P. stygia*) showed lower *P*-values than the three *Troglohyphantes* species as computed with a Mantel test, the slopes and intercepts were higher in the latter species (**Table 3**, **Figure 4**). Analyses conducted on the subset of populations not showing complete

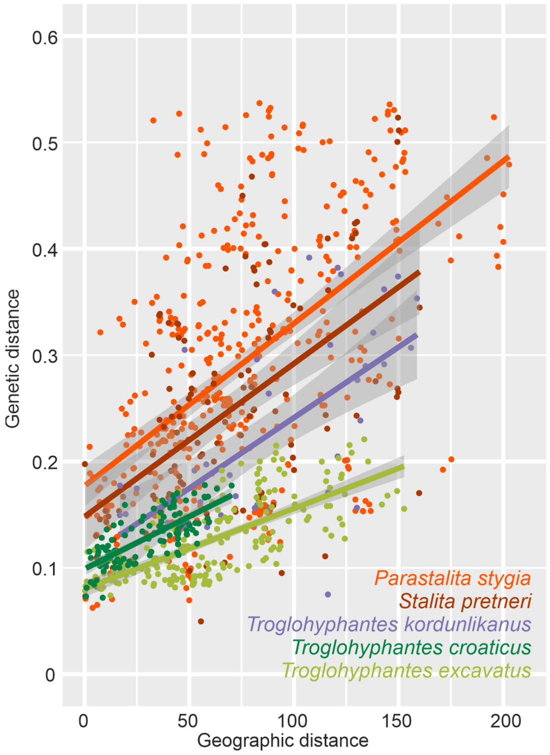


FIGURE 4
Plot showing the overall relationship between pairwise Nei's genetic and linear geographical distances for each of the five studied species. A linear model is represented for each species, with corresponding information, adjusted R-squared (R^2) and linear equations indicated in Table 3.

geographic and genetic isolation, mostly agreed with the whole-population analyses, with few exceptions (Supplementary Figure 6). In the case of the *P. stygia* central group (light blue on Figures 3A,B) the slope and the intercept were both lower than for the southern group (orange on Figures 3A,B). In *T. croaticus* the western group from the K = 2 analysis (orange on Figure 3G) showed a higher slope than the eastern group (blue on Figure 3G), while the western group found in *T. kordunlikanus* (blue on Figure 3M) showed no correlation between genetic and geographic distance (also visible from the structure analyses, Figure 3N). In the case of *T. kordunlikanus*, the sampling also included two easternmost populations (not hosting any of the other species)–interestingly, when these two populations were removed, isolation by distance was found to be non-significant.

Troglomorphy

The *Troglohyphantes* species exhibited different levels of troglomorphic adaptation (Supplementary Table 1B). *Troglohyphantes excavatus* typically exhibited partial pigmentation and normal eyes, and only a few depigmented individuals were found. In the case of *T. kordunlikanus*, all specimens were depigmented, while their eyes were normally developed, and had thick or thin pigment rings around them. Interestingly, *T. croaticus* showed greater diversity, and a clear geographical segregation of troglomorphic traits. All individuals in the Kordun region (eastern part of the species' distribution) were depigmented, most of them with eyes reduced to tiny white spots. Only a few individuals had normally developed eyes with thin black rings around them. In contrast, specimens from the western part (Gorski kotar region) were mostly depigmented, but with normally developed eyes with thin or thick black rings around them. Additionally, few individuals were strongly pigmented and with normally developed eyes, similar to what is found in *T. excavatus*.

Discussion

Museomics as a powerful tool to investigate "rare" species

Our study was made possible thanks to the application of museomics to historical samples. Indeed, as the availability of samples in the field is limited and our target organisms are difficult to sample *in natura*, the progressive accumulation of samples in the CBSS, ZCSL, and ROC natural history collections has made it possible to obtain enough samples to allow the comparative phylogeography inferences produced in this study. This opportunity to exploit tissues or even only DNA molecules from historical samples is associated with recent developments in museomics that allow the extraction and sequencing of degraded and low-concentration DNA from collection samples. Whereas it was previously only possible to recover specific gene sequences (usually short barcodes) through PCR-based approaches (Raxworthy and Smith, 2021), the development of innovative hybridization-capture methods such as HyRAD now allows the recovery of a large number of loci along the genome, and a more detailed investigation in a vast array of evolutionary questions (Suchan et al., 2016). The HyRAD method is based on the construction of probes from a ddRAD library followed by capture of these loci in the historical DNA. This approach allows capturing only the loci of interest and avoids inclusion of possible contaminants linked to historical samples.

Thus, in our study, we integrated 106 samples from the 117 samples initially sampled (91% success). At the scale of our study, the ability to capture DNA in a sample does not appear to be related to its age or to the concentration of DNA initially retrieved (Supplementary Table 2 and Supplementary Figure 7), as previously observed in a study on ground beetles using a similar method from the HyRAD family (Toussaint et al., 2021). Anecdotally, the oldest sample in the analysis collected in 1974 (tcro_3422) yielded enough information to be included in the analysis, which enabled us to detect a labeling mistake. The locality on the label was 200 km from the closest *T. croaticus* locality, but this sample grouped with a sample from the Vrelo cave (number 147) (**Supplementary Figure 2**), which is located in the middle of the species distribution. This, combined with the fact that the collector, the famous Dutch arachnologist Christa Deeleman-Reinhold, visited Vrelo cave just a few days after collecting sample tcro_3422, was sufficient evidence to demonstrate that the label was wrong and that the species distribution is not drastically different than previously known. Furthermore, thanks to the HyRAD approach, we were able to recover 9,212 loci (42% of the target loci), 8,356 loci (36% of the target loci), 11,883 loci (52% of the target loci), 9,656 loci (42% of the target loci), and 7,418 loci (35% of the target loci) for *P. stygia, S. pretneri, T. croaticus, T. excavatus,* and *T. kordunlikanus*, respectively. This is a larger amount of genetic information than what has been reported using UCEs, an alternative approach for retrieving molecular information from historical samples (Derkarabetian et al., 2019). The downside of the HyRAD strategy is that it requires the construction of a specific probe set for each group of closely related taxa, while UCE probes may enrich historical samples across a wider taxonomic range. As a matter of comparison, a recent UCE-based study, also on cave dwelling arachnids, recovered 289 loci (Derkarabetian et al., 2022). Meanwhile, HyRAD does not require any previous knowledge of genome sequences to produce orthologous data across samples. Moreover, HyRAD loci, which are derived from a ddRAD template, are potentially more informative at the intraspecific evolutionary scale, as exemplified by the present study. Whereas Derkarabetian et al. (2022) recovered a total of 1,277 SNPs from UCEs, we identified here a number of genetic polymorphisms larger by an order of magnitude, ranging between 13,621 and 27,206 SNPs, which enabled us to provide fine-scale insights into the population genetic structuring of our target species.

Geography, climate, and topography as possible causes for observed boundaries

In general, the distribution of genetic groups for all five species cannot be explained by current climate or habitat suitability that was modeled for the two Dysderidae species (Pavlek and Mammola, 2021; **Supplementary Figure 8**), and the reasons probably lie in the complex geological and climatic history, both of which are still not sufficiently explored for the area in question, and for the Dinarides in general.

Comparison of the spatial genetic structuring in the five species revealed common phylogeographic patterns except for *T. kordunlikanus*, whose different spatial distribution with only a narrow overlap with that of the four other species makes it difficult to identify common genetic breaks. A common barrier to gene flow has been observed for four of the five species (*P. stygia, S. pretneri, T. croaticus*, and *T. excavatus*) in the eastern part of the Gorski kotar region. One of the hypotheses to explain this barrier resides in differences in climatic conditions on both sides of the barrier. Indeed, the south-western side of this break shows a lower average temperature and higher precipitation levels than the other side (**Supplementary Figure 8**). For two species (*S. pretneri* and *T. croaticus*), no gene flow was identified across this barrier. Such strong genetic isolation between populations could explain morphological differences found on both sides of the genetic boundary for *T. croaticus* (see below). In the case of *T. excavatus*, a species able to use surface habitats for dispersal, some hybrid specimens were observed in this region, indicating the existence of detectable gene flow. Conversely, *P. stygia* seems to be able to disperse southwards across the break line, demonstrating good dispersal abilities also revealed by the large distribution observed for the central *P. stygia* lineage (blue pie charts in **Figures 3A,B**). Two other breaks are shared between the two Dysderidae species: one in the northern Lika region (dashed line in **Figures 3A,B,D,E**) that separates northern and southern clusters in both species, and a second one at the Ogulin-Plaški valley (wavy line in **Figures 3A,B,D,E**). While we were not able to identify the topographic features associated with the northern Lika break line, the second break could be explained by the specific geology of this area. The Ogulin-Plaški valley is a relatively narrow valley made from a less permeable rock (dolomites) than the surrounding area (Velić et al., 1980), resulting in surface streams that spring on one side, and sink on the other side of the valley. Our results might indicate that there is no underground system of crevices, which could be used for migration by the two Dysderidae species studied here, thus effectively isolating populations from the opposing sides of the valley. A similar explanation for the distribution throughout an area of dolomite deposits was invoked to explain the lower dispersal and deeper isolation of a lineage of the troglobiotic beetle *Troglocharinus ferreri* (Reitter, 1908) from Catalonia, when compared to another lineage distributed in a more permeable, and thus more connected limestone rock (Rizzo et al., 2017). The deep geographical structuring found in both Dysderidae species is comparable to that reported in North America cave beetles and harvesters for which geographic distance and landscape features each contributed to the formation of distinct genetic clusters (Kane and Brunner, 1986; Boyd et al., 2020; Derkarabetian et al., 2022). Conversely, no evidence of gene flow was detected among populations of troglobiotic spider and beetle species in caves less than 15 km apart (Balogh et al., 2020).

Links between genetic structure, isolation by distance, species biology, and ecology

Comparison of isolation by distance patterns shows that two factors, among others, could influence the genetic structuring of a given species, namely the level of troglomorphism and the foraging strategy (ground-dwellers vs. web-builders). Both Dysderidae species analyzed here are blind, depigmented, and restricted to cave habitats, while the three studied *Troglohyphantes* species display different levels of eye reduction, depigmentation, and the possibility, for two species at least, to disperse through surface habitats. The former observation would hint at a lower dispersal ability of Dysderidae and hence more structured populations. However, *Troglohyphantes* species build webs on which they spend most of their lives, while Dysderidae are active hunters which move swiftly around the cave in search for prey or mates, which, alternatively, would suggest better chances for dispersal through a well-connected underground system of crevices (Barr, 1967; Culver et al., 2004) for Dysderidae than for *Troglohyphantes*. In order to confront the two hypotheses, we examined IBDs in all five studied species and found that the slope and intercept of the correlation between geographic and genetic distances were always larger in Dysderidae than in *Troglohyphantes*. While the slopes and intercepts for one species of *Troglohyphantes*, i.e., *T. kordunlikanus*, are only marginally below those of the two species of Dysderidae, the pattern seems clearer still when the two easternmost populations of *T. kordunlikanus* (i.e., those that do not overlap with the other species' distributions) are removed, as at this point no significant isolation by distance is retrieved. Altogether, our results suggest that lower levels of cave adaptation are associated with higher dispersal ability, regardless of the hunting strategy (**Table 3** and **Figure 4**). Our study also reinforces the hypothesis that web-building species disperse better (*T. kordunlikanus* and *T. excavatus* even being found outside caves occasionally) and corroborate recent findings in sympatric cave springtails of the Salem Plateau in North America, which revealed stronger correlation between genetic and geographic distance in troglobionts when compared with troglophiles (Katz et al., 2018). Whereas differences in the intercept might reveal other scenarios such as different lineage age or timing of colonization in the studied clades (older species or older cave colonizers would demonstrate steeper slopes in the relationship between geographic and genetic distances), we currently lack data for alternative scenario testing. Different patterns were also unveiled among the *Troglohyphantes* species. Within the *croaticus* group, *T. croaticus*, which is the most troglomorphic species in the *Troglohyphantes* genus, and presumably the most dependent on its stable conditions—it was never collected outside caves—shows a steeper IBD than *T. excavatus*, indicating lower dispersal ability. Interestingly, genetic clusters in *T. croaticus* matched the morphological variation observed in the distribution of troglomorphic traits and in the levels of IBD, which warrants a re-examination of the species status across different populations. Indeed, the eastern group included mostly anophthalmic and depigmented individuals, while in the western group, which is genetically much more heterogeneous (**Figures 3H,I** and **Supplementary Table 1B**), a high degree of polymorphisms in troglomorphic traits was observed. This is also reflected in a higher IDB of the western group compared to the eastern one (**Supplementary Figure 6**). It is worth noting that no admixture was found between these two lineages. Contrastingly, *T. excavatus* showed a shallower IBD slope (**Figure 4**), suggesting higher connectivity among populations, which fits well with the fact that this species shows the lowest levels of troglomorphism among the five studied here. As mentioned above, results obtained for *T. kordunlikanus* are difficult to compare with those of the other species given the fact that its distribution only partly overlaps with that of the two other *Troglohyphantes* species as well as with the two Dysderidae species.

The fact that both Dysderidae species showed more marked IBD patterns was coherent with their blind and depigmented habitus, and the fact that they were never found outside caves—features we hypothesized would cause isolation of distant populations. Interestingly, both Dysderidae species demonstrate very similar values of Mantel and Spearman statistics, and geographical patterning of genetic lineages. Their ecological niche is also very similar, their distribution almost completely overlaps, and they can often be found in sympatry (Pavlek and Mammola, 2021). The only obvious difference is in the cheliceral morphology—*P. stygia* has elongated chelicerae, a trait that has been associated with trophic specialization (oniscophagy, i.e., feeding on woodlice) in other Dysderidae species (e.g., genus *Dysdera*, Rezaè et al., 2008)—which could imply that a segregation in diet may underlie and promote co-existence of the two species in sympatry.

Contrasting life-history traits of the two spider families are, as we predicted, reflected in their spatial population structures. Highly adapted *Parastalita stygia* and *Stalita pretneri* species show higher values of slope and intercept in the isolation by distance regression, suggesting lower dispersal ability, despite of their non-stationary and active way of life in caves. In contrast, lower dependency on cave conditions, as indicated by lower levels of troglomorphy, seems to be associated with higher dispersal rates in *Trologyphantes* species, regardless of their more stationary lifestyle. Spatial distribution of genetic groups revealed some common underlying breaks in the spatial genetic structure, indicating that geographic and climatic features probably influence dispersal potential of whole communities. Lastly, the existence of distinct genetic groups in all five studied species and their distinctive geographical distribution, should be considered when making management and conservation decisions.

Data availability statement

Raw reads are available on the NCBI BioProject PRJNA827597. The whole pipeline including custom scripts are available in the Github repository https://github.com/JeremyLGauthier/PHyRAD.

Author contributions

MP, NA, and MA designed the study. MP performed the sampling. MP and JB performed the lab work. JG and MP analyzed the molecular data, with contributions from VT. All authors took part in discussions concerning the analyses and result interpretations. MP, JG, and NA wrote the manuscript, with contributions from all authors.

Funding

This work was financially supported by a Marie Skłodowska-Curie Individual Fellowship (Grant agreement: 749867) (MP). Additional funding was provided by the Spanish Ministry of Economy and Competitiveness Grant CGL CGL2016-80651-P and the Agency for Management of University and Research Grants of Catalonia (2017SGR83) (both to MA). Open access funding was provided by the University of Geneva.

Acknowledgments

We sincerely thank all members of the Croatian Biospeleological Society, to Slavko Polak, Roman Ozimec, and all other speleologists and colleagues who collected the spider material. We also thank to SubBioLab from the University of Ljubljana for giving access to their collection. Thanks to Alba Enguídanos for her help with DNA isolation, to Tin Rožman for help with adjusting photographs, and also to Mladen Kuhta for help in interpreting the geological setup of the Ogulin-Plaški valley. The analyses were done on the Hercules server at the University of Barcelona and the Pyrgus server hosted by the City of Geneva.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fevo.2022.910084/full#supplementary-material

SUPPLEMENTARY FIGURE 1
For each species, principal component analysis (PCA) plots integrating eigenvalues.

SUPPLEMENTARY FIGURE 2
Phylogenetic trees for **(A)** *Parastalita stygia* and *Stalita pretneri*, **(B)** *Troglohyphantes croaticus*, *Troglohyphantes excavatus*, and *Troglohyphantes kordunlikanus*. Branch support, SH-aLRT support/ultrafast bootstrap support (%), is shown for all branches.

SUPPLEMENTARY FIGURE 3
Population structure of the five studied species as inferred by conStruct. For each species cross-validation results, layer contribution for each *K* value examined and admixture proportions pie charts are shown for spatial and non-spatial models. The colors in the bars show the contribution of each cluster or layer to the total covariance for each *K*. Each pie represents an individual. The color of the pie shows the proportion of the individual's genome that is assigned to each of the *K* layers.

SUPPLEMENTARY FIGURE 4
For each species, spatial genetic structuring for several *K* values (different for each species, based on STRUCTURE statistics) with pie charts corresponding to the Bayesian assignment probabilities per sampling locality exemplified for the best run of each *K* value. Summed results of all runs for each *K* value and for each species are shown as bar plots next to a map with the corresponding best *K* value or below the maps for the rest of *K* values. **(A)** *Parastalita stygia*, **(B)** *Stalita pretneri*, **(C)** *Troglohyphantes croaticus*, **(D)** *Troglohyphantes excavatus*, **(E)** *Troglohyphantes kordunlikanus*.

SUPPLEMENTARY FIGURE 5
Stairway plots for each population, with population one in blue and population two in orange, in each species.

SUPPLEMENTARY FIGURE 6
(A–C) Plots showing the relationship between genetic and simple geographical distance (expressed in km). **(A)** *Parastalita stygia*, plot for all samples together (black line), samples from central and southern group as in the *K* = 3 clustering analyzed together (green line), samples from southern (orange), and central (blue) group as in the *K* = 3 clustering. **(B)** *Troglohyphantes croaticus*, plot for all samples together (black line), samples from western (orange) and eastern (blue) groups as in the *K* = 2 clustering. **(C)** *Troglohyphantes kordunlikanus*, plot for all samples together (black line), and samples from western (blue) group as in the *K* = 2 clustering. **(D)** Correlations among pairwise genetic and geographic distances for groups within *P. stygia*, *T. croaticus*, and *T. kordunlikanus* based on Mantel tests and Spearman's rho correlation coefficient.

SUPPLEMENTARY FIGURE 7
Plots representing the relationship between the age of the specimens and the amount of reads and loci recovered (the samples included in the study are in black and the samples excluded in gray).

SUPPLEMENTARY FIGURE 8
Distribution of each of the five species as a function of climatic factors. **(A–C)** *Parastalita stygia*; **(D–F)** *Stalita preneri*; **(G,J)** *Troglohyphantes croaticus*, **(H,K)** *Troglohyphantes excavatus*, **(I,L)** *Troglohyphantes*

kordunlikanus. **(A)** Records of *P. stygia* overlapped with modeled habitat suitability taken from Pavlek and Mammola (2021) (white patches presenting areas of high suitability). **(D)** Records of *S. pretneri* overlapped with modeled habitat suitability taken from Pavlek and Mammola (2021) (white patches presenting areas of high suitability). **(B,E,G–I)** Species records overlapped with annual temperature values. **(C,F,J–L)** Species records overlapped with annual precipitation values. Climatic variables (annual precipitation and temperature) were taken from the WorldClim website (Fick and Hijmans (2017). Worldclim 2: New 1 km spatial resolution climate surfaces for global land areas. International Journal of Climatology, 37 (12), 4302–4315.).

SUPPLEMENTARY TABLE 1

(A) Samples information. **(B)** Troglomorphism measurements.

SUPPLEMENTARY TABLE 2

Sequencing statistics for each sample: raw reads number, clean reads number, raw number of mapped reads, % of mapped reads, number of mapped reads after cleaning, number of loci, and number of SNPs.

SUPPLEMENTARY TABLE 3

Structure statistics for all five species. **(A)** *Parastalita stygia*, **(B)** *Stalita pretneri*, **(C)** *Troglohyphantes croaticus*, **(D)** *Troglohyphantes excavatus*, **(E)** *Troglohyphantes kordunlikanus*.

References

Balogh, A., Ngo, L., Zigler, K. S., and Dixon, G. (2020). Population genomics in two cave-obligate invertebrates confirms extremely limited dispersal between caves. *Sci. Rep.* 10:17554. doi: 10.1038/s41598-020-74508-9

Barr, T. C. (1967). Observations on the ecology of caves. *Am. Nat.* 101, 475–491.

Barr, T. C. (1968). Cave ecology and the evolution of troglobites. *Evol. Biol.* 2, 35–102.

Barr, T. C., and Holsinger, J. R. (1985). Speciation in cave faunas. *Annu. Rev. Ecol. Syst.* 16, 313–337.

Borowiec, M. L. (2016). AMAS: a fast tool for alignment manipulation and computing of summary statistics. *PeerJ* 4:e1660. doi: 10.7717/peerj.1660

Borregaard, M. K., Amorim, I. R., Borges, P. A. V., Cabral, J. S., Fernández-Palacios, J. M., Field, R., et al. (2017). Oceanic island biogeography through the lens of the general dynamic model: assessment and prospect. *Biol. Rev.* 92, 830–853. doi: 10.1111/brv.12256

Boyd, O. F., Philips, T. K., Johnson, J. R., and Nixon, J. J. (2020). Geographically structured genetic diversity in the cave beetle *Darlingtonea kentuckensis* Valentine, 1952 (Coleoptera, Carabidae, Trechini, Trechina). *Subterr. Biol.* 34, 1–23. doi: 10.3897/subtbiol.34.46348

Bradburd, G. S., Coop, G. M., and Ralph, P. L. (2018). Inferring continuous and discrete population genetic structure across space. *Genetics* 210, 33–52. doi: 10.1534/genetics.118.301333

Bregović, P., and Zagmajster, M. (2016). Understanding hotspots within a global hotspot - identifying the drivers of regional species richness patterns in terrestrial subterranean habitats. *Insect Conserv. Divers.* 9, 268–281. doi: 10.1111/icad.12164

Caccone, A. (1985). Gene flow in cave arthropods: a qualitative and quantitative approach. *Evolution* 39, 1223–1235. doi: 10.1111/j.1558-5646.1985.tb05688.x

Christiansen, K. (2012). "Morphological adaptations," in *Encyclopedia of Caves*, eds W. B. White and D. C. Culver (Oxford: Elsevier Academic Press), 517–528.

Culver, D. C., Christman, M. C., Šereg, I., Trontelj, P., and Sket, B. (2004). The location of terrestrial species-rich caves in a cave-rich area. *Subterr. Biol.* 2, 27–32.

Culver, D. C., Deharveng, L., Bedos, A., Lewis, J. J., Madden, M., Reddell, J. R., et al. (2006). The mid-latitude biodiversity ridge in terrestrial cave fauna. *Ecography* 29, 120–128.

Curl, R. L. (1986). Fractal dimensions and geometries of caves. *Math. Geol.* 18, 765–783. doi: 10.1007/BF00899743

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., et al. (2011). The variant call format and VCFtools. *Bioinformatics* 27, 2156–2158. doi: 10.1093/bioinformatics/btr330

Deeleman-Reinhold, C. L. (1971). Beitrag zur Kenntnis höhlenbewohnender Dysderidae (Araneida) aus Jugoslawien. *Razpr. Slov. Akad. Znan. Umet.* 14, 95–120.

Deeleman-Reinhold, C. L. (1978). Revision of the cave-dwelling and related spiders of the genus Troglohyphantes Joseph (Linyphiidae), with special reference to the *Yugoslav species. Razpr. slov. Akad. Znan. Umet.* 23, 1–220.

Derkarabetian, S., Benavides, L. R., and Gonzalo, G. (2019). Sequence capture phylogenomics of historical ethanol-preserved museum specimens: unlocking the rest of the vault. *Mol. Ecol. Res.* 19, 1531–1544. doi: 10.1111/1755-0998.13072

Derkarabetian, S., Paquin, P., Reddell, J., and Hedin, M. (2022). Conservation genomics of federally endangered *Texella harvester* species (Arachnida, Opiliones, Phalangodidae) from cave and karst habitats of central Texas. *Conserv. Genet.* 23, 401–416. doi: 10.1007/s10592-022-01427-9

Earl, D. A., and VonHoldt, B. M. (2012). Structure harvester: a website and program for visualizing structure output and implementing the Evanno method. *Conserv. Genet. Resour.* 4, 359–361. doi: 10.1007/s12686-011-9548-7

Eaton, D. A. R., and Overcast, I. (2020). ipyrad: interactive assembly and analysis of RADseq datasets. *Bioinformatics* 36, 2592–2594. doi: 10.1093/bioinformatics/btz966

Eme, D., Malard, F., Konecny-Dupré, L., Lefébure, T., and Douady, C. J. (2013). Bayesian phylogeographic inferences reveal contrasting colonization dynamics among European groundwater isopods. *Mol. Ecol.* 22, 5685–5699. doi: 10.1111/mec.12520

Evanno, G., Regnaut, S., and Goudet, J. (2005). Detecting the number of clusters of individuals using the software structure: a simulation study. *Mol. Ecol.* 14, 2611–2620. doi: 10.1111/j.1365-294X.2005.02553.x

Fage, L. (1919). Etudes sur les araignées cavernicoles. III. Sur le genre Troglohyphantes. Biospelogica XL. *Arch. Zool. Exp. Gén.* 55, 55–148.

Fick, S. E., and Hijmans, R. J. (2017). Worldclim 2: new 1-km spatial resolution climate surfaces for global land areas. *Int. J. Climatol.* 37, 4302–4315.

Garrison, E., and Marth, G. (2012). Haplotype-based variant detection from short-read sequencing. *arXiv* [Preprint]. doi: 10.48550/arXiv.1207.3907

Gauthier, J., Pajkovic, M., Neuenschwander, S., Kaila, L., Schmid, S., Orlando, L., et al. (2020). Museomics identifies genetic erosion in two butterfly species across the 20th century in Finland. *Mol. Ecol. Resour.* 20, 1191–1205. doi: 10.1111/1755-0998.13167

Goudet, J. (2005). Hierfstat, a package for R to compute and test hierarchical F-statistics. *Mol. Ecol. Notes* 5, 184–186. doi: 10.1111/j.1471-8286.2004.00828.x

Hedin, M. (1997). Molecular phylogenetics at the population/species interface in cave spiders of the southern *Appalachians* (Araneae: Nesticidae: Nesticus). *Soc. Mol. Biol. Evol.* 14, 309–324. doi: 10.1093/oxfordjournals.molbev.a025766

Howarth, F. G. (1987). Evolutionary ecology of aeolian and subterranean habitats in Hawaii. *Tree* 2, 220–223. doi: 10.1016/0169-5347(87)90025-5

Isaia, M., Mammola, S., Mazzuca, P., Arnedo, M. A., and Pantini, P. (2017). Advances in the systematics of the spider genus *Troglohyphantes* (Araneae, Linyphiidae). *Syst. Biodivers.* 15, 307–326. doi: 10.1080/14772000.2016.1254304

Jakobsson, M., and Rosenberg, N. A. (2007). CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. *Bioinformatics* 23, 1801–1806. doi: 10.1093/bioinformatics/btm233

Jalžić, B., Bedek, J., Bilandžija, H., Bregović, P., Cvitanović, H., Èuković, T., et al. (2013). *The Cave Type Localities Atlas of Croatian Fauna, volume 2*. Zagreb: CBSS.

Jombart, T., and Ahmed, I. (2011). adegenet 1.3-1: new tools for the analysis of genome-wide SNP data. *Bioinformatics* 27, 3070–3071. doi: 10.1093/bioinformatics/btr521

Jónsson, H., Ginolhac, A., Schubert, M., Johnson, P. L. F., and Orlando, L. (2013). mapDamage2.0: fast approximate Bayesian estimates of ancient DNA damage parameters. *Bioinformatics* 29, 1682–1684. doi: 10.1093/bioinformatics/btt193

Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., Von Haeseler, A., and Jermiin, L. S. (2017). ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat. Methods* 14, 587–589. doi: 10.1038/nmeth.4285

Kane, T. C., and Brunner, G. D. (1986). Geographic variation in the cave beetle *Neaphaenops tellkampfi* (Coleoptera: Carabidae). *Psyche* 93, 231–251. doi: 10.1155/1986/86164

Katoh, K., Misawa, K., Kuma, K., and Miyata, T. (2002). MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* 30, 3059–3066. doi: 10.1093/nar/gkf436

Katz, A. D., Taylor, S. J., and Davis, M. A. (2018). At the confluence of vicariance and dispersal: phylogeography of cavernicolous springtails (Collembola: Arrhopalitidae, Tomoceridae) codistributed across a geologically complex karst landscape in Illinois and Missouri. *Ecol. Evol.* 8, 10306–10325. doi: 10.1002/ece3.4507

Keightley, D. P., Ness, R. W., Halligan, D. L., and Haddrill, P. R. (2014). Estimation of the spontaneous mutation rate per nucleotide site in a *Drosophila melanogaster* Full-Sib family. *Genetics* 196, 313–320. doi: 10.1534/genetics.113.158758

Lanfear, R., Frandsen, P. B., Wright, A. M., Senfeld, T., and Calcott, B. (2017). Partitionfinder 2: new methods for selecting partitioned models of evolution for molecular and morphological phylogenetic analyses. *Mol. Biol. Evol.* 34, 772–773. doi: 10.1093/molbev/msw260

Lefébure, T., Douady, C. J., Gouy, M., Trontelj, P., Briolay, J., and Gibert, J. (2006). Phylogeography of a subterranean amphipod reveals cryptic diversity and dynamic evolution in extreme environments. *Mol. Ecol.* 15, 1797–1806. doi: 10.1111/j.1365-294X.2006.02888.x

Li, C., Hofreiter, M., Straube, N., Corrigan, S., and Naylor, G. J. P. (2013). Capturing protein-coding genes across highly divergent species. *Biotechniques* 54, 321–326. doi: 10.2144/000114039

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25, 1754–1760. doi: 10.1093/bioinformatics/btp324

Liu, X., and Fu, Y.-X. (2020). Stairway Plot 2: demographic history inference with folded SNP frequency spectra. *Genome Biol.* 21:280. doi: 10.1186/s13059-020-02196-9

Mammola, S., and Isaia, M. (2017). Spiders in caves. *Proc. R. Soc. B Biol. Sci.* 284:20170193. doi: 10.1098/rspb.2017.0193

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet J.* 17:10. doi: 10.14806/ej.17.1.200

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., et al. (2010). The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* 20, 1297–1303. doi: 10.1101/gr.107524.110

Mihevc, A., Prelošek, M., and Zupan Hajna, N. (2010). *Introduction to the Dinaric Karst*. Postojna: Karst Research Institute and Research Centre of the Slovenian Academy of Sciences and Arts.

Minh, B. Q., Schmidt, H. A., Chernomor, O., Schrempf, D., Woodhams, M. D., von Haeseler, A., et al. (2020). IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. *Mol. Biol. Evol.* 37, 1530–1534. doi: 10.1093/molbev/msaa015

Morelli, T. L., Smith, A. B., Mancini, A. N., Balko, E. A., Borgerson, C., Dolch, R., et al. (2020). The fate of Madagascar's rainforest habitat. *Nat. Clim. Change* 10, 89–96. doi: 10.1038/s41558-019-0647-x

Oksanen, J., Blanchet, F. G., Friendly, M., Kindt, R., Legendre, P., McGlinn, D., et al. (2020). *vegan: Community Ecology Package. R Packag. version 2.5-6.*

Pavlek, M., and Mammola, S. (2021). Niche-based processes explaining the distributions of closely related subterranean spiders. *J. Biogeogr.* 48, 118–133. doi: 10.1111/jbi.13987

Pritchard, J. K., Stephens, M., and Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics* 155, 945–959. doi: 10.1093/genetics/155.2.945

Raxworthy, C. J., and Smith, B. T. (2021). Mining museums for historical DNA: advances and challenges in museomics. *Trends Ecol. Evol.* 36, 1049–1060. doi: 10.1016/j.tree.2021.07.009

Rezaè, M., Pekar, S., and Lubin, Y. (2008). How oniscophagous spiders overcome woodlouse armour. *J. Zool.* 275, 64–71.

Ribera, I., Cieslak, A., Faille, A., and Fresneda, J. (2018). "Historical and ecological factors determining cave diversity," in *Cave Ecology*, eds O. T. Moldovan, L. Kováè, and S. Halse (Berlin: Springer), 229–252. doi: 10.1007/978-3-319-98852-8_10

Rizzo, V., Sánchez-Fernández, D., Alonso, R., Pastor, J., and Ribera, I. (2017). Substratum karstificability, dispersal and genetic structure in a strictly subterranean beetle. *J. Biogeogr.* 44, 2527–2538. doi: 10.1111/jbi.13074

Rosenberg, N. A. (2003). distruct: a program for the graphical display of population structure. *Mol. Ecol. Notes* 4, 137–138. doi: 10.1046/j.1471-8286.2003.00566.x

Salces-Castellano, A., Patiño, J., Alvarez, N., Andújar, C., Arribas, P., Braojos-Ruiz, J. J., et al. (2020). Climate drives community-wide divergence within species over a limited spatial scale: evidence from an oceanic island. *Ecol. Lett.* 23, 305–315. doi: 10.1111/ele.13433

Sbordoni, V., Allegrucci, G., and Cesaroni, D. (2000). Population genetic structure, speciation and evolutionary rates in cave-dwelling organisms. *Subterr. Ecosyst.* 24, 459–483. doi: 10.1093/jhered/esv078

Sket, B. (2012). "Diversity patterns in the Dinaric Karst," in *Encyclopedia of Caves*, eds W. B. White and D. C. Culver (Oxford: Elsevier Academic Press), 228–238.

Sket, B., Paragamian, K. K., and Trontelj, P. (2004). A census of the obligate subterranean fauna of the Balkan Peninsula. *Balk. Biodivers. Pattern Process Eur. Hotspot* 1540, 309–322. doi: 10.1007/978-1-4020-2854-0_18

Suchan, T., Kusliy, M. A., Khan, N., Chauvey, L., Tonasso-Calvière, L., Schiavinato, S., et al. (2022). Performance and automation of ancient DNA capture with RNA hyRAD probes. *Mol. Ecol. Resour.* 22, 891–907. doi: 10.1111/1755-0998.13518

Suchan, T., Pitteloud, C., Gerasimova, N. S., Kostikova, A., Schmid, S., Arrigo, N., et al. (2016). Hybridization capture using RAD probes (hyRAD), a new tool for performing genomic analyses on collection specimens. *PLoS One* 11:e0151651. doi: 10.1371/journal.pone.0151651

Thaler, K. (1986). Über einige Funde von Troglohyphantes-Arten in Kärnten (Österreich) (Arachnida, Aranei: Linyphiidae). *Carinthia II* 176, 287–302.

Toussaint, E. F. A., Gauthier, J., Bilat, J., Gillett, C. P. D. T., Gough, H. M., Lundkvist, H., et al. (2021). HyRAD-X exome capture museomics unravels giant ground beetle evolution. *Genome Biol. Evol.* 13, 1–18. doi: 10.1093/gbe/evab112

Trontelj, P. (2018). "Structure and genetics of cave populations," in *Cave Ecology*, eds O. T. Moldovan, L. Kováè, and S. Halse (Berlin: Springer), 269–292.

Velić, I., Sokaè, B., and Šćavnièar, B. (1980). Tumaè za list Ogulin. *Osnovna Geološka Karta SFRJ* 1:100.

Zagmajster, M., Culver, D. C., Christman, M. C., and Sket, B. (2010). Evaluating the sampling bias in pattern of subterranean species richness: combining approaches. *Biodivers. Conserv.* 19, 3035–3048. doi: 10.1007/s10531-010-9873-2

Zimmerman, E. C. (1949). *Insects of Hawaii*. Honolulu: University of Hawai'i Press.

Zupan Hajna, N. (2019). "Dinaric karst—Geography and geology," in *Encyclopedia of Caves*, eds C. C. David and W. B. White (Cambridge, MA: Academic Press), 353–362.

frontiers | Frontiers in Ecology and Evolution

TYPE Systematic Review
PUBLISHED 24 October 2022
DOI 10.3389/fevo.2022.966605

Cross-sectional use of barcode of life data system and GenBank as DNA barcoding databases for the advancement of museomics

Takeru Nakazato[1]* and Utsugi Jinbo[2]

[1]Database Center for Life Science (DBCLS), Joint Support-Center for Data Science Research (ROIS-DS), Research Organization of Information and Systems (ROIS), Mishima, Japan, [2]Center for Collections, National Museum of Nature and Science, Tsukuba, Japan

OPEN ACCESS

EDITED BY
Jonathan J. Fong,
Lingnan University, Hong Kong SAR, China

REVIEWED BY
Marta Álvarez-Presas,
University of Bristol, United Kingdom
Solange Sotuyo,
Department of Botany, Institute of Biology, National Autonomous University of Mexico, Mexico

*CORRESPONDENCE
Takeru Nakazato
nakazato@dbcls.rois.ac.jp

SPECIALTY SECTION
This article was submitted to Phylogenetics, Phylogenomics, and Systematics, a section of the journal Frontiers in Ecology and Evolution

RECEIVED 11 June 2022
ACCEPTED 26 September 2022
PUBLISHED 24 October 2022

CITATION
Nakazato T and Jinbo U (2022) Cross-sectional use of barcode of life data system and GenBank as DNA barcoding databases for the advancement of museomics. *Front. Ecol. Evol.* 10:966605. doi: 10.3389/fevo.2022.966605

Museomics is an approach to the DNA sequencing of museum specimens that can generate both biodiversity and sequence information. In this study, we surveyed both the biodiversity information-based database BOLD (Barcode of Life System) and the sequence information database GenBank, by using DNA barcoding data as an example, with the aim of integrating the data from these two databases. DNA barcoding is a method of identifying species from DNA sequences by using short genetic markers. We surveyed how many entries had biodiversity information (such as links to BOLD and specimen IDs) by downloading all fish, insect, and flowering plant data available from the GenBank Nucleotide, and BOLD ID was assigned to 26.2% of entries for insects. In the same way, we downloaded the respective BOLD data and checked the status of links to sequence information. We also investigated how many species do these databases cover, and 7,693 species were found to exist only in BOLD. In the future, as museomics develops as a field, the targeted sequences will be extended not only to DNA barcodes, but also to mitochondrial genomes, other genes, and genome sequences. Consequently, the value of the sequence data will increase. In addition, various species will be sequenced and, thus, biodiversity information such as the evidence specimen photographs used as a basis for species identification, will become even more indispensable. This study contributes to the acceleration of museomics-associated research by using databases in a cross-sectional manner.

KEYWORDS

sequencing data, voucher specimen, DNA barcode, biodiversity information, taxonomy database

Introduction

Museomics is, in very simple terms, an approach to DNA sequencing on museum specimens (Raxworthy and Smith, 2021). Museomics research generates both biodiversity and sequence information. Therefore, it is necessary to use these two data in an integrated manner. In this study, we surveyed both BOLD (Barcode of Life System; a biodiversity-based database)[1] (Ratnasingham and Hebert, 2007) and GenBank[2] (a sequence information database) (Sayers et al., 2022b), by using DNA barcoding data as an example, and attempted to merge the data obtained from these two databases.

DNA barcoding technology has been used in order to identify species from DNA sequences as short genetic markers (Hebert et al., 2003). The most commonly used barcode region for animals is a portion of the cytochrome c oxidase I (COI or COX1) gene, found in mitochondrial DNA. Other genes suitable for DNA barcoding are the internal transcribed spacer (ITS) rRNA (often used for fungi) and RuBisCO (used for plants). In addition, the development of massively parallel sequencing technology, also called "next-generation sequencing technology" (NGS), has also made it possible to comprehensively identify the biological flora in the observed environment (Buerki and Baker, 2016; Miya, 2022). Metagenome analysis is a technique used for profiling 16S rRNA and detecting functional genes by sequencing environmental samples on a large scale, without isolating or culturing the microorganisms contained in the samples. For animals, plants, and fungi, a method of large-scale detection of DNA barcodes with NGS can also be used in the form of metabarcoding, by combining DNA barcoding and NGS (Adamowicz et al., 2019; DeSalle and Goldstein, 2019). DNA barcoding technology interests not only biodiversity researchers such as taxonomists and phylogeneticists, but also molecular biologists and bioinformaticians involved in the performance of metagenomics.

DNA barcoding requires a database for querying sequences of DNA barcodes as genetic markers, and the species information identified by the DNA barcode (or the specimen information required in order to identify the species). BOLD (see text footnote 1) is a popular database of DNA barcodes for animals and plants (Ratnasingham and Hebert, 2007), and so is UNITE[3] for fungi (Nilsson et al., 2019). DNA barcodes also include DNA sequence aspects; thus, DNA barcodes have also been deposited in the NCBI (National Center for Biotechnology Information, US) GenBank Nucleotide (see text footnote 2); a database of nucleotide sequences. BOLD and GenBank Nucleotide collect DNA barcode data separately, and import the data from each other. However, the contents are different due to the difference in their backgrounds.

BOLD is an informatics workbench aiding the acquisition, storage, analysis, and publication of DNA barcode records; it was launched in 2005 (Ratnasingham and Hebert, 2007). BOLD provides about 11 million barcodes, thereby indexing 239,000 animals, 71,000 plants, and 24,000 fungi and other species as of May 2022. BOLD requires data with the following seven elements in order for them to qualify as a specimen record with a formal DNA barcode status: (i) species name, (ii) voucher data (catalog number and institution storing), (iii) collection record (collector, collection date, and location with GPS coordinates), (iv) identifier of the specimen, (v) barcode sequence, (vi) PCR primers used in order to generate the amplicon, and (vii) trace files (Ratnasingham and Hebert, 2007). BOLD has been widely used, especially by taxonomists and phylogeneticists, for the referencing of biodiversity information assigned to DNA barcoding due to the large archive of photographic data of evidence specimens and the richness of information on specimens enabling the user to identify or to review for identification.

DNA sequences have been collected for more than 30 years by International Nucleotide Sequence Database Collaboration (INSDC)[4] (Arita et al., 2021), that consists of NCBI[5], the European Bioinformatics Institute (EBI)[6], and the DNA Data Bank of Japan (DDBJ)[7], and are provided as databases in the NCBI GenBank (Sayers et al., 2022b), the European Nucleotide Archive (ENA)[8], and DDBJ[9], respectively. In recent years, DNA barcodes, mitochondrial genomes, whole genomes, and other gene sequences have been obtained for various organisms, and sequence information has been archived in these databases. In addition, NGS data (including metagenomics and metabarcoding) are also collected by INSDC in the form of a Sequence Read Archive (SRA)[10] (Sayers et al., 2022a). Molecular biologists and bioinformaticians usually perform their research from DNA sequence aspects, and make extensive use of the NCBI services dealing with DNA sequences.

Recently, it has become possible to register occurrence information based on sequences such as environmental DNA (eDNA) in the Global Biodiversity Information Facility (GBIF)[11]; the major database of biodiversity information (Andersson et al., 2020). In addition, GenBank is now also able to record much biodiversity information. In this study, we focus on DNA barcode data as an actual use scene of museomics,

[1] https://www.boldsystems.org/

[2] https://www.ncbi.nlm.nih.gov/nuccore

[3] https://unite.ut.ee/

[4] https://www.insdc.org/

[5] https://ncbi.nlm.nih.gov/

[6] https://www.ebi.ac.uk/

[7] https://www.ddbj.nig.ac.jp/

[8] https://www.ebi.ac.uk/ena/browser/

[9] https://www.ddbj.nig.ac.jp/ddbj/

[10] https://ncbi.nlm.nih.gov/sra

[11] https://www.gbif.org/

and point out the necessity of integrated use of BOLD and GenBank and the associated problems. We also propose that GenBank will become a useful resource for species identification by gene sequences other than the current DNA barcode region in the future. We believe that our work will accelerate future life science- and museomics-associated research employing biodiversity and sequence data (Groom et al., 2021).

Methods

Downloading GenBank data from national center for biotechnology information

We obtained all data on the base sequence of fish from NCBI. NCBI GenBank provides data on other vertebrates except mammals in the form of VRT divisions (for reference, they are distributed as HUM for humans, ROD for rodents, and MAM for mammals). We downloaded all VRT division data (gbvrt###.seq.gz, ### = 1–277) from the GenBank FTP site[12] (as of December 2021). The files are distributed in FASTA format (**Supplementary Figure 1**). Subsequently, we extracted the entry containing "Actinopterygii" in the taxonomy hierarchy from the downloaded files, and created the entire data of the base sequence of fish.

As with fish, we downloaded the invertebrate data file provided as an INV division (gbinv###.seq.gz, ### = 1–461), as well as the plant and fungi data file distributed as a PLN division (gbpln###.seq.gz, ### = 1–723) from the NCBI FTP site. Subsequently, we extracted only the entries containing "Insecta" and "Magnoliopsida" in taxonomy tree from the downloaded data, respectively, and used them as insect and flowering plants data for the subsequent analyses.

Data extraction from GenBank

Data submitters can label their sequence as DNA barcoding data by describing "BARCODE" in the KEYWORD field of the GenBank entry (**Supplementary Figure 1**). We counted the number of data containing this description.

Moreover, the BOLD ID is listed in the db_xref qualifier in the "Features" field as the ID of the external database (**Supplementary Figure 1**). We extracted such BOLD IDs from the downloaded GenBank files. The BOLD ID is written after the description of "BOLD." We looked at the number of BOLD IDs mentioned in GenBank and compared them with the BOLD data. GenBank provides qualifiers in order to record biodiversity information for the registration of gene sequences derived from specimens: voucher_specimen, lat_lon (latitude and longitude), altitude, collection_date, collected_by, identified_by, and country (**Supplementary Figure 1**). We surveyed how many entries were given these qualifiers related to biodiversity information. Finally, we especially extracted the specimen IDs listed in the specimen_voucher qualifier of the "Features" field.

Downloading barcode of life system data

We downloaded the public data of the DNA barcode of fish (Animals; Chordata; Actinopterygii)[13] from the BOLD database. Herein, we downloaded the combined data in a tab-delimited format, containing both specimen and sequence data.

As with fish, we downloaded the data of flowering plants (Plants; Magnoliophyta; Magnoliopsida, see text footnote 13).

In addition, we attempted to obtain data on insects (Animals; Arthropoda; Insecta, see text footnote 13). However, BOLD's web pages and APIs are so slow to respond, and the insect data are so extensive that it often seemed that the process had finished before all the data were downloaded. Therefore, we downloaded the specimen and sequence data separately instead of downloading them in the form of combined data. We, herein, attempted to download the data twice, and after confirming that the same data were obtained, the subsequent analysis was performed.

Data extraction from barcode of life system

From the downloaded BOLD data, we extracted BOLD IDs (Specimen ID, Sequence ID), data sources, taxonomic classifications (such as species_name and genus), linked GenBank IDs, and gene names. Especially, BOLD has imported data from GenBank and has labeled them as "Mined from GenBank, NCBI" in the institution_storing field.

Comparison of referring status to each other's IDs for barcode of life system and GenBank

We created pairs of GenBank Accession numbers and BOLD IDs described in the db_xref qualifier from the GenBank data on fish. We also created pairs of BOLD sequence IDs and referring GenBank Accession numbers from the BOLD data. We then

[12] http://ftp.ncbi.nlm.nih.gov/genbank/

[13] https://www.boldsystems.org/index.php/Taxbrowser_Taxonpage?taxid=77

compared these two groups of pairs in order to investigate whether the GenBank and the BOLD data refer to each other (Figure 2). In BOLD, the barcode sequences of multiple different genes obtained from one specimen are often registered. In order to distinguish these, the "specimen ID.gene name" style was used as the ID of BOLD (e.g., BCF519-07.COI-5P), but some GenBank entries refer to BOLD by only the specimen ID. We, therefore, extracted the gene names in addition to the BOLD IDs from GenBank, and restored the "specimen ID.gene name" style ID.

Comparison of biological classifications between barcode of life system and GenBank

National center for biotechnology information GenBank uses NCBI Taxonomy as Taxonomy data, and BOLD seems to be based on the GBIF Backbone Taxonomy. We downloaded both these data. We downloaded the new_taxdump.tar.gz file from the FTP site as NCBI Taxonomy data. We used names.dmp, rankedlineage.dmp, and nodes.dmp files among the uncompressed files, and extracted the scientific name, the taxonomy ID, and the taxonomy tree information. We also downloaded the GBIF Backbone Taxonomy from the GBIF website (GBIF Secretariat, 2021). The file is distributed in the form of a tab-delimited format, and we used TaxonID, scientific name, and taxonomy classification information.

Subsequently, we compared the biological taxonomy information described in BOLD and GenBank, and the identified level of classification (such as species, genus, and class). The description written as a species name may include sp. (species: no valid published scientific description or lack of information), aff. (affinis: the identity of a distinct biological species is unknown, but it has a striking similarity or close relation with a known species), or cf. (confer: the specimen resembles the named species very closely, but has certain minor features not found on the type specimens). Since these have not been identified as a species level, we excluded species names containing these suffixes, and treated such data as species level names.

We also surveyed how much of those data accounted for in the taxonomy database. As a biological taxonomy database, GenBank uses NCBI Taxonomy, and BOLD uses a GBIF Backbone Taxonomy-based classification.

Extraction of new DNA barcode candidates from the GenBank data

There are many entries in GenBank that do not have a BOLD ID, but have a sample ID in voucher_specimen. We regarded these sequences as new candidates for DNA barcodes, and extracted these data. We extracted data from voucher_specimen, but without the BOLD ID from db_xref from the sequence data of all fish and flowering plants previously created from the GenBank Nucleotide. In the GenBank data, the gene name is written in the gene qualifier in the "Features" field (Supplementary Figure 1). We have summarized the generated DNA barcode candidate gene data by gene name. Since the described gene name could be freely described by the submitters, there were cases where the same gene had a different description (e.g., COI, COX1, and CO1). Text mining technology can solve this problem, but this time we have simply listed the genes described without it. In addition, species names were extracted from these candidate data, and were compared with the list of species covered by existing DNA barcode data.

Results

DNA barcode data in GenBank

GenBank Nucleotide is originally a database of DNA sequences, and DNA barcoding data are also registered in GenBank as they are nucleotide sequences. DNA barcode data are increasingly being used in order to monitor fish as "environmental DNA" (Miya, 2022). In this study, we obtained all GenBank Nucleotide data for fish and extracted the DNA barcoding data for trend analysis and comparison with those of BOLD. In addition, a large amount of DNA barcode data has been accumulated for insects. On the other hand, BOLD collects not only animal data, but plant data as well. Thus, similar analyses were also performed for insects and flowering plants.

All GenBank Nucleotide data used in this study consisted of 1,272,272 entries for fish, 7,010,856 entries for insects, and 1,356,592 entries for flowering plants. There is a way to write the "BARCODE" description in the KEYWORD section so as to indicate that the entry refers to DNA barcoding data in GenBank (Supplementary Figure 1). We extracted this description from fish, insect, and flowering plant data, and found that it was present in 50,373 (4.0%), 768,010 (11.0%), and 17,377 (0.8%) of the entries, respectively (Figure 1B). In addition, there are entries in GenBank that provide a more direct link to BOLD data. The BOLD ID can be found in the db_xref qualifier in the Features field of GenBank (Supplementary Figure 1). We surveyed how many GenBank entries referred to BOLD IDs: 90,927 (7.1%) for fish, 1,836,440 (26.2%) for insects, and 10,249 (0.8%) for flowering plants (Figure 1). The most major data registration source was iBOL (International Barcode of Life): 9,070 entries (10.0% of entries with BOLD ID) for fish, 283,215 entries (15.4%) for insects, and 485 entries (4.7%) for flowering plants.

In addition to the nucleotide sequence, the specimen information as the basis for identification is essential for DNA barcoding data. GenBank has several qualifiers for describing

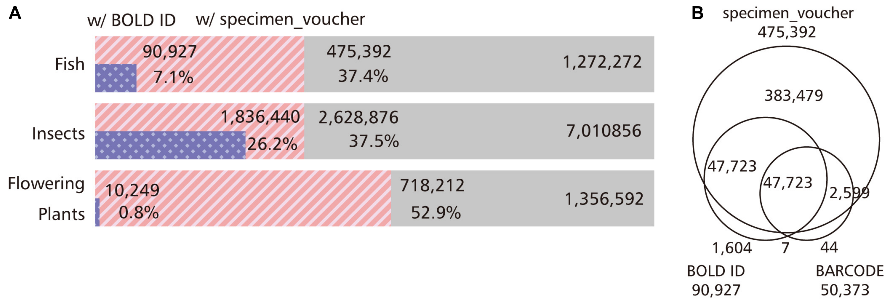


FIGURE 1
Status of entries with information related to the DNA barcoding in GenBank. By using GenBank, we extracted the BARCODE from the KEYWORD section, the BOLD ID referenced as the ID of the external database in the db_xref qualifier, and the sample ID written in the specimen_voucher qualifier as information related to DNA barcoding. **(A)** Percentage of entries with BOLD ID and the specimen_voucher qualifier. We examined the proportion of entries with BOLD ID and the specimen_voucher information in fish, insects, and flowering plants. In insects, a quarter of the entries correspond to barcode sequences with links to BOLD, and rich barcode information can be obtained from GenBank. Plants, on the other hand, have poor links to BOLD, but half of the entries are assigned specimen IDs, and DNA barcode candidates may be hidden in these entries. **(B)** Venn diagram of entries with BARCODE keyword, BOLD ID, and specimen_voucher qualifier. We examined the overlap of entries with the BARCODE keyword, BOLD ID, and sample ID in the GenBank fish data. In order to extract the entry corresponding to the DNA barcode, not only the BARCODE in the KEYWORD section must be extracted, but also the entry with the BOLD ID as the external database ID.

biodiversity information such as altitude, collection_date, and country (**Supplementary Figure 1**). GenBank has a specimen_voucher qualifier for entering the sample ID, and if a re-identification is required, it is theoretically possible to trace the sample information based on this qualifier. We examined the number of entries with specimen_voucher information in GenBank. We found 475,392 (37.4%) entries for fish, 2,628,876 (37.5%) entries for insects, and 718,212 (52.9%) entries for flowering plants (**Figure 1A**). These are more than the entries identified with the use of the "BARCODE" description in the KEYWORD section (**Figure 1B**).

Link to GenBank in barcode of life system data

We obtained Public Data from the BOLD website and examined the links to GenBank for fish, insects, and flowering plants. The total number of specimens was 274,717 for fish, 7,122,873 for insects, and 258,436 for flowering plants.

We counted the data imported from GenBank by checking the description of those "Mined from GenBank, NCBI" in the institution_storing field, and we identified 138,050 (50.3%) fish, 542,035 (7.6%) insects, and 180,146 (69.7%) flowering plants indexed for such data.

Of the fish data registered in BOLD, GenBank IDs were assigned to 234,491 sequences in 213,088 specimens. These correspond to 215,806 GenBank entries.

The number is reduced because multiple sequence entries from the same specimen (e.g., GBMTG999-16.COI-5P, GBMTG999-16.ND5-0, and GBMTG999-16.CYTB) refer to the same GenBank entry (e.g., NC_008679: Schistura balteata mitochondrion, complete genome).

Of the 234,491 sequences with GenBank IDs, 4,330 GenBank entries were in the "suppressed state" (e.g., HM379807). NCBI labels the data as a "suppressed state" in cases where there is doubt or inadequate registration. We were not able to find corresponding data in the "suppressed state" by keyword search, but we were able to see them by specifying the Accession ID. In this case, these data did not qualify for the iBOL/GenBank early release agreement due to the lack of tentative taxonomic identifications (National Library of Medicine [NLM], 2009). In addition, some GenBank entries have been assigned with the "WITHDRAWN" label.

Mismatch between barcode of life system data and GenBank data

We compared these two groups of pairs in order to investigate whether the GenBank and the BOLD data refer to each other by creating GenBank-BOLD ID pairs from GenBank and BOLD data, respectively.

We present the obtained referring status in **Figure 2**. Of the 234,491 ID pairs with GenBank Accession linked from BOLD, 80,878 GenBank entries (34.5%) contained a description of the BOLD ID. Of these, 71,314 pairs referred to each other for the same ID in both GenBank and BOLD (**Figure 2A**). The 9,564 pairs had different BOLD IDs that refer to GenBank, and BOLD IDs that refer to GenBank. Herein, we found that IDs for specimens (e.g., FOA941-05) and IDs for barcode sequences (e.g., FOA941-05.COI-5P) coexist in the writing style of GenBank. For fish, 9,162 specimens were written in the former style and 81,676 specimens were written in the latter. Of 9,567 pairs, 8,720 BOLD IDs in GenBank were written in a BOLD Specimen ID format, so the IDs were actually the same. It can be said that they actually refer to each other (**Figure 2B**).

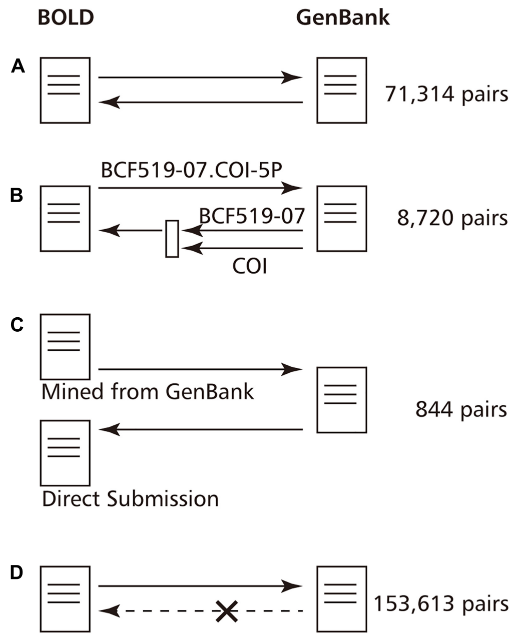


FIGURE 2
Reference status of the BOLD entry and the GenBank entry to each other. Since GenBank and BOLD import data from each other, their IDs are often mentioned in their entries. However, the references may not be reciprocal due to ambiguity in the description or duplicate registrations. **(A)** ID references are reciprocal. **(B)** GenBank refers to the ID of the specimen from which the DNA barcode is derived, but the BOLD Barcode ID can be recovered by extracting the gene name as well. **(C)** The case where the ID of a direct submission and the import data exist because the data were registered in both GenBank and BOLD. In past examples, these will be unified later. **(D)** BOLD ID is not described in GenBank.

For the 844 pairs, the BOLD ID of the reference source and the BOLD ID of the reference target are completely different. For example, BOLD: ANGBF29940-19.COI-5P refers to GenBank: KY570698, but GenBank: KY570698 refers to GAMBA659-12.COI-5P. This seems to be a case where both direct submission data and imported data from GenBank exist in BOLD because the researchers submitted the same data in both the BOLD and the GenBank databases (Figure 2C). In past cases, these duplications have been resolved, and it is assumed that BOLD is taking some action regarding this issue.

Species covered by barcode of life system and GenBank

We investigated the number of species covered by BOLD data in the case of fish. There were 274,717 sample data entries in total, of which 238,633 entries (86.9%) had data as species_name. This number corresponds to 20,660 types of descriptions, but it has not been identified down to the species level, and contains an entry with the genus name followed by sp./aff./cf. By excluding these, 224,742 entries (81.8%) corresponding to 15,882 species of fish were identified down to the species level (Figure 3). Similarly, there were 251,540 (91.6%) entries at the genus level, 255,094 (92.9%) at the family level, and 270,539 (98.5%) entries at the order level. We also looked at the number of species covered by the GenBank entries referencing the BOLD ID. As a result, 12,251 types of descriptions were found in GenBank. After excluding sp./aff./cf. from here, GenBank covered 8,744 species of fish with its DNA barcoding data (Figure 3). In addition, we found that some GenBank data were described up to the species level or the species name with a BIN ID (e.g., *Platycephalus* sp. 1 BOLD:ACT2912) (Ratnasingham and Hebert, 2013), whereas BOLD data were described up to the genus level.

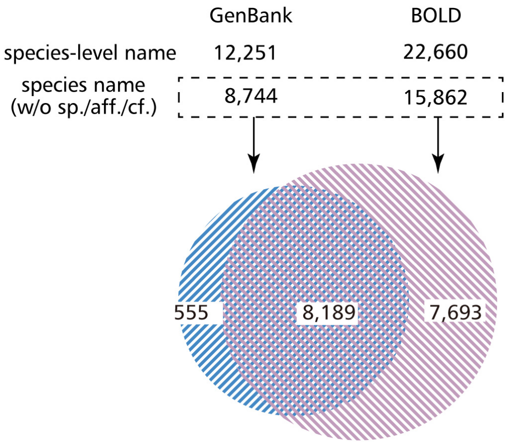


FIGURE 3
Venn diagram of overlapping species covered by GenBank and BOLD DNA barcode entries. We investigated how many species do BOLD and GenBank cover, and how many species overlap in these databases with regard to the fish DNA barcode data. Even if there are data in the species field, they are often not identified to the species level (such as sp. or aff.). Most of the species covered by GenBank are also covered by BOLD, which may be because BOLD imports GenBank data.

We then compared how many species did these species cover in the NCBI Taxonomy[14] (Schoch et al., 2020) and the GBIF Backbone Taxonomy[15] in the case of the fish. NCBI Taxonomy includes 22,041 species and subspecies. By comparing these with the 15,882 species that appear in BOLD, the descriptions of 14,505 matched (Table 1). Among the 1,377 descriptions that did not match, there were some that did not produce a hit because their description in BOLD was synonym, and so the percentage of matches was actually higher. If this synonym is not taken into account, then one could say that BOLD covers 65.8% of the NCBI Taxonomy species. In addition,

[14] https://www.ncbi.nlm.nih.gov/taxonomy

[15] https://www.gbif.org/dataset/d7dddbf4-2cf0-4f39-9b2a-bb099caae36c

GenBank covers 39.7%. The GBIF Backbone Taxonomy includes 104,767 species and subspecies of fish. By comparing these with the 15,882 species that appear in BOLD, the descriptions of 15,608 matched (Table 1). This is simply equivalent to 14.9% of the GBIF Backbone Taxonomy (Table 1). Of the 274 descriptions that did not produce a hit, 69 referred to hybrids. Moreover, of the 8,744 species that appeared in GenBank, 8,668 descriptions matched those of the GBIF Backbone Taxonomy. This is equivalent to 8.3% of the GBIF Backbone Taxonomy entries (Table 1).

Subsequently, we compared the similarities and the differences of the organisms covered by BOLD and GenBank. By comparing the list of species identified down to the species level, 8,189 descriptions were found to be common between BOLD and GenBank, and 7,693 descriptions were found to exist only in BOLD. Moreover, GenBank covered the description of 555 descriptions on its own (Figure 3).

New DNA barcode sequence candidates in GenBank

As mentioned earlier, a specimen_voucher qualifier is provided in GenBank in order to record the specimen ID. The number of entries with data here is much larger than the number of entries that refer to BOLD IDs (Figure 1B). Such entries without BOLD IDs but with specimen_voucher are potential candidates for new DNA barcode sequences, and we extracted these data from GenBank Nucleotide. There were 386,078 GenBank entries for fish with no BOLD IDs, but with data in the specimen_voucher field. Table 2 shows a list of candidate genes for DNA barcodes extracted from these entries, including COI, ND2, and RAG1. The candidate gene list in Table 2 contains descriptions representing the same genes because the various patterns of the gene names described by the submitters are not unified by text mining (e.g., COI and COX1, and cytb and Cytb).

Moreover, we surveyed how these candidate data would increase the species coverage. There are data on candidate barcode genes for 4,089 new organisms when compared to GenBank entries with BOLD IDs and BOLD data, and for 685 organisms when limited to data with gene name as "COI." In addition, the same analysis was performed on flowering plants, and entries for genes such as matK, rbcL, and trnL were obtained.

TABLE 1 Species coverage by the GenBank and BOLD DNA barcode entries.

		GenBank	BOLD
	Total	8,744	15,882
NCBI taxonomy	22,041	8,744 (39.7%)	14,505 (65.8%)
GBIF backbone taxonomy	104,767	8,668 (8.3%)	15,608 (14.9%)

We investigated how many species of the GenBank fish DNA barcode data cover the NCBI taxonomy and the GBIF backbone taxonomy. Species not covered here will be candidates for a new DNA barcode research in the future.

TABLE 2 Candidate list of DNA barcodes in GenBank.

(A) Fish

Gene name	Number of entries
COI	42,602
cytb	42,068
COX1	13,620
RAG1	9,062
ND2	5,932
S7	4,169
Cytb	3,771
myh6	3,349
zic1	3,043
RAG2	2,794

(B) Flowering plant

Gene name	Number of entries
matK	62,173
rbcL	46,917
trnL	34,273
psbA	25,761
rps16	23,473
trnK	20,068
ndhF	19,049
rpl16	12,635
rpl32	12,573
trnF	11,197

We picked up entries from GenBank with a sample ID but no BOLD ID, and extracted the gene names. Since we have not processed them by text mining, the same genes exist in the list with different spellings.

Discussion

Data import between barcode of life system and GenBank

Researchers can use BOLD and GenBank Nucleotide as databases for DNA barcodes. However, the two are different in nature: BOLD is the workbench for DNA barcoding projects, while INSDC (including GenBank Nucleotide) is a public repository of DNA data. In addition, the use of these databases differs between biodiversity researchers focusing on specimens, and molecular biologists focusing on nucleotide sequences.

Researchers often submit the same data in both databases. This should not be prohibited, and the Earth BioGenome Project[16] (Lewin et al., 2022) recommends submitting data

[16] https://www.earthbiogenome.org/

to both databases (Lawniczak et al., 2022). This suggests the convenience of using the two databases in an integrated manner, and emphasizes the differences in format and description.

BOLD imports DNA barcode sequences from GenBank. For the submission to both databases, we found an example where both the direct submission data and the imported data from GenBank for the same DNA barcode exist in the BOLD database (e.g., GAMBA659-12.COI-5P and ANGBF29940-19.COI-5P). We have previously reported examples of data directly being registered with BOLD (JBOL054-11) and data imported from GenBank (GBDP15012-14) (Nakazato, 2019). In these data, there was a difference in the description contents due to the differences in the formats applied by BOLD and by GenBank (**Supplementary Figure 2**). However, these duplicated data have now been resolved and unified to JBOL054-11.

Differences between barcode of life system and GenBank descriptions

GenBank utilizes BOLD IDs, but two types of writing style coexist: IDs for specimens (e.g., FOA941-05) and IDs for barcode sequences (e.g., FOA941-05.COI-5P). Specimen IDs often have multiple barcode genes assigned, thus GenBank should probably refer to IDs in the style of IDs for barcode sequences (i.e., FOA941-05.COI-5P). In order to solve this problem, the INSDC may need to check the format upon submission, or a secondary integration site may be required to do so. BOLD records data on a Darwin Core[17] (Wieczorek et al., 2012) basis, while GenBank records data in its own format; biodiversity information can also be described within GenBank: voucher_specimen, lat_lon (latitude and longitude), altitude, collection_date, collected_by, idetified_by, and country (**Supplementary Figure 1**). GenBank has been collecting sequences for over 30 years (Sayers et al., 2022b), so it will be difficult to comply with the Darwin Core anytime soon. NCBI and other bioinformatics organizations are working on data standardization and Semantic Web activities, thereby including data integration with the biodiversity field (Chawuthai et al., 2016; Groom et al., 2021; Nakazato, 2021).

Differences in taxonomy between barcode of life system and GenBank

GenBank and BOLD have different taxonomies: sequencing data-indexing GenBank uses the NCBI Taxonomy as its species list, while biodiversity databases such as GBIF and BOLD usually use the GBIF Backbone Taxonomy. For example, in BOLD, one level above Magnoliopsida as a class one will find Magnoliophyta (flowering plants) as a phylum, while in GenBank, the phylum is Streptophyta (green plants), and there are several hierarchical terms designed between the phylum and the class. We used Actinopterygii for fish, Insecta for insects, and Magnoliopsida for flowering plants in this study, which was the result of a careful selection of a common biological classification group for both BOLD and GenBank.

It should also be noted that the NCBI Taxonomy is a list of organisms for which sequences have been archived in INSDC, and it is not intended to cover all species. In addition, the NCBI Taxonomy may have the wrong species name because the submitter made a mistake when submitting the sequence (e.g., *Scarabaeus typhon* with Taxonomy ID: 1685123 should have been *Scarabaeus typhon*).

In this study, we have not normalized the descriptions of the NCBI Taxonomy and the GBIF Backbone Taxonomy. This is because the two databases are so different that the integrating of their data would be a big project by itself. However, we are able to assign the species that the DNA barcode indicates to the species in each database, and we have compared the assigned species. Currently, taxonomic information can only be confirmed by NCBI Taxonomy in GenBank and by the GBIF Backbone Taxonomy in BOLD. The integration of GenBank and BOLD data will make it easier to confirm the taxonomy of organisms in both the NCBI Taxonomy and the GBIF Backbone Taxonomy, and will enrich the information on the species indicated by the DNA barcode.

Further usefulness of using GenBank for the mining of DNA barcode data

In the field of DNA barcoding, DNA metabarcoding by using NGS is also performed (Adamowicz et al., 2019; Miya, 2022), and these data are archived in the SRA. The sequences assembled from these results will also be deposited in the NCBI database. Moreover, GenBank has rich literature information, and the use of this information is another advantage of data integration.

In addition, GenBank data were described up to the species level or the species name with BIN IDs (e.g., *Platycephalus* sp. 1 BOLD:ACT2912) (Ratnasingham and Hebert, 2013), whereas BOLD data were described up to the genus level. GenBank may be less reliable in identifying species than BOLD, since GenBank data are usually submitted by molecular biologists who are not experts in taxonomy (Leray et al., 2019; Meiklejohn et al., 2019; Pentinsaari et al., 2020). However, combining data from BOLD and GenBank would generate more detailed data that would complement each other. In this case, the identified_by field may increase the reliability of the obtained data.

Some of these may contain sequences corresponding to BOLD simply because there is no link from GenBank to BOLD, but they represent new possibilities for GenBank.

[17] https://dwc.tdwg.org/

In animals, COI genes are currently used primarily as barcodes, but in the future other genes, mitochondrial genomes, and whole genomes will be used as sources of barcodes. The value of using GenBank entries other than COI genes with specimen IDs will also increase. In fact, many gene sequences with specimen_voucher information are archived in GenBank, and it is expected that more data will be added through future museomics-associated research.

In this study, it was very difficult for us to download insect DNA barcoding data from BOLD in bulk. In order to solve this, there is a way to allow BOLD data downloadable from FTP sites for each taxonomic group. Alternatively, a further collaboration with GenBank would make it easier to do the research we have done here by processing the data provided by NCBI.

The Earth BioGenome Project (see text footnote 16) (Lewin et al., 2022) is another example that produces both sequences containing genomes and biodiversity information. They provide reports on various standards on their web page (see text footnote 16), and the information regarding the data registration in "IT and Informatics Standards[18]" is a particularly useful resource. The summaries of those reports have also been published in the form of a journal article (Lawniczak et al., 2022).

DNA barcoding research accelerates museomics

Museomics is a method of obtaining gene sequences from museum specimens. Museomics makes it possible to ascertain the phenotype (such as morphology and color), and the genotype (by gene sequence) of an organism of interest without the need to sample at the right time and place so as to obtain a living organism. In addition, sequence information can be used in order to distinguish between species and populations of organisms, which was not previously known from morphology. By sequencing older specimens, one gains the ability to perform a phylogenetic analysis of how evolution and differentiation occurred from both morphological and genetic aspects. This way, gene sequences are now an indispensable resource even in the field of biodiversity. The development of molecular biology in the last half-century may have brought about an unfortunate division in the life science fields: DNA-central molecular biology and bioinformatics, and non-DNA-central ecology and taxonomy. Museomics can fill these gaps, and DNA barcoding is also an important technology that bridges these two fields. The integration of biodiversity and sequence data will make these studies easier, and our current study will facilitate the application of museomics and bring the biological world together.

[18] https://www.earthbiogenome.org/it-and-informatics-standards

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found here: https://ftp.ncbi.nlm.nih.gov/genbank/, https://www.boldsystems.org/index.php/TaxBrowser_Home, https://ftp.ncbi.nlm.nih.gov/pub/taxonomy/.

Author contributions

TN conceived the idea, carried out the analyses, and wrote the first draft of the manuscript. UJ supervised the analyses and provided critical feedback on the manuscript. Both authors contributed to the article and approved the submitted version.

Funding

This work was supported by ROIS-DS-JOINT (004RM2017 and 009RM2018).

Acknowledgments

We would like to thank Hiromi Matsumae, Tsuyoshi Hosoya, Masaaki Kotera, and Ai Muto-Fujita for their very meaningful discussions regarding the writing of this manuscript.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fevo.2022.966605/full#supplementary-material

References

Adamowicz, S. J., Boatwright, J. S., Chain, F., Fisher, B. L., Hogg, I. D., Leese, F., et al. (2019). Trends in DNA barcoding and metabarcoding. *Genome* 62:v–viii. doi: 10.1139/gen-2019-0054

Andersson, A. F., Bissett, A., Finstad, A. G., Fossøy, F., Grosjean, M., and Hope, M. (2020). *Publishing Sequence-Derived Data Through Biodiversity Data Platforms. V1.0*. Copenhagen: GBIF Secretariat, doi: 10.35035/doc-vf1a-nr22

Arita, M., Karsch-Mizrachi, I., and Cochrane, G. (2021). The international nucleotide sequence database collaboration. *Nucl. Acids Res.* 49:D121–D124. doi: 10.1093/nar/gkaa967

Buerki, S., and Baker, W. J. (2016). Collections-based research in the genomic era. *Biol. J. Linn. Soc.* 117, 5–10. doi: 10.1111/bij.12721

Chawuthai, R., Takeda, H., Wuwongse, V., and Jinbo, U. (2016). Presenting and preserving the change in taxonomic knowledge for linked data. *Semant. Web* 7, 589–616. doi: 10.3233/SW-150192

DeSalle, R., and Goldstein, P. (2019). Review and interpretation of trends in DNA barcoding. *Front. Ecol. Evol.* 7:302. doi: 10.3389/fevo.2019.00302

GBIF Secretariat (2021). *GBIF Backbone Taxonomy*. Copenhagen: GBIF Secretariat, doi: 10.15468/39omei

Groom, Q. J., Dillen, M., Huybrechts, P., Johaadien, R., Kyriakopoulou, N., and Fernandez, F. J. Q. (2021). Connecting molecular sequences to their voucher specimens. *BioHackrXiv* [Preprint]. doi: 10.37044/osf.io/93qf4

Hebert, P. D., Cywinska, A., Ball, S. L., and deWaard, J. R. (2003). Biological identifications through DNA barcodes. *Proc. Biol. Sci.* 270, 313–321. doi: 10.1098/rspb.2002.2218

Lawniczak, M., Durbin, R., Flicek, P., Lindblad-Toh, K., Wei, X., Archibald, J. M., et al. (2022). Standards recommendations for the Earth BioGenome Project. *Proc. Natl. Acad. Sci. U.S.A.* 119:e2115639118. doi: 10.1073/pnas.2115639118

Leray, M., Knowlton, N., Ho, S. L., Nguyen, B. N., and Machida, R. J. (2019). GenBank is a reliable resource for 21st century biodiversity research. *Proc. Natl. Acad. Sci. U.S.A.* 116, 22651–22656. doi: 10.1073/pnas.1911714116

Lewin, H. A., Richards, S., Lieberman Aiden, E., Allende, M. L., Archibald, J. M., Bálint, M., et al. (2022). The earth BioGenome project 2020: starting the clock. *Proc. Natl. Acad. Sci. U.S.A.* 119:e2115635118. doi: 10.1073/pnas.2115635118

Meiklejohn, K. A., Damaso, N., and Robertson, J. M. (2019). Assessment of BOLD and GenBank - Their accuracy and reliability for the identification of biological materials. *PLoS One* 14:e0217084. doi: 10.1371/journal.pone.0217084

Miya, M. (2022). Environmental DNA metabarcoding: a novel method for biodiversity monitoring of marine fish communities. *Annu. Rev. Mar. Sci.* 14, 161–185. doi: 10.1146/annurev-marine-041421-082251

Nakazato, T. (2019). Current situation of DNA Barcoding data in biodiversity and genomics databases and data integration for museomics. *Biodivers. Inf. Sci. Stand.* 3:e35165. doi: 10.3897/biss.3.35165

Nakazato, T. (2021). knowledge extraction from specimen-derived data from GenBank to enrich biodiversity information. *Biodivers. Inf. Sci. Stand.* 5:e73787. doi: 10.3897/biss.5.73787

National Library of Medicine [NLM] (2009). *iBOL/GenBank/Genome Canada Letter of Cooperation*. Available Online at: https://www.ncbi.nlm.nih.gov/core/assets/genbank/files/iBol-Letter-of-Cooperation.pdf (accessed June 4, 2022).

Nilsson, R. H., Larsson, K. H., Taylor, A., Bengtsson-Palme, J., Jeppesen, T. S., Schigel, D., et al. (2019). The UNITE database for molecular identification of fungi: handling dark taxa and parallel taxonomic classifications. *Nucl. Acids Res.* 47:D259–D264. doi: 10.1093/nar/gky1022

Pentinsaari, M., Ratnasingham, S., Miller, S. E., and Hebert, P. (2020). BOLD and GenBank revisited - Do identification errors arise in the lab or in the sequence libraries? *PLoS One* 15:e0231814. doi: 10.1371/journal.pone.0231814

Ratnasingham, S., and Hebert, P. D. (2007). BOLD: the barcode of life data system (http://www.barcodinglife.org). *Mol. Ecol. Notes* 7, 355–364. doi: 10.1111/j.1471-8286.2007.01678.x

Ratnasingham, S., and Hebert, P. D. N. (2013). A DNA-based registry for all animal species: the Barcode Index Number (BIN) system. *PLoS One* 8:e66213. doi: 10.1371/journal.pone.0066213

Raxworthy, C. J., and Smith, B. T. (2021). Mining museums for historical DNA: advances and challenges in museomics. *Trends Ecol. Evol.* 36, 1049–1060. doi: 10.1016/j.tree.2021.07.009

Sayers, E. W., Cavanaugh, M., Clark, K., Pruitt, K. D., Schoch, C. L., Sherry, S. T., et al. (2022b). GenBank. *Nucl. Acids Res.* 50:D161–D164. doi: 10.1093/nar/gkab1135

Sayers, E. W., Bolton, E. E., Brister, J. R., Canese, K., Chan, J., Comeau, D. C., et al. (2022a). Database resources of the national center for biotechnology information. *Nucl. Acids Res.* 50:D20–D26. doi: 10.1093/nar/gkab1112

Schoch, C. L., Ciufo, S., Domrachev, M., Hotton, C. L., Kannan, S., Khovanskaya, R., et al. (2020). NCBI Taxonomy: a comprehensive update on curation, resources and tools. *Database* 2020:baaa062. doi: 10.1093/database/baaa062

Wieczorek, J., Bloom, D., Guralnick, R., Blum, S., Döring, M., Giovanni, R., et al. (2012). Darwin core: an evolving community-developed biodiversity data standard. *PLoS One* 7:e29715. doi: 10.1371/journal.pone.0029715

frontiers | Frontiers in Ecology and Evolution

TYPE Original Research
PUBLISHED 10 November 2022
DOI 10.3389/fevo.2022.943361

OPEN ACCESS

EDITED BY
Jonathan J. Fong,
Lingnan University,
China

REVIEWED BY
Pedro Ribeiro,
Academy of Sciences of the
Czech Republic (ASCR), Czechia
Lauren Eserman-Campbell,
Atlanta Botanical Garden,
United States

*CORRESPONDENCE
David J. Lohman
dlohman@ccny.cuny.edu

†PRESENT ADDRESS
Caroline Storer,
Pacific Biosciences,
Menlo Park, CA, United States

SPECIALTY SECTION
This article was submitted to
Phylogenetics, Phylogenomics, and
Systematics, a section of the journal
Frontiers in Ecology and Evolution

RECEIVED 13 May 2022
ACCEPTED 04 October 2022
PUBLISHED 10 November 2022

CITATION
Nunes R, Storer C, Doleck T, Kawahara AY, Pierce NE and Lohman DJ (2022) Predictors of sequence capture in a large-scale anchored phylogenomics project. *Front. Ecol. Evol.* 10:943361. doi: 10.3389/fevo.2022.943361

Predictors of sequence capture in a large-scale anchored phylogenomics project

Renato Nunes[1,2], Caroline Storer[3†], Tenzing Doleck[1,2], Akito Y. Kawahara[3,4,5], Naomi E. Pierce[6] and David J. Lohman[1,2,7*]

[1]Biology Department, City College of New York, City University of New York, New York, NY, United States, [2]PhD Program in Biology, Graduate Center, City University of New York, New York, NY, United States, [3]McGuire Center for Lepidoptera and Biodiversity, Florida Museum of Natural History, University of Florida, Gainesville, FL, United States, [4]Entomology and Nematology Department, University of Florida, Gainesville, FL, United States, [5]Department of Biology, University of Florida, Gainesville, FL, United States, [6]Department of Organismic and Evolutionary Biology and Museum of Comparative Zoology, Harvard University, Cambridge, MA, United States, [7]Entomology Section, National Museum of Natural History, Manila, Philippines

Next-generation sequencing (NGS) technologies have revolutionized phylogenomics by decreasing the cost and time required to generate sequence data from multiple markers or whole genomes. Further, the fragmented DNA of biological specimens collected decades ago can be sequenced with NGS, reducing the need for collecting fresh specimens. Sequence capture, also known as anchored hybrid enrichment, is a method to produce reduced representation libraries for NGS sequencing. The technique uses single-stranded oligonucleotide probes that hybridize with pre-selected regions of the genome that are sequenced *via* NGS, culminating in a dataset of numerous orthologous loci from multiple taxa. Phylogenetic analyses using these sequences have the potential to resolve deep and shallow phylogenetic relationships. Identifying the factors that affect sequence capture success could save time, money, and valuable specimens that might be destructively sampled despite low likelihood of sequencing success. We investigated the impacts of specimen age, preservation method, and DNA concentration on sequence capture (number of captured sequences and sequence quality) while accounting for taxonomy and extracted tissue type in a large-scale butterfly phylogenomics project. This project used two probe sets to extract 391 loci or a subset of 13 loci from over 6,000 butterfly specimens. We found that sequence capture is a resilient method capable of amplifying loci in samples of varying age (0–111years), preservation method (alcohol, papered, pinned), and DNA concentration (0.020ng/µl - 316ng/ul). Regression analyses demonstrate that sequence capture is positively correlated with DNA concentration. However, sequence capture and DNA concentration are negatively correlated with sample age and preservation method. Our findings suggest that sequence capture projects should prioritize the use of alcohol-preserved samples younger than 20years old when available. In the absence of such specimens, dried samples of any age can yield sequence data, albeit with returns that diminish with increasing age.

KEYWORDS

anchored hybrid enrichment, historical DNA, hybrid capture, Lepidoptera, museomics, archival DNA, Papilionoidea, phylogenomics

Introduction

Next-generation sequencing (NGS) has revolutionized phylogenomics by drastically decreasing the cost and time required to generate large datasets of genome-wide genetic markers. However, while NGS technologies were developed to sequence whole genomes, entire assemblies are generally not preferred for systematics because the surfeit of data is unwieldy. Data files are large, requiring high performance computer clusters and much time for bioinformatics and phylogenetic analysis. In addition, gene duplication and chromosomal arrangements complicate assessment of homology between species and make alignment of whole assemblies difficult (Armstrong et al., 2019). Low-coverage whole genome sequencing is an alternative to traditional high-coverage genome sequencing that shows promise for use in phylogenomics and population genetics (Zhang et al., 2019a; Lou et al., 2021). This method can be used in both model and non-model organisms and for species with relatively small genomes it can be a powerful and cost-effective approach (Zhang et al., 2019b). Low-coverage whole genome sequencing has been used to study evolution of the butterfly family Papilionidae by extracting loci with BLAST-based orthology searches (Allio et al., 2020). There is also potential for combining low-coverage whole genome data with other methods to increase genetic and taxonomic sampling in phylogenetic studies (Ribeiro et al., 2021; Talavera et al., 2021). Despite this, low-coverage whole genome sequencing still retains some limitations of whole genome sequencing, including dependency on existing reference genomes and genomic resources. To overcome these limitations, several reduced representation methods have been developed to target and sequence only homologous loci (Davey et al., 2011). These methods still require high performance computers, but the computational power needed is lower than for assembly of whole genomes. The most common reduced representation methods used in phylogenetics might be divided into three categories: enzymatic digestion methods such as RADseq (Baird et al., 2008); sequence capture including capture and sequencing of ultraconserved elements (UCEs; Faircloth et al., 2012; McCormack et al., 2012), which targets a specific category of genomic areas; and transcriptomics. Transcriptomes, another source of genome-wide markers from protein-coding genes that can be used for phylogenomic reconstruction (Grabherr et al., 2011; Kawahara and Breinholt, 2014; Kawahara et al., 2019). There are costs and benefits of each method (Table 1).

Reduced representation methods

Complete taxon sampling is desirable to provide accurate estimates of diversification through time and other questions in macroecology and evolution (Morlon et al., 2011; Jetz et al., 2012). Increased taxon sampling also increases the accuracy of phylogenetic inference by breaking up long branches and minimizing the effects of coalescent stochasticity (Zwickl and Hillis, 2002; Huang et al., 2010). Comprehensive phylogenetic studies that aim to include samples from all described taxa within a group or samples from a geographically broad area are frequently hampered by lack of samples with high quality DNA. Many species are rare, have limited geographic distributions, are protected from collecting by legislation, or live in a part of the world where research permission is difficult to obtain (Rabinowitz, 1981; Prathapan et al., 2018; Wells et al., 2019). Thus, more comprehensive sampling can be achieved by incorporating existing genetic data, such as DNA barcodes or other Sanger data. These pre-existing data cannot usually be combined with UCEs or RADseq data because they rarely have any homologous loci in common (Table 1; Harvey et al., 2016; Toussaint et al., 2021c). However, loci with ample pre-existing data can be targeted by sequence capture. In addition, DNA can be sequenced from museum or herbarium specimens that were not collected specifically for genetic research (Bi et al., 2013; Staats et al., 2013). Following recent usage, we refer to DNA extracted from such specimens as historical DNA or hDNA (Billerman and Walsh, 2019; Raxworthy and Smith, 2021). Historical DNA is typically degraded and fragmented after years of storage at ambient temperatures. Prior to NGS, specimens collected within a few decades could sometimes yield sequence data by labor-intensive means: designing taxon-specific primers to amplify short, overlapping DNA segments usually under 200 bp (Eastwood and Hughes, 2003; Lohman et al., 2008). Fortuitously, preparation of DNA for short-read NGS requires that it be fragmented into short pieces, so specimens collected in the 20th century frequently yield NGS sequence data.

TABLE 1 Advantages and disadvantages of several reduced representation methods for obtaining phylogenomic datasets.

Attributes	RADseq	UCEs	PCR/Sanger	Transcriptomes	Target Capture
Can efficiently sequence hundreds or thousands of loci	X	X		X	X
Ease of combining with Sanger data, including DNA barcodes			X	X	X
Ease of extracting homologous loci from genome assemblies		X	X	X	X
Can easily sequence DNA from museum specimens		X			X
Targets pre-selected genomic regions			X	X	X
May require investment in probe design					X

RADseq and allied methods use enzymes to cut high molecular weight genomic DNA into fragments that are then selected based on their size. If the only sample available for a particular taxon is from a decades-old museum specimen with degraded hDNA, the technique will likely not work because the DNA has already been fragmented randomly over time before digestion with site-specific enzymes. Thus, fragments of a given length may not be homologous among samples, and sequence quality may be poor (Graham et al., 2015). While it is possible to map short NGS reads of hDNA to existing RADseq loci or develop sequence capture probes matching the RAD fragments (Tin et al., 2014; Ali et al., 2016; Hoffberg et al., 2016; Suchan et al., 2016; Lang et al., 2020), these methods are more expensive and complex. In addition, it is difficult to distinguish orthologs from paralogs and assess potential linkage disequilibrium with RADseq data (Rubin et al., 2012).

Both UCEs and target capture can use short-read NGS and are thus amenable to sequencing hDNA from museum specimens (Bailey et al., 2016; Blaimer et al., 2016; McCormack et al., 2016). However, target capture has a few advantages over UCEs: Sanger sequences are available for a greater diversity of species because the techniques have been around longer (Table 1). In addition, the function of UCEs and the evolutionary mechanism for their invariance among distantly related taxa are poorly understood (Dermitzakis et al., 2005; Ahituv et al., 2007). Some researchers are therefore reluctant to apply evolutionary models to stretches of DNA flanking the UCE sites, which may evolve in an atypical fashion. With target capture, loci with known evolutionary rates can be targeted to resolve either deep or shallow relationships (Leaché and Rannala, 2011; Townsend and Leuenberger, 2011; Grover et al., 2012; Hamilton et al., 2016). A possible disadvantage of target capture is the time and money that needs to be invested in identifying target loci and developing probes for them (Faircloth, 2017), but probe sets for numerous taxa already exist (Andermann et al., 2020), or can be designed with the help of software packages including MrBait and others (Chamala et al., 2015; Mayer et al., 2016; Faircloth, 2017; Campana, 2018; Chafin et al., 2018). Thus, target capture is frequently the method of choice for phylogenomics projects, especially those that incorporate hDNA from museum and herbarium samples (Jones and Good, 2016). The method has been used to investigate relationships among many taxa including bats (Bailey et al., 2016), birds (Prum et al., 2015), frogs (Hime et al., 2021), spiders (Hamilton et al., 2016; Wood et al., 2018), harvestmen (Derkarabetian et al., 2019), odonates (Bybee et al., 2021), butterflies (Breinholt et al., 2018; Espeland et al., 2018; Kawahara et al., 2018; Ma et al., 2020), moths (Hamilton et al., 2019; Homziak et al., 2019; Dowdy et al., 2020; Zhang et al., 2020), and a variety of plants (Johnson et al., 2019; Eserman et al., 2021; Acha and Majure, 2022).

Sequence capture: How it works

Sequence capture, also known as target capture, target sequence capture, target enrichment, or anchored hybrid enrichment, is an *in vitro* process that separates pre-selected loci of interest from other genomic regions (Lemmon et al., 2012). First, genomic regions are selected and single-stranded, oligonucleotide probes complementary to the target sequences are designed using existing genomes (Gnirke et al., 2009). If the probes target exons, the process is sometimes called exon capture (Bragg et al., 2016), and if all of the protein-coding loci in the genome are sequenced, the end result is called an exome. The probes are only ca. 100–200 bp in length, but longer genomic regions can be targeted by overlapping or "tiling" multiple probes to span the desired probe region (Bertone et al., 2006). The success of sequence capture depends on the similarity of the probe sequence to the target sequence, which declines with decreasing relatedness between the taxon used to design the probes and the taxon being enriched. Tiling probes from more than one species' genome can increase the taxonomic breadth with which the probes can be used.

Probes can be synthesized commercially or be made from the modified PCR products of high-quality genomic DNA (Maricic et al., 2010; Peñalba et al., 2014; Knyshov et al., 2019; Zhang et al., 2019a, 2019b). One advantage of PCR-generated probes is that a reference genome is not required to design the probes, and sequence capture may therefore be used in taxa that lack genomic resources (Jones and Good, 2016). The probes are then biotinylated and combined with streptavidin-coated magnetic beads. Ratios of different probes should be carefully controlled so that sequencing coverage will be equal for all loci, which requires reducing the concentration of probes for organellar DNA in relation to nuclear DNA because it is more abundant in DNA extracts (Peñalba et al., 2014).

To prepare specimens for sequence capture, genomic DNA is extracted from each sample and transmogrified into a "library" by chopping it into short pieces with ultrasound or enzymes, then ligating sequencing adapters and sample-specific indexes (a.k.a. barcodes) to the ends of the DNA fragments (Bronner and Quail, 2019). At this stage, multiple libraries can be multiplexed by combining them and sequencing them together (Meyer and Kircher, 2010). Next, the probes and libraries are combined in a solution hot enough to denature double-stranded library fragments, and the temperature is lowered so that target sequences anneal to their complementary probes. The biotin within the probe then irreversibly binds to the streptavidin on the magnetic beads. A neodymium magnet is placed near the tube, causing the targeted fragments, now bound to the magnetic beads, to adhere to the sides (Paijmans et al., 2015). The fluid is then removed from the tube along with non-target DNA in solution. After a purification step, the tube is re-filled with buffer, heated so the hydrogen bonds binding the target DNA to the probes break, thus releasing the targeted library fragments from the probes and into solution, and—with the magnet still in place—the buffer perfused with DNA fragments from targeted regions is removed and sequenced on a short-read NGS platform such as Illumina. Libraries can be PCR amplified before and/or after the hybridization step. The resulting short reads are bioinformatically demultiplexed, quality-controlled, and assembled.

First, low quality reads and sequence contaminants including adapters are removed. Next, the filtered reads are assembled in one of several ways: *de novo*, with reference sequences, or *via* reference-guided assembly (Allen et al., 2017). Paralogs are then removed, and consensus sequences are extracted (Andermann et al., 2020). Several bioinformatic pipelines for assembling short of target loci are available (Faircloth, 2016; Johnson et al., 2016; Allen et al., 2017; Andermann et al., 2018). The final product is a set of homologous sequences for a group of taxa.

Sample preservation and DNA quality

Decades of research have identified best practices for preserving tissues for genetic and other molecular research. The high molecular weight nucleic acids present in the nuclei of living tissues quickly degrade into ever-smaller fragments as the post-mortem interval increases (Ludes et al., 1993; Camacho-Sanchez et al., 2013). When genetic data became more commonplace in evolutionary and systematic studies in the late 1980s, it was apparent that standard methods of specimen preservation, such as pinning insects and preparing vertebrate skins, was not ideal for preserving DNA. Conventional wisdom held that thin insect legs dried quickly and often yielded DNA suitable for PCR, but drying, relaxing, spreading, and re-drying Lepidoptera specimens accelerated DNA fragmentation. Experiments to find the best DNA preservation methods ensued (Arctander, 1988; Pyle and Adams, 1989; Post et al., 1993) and continue to be tested as new preservatives are developed (Dillon et al., 1996; Dawson et al., 1998; Camacho-Sanchez et al., 2013; Moreau et al., 2013). The current consensus affirms that cryopreservation in liquid nitrogen or −80°C storage is the preservation method of choice for animal tissues because it preserves DNA, RNA, and proteins indefinitely if the cold chain remains unbroken (Prendini et al., 2002). However, it is often not feasible to lug a nitrogen vapor shipper into the field, keep it charged with liquid nitrogen, and convince airline staff that the bomb-shaped container is safe to bring on an airplane. Thus, fieldwork-friendly alternatives are required. Comparative studies on vertebrate tissues find that some buffers can preserve RNA and DNA at room temperature for long periods of time (Camacho-Sanchez et al., 2013), while a dimethylsulfoxide-sodium solution works well for marine invertebrates (Dawson et al., 1998). Strong (95–100%) ethanol is the favored preservative for insects (Quicke et al., 1999; King and Porter, 2004; Moreau et al., 2013), and drying specimens quickly using silica gel also works well for preserving insect DNA (Post et al., 1993; Dillon et al., 1996). Other types of alcohol, such as methanol and propanol, are not as effective as ethanol for DNA preservation (Post et al., 1993). Killing insects with ethyl acetate seems to degrade DNA (Dillon et al., 1996), and should therefore be avoided. Since the scaly wings of Lepidoptera would be disfigured if immersed in ethanol, making them difficult to identify, one or both forewing-hindwing pairs are removed and placed in a glassine envelope or coin holder before the body is placed in a tube of ethanol (Supplementary Figure S1; Cho et al., 2016). With their cell walls and enzyme-inhibiting secondary metabolites, preservation conditions differ for plants. Early research suggested that ethanol is a poor preservative of plant DNA (Doyle and Dickson, 1987), and drying leaf tissue rapidly in silica gel is generally the preferred method (Pyle and Adams, 1989; Chase and Hills, 1991).

Sample preservation and sequence capture success

As studies incorporating hDNA become increasingly common (Colella et al., 2020; Toussaint et al., 2021c; Garg et al., 2022), researchers will be faced with decisions regarding sample selection. Should an ethanol-preserved specimen always be extracted if a museum specimen is available? If an irreplaceable specimen is destructively sampled to extract DNA, how likely is sequence capture success? What body parts are most likely to yield high quality DNA? We took advantage of sample metadata collected from a large-scale sequence capture project aimed at investigating the evolutionary history of butterflies to identify relationships among several measures of sequencing success and sample age, preservation method, and extracted tissue type. Our results are summarized to provide a decision tree to aid sample selection. While our results are derived exclusively from butterfly samples, they will apply to other insects and dried specimens stored at ambient temperatures.

Materials and methods

Samples

We analyzed metadata associated with 6,146 butterfly specimens from six families that were subjected to sequence capture for several phylogenetic studies undertaken as part of ButterflyNet (Espeland et al., 2018; Kawahara et al., 2018; Toussaint et al., 2018; Toussaint et al., 2019; Braby et al., 2020; Carvalho et al., 2020; Valencia-Montoya et al., 2021; Toussaint et al., 2021a; Toussaint et al., 2021b; Kawahara et al., 2022). This NSF-funded collaborative network aims to infer the phylogeny of butterflies and aggregate data on species distributions (Pinkert et al., 2022) and traits (Shirey et al., 2022; butterflynet.org). The phylogenomic component of the project used two sequence capture probe sets. The first of these, BUTTERFLY1.0, targets 390 single-copy, protein-coding nuclear loci and a single mitochondrial locus: the DNA barcoding fragment of cytochrome c oxidase I (COI; Breinholt et al., 2018; Espeland et al., 2018). We refer to this as the “391-locus probe set”. We aimed to sequence at least one species from each of the *ca*. 1900 valid butterfly genera (Lamas, 2015) with the BUTTERFLY1.0 probe set (Kawahara

TABLE 2 Sample predictor variables that may impact sequence capture success.

Variable	Type	Unit/Value	N	Mean	Median	Range
Age	Continuous	years	5,273	8.7	5	0–111
Concentration	Continuous	ng/µl	5,525	43.2	37.5	0–316
Preservation	Categorical	ethanol	1779			
Preservation	Categorical	papered	1,440			
Preservation	Categorical	pinned	430			
Tissue	Categorical	abdomen	2,372			
Tissue	Categorical	leg	671			
Tissue	Categorical	thorax	1,605			
ProbeSet	Categorical	13/391	6,146			
Family	Categorical	Hesperiidae	422			
Family	Categorical	Lycaenidae	1,201			
Family	Categorical	Nymphalidae	1,026			
Family	Categorical	Papilionidae	78			
Family	Categorical	Pieridae	483			
Family	Categorical	Riodinidae	121			

Sample sizes (N) indicate the number of samples with data that could be included in analyses. Fractional years were used in the analyses, and Concentration was also used as a response variable.

et al., 2022); the type species of each genus was sequenced if available. Sequences from the remaining specimens were captured with the BUTTERFLY2.0 probe set (Kawahara et al., 2018), which targets 13 loci found in BUTTERFLY1.0 that are often used in butterfly phylogenetics, (Wahlberg and Wheat, 2008) including COI. We call this the "13-locus probe set".

The 13-locus probe set and the 391-locus probe set have successfully generated data to resolve evolutionary relationships at varying taxonomic levels. The BUTTERFLY 2.0 13-locus dataset has resolved relationships within the family Hedylidae providing robust support for 80% of nodes (Kawahara et al., 2018). Data generated with this probe set has also been used to recover tribal level relationships in the Acraeini (Carvalho et al., 2020), Baorini (Toussaint et al., 2019), and Candalidini (Braby et al., 2020). The larger BUTTERFLY 1.0 probe set has most notably been used in creating comprehensive and dated phylogenies of the superfamily Papilionoidea (butterflies) including 98% of all tribes (Espeland et al., 2018) and 84% of all genera (Kawahara et al., 2022). The loci in this set have also been use to generate phylogenetic backbones for the subtribe Euptychiina (Espeland et al., 2019) and the tribe Eumaeini (Valencia-Montoya et al., 2021). Some studies have even combined both sets to further increase phylogenetic resolution in the subfamily Coeliadinae (Toussaint et al., 2021a, 2021b, 2021c) and in the subfamily Heteropterinae (Toussaint et al., 2021a, 2021b, 2021c). Data generated with these sets also have applications beyond systematics and have been applied to study butterfly phylogenetic diversity (Earl et al., 2021).

We recorded specimen variables that might predict sequencing success: DNA concentration; type of tissue extracted; preservation method; sample age; and family. We refer to these variables as Concentration, Tissue, Preservation, Age, and Family, respectively (Table 2). Values for Preservation were "ethanol" for samples in which wingless bodies were preserved in a tube of 95–100% ethanol specifically for genetic research, "papered" for specimens that were dried with their wings folded and stored in a paper envelope—a common method of preservation in the field, and "pinned" to indicate specimens that had been skewered on a pin and prepared for a dry specimen collection (Supplementary Figure S1). Most pinned samples were likely dried and papered in the field, then relaxed in a sealed, humid container for *ca.* 3–24 h before being pinned and spread. The length of time between collection and relaxing/spreading/pinning is unknown and likely varies among samples. Pinned and papered specimens were obtained from the Museum of Comparative Zoology at Harvard University, the McGuire Center for Lepidoptera and Biodiversity at the University of Florida, the City College of New York, and the American Museum of Natural History. Pinned and papered specimens are common in museum collections and were not preserved with the intention of using the samples for genetic research (Kassambara, 2020). There were 654 samples sequenced with the 391-locus probe set and 2,645 samples sequenced with the 13-locus probe set that had complete metadata. Thousands of other samples had some but not all metadata. Missing metadata meant that analyses were conducted with different numbers of samples (Table 2).

We used these predictor variables to assess several measures of sequence capture success: DNA concentration (which is a response variable in some analyses); the fragment length of extracted DNA before library preparation; the probe set used; the number of loci captured with each probe set; and the sequence quality (Table 3). Average DNA fragment length after extraction but prior to library preparation was assessed by running *ca.* 3 µl of each extracted DNA sample on a 2% agarose gel. This index of DNA quality, which we called "Fragmentation," was scored in a binary manner depending on whether most fragments were greater than or less than 1,000 bp in relation to a standard DNA ladder. After the raw reads for each sample were processed in accordance with uniform quality control measures described

TABLE 3 Response variables used as indicators of successful sequence capture.

Variable	Type	Unit/Value	N	Mean	Median	Range
LociCaptured13	Ordinal	integer (0–13)	3,741	12.5	13	0–13
LociCaptured391	Ordinal	integer (0–391)	1873	350.4	381	0–391
Fragmentation	Binary	1kbp	2,771			
Quality	Continuous	integer	3,586	0.412	0	1–144

below, we assessed sequencing success as the number of loci captured (variable names: LociCaptured13 and LociCaptured391), depending on the probe set (13 or 391) and assessed sequence quality by calculating the number of IUPAC ambiguities in the 657 bp sequence of COI from each specimen (variable name: Quality). This mitochondrial gene is maternally inherited and should be wholly homozygous within a single individual. Any ambiguities therefore represent uncertainty in the assembly associated with poor sequence quality. Ambiguous bases might represent truly heterozygous sites in nuclear genes, but not in mitochondrial genes, which is why we only used COI.

DNA extraction

DNA was extracted with OmniPrep™ Genomic DNA Purification Kits for Tissue.[1] Tissue samples were not weighed before extraction. Ethanol preserved specimens were extracted following the methods in Espeland et al. (2018), while papered and pinned specimens were extracted following methods described in St Laurent et al. (2018). Genitalia at the tip of the abdomen were never extracted. If abdominal tissue from a pinned specimen was extracted non-destructively by macerating it in extraction buffer, the distal end of the abdomen was placed in a clear gelatin capsule that was then pierced with the specimen pin (Supplementary Figure S1). DNA extracts were quantified using a Qubit 3 Fluorometer using dsDNA HS and BR Assay kits.[2] To minimize sequencing failure, samples with a DNA concentration less than 4 ng/μl were rarely subjected to capture and sequencing, and overly concentrated extracts were often diluted to be less than 150 ng/μl to prevent problems with multiplexing.

Library preparation, target enrichment, and sequencing

Quantified extracts were submitted to RAPiD Genomics[3] for library preparation, hybrid enrichment, and sequencing. Libraries were generated by first mechanically shearing DNA to a size of 300 bp. Once sheared, adenine residues were ligated to the 3′ end of the blunt-end fragments to allow for the ligation of barcoded adapters and the PCR-amplification of the library (Breinholt et al., 2018; Espeland et al., 2018; Kawahara et al., 2018). Agilent SureSelect probes[4] were then used for solution-based target enrichment of pools containing 16 libraries. Enrichment of these libraries followed the SureSelect^{XT} Target Enrichment System for Illumina Paired-End Multiplexed Sequencing Library protocol (Breinholt et al., 2018; Espeland et al., 2018; Kawahara et al., 2018). These enriched libraries were then multiplexed and sequenced with an Illumina HiSeq 3,000 producing paired-end 100-bp reads (Espeland et al., 2018; Kawahara et al., 2018).

Locus assembly

An existing pipeline for anchored phylogenomics was used to assemble raw Illumina reads (Breinholt et al., 2018). First, paired-end Illumina data were cleaned, and adapters were removed using Trim Galore! 0.4.0.[5] Selected reads had a minimum read size of 30 bp and bases with a Phred score above 20 (Breinholt et al., 2018). Loci were then assembled using an iterative baited assembly (IBA) process that used reads with a forward and reverse read that passed prior filtering (Breinholt et al., 2018; Espeland et al., 2018; Kawahara et al., 2018). The assembly process uses the custom python script IBA.py available on Dryad (Breinholt et al., 2017), which uses USEARCH v7.0 (Edgar, 2010) to find raw reads that matches the probe region of the reference taxa. These assembled reads were then filtered using the python script s_hit_checker.py available on Dryad (Breinholt et al., 2017). This script searched assembled reads against a *Danaus plexippus* reference genome and these results were used for single hit and orthology filtering with a bit score threshold of 0.90 (Breinholt et al., 2018; Espeland et al., 2018; Kawahara et al., 2018). Orthologs were then screened for contamination by identifying and removing sequences that were identical or nearly identical at different taxonomic levels (Breinholt et al., 2018; Espeland et al., 2018; Kawahara et al., 2018).

Statistical analyses

Data were cleaned in the tidyverse (Wickham et al., 2019) and visualized with ggplot (Wickham 2016; Kassambara, 2020). First, we modeled Concentration as a response variable with Age, Preservation, Tissue, and Family as the explanatory

[1] gbiosciences.com

[2] thermofisher.com

[3] rapid-genomics.com

[4] agilent.com

[5] bioinformatics.babraham.ac.uk/projects/trim_galore/

variables (Table 2). We considered interactions between Age and Preservation to determine whether Preservation had age-dependent effects on DNA concentration. We log-transformed Concentration and generated generalized linear models (GLM) in R (RStudio Team, 2020; R Core Team, 2021) using the lme4 package (Bates et al., 2015).

Next, we modeled LociCaptured13 and LociCaptured391 (Table 3) with Age, Preservation, and Tissue as explanatory variables (Table 2). Family was initially used as an explanatory variable but was removed from the final model due to its lack of significance. We considered interactions between Age and Preservation to determine whether Preservation had age-dependent effects on locus capture. We generated GLMs in R using the MASS package (Venables and Ripley, 2002) with a quasi-Poisson distribution to model LociCaptured13 and LociCaptured391 while accounting for overdispersion. To determine whether the proportion of loci captured was different between probe sets, we calculated the proportion of loci captured as the ratio of loci captured over the targeted number of loci. We used a nonparametric Kruskal-Wallis test to determine if the proportion of loci captured was significantly different between probe sets.

To understand how sequence capture and concentration varied in relation to age for each combination of Preservation and Tissue, we calculated Spearman rank correlations between LociCaptured13, LociCaptured391 and Concentration versus sample age across the 9 unique combinations of Preservation and Tissue type possible. To explore the relationship between sequence capture and butterfly family we plotted LociCaptured13 and LociCaptured391 versus sample age across the unique combinations of family and preservation method. We also calculated Spearman rank correlations between the numerical variables in our dataset for each probe set, which included combinations of Age:Concentration, Age:LociCaptured, Age:LociCaptured13, Age:LociCaptured391 and Concentration:LociCaptured (Table 3). Spearman rank correlations were calculated in R using the correlation package (Makowski et al., 2020).

To determine whether some Preservation methods or Tissue types led to higher LociCaptured13, higher LociCaptured391, or longer DNA fragment lengths, we used Pearson chi-square tests. We compared the number of ethanol, papered, and pinned samples that failed or succeeded to capture 50% or more of the loci targeted by the probe set, which is how we coded "successful" locus capture. We performed a similar analysis comparing numbers of samples with average DNA Fragment sizes over 1,000 bp vs. under 1,000 bp in relation to their method of Preservation. We then assessed failed vs. successful sequence capture as a function of the Tissue that was extracted: legs, thorax, or abdomen. Since the majority of samples that we analyzed were ethanol samples, we suspected that these might drive the result, so we excluded them and repeated the analysis with data from papered and pinned specimens only

Results

Determinants of DNA concentration

Age, Preservation, Tissue, and Family were significant predictors of DNA Concentration. Additionally, there were significant interactions between Age and Preservation suggesting that Age impacted Concentration differently depending on the Preservation method (Supplementary Figure S2). The concentration of extracted DNA declines with specimen age when data from all sample preservation types are aggregated ($\rho = -0.071$, $p = 3.07\text{e-}07$; Table 3). Throughout this paper ρ = the Greek letter rho, which is the Spearman rank correlation test statistic, and p, an abbreviation for probability, is the Latin lowercase letter P. However, the effect is only significant in papered ($\rho = -0.1$, $p = 0.00012$) and pinned specimens ($\rho = -0.26$, $p = 1\text{e-}06$), which were not preserved for molecular research (Figure 1A; Supplementary Figure S2). There was no relationship between age and DNA concentration in ethanol preserved tissues ($\rho = 0.022$, $p = 0.37$), but the oldest such sample that we included was 26.83 years old because preservation of Lepidoptera in ethanol for genetic research began only around three decades ago. The type of tissue extracted had a strong effect on DNA concentration. For papered and pinned specimens, the rank order from highest to lowest concentration was abdomen > thorax > legs, while for ethanol-preserved specimens, the order was thorax > abdomen > legs (Figure 2A). Within each tissue type, the rank order of DNA concentration was always ethanol > papered > pinned, though the differences were negligible when legs were extracted (Figure 2A).

DNA fragmentation and sequence quality

Fragment length depends on Preservation method ($\chi^2 = 19.12$; $p = 7.05\text{E-}05$). Ethanol-preserved specimens had more samples with fragment lengths over 1,000 bp (93%), followed by papered (72%) and pinned (56%) samples. Ethanol-preserved and papered specimens had significantly more samples with fragment lengths over 1,000 bp than would be expected by chance ($p = 5.75\text{E-}163$ and $p = 3.67\text{E-}36$; Supplementary Figure S3A). Remarkably, there were no significant relationships between age and fragment length in any Preservation method (Figure 3A). Out of 3,586 COI mitochondrial sequences, only 210 (~6%) had at least one ambiguity. The modal number of ambiguities per sequence was 2 (68 samples), and the highest number of ambiguities per sequence was 144. When disaggregated by Preservation method and plotted against sample Age, there were no apparent relationships (Figure 3B).

Determinants of sequence capture success

The 13-locus and 391-locus probe sets successfully captured loci from samples of varying Age, Concentration, Preservation,

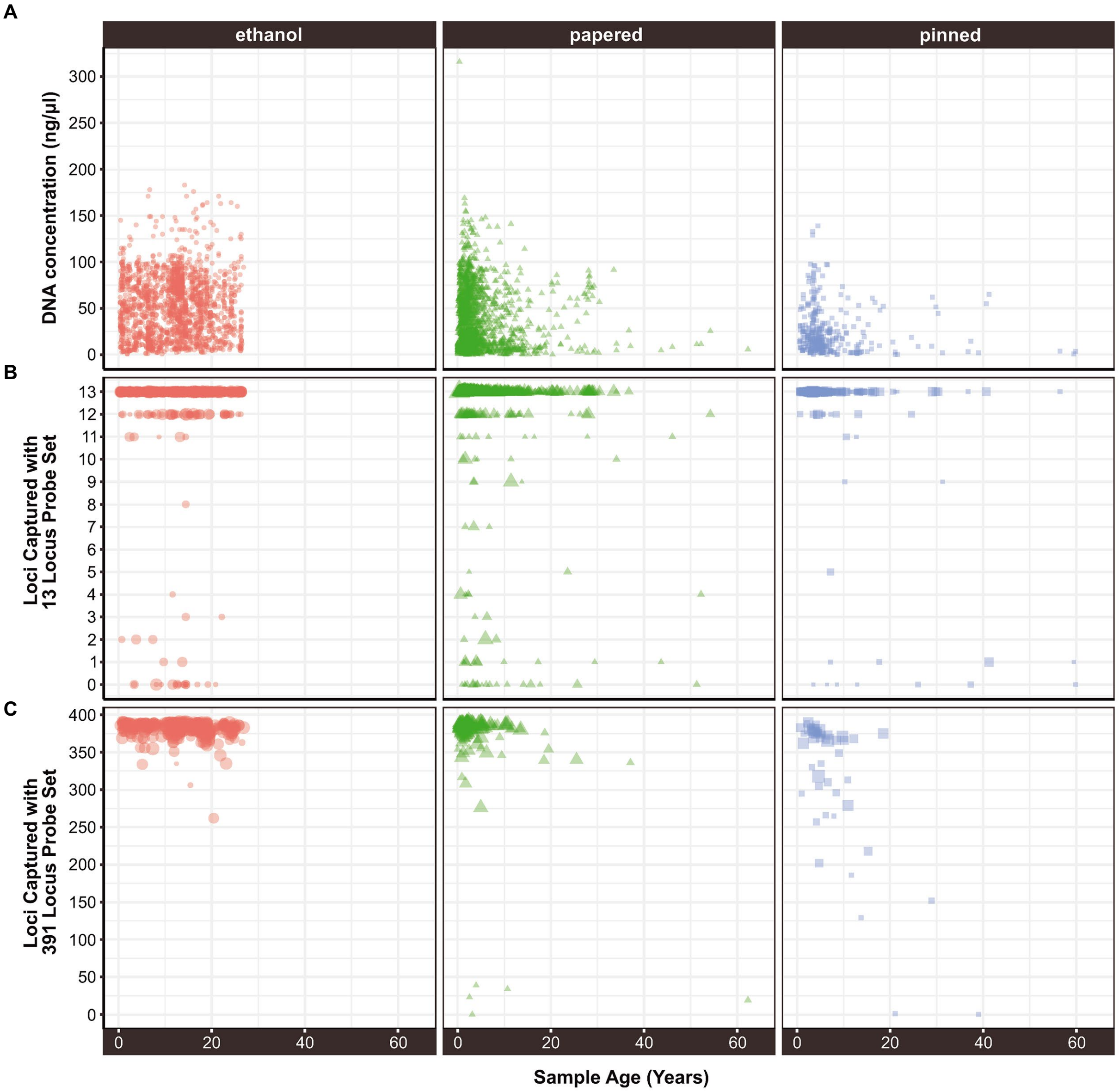


FIGURE 1
The relationship between sample age and **(A)** extracted DNA concentration, **(B)** locus capture with a 13-locus probe set, and **(C)** locus capture with a 391-locus probe set. In **(B,C)** the size of each point is proportional to its concentration.

Tissue, and Family. Family had no significant effect on LociCaptured13 or LociCaptured391, but all other variables did. There are no significant relationships between Family and LociCaptured with either ProbeSet or any Preservation method (Supplementary Figures S4, S5). The variable "Family" was therefore removed from the models. Age, Concentration, and Preservation were significant predictors of both LociCaptured13 and LociCaptured391. However, while Tissue was not a significant predictor of LociCaptured13, it was a significant predictor of LociCaptured391. Interactions between Age: Preservation were significant, suggesting that Age impacts locus capture differently depending on the sample preservation method (Supplementary Figures S6, S7). Ethanol preserved specimens have higher average locus capture as Age increases when the other predictors are held constant, followed by papered specimens, and then pinned specimens.

The BUTTERFLY1.0 probe set recovered 100% of 391 targeted loci in some samples, with a mean of 352.68 loci (mode = 385) and the BUTTERFLY2.0 probe set captured a mean of 12.53 loci (median and mode = 13; Table 3). Remarkably, this probe set captured 100% of the 13 targeted loci from the oldest sample in our dataset (111 years). Across all 6,146 samples, we recovered more than 50% of targeted loci in 5879 samples (391-locus probe set = 1888 samples; 13-locus probe set = 3,991 samples), and less

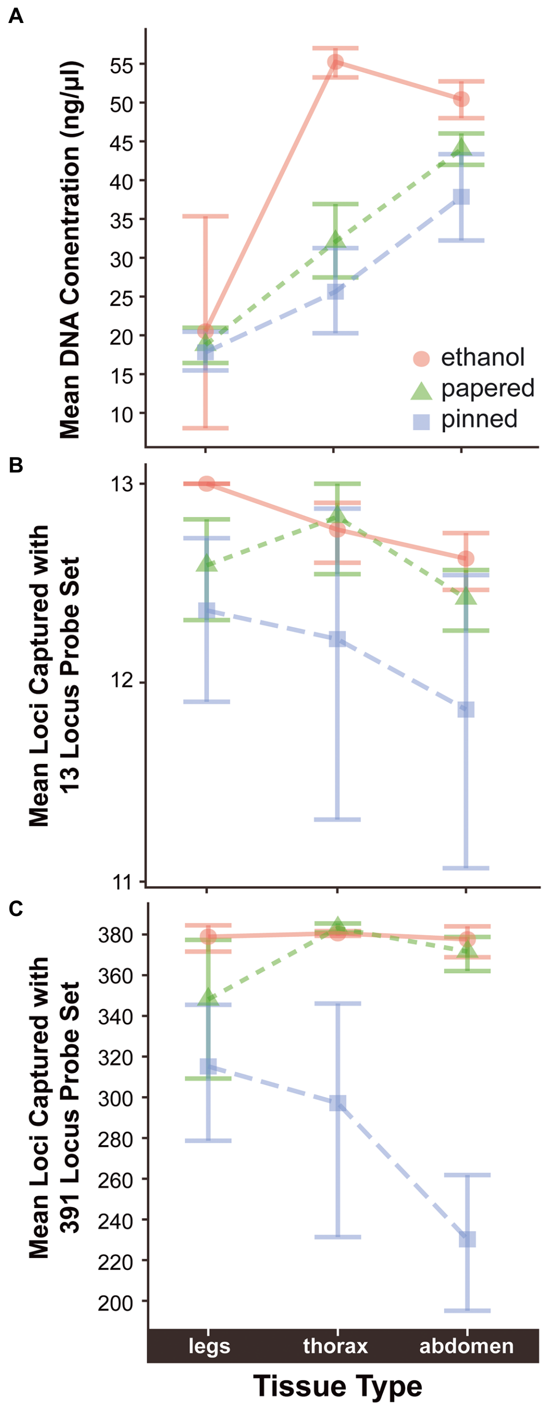


FIGURE 2
Differences in **(A)** mean DNA concentration and **(B)** mean number of loci captured with 13-locus and **(C)** 391-locus probe sets in relation to preservation method and type of tissue extracted. Jitter has been added to points in **(A)** so their distribution can be better assessed visually.

than 50% of targeted loci from 267 samples (391-locus probe set = 137 samples; 13-locus probe set = 130 samples), including at least 82 samples that failed to recover any loci (391-locus probe set = 9 samples, 13-locus probe set = 73 samples). The median proportion of locus capture (ratio of loci captured over the number of loci targeted) of the 13-locus probe set was significantly higher than the median proportion of locus capture of the 391-locus probe set (H = 3561.3, df = 1, p = <2.2e–16; Supplementary Figure S8).

LociCaptured13, LociCaptured391, and Concentration are negatively correlated with sample Age, and, while the direction of the correlations is consistent between the probe sets, the strength of the correlations varies (Figures 1B,C; Supplementary Figures S6, S7). The number of loci captured is negatively correlated with Age, and this effect is stronger for the 391-locus probe set (LociCaptured391) than the 13-locus probe set (LociCaptured13; ρ_{391} = –0.25, p = 8.12E-24; ρ_{13} = –0.13, p = 9.24E-15; Table 4). There was an exception to this pattern when looking at the unique combinations of Preservation and Tissue: LociCaptured391 was not affected by the Age of papered specimens, as there were several young and old specimens that failed to capture (Supplementary Figure S7). Across all sample tissues and preservation methods, a negative trend between locus capture and age is apparent although not always significant. The strength of the relationship between sample age and loci captured was weak for ethanol-preserved samples (ρ_{391} = –0.19; p = 3.4e-05; ρ_{13} = –0.07, p = 0.013), strongest for pinned samples (ρ_{391} = –0.64; p = 1.2e-06; ρ_{13} = –0.31, p = 1.4e-06), and intermediate for papered samples (ρ_{391} = –0.024; p = 0.74; ρ_{13} = –0.12, p = 3.1e-0.5). Age and LociCaptured for papered and pinned specimens generally had significant negative correlation coefficients (Supplementary Figures S6, S7). Age-dependent capture was strongly affected by tissue type and ProbeSet (Supplementary Figures S6, S7). This trend of decreasing locus capture with age is more clearly seen with both probe sets in pinned samples regardless of Tissue extracted, although the decrease in LociCaptured vs. Age is more apparent in the 391-locus probe set.

LociCaptured is positively correlated with Concentration, and this effect is stronger for the 391-locus than the 13-locus probe set (ρ_{391} = 0.22, p = 1.03E-18; ρ_{13} = 0.16, p = 1.09E-23). When including LociCaptured for both probe sets, Age is negatively correlated with LociCaptured (ρ = –0.053, p = 0.00011); Concentration and LociCaptured are positively correlated (ρ = 0.150, p = 1.85E-27; Table 4).

The incidence of sequence capture failure was low, but there was again a clear rank order of success. Ethanol samples had the highest capture rate (98%) followed by papered (96%), then pinned specimens (94%; Supplementary Figure S3B). The type of tissue extracted had a similarly negligible effect on capture success. Extractions from abdominal tissue were most successful (98%), followed by thorax tissue (97%), followed by legs (96%). These values were lower by 1–2% when ethanol samples were excluded from the analysis (Supplementary Figures S3C,D).

Discussion

Sample preservation

Of the three methods we analyzed, immersion in absolute ethanol is the best way to preserve sample DNA for sequencing. If

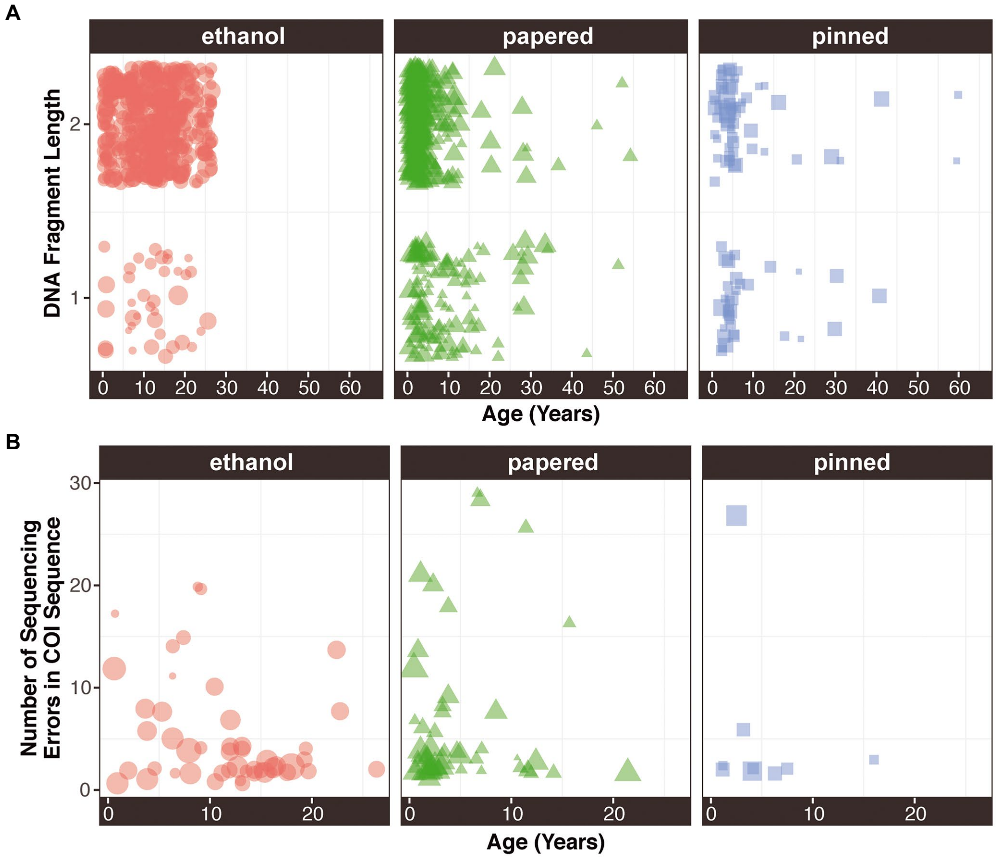


FIGURE 3
(A) DNA fragment length and **(B)** sequence quality as a function of age preservation method. The size of each point is proportional to its concentration.

TABLE 4 Spearman rank correlations between sample age, DNA extract concentration, and the number of loci captured with two probe sets targeting 13 or 391 loci.

	Age	*p* value	Concentration	*p* value
Concentration	−0.071	3.07E-07		
LociCaptured13	−0.13	9.24E-15	0.16	1.09E-23
LociCaptured391	−0.25	8.12E-24	0.22	1.03E-18
LociCapturedBoth	−0.053	0.00011	0.15	1.85E-27

ethanol preserved samples are not available, dry papered specimens generally have better results than pinned specimens. The concentration of DNA extracted from ethanol-preserved specimens did not decline with sample age (Figures 1A, 2A; Supplementary Figure S2), as it did with papered and pinned specimens. The fragment length of extracted DNA was also generally longer (Figure 3A; Supplementary Figure S3A). While this is not crucial for sequence capture, which requires fragmented DNA for short-read sequencing, it is essential for other sequencing platforms such as PacBio HiFi and Oxford Nanopore Technologies long-read sequencing (Whibley et al., 2021; Lawniczak et al., 2022). Thus, preserving samples in ethanol allows them to be used with a broader range of genetic/genomic techniques.

We found no relationship between Preservation type and sequence quality. Although we found a non-significant trend for declining sequence quality with sample age in ethanol-preserved samples but no other sample types (Figure 3B), this might have been an artifact of how we plotted these data. We removed samples with perfect sequence quality (no ambiguities in COI), which comprised most samples, prior to plotting the data. There were thousands more ethanol samples than other sample types (Table 2), so the true impact of age on sequence quality is likely negligible. A greater proportion of loci were captured from ethanol preserved samples than from papered or pinned samples (Table 4; Figures 1B,C, 2B,C). The labs that provided the ethanol-preserved specimens sequenced for this study follow best practices that may improve DNA preservation: 1) Specimens are immersed in 100% ethanol immediately after being killed by pinching the thorax and having their wings removed. No chemical killing agents are used that could compromise DNA quality, and dead specimens are not allowed to air dry (and potentially decay) before ethanol preservation. 2) Several weeks after returning from the field, the ethanol in each tube is discarded and replaced with fresh 100% ethanol. Water in the specimen leaches into the ethanol and dilutes its concentration over time. 3) Ethanol samples are stored in ultracold −80° C freezers.

There are other insect preservation methods not evaluated in this study. For example, we had no access to tissues stored at −140°C in liquid nitrogen vapor. While it is an excellent method for preserving biological molecules, it is impractical to use in many field situations. We extracted *ca.* ten samples preserved in RNAlater, but these rarely yielded DNA that was sufficiently concentrated for sequencing (>4 ng/μl). These samples were immersed in the preservative immediately after specimens were killed and torn into pieces because aqueous solutions such as RNAlater cannot easily penetrate the hydrophobic cuticle of insects, and thus can fail to preserve tissues suspended in preservative unless the cuticle is ruptured (Evans et al., 2013). In accordance with the manufacturer's instructions, these specimens were kept as cold as possible in a thermos with ice in the field, and frozen upon return to the lab. There were too few RNAlater preserved specimens to include in our statistical analyses, but we anecdotally conclude that RNAlater is a poor DNA preservative, consistent with the findings of others (Moreau et al., 2013). A study comparing nucleic acid preservation methods for mammal tissues stored at room temperature found that nucleic acid preservation (NAP) buffer was better than 100% ethanol and cryopreservation for preserving DNA and better than RNAlater for preserving RNA after several months of storage (Camacho-Sanchez et al., 2013) at ambient temperatures. Future comparative work should investigate preservation of insect tissues with NAP buffer under ambient conditions, as this buffer has additional advantages of being inexpensive, non-flammable, and stable at ambient temperatures.

Sample age

Sample age has miniscule effects on DNA concentration (Supplementary Figure S2) and sequence capture (Supplementary Figures S6, S7) of ethanol-preserved tissues, regardless of tissue type. The concentration of DNA extracts declines with sample age in papered and pinned specimens, but the type of tissue extracted affects this pattern. The negative correlation is strongest and most significant in abdominal tissues, but weak and not significant (or marginally significant) in extracts from legs or thoraxes. However, extracts from abdomens are generally more concentrated than extracts from other tissues (Supplementary Figure S2). The relationships between Age and LociRecovered13 and LociRecovered391 are significantly negative for pinned specimens, but the relationship is weak for papered specimens (Supplementary Figures S6, S7). In sum, ethanol preserved specimens do not degrade over time, but if one must use papered or pinned specimens, younger specimens yield better results—especially for pinned specimens.

These results bolster results from other research taxa, demonstrating that plant specimens up to 204 years old are amenable to hybrid capture (Brewer et al., 2019). While McGaughran (2020) found that older moth samples have the poorest capture success, Toussaint et al. (2021c) found that sequence coverage was not linked to the age of beetle specimens.

DNA concentration

Hybrid capture requires more DNA than PCR (Chung et al., 2016). While PCR can proceed if there are just a few strands of DNA that are not fragmented between the binding sites of the two primers, the commercial laboratory that we contracted to perform sequence capture and sequencing (see footnote 3) recommends a minimum of *ca.* 132 ng of DNA per sample (4 ng/μl x 33 μl), though we successfully sequenced samples with less DNA. Since DNA concentration generally decreases with age in pinned and papered specimens (Figure 1A; Supplementary Figure S2), it is best to select the youngest available specimens if there are several of varying ages. The small size of many insects constrains the amount of DNA that can be extracted from them. The amount of DNA that can be extracted is further diminished as papered and pinned specimens age at ambient temperatures (Supplementary Figure S2).

DNA concentration can affect sequence capture below a threshold concentration that is difficult to estimate (perhaps *ca.* 2–5 ng/μl), but above that, it has a negligible impact on the number of loci captured. We captured 100% of loci from samples with DNA concentrations as low as 0.020 ng/μl and 10.60 ng/μl (13-locus and 391-locus probe sets, respectively), and large numbers of loci were captured with the 391-locus probe set from samples with much lower concentrations, including a sample with a DNA concentration of 2.4 ng/μl that captured 386 loci. These results demonstrate that high sequence capture success can be achieved with surprising small amounts of DNA, albeit not consistently. Conversely, samples with high DNA concentrations do not always guarantee sequence capture. Samples with concentrations of 144 ng/μl and 167 ng/μl failed to recover any loci with the 13-locus and 391-locus probe sets, respectively. Higher DNA concentrations do not guarantee locus capture or higher numbers of captured loci. Further, high DNA concentrations can adversely affect the sequencing depth of other samples multiplexed in the same run by using a disproportionately large number of sequencing reads.

While tissue type is a significant determinant of DNA Concentration, it has little impact on the number of loci captured (Supplementary Figures S6, S7). Therefore, destructively sampling a specimen's thorax or abdomen only needs to be undertaken when the minimum DNA concentration threshold cannot be met by extracting legs. The value of this threshold will likely depend on the requirements of the PCR hybridization and amplification steps employed in the sequence capture protocol. We used a standard number of PCR cycles during the hybridization step for every sample, but increasing the number of PCR cycles might increase locus capture success of samples with low DNA concentrations. This strategy might increase the likelihood of successful sequence capture of rare or endangered species that can only be obtained as old museum samples.

Degradation

Preservation method seems to be an important determinant of both DNA concentration and locus capture since alcohol preserved specimens had consistently high average concentrations and locus capture regardless of age, while papered and pinned samples had gradual decreases in concentration and loci capture versus sample Age. This is likely due to the ability of different preservation methods to stabilize DNA and prevent degradation.

Short-read next-generation sequencing methods require short fragments of DNA and can sequence DNA from old specimens. Thus, NGS has become a common alternative to PCR and Sanger sequencing, enabling incorporation of museum and herbarium samples in projects that require DNA sequencing (McGaughran, 2020; Mayer et al., 2021; Raxworthy and Smith, 2021). However, severe degradation that produces fragment lengths below the target length of the sequencing method will likely prevent a sample from being captured. The magnitude of these effects depends on the probe length and sequence target length of the library preparation step. Increasing the probe tiling depth and length of the probed region will likely aid capture of degraded samples.

Stochastic variation

We analyzed thousands of samples—one to two orders of magnitude more than similar comparative studies investigating the relationship between sample type and sequencing success (McGaughran, 2020; Mayer et al., 2021). Several samples that were expected to perform well failed to recover many (or any) loci. Given our large sample size, outliers are likely, and may have resulted from unrecorded sample properties that would be important for determining the amount of DNA degradation such as storage temperature, humidity, sample history (specimens shipped as loans, extractions being repeatedly frozen/thawed, extracts kept at ambient temperature for too long, etc.). Additionally, this could also be the result of human or laboratory error. Competition for sequencing within pooled runs could also explain some of this variation, but we did not have information to include that factor in our models.

The sequence quality metadata in this study are a byproduct of multiple phylogenetics studies and many steps were taken to maximize the likelihood of successful locus capture. Therefore, our dataset has a disproportionate number of younger samples, meaning that the smaller number of older samples that happen to have been successful have a strong effect on the relationships that we explore. We excluded no outliers in our analyses. Including old samples that captured successfully sometimes created weakly positive relationships between locus capture and sample age, when this relationship is expected to be negative. However, removal of these outlier samples could erroneously create models that confirm *a priori* assumptions about locus capture.

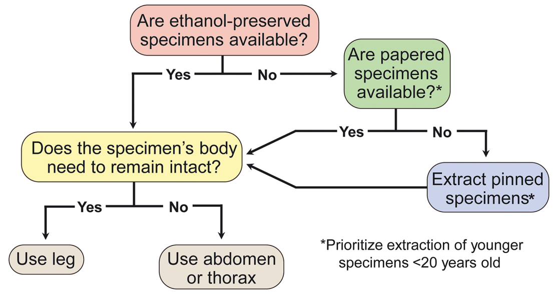

Are ethanol-preserved specimens available?
Yes
No
Are papered specimens available?*
Yes
No
Does the specimen's body need to remain intact?
Extract pinned specimens*
Yes
No
Use leg
Use abdomen or thorax
*Prioritize extraction of younger specimens <20 years old

FIGURE 4
Decision tree for selection of insect specimens most likely to yield optimal sequence capture results.

Conclusion

Sequence capture is a remarkably resilient method for obtaining sequence data for phylogenomic analysis. We find that DNA from insect specimens stored under less-than-ideal conditions and over a century old can be sequenced successfully. However, success is more likely under certain conditions, and we use our results to provide recommendations for sample selection and preservation (Figure 4). We find higher DNA concentrations are correlated with greater locus capture, but the difference between loci captured is small across samples with low and high concentrations. Sample age is negatively correlated with locus capture, although many or all loci can be captured from older samples. Sample preservation type plays an important role for determining locus capture, with ethanol-preserved samples performing better than papered and pinned samples in our models and correlation analyses. However, samples preserved with any of the methods we investigated can capture a large proportion of targeted loci. The effect that age has on locus capture appears to depend on preservation method, and pinned samples have the steepest decline in locus capture vs. age. By comparing the proportion of loci captured with the number of targeted loci for each probe set, we find that the probe set with fewer targeted loci not only performs better, it also appears to be resistant to decreases in locus capture associated with Age, Concentration, Preservation, and Tissue. We conclude that sequence capture is a robust method that can be used to include historical samples in contemporary phylogenetic and population genetic studies with relatively low risk of failure and marginally diminishing returns when using older and non-ethanol-preserved samples, regardless of the tissue type used for DNA extraction.

Data availability statement

Supplementary figures and the dataset analyzed in this paper are provided in the Supplementary material. Further inquiries can be directed to the corresponding author.

Author contributions

DL conceived of this project. RN performed statistical analyses. CS and TD performed lab work. CS undertook bioinformatic analyses of NGS data. AK and NP provided samples for analysis and conceived of the ButterflyNet project with DL. RN and DL wrote the first draft of the manuscript and prepared the figures. All authors contributed to the article and approved the submitted version.

Funding

This project was supported by an NSF GoLife collaborative grant “ButterflyNet”: DEB-1541500 to AK and Robert Guralnick; DEB-1541560 to NP; and DEB-1541557 to DL. Published by a grant from the Wetmore Colles Fund of the Museum of Comparative Zoology.

Acknowledgments

We thank Y-Lan Nguyen and Kelly M. Dexter for performing lab work, and are grateful to David Grimaldi, Andrew D. Warren, Crystal A. Maier, and Rachel L. Hawkins Sipe for facilitating access to specimens under their care. We are grateful for the improvements suggested by two reviewers.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher’s note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors, or the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fevo.2022.943361/full#supplementary-material

References

Acha, S., and Majure, L. C. (2022). A new approach using targeted sequence capture for phylogenomic studies across Cactaceae. *Genes* 13:350. doi: 10.3390/genes13020350

Ahituv, N., Zhu, Y., Visel, A., Holt, A., Afzal, V., Pennacchio, L. A., et al. (2007). Deletion of ultraconserved elements yields viable mice. *PLoS Biol.* 5:e234. doi: 10.1371/journal.pbio.0050234

Ali, O. A., O’Rourke, S. M., Amish, S. J., Meek, M. H., Luikart, G., Jeffres, C., et al. (2016). RAD capture (rapture): flexible and efficient sequence-based genotyping. *Genetics* 202, 389–400. doi: 10.1534/genetics.115.183665

Allen, J. M., Boyd, B., Nguyen, N.-P., Vachaspati, P., Warnow, T., Huang, D. I., et al. (2017). Phylogenomics from whole genome sequences using aTRAM. *Syst. Biol.* 66, syw105–syw798. doi: 10.1093/sysbio/syw105

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

Andermann, T., Cano, Á., Zizka, A., Bacon, C., and Antonelli, A. (2018). SECAPR—a bioinformatics pipeline for the rapid and user-friendly processing of targeted enriched Illumina sequences, from raw reads to alignments. *PeerJ* 6:e5175. doi: 10.7717/peerj.5175

Andermann, T., Torres Jiménez, M. F., Matos-Maraví, P., Batista, R., Blanco-Pastor, J. L., Gustafsson, A. L. S., et al. (2020). A guide to carrying out a phylogenomic target sequence capture project. *Front. Genet.* 10:1407. doi: 10.3389/fgene.2019.01407

Arctander, P. (1988). Comparative studies of avian DNA by restriction fragment length polymorphism analysis: convenient procedures based on blood samples from live birds. *J. Ornithol.* 129, 205–216. doi: 10.1007/BF01647289

Armstrong, J., Fiddes, I. T., Diekhans, M., and Paten, B. (2019). Whole-genome alignment and comparative annotation. *Annu. Rev. Anim. Biosci.* 7, 41–64. doi: 10.1146/annurev-animal-020518-115005

Bailey, S. E., Mao, X., Struebig, M., Tsagkogeorga, G., Csorba, G., Heaney, L. R., et al. (2016). The use of museum samples for large-scale sequence capture: a study of congeneric horseshoe bats (family Rhinolophidae). *Biol. J. Linn. Soc.* 117, 58–70. doi: 10.1111/bij.12620

Baird, N. A., Etter, P. D., Atwood, T. S., Currey, M. C., Shiver, A. L., Lewis, Z. A., et al. (2008). Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS One* 3:e3376. doi: 10.1371/journal.pone.0003376

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015). Fitting linear mixed-effects models using lme4. *J. Stat. Softw.* 67, 1–48. doi: 10.18637/jss.v067.i01

Bertone, P., Trifonov, V., Rozowsky, J. S., Schubert, F., Emanuelsson, O., Karro, J., et al. (2006). Design optimization methods for genomic DNA tiling arrays. *Genome Res.* 16, 271–281. doi: 10.1101/gr.4452906

Bi, K., Linderoth, T., Vanderpool, D., Good, J. M., Nielsen, R., and Moritz, C. (2013). Unlocking the vault: next-generation museum population genomics. *Mol. Ecol.* 22, 6018–6032. doi: 10.1111/mec.12516

Billerman, S. M., and Walsh, J. (2019). Historical DNA as a tool to address key questions in avian biology and evolution: a review of methods, challenges, applications, and future directions. *Mol. Ecol. Resour.* 19, 1115–1130. doi: 10.1111/1755-0998.13066

Blaimer, B. B., Lloyd, M. W., Guillory, W. X., and Brady, S. G. (2016). Sequence capture and phylogenetic utility of genomic ultraconserved elements obtained from pinned insect specimens. *PLoS One* 11:e0161531. doi: 10.1371/journal.pone.0161531

Braby, M. F., Espeland, M., Müller, C. J., Eastwood, R., Lohman, D. J., Kawahara, A. Y., et al. (2020). Molecular phylogeny of the tribe Candalidini (Lepidoptera: Lycaenidae): systematics, diversification and evolutionary history. *Syst. Entomol.* 45, 703–722. doi: 10.1111/syen.12432

Bragg, J. G., Potter, S., Bi, K., and Moritz, C. (2016). Exon capture phylogenomics: efficacy across scales of divergence. *Mol. Ecol. Resour.* 16, 1059–1068. doi: 10.1111/1755-0998.12449

Breinholt, J. W., Earl, C., Lemmon, A. R., Lemmon, E. M., Xiao, L., and Kawahara, A. Y. (2017). Data from: resolving relationships among the megadiverse butterflies and moths with a novel pipeline for anchored phylogenomics. *Syst. Biol.* 67, 78–93. doi: 10.5061/DRYAD.RF7G5

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

Brewer, G. E., Clarkson, J. J., Maurin, O., Zuntini, A. R., Barber, V., Bellot, S., et al. (2019). Factors affecting targeted sequencing of 353 nuclear genes from herbarium specimens spanning the diversity of angiosperms. *Front. Plant Sci.* 10:1102. doi: 10.3389/fpls.2019.01102

Bronner, I. F., and Quail, M. A. (2019). Best practices for Illumina library preparation. *Curr. Protoc. Hum. Genet.* 102:e86. doi: 10.1002/cphg.86

Bybee, S. M., Kalkman, V. J., Erickson, R. J., Frandsen, P. B., Breinholt, J. W., Suvorov, A., et al. (2021). Phylogeny and classification of Odonata using targeted genomics. *Mol. Phylogenet. Evol.* 160:107115. doi: 10.1016/j.ympev.2021.107115

Camacho-Sanchez, M., Burraco, P., Gomez-Mestre, I., and Leonard, J. A. (2013). Preservation of RNA and DNA from mammal samples under field conditions. *Mol. Ecol. Resour.* 13, 663–673. doi: 10.1111/1755-0998.12108

Campana, M. G. (2018). BaitsTools: software for hybridization capture bait design. *Mol. Ecol. Resour.* 18, 356–361. doi: 10.1111/1755-0998.12721

Carvalho, A. P. S., St Laurent, R. A., Toussaint, E. F. A., Storer, C., Dexter, K. M., Aduse-Poku, K., et al. (2020). Is sexual conflict a driver of speciation? A case study with a tribe of brush-footed butterflies. *Syst. Biol.* 70, 413–420. doi: 10.1093/sysbio/syaa070

Chafin, T. K., Douglas, M. R., and Douglas, M. E. (2018). MrBait: universal identification and design of targeted-enrichment capture probes. *Bioinformatics* 34, 4293–4296. doi: 10.1093/bioinformatics/bty548

Chamala, S., García, N., Godden, G. T., Krishnakumar, V., Jordon-Thaden, I. E., De Smet, R., et al. (2015). MarkerMiner 1.0: a new application for phylogenetic marker development using angiosperm transcriptomes. *Appl. Plant Sci.* 3:1400115. doi: 10.3732/apps.1400115

Chase, M. W., and Hills, H. H. (1991). Silica gel: an ideal material for field preservation of leaf samples for DNA studies. *Taxon* 40, 215–220. doi: 10.2307/1222975

Cho, S., Epstein, S. W., Mitter, K., Hamilton, C. A., Plotkin, D., Mitter, C., et al. (2016). Preserving and vouchering butterflies and moths for large-scale museum-based molecular research. *PeerJ* 4:e2160. doi: 10.7717/peerj.2160

Chung, J., Son, D.-S., Jeon, H.-J., Kim, K.-M., Park, G., Ryu, G. H., et al. (2016). The minimal amount of starting DNA for Agilent's hybrid capture-based targeted massively parallel sequencing. *Sci. Rep.* 6:26732. doi: 10.1038/srep26732

Colella, J. P., Tigano, A., and MacManes, M. D. (2020). A linked-read approach to museomics: higher quality de novo genome assemblies from degraded tissues. *Mol. Ecol. Resour.* 20, 856–870. doi: 10.1111/1755-0998.13155

Davey, J. W., Hohenlohe, P. A., Etter, P. D., Boone, J. Q., Catchen, J. M., and Blaxter, M. L. (2011). Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nat. Rev. Genet.* 12, 499–510. doi: 10.1038/nrg3012

Dawson, M. N., Raskoff, K. A., and Jacobs, D. K. (1998). Field preservation of marine invertebrate tissue for DNA analyses. *Mol. Mar. Biol. Biotechnol.* 7, 145–152.

Derkarabetian, S., Benavides, L. R., and Giribet, G. (2019). Sequence capture phylogenomics of historical ethanol-preserved museum specimens: unlocking the rest of the vault. *Mol. Ecol. Resour.* 19, 1531–1544. doi: 10.1111/1755-0998.13072

Dermitzakis, E. T., Reymond, A., and Antonarakis, S. E. (2005). Conserved non-genic sequences — an unexpected feature of mammalian genomes. *Nat. Rev. Genet.* 6, 151–157. doi: 10.1038/nrg1527

Dillon, N., Austin, A. D., and Bartowsky, E. (1996). Comparison of preservation techniques for DNA extraction from hymenopterous insects. *Insect Mol. Biol.* 5, 21–24. doi: 10.1111/j.1365-2583.1996.tb00036.x

Dowdy, N. J., Keating, S., Lemmon, A. R., Lemmon, E. M., Conner, W. E., Scott Chialvo, C. H., et al. (2020). A deeper meaning for shallow-level phylogenomic studies: nested anchored hybrid enrichment offers great promise for resolving the tiger moth tree of life (Lepidoptera: Erebidae: Arctiinae). *Syst. Entomol.* 45, 874–893. doi: 10.1111/syen.12433

Doyle, J. J., and Dickson, E. E. (1987). Preservation of plant samples for DNA restriction endonuclease analysis. *Taxon* 36, 715–722. doi: 10.2307/1221122

Earl, C., Belitz, M. W., Laffan, S. W., Barve, V., Barve, N., Soltis, D. E., et al. (2021). Spatial phylogenetics of butterflies in relation to environmental drivers and angiosperm diversity across North America. *iScience* 24:102239. doi: 10.1016/j.isci.2021.102239

Eastwood, R., and Hughes, J. (2003). Molecular phylogeny and evolutionary biology of *Acrodipsas* (Lepidoptera: Lycaenidae). *Mol. Phylogenet. Evol.* 27, 93–102. doi: 10.1016/s1055-7903(02)00370-6

Edgar, R. C. (2010). Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* 26, 2460–2461. doi: 10.1093/bioinformatics/btq461

Eserman, L. A., Thomas, S. K., Coffey, E. E. D., and Leebens-Mack, J. H. (2021). Target sequence capture in orchids: developing a kit to sequence hundreds of single-copy loci. *Appl. Plant Sci.* 9:e11416. doi: 10.1002/aps3.11416

Espeland, M., Breinholt, J. W., Barbosa, E. P., Casagrande, M. M., Huertas, B., Lamas, G., et al. (2019). Four hundred shades of brown: higher level phylogeny of the problematic Euptychiina (Lepidoptera, Nymphalidae, Satyrinae) based on hybrid enrichment data. *Mol. Phyloget. Evol.* 131, 116–124. doi: 10.1016/j.ympev.2018.10.039

Espeland, M., Breinholt, J., Willmott, K. R., Warren, A. D., Vila, R., Toussaint, E. F. A., et al. (2018). A comprehensive and dated phylogenomic analysis of butterflies. *Curr. Biol.* 28, 770–778.e5. doi: 10.1016/j.cub.2018.01.061

Evans, J. D., Schwarz, R. S., Chen, Y. P., Budge, G., Cornman, R. S., De La Rua, P., et al. (2013). Standard methods for molecular research in *Apis mellifera*. *J. Apic. Res.* 52, 1–54. doi: 10.3896/ibra.1.52.4.11

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

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

Faircloth, B. C., McCormack, J. E., Crawford, N. G., Harvey, M. G., Brumfield, R. T., and Glenn, T. C. (2012). Ultraconserved elements anchor thousands of genetic markers spanning multiple evolutionary timescales. *Syst. Biol.* 61, 717–726. doi: 10.1093/sysbio/sys004

Garg, K. M., Chattopadhyay, B., Cros, E., Tomassi, S., Benedick, S., Edwards, D. P., et al. (2022). Island biogeography revisited: museomics reveals affinities of shelf island birds determined by bathymetry and paleo-rivers, not by distance to mainland. *Mol. Biol. Evol.* 39:msab340. doi: 10.1093/molbev/msab340

Gnirke, A., Melnikov, A., Maguire, J., Rogov, P., Leproust, E. M., Brockman, W., et al. (2009). Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. *Nat. Biotechnol.* 27, 182–189. doi: 10.1038/nbt.1523

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

Graham, C. F., Glenn, T. C., McArthur, A. G., Boreham, D. R., Kieran, T., Lance, S., et al. (2015). Impacts of degraded DNA on restriction enzyme associated DNA sequencing (RADSeq). *Mol. Ecol. Resour.* 15, 1304–1315. doi: 10.1111/1755-0998.12404

Grover, C. E., Salmon, A., and Wendel, J. F. (2012). Targeted sequence capture as a powerful tool for evolutionary analysis. *Am. J. Bot.* 99, 312–319. doi: 10.3732/ajb.1100323

Hamilton, C. A., Lemmon, A. R., Lemmon, E. M., and Bond, J. E. (2016). Expanding anchored hybrid enrichment to resolve both deep and shallow relationships within the spider tree of life. *BMC Evol. Biol.* 16:212. doi: 10.1186/s12862-016-0769-y

Hamilton, C. A., St Laurent, R. A., Dexter, K., Kitching, I. J., Breinholt, J. W., Zwick, A., et al. (2019). Phylogenomics resolves major relationships and reveals significant diversification rate shifts in the evolution of silk moths and relatives. *BMC Evol. Biol.* 19:182. doi: 10.1186/s12862-019-1505-1

Harvey, M. G., Smith, B. T., Glenn, T. C., Faircloth, B. C., and Brumfield, R. T. (2016). Sequence capture versus restriction site associated DNA sequencing for shallow systematics. *Syst. Biol.* 65, 910–924. doi: 10.1093/sysbio/syw036

Hime, P. M., Lemmon, A. R., Lemmon, E. C. M., Prendini, E., Brown, J. M., Thomson, R. C., et al. (2021). Phylogenomics reveals ancient gene tree discordance in the amphibian tree of life. *Syst. Biol.* 70, 49–66. doi: 10.1093/sysbio/syaa034

Hoffberg, S. L., Kieran, T. J., Catchen, J. M., Devault, A., Faircloth, B. C., Mauricio, R., et al. (2016). RADcap: sequence capture of dual-digest RADseq libraries with identifiable duplicates and reduced missing data. *Mol. Ecol. Resour.* 16, 1264–1278. doi: 10.1111/1755-0998.12566

Homziak, N. T., Breinholt, J. W., Branham, M. A., Storer, C. G., and Kawahara, A. Y. (2019). Anchored hybrid enrichment phylogenomics resolves the backbone of erebine moths. *Mol. Phylogenet. Evol.* 131, 99–105. doi: 10.1016/j.ympev.2018.10.038

Huang, H., He, Q., Kubatko, L. S., and Knowles, L. L. (2010). Sources of error inherent in species-tree estimation: impact of mutational and coalescent effects on accuracy and implications for choosing among different methods. *Syst. Biol.* 59, 573–583. doi: 10.1093/sysbio/syq047

Jetz, W., Thomas, G. H., Joy, J. B., Hartmann, K., and Mooers, A. O. (2012). The global diversity of birds in space and time. *Nature* 491, 444–448. doi: 10.1038/nature11631

Johnson, M. G., Gardner, E. M., Liu, Y., Medina, R., Goffinet, B., Shaw, A. J., et al. (2016). HybPiper: extracting coding sequence and introns for phylogenetics from high-throughput sequencing reads using target enrichment. *Appl. Plant Sci.* 4:1600016. doi: 10.3732/apps.1600016

Johnson, M. G., Pokorny, L., Dodsworth, S., Botigué, L. R., Cowan, R. S., Devault, A., et al. (2019). A universal probe set for targeted sequencing of 353 nuclear genes from any flowering plant designed using k-medoids clustering. *Syst. Biol.* 68, 594–606. doi: 10.1093/sysbio/syy086

Jones, M. R., and Good, J. M. (2016). Targeted capture in evolutionary and ecological genomics. *Mol. Ecol.* 25, 185–202. doi: 10.1111/mec.13304

Kassambara, A. (2020). *ggpubr: 'ggplot2' bvsed publication ready plots. R package version 0.4.0.* Available at: https://CRAN.R-project.org/package=ggpubr (Accessed May 13, 2022).

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

Kawahara, A. Y., Breinholt, J. W., Espeland, M., Storer, C., Plotkin, D., Dexter, K. M., et al. (2018). Phylogenetics of moth-like butterflies (Papilionoidea: Hedylidae) based on a new 13-locus target capture probe set. *Mol. Phylogenet. Evol.* 127, 600–605. doi: 10.1016/j.ympev.2018.06.002

Kawahara, A. Y., Plotkin, D., Espeland, M., Meusemann, K., Toussaint, E. F. A., Donath, A., et al. (2019). Phylogenomics reveals the evolutionary timing and pattern of butterflies and moths. *Proc. Natl. Acad. Sci.* 116, 22657–22663. doi: 10.1073/pnas.1907847116

Kawahara, A. Y., Storer, C., Carvalho, A. P. S., Plotkin, D. M., Condamine, F., Braga, M. P., et al. (2022). Evolution and diversification dynamics of butterflies. *BioRxiv*. 1–27. doi: 10.1101/2022.05.17.491528

King, J. R., and Porter, S. D. (2004). Recommendations on the use of alcohols for preservation of ant specimens (Hymenoptera, Formicidae). *Insect. Soc.* 51, 197–202. doi: 10.1007/s00040-003-0709-x

Knyshov, A., Gordon, E. R. L., and Weirauch, C. (2019). Cost-efficient high throughput capture of museum arthropod specimen DNA using PCR-generated baits. *Methods Ecol. Evol.* 10, 841–852. doi: 10.1111/2041-210x.13169

Lamas, G. (2015). Catalog of the butterflies (Papilionoidea). Available from the author.

Lang, P. L. M., Weiß, C. L., Kersten, S., Latorre, S. M., Nagel, S., Nickel, B., et al. (2020). Hybridization ddRAD-sequencing for population genomics of non-model plants using highly degraded historical specimen DNA. *Mol. Ecol. Resour.* 20, 1228–1247. doi: 10.1111/1755-0998.13168

Lawniczak, M. K. N., Durbin, R., Flicek, P., Lindblad-Toh, K., Wei, X., Archibald, J. M., et al. (2022). Standards recommendations for the earth BioGenome project. *Proc. Natl. Acad. Sci.* 119:e2115639118. doi: 10.1073/pnas.2115639118

Leaché, A. D., and Rannala, B. (2011). The accuracy of species tree estimation under simulation: a comparison of methods. *Syst. Biol.* 60, 126–137. doi: 10.1093/sysbio/syq073

Lemmon, A. R., Emme, S. A., and Lemmon, E. M. (2012). Anchored hybrid enrichment for massively high-throughput phylogenomics. *Syst. Biol.* 61, 727–744. doi: 10.1093/sysbio/sys049

Lohman, D. J., Peggie, D., Pierce, N. E., and Meier, R. (2008). Phylogeography and genetic diversity of a widespread Old World butterfly, *Lampides boeticus* (Lepidoptera: Lycaenidae). *BMC Evol. Biol.* 8:301. doi: 10.1186/1471-2148-8-301

Lou, R. N., Jacobs, A., Wilder, A. P., and Therkildsen, N. O. (2021). A beginner's guide to low-coverage whole genome sequencing for population genomics. *Mol. Ecol.* 30, 5966–5993. doi: 10.1111/mec.16077

Ludes, B., Pfitzinger, H., and Mangin, P. (1993). DNA fingerprinting from tissues after variable postmortem periods. *J. Forensic Sci.* 38, 686–690. doi: 10.1520/JFS13456J

Ma, L., Zhang, Y., Lohman, D. J., Wahlberg, N., Ma, F., Nylin, S., et al. (2020). A phylogenomic tree inferred with an inexpensive PCR-generated probe kit resolves higher-level relationships among *Neptis* butterflies (Nymphalidae: Limenitidinae). *Syst. Entomol.* 45, 924–934. doi: 10.1111/syen.12435

Makowski, D., Ben-Shachar, M., Patil, I., and Lüdecke, D. (2020). Methods and algorithms for correlation analysis in R. *J. Open Source Softw.* 5:2306. doi: 10.21105/joss.02306

Maricic, T., Whitten, M., and Pääbo, S. (2010). Multiplexed DNA sequence capture of mitochondrial genomes using PCR products. *PLoS One* 5:e14004. doi: 10.1371/journal.pone.0014004

Mayer, C., Dietz, L., Call, E., Kukowka, S., Martin, S., and Espeland, M. (2021). Adding leaves to the Lepidoptera tree: capturing hundreds of nuclear genes from old museum specimens. *Syst. Entomol.* 46, 649–671. doi: 10.1111/syen.12481

Mayer, C., Sann, M., Donath, A., Meixner, M., Podsiadlowski, L., Peters, R. S., et al. (2016). BaitFisher: a software package for multispecies target DNA enrichment probe design. *Mol. Biol. Evol.* 33, 1875–1886. doi: 10.1093/molbev/msw056

McCormack, J. E., Faircloth, B. C., Crawford, N. G., Gowaty, P. A., Brumfield, B. T., and Glenn, T. C. (2012). Ultraconserved elements are novel phylogenomic markers that resolve placental mammal phylogeny when combined with species tree analysis. *Genome Res.* 22, 746–754. doi: 10.1101/gr.125864.111

McCormack, J. E., Tsai, W. L. E., and Faircloth, B. C. (2016). Sequence capture of ultraconserved elements from bird museum specimens. *Mol. Ecol. Resour.* 16, 1189–1203. doi: 10.1111/1755-0998.12466

McGaughran, A. (2020). Effects of sample age on data quality from targeted sequencing of museum specimens: what are we capturing in time? *BMC Genomics* 21:188. doi: 10.1186/s12864-020-6594-0

Meyer, M., and Kircher, M. (2010). Illumina sequencing library preparation for highly multiplexed target capture and sequencing. *Cold Spring Harb. Protoc.* 2010:pdb.prot5448. doi: 10.1101/pdb.prot5448

Moreau, C. S., Wray, B. D., Czekanski-Moir, J. E., and Rubin, B. E. R. (2013). DNA preservation: a test of commonly used preservatives for insects. *Invertebr. Syst.* 27, 81–86. doi: 10.1071/IS12067

Morlon, H., Parsons, T. L., and Plotkin, J. B. (2011). Reconciling molecular phylogenies with the fossil record. *Proc. Natl. Acad. Sci.* 108, 16327–16332. doi: 10.1073/pnas.1102543108

Paijmans, J. L. A., Fickel, J., Courtiol, A., Hofreiter, M., and Förster, D. W. (2015). Impact of enrichment conditions on cross-species capture of fresh and degraded DNA. *Mol. Ecol. Resour.* 16, 42–55. doi: 10.1111/1755-0998.12420

Peñalba, J. V., Smith, L. L., Tonione, M. A., Sass, C., Hykin, S. M., Skipwith, P. L., et al. (2014). Sequence capture using PCR-generated probes: a cost-effective method of targeted high-throughput sequencing for nonmodel organisms. *Mol. Ecol. Resour.* 14, 1000–1010. doi: 10.1111/1755-0998.12249

Pinkert, S., Barve, V., Guralnick, R., and Jetz, W. (2022). Global geographical and latitudinal variation in butterfly species richness captured through a comprehensive country-level occurrence database. *Glob. Ecol. Biogeogr.* 31, 830–839. doi: 10.1111/geb.13475

Post, R. J., Flook, P. K., and Millest, A. L. (1993). Methods for the preservation of insects for DNA studies. *Biochem. Syst. Ecol.* 21, 85–92. doi: 10.1016/0305-1978(93)90012-g

Prathapan, K. D., Pethiyagoda, R., Bawa, K. S., Raven, P. H., and Rajan, P. D. (2018). When the cure kills—CBD limits biodiversity research. *Science* 360, 1405–1406. doi: 10.1126/science.aat9844

Prendini, L., Hanner, R., and Desalle, R. (2002). "Obtaining, storing and archiving specimens and tissue samples for use in molecular studies," in *Techniques in molecular systematics and evolution* (Basel: Birkhäuser), 176–248.

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

Pyle, M. M., and Adams, R. P. (1989). *In situ* preservation of DNA in plant specimens. *Taxon* 38, 576–581. doi: 10.2307/1222632

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

R Core Team (2021). R: A language and environment for statistical computing. R foundation for statistical computing, Vienna, Austria. Available at: https://www.R-project.org (Accessed May 13, 2022).

RStudio Team (2020). *RStudio: Integrated development for R. PBC*, Boston, MA: RStudio. Available at: http://www.rstudio.com/. (Accessed May 13, 2022).

Rabinowitz, D. (1981). "Seven forms of rarity," in *The biological aspects of rare plant conservation*. ed. H. Synge (Chichester: John Wiley & Sons Ltd.), 205–217.

Raxworthy, C. J., and Smith, B. T. (2021). Mining museums for historical DNA: advances and challenges in museomics. *Trends Ecol. Evol.* 36, 1049–1060. doi: 10.1016/j.tree.2021.07.009

Ribeiro, P., Torres Jiménez, M. F., Andermann, T., Antonelli, A., Bacon, C. D., and Matos-Maraví, P. (2021). A bioinformatic platform to integrate target capture and whole genome sequences of various read depths for phylogenomics. *Mol. Ecol.* 30, 6021–6035. doi: 10.1111/mec.16240

Rubin, B. E. R., Ree, R. H., and Moreau, C. S. (2012). Inferring phylogenies from RAD sequence data. *PLoS One* 7:e33394. doi: 10.1371/journal.pone.0033394

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

St Laurent, R. A., Hamilton, C. A., and Kawahara, A. Y. (2018). Museum specimens provide phylogenomic data to resolve relationships of sack-bearer moths (Lepidoptera, Mimallonoidea, Mimallonidae). *Syst. Entomol.* 43, 729–761. doi: 10.1111/syen.12301

Staats, M., Erkens, R. H. J., van de Vossenberg, B., Wieringa, J. J., Kraaijeveld, K., Stielow, B., et al. (2013). Genomic treasure troves: complete genome sequencing of herbarium and insect museum specimens. *PLoS One* 8:e69189. doi: 10.1371/journal.pone.0069189

Suchan, T., Pitteloud, C., Gerasimova, N. S., Kostikova, A., Schmid, S., Arrigo, N., et al. (2016). Hybridization capture using RAD probes (hyRAD), a new tool for performing genomic analyses on collection specimens. *PLoS One* 11:e0151651. doi: 10.1371/journal.pone.0151651

Talavera, G., Lukhtanov, V., Pierce, N. E., and Vila, R. (2021). DNA barcodes combined with multi-locus data of representative taxa can generate reliable higher-level phylogenies. *Syst. Biol.* 71, 382–395. doi: 10.1093/sysbio/syab038

Tin, M. M.-Y., Economo, E. P., and Mikheyev, A. S. (2014). Sequencing degraded DNA from non-destructively sampled museum specimens for RAD-tagging and low-coverage shotgun phylogenetics. *PLoS One* 9:e96793. doi: 10.1371/journal.pone.0096793

Toussaint, E. F. A., Breinholt, J. W., Earl, C., Warren, A. D., Brower, A. V. Z., Yago, M., et al. (2018). Anchored phylogenomics illuminates the skipper butterfly tree of life. *BMC Evol. Biol.* 18:101. doi: 10.1186/s12862-018-1216-z

Toussaint, E. F. A., Chiba, H., Yago, M., Dexter, K. M., Warren, A. D., Storer, C., et al. (2021a). Afrotropics on the wing: phylogenomics and historical biogeography of awl and policeman skippers. *Syst. Entomol.* 46, 172–185. doi: 10.1111/syen.12455

Toussaint, E. F. A., Ellis, E. A., Gott, R. J., Warren, A. D., Dexter, K. M., Storer, C., et al. (2021b). Historical biogeography of Heteropterinae skippers via Beringian and post-Tethyan corridors. *Zool. Scr.* 50, 100–111. doi: 10.1111/zsc.12457

Toussaint, E. F. A., Gauthier, J., Bilat, J., Gillett, C. P. D. T., Gough, H. M., Lundkvist, H., et al. (2021c). HyRAD-X exome capture museomics unravels giant ground beetle evolution. *Genome Biol. Evol.* 13, 13:evab112. doi: 10.1093/gbe/evab112

Toussaint, E. F. A., Vila, R., Yago, M., Chiba, H., Warren, A. D., Aduse-Poku, K., et al. (2019). Out-of-orient: post-Tethyan transoceanic and trans-Arabian routes fostered the spread of Baorini skippers in the Afrotropics. *Syst. Entomol.* 44, 926–938. doi: 10.1111/syen.12365

Townsend, J. P., and Leuenberger, C. (2011). Taxon sampling and the optimal rates of evolution for phylogenetic inference. *Syst. Biol.* 60, 358–365. doi: 10.1093/sysbio/syq097

Valencia-Montoya, W. A., Quental, T. B., Tonini, J. F. R., Talavera, G., Crall, J. D., Lamas, G., et al. (2021). Evolutionary trade-offs between male secondary sexual traits revealed by a phylogeny of the hyperdiverse tribe Eumaeini (Lepidoptera: Lycaenidae). *Proc. R. Soc. B* 288:20202512. doi: 10.1098/rspb.2020.2512

Venables, W.N., and Ripley, B.D. (2002). *Modern applied statistics with S*. New York: Springer, doi: 10.1007/978-0-387-21706-2.

Wahlberg, N., and Wheat, C. W. (2008). Genomic outposts serve the phylogenomic pioneers: designing novel nuclear markers for genomic DNA extractions of Lepidoptera. *Syst. Biol.* 57, 231–242. doi: 10.1080/10635150802033006

Wells, A., Johanson, K. A., and Dostine, P. (2019). Why are so many species based on a single specimen? *Zoosymposia* 14, 32–38. doi: 10.11646/zoosymposia.14.1.5

Whibley, A., Kelley, J., and Narum, S. (2021). The changing face of genome assemblies: guidance on achieving high-quality reference genomes. *Mol. Ecol. Resour.* 21, 641–652. doi: 10.1111/1755-0998.13312

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. New York: Springer-Verlag. ISBN 978-3-319-24277-4. Available at: https://ggplot2.tidyverse.org. (Accessed May 13, 2022).

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., et al. (2019). Welcome to the tidyverse. *J. Open Source Softw.* 4:1686. doi: 10.21105/joss.01686

Wood, H. M., González, V. L., Lloyd, M., Coddington, J., and Scharff, N. (2018). Next-generation museum genomics: phylogenetic relationships among palpimanoid spiders using sequence capture techniques (Araneae: Palpimanoidea). *Mol. Phylogenet. Evol.* 127, 907–918. doi: 10.1016/j.ympev.2018.06.038

Zhang, Y., Deng, S., Liang, D., and Zhang, P. (2019a). Sequence capture across large phylogenetic scales by using pooled PCR-generated baits: a case study of Lepidoptera. *Mol. Ecol. Resour.* 19, 1037–1051. doi: 10.1111/1755-0998.13026

Zhang, F., Ding, Y., Zhu, C.-D., Zhou, X., Orr, M. C., Scheu, S., et al. (2019b). Phylogenomics from low-coverage whole-genome sequencing. *Methods Ecol. Evol.* 10, 507–517. doi: 10.1111/2041-210X.13145

Zhang, Y., Huang, S., Liang, D., Wang, H., and Zhang, P. (2020). A multilocus analysis of Epicopeiidae (Lepidoptera, Geometroidea) provides new insights into their relationships and the evolutionary history of mimicry. *Mol. Phylogenet. Evol.* 149:106847. doi: 10.1016/j.ympev.2020.106847

Zwickl, D. J., and Hillis, D. M. (2002). Increased taxon sampling greatly reduces phylogenetic error. *Syst. Biol.* 51, 588–598. doi: 10.1080/10635150290102339

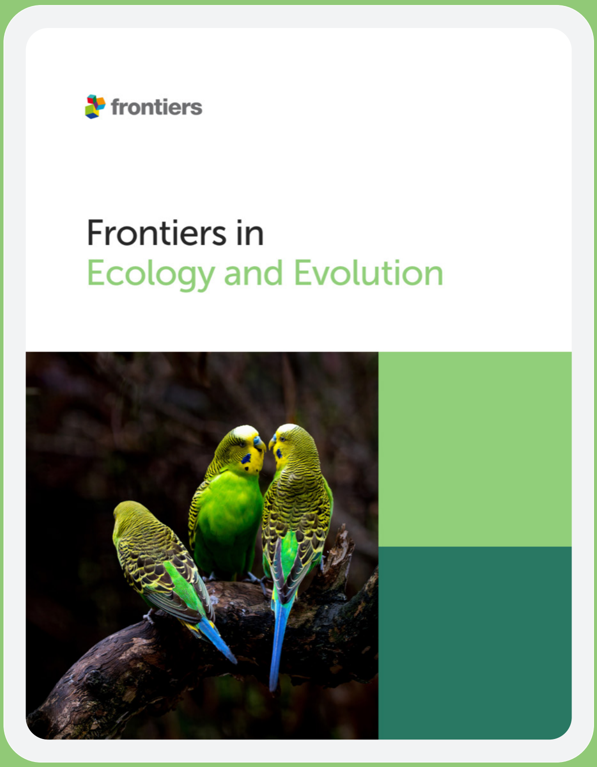
Frontiers in
Ecology and Evolution

Ecological and evolutionary research into our natural and anthropogenic world

This multidisciplinary journal covers the spectrum of ecological and evolutionary inquiry. It provides insights into our natural and anthropogenic world, and how it can best be managed.

Discover the latest Research Topics

See more →



Frontiers

Avenue du Tribunal-Fédéral 34
1005 Lausanne, Switzerland
frontiersin.org

Contact us

+41 (0)21 510 17 00
frontiersin.org/about/contact