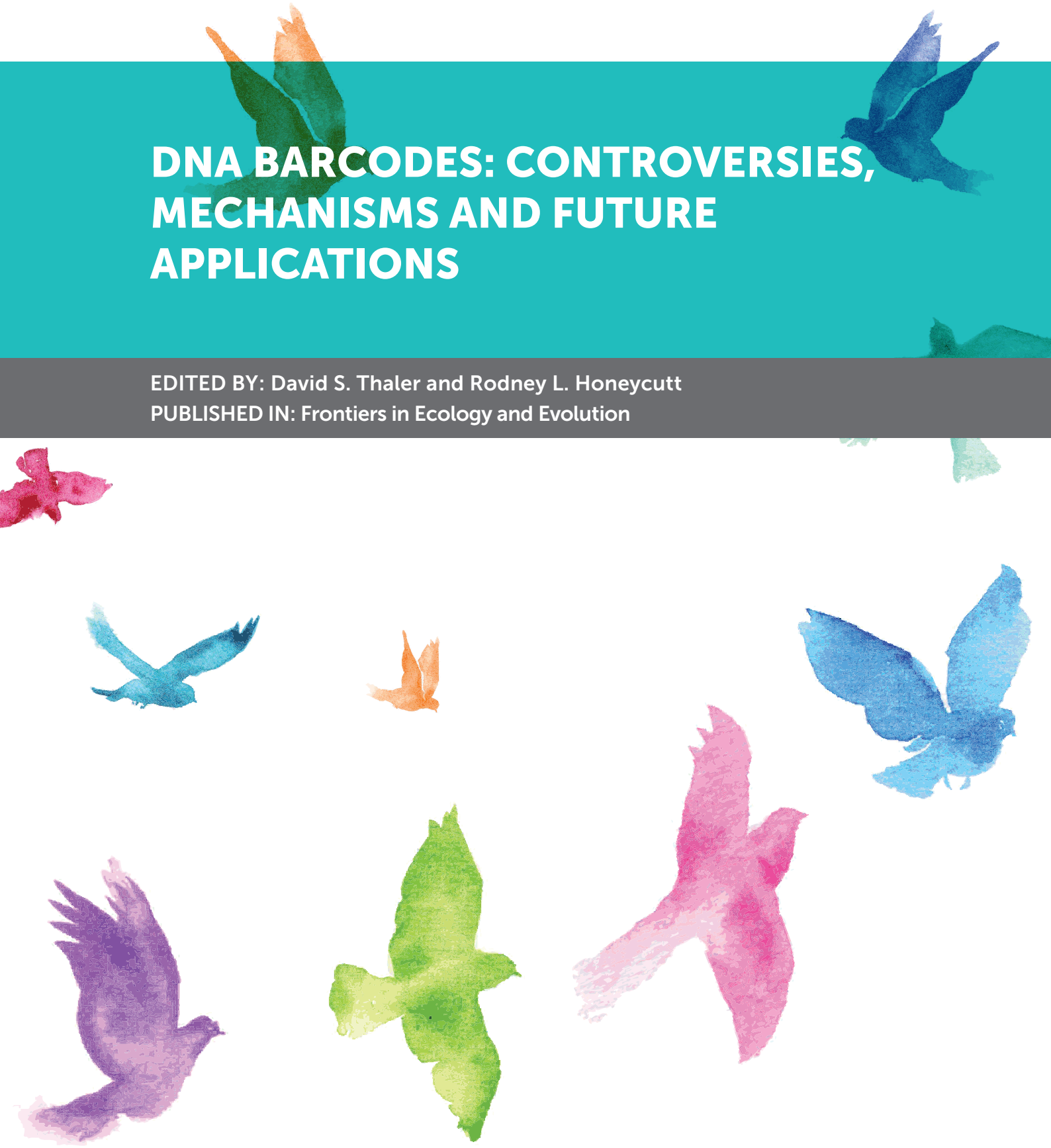
DNA BARCODES: CONTROVERSIES, MECHANISMS AND FUTURE APPLICATIONS

EDITED BY: David S. Thaler and Rodney L. Honeycutt
PUBLISHED IN: Frontiers in Ecology and Evolution



frontiers Research Topics

frontiers

ISSN 1664-8714
ISBN 978-2-88971-290-8
DOI 10.3389/978-2-88971-290-8

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

DNA BARCODES: CONTROVERSIES, MECHANISMS AND FUTURE APPLICATIONS

Topic Editors:
David S. Thaler, University of Basel, Switzerland
Rodney L. Honeycutt, Pepperdine University, United States

Citation: Thaler, D. S., Honeycutt, R. L., eds. (2021). DNA Barcodes: Controversies, Mechanisms and Future Applications. Lausanne: Frontiers Media SA.
doi: 10.3389/978-2-88971-290-8

Table of Contents

frontiers in Ecology and Evolution

EDITORIAL
published: 02 July 2021
doi: 10.3389/fevo.2021.718865

Editorial: DNA Barcodes: Controversies, Mechanisms, and Future Applications

*Rodney L. Honeycutt**

Natural Science Division, Pepperdine University, Malibu, CA, United States

Keywords: DNA barcodes, metabarcoding, taxonomy, biosecurity and regulation, microbial diversity, selection, mtDNA

Editorial on the Research Topic

DNA Barcodes: Controversies, Mechanisms, and Future Applications

Biodiversity provides ecosystem services and direct and indirect benefits to society. Unfortunately, human activities are now accelerating the extinction rate of biodiversity at an alarming rate, and in many cases, species will disappear before their discovery. Except for vertebrates and plants, knowledge about the number of species for many groups of organisms and the biogeographic regions harboring high levels of biodiversity is lacking (Honeycutt et al., 2010). Traditional taxonomy alone cannot achieve an all-species inventory, but the integration of conventional taxonomy, DNA-based technology, and bioinformatics increases the feasibility of filling in the knowledge gaps.

In 2003, Hebert et al. (2003a,b) proposed using an ∼650 bp sequence from the mitochondrial cytochrome c oxidase subunit I gene (COI) as a valuable barcode for identifying species in the kingdom Animalia. Given the rate of change and selection patterns, mitochondrial genes like COI demonstrate patterns of change conducive to their use as DNA barcodes (Shtolz and Mishmar). BOLD (Barcode of Life Data System) is a web-based reference system of COI sequences developed to allow species-level identification (Ratnasingham and Hebert, 2007). BOLD is now international and is cataloging sequences of species at a rapid rate.

What are some trends in the use of DNA barcodes? DeSalle and Goldstein present a summary of DNA barcoding papers published over the last 15 years. Throughout this period, the primary focus has been alpha taxonomy (the identification and delimitation of species) and the discovery of cryptic species not easily diagnosed by morphology. A parallel survey by Yang et al. indicates that in China, the primary use of species identification emphasized food safety, control of pests and invasive species, and traditional medicine. More recently, researchers in China are also using barcodes to discover cryptic species and create biodiversity inventories. Given continued global threats to biodiversity, species discovery is likely to remain a primary use for barcodes.

The concept of DNA barcoding has revolutionized fields of science interested in inventorying biodiversity (Jansen and Hallwachs, 2016), ecological studies of species interactions and community structure (Valentini et al., 2008; Joly et al., 2013), conservation biology (Shapcott et al., 2015), assessment of biosecurity risks from invasive species (Molnar et al., 2008; Madden et al., 2019), and forensics (Mwale et al., 2017).

Several papers in this series highlight the use of barcodes related to food safety and biosecurity in marine ecosystems. Silva et al. examined the accuracy of mini-barcodes to identify mixed species of fish included in processed fish balls and cakes. One finding was that not all species are equally identifiable. As a result, these authors recommend cloning of PCR products and

Edited and reviewed by:
Mark A. Elgar,
The University of Melbourne, Australia

***Correspondence:**
Rodney L. Honeycutt
rodney.honeycutt@pepperdine.edu

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

Received: *01 June 2021*
Accepted: *11 June 2021*
Published: *02 July 2021*

Citation:
Honeycutt RL (2021) Editorial: DNA Barcodes: Controversies, Mechanisms, and Future Applications. Front. Ecol. Evol. 9:718865. doi: 10.3389/fevo.2021.718865

next-generation sequencing. Suarez-Menendez et al. describe the use of eDNA (environmental DNA), extracted from 6 L of water, Illumina sequencing, and metabarcoding to identify invasive species and indicators of loss of environmental quality in coastal lagoons of the Mediterranean. This approach proved helpful in identifying habitats threatened by loss of environmental quality. Finally, Garcia-Vazquez et al. used a barcode approach for establishing biotic surveys of ports vulnerable to the importation of invasive or alien species. These ports showed differences in their susceptibility to infiltration of non-native species, and the authors offer several explanations for these differences.

Technological advances allow for faster acquisition of sequences at less cost (Hebert et al., 2018; Knot et al., 2020). Multiplex sequencing with Illumina MiSeq platform is a good example of such advancements. Using dried museum specimens of saproxylic beetles and DNA barcodes, Sire et al. compare the effectiveness of both Illumina sequencing and traditional Sanger sequencing. Recovery of barcode sequences was similar for the two methods, with the cost per sequence considerably less for the Illumina method. The paper by Porter and Hajibabaei published in this series provides an overview of current methods and implications for species identification in challenging groups. They present a review of the application of metabarcoding for sampling whole communities. Additionally, they discuss the bioinformatic approach needed to process metabarcode data.

COI is not the only useful barcode marker. Chloroplast and nuclear genes represent barcode markers for plant species identification (Kress, 2017), ribosomal ITS for fungi (Lücking et al., 2020), 18S rDNA, 28S rDNA, and ITS for protists (Pawlowski et al., 2012), and 16S rDNA for bacteria and archaebacteria (Lebonah et al., 2014). In each of these cases, databases exist, which is a requisite for any barcode marker. In this series, Pierce discusses the limitations of using only COI and argues for the inclusion of ultraconserved elements (UCEs) from the nuclear genome, which he suggests would strengthen species identification across divergent groups of taxa. Unfortunately, one constraint of UCEs is the lack of an adequate database.

Is DNA barcoding a challenge to systematic biology? Taxonomy is a component of systematics that emphasizes identification, delimitation, and description of species. The importance of taxonomy is evident from the specimens collected over the centuries and housed in natural history museums. These specimens serve as a valuable resource for those employing DNA barcode technology for species identification and the establishment of databases. Therefore, natural history museums and expert taxonomists are essential for cataloging biodiversity (Pinheiro et al., 2019). Cognato et al.'s research on ambrosia beetles demonstrates that DNA barcodes cannot replace taxonomic expertise. For some species of beetles, the authors did not observe a barcode gap, and confidence in species identification decreased with an increase in percent divergence. Phylogenetics represents another component of systematics used to determine relationships among organisms. The resultant phylogeny is useful for comparative studies and the derivation of classifications. As noted by DeSalle and Goldstein, single DNA barcode markers and the distance-based approach for identifying species are insufficient for diagnosing phylogenetic relationships. Researchers now use sequences from multiple genes and genomes to examine evolutionary divergence within and between species. An excellent example in this series is the study by Mankga et al., who used both plant barcode and nuclear gene sequences to study diversification and phylogenetic relationships of cycads.

As Thaler notes, our current knowledge of microbial biodiversity is unknown and difficult to discern with approaches used for multicellular organisms. As a result, early efforts at understanding microbial diversity emphasized the use of molecular markers (Pace, 1997). Today, more advanced methods are beginning to estimate the number of eubacteria and archaebacteria species (Louca et al., 2019). Other problems with microbial diversity noted by Thaler are the generation time or rate at which microbes (single-celled organisms and viruses) evolve, the difficulty in identifying OTUs (operational taxonomic units), and the exchange of genes between microbes via horizontal gene transfer. All these factors complicate our ability to evaluate microbial diversity in space and time.

The breadth of papers presented in this series and the plethora of DNA barcode papers published each year indicate that barcoding will continue to be an important tool for realizing an all species inventory. Advancements in genomics and bioinformatics continue to be developed, and these advancements offer accessibility of these tools to a broad range of researchers interested in the application of barcodes.

AUTHOR CONTRIBUTIONS

The author confirms being the sole contributor of this work and has approved it for publication.

ACKNOWLEDGMENTS

The author would like to acknowledge David Thaler for his contributions as topic editor to this Research Topic. I wish to thank the authors and reviewers who participated in this series on DNA barcodes.

REFERENCES

Hebert, P. D. N., Braukmann, T. W. A., Prosser, S. W. J., Ratnasingham, S., deWaard, J. R., Ivanova, N. V., et al. (2018). A sequel to sanger: amplicon sequencing that scales. *BMC Genomics* 19:219. doi: 10.1186/s12864-018-4611-3

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

Hebert, P. D. N., Ratnasingham, S., and deWaard, J. R. (2003b). Barcoding animal life: cytochrome c oxidase subunit 1 divergences among closely related species. *Proc. R. Soc. Lond. B* 270(Suppl.), S96–S99. doi: 10.1098/rsbl.2003.0025

Honeycutt, R. L., Hillis, D. M., and Bickham, J. W. (2010). "Biodiversity discovery and its importance to conservation," in *Molecular Approaches in Natural Resource Conservation and Management*, eds J. A. DeWoody, J. W. Bickham, C. H. Michler, K. M. Nichols, G. E. Rhodes, and K. E. Woeste (Cambridge:

Cambridge University Press), 1–35. doi: 10.1017/CBO9780511777592.002

Jansen, D. H., and Hallwachs, W. (2016). DNA barcoding the Lepidoptera inventory of a large complex tropical conserved wildland, Area de Conservacion Guanacaste, northwestern Costa Rica. *Genome* 59, 641–660. doi: 10.1139/gen-2016-0005

Joly, S., Davies, T. J., Archambault, A., Bruneau, A., Derry, A., Kembel, S. W., et al. (2013). Ecology in the age of DNA barcoding: the resource, the promise and the challenges ahead. *Mol. Ecol. Resour.* 14, 221–232. doi: 10.1111/1755-0998.12173

Knot, I. E., Zouganelis, G. D., Weedall, G. D., Wich, S. A., and Rae, R. (2020). DNA barcoding of nematodes using MinION. *Front. Ecol. Evol.* 8:100. doi: 10.3389/fevo.2020.00100

Kress, W. J. (2017). Plant DNA barcodes: applications today and in the future. *J. Sys. Evol.* 55, 291–307. doi: 10.1111/jse.12254

Lebonah, D. E., Dileep, A., Chandrasekhar, K., Sreevani, S., Sreedevi, B., and, J. P., et al. (2014). DNA barcoding on bacteria: a review. *Adv. Biol.* 2014:541787. doi: 10.1155/2014/541787

Louca, S., Mazel, F., Doebeli, M., and Parfrey, L. W. (2019). A census-based estimate of Earth's bacterial and archaeal diversity. *PLoS Biol.* 17:e3000106. doi: 10.1371/journal.pbio.3000106

Lücking, R., Aime, M. C., Robbertse, B., Miller, A. N., Ariyawansa, H. A., Aoki, T., et al. (2020). Unambiguous identification of fungi: where do we stand and how accurate and precise is fungal DNA barcoding? *IMA Fungus* 11:14. doi: 10.1186/s43008-020-00033-z

Madden, M. J. L., Young, R. G., Brown, J. W., Miller, S. E., Frewin, A. J., and Hanner, R. H. (2019). Using DNA barcoding to improve invasive pest identification at U.S. ports-of-entry. *PLoS ONE* 14:e0222291. doi: 10.1371/journal.pone.0222291

Molnar, J. L., Gamboa, R. L., Revenga, C., and Spalding, M. D. (2008). Assessing the global threat of invasive species to marine biodiversity. *Front. Ecol. Environ.* 6, 485–492. doi: 10.1890/070064

Mwale, M., Dalton, D. L., Jansen, R., De Bruyn, M., Pietersen, D., Mokgokong, P. S., et al. (2017). Forensic application of DNA barcoding for identification of illegally traded African pangolin scales. *Genome* 60, 272–284. doi: 10.1139/gen-2016-0144

Pace, N. R. (1997). A molecular view of microbial diversity and the biosphere. *Science* 276, 734–740. doi: 10.1126/science.276.5313.734

Pawlowski, J., Audic, S., Adl, S., Bass, D., Belbahri, L., Berney, C., et al. (2012). CBOL protist working group: barcoding eukaryotic richness beyond the animal, plant, and fungal kingdoms. *PLoS Biol.* 10:e1001419. doi: 10.1371/journal.pbio.1001419

Pinheiro, H. T., Moreau, C. S., Daly, M., and Rocha, L. A. (2019). Will DNA barcoding meet taxonomic needs? *Science* 365, 873–874. doi: 10.1126/science.aay7174

Ratnasingham, S., and Hebert, P. D. (2007). BOLD: the barcode of life data system (http://www.barcodinglife.org/) Ecol. Notes 7, 355–364. doi: 10.1111/j.1471-8286.2007.01678.x

Shapcott, A., Forster, P. I., Guymer, G. P., McDonald, W. J. F., Faith, D. P., et al. (2015). Mapping biodiversity and setting conservation priorities for SE Queensland's rainforests using DNA barcoding. *PLoS ONE* 10:e0122164. doi: 10.1371/journal.pone.0122164

Valentini, A., Pompanon, F., and Taberlet, P. (2008). DNA barcoding for ecologists. *Trends Ecol. Evol.* 24, 110–117. doi: 10.1016/j.tree.2008.09.011

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

frontiers in Ecology and Evolution

SYSTEMATIC REVIEW
published: 10 September 2019
doi: 10.3389/fevo.2019.00302

Review and Interpretation of Trends in DNA Barcoding

Rob DeSalle[1]* *and Paul Goldstein*[2]

[1] Sackler Institute for Comparative Genomics, American Museum of Natural History, New York, NY, United States,
[2] Systematic Entomology Laboratory, USDA, National Museum of Natural History, Washington, DC, United States

Interpretations and analytical practices surrounding DNA barcoding are examined using a compilation of 3,756 papers (as of December 31, 2018) with "DNA Barcode" in the abstract published since 2004. By examining the rise of DNA barcoding in natural history and biodiversity science over this period, we hope to detect the extent to which its purposes, premises, rationale and application have evolved. The number of studies involving identification, taxonomic decisions and the discovery of cryptic species has grown rapidly and appears to have driven much of the publication activity of DNA barcode studies overall. Forensic studies and papers on biological conservation involving DNA barcodes have loosely tracked the ensemble number of studies but appear to have risen sharply in 2017. Although analytical paradigms have diversified, particularly following the growing availability of tools in BoLD, neighbor-joining and graphic (tree-based) criteria for species delimitation remain preeminent. We conclude that the practices and paradigms of DNA barcoding data are likely to persist and, in groups such as Lepidoptera, remain a widely used tool in taxonomic science.

Keywords: DNA barcode, phylogenetics, diagnosis, species delimitation, specimen identification

OPEN ACCESS

Edited by:
David S. Thaler,
Universität Basel, Switzerland

Reviewed by:
Mark Stoeckle,
The Rockefeller University, United States
Rodney L. Honeycutt,
Pepperdine University, United States

***Correspondence:**
Rob DeSalle
desalle@amnh.org

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

Received: *15 March 2019*
Accepted: *26 July 2019*
Published: *10 September 2019*

Citation:
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

The doing is often more important than the outcome.

—Arthur Ashe

INTRODUCTION

Widely heralded as a revolutionary taxonomic discovery tool, DNA barcoding represents perhaps the most reliable framework available for organizing specimens and specimen-based data for systematic research. Arranging specimens by barcode haplotype early in the study process allows for efficient inspection of material, and facilitates the organization and management of a wealth of character data and life history information, depending on how much is available for the barcoded specimens. While DNA sequences have been used to identify specimens or parts of specimens since the 1980's, their use as a broader natural history tool was not formalized until 2003. Three organizational meetings sponsored by the Sloan Foundation at the Banbury Center at Cold Spring Harbor and seminal publications that year (Hebert et al., 2003a,b; Stoeckle, 2003) christened DNA barcoding and launched the program that would globalize its application. Since then, over 3,700 peer-reviewed papers have been published with "DNA barcoding" in their title. These studies range from taxonomic works in which DNA barcodes are used to elucidate cryptic species, to surveys of environmental samples (e.g., marine sediments, ocean water) that feature estimates of phyletic diversity and regional comparisons of genetic variation, and finally to forensic and conservation applications. Many of the early papers can be characterized as proof-of-concept studies in which

the utility of the COI barcoding region was being tested for particular taxonomic groups or in different study designs. To the extent controversy emerged around barcode data, it was generally associated with the taxonomic interpretation and applicability of their analyses. These included the uniformity and generalizability of criteria for circumscribing species, the phylogenetic implications of dendrograms, and the proliferation of informal specific epithets in reference to species that were discovered through DNA barcodes but which remained undescribed. Many of these concerns were mitigated by increasingly sophisticated treatments that incorporated barcodes with morphological, behavioral and ecological data under the rubric of integrative taxonomy and, for groups such as Lepidoptera in which extensive taxonomic coverage has been achieved (Hajibabaei et al., 2006; Hausmann et al., 2016; Zahiri et al., 2017), barcode data have become commonplace if not critical to taxonomic revisionary works.

As a paradigm, DNA barcoding engendered a democratization of molecular data (or at least metadata) by automating analytical steps that might otherwise have deterred may some practicing taxonomists. This quickened the pace of alpha taxonomy by enabling the rapid and unambiguous discovery of new species in many groups. One possible drawback has been that in coopting the terminology of phylogenetics, DNA barcode endeavors may have inadvertently broadened the meaning of or even re-branded terminology in a manner inconsistent with its formal interpretation. Taxonomic papers incorporating DNA barcode data routinely present metrics or tree graphics as self-evident while conflating descriptions with diagnoses or barcode trees with phylogenies. Semantics aside, we wished to understand whether such usage reflected a manifestation of some trend in how systematics is perceived by the scientific community at large.

The rapid growth of the DNA barcode paradigm thus invites an examination of how, during a 15-year period, its ontology and application developed with respect to technological, analytical, and terminological preferences that had until only recently fallen exclusively within the purview of molecular systematists. Our purpose here is to examine the development of DNA barcoding through a coarse examination of search terms and explore whether they reflect trends in how DNA barcoding practices may have evolved to accommodate analytical and practical considerations. To the extent they have not, we highlight those considerations at the empirical intersection of DNA barcoding, taxonomy and phylogenetics that are not simply semantic.

A CONCEPTUAL FRAMEWORK FOR EXAMINING THE ONTOLOGY OF DNA BARCODING

For clarity and transparency both, it is necessary to establish a conceptual framework on which to arrange this discussion. DNA barcoding intersects with systematics most conspicuously at the level of alpha taxonomy, that is in the discovery, diagnosis, and description of new species. "Description" and "diagnosis" are formal terms defined in nomenclatural codes (e.g., ICZN) that govern the naming of species and other taxa and the means of tracking and stabilizing taxonomic nomenclature. They represent components of taxonomic refinement and formalized nomenclatural change, and correspond to the character-based empirical work of substantiating named groups as historical or natural entities. It is generally understood that taxonomic rank does not of itself confer natural comparability: Any rank above species is a function of convention and discretion as well as actual data, and as long as monophyletic groups are recognized the fact that families or tribes are not uniformly or evolutionarily equivalent does not hamper studies unless they make the mistake of treating such groups, e.g., by inferring evolutionary trends from numbers of genera, families, etc. A named species, on the other hand, is a different sort of construct that may correspond to a range of biological entities consistent with historical, reproductive, or genetic criteria. Biological or historical comparability is perhaps more easily justified for species than for higher taxa because their identity as species can at least be tested by universal criteria, namely the establishment of diagnostic characters. At supra-specific taxonomic levels, in contrast, common ancestry is depicted hierarchically and articulated with reference to apomorphy, and independently derived diagnostic characters recognized as synapomorphies provide evidence both for a given species' inclusion in a given group and for that group's monophyly.

However, the usage of monophyly has been broadened to include its graphic depiction on trees, just as the traditional use of "phylogeny" as an abstract term for evolutionary history has been expanded and pluralized to include any tree-like graphics ("phylogenies"). At least one general consequence of this usage bears directly on the practice of DNA barcoding: the perception that species be legitimately represented and expected to appear as monophyletic. Whether one disputes this on the grounds that individual organisms are not related hierarchically even if mitochondria are (Doyle, 1995), or on the grounds that species often appear paraphyletic (Funk and Omland, 2003), the disconnection between the graphic representation of a monophyletic group and the characters underlying it is amplified when trees are treated as arbiters of species boundaries. When phylogenetics began to enjoy popularity, it was because there was consensus that empirical phylogenetic considerations were important to classification and evolutionary biology, but there remained strong methodological debates to the point where trees were judged less by what they said than how they were generated. The opposite experience seems to characterize DNA barcoding as a field. How barcode data—or any sequence data—are analyzed to generate trees bears directly on how those trees may be interpreted and on the scope of how DNA barcode data are ultimately used.

The ∼3,700 DNA barcoding studies published over the past 15 years represent a prodigious record of peer-reviewed research, notwithstanding the variance in their intent or in the analyses and interpretations espoused. By examining the cohort of natural history and biodiversity science that incorporated DNA barcodes over this period, we explored the extent to which their purposes, premises, rationale and application have evolved.

3756 BARCODING PAPERS SINCE 2004

We compiled a glossary of terms used in DNA barcoding from our knowledge of the literature. We attempted to be as inclusive as possible with these terms and even included some from the literature on species boundaries and, speciation mechanisms. We next used the PubMed at NCBI (https://www.ncbi.nlm.nih.gov/pubmed/) to search for peer-reviewed papers with abstracts published since 2003. We used December 31, 2018 as a cutoff for inclusion in our database. In all, we compiled the abstracts from the 3,756 peer-reviewed papers with "DNA Barcode" as a query (**Figure 1A**), and used the resulting database (**Supplementary Folder 1**) to track the usage of specific terms as described below. Perhaps naïvely, all papers retrieved by the search are assumed to have been peer-reviewed as they are included in the PubMed database. Papers were cataloged by year from 2005 to 2018 since only a few papers appeared in 2003 and 2004. Hence, we combine 2003, 2004, and 2005 into a single data point. Abstracts from each of the papers were compiled in text files by year. Word searches were done in BBedit, an efficient textline editor, that retrieves the number and location of search terms. The location of the search term hit allowed us to eliminate duplicate hits in single papers. The number of hits for each search term (or combination of terms) were compiled in excel spreadsheets. Each of the terms in the glossary (**Table 1**) were searched and tabulated. **Figure 1** provides more detail on the search strategies for the terms we used for generating graphs. For example, the raw number of hits for the general category "Neighbor Joining" was a combination of searches for "neighbor joining" plus "NJ."

An eclectic lexicon has grown around DNA barcoding, comprising a range of terms from taxonomy, phylogenetic and molecular systematics, and population genetics as well as a smattering of neologisms. The database we developed was queried for 29 terms based on our own extensive reading of the barcode literature. These terms span a range of purposes and methods, which we grouped according to (1) general disciplines (conservation/conservation biology/conservation genetics, forensic, taxonomy/systematics/integrative taxonomy, phylogeography); (2) biological terms (character, crypsis/cryptic species, fixation/fixed character, population); (3) graphic terms (clade, cluster, tree); (4) tree-building methods (Bayesian, likelihood, neighbor-joining, parsimony); (5) general purpose operational terms (diagnosis, species circumscription/delimitation/delineation, species description, species discovery, specimen identification/determination, flag); and finally (6) tools and metrics (barcode gap, BIN, BLAST, bootstrap, phylogenetic support). The queried terms comprise a combination of rudimentary verbiage commonly used in systematics and molecular evolution, with that specific to DNA barcoding. Neither their groupings nor the underlying terms are mutually exclusive, but we have tried to arrange the terms as coherently as possible. We did not account for context or whether the terms were used correctly or with approbation. In some cases, to facilitate broader comparisons we combined counts for intrinsically related terms such as similarity/distance, or terms used interchangeably such as species delimitation, circumscription and delineation. These are detailed in **Figure 1**, **Table 1**, and in **Supplementary File 1**.

Inevitably, this exercise is influenced by our own perspective which favors an integrative taxonomic approach to corroborating the results of barcode analyses with other observations. It is our impression that this perspective is reasonably widespread. In general, we prefer to think of DNA barcode variation as having the potential to reveal corroborating patterns in morphology and behavior than as necessary or sufficient requirements for discovering species or as means of generating universal distance thresholds as criteria for demarcating them. Our choice of queried terms also, therefore, reflects the distinction between indirect or tree-based interpretations that rely on inspecting dendrograms, and direct analyses of diagnostic characters. To the extent that trends may be evinced from our seemingly chimeric exploration of language, we hope that occasional inventories such as this serve to take stock of and even illuminate the direction of a field regardless of perspective.

We present the results in two ways: (1) in the form of raw counts by year to track raw usage (**Figure 1**; search terms themselves in **Supplementary File 1**) and; (2) as scaled percentages of the occurrence of all terms per year (**Supplementary File 1**). Although crude, this approach affords context for cross-comparison of year-to-year usage; we suspect more complex analysis of data such as these would simply obfuscate any observable trends.

TRENDS IN DNA BARCODING BASED ON ITS VOCABULARY

Characters, Distance Measures, and Tree-Building Functions

An important comparison concerns the use of direct character information, which corresponds to the empirical treatment of observable data, vs. lumped (phenetic) summaries in the form of similarity or distance measures. By compressing character state information into a single measure of genetic similarity, distance measures mask changes in specific loci. As such, they do not enable one to discriminate homologous character state changes, much the way a mathematical average hides partitioned variation. For this reason, such methods have been eschewed in phylogenetic reconstruction for several decades and represent perhaps the most contentious points of discussion surrounding DNA barcodes.

The explosion of DNA barcode data and distance-based dendrograms did occasion certain remedial presentations (e.g., Prendini, 2005) of such methodological issues that had been debated and largely settled in the early decades of phylogenetic systematics. From our perspective, tree-building methods in the context of DNA barcoding are not, as they are in systematics, at issue on the grounds of their legitimacy as phylogenetic inference tools, if only because most studies suggest that COI analyzed in isolation is a fundamentally insufficient source of decisive phylogenetic information. Rather, distance methods fall short specifically in the realm of identification and diagnosis. The practical implications are (1) that above the level of

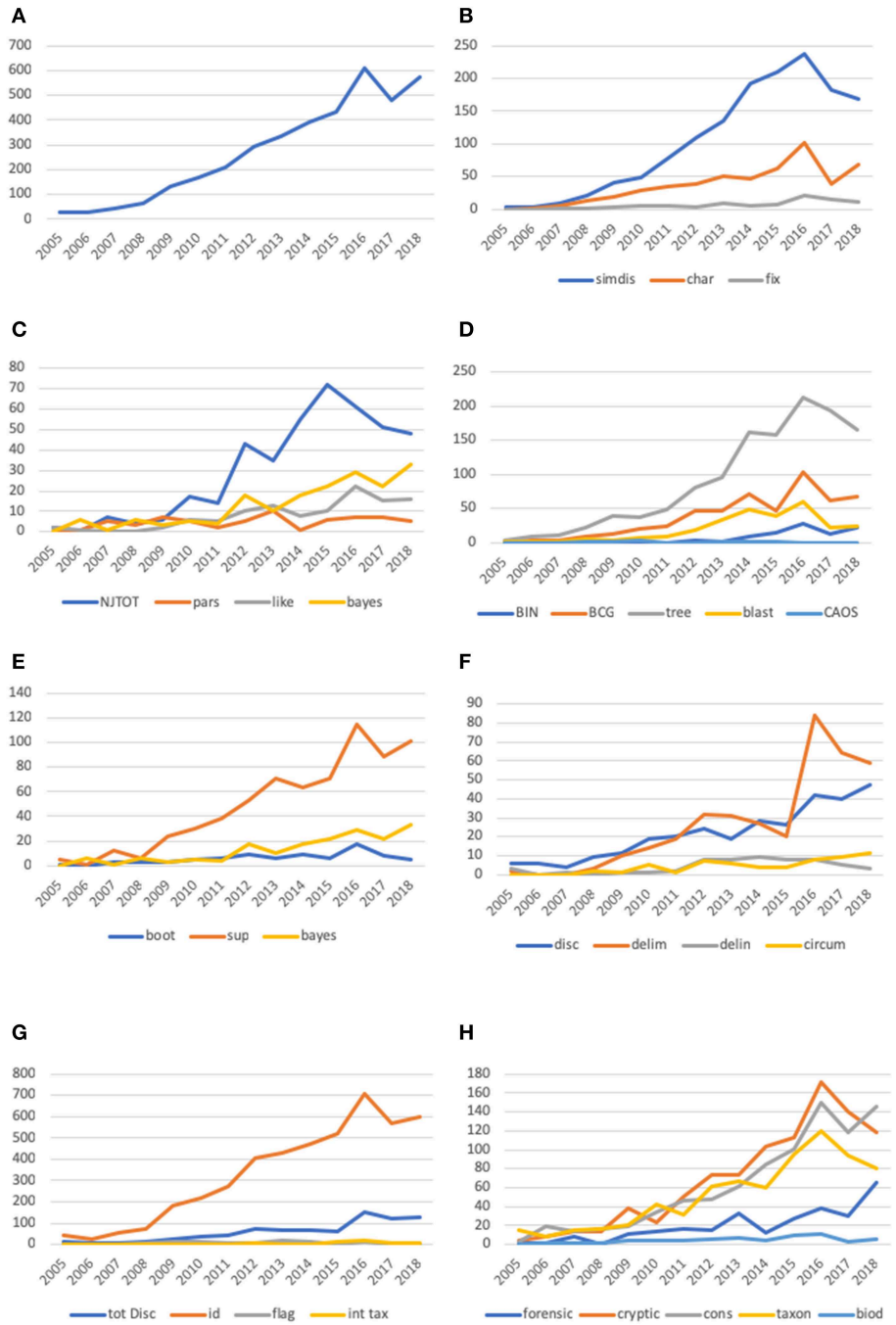


FIGURE 1 | Line plots of number of "hits" for keywords in the DNA barcode vocabulary subcategories established in the text. In all graphs the number of citations is given on the Y-axis and year is given on the X-axis. We also computed relative percentage of citations per year and these results are shown in **Supplemental Figure 1**. **(A)** Graph of the occurrence of scientific papers with the search word "DNA barcoding" in the title from 2003 to 2018. The "blip" in number of papers in 2016 that disrupts an otherwise smooth increase in number of papers by year might represent an increase in reports for the several international meetings that occurred in 2015. **(B)** The results of this analysis compare character based approaches to similarity/distance approaches. For this analysis we also use fixation as a character based term and show its usage in the graph. Search terms: "similarity" and "distance" combined into "simdis" and "character" and "fixation" combined into "char." We show the usage of "fixation" alone to demonstrate that this term is rarely used. **(C)** The results of this analysis compare the three major criteria for phylogenetic analysis—distance, parsimony and likelihood. Search terms: "NJ" and "neighbor joining" combined into "NJTOT," "parsimony" listed as "pars," likelihood listed as "like." Bayesian phylogenetic inference methods have also been used and these are listed under "bayes." **(D)** This figure shows comparison of the usage of terms that imply an examination of the robustness of the DNA barcode analysis. Such measures of robustness can be metrics such as bootstrap, or posterior probabilities such as in Bayesian phylogenetic inference. We also Search terms: "bootstrap" listed as "boot," "support" listed as sup, statistic, bayes. **(E)** The figure

(Continued)

FIGURE 1 | compares various methods of treating DNA barcode data. We include tree to demonstrate the use of tree relative to these other approaches. Search terms: barcode index "number" and "BIN" combined into "BIN," "barcode gap" listed as "BCG," "tree" listed as "tree," "blast" listed as "blast" and "character aggregation organization system" and "CAOS" combined into "CAOS." **(F)** This figure shows the usage of species discovery vocabulary in DNA barcoding. As we point out in the text, species description is a technical term used in taxonomy, while other terms like circumscription, delimitation and delineation are terms used by biologists studying speciation and species boundaries. Search terms: "species discovery" listed as "disc," "species delimitation" listed as "delim," "species delineation" listed as "delin" and "species circumscription" listed as "circum." **(G)** This figure compares the usage of "species discovery" terms with "specimen identification." We also compare the usage of "flagging" listed as "flag" and "integrative taxonomy" listed as "inttax." Search terms: "species discovery"or "totdisc" is the sum of counts for "species discovery," "species delineation," "species delimitation" and "species circumscription." **(H)** This figure compares the focus of papers in five areas that are generally listed by DNA barcode studies. DNA barcoding has been used in forensic studies, biodiversity studies, taxonomy, cryptic species studies and conservation biology. Search terms: "forensic" listed as "forensic." "cryptic" listed as "cryptic," "conservation" listed as "cons," "taxonomy" listed as "taxon" and "biodiversity" listed as "biod".

very closely related species, the COI gene typically realizes its greatest contribution to phylogenetic matrices that include a combination of other organellar and nuclear genes (Cameron et al., 2007; Leavitt et al., 2013) and (2) that no level of parameterization can compensate for the levels of saturation that inevitably appear in datasets with distantly related species or particularly in datasets with more terminals than characters. The immediate concern for the purposes of DNA barcoding is not that COI is necessarily inadequate as a sole phylogenetic marker, but that the ability of any data analyzed via distance is equally impeded in serving the goals of DNA barcoding as it is in phylogeny reconstruction. This is a function of the incompatibility of distance data with the transmission of diagnostic information. Simply put, a properly rooted parsimoniously optimized tree represents the most efficient summary possible of the available data, and enables the direct diagnosis of would-be species based on observable character state changes. This is a matter of mathematics, not opinion (Farris, 1980). The ostensible advantage of Neighbor-joining is its computational ease and straightforward presentation (a single tree is generated). Interpretive issues may arise only if such analyses are accepted as decisive without further exploration.

Figure 1B compares the occurrence of the search terms "character" and "similarity+distance" and suggests a consistent preference for Neighbor-joining (NJ) a tree-building algorithm. This is of course at least in part a function of the tools available in BoLD (Ratnasingham and Hebert, 2007), and we do not suggest that these analyses are all interpreted identically or for the same purposes. Two empirically linked search terms "fixed" and "character" align with diagnostic approaches and track their usage (**Figure 1B**).

Explicit mention of other methods of sequence analysis, Neighbor-joining (NJ), parsimony or "maximum parsimony" (MP), maximum likelihood (ML), and Bayesian (**Figure 1C**), appear erratically prior to 2008. Since then, the mentions of ML and Bayesian analysis have risen but not approached those of NJ, with parsimony (MP) appearing least frequently. This result is not surprising given the initial availability of NJ as the *prima facie* tool in the Barcode of Life Database (BoLD) system.

Visualization and Interpretation of Trees

In our reading of the barcode literature we noted many cases where taxonomic decisions were based either directly on distance measures (e.g., the barcode gap, discussed below) or on trees generated by such measures, but effectively decoupled from justification or discussion of those methods. Following Goldstein and DeSalle (2011), we distinguish the strictly graphic, tree-based approaches from tree-independent approaches, among which we further differentiate distance-based (e.g., BIN, barcode gap, BLAST searches) from diagnostic (e.g., CAOS; **Figure 1D**). Despite occasional papers in which barcode NJ trees are referred to as phylogenies, many authors have been careful to stress the utility of DNA barcoding for identification and discovery, and not as explicit phylogenetic statements. To be clear, tree-based approaches are valuable both as inferential tools for visualizing prospective species delimitation, and as provisional road maps of where to direct further research in delimiting species boundaries.

The interpretation of a barcode tree as a visual first pass for demarcating species vs. a phylogeny properly focuses attention on the integrity of the species themselves rather than the groups to which they belong (see Introduction), and perhaps for this reason—as well as the nature of variation within the COI gene, the often high number of individual sequences under analysis, and the types of analysis employed—measures of nodal support tend to find limited relevance in typical barcode analyses. Measures of nodal support have been presented with increasing frequency among DNA barcoding studies (**Figure 1E**), but in our survey the search terms reflecting such use (bootstrap, Bayes and statistic) appear less than a fifth as frequently as the term "support" itself.

Tree graphics and BLAST searches have each been used steadily since the inception of DNA barcoding **Figure 1D**. The term "barcode gap" (BCG), first coined in 2005 (Meyer and Paulay, 2005 and reiterated by Wiemers and Fiedler, 2007), appears steadily after 2009 and is the most frequently used of the terms referring to tree-independent analytics. The most recently minted tree-independent approach (BIN; Ratnasingham and Hebert, 2013), is unique to DNA barcoding and its use has increased slightly since its introduction in 2010. In our survey there appears to be a preference for tree-based approaches accompanying the preference for NJ trees, and limited growth in the use of tree-independent terms (even distance-based ones) after 2015. Diagnostic algorithms (e.g., CAOS, Sarkar et al., 2008) appear rarely, consistent with the infrequent reliance on character-based tree-independent approaches relative to BIN, BLAST, and BCG. **Table 2** summarizes the intersection between tree- and character-based (diagnostic) methods.

Specimen Identification and Species Delimitation

At the inception of DNA barcoding, two of its most frequently stressed benefits were specimen identification (or determination) and species discovery (**Figure 1F**). Specimen identification has

TABLE 1 | A glossary of DNA barcoding terms.

DISCIPLINES	
1.	*Conservation/conservation genetics/conservation biology*—A crisis discipline that employs multiple lines of evidence to prioritize and manage populations and assemblages of organisms and the natural areas they inhabit. 'Conservation genetics' refers to the subdiscipline of conservation biology that draws on genetic data for empirical solutions to conservation problems. One of the explicitly articulated applications of DNA barcoding is in conservation biology/genetics as it applies both to the discovery of new species and their management.
2.	*Forensic study*—Broadly, that which employs scientific methods to examine criminal activity. DNA barcoding may be used to evaluate the origins of commercial products, the presence of illegally obtained species, or factors related to decomposition, especially when other evidence is fragmentary and holomorphological inspection impossible.
3.	*Phylogeography*—Term introduced by John Avise and colleagues (Avise et al., 1987) to refine the focus of population level research in concert with geographic data. The approach is anchored in population-level analyses of molecular genetic data, traditionally mitochondrial or other uniparentally inherited markers.
4.	*Taxonomy/systematics*—The science of classifying biological organisms for purposes of efficient communication and the exploration of their evolutionary history. To be distinguished from nomenclature, which is a formalized aspect of taxonomy, and systematics, which encompasses and connotes a phylogenetic dimension. Taxonomy is an empirical (hypothetico-deductive) endeavor whereby hypotheses of species and higher taxa are tested (corroborated or falsified) with observational character data from multiple sources (morphological, molecular-genetic, behavioral, etc.). Integrative taxonomy is a term coined to encourage the integration of multiple sources of data with taxonomic practice.
BIOLOGICAL TERMS	
5.	*Character/character-based*—Characters are those features of organisms reflected in classification or phylogeny reconstruction. "Character-based" may refer to phylogenetic inference methods such as parsimony, likelihood, or Bayesian inference or to diagnoses as opposed to distance metrics. Davis and Nixon (1992) articulated Population Aggregation Analysis (PAA) which provides an example of how one might extract fixed characters from DNA sequences and thereby delimit diagnosable populations or species.
6.	*Cryptic/cryptic species/crypsis*—Difficult to detect and, in reference to species, referring to difficulty in diagnosing or recognizing morphologically indistinguishable species without DNA barcode data. One of the explicitly targeted applications of DNA barcoding is that of detecting cryptic species.
7.	*Fixed (character)/fixation*—A descriptor of character state as universally distributed within a given set or population. In the context of DNA sequences positions, a site is fixed when it bears the same base pair (A, C, G, T) for all individuals examined or, by inference, all members of a population. Fixation is used in PAA (see above), a tree-independent character-based approach.
8.	*Population*—A group of organisms that have the capacity to interbreed freely with one another, usually circumscribed geographically.
GRAPHIC TERMS	
9.	*Clade*—This term refers to a monophyletic (natural) group, namely a hypothetical common ancestor and all its descendants, as identified by uniquely derived and unreversed synapomorphies. A clade is visualized on a cladogram as a node and all its subtended terminals.
10.	*Cluster*—A group of individuals or genes visualized as terminals on a tree or dendrogram and used in place of "clade" whenever analyses are conducted below the species level. A group of organisms is said to cluster in an analysis when they share an exclusive node. Because clustering algorithms may be applied below the species level where relationships are not strictly nested, clusters are not monophyletic in the strict sense, only a graphic one. Cluster is also a term used to define closely related organisms in principal components analysis (e.g., Jombart et al., 2010) or STRUCTURE (Pritchard et al., 2003).
11.	*Tree/phylogenetic tree*—Any bifurcating graphic or dendrogram intended to summarize comparative data and interpreted to reflect common ancestry. Since "tree" refers to the graphic, it is not strictly synonymous with "phylogeny" but may be treated equivalently under the explicit assumptions of an underlying nested hierarchy generated by descent with modification. Trees based on recombinant elements of individual conspecific organisms may violate these assumptions but are still be used as provisional tools for approximating species boundaries. Phylogenetic trees can be generated using any number of methods as described above; the term "clade" is properly used with reference to derived or diagnostic characters and thus "cladogram" is generally reserved for trees generated under parsimony.
TREE-BUILDING METHODS	
12.	*Bayes/Bayesian*—A class of phylogenetic inference methods that employs the use of posterior probabilities first made widely available by the release of MrBayes (Ronquist and Huelsenbeck, 2003; in the same year DNA barcoding was proposed). "Bayesian" may also refer to species delimitation methods such as those proposed by Yang and Rannala (2010) and Fujita et al. (2012).
13.	*Likelihood/Maximum Likelihood/ML*—A class of parameterized tree-building approaches that incorporates probabilities of character state change based on frequentist statistics among different classes of character data (e.g., transitions vs. transversions, codon positions, etc.). The likelihood of the data given a tree and a model is computed to find an optimal tree for a dataset.
14.	*Neighbor Joining/NJ*—A numerical procedure using a distance (similarity) matrix to generate a dendrogram depicting distances among individuals. The matrix may be generated using a range of distance measures and parameters. Most NJ trees published from DNA barcode data employ the K2P distances.
15.	*Parsimony/Maximum Parsimony/MP*—The principle of parsimony is an empirical fundamental that equates scientific corroboration with the minimization of *ad hoc* hypotheses required to explain observations (data). In the context of tree-building algorithms, it is represented as an optimality criterion that minimizes the number of steps (character state changes) required by a cladogram. In this paradigm, the most parsimonious tree or set of trees for a given data set is simultaneously the most strictly supported hypothesis of relative recency of common ancestry and, as in the case of most DNA barcode analyses (which are not phylogenetic in the strict stense), the most efficient summary of character state distributions. Although early variants of parsimony have been widely abandoned, "maximum parsimony" is a neologism intended to convey empirical symmetry with maximum likelihood.
GENERAL PURPOSE OPERATIONAL TERMS	
16.	*Diagnose/diagnostic/diagnosis*—Diagnosis of putative species by means of unique, observable, and ostensibly fixed characters is a formal requirement of taxonomic nomenclature stipulated by the ICZN. With respect to DNA barcoding, diagnosis may be realized by demonstrating unique suites of base pairs.
17.	*Species circumscription/delimitation/delineation*—The iterative process of collating potentially diagnostic character data to proscribe observational boundaries between two or more species. Species delimitation methods are broad and require a criterion specified a priori (De Queiroz, 2007). Delimitation is used interchangeably with delineation, circumscription and demarcation.

(Continued)

TABLE 1 | Continued

18.	*Species description*—A formal description of a species based on comparative examination of specimens, ideally including detailed anatomical, behavioral and biogeographic data, and accompanied by formal naming and diagnosis from similar species.
19.	*Species discovery*—The conclusion drawn from collated character data that specimens cannot be assigned to described species.
20.	*Specimen identification/determination*—The process of using morphological or molecular diagnostics or other organismal attributes to assign biological specimens taxonomic names. Not to be confused with species delimitation or discovery (DeSalle, 2006; Rubinoff, 2006a,b; Goldstein and DeSalle, 2011).
21.	*Flag*—The annotation of an item, individual organism, group of organisms, or haplotype for subsequent study. In the context of DNA barcoding, specimens are flagged as potentially novel or cryptic species following provisional analyses.
TOOLS AND METRICS	
22.	*Barcode gap/BCG**–Presupposing accurate determination of the taxonomic rank for specimens under examination, the barcode gap s is the difference between the largest intraspecific distance and the smallest interspecific distance.
23.	*BIN/BIN system*—The barcode identification number (BIN; Ratnasingham and Hebert, 2013) is part of a system that clusters sequences using distance algorithms to produce identify operational taxonomic units (OTUs) for possible taxonomic designation.
24.	*BLAST*—The Basic Local Alignment Search Tool uses a query sequence and large database to find regions of local similarity between sequences. The program is at the heart of the National Center for Biotechnology Information's sequence search engine, compares nucleotide or protein sequences to the ever-growing sequence databases and estimates the statistical significance of matches.
25.	*Bootstrap/bootstrap support*—The bootstrap is a statistical tool for estimating confidence intervals that was developed for phylogenetics by Felsenstein (1985), although in this context it is not considered a confidence interval so much as a comfort index. It involves multi-replicate random resampling with replacement of individual columns of character data to generate bootstrap percentages for each node in a phylogenetic tree used as surrogates for support (see below).
26.	*CAOS (Character Aggregation Organization System)*—Sarkar et al. (2008) developed this program for discovering DNA sequence diagnostics using population level datasets. Jörger and Schrödl (2013) have articulated how the software can be used to generate diagnostics for taxonomic research.
27.	*Population Aggregation Analysis (PAA)*—This character based approach discovers diagnostics of different aggregates of individuals in a population level analysis. First articulated by Davis and Nixon (1992), this approach is used in the CAOS algorithm and software (see above). Variations of the PAA approach have been developed by several authors. These include the Cladistic Haplotype Analysis (CHA; Brower, 1999) and multilocus field for recombination (ml-FFR; Doyle, 1995).
28.	*(Genetic) Distance*/similarity— A phenetic measure of comparison which represents the overall similarity of two organisms. Operationally, a pairwise measure generated from sequence data, most commonly via the Kimura two parameter (K2P) model which specifies probabilities of different kinds of character state (base pair) change. The lack of equiprobability is used to correct the distance measure for rate heterogeneity of sequence change.
29.	*(Phylogenetic) Support*—The strength of inference for nodes in a phylogenetic tree are assessed using support measures. Higher the support measures connote greater reliability for a given hypothesized relationship. Bremer support (maximum parsimony based), bootstrap (distance, parsimony, likelihood) and Bayesian posteriors are all different kinds of support measures used in phylogenetic analysis.

TABLE 2 | A (not-exhaustive) categorization of the analysis space for DNA barcoding.

	Character-explicit	Distance-based
Tree-based	MP, ML, BPP BEAST[1]	NJ*, minimum evolution
Tree-independent	CAOS[2], PTP[3] and bPTP[3] GMYC[4]	BCG[5], BIN[6], BLAST[7] STRUCTURE[8]; PCA[9] (principal icomponents) ABGD[10] (automated BCG discovery), BAPS[11]

References for the methods mentioned in this table for MP, ML, and NJ are classic ones. References for other methods are as follows: 1. Drummond and Rambaut (2007); 2. Sarkar et al. (2008) and Jörger and Schrödl (2013); 3. Zhang et al. (2013) and Yang and Rannala (2010); 4. Monaghan et al. (2009); 5. Fujita et al. (2012); 6. Ratnasingham and Hebert (2013); 7. Johnson et al. (2008); 8. Pritchard et al. (2003); 9. Jombart et al. (2010); 10. Puillandre et al. (2012); 11. Cheng et al. (2013).

been used interchangeably with "species identification" in some publications, as have a number of terms related to identification and discovery. DeSalle (2006) used the term "identification" only in the context of assigning taxonomic information. Although in the present paper we refer to this as "determination" (of specimens, not species), the published usage is too broad in intent to be parsed with any great deal of precision. Since the power of DNA barcoding resides in the coverage of the available database, the conclusion that a given species is new to science for example, is a function of whether a queried sequence corresponds to those from authoritatively identified specimens. The discovery of species new to science is thus a function of failure to assign a valid name to a given sequence under the assumption that identical (or highly similar) available sequences represent conspecific individuals. As such, "discovery" has for some authors been more controversial than identification (Matz and Nielsen, 2005), and that controversy may easily be amplified by the use of barcoding to estimate species richness in bulk samples (Andersen et al., 2012; Shokralla et al., 2012; Kress et al., 2015; Sickel et al., 2015). Specimen identification, particularly for thoroughly studied and well-sampled groups, holds broader appeal, particularly outside the academic community.

Incorporating DNA barcoding with taxonomy has been discussed and widely adopted as a form of integrative taxonomy, which simply refers to simultaneous analysis of disparate sources of data (**Figure 1G**). DNA barcodes are among the more readily got and appealing forms of data that may be used to flag specimens as warranting taxonomic attention (Goldstein and DeSalle, 2011). Based on their occurrences summarized in **Figure 1F**, "integrative taxonomy" and "flag" are not often used explicitly in connection with species "discovery." This may suggest a disconnect between the appeal of species discovery in the abstract and its actual undertaking. If so, it highlights the important point that cryptic species discovered from DNA barcodes are not always accompanied by taxonomic revisionary work.

Since its inception, DNA barcoding has been bolstered by its utility for discovering cryptic species specifically as well as in taxonomic revision, forensics, conservation and biodiversity studies generally. Recognizing the potential bearing of cryptic species on each of these fields, **Figure 1H** illustrates that the study of cryptic species has consistently played a focal role in a range of fields over the 15-year period we examined, with explicit mention of conservation and taxonomy appearing with less frequent emphasis, followed by "forensic" and "biodiversity."

MEANING

Examinations of word usage are productive only to the degree that common ground in both meaning and intent is well-understood, and inferences from any compendium of word usage are only as good as the precision with which the search terms were originally used. Loose usage of terms like "diagnosis" or "tree" seem inevitable as barcoding tools become increasingly accessible. As genomic data are generated with increasing ease, it remains to be seen whether the enthusiasm for DNA as it is currently practiced will transition to the larger endeavor of archiving accessible genomic data.

The most obvious and important result of the exercises performed here is that distance or phenetic approaches have prevailed in DNA barcoding practices for reasons that appear to be more practical than scientific. Conflating distance data with diagnoses and algorithms with tree graphics are not uncommon mistakes in the taxonomic literature. Although the use of NJ trees or distances to diagnose species appears in the literature, we would argue that doing so obviates the real diagnostic value of barcode data that would meet the requirements of diagnoses set forth in the ICZN and elsewhere.

Distance-based methods have a well-established place in population genetics, where they play important roles in evaluating raw divergence among related individuals or populations. In the context of phylogenetic inference, however, clustering operations based on phenetic similarity have for several decades been rejected by systematists for empirical and statistical reasons, not the least of which is that since they combine available character data into a single ensemble metric, they cannot test or summarize specific character homologies that would otherwise contribute to a diagnosis (Ferguson, 2002; DeSalle, 2007; Little and Stevenson, 2007). Distance metrics are nevertheless easy to calculate and methods such as NJ generate dendrograms with a seeming minimum of ambiguity. The development of DNA barcode databases hinged on the ease of NJ precisely because of this computational ease, because any lack of decisiveness among the data is not transparent in seemingly unambiguous single tree that obtains from every NJ analysis.

There exists quite a bit of variation in the handling of dendrograms (distance based figures) generated by DNA barcodes for purposes following the organization of specimens. Many draw empirical conclusions directly from a given NJ tree instead of using it recursively to examine/interpret other characters or pieces of information. But how researchers use the tree to summarize variation and evaluate actual support for would-be relationships varies considerably. Phenetic trees, rapidly generated as they are, risk yielding spurious representations of data, and represent liabilities to the extent that apparent tree structure is uncorroborated.

Clustering algorithms and dendrograms are used throughout biology for purposes ranging from ecological community analysis to visualizing gene expression data. The use of trees in phylogenetic science is distinguished from other applications by the implied superposition of a temporal dimension that enables testing hypotheses of character evolution. At its simplest, this is achieved by establishing polarity, or the direction of character state change, through the operation of rooting, followed by optimization of hypothetical character states at nodes. Regardless of whether scientists imagine distance-generated trees to be "phylogenies," neither of these operations is possible on such trees without violating the fundamental assumptions of rooting and optimization. A raw dendrogram, however it is generated, is simply a form of metadata that summarizes similarity using a given metric or optimality criterion; it cannot by itself serve to "diagnose" anything with reference to observable character states much less evaluate synapomorphy, establish monophyly, or test ideas of character evolution.

To the credit DNA barcoding's architects, it has been stressed that barcode trees are not intended to serve as phylogenies, and as the menu of tools available on BOLD has expanded to include features that enable proper diagnoses, it is our hope that the number of taxonomic papers perpetuating that error will one day subside. Our purpose is not to belabor this any further, but to stress that despite their computational ease, NJ trees render barcode data under-utilized.

DISCUSSION

Inevitably, whenever a new tool is developed that expedites a set of tasks, the training required prior to that development becomes at least partly obsolete, and it becomes easy to overlook standards—obsolete or not—that went along with it. In this case those standards range from matters as straightforward as species diagnosis to the more nuanced interpretation of molecular phylogenetic trees. It has at times appeared as though the antiquated view of systematics as an exercise in naming things, rather than an empirical endeavor to reconcile classifications with evolutionary hypotheses, has persisted. Graphic summary statements of phylogenetic data are rarely as decisive as they appear when stripped of their analytical details, and from the taxonomy-as-nomenclature perspective, systematics is seen as a pedantic holdover of Victorian pseudo-science, its practices the relics of a bygone era, and the very existence of undescribed species or unstable classification the function of some intrinsic psycho-intellectual flaw known collectively as the "taxonomic impediment" rather than a reflection of the raw magnitude of biodiversity. Similar brands of taxonomic naïvete have manifested elsewhere, as in recent debates over wisdom of taxonomic descriptions using photographs as "types."

(Garraffoni and Freitas, 2017; see also Amorim et al., 2016, Ceríaco et al., 2016, Pape, 2016, Santos et al., 2016). Although hailed as a possible solution to the taxonomic impediment, DNA barcoding performed uncritically risks the encumbrance of subsequent efforts and defeats its own purpose.

It seems generally accepted that, with exceptions in various groups ranging from genera to families, conventional barcode analyses work quite well in circumscribing potentially recognizable species that can be further corroborated with other characters. Why then be concerned about using distance measures as arbiters of identity? Although this paper is no place to resurrect a discussion on species concepts, there is nothing mysterious about the fact that barcode analyses tend to predict species that are ultimately recognizable by other means—certainly the rigorous evaluation of candidate loci undertaken before settling on COI has resolved that much. But it is important to separate the statement that NJ analyses "work" to identify species from the supposition that they allow us to infer anything about species in the abstract. The premise of the claim that NJ works to identify species united by some abstracted metaphysical property is that the species criterion is unspecified. This is not mere sophistry: Without establishing or allowing for an independent criterion for corroboration, there can be no means of evaluating what works and what does not because the claim is fundamentally unfalsifiable. If we adopt the perspective that species—whatever evolutionary concepts to which they may or may not conform—can be palatably recognized by congruent character data, then accepting provisional clusters as working hypotheses subject to further corroboration is quite reasonable. In other words, the fact that a very high proportion of diagnosable species are captured by NJ analyses is encouraging, but not sufficient. We maintain simply that even a small a small percentage of species overlooked or misdiagnosed warrant acknowledgment and the arbitrariness of inferring a universal distance measure is unnecessary when the means exist for quantifying diagnostic features directly.

DNA barcoding represents a tool with a range of empirical uses as broad as the array of taxa and available specimens with accompanying barcodes. Although these empirical uses do not extend to rigorous phylogenetic testing, barcode data realize their greatest potential throughout the recursive process of taxonomic investigation. In our view, the coupling of DNA barcoding with distance methods rendered its potential as a taxonomic tool under-realized. Although we actively embrace DNA barcoding in our own taxonomic research and as a near-universal advance for taxonomic research in general, we reject the premise that DNA barcoding serves to repair some inherent flaw in the practice of systematics. We view the taxonomic impediment not as a manifestation of human-induced shortcomings but as a reflection of the magnitude of global species richness.

We hope to have distinguished methodological issues from semantic ones, by pointing out, for example, the percent differences are by definition mathematically non-diagnostic. But our primary is not to redress common practices, but to suggest that more could be gained from additional analyses that would serve the formal taxonomic goals of diagnosis. It is not our intent to cast a pall over the use of barcode data to uncover diversity at fine scales, but to articulate how those data may continue to be enhanced. We stress the importance of not overstating the implications of a word survey; our hope is merely to have provided a crude calibration of how quickly we might reasonably expect to see significant shifts in how barcode data are analyzed. A conclusion of this exercise is that researchers are more likely to follow the examples of their peers and use the tools most readily available than they are to ponder the minutiae of evolutionary analyses.

AUTHOR CONTRIBUTIONS

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

ACKNOWLEDGMENTS

RD acknowledges the Institute for Comparative Genomics at the AMNH (ICG-AMNH) and the Lewis and Dorothy Cullman Program in Molecular Systematics and the Korein Family for continued support. Mention of trade names or commercial products in this publication is solely for the purpose of providing specific information and does not imply recommendation or endorsement by the USDA; USDA is an equal opportunity provider and employer.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fevo.2019.00302/full#supplementary-material

REFERENCES

Amorim, D. S., Santos, C. M., Krell, F. T., Dubois, A., Nihei, S. S., Oliveira, O. M., et al. (2016). Timeless standards for species delimitation. *Zootaxa* 4137, 121–128. doi: 10.11646/zootaxa.4137.1.9

Andersen, K., Bird, K. L., Rasmussen, M., Haile, J., Breuning-Madsen, H., Kjaer, K. H., et al. (2012). Meta-barcoding of 'dirt'DNA from soil reflects vertebrate biodiversity. *Mol. Ecol.* 21, 1966–1979. doi: 10.1111/j.1365-294X.2011.05261.x

Avise, J. C., Arnold, J., Ball, R. M., Bermingham, E., Lamb, T., Neigel, J. E., et al. (1987). Bridge between population, genetics and systematics. *Ann. Rev. Ecol. Syst.* 18, 489–522. doi: 10.1146/annurev.es.18.110187.002421

Brower, A. V. Z. (1999). Delimitation of phylogenetic species with DNA sequences: a critique of Davis and Nixon's population aggregation analysis. *Syst. Biol.* 48, 199–213. doi: 10.1080/106351599260535

Cameron, S. L., Lambkin, C. L., Barker, S. C., and Whiting, M. F. (2007). A mitochondrial genome phylogeny of Diptera: Whole genome sequence data accurately resolve relationships over broad timescales with

high precision. *Syst. Entomol.* 32, 40–59. doi: 10.1111/j.1365-3113.2006.00355.x

Ceríaco, L. M., Gutiérrez, E. E., and Dubois, A. (2016). Photography-based taxonomy is inadequate, unnecessary, and potentially harmful for biological sciences. *Zootaxa* 4196, 435–445. doi: 10.11646/zootaxa.4196.3.9

Cheng, L., Connor, T. R., Sirén, J., Aanensen, D. M., and Corander, J. (2013). Hierarchical and spatially explicit clustering of DNA sequences with BAPS software. *Mol. Biol. Evol.* 30, 1224–1228. doi: 10.1093/molbev/mst028

Davis, J. I., and Nixon, K. C. (1992). Populations, genetic variation, and the delimitation of phylogenetic species. *Syst. Biol.* 41, 421–435. doi: 10.1093/sysbio/41.4.421

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

DeSalle, R. (2006). Species discovery versus species identification in DNA barcoding efforts: response to Rubinoff. *Conserv. Biol.* 20, 1545–1547. doi: 10.1111/j.1523-1739.2006.00543.x

DeSalle, R. (2007). Phenetic and DNA taxonomy; a comment on Waugh. *Bioessays* 29, 1289–1290. doi: 10.1002/bies.20667

Doyle, J. J. (1995). The irrelevance of allele tree topologies for species delimitation, and a non-topological alternative. *Syst. Bot.* 20, 574–588.

Drummond, A. J., and Rambaut, A. (2007). BEAST: bayesian evolutionary analysis by sampling trees. *BMC Evol. Biol.* 7, 214–217. doi: 10.1186/1471-2148-7-214

Farris, J. S. (1980). The efficient diagnoses of the phylogenetic system. *Syst. Zool.* 29, 386–401. doi: 10.2307/2992344

Felsenstein, J. (1985). Confidence limits on phylogenies: an approach using the bootstrap. *Evolution* 39, 783–791.

Ferguson, J. W. H. (2002). On the use of genetic divergence for identifying species. *Biol. J. Linn. Soc.* 75, 509–516. doi: 10.1046/j.1095-8312.2002.00042.x

Fujita, M. K., Leaché, A. D., Burbrink, F. T., McGuire, J. A., and Moritz, C. (2012). Coalescent-based species delimitation in an integrative taxonomy. *Trends Ecol. Evol.* 27, 480–488. doi: 10.1016/j.tree.2012.04.012

Funk, D. J., and Omland, K. E. (2003). Species-level paraphyly and polyphyly: Frequency, causes, and consequences, with insights from animal mitochondrial DNA. *Annu. Rev. Ecol. Evol. Syst.* 34, 397–423. doi: 10.1146/annurev.ecolsys.34.011802.132421

Garraffoni, A. R., and Freitas, A. V. (2017). Photos belong in the taxonomic Code. *Science* 355, 805–805. doi: 10.1126/science.aam7686

Goldstein, P. Z., and DeSalle, R. (2011). Integrating DNA barcode data with taxonomic practice: Determination, discovery, and description. *Bioessays* 33, 135–147. doi: 10.1002/bies.201000036

Hajibabaei, M., Janzen, D. H., Burns, J. M., Hallwachs, W., and Hebert, P. D. (2006). DNA barcodes distinguish species of tropical Lepidoptera. *Proc Natl Acad Sci U.S.A.* 103, 968–971. doi: 10.1073/pnas.0510466103

Hausmann, A., Miller, S. E., Holloway, J. D., deWaard, J. R., Pollock, D., Prosser, S. W., et al. (2016). Calibrating the taxonomy of a megadiverse insect family: 3000 DNA barcodes from geometrid type specimens (Lepidoptera, Geometridae). *Genome* 59, 671–684. doi: 10.1139/gen-2015-0197

Hebert, P. D., Cywinska, A., Ball, S. L., and deWaard, J. R. (2003a). Biological identifications through DNA barcodes. *Proceedings of the Royal Society of London. Series B: Biological Sciences* 270, 313–321. doi: 10.1098/rspb.2002.2218

Hebert, P. D., Ratnasingham, S., and deWaard, J. R. (2003b). Barcoding animal life: cytochrome c oxidase subunit 1 divergences among closely related species. *Proc. R. Soc. Lond. Ser. B Biol. Sci.* 270, S96–S99. doi: 10.1098/rsbl.2003.0025

Johnson, M., Zaretskaya, I., Raytselis, Y., Merezhuk, Y., McGinnis, S., and Madden, T. L. (2008). NCBI BLAST: a better web interface. *Nucleic Acids Res.* 36, W5–W9. doi: 10.1093/nar/gkn201

Jombart, T., Devillard, S., and Balloux, F. (2010). Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. *BMC Genet.* 11:94. doi: 10.1186/1471-2156-11-94

Jörger, K. M., and Schrödl, M. (2013). How to describe a cryptic species? Practical challenges of molecular taxonomy. *Front. Zool.* 10:59. doi: 10.1186/1742-9994-10-59

Kress, W. J., García-Robledo, C., Uriarte, M., and Erickson, D. L. (2015). DNA barcodes for ecology, evolution, and conservation. *Trends Ecol. Evol.* 30, 25–35. doi: 10.1016/j.tree.2014.10.008

Leavitt, J. R., Hiatt, K. D., Whiting, M. F., and Song, H. (2013). Searching for the optimal data partition- ing strategy in mitochondrial phylogenomics: a phylogeny of Acridoidea (Insecta: Orthoptera: Caelifera) as a case study. *Mol. Phylogenet. Evol.* 67, 494–508. doi: 10.1016/j.ympev.2013.02.019

Little, D. P., and Stevenson, D. W. (2007). A comparison of algorithms for the identification of specimens using DNA barcodes: examples from gymnosperms. *Cladistics* 23, 1–21. doi: 10.1111/j.1096-0031.2006.00126.x

Matz, M. V., and Nielsen, R. (2005). A likelihood ratio test for species membership based on DNA sequence data. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 360, 1969–1974. doi: 10.1098/rstb.2005.1728

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

Monaghan, M. T., Wild, R., Elliot, M., Fujisawa, T., Balke, M., Inward, D. J., et al. (2009). Accelerated species inventory on Madagascar using coalescent-based models of species delineation. *Syst. Biol.* 58, 298–311. doi: 10.1093/sysbio/syp027

Pape, T. (2016). Taxonomy: species can be named from photos. *Nature* 537:307. doi: 10.1038/537307b

Prendini, L. (2005). Comment on "Identifying spiders through DNA barcodes". *Can. J. Zool.* 83, 498–504. doi: 10.1139/z05-025

Pritchard, J. K., Wen, W., and Falush, D. (2003). *STRUCTURE. Documentation for Structure Software: Version* 2. Available online at: http://pritch.bsd.uchicago.edu

Puillandre, N., Lambert, A., Brouillet, S., and Achaz, G. (2012). ABGD, Automatic Barcode Gap Discovery for primary species delimitation. *Mol. Ecol.* 21, 1864–1877. doi: 10.1111/j.1365-294X.2011.05239.x

Ratnasingham, S., and Hebert, P. D. (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

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

Ronquist, F., and Huelsenbeck, J. P. (2003). Mrbayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* 19, 1572–1574. doi: 10.1093/bioinformatics/btg180

Rubinoff, D. (2006a). Utility of mitochondrial DNA barcodes in species conservation. *Conserv. Biol.* 20: 1026–1033. doi: 10.1111/j.1523-1739.2006.00372.x

Rubinoff, D. (2006b). Barcodes, integrated. DNA barcoding evolves into the familiar.*Conserv. Biol.* 20, 1548–1549. doi: 10.1111/j.1523-1739.2006.00542.x

Santos, C. M. D., Amorim, D. S., Klassa, B., Fachin, D. A., Nihei, S. S., De Carvalho, C. J. B., et al. (2016). On typeless species and the perils of fast taxonomy. *Syst. Entomol.* 41, 511–515. doi: 10.1111/syen.12180

Sarkar, I. N., Planet, P. J., and Desalle, R. (2008). CAOS software for use in character-based DNA barcoding. *Mol. Ecol. Resour.* 8, 1256–1259. doi: 10.1111/j.1755-0998.2008.02235.x

Shokralla, S., Spall, J. L., Gibson, J. F., and Hajibabaei, M. (2012). Next-generation sequencing technologies for environmental DNA research. *Mol. Ecol.* 21, 1794–1805. doi: 10.1111/j.1365-294X.2012.05538.x

Sickel, W., Ankenbrand, M. J., Grimmer, G., Holzschuh, A., Härtel, S., and Lanzen, J., et al. (2015). Increased efficiency in identifying mixed pollen samples by meta-barcoding with a dual-indexing approach. *BMC Ecol.* 15:20. doi: 10.1186/s12898-015-0051-y

Stoeckle, M. (2003). Taxonomy, DNA, and the bar code of life. *Bioscience* 53, 796–797. doi: 10.1641/0006-3568(2003)053[0796:TDATBC]2.0.CO;2

Wiemers, M., and Fiedler, K. (2007). Does the DNA barcoding gap exist?–a case study in blue butterflies (Lepidoptera: Lycaenidae). *Front. Zool.* 4:8. doi: 10.1186/1742-9994-4-8

Yang, Z., and Rannala, B. (2010). Bayesian species delimitation using multilocus sequence data. *Proc. Natl Acad Sci. U.S.A.* 107, 9264–9269. doi: 10.1073/pnas.0913022107

Zahiri, R., Lafontaine, J. D., Schmidt, B. C., deWaard, J. R., Zakharov, E. V., and Hebert, P. D. N. (2017). Probing planetary biodiversity with DNA barcodes: The Noctuoidea of North America. *PLoS ONE* 12:e0178548. doi: 10.1371/journal.pone.0178548

Zhang, J., Kapli, P., Pavlidis, P., and Stamatakis, A. (2013). A general species delimitation method with applications to phylogenetic placements. *Bioinformatics* 29, 2869–2876. doi: 10.1093/bioinformatics/Complex btt499

Disclaimer: The authors are solely responsible for the writing of this paper.

Conflict of Interest Statement: 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.

The handling editor and reviewer, RH, declared their involvement as co-editors in the Research Topic, and confirm the absence of any other collaboration.

frontiers in Ecology and Evolution

REVIEW
published: 18 September 2019
doi: 10.3389/fevo.2019.00342

The Mitochondrial Genome–on Selective Constraints and Signatures at the Organism, Cell, and Single Mitochondrion Levels

*Noam Shtolz and Dan Mishmar**

Department of Life Sciences, Ben-Gurion University of the Negev, Beer Sheva, Israel

Natural selection acts on the phenotype. Therefore, many mistakenly expect to observe its signatures only in the organism, while overlooking its impact on tissues, cells and subcellular compartments. This is particularly crucial in the case of the mitochondrial genome (mtDNA), which, unlike the nucleus, resides in multiple cellular copies that may vary in sequence (heteroplasmy) and quantity among tissues. Since the mitochondrion is a hub for cellular metabolism, ATP production, and additional activities such as nucleotide biosynthesis and apoptosis, mitochondrial dysfunction leads to both tissue-specific and systemic disorders. Therefore, strong selective pressures act to maintain mitochondrial function via removal of deleterious mutations via purifying (negative) selection. In parallel, selection also acts on the mitochondrion to allow adaptation of cells and organisms to new environments and physiological conditions (positive selection). Nevertheless, unlike the nuclear genetic information, the mitochondrial genetic system incorporates closely interacting bi-genomic factors (i.e., encoded by the nuclear and mitochondrial genomes). This is further complicated by the order of magnitude higher mutation rate of the vertebrate mtDNA as compared to the nuclear genome. Such mutation rate difference generates a generous mtDNA mutational landscape for selection to act, but also requires tight mito-nuclear co-evolution to maintain mitochondrial activities. In this essay we will consider the unique mitochondrial signatures of natural selection at the organism, tissue, cell, and single mitochondrion levels.

Keywords: mitochondria, mtDNA, selection, single cell, single mitochondrion, evolution

OPEN ACCESS

Edited by:
David S. Thaler, Biozentrum, Universität Basel, Switzerland

Reviewed by:
Geoffrey E. Hill, Auburn University, United States
Mark Stoeckle, The Rockefeller University, United States

****Correspondence:***
Dan Mishmar
dmishmar@bgu.ac.il

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

Received: *28 May 2019*
Accepted: *26 August 2019*
Published: *18 September 2019*

Citation:
Shtolz N and Mishmar D (2019) The Mitochondrial Genome–on Selective Constraints and Signatures at the Organism, Cell, and Single Mitochondrion Levels. Front. Ecol. Evol. 7:342. doi: 10.3389/fevo.2019.00342

INTRODUCTION

All cells require ATP, as it is the most common cellular currency to do work. Although glycolysis provides the means to produce ATP when glucose is available, an order of magnitude and more efficient energy-production system emerged ∼2.5 billion years ago in eukaryotes through endosymbiosis between the ancestor of the mitochondria and the progenitor of eukaryotic cells (Sagan, 1967). Ever since, genetic material migrated from the genome of the former free-living alpha proteo-bacterium to the host nucleus (the formation of which will not be discussed here). This apparent lateral gene transfer created interdependence between the host and tenant, not only due to the cellular reliance on mitochondrial ATP production, but also due to the involvement of mitochondria in many other major activities, such as nucleotide biosynthesis, the generation

of Iron-Sulfur protein clusters, apoptosis etc. Consequently, mitochondria cannot be grown outside of the eukaryotic cells, and the vast majority of eukaryotic cells cannot survive without their mitochondria.

Such relocation of genetic material from the cytoplasm to the cell nucleus required adaptation of the former mitochondrial DNA (mtDNA)-encoded genes to the nuclear genetic code and translation machinery, assimilation of the "new" genetic immigrants into the nuclear mode of gene regulation, which respond to chromatin remodeling, and finally required the acquisition of the protein properties that allow their re-import into the mitochondria to maintain their function. This became further complicated by the emergence of metazoans, which required differential energy expenditure per tissues and cell types (Lane and Martin, 2010). Hence, strong selective constraints should have been inflicted to preserve the activity (via negative selection) of the factors which generate such energy. In parallel, positive selection likely also acted to enable adaptation of the energy metabolism system to a variety of environments, and possibly allows each cell type to incorporate its specific activities within the tissues of the organism. In the current essay, we will discuss evidence supporting selective signatures that marked the mitochondrial genetic system after its emergence. We will demonstrate such signatures at three different levels—the organism, the cell and the individual mitochondrion. As much literature studied and discussed the organism level, we will put more emphasis on the cellular and mitochondrial levels.

Selection Acts on the Phenotype—the Various Levels of Mitochondrial Phenotypic Expression

In 1983 the neutral theory of molecular evolution was put forward by Motoo Kimura, who argued that most population genetic variants result from mutations which propagate via genetic drift and not selection, and therefore, different alleles generally do not affect the individual's fitness (Kimura, 1983). This theory is frequently misinterpreted, as it does not imply that organisms are not adapted to their environments, and it does not state that natural selection is negligible in shaping genetic variation and genomes. Although geneticists and their next generation of scientists, genomicists, are keen to identify the signatures of selection readily written as an epitaph in the genetic material, selection acts first on the phenotype and (indirectly) on the genotype. Indeed, although many of the traits of a given individual, in any given metazoan species, are mostly the result of changes in the inherited matter, many traits may vary in their appearance due to interactions between two or more genetic factors (i.e., epistasis), and due to variable interactions with environmental conditions. While considering the mitochondria, this scheme of interaction is further complicated by the involvement of factors encoded by both the nuclear and the mitochondrial genomes (G × G) (Levin et al., 2014). Therefore, when also considering the environment, mitochondrial-encoded traits are influenced by interactions between two genomes and a variety of environments and physiological conditions (G × G × E) (Zhu et al., 2014). This scheme experienced another level of complexity, due to the existence of multiple mitochondria per cell, the mtDNAs of which may vary in sequence (heteroplasmy). Specifically, heteroplasmy patterns and their phenotypic consequences notably differs between dividing (mitotic) and post-mitotic tissues (Kowald and Kirkwood, 2013; Filograna et al., 2019). Therefore, mitochondrial phenotypic implications should be considered not only at the level of the organism which has been widely discussed in the past (Meiklejohn et al., 2007; Levin et al., 2014), but also at the level of the cell, and even in the single mitochondrion (**Figure 1**). At first, analysis of a single mitochondrion was limited by the technological resolution of mutational detection (Reiner et al., 2010). However, a recent study investigated mtDNA sequence variation within 118 isolated mitochondria from mouse neurons and astrocytes (Morris et al., 2017). Whole mtDNA sequencing of each of the isolated mitochondria revealed an average of 3.9 (± 5.71 SD) single heteroplasmic nucleotide variants (SNVs) in the tested mitochondrial population per nucleotide position. Hence, despite the fact that all samples originate from the same mouse strain (C57BL/6N), a notable repertoire of mtDNA heteroplasmic SNVs was discovered. While analyzing this mutational repertoire, these researchers plotted the most frequent non-reference alleles against log2 appearance counts. Unexpectedly, they observed a distribution that did not match the symmetric U-shaped distribution, which is the expected distribution of alleles, under the assumption of neutrality (Birky et al., 1983). In addition to the distribution of non-reference alleles, the authors identified three specific SNVs with high potential impact. They found a negative correlation between the predicted impact of the specific mutations and their degree of variability among the tested samples. These pieces of evidence suggest that selection likely shaped the mutational distribution pattern at the single mitochondrion level and was responsible for the divergence from a random pattern.

While considering intercellular patterns of heteroplasmy, under the assumption of neutrality, with genetic drift being the primary force that shapes genetic variation over-time, it is likely that certain cells will develop a heterogenous mitochondrial population, without any phenotypic implications. Such a scenario predicts stochastic accumulation of mutations during each replication cycle, and random assortment of mtDNA molecules between daughter cells. To test such prediction, heteroplasmy patterns were studied in colonies derived from single cells of two human cell lines—MDA-MB-157 (breast cancer) and U20S (osteosarcoma) (Jayaprakash et al., 2015). Jayaprakash et al. found stably maintained heteroplasmy in daughter cells over multiple passages. This finding slightly departs from the expected random mitochondrial segregation, which predicts diverse levels of heteroplasmy regardless of the nature of the identified mtDNA mutations. Although these findings reveal constant levels of heteroplasmy regardless of the mtDNA position in which the mutations occurred, one cannot easily explain the results either by random forces, or by selective constraints. However, a recent study of heteroplasmy in single cells identified consistent cell-lineage-specific segregation of heteroplasmic mtDNA mutations during differentiation of hematopoietic cells, which did not comply with random

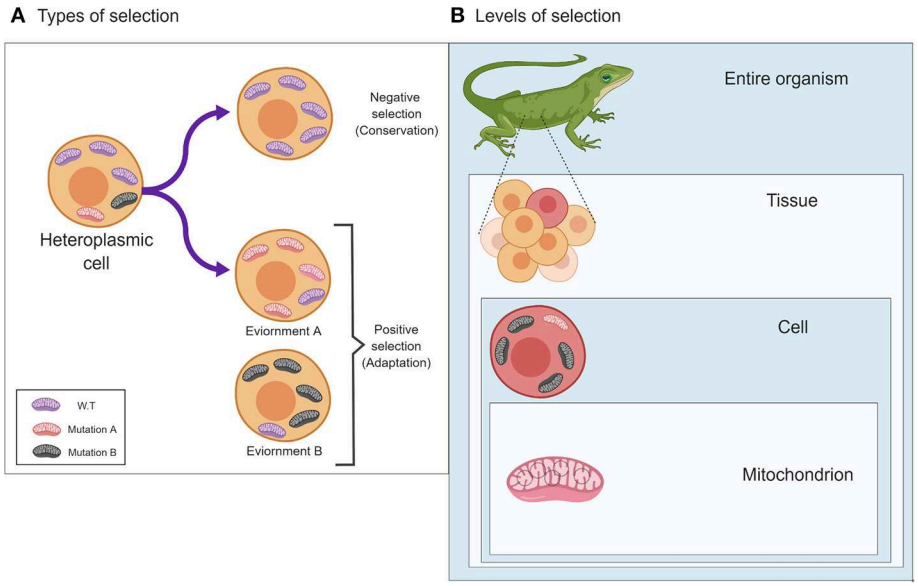


FIGURE 1 | The two main types of selection, Negative and Positive, and their effects on patterns of mitochondrial genetic variability. **(A)** Different types of selection (negative or positive) affect the pattern of mitochondrial mutations in response to different environments (lower panels) or will undergo a selective sweep (negative selection to maintain function (upper panel). **(B)** The different levels in which selection may act on the mitochondria, i.e., (down-up) the single mitochondrion, cells and the entire organism. The figure was designed using ©BioRender—www.biorender.com.

segregation of mutations during cell division (Ludwig et al., 2019). The growing availability of single cell genomics data holds much promise in future investigation of the forces that govern patterns of mtDNA heteroplasmy during cell division and differentiation (Ludwig et al., 2019).

Deep sequencing of many human individuals (Li et al., 2010; Payne et al., 2013; Ye et al., 2014), family members (Goto et al., 2011) and identical twins (Avital et al., 2012) attest for the common phenomenon of heteroplasmy, its inheritance and accumulation during the life of the individual. Notably, the distribution of these mutations across the mtDNA was not random, with much higher incidence of heteroplasmy in non-coding mtDNA sequences than expected by chance (Avital et al., 2012). Thus, it can be inferred that heteroplasmic mutations are likely subjected to selective constraints. We previously found that such non-random mutational distribution throughout the mtDNA occurred regardless of the heteroplasmy levels in two cell populations (blood and skeletal muscle) in identical twins, thus attesting for the impact of selection, not only at the level by which phenotypic consequences are expected for the organism, but also at the cellular level (Avital et al., 2012). Nevertheless, while investigating very low-level heteroplasmic mutations, it has been argued that known mtDNA disease-causing mutations could be present in healthy individuals in the population, if their heteroplasmy is maintained at low levels (Ye et al., 2014). A growing body of evidence suggests that since the mitochondria within cells are interconnected in a network, pathological mutations could survive due to inter-mitochondrial functional complementation, and exchange of nutrients (Schon and Gilkerson, 2010). As mitochondria go through cycles of fission and fusion, which build such network (fusion), or disconnect it (fission), dysfunctional mitochondria could be removed by mitophagy (Twig and Shirihai, 2011). When the latter is compromised in model organisms, the level of deleterious mutations elevates to a degree that may have phenotypic consequences (Valenci et al., 2015). Accordingly, Ferree et al. (2013) showed that knockout of mitophagy factors led to an accumulation of partly dysfunctional, aged mitochondria. Hence, mitochondria within single mammalian cells are frequently not uniform in function (Aryaman et al., 2018), and hence will be under differential selective constraints between cell types, tissues, and cells in the same tissue. Taken together, functional differences in mitochondrial activities are observed not only at the organism level, but also between and within cells.

Selection and Drift Shape the Mitochondrial Population During Female Germline Formation

Variation in heteroplasmy patterns between tissues could either result from somatic accumulation of mutations during the lifetime of an individual, or could be inherited. Therefore, one has to consider patterns of heteroplasmy already in the ovum, and specifically during the emergence of the maternal germ line. Indeed, much discussion revolved around the nature of mitochondrial bottleneck in the primordial maternal germ line in mouse and humans (Cree et al., 2008; Cao et al., 2009; Rebolledo-Jaramillo et al., 2014). A ∼1000-fold reduction in mtDNA content was estimated during the development of the human female germ cell (Freyer et al., 2012; Floros et al., 2018), followed by intense replication as cells migrate to form the gonad. Although neutral effects (genetic drift) are widely thought

to govern this process, one has to take into account that the female germline formation requires OXPHOS activity (Ginsburg et al., 1990). Therefore, it is logical in addition to neutrality that two types of selection come into play: firstly, positive selection, which prefers highly replicating variants, and secondly negative selection acts to remove variants with reduced ATP production (Wei et al., 2019). A relatively straightforward way to study the signatures of selection in the mtDNA of the germline is by comparing heteroplasmy patterns between mothers and offspring. This approach has recently been taken by Wei et al. by analyzing 1,526 human mother-offspring pairs (Wei et al., 2019). Firstly, higher levels of heteroplasmy across the entire mtDNA were observed in the mothers as compared to their offspring. Secondly, they identified, from the distribution of heteroplasmic mutations across the mtDNA sequence of the offspring, that these mutations were more frequent in the first or second codon positions as compared to the third position. Furthermore, while considering the entire coding mtDNA sequence, evolutionary conserved positions tended to retain heteroplasmy, in contrast to the tendency toward homoplasmy in positions with lesser conservation. Third, homoplasmic mutations were seldom found in evolutionary conserved positions. Fourth, most of the heteroplasmic mutations were characterized as *de novo* (found only in the offspring) or lost (found only in the mother), which again was best interpreted as the result of negative selection in the maternal germline. Finally, analysis of the mtDNA control region (D-loop) revealed reduced frequency of heteroplasmic positions within regions of regulatory importance for replication and transcription. Taken together, these findings support the action of negative selection during maternal germ line formation in the human mtDNA sequence. Although the mitochondrial population in the maternal germline determines the initial mitochondrial repertoire in the zygote, such population is still prone to the effects of evolutionary forces during differentiation, and in the lifetime of the organism.

Natural Selection Acts on the Mitochondria at the Whole Organism Level

The potential phenotypic impact, and hence the signatures of natural selection, on mitochondrial changes is mostly considered at the organism level (Stewart et al., 2008; Castellana et al., 2011). Probably the best example for such is disease-causing mutations, as thoroughly reviewed previously (Dowling, 2014), and hence will be discussed here only briefly. In 1988, the research group led by Douglas C Wallace was the first to have discovered a clear association between high levels of heteroplasmy of an mtDNA mutation and the tendency to develop a disease—Leber's hereditary optic neuropathy (LHON) (Wallace et al., 1988). In the same year, Ian Holt and others showed that a high level heteroplasmic mtDNA deletion led to mitochondrial myopathy (Holt et al., 1988). Ever since these discoveries, the association between certain threshold levels of heteroplasmy (∼85% for point mutations) with expression of disease phenotypes was found in multiple mitochondrial disorders (Stewart and Chinnery, 2015; Wallace, 2018). As mitochondrial diseases are relatively rare (Craven et al., 2017), the most logical explanation is that negative selection acts to remove such mutations from the population both in the mtDNA, and in mitochondrial genes encoded by the nucleus, although, in the latter, mutations could be retained in the population due to recessive modes of inheritance, and/or due to partial penetrance. Supportive evidence for such negative selection, while considering the mtDNA, came from large-scale deep sequence analysis of multiple individuals (mentioned above), which revealed that, although deleterious mutations are relatively prevalent in the human population, they only appear at very low heteroplasmy levels (Payne et al., 2013; Ye et al., 2014). Negative selection is also exemplified by the strong mutational bias in population genetics mtDNA variants (Gu et al., 2019). The differential accumulation pattern of mtDNA deleterious mutations in Drosophila males as compared to females, termed the "mother's curse", also supports the impact of natural selection on mtDNA sequences at the organism level (Innocenti et al., 2011), since in general, the mtDNA is maternally inherited. The signature of negative selection is also evident at the very early stages of embryo development, i.e., during oogenesis, as discussed above (De Fanti et al., 2017). Although bottleneck of mitochondrial transmission has been suggested to occur during the development of female germ cells in mice (Cree et al., 2008; Floros et al., 2018) and humans (Rebolledo-Jaramillo et al., 2014), divergence between mtDNA mutational patterns between tissues support the impact of natural selection (Latorre-Pellicer et al., 2016). Finally, certain inherited mtDNA mutations associate with altered tendency to develop age-related disorders (Marom et al., 2017), i.e., diseases whose onset is during post-reproductive age; similarly, large mtDNA deletions tend to accumulate preferentially in aging humans, again during post-reproductive ages (Arnheim and Cortopassi, 1992; Simonetti et al., 1992), hence with little effect on fitness. These pieces of evidence clearly attest for the impact of negative selection on mitochondrial function at the organism level.

Unlike negative selection, adaptive selection (i.e., positive selection) is less obvious to observe. The first evidence for signatures of positive selection in the human mtDNA has been demonstrated by the analysis of multiple whole mitochondrial genomes from individuals representing all major global populations (Mishmar et al., 2003; Ruiz-Pesini et al., 2004). These analyses correlated the pattern of ancient mtDNA mutations with global geographic distribution of mtDNA haplotypes, and argued for differential advantage of mtDNA haplogroups to survive in different climatic conditions (Mishmar et al., 2003). These predictions gained recent experimental support from the identification of sharp differentiation in the geographic distribution of Drosophila mtDNA haplotypes between cold and warm latitudinal regions in Australia: the haplotype that predominated low (subtropical) latitudes displayed greater resilience to heat than to cold stresses, as compared to haplotypes predominating higher (temperate) latitudes (Camus et al., 2017; Lajbner et al., 2018). Nevertheless, as mitochondrial function depends on interactions between mtDNA and nuclear DNA-encoded factors, one still awaits testing the impact of each of the tested mtDNA haplotypes in different nuclear genetic backgrounds.

Mito-Nuclear Interactions: Corresponding Signatures of Selection in Both the mtDNA and the nDNA

As mentioned above, one unique characteristic of the mitochondrial genomic system is the G x G interaction, or in other words, the interaction between factors encoded by the nuclear and mitochondrial genomes. The possible, and actual, impact of compatible vs. incompatible mito-nuclear genotypes were investigated in cell culture harboring different combinations of mtDNAs and nDNAs (cybrids) (Suissa et al., 2009; Gomez-Duran et al., 2010; Ji et al., 2012; Kenney et al., 2014; Crawford et al., 2018), and from repeated backcrossing of model organisms for the sake of generating animals with differential combinations of mito-nuclear genotypes (conplastic animals) originating from different strains of model organisms (Dingley et al., 2014; Latorre-Pellicer et al., 2016). Additionally, population genetics studies from species and different population isolates from the same species in the copepod *Tigriopus californicus* (Burton et al., 2006; Ellison and Burton, 2006), in reptiles (chameleons) (Bar-Yaacov et al., 2015) and in birds (sparrows) (Trier et al., 2014), revealed hybrid incompatibility, which constitutes an important step toward speciation. These pieces of evidence provide strong support for the importance of genetic compatibility between the nuclear and mitochondrial genomes, that when interfered with can either lead to diseases (Gershoni et al., 2014) or lead to the creation of reproductive barriers in both invertebrates and in vertebrates (Gershoni et al., 2009; Trier et al., 2014; Telschow et al., 2019; Tobler et al., 2019). The underlying mechanism of mito-nuclear genetic compatibility has previously been thoroughly discussed at the protein-protein, RNA-protein and protein-mtDNA levels at the whole organism level (Bar-Yaacov et al., 2012; Levin et al., 2014; Hill, 2019; Hill et al., 2019). However, although mito-nuclear interactions occur at the inter-cellular and intracellular levels, they are currently more technically challenging to identify. One good example that might reflect possible intra-cellular mito-nuclear incompatibilities is the establishment of heteroplasmic animals harboring a mixed population of mitochondria from two different strains (Sharpley et al., 2012). Although the phenotypes of such heteroplasmic mice could result from defective interactions between mitochondria with different haplotypes in the cells, they could also result from differential interactions of the two mtDNA genotypes with products coming from the nuclear genome, as reflected in a similar study (Latorre-Pellicer et al., 2016). Finally, it has been shown in yeast that the mixture of mtDNA mulecules from two different strains have slowly "drifted" into a situation where the original mtDNA had outcompeted the donor mtDNA to reconstruct the original mtDNA-nDNA genotype combination (Lee et al., 2008). Taken together, genotype compatibility between the mitochondrial and nuclear genomes are likely important for life. Nevertheless, one has to consider the fact that human offspring that were born from fertilized eggs in which the nuclei were transferred to an enucleated donor ovum (i.e., the mitochondrial replacement therapy) had no apparent phenotypes (Hyslop et al., 2016). Although mito-nuclear incompatibility has been shown in mitochondrial replacement experiments in multiple species (Dobler et al., 2018), some have questioned its wide impact on mitochondrial activity (Eyre-Walker, 2017). Interestingly, a recent analysis of human admixed populations revealed that mtDNA copy number decreases with increasing discordance between nuclear and mtDNA ancestry (Zaidi and Makova, 2019). However, since the lack of phenotype in human offspring that were born from mitochondrial replacement therapy was only demonstrated for generation F1, and as backcrossing is not possible in humans, the long-term impact of mito-nuclear incompatibility in humans still awaits a longitudinal study in the years to come.

Identifying the Signatures of Selection—Assessing the Functional Potential of Changes in DNA Sequences

How can one predict the effects of selection by mere sequence analysis? As protein-coding genes served as the focus of many studies, and since the genetic code and mutations in protein-coding genes are well-investigated, many bioinformatics computational tools were developed to assess the potential functional implications in protein-coding genes. The logic underlying such analyses stem from the notion that functional potential of changes in protein-coding sequences range from missense mutations, which alter the amino-acid composition of the translation product, to non-sense mutations and splice variants (both exon inclusion or skipping). This not only changes the amino acid composition, but also alters the length of the resultant protein. The latter may alter protein domains, and hence may have the potential to fundamentally affect the activity of a given protein, and even its subcellular localization. In contrast, it is believed that by-and-large, lesser functional impact is caused by sequence alterations in codon positions that do not change the amino acid sequence (Stern et al., 2007). Nevertheless, such so-called synonymous mutations change the choice of tRNAs during the translation process, which is directly influenced by differences in the abundance of different tRNAs that recognize different codons of the same amino acid (Levin et al., 2013). Previously, we have shown that, in the human mtDNA, there is preference for codons recognized by mtDNA-encoded tRNAs, which are more prevalent in the mitochondria as compared to imported tRNAs (Levin et al., 2013). Notably, as some mtDNA protein-coding genes harbor codons which are not recognized by the mtDNA-encoded tRNAs, nuclear-DNA encoded tRNA come into play and are imported into the mitochondria; such import depends on the varying tRNA gene contents of the mitochondrial genomes of different organisms, such as yeast, plants and mammals [reviewed in: (Rubio and Hopper, 2011; Schneider, 2011)]. For example, $tRNA^{Gln}_{CUG}$, which is not encoded by both mouse and human mtDNAs, was identified *in-vivo* within the mitochondria of cells from both organisms, where it was incorporated into the translation machinery (Rubio et al., 2008). It was suggested that the import of such nuclear tRNA into the mitochondria occurs via an ATP-dependent mechanism, which is likely distinct from the general mitochondrial protein import machinery. Can the usage of a rare tRNA affect protein function? Synonymous mutations

in the nuclear DNA-encoded Multidrug Resistance 1(MDR1) gene affected protein folding (Kimchi-Sarfaty et al., 2007). Specifically, the authors argued that the slow incorporation of a rare $tRNA^{Gly}$ led to slight hindrance of translation, which in turn affected protein folding, and its subsequent functionality. tRNA sequences themselves are under selective constraints, yet it was previously argued that tRNA mutations in the mtDNA that cause diseases in humans can appear as polymorphic variants in other species (i.e. recurrent mutations), thus supporting the putative impact of epistatic interactions acting as functional compensation (Kern and Kondrashov, 2004; Breen et al., 2012; Levin et al., 2013; Levin and Mishmar, 2017). Accordingly, experiments in Drosophila suggested that, in some cases, tRNA synthetase undergoes coordinated changes alongside their corresponding tRNAs to maintain function (Meiklejohn et al., 2013; Holmbeck et al., 2015). The availability of ribosome profiling as a highly quantitative approach to assess rates of translation, and its recent adaptation to the mtDNA (Rooijers et al., 2013; Gao et al., 2017), serves as a promising future experimental approach to compare rates of mtDNA translation between samples differing in synonymous mtDNA mutations.

As mutations in protein-coding genes are easier to interpret, most computational tools that screen for the signatures of natural selection calculate the ratios of missense (non-synonymous) to synonymous mutations (Ka/Ks) in genes of interest. Another measurable parameter that may reflect signatures of selection is whether the mutations occurred in evolutionary conserved sequence positions, and physico-chemical properties of amino acids. Tools (Bank et al., 2014), such as SELECTON (Stern et al., 2007), PAML (Yang, 1997), and PANTHER (Thomas et al., 2003), are designed to perform such calculations. Sequence conservation is also a useful parameter to assess the functional impact of mutations in RNA genes, i.e., those transcripts that are not ultimately translated, including tRNAs, rRNAs and the various types of non-coding RNAs (e.g., long, small and microRNAs). Human mtDNA encodes for 2 rRNAs, 22 tRNAs, and few relatively recently discovered long and microRNAs (Mercer et al., 2011; Rackham et al., 2011). As such transcripts overlap the coding regions of protein-coding genes, it is difficult to distinguish the potential functional impact of mutations in such elements. Despite all of the above, although it is possible to assess the potential impact of mutations that alter codons but not the amino acid code (the so-called silent mutations) by assessing codon bias index (Levin et al., 2013), only minor phenotypic effects have been demonstrated regarding such in the mitochondria, leading some researchers to argue for lack of selective signatures in most mtDNA mutations (Stoeckle and Thaler, 2018). Nevertheless, phenotypic impact of a mutation is not only limited to the subsequent gene function—an apparently silent mutation could alter a previously un-noticed regulatory element (Blumberg et al., 2018). Indeed, it has previously been demonstrated that certain coding sequences could also act as regulatory elements both in the nucleus (Birnbaum et al., 2012) and in the mitochondria (Blumberg et al., 2014). However, unlike the functional potential of mutations in genes, which could be assessed using the above-mentioned parameters, the functionality of mutations in mtDNA regulatory elements is much harder to predict, mainly since the "language and grammar" of these mtDNA regulatory elements are yet to be deciphered. In the subsequent sub-section, we will discuss efforts to assess the impact of mtDNA sequence variation on mtDNA regulation.

The Impact of mtDNA Mutations on Regulation of Gene Expression and mtDNA Copy Number

As mentioned above, although changes in a given sequence will not affect the amino acid sequence, or the activity of an RNA gene, a regulatory element which might reside within the same sequence could suffer from the same mutation and lead to altered transcription (Blumberg et al., 2014, 2017, 2018), replication, and possibly recombination or repair. Sequencing the human mtDNA in multiple tissues (Samuels et al., 2013) revealed that certain tissues (kidney, liver and skeletal muscle), shared the same recurrent heteroplasmic mutations, all in regulatory mtDNA regions, which were undetectable in other tissues in the same individuals, supporting non-random mutational segregation. Previously, ourselves and others showed that changes in mtDNA regulatory elements affect mtDNA regulation *in vitro* (Asari et al., 2007; Suissa et al., 2009) and in cytoplasmic hybrids (cybrids) (Gomez-Duran et al., 2010; Kenney et al., 2014). RNA-seq analysis of multiple lymphoblastoid cell lines demonstrated an association of ancient mtDNA SNPs and linked sets of mutations (haplogroups) with altered mtDNA gene expression (Cohen et al., 2016). In the latter study, DNA SNPs in the nucleus marked nuclear regulators of mtDNA gene expression as candidate modulators of such association. Similarly, a recent study used expression SNPs (eSNP) association in the nuclear genome to identify nuclear regulators of mtDNA gene expression, while using mtDNA genetic backgrounds as co-variants (Ali et al., 2019). These association studies with gene expression suggest that co-adaptation of the nuclear and mitochondrial genomes should also be extended to co-regulation of the two genomes. Indeed, analysis of ∼8500 RNA-seq experiments from multiple human individuals sampled from 48 different human body sites demonstrated mito-nuclear co-expression and co-regulation of gene expression (Barshad et al., 2018). Hence, compatibility of the two genomes should be extended beyond the direct interactions of proteins within the OXPHOS system to a regulatory cross talk. In summary, as a growing set of functional genomics tools are being adapted to investigate mitochondrial regulation (mtDNA gene expression, replication, translation and even repair), one will be able to directly assess the impact of mtDNA SNPs on mtDNA regulation in cells and organisms in the near future.

SUMMARY AND CONCLUSIONS

In the current essay, we argued that selection is a significant force that shaped, and still shapes, the landscape of mtDNA variation at the organismal, cellular, and single mitochondrial levels. To support this argument, we brought ample evidence from a variety of experimental systems, and demonstrated newly-generated evidence from tissues and cells, enabled by the adaptation of next generation sequencing techniques to investigate functional

genomics of the mitochondrion. At the organism level, it is worth mentioning the multiple disease-association studies that identified association of mtDNA genetic variants with altered susceptibility to develop genetic disorders, which strongly attest for the phenotypic impact of mtDNA mutations and hence their "visibility" to selective constraints (Marom et al., 2017). The fact that such association studies reveal positive association in certain populations, but not in others, supports the likelihood of an epistatic impact on nuclear genotypes as modifying factors, which compensate for the phenotypic impact of mtDNA variants, thus supporting the importance of mito-nuclear genotype compatibility. The three levels of functional impact on the mitochondria, e.g., the organism, cell, and single mitochondrion, set forth another dimension for considering mitochondrial phenotypes and assessment of the functionality of mitochondrial mutations. The availability of single cell functional genomics data, and consideration of the population of RNA transcripts per cell suggests, that the logic presented in the current essay may also imply to discussion of phenotypes caused by changes in nuclear genes in other, non-mitochondrial, systems.

AUTHOR CONTRIBUTIONS

DM and NS wrote the paper. DM conceived the idea and supervised the analyses and literature.

FUNDING

The study was funded by grants from the Israeli Science Foundation (372/17) and by the Life Sciences division of the US army (LS67993), awarded to DM. Both agents were not involved and did not influence the writing of the current manuscript.

ACKNOWLEDGMENTS

This study was supported by research grants from the Israel Science Foundation (ISF grant 372/17) and by the US army life sciences division grant (LS67993) awarded to DM.

REFERENCES

Ali, A. T., Boehme, L., Carbajosa, G., Seitan, V. C., Small, K. S., and Hodgkinson, A. (2019). Nuclear genetic regulation of the human mitochondrial transcriptome. *Elife* 8:41927. doi: 10.7554/eLife.41927

Arnheim, N., and Cortopassi, G. (1992). Deleterious mitochondrial DNA mutations accumulate in aging human tissues. *Mutat. Res.* 275, 157–167. doi: 10.1016/0921-8734(92)90020-P

Aryaman, J., Johnston, I. G., and Jones, N. S. (2018). Mitochondrial Heterogeneity. *Front. Genet.* 9:718. doi: 10.3389/fgene.2018.00718

Asari, M., Tan, Y., Watanabe, S., Shimizu, K., and Shiono, H. (2007). Effect of length variations at nucleotide positions 303-315 in human mitochondrial DNA on transcription termination. *Biochem. Biophys. Res. Commun.* 361, 641–644. doi: 10.1016/j.bbrc.2007.07.055

Avital, G., Buchshtav, M., Zhidkov, I., Tuval Feder, J., Dadon, S., Rubin, E., et al. (2012). Mitochondrial DNA heteroplasmy in diabetes and normal adults: role of acquired and inherited mutational patterns in twins. *Hum. Mol. Genet.* 21, 4214–4224. doi: 10.1093/hmg/dds245

Bank, C., Ewing, G. B., Ferrer-Admettla, A., Foll, M., and Jensen, J. D. (2014). Thinking too positive? Revisiting current methods of population genetic selection inference. *Trends Genet.* 30, 540–546. doi: 10.1016/j.tig.2014.09.010

Barshad, G., Blumberg, A., Cohen, T., and Mishmar, D. (2018). Human primitive brain displays negative mitochondrial-nuclear expression correlation of respiratory genes. *Genome Res.* 28, 952–967. doi: 10.1101/gr.226324.117

Bar-Yaacov, D., Blumberg, A., and Mishmar, D. (2012). Mitochondrial-nuclear co-evolution and its effects on OXPHOS activity and regulation. *Biochim. Biophys. Acta* 1819, 1107–1111. doi: 10.1016/j.bbagrm.2011.10.008

Bar-Yaacov, D., Hadjivasiliou, Z., Levin, L., Barshad, G., Zarivach, R., Bouskila, A., et al. (2015). Mitochondrial Involvement in Vertebrate Speciation? The case of mito-nuclear genetic divergence in Chameleons. *Genome Biol. Evol.* 7, 3322–3336. doi: 10.1093/gbe/evv226

Birky, C. W., Maruyama, T., and Fuerst, P. (1983). An approach to population and evolutionary genetic theory for genes in mitochondria and chloroplasts, and some results. *Genetics* 103, 513–527.

Birnbaum, R. Y., Clowney, E. J., Agamy, O., Kim, M. J., Zhao, J., Yamanaka, T., et al. (2012). Coding exons function as tissue-specific enhancers of nearby genes. *Genome Res.* 22, 1059–1068. doi: 10.1101/gr.133546.111

Blumberg, A., Danko, C. G., Kundaje, A., and Mishmar, D. (2018). A common pattern of DNase I footprinting throughout the human mtDNA unveils clues for a chromatin-like organization. *Genome Res.* 28, 1158–1168. doi: 10.1101/gr.230409.117

Blumberg, A., Rice, E. J., Kundaje, A., Danko, C. G., and Mishmar, D. (2017). Initiation of mtDNA transcription is followed by pausing, and diverges across human cell types and during evolution. *Genome Res.* 27, 362–373. doi: 10.1101/gr.209924.116

Blumberg, A., Sailaja, B. S., Kundaje, A., Levin, L., Dadon, S., Shmorak, S., et al. (2014). Transcription factors bind negatively-selected sites within human mtDNA genes. *Genome Biol. Evol.* 6, 2634–2646. doi: 10.1093/gbe/evu210

Breen, M. S., Kemena, C., Vlasov, P. K., Notredame, C., and Kondrashov, F. A. (2012). Epistasis as the primary factor in molecular evolution. *Nature* 490, 535–538. doi: 10.1038/nature11510

Burton, R. S., Ellison, C. K., and Harrison, J. S. (2006). The sorry state of F2 hybrids: consequences of rapid mitochondrial DNA evolution in allopatric populations. *Am. Nat.* 168(Suppl. 6), S14–24. doi: 10.1086/509046

Camus, M. F., Wolff, J. N., Sgro, C. M., and Dowling, D. K. (2017). Experimental support that natural selection has shaped the latitudinal distribution of mitochondrial haplotypes in australian *Drosophila melanogaster*. *Mol. Biol. Evol.* 34, 2600–2612. doi: 10.1093/molbev/msx184

Cao, L., Shitara, H., Sugimoto, M., Hayashi, J., Abe, K., and Yonekawa, H. (2009). New evidence confirms that the mitochondrial bottleneck is generated without reduction of mitochondrial DNA content in early primordial germ cells of mice. *PLoS Genet.* 5:e1000756. doi: 10.1371/journal.pgen.1000756

Castellana, S., Vicario, S., and Saccone, C. (2011). Evolutionary patterns of the mitochondrial genome in Metazoa: exploring the role of mutation and selection in mitochondrial protein coding genes. *Genome Biol. Evol.* 3, 1067–1079. doi: 10.1093/gbe/evr040

Cohen, T., Levin, L., and Mishmar, D. (2016). Ancient out-of-africa mitochondrial DNA variants associate with distinct mitochondrial gene expression patterns. *PLoS Genet.* 12:e1006407. doi: 10.1371/journal.pgen.1006407

Craven, L., Alston, C. L., Taylor, R. W., and Turnbull, D. M. (2017). Recent advances in mitochondrial disease. *Annu. Rev. Genomics Hum. Genet.* 18, 257–275. doi: 10.1146/annurev-genom-091416-035426

Crawford, N., Prendergast, D., Oehlert, J. W., Shaw, G. M., Stevenson, D. K., Rappaport, N., et al. (2018). Divergent patterns of mitochondrial and nuclear ancestry are associated with the risk for preterm birth. *J. Pediatr.* 194, 40–46 e4. doi: 10.1016/j.jpeds.2017.10.052

Cree, L. M., Samuels, D. C., de Sousa Lopes, S. C., Rajasimha, H. K., Wonnapinij, P., Mann, J. R., et al. (2008). A reduction of mitochondrial DNA molecules during embryogenesis explains the rapid segregation of genotypes. *Nat. Genet.* 40, 249–254. doi: 10.1038/ng.2007.63

De Fanti, S., Vicario, S., Lang, M., Simone, D., Magli, C., Luiselli, D., et al. (2017). Intra-individual purifying selection on mitochondrial DNA variants

during human oogenesis. *Hum. Reprod.* 32, 1100–1107. doi: 10.1093/humrep/dex051

Dingley, S. D., Polyak, E., Ostrovsky, J., Srinivasan, S., Lee, I., Rosenfeld, A. B., et al. (2014). Mitochondrial DNA variant in COX1 subunit significantly alters energy metabolism of geographically divergent wild isolates in Caenorhabditis elegans. *J. Mol. Biol.* 426, 2199–2216. doi: 10.1016/j.jmb.2014.02.009

Dobler, R., Dowling, D. K., Morrow, E. H., and Reinhardt, K. (2018). A systematic review and meta-analysis reveals pervasive effects of germline mitochondrial replacement on components of health. *Hum. Reprod. Update* 24, 519–534. doi: 10.1093/humupd/dmy018

Dowling, D. K. (2014). Evolutionary perspectives on the links between mitochondrial genotype and disease phenotype. *Biochim. Biophys. Acta* 1840, 1393–1403. doi: 10.1016/j.bbagen.2013.11.013

Ellison, C. K., and Burton, R. S. (2006). Disruption of mitochondrial function in interpopulation hybrids of Tigriopus californicus. *Evol. Int. J. Org. Evol.* 60, 1382–1391. doi: 10.1111/j.0014-3820.2006.tb01217.x

Eyre-Walker, A. (2017). Mitochondrial replacement therapy: are mito-nuclear interactions likely to be a problem? *Genetics* 205, 1365–1372. doi: 10.1534/genetics.116.196436

Ferree, A. W., Trudeau, K., Zik, E., Benador, I. Y., Twig, G., Gottlieb, R. A., et al. (2013). MitoTimer probe reveals the impact of autophagy, fusion, and motility on subcellular distribution of young and old mitochondrial protein and on relative mitochondrial protein age. *Autophagy* 9, 1887–1896. doi: 10.4161/auto.26503

Filograna, R., Koolmeister, C., Upadhyay, M., Pajak, A., Clemente, P., Wibom, R., et al. (2019). Modulation of mtDNA copy number ameliorates the pathological consequences of a heteroplasmic mtDNA mutation in the mouse. *Sci. Adv.* 5:eaav9824. doi: 10.1126/sciadv.aav9824

Floros, V. I., Pyle, A., Dietmann, S., Wei, W., Tang, W. C. W., Irie, N., et al. (2018). Segregation of mitochondrial DNA heteroplasmy through a developmental genetic bottleneck in human embryos. *Nat. Cell Biol.* 20, 144–151. doi: 10.1038/s41556-017-0017-8

Freyer, C., Cree, L. M., Mourier, A., Stewart, J. B., Koolmeister, C., Milenkovic, D., et al. (2012). Variation in germline mtDNA heteroplasmy is determined prenatally but modified during subsequent transmission. *Nat. Genet.* 44, 1282–1285. doi: 10.1038/ng.2427

Gao, F., Wesolowska, M., Agami, R., Rooijers, K., Loayza-Puch, F., Lawless, C., et al. (2017). Using mitoribosomal profiling to investigate human mitochondrial translation. *Wellcome Open Res.* 2:116. doi: 10.12688/wellcomeopenres.13119.1

Gershoni, M., Levin, L., Ovadia, O., Toiw, Y., Shani, N., Dadon, S., et al. (2014). Disrupting mitochondrial-nuclear co-evolution affects OXPHOS complex I integrity and impacts human health. *Genome Biol. Evol.* 6, 2665–2680. doi: 10.1093/gbe/evu208

Gershoni, M., Templeton, A. R., and Mishmar, D. (2009). Mitochondrial bioenergetics as a major motive force of speciation. *Bioessays* 31, 642–650. doi: 10.1002/bies.200800139

Ginsburg, M., Snow, M. H., and McLaren, A. (1990). Primordial germ cells in the mouse embryo during gastrulation. *Development* 110, 521–528.

Gomez-Duran, A., Pacheu-Grau, D., Lopez-Gallardo, E., Diez-Sanchez, C., Montoya, J., Lopez-Perez, M. J., et al. (2010). Unmasking the causes of multifactorial disorders: OXPHOS differences between mitochondrial haplogroups. *Hum. Mol. Genet.* 19, 3343–3353. doi: 10.1093/hmg/ddq246

Goto, H., Dickins, B., Afgan, E., Paul, I. M., Taylor, J., Makova, K. D., et al. (2011). Dynamics of mitochondrial heteroplasmy in three families investigated via a repeatable re-sequencing study. *Genome Biol.* 12:R59. doi: 10.1186/gb-2011-12-6-r59

Gu, X., Kang, X., and Liu, J. (2019). Mutation signatures in germline mitochondrial genome provide insights into human mitochondrial evolution and disease. *Hum. Genet.* 138, 613–624. doi: 10.1007/s00439-019-02009-5

Hill, G. E. (2019). Reconciling the mitonuclear compatibility species concept with rampant mitochondrial introgression. *Integr. Comp. Biol.* doi: 10.1093/icb/icz019

Hill, G. E., Havird, J. C., Sloan, D. B., Burton, R. S., Greening, C., and Dowling, D. K. (2019). Assessing the fitness consequences of mitonuclear interactions in natural populations. *Biol. Rev. Camb. Philos. Soc.* 94, 1089–1104. doi: 10.1111/brv.12493

Holmbeck, M. A., Donner, J. R., Villa-Cuesta, E., and Rand, D. M. (2015). A Drosophila model for mito-nuclear diseases generated by an incompatible interaction between tRNA and tRNA synthetase. *Dis. Model Mech.* 8, 843–854. doi: 10.1242/dmm.019323

Holt, I. J., Harding, A. E., and Morgan-Hughes, J. A. (1988). Deletions of muscle mitochondrial DNA in patients with mitochondrial myopathies. *Nature* 331, 717–719. doi: 10.1038/331717a0

Hyslop, L. A., Blakeley, P., Craven, L., Richardson, J., Fogarty, N. M., Fragouli, E., et al. (2016). Towards clinical application of pronuclear transfer to prevent mitochondrial DNA disease. *Nature* 534, 383–386. doi: 10.1038/nature18303

Innocenti, P., Morrow, E. H., and Dowling, D. K. (2011). Experimental evidence supports a sex-specific selective sieve in mitochondrial genome evolution. *Science* 332, 845–848. doi: 10.1126/science.1201157

Jayaprakash, A. D., Benson, E. K., Gone, S., Liang, R., Shim, J., Lambertini, L., et al. (2015). Stable heteroplasmy at the single-cell level is facilitated by intercellular exchange of mtDNA. *Nucleic Acids Res.* 43, 2177–2187. doi: 10.1093/nar/gkv052

Ji, F., Sharpley, M. S., Derbeneva, O., Alves, L. S., Qian, P., Wang, Y., et al. (2012). Mitochondrial DNA variant associated with Leber hereditary optic neuropathy and high-altitude Tibetans. *Proc. Natl. Acad. Sci. U.S.A.* 109, 7391–7396. doi: 10.1073/pnas.1202484109

Kenney, M. C., Chwa, M., Atilano, S. R., Falatoonzadeh, P., Ramirez, C., Malik, D., et al. (2014). Molecular and bioenergetic differences between cells with African versus European inherited mitochondrial DNA haplogroups: implications for population susceptibility to diseases. *Biochim. Biophys. Acta* 1842, 208–219. doi: 10.1016/j.bbadis.2013.10.016

Kern, A. D., and Kondrashov, F. A. (2004). Mechanisms and convergence of compensatory evolution in mammalian mitochondrial tRNAs. *Nat. Genet.* 36, 1207–1212. doi: 10.1038/ng1451

Kimchi-Sarfaty, C., Oh, J. M., Kim, I. W., Sauna, Z. E., Calcagno, A. M., Ambudkar, S. V., et al. (2007). A "silent" polymorphism in the MDR1 gene changes substrate specificity. *Science* 315, 525–528. doi: 10.1126/science.1135308

Kimura, M. (1983). *The Neutral Theory of Molecular Evolution*. Cambridge: Cambridge University Press.

Kowald, A., and Kirkwood, T. B. (2013). Mitochondrial mutations and aging: random drift is insufficient to explain the accumulation of mitochondrial deletion mutants in short-lived animals. *Aging Cell* 12, 728–731. doi: 10.1111/acel.12098

Lajbner, Z., Pnini, R., Camus, M. F., Miller, J., and Dowling, D. K. (2018). Experimental evidence that thermal selection shapes mitochondrial genome evolution. *Sci. Rep.* 8:9500. doi: 10.1038/s41598-018-27805-3

Lane, N., and Martin, W. (2010). The energetics of genome complexity. *Nature* 467, 929–934. doi: 10.1038/nature09486

Latorre-Pellicer, A., Moreno-Loshuertos, R., Lechuga-Vieco, A. V., Sanchez-Cabo, F., Torroja, C., Acin-Perez, R., et al. (2016). Mitochondrial and nuclear DNA matching shapes metabolism and healthy ageing. *Nature* 535, 561–565. doi: 10.1038/nature18618

Lee, H. Y., Chou, J. Y., Cheong, L., Chang, N. H., Yang, S. Y., and Leu, J. Y. (2008). Incompatibility of nuclear and mitochondrial genomes causes hybrid sterility between two yeast species. *Cell* 135, 1065–1073. doi: 10.1016/j.cell.2008.10.047

Levin, L., Blumberg, A., Barshad, G., and Mishmar, D. (2014). Mito-nuclear co-evolution: the positive and negative sides of functional ancient mutations. *Front. Genet.* 5:448. doi: 10.3389/fgene.2014.00448

Levin, L., and Mishmar, D. (2017). The genomic landscape of evolutionary convergence in mammals, birds and reptiles. *Nat. Ecol. Evol.* 1:0041. doi: 10.1038/s41559-016-0041

Levin, L., Zhidkov, I., Gurman, Y., Hawlena, H., and Mishmar, D. (2013). Functional recurrent mutations in the human mitochondrial phylogeny - dual roles in evolution and disease. *Genome Biol. Evol.* 5, 876–890. doi: 10.1093/gbe/evt058

Li, M., Schonberg, A., Schaefer, M., Schroeder, R., Nasidze, I., and Stoneking, M. (2010). Detecting heteroplasmy from high-throughput sequencing of complete human mitochondrial DNA genomes. *Am. J. Hum. Genet.* 87, 237–249. doi: 10.1016/j.ajhg.2010.07.014

Ludwig, L. S., Lareau, C. A., Ulirsch, J. C., Christian, E., Muus, C., Li, L. H., et al. (2019). Lineage tracing in humans enabled by mitochondrial mutations and single-cell genomics. *Cell* 176, 1325–1339 e22. doi: 10.1016/j.cell.2019.01.022

Marom, S., Friger, M., and Mishmar, D. (2017). MtDNA meta-analysis reveals both phenotype specificity and allele heterogeneity: a model for differential association. *Sci. Rep*. 7:43449. doi: 10.1038/srep43449

Meiklejohn, C. D., Holmbeck, M. A., Siddiq, M. A., Abt, D. N., Rand, D. M., and Montooth, K. L. (2013). An Incompatibility between a mitochondrial tRNA and its nuclear-encoded tRNA synthetase compromises development and fitness in Drosophila. *PLoS Genet*. 9:e1003238. doi: 10.1371/journal.pgen.1003238

Meiklejohn, C. D., Montooth, K. L., and Rand, D. M. (2007). Positive and negative selection on the mitochondrial genome. *Trends Genet*. 23, 259–263. doi: 10.1016/j.tig.2007.03.008

Mercer, T. R., Neph, S., Dinger, M. E., Crawford, J., Smith, M. A., Shearwood, A. M., et al. (2011). The human mitochondrial transcriptome. *Cell* 146, 645–658. doi: 10.1016/j.cell.2011.06.051

Mishmar, D., Ruiz-Pesini, E., Golik, P., Macaulay, V., Clark, A. G., Hosseini, S., et al. (2003). Natural selection shaped regional mtDNA variation in humans. *Proc Natl Acad Sci U.S.A*. 100, 171–176. doi: 10.1073/pnas.0136972100

Morris, J., Na, Y. J., Zhu, H., Lee, J. H., Giang, H., Ulyanova, A. V., et al. (2017). Pervasive within-mitochondrion single-nucleotide variant heteroplasmy as revealed by single-mitochondrion sequencing. *Cell Rep*. 21, 2706–2713. doi: 10.1016/j.celrep.2017.11.031

Payne, B. A., Wilson, I. J., Yu-Wai-Man, P., Coxhead, J., Deehan, D., Horvath, R., et al. (2013). Universal heteroplasmy of human mitochondrial DNA. *Hum. Mol. Genet.* 22, 384–390. doi: 10.1093/hmg/dds435

Rackham, O., Shearwood, A. M., Mercer, T. R., Davies, S. M., Mattick, J. S., and Filipovska, A. (2011). Long noncoding RNAs are generated from the mitochondrial genome and regulated by nuclear-encoded proteins. *RNA* 17, 2085–2093. doi: 10.1261/rna.029405.111

Rebolledo-Jaramillo, B., Su, M. S., Stoler, N., McElhoe, J. A., Dickins, B., Blankenberg, D., et al. (2014). Maternal age effect and severe germ-line bottleneck in the inheritance of human mitochondrial DNA. *Proc Natl Acad Sci U.S.A*. 111, 15474–15479. doi: 10.1073/pnas.1409328111

Reiner, J. E., Kishore, R. B., Levin, B. C., Albanetti, T., Boire, N., Knipe, A., et al. (2010). Detection of heteroplasmic mitochondrial DNA in single mitochondria. *PLoS ONE*. 5:e14359. doi: 10.1371/journal.pone.0014359

Rooijers, K., Loayza-Puch, F., Nijtmans, L. G., and Agami, R. (2013). Ribosome profiling reveals features of normal and disease-associated mitochondrial translation. *Nat. Commun*. 4:2886. doi: 10.1038/ncomms3886

Rubio, M. A., and Hopper, A. K. (2011). Transfer RNA travels from the cytoplasm to organelles. *Wiley Interdiscip. Rev. RNA* 2, 802–817. doi: 10.1002/wrna.93

Rubio, M. A., Rinehart, J. J., Krett, B., Duvezin-Caubet, S., Reichert, A. S., Soll, D., et al. (2008). Mammalian mitochondria have the innate ability to import tRNAs by a mechanism distinct from protein import. *Proc. Natl. Acad. Sci. U.S.A*. 105, 9186–9191. doi: 10.1073/pnas.0804283105

Ruiz-Pesini, E., Mishmar, D., Brandon, M., Procaccio, V., and Wallace, D. C. (2004). Effects of purifying and adaptive selection on regional variation in human mtDNA. *Science* 303, 223–226. doi: 10.1126/science.1088434

Sagan, L. (1967). On the origin of mitosing cells. *J. Theor. Biol*. 14, 255–274. doi: 10.1016/0022-5193(67)90079-3

Samuels, D. C., Li, C., Li, B., Song, Z., Torstenson, E., Boyd Clay, H., et al. (2013). Recurrent tissue-specific mtDNA mutations are common in humans. *PLoS Genet*. 9:e1003929. doi: 10.1371/journal.pgen.1003929

Schneider, A. (2011). Mitochondrial tRNA import and its consequences for mitochondrial translation. *Annu. Rev. Biochem*. 80, 1033–1053. doi: 10.1146/annurev-biochem-060109-092838

Schon, E. A., and Gilkerson, R. W. (2010). Functional complementation of mitochondrial DNAs: mobilizing mitochondrial genetics against dysfunction. *Biochim. Biophys. Acta* 1800, 245–249. doi: 10.1016/j.bbagen.2009.07.007

Sharpley, M. S., Marciniak, C., Eckel-Mahan, K., McManus, M., Crimi, M., Waymire, K., et al. (2012). Heteroplasmy of mouse mtDNA is genetically unstable and results in altered behavior and cognition. *Cell* 151, 333–343. doi: 10.1016/j.cell.2012.09.004

Simonetti, S., Chen, X., DiMauro, S., and Schon, E. A. (1992). Accumulation of deletions in human mitochondrial DNA during normal aging: analysis by quantitative PCR. *Biochim. Biophys. Acta* 1180, 113–122. doi: 10.1016/0925-4439(92)90059-V

Stern, A., Doron-Faigenboim, A., Erez, E., Martz, E., Bacharach, E., and Pupko, T. (2007). Selecton 2007: advanced models for detecting positive and purifying selection using a Bayesian inference approach. *Nucleic Acids Res*. 35, W506–W511. doi: 10.1093/nar/gkm382

Stewart, J. B., and Chinnery, P. F. (2015). The dynamics of mitochondrial DNA heteroplasmy: implications for human health and disease. *Nat. Rev. Genet*. 16, 530–542. doi: 10.1038/nrg3966

Stewart, J. B., Freyer, C., Elson, J. L., Wredenberg, A., Cansu, Z., Trifunovic, A., et al. (2008). Strong purifying selection in transmission of mammalian mitochondrial DNA. *PLoS Biol*. 6:e10. doi: 10.1371/journal.pbio.0060010

Stoeckle, M. Y., and Thaler, D. S. (2018). Why should mitochondria define species? *BioRxiv* doi: 10.1101/276717

Suissa, S., Wang, Z., Poole, J., Wittkopp, S., Feder, J., Shutt, T. E., et al. (2009). Ancient mtDNA genetic variants modulate mtDNA transcription and replication. *PLoS Genet*. 5:e1000474. doi: 10.1371/journal.pgen.1000474

Telschow, A., Gadau, J., Werren, J. H., and Kobayashi, Y. (2019). Genetic incompatibilities between mitochondria and nuclear genes: effect on gene flow and speciation. *Front. Genet*. 10:62. doi: 10.3389/fgene.2019.00062

Thomas, P. D., Campbell, M. J., Kejariwal, A., Mi, H., Karlak, B., Daverman, R., et al. (2003). PANTHER: a library of protein families and subfamilies indexed by function. *Genome Res*. 13, 2129–2141. doi: 10.1101/gr.772403

Tobler, M., Barts, N., and Greenway, R. (2019). Mitochondria and the origin of species: bridging genetic and ecological perspectives on speciation processes. *Integr. Comp. Biol*. doi: 10.1093/icb/icz025

Trier, C. N., Hermansen, J. S., Saetre, G. P., and Bailey, R. I. (2014). Evidence for mito-nuclear and sex-linked reproductive barriers between the Hybrid Italian sparrow and its parent species. *PLoS Genet*. 10:e1004075. doi: 10.1371/journal.pgen.1004075

Twig, G., and Shirihai, O. S. (2011). The interplay between mitochondrial dynamics and mitophagy. *Antioxidants Redox Signal.* 14, 1939–1951. doi: 10.1089/ars.2010.3779

Valenci, I., Yonai, L., Bar-Yaacov, D., Mishmar, D., and Ben-Zvi, A. (2015). Parkin modulates heteroplasmy of truncated mtDNA in Caenorhabditis elegans. *Mitochondrion* 20, 64–70. doi: 10.1016/j.mito.2014.11.001

Wallace, D. C. (2018). Mitochondrial genetic medicine. *Nat. Genet*. 50, 1642–1649. doi: 10.1038/s41588-018-0264-z

Wallace, D. C., Singh, G., Lott, M. T., Hodge, J. A., Schurr, T. G., Lezza, A. M., et al. (1988). Mitochondrial DNA mutation associated with Leber's hereditary optic neuropathy. *Science* 242, 1427–1430. doi: 10.1126/science.3201231

Wei, W., Tuna, S., Keogh, M. J., Smith, K. R., Aitman, T. J., Beales, P. L., et al. (2019). Germline selection shapes human mitochondrial DNA diversity. *Science* 364. doi: 10.1126/science.aau6520

Yang, Z. (1997). PAML: a program package for phylogenetic analysis by maximum likelihood. *CABIOS* 13, 555–556. doi: 10.1093/bioinformatics/13.5.555

Ye, K., Lu, J., Ma, F., Keinan, A., and Gu, Z. (2014). Extensive pathogenicity of mitochondrial heteroplasmy in healthy human individuals. *Proc. Natl. Acad. Sci. U.S.A*. 111, 10654–10659. doi: 10.1073/pnas.1403521111

Zaidi, A. A., and Makova, K. D. (2019). Investigating mitonuclear interactions in human admixed populations. *Nat. Ecol. Evol*. 3, 213–222. doi: 10.1038/s41559-018-0766-1

Zhu, C. T., Ingelmo, P., and Rand, D. M. (2014). GxGxE for lifespan in drosophila: mitochondrial, nuclear, and dietary interactions that modify longevity. *PLoS Genet*. 10:e1004354. doi: 10.1371/journal.pgen.1004354

Conflict of Interest Statement: 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.

frontiers in Ecology and Evolution

PERSPECTIVE
published: 04 December 2019
doi: 10.3389/fevo.2019.00469

Filling in the Gaps: Adopting Ultraconserved Elements Alongside COI to Strengthen Metabarcoding Studies

*Mac P. Pierce**

School of Public Health, University of Hong Kong, Hong Kong, Hong Kong

Metabarcoding is rapidly gaining popularity as a means of conducting biodiversity studies. Using DNA barcodes to identify and catalog biodiversity has many advantages, and compares favorably with traditional methods based on morphological examination. Ease of use, taxonomic coverage, and increased efficiency are qualities that make metabarcoding a valuable ecological tool, particularly in light of the drastic anthropogenically induced ecosystem changes currently underway. However, limitations and challenges pertaining to existing barcodes create gaps from which inaccuracies can arise, contributing to skepticism regarding the value of metabarcoding based methods. Developing novel ways to address these limitations is crucial to improve metabarcoding methods and dispel doubt about their utility. Ultraconserved genomic elements (UCEs), genetic markers that have been used successfully in the field of phylogenomics, possess advantageous qualities that may be applied to fill in the gaps of existing metabarcoding methods. Here, I outline the strengths of UCEs and discuss their potential for complementing and strengthening existing metabarcoding methods based on the mitochondrial marker cytochrome oxidase I (COI).

Keywords: biomonitoring, DNA barcodes, marker multiplexing, metabarcoding, ultraconserved elements, biodiversity surveys

OPEN ACCESS

Edited by:
Rodney L. Honeycutt,
Pepperdine University, United States

Reviewed by:
Brant Faircloth,
Louisiana State University, United States
Hugo J. De Boer,
University of Oslo, Norway

***Correspondence:**
Mac P. Pierce
macppierce@gmail.com

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

Received: *01 July 2019*
Accepted: *21 November 2019*
Published: *04 December 2019*

Citation:
Pierce MP (2019) Filling in the Gaps: Adopting Ultraconserved Elements Alongside COI to Strengthen Metabarcoding Studies. Front. Ecol. Evol. 7:469. doi: 10.3389/fevo.2019.00469

INTRODUCTION

Researchers are increasingly using metabarcoding to address questions across a wide range of scientific fields. For example, metabarcoding has been used in recent studies to assess parasitism in an invasive species (Kitson et al., 2018), characterize hidden cryptic diversity in a reef ecosystem (Carvalho et al., 2019), and identify dietary choices of a prairie bird (Sullins et al., 2018), to name just a few. Over recent years, studies evaluating the performance of these methods have consistently demonstrated that metabarcoding can match, and in many cases exceed, the performance of traditional morphology-based methods (Ji et al., 2013; Deiner et al., 2017; Bush et al., 2019). Particular strengths of metabarcoding include taxonomic comprehensiveness and resolution, independence from taxonomic expertise, ability to overcome misidentifications, and efficiency in terms of time, manpower, and cost (Ji et al., 2013; Bush et al., 2019).

However, significant limitations and challenges to metabarcoding remain (Zinger et al., 2019). These include inherent issues like estimating abundance (Piñol et al., 2018), as well as logistical challenges such as selecting robust barcodes that work accurately across a wide taxonomic range

(Kress et al., 2015). Barcodes must meet certain criteria (Taberlet et al., 2007), and no universal genetic marker meeting these criteria has yet been identified (Valentini et al., 2009). Consequently, a range of markers has emerged, each utilized by researchers focusing on different taxonomic groups (Porter and Hajibabaei, 2018). For animals, the mitochondrial gene cytochrome c oxidase subunit 1 (COI) has emerged as the most commonly used marker for barcoding. This marker choice has many advantages, as reflected by the extent to which it is used and its thorough coverage in reference databases, a critical point for effective metabarcoding studies (Andújar et al., 2018).

Unfortunately, several issues associated with COI create potential sources for error, including incomplete lineage sorting, heteroplasmy, introgression, and the presence of pseudogenes (Rubinoff et al., 2006). More importantly, COI does not amplify equally well across all animal taxa, a major limitation for metabarcoding surveys aiming to achieve maximal taxonomic coverage (Kress et al., 2015). Overcoming these obstacles is therefore an important goal, both to increase metabarcoding accuracy and to dispel skepticism inhibiting a more widespread adoption of metabarcoding methods. Solutions that are relatively easy to incorporate into existing pipelines, such as marker multiplexing (Zhang et al., 2018), will be especially valuable. A phylogenomic approach of growing popularity utilizing ultraconserved elements (UCEs) may provide a complementary approach. Primarily used for reconstructing evolutionary relationships, several unique qualities of UCEs make them promising candidates for complementing and strengthening COI-based metabarcoding studies.

ULTRACONSERVED ELEMENTS

UCEs are conserved genomic regions found in large numbers throughout the genome. They consist of a highly-conserved core region flanked by more variable sequence (Faircloth et al., 2012), and have been identified in a wide range of eukaryotic groups, including plants, fungi, invertebrates, and vertebrates (Siepel et al., 2005; Reneker et al., 2012). UCEs are identified by aligning two or more genomes and scanning for regions of high fidelity, from which bait sets are then designed to extract DNA fragments containing UCE regions during targeted enrichment (Faircloth, 2017). Several advantages of UCEs have made them valuable tools in phylogenomics, where they have been used to high success in a number of animal groups (McCormack et al., 2013; Branstetter et al., 2017a; Alfaro et al., 2018). These include their high level of sequence conservation, robustness to duplication, strong phylogenetic signal, and the large number of alternate UCE loci present in the nuclear genome (Derti et al., 2006; Stephen et al., 2008; Faircloth et al., 2012). These same advantages of UCEs have application to metabarcoding, and may help fill in the gaps created by the limitations of COI.

UCEs AND BARCODE CRITERIA

Not all genetic markers can be used as barcodes, and not all barcodes work equally well across or within taxonomic groups (Kress and Erickson, 2008). Selecting appropriate barcoding regions is critically important for biodiversity surveys, with major implications for the organismal groups that can be studied (Deagle et al., 2014), and candidate genetic markers should meet certain criteria (Taberlet et al., 2007; Valentini et al., 2009). Furthermore, the best choice of barcode will depend on the individual priorities and goals of a given study. Metabarcoding studies prioritize accurate, high throughput species recovery from samples of unknown taxonomic composition, typically containing degraded DNA, and therefore should use genetic markers with strengths in these areas (Taberlet et al., 2012). Although COI-based metabarcoding has been shown to work well-compared to traditional morphological approaches, its weaknesses limit the ability of metabarcoding studies to accurately recover the full range of animal species present in an environment. Incorporating other markers with complementary strengths will increase the accuracy and reliability of existing metabarcoding methods. Below, I summarize the relative strengths and weakness of UCEs and COI in the context of metabarcoding based on previously suggested criteria (Taberlet et al., 2007; Valentini et al., 2009), and discuss how UCEs may be used to complement and strengthen metabarcoding methods.

Species Discrimination

DNA barcodes should discriminate species effectively, having high intraspecific fidelity while being variable between species. COI has been used to successfully identify and differentiate species in many groups (Hebert et al., 2003), particularly when used as part of an integrative approach to taxonomy (Will et al., 2005; Janzen et al., 2009). However, issues like incomplete lineage sorting, heteroplasmy, introgression, and the existence of pseudogenes, may result in incongruence between the number and identity of COI sequences and species or populations represented in a sample, resulting in false estimates and misidentifications (Moritz and Cicero, 2004; Will et al., 2005; Rubinoff et al., 2006). Additionally, single-locus methods are vulnerable to overlapping character variation (Will et al., 2005). These issues limit the ability of COI to accurately and reliably differentiate species, particularly uncharacterized taxa. This is especially problematic for metabarcoding studies where using additional verification methods is generally not desirable. Conversely, UCEs are robust to such issues. UCE loci have been found to be depleted from duplicated gene regions, are present in high numbers throughout the genome, and the bait design workflow removes loci deemed likely to be paralogs (Derti et al., 2006; Stephen et al., 2008; Faircloth, 2017). Though UCEs may be occasionally duplicated in some taxa or missing in others, the large number of UCEs available can provide consensus estimates, and problematic UCE loci can be pruned from bait sets as these become more refined through the increasing availability of sequenced genomes.

Universal Standardization

A truly universal barcode will have functionality across the Tree of Life, working equally well across and within all groups. Because no genetic marker fitting this criterion has yet been identified, utilizing barcodes that work well across different

taxonomic ranges is the best possible alternative. As such, different *de facto* universal barcodes have emerged for different taxonomic groups (Taberlet et al., 2007). COI has become the *de facto* universal barcode for the metazoa. However, it is not equally effective across or within all metazoan lineages (Deagle et al., 2014). Some groups will require different barcodes, resulting in fragmentation of metabarcoding methodologies and sequence databases. Although no single UCE locus is likely to be universal, comprehensive UCE bait sets can be designed with wide taxonomic coverage, as has been demonstrated for diverse groups such as amniotes (Faircloth et al., 2012), fish (Faircloth et al., 2013; Alfaro et al., 2018), and several hyper-diverse invertebrate groups (Starrett et al., 2016; Baca et al., 2017; Branstetter et al., 2017b; Quattrini et al., 2018). While the number of orthologous UCE loci drops as a function of phylogenetic distance between taxa, hundreds to thousands of UCEs are still available covering groups separated by hundreds of millions of years of evolution (Faircloth et al., 2012). Eventually, UCE bait sets with universal coverage of metazoan groups may be designed to consistently recover the full range of species represented in environmental and bulk samples.

Phylogenetic Signal

Barcodes must contain a sufficient phylogenetic signal to assign taxonomy to recovered sequences. Ultimately this requires the availability of comprehensive open-access databases of taxonomically verified sequences for comparison. Such databases already exist for COI, and this is arguably the greatest strength of this marker as a barcode choice (Andújar et al., 2018). Because UCEs are flanked by regions of increasing variability, they are useful for resolving both deep (Crawford et al., 2012) and shallow relationships (Smith et al., 2014). This suggests that UCEs would be effective for both discriminating species and assigning taxonomy in metabarcoding studies. However, because taxonomically comprehensive databases of UCE sequences from a wide range of species do not exist at present, assigning taxonomy at lower levels (e.g., family and below) to UCEs recovered during metabarcoding represents a challenge. Combining both marker types would allow users to generate consensus diversity estimates and pinpoint possible sources of error, while leveraging the taxonomic coverage of COI reference databases.

Robustness and Recoverability

For barcodes to be effective they must be reliably amplified, containing both highly conserved and variable regions (Taberlet et al., 2007). Sequence conservation is especially important for metabarcoding, which uses DNA extracted from environmental and bulk samples containing a wide taxonomic range of species of an unknown composition. However, the conserved region of COI, necessary for effective primer binding, is not sufficiently conserved to work equally well across or within all animal groups (Deagle et al., 2014). Because of this, the amplification step may introduce biases in both copy number and taxonomic representation. The core regions of UCEs are, as the name implies, highly conserved, and are reliably recovered using targeted enrichment (Gnirke et al., 2009; Faircloth et al., 2012). The targeted enrichment approach does not require amplification with universal primers, reducing the opportunity for bias. Moreover, UCEs are identified by comparing the genomes of highly divergent taxa, but importantly the bait sets designed to target them have been demonstrated to work well on a multitude of intermediate taxa, an underpinning of the approach (Faircloth et al., 2012).

Environmental and Degraded DNA

Environmental DNA is a major component of metabarcoding studies (Deiner et al., 2017), which obtain DNA from diverse sources like seawater (Boussarie et al., 2018) and fecal samples (Sullins et al., 2018), and which may be from either modern or ancient environments (Thomsen and Willerslev, 2015). Environmental DNA is generally degraded, and the proportion of amplifiable fragments drops off with increasing amplicon size (Deagle et al., 2006). Longer barcodes will be difficult to amplify, and it has been recommended that markers used for amplifying degraded DNA be no longer than 150 bp (Valentini et al., 2009). COI is several times this length, creating a need for developing shorter barcodes for use with degraded DNA (Hajibabaei et al., 2006). By contrast, UCE loci have a wide range of lengths (Bejerano et al., 2004), bait sets targeting shorter UCE loci can be specified, and at 120 bp, the baits used to enrich UCE loci fit within the commended maximum length. Furthermore, UCEs have been demonstrated to work successfully with old and degraded DNA such as that obtained from museum specimens stored in suboptimal conditions (Blaimer et al., 2016; McCormack et al., 2016).

DISCUSSION

COI does not make a perfect metabarcode, and its widespread use reflects the lack of apposite substitutes rather than its suitability as a marker. As noted by Deagle et al. (2014), even the best metabarcoding studies using COI have pointed out its limitations, underscoring the importance of developing alternative markers. At the heart of this lies the fact that COI cannot be reliably and consistently amplified from all animal groups or from environmental samples containing degraded DNA, both of which are crucial points for metabarcoding. Utilizing multiple barcoding regions and markers better suited for use with degraded DNA will likely become a matter of routine, as multiplexing markers can improve the accuracy and reliability of species recovery (De Barba et al., 2014; Zhang et al., 2018). Taberlet et al. (2012) also discuss potential methods for overcoming amplification bias. Direct sequencing methods, similar to genome skimming (Dodsworth, 2015) or metagenomics (Quince et al., 2017), are one possible solution. Direct sequencing methods produce large amounts of data without the bias introduced during amplification. However, most of the data is likely to be taxonomically unassignable or prokaryotic in origin, and direct sequencing has been shown to significantly underperform compared to metabarcoding in regard to evaluating eukaryotic diversity (Stat et al., 2017). Another solution identified by Taberlet et al. (2012) involves sequence capture, using hundreds of baits to target different taxonomic groups. Targeted enrichment using UCEs fits this

proposed solution neatly, with the significant advantage of established bait sets and open-access workflows, minimizing the cost and effort required to adapt these methods to metabarcoding studies.

It is important to note that the limitations of COI described here apply to metabarcoding studies that utilize DNA extracted directly from bulk community and environmental samples containing an unknown species composition (*sensu* Stat et al., 2017; Ritter et al., 2019). In regard to standard barcoding studies, which benefit from a narrower approach focusing on single specimens and usually complemented by alternative methods like morphological examination, COI has been used to a high degree of success (Janzen et al., 2009). Even with its limitations, COI has been used successfully in a variety of metabarcoding applications and retains several advantageous qualities. Chief among these is the public availability of millions of taxonomically verified COI sequences from hundreds of thousands of species (Ratnasingham and Hebert, 2007, 2013), which alone is sufficient to justify the continued usefulness of COI in metabarcoding studies (Andújar et al., 2018). The way forward lies not in replacing COI as a metabarcode, but rather in developing suites of markers to use in parallel, which can then complement one another's strengths and shortcomings.

UCEs may offer a way to strengthen results obtained using COI by redressing its limitations, such as the amplification step, as well as by providing replication. Furthermore, their utility may stretch to fungi and plants (Siepel et al., 2005; Reneker et al., 2012), groups beyond the reach of COI (Kress et al., 2015). Already available are bait sets covering a number of broad taxonomic groups, and open-source workflows for identifying UCEs in other taxa (Faircloth, 2017). Used together, UCEs may be able to provide a way to generate comprehensive bait sets that can reliably recover species from across the tree of life, for parallel use with standard barcodes like COI that can leverage the verified taxonomic coverage available in standard barcode databases. Obtaining both UCE and COI data simultaneously is efficient and cost-effective, as mitochondrial DNA is captured concomitantly during targeted enrichment of UCEs as "bycatch" (Raposo do Amaral et al., 2015), as demonstrated in several phylogenomic studies utilizing UCEs (Pierce et al., 2017; Zarza et al., 2018; Branstetter and Longino, 2019). The relative ease with which these markers have been used together in phylogenomics suggests a similar feasibility for metabarcoding. Despite the many possible advantages, thorough testing will be required to determine the feasibility and advantage of utilizing UCEs in a metabarcoding context.

Several distinct challenges would need to be overcome to obtain the full added benefit of implementing UCEs in metabarcoding. Chief among these is the lack of a comprehensive UCE reference database, posing a challenge to taxonomic assignment. Creating such a database will be time and resource intensive, and in the meantime species identification of UCE loci will be limited based on available GenBank data. This underscores the importance of building on the COI framework and utilizing its extensive database, and using UCEs to provide replication, fill in gaps where COI does not work well, and identify potential sources of error. Given the multi-locus nature of UCEs, combining UCE data from single organisms in a mixed sample would represent another major challenge. Without linking data from intra-individual UCE loci, each locus would act independently as a barcode sequence, limiting the added value of UCEs to providing support and validation for conventional metabarcoding methods. Though still valuable, linking the combined multi-locus data would be able to provide much stronger phylogenetic signal, potentially allowing elucidation of evolutionary relationships or population level analysis from mixed samples, greatly enhancing the added benefit of UCEs to metabarcoding analyses. The linkage problem is further compounded for degraded DNA, given that shorter DNA fragments are likely to contain less phylogenetic signal, further limiting the usefulness of unlinked loci. Potential solutions to the linkage problem include progressive match-based and/or distance-based binning, and machine or deep learning algorithms based on input data from taxon-specific studies. Whether these or other possible solutions will work is unclear, and solving this analytical problem will be an important but challenging goal.

CONCLUDING REMARKS

As we move deeper into the Anthropocene, global biodiversity faces an unparalleled and worsening crisis. Scientists tasked with cataloging global diversity face two monumental challenges: the profound biological and ecological diversity of life, and the precipitous rate at which humanity is destroying and altering the environment. In the midst of anthropogenically induced mass extinction (Ceballos et al., 2015), environmental upheaval (Newbold et al., 2015; Seebens et al., 2018), and climate change (Bellard et al., 2012), the number of species on Earth remains unknown (Caley et al., 2014), the majority of species remain undiscovered (Mora et al., 2011), and their evolutionary histories and ecological interactions uncharacterized. In the face of these challenges, scientists must adopt and improve the most effective methods available to discover, catalog, and monitor biodiversity.

Of available methods for surveying biodiversity, metabarcoding provides the greatest balance of taxonomic coverage and resolution, sampling depth, accuracy, efficiency, and ease of use (Ji et al., 2013; Bush et al., 2019). However, there remain significant hurdles to overcome to improve its accuracy and reliability. Fundamental to this is the need to identify alternative barcodes that can fill in the gaps of standard barcodes (Deagle et al., 2014). Ultraconserved elements are one possible solution. Their many strengths, particularly in areas where standard barcodes have demonstrated weaknesses, make them promising candidates that merit consideration.

AUTHOR CONTRIBUTIONS

MP developed and wrote the manuscript.

REFERENCES

Alfaro, M. E., Faircloth, B. C., Harrington, R. C., Sorenson, L., Friedman, M., Thacker, C. E., et al. (2018). Explosive diversification of marine fishes at the Cretaceous–Palaeogene boundary. *Nat. Ecol. Evol.* 2, 688–696. doi: 10.1038/s41559-018-0494-6

Andújar, C., Arribas, P., Yu, D. W., Vogler, A. P., and Emerson, B. C. (2018). Why the COI barcode should be the community DNA metabarcode for the metazoa. *Mol. Ecol.* 27, 3968–3975. doi: 10.1111/mec.14844

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

Bejerano, G., Pheasant, M., Makunin, I., Stephen, S., Kent, W. J., Mattick, J. S., et al. (2004). Ultraconserved elements in the human genome. *Science* 304, 1321–1325. doi: 10.1126/science.1098119

Bellard, C., Bertelsmeier, C., Leadley, P., Thuiller, W., and Courchamp, F. (2012). Impacts of climate change on the future of biodiversity: biodiversity and climate change. *Ecol. Lett.* 15, 365–377. doi: 10.1111/j.1461-0248.2011.01736.x

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

Boussarie, G., Bakker, J., Wangensteen, O. S., Mariani, S., Bonnin, L., Juhel, J.-B., et al. (2018). Environmental DNA illuminates the dark diversity of sharks. *Sci. Adv.* 4:eaap9661. doi: 10.1126/sciadv.aap9661

Branstetter, M. G., Danforth, B. N., Pitts, J. P., Faircloth, B. C., Ward, P. S., Buffington, M. L., et al. (2017a). Phylogenomic insights into the evolution of stinging wasps and the origins of ants and bees. *Curr. Biol.* 27, 1019–1025. doi: 10.1016/j.cub.2017.03.027

Branstetter, M. G., and Longino, J. T. (2019). Ultra-conserved element phylogenomics of new world ponera (hymenoptera: formicidae) illuminates the origin and phylogeographic history of the endemic exotic ant *Ponera exotica*. *Insect Syst. Divers.* 3, 1–13. doi: 10.1093/isd/ixz001

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

Bush, A., Compson, Z. G., Monk, W. A., Porter, T. M., Steeves, R., Emilson, E., et al. (2019). Studying ecosystems with DNA metabarcoding: lessons from biomonitoring of aquatic macroinvertebrates. *Front. Ecol. Evol.* 7:434. doi: 10.3389/fevo.2019.00434

Caley, M. J., Fisher, R., and Mengersen, K. (2014). Global species richness estimates have not converged. *Trends Ecol. Evol.* 29, 187–188. doi: 10.1016/j.tree.2014.02.002

Carvalho, S., Aylagas, E., Villalobos, R., Kattan, Y., Berumen, M., and Pearman, J. K. (2019). Beyond the visual: using metabarcoding to characterize the hidden reef cryptobiome. *Proc. R. Soc. B Biol. Sci.* 286:20182697. doi: 10.1098/rspb.2018.2697

Ceballos, G., Ehrlich, P. R., Barnosky, A. D., García, A., Pringle, R. M., and Palmer, T. M. (2015). Accelerated modern human–induced species losses: entering the sixth mass extinction. *Sci. Adv.* 1:e1400253. doi: 10.1126/sciadv.1400253

Crawford, N. G., Faircloth, B. C., McCormack, J. E., Brumfield, R. T., Winker, K., and Glenn, T. C. (2012). More than 1000 ultraconserved elements provide evidence that turtles are the sister group of archosaurs. *Biol. Lett.* 8, 783–786. doi: 10.1098/rsbl.2012.0331

De Barba, M., Miquel, C., Boyer, F., Mercier, C., Rioux, D., Coissac, E., et al. (2014). DNA metabarcoding multiplexing and validation of data accuracy for diet assessment: application to omnivorous diet. *Mol. Ecol. Resour.* 14, 306–323. doi: 10.1111/1755-0998.12188

Deagle, B. E., Eveson, J. P., and Jarman, S. N. (2006). Quantification of damage in DNA recovered from highly degraded samples – a case study on DNA in faeces. *Front. in Zool.* 3:11. doi: 10.1186/1742-9994-3-11

Deagle, B. E., Jarman, S. N., Coissac, E., Pompanon, F., and Taberlet, P. (2014). DNA metabarcoding and the cytochrome c oxidase subunit I marker: not a perfect match. *Biol. Lett.* 10:20140562. doi: 10.1098/rsbl.2014.0562

Deiner, K., Bik, H. M., Mächler, E., Seymour, M., Lacoursière-Roussel, A., Altermatt, F., et al. (2017). Environmental DNA metabarcoding: transforming how we survey animal and plant communities. *Mol. Ecol.* 26, 5872–5895. doi: 10.1111/mec.14350

Derti, A., Roth, F. P., Church, G. M., and Wu, C. (2006). Mammalian ultraconserved elements are strongly depleted among segmental duplications and copy number variants. *Nat. Genet.* 38, 1216–1220. doi: 10.1038/ng1888

Dodsworth, S. (2015). Genome skimming for next-generation biodiversity analysis. *Trends Plant Sci.* 20, 525–527. doi: 10.1016/j.tplants.2015.06.012

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

Faircloth, B. C., Sorenson, L., Santini, F., and Alfaro, M. E. (2013). A phylogenomic perspective on the radiation of ray-finned fishes based upon targeted sequencing of ultraconserved elements (UCEs). *PLoS ONE* 8:e65923. doi: 10.1371/journal.pone.0065923

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

Hajibabaei, M., Smith, M. A., Janzen, D. H., Rodriguez, J. J., Whitfield, J. B., and Hebert, P. D. N. (2006). A minimalist barcode can identify a specimen whose DNA is degraded: BARCODING. *Mol. Ecol. Notes* 6, 959–964. doi: 10.1111/j.1471-8286.2006.01470.x

Hebert, P. D., Ratnasingham, S., and de Waard, J. R. (2003). Barcoding animal life: cytochrome c oxidase subunit 1 divergences among closely related species. *Proc. R. Soc. Lond. Ser. B Biol. Sci.* 270(Suppl 1), S96–S99. doi: 10.1098/rsbl.2003.0025

Janzen, D. H., Hallwachs, W., Blandin, P., Burns, J. M., Cadiou, J.-M., Chacon, I., et al. (2009). Integration of DNA barcoding into an ongoing inventory of complex tropical biodiversity. *Mol. Ecol. Resour.* 9, 1–26. doi: 10.1111/j.1755-0998.2009.02628.x

Ji, Y., Ashton, L., Pedley, S. M., Edwards, D. P., Tang, Y., Nakamura, A., et al. (2013). Reliable, verifiable and efficient monitoring of biodiversity via metabarcoding. *Ecol. Lett.* 16, 1245–1257. doi: 10.1111/ele.12162

Kitson, J. J. N., Hahn, C., Sands, R. J., Straw, N. A., Evans, D. M., and Lunt, D. H. (2018). Detecting host–parasitoid interactions in an invasive Lepidopteran using nested tagging DNA metabarcoding. *Mol. Ecol.* 28, 471–483. doi: 10.1111/mec.14518

Kress, W. J., and Erickson, D. L. (2008). DNA barcodes: genes, genomics, and bioinformatics. *Proc. Natl. Acad. Sci. U.S.A.* 105, 2761–2762. doi: 10.1073/pnas.0800476105

Kress, W. J., García-Robledo, C., Uriarte, M., and Erickson, D. L. (2015). DNA barcodes for ecology, evolution, and conservation. *Trends Ecol. Evol.* 30, 25–35. doi: 10.1016/j.tree.2014.10.008

McCormack, J. E., Harvey, M. G., Faircloth, B. C., Crawford, N. G., Glenn, T. C., and Brumfield, R. T. (2013). A phylogeny of birds based on over 1,500 loci collected by target enrichment and high-throughput sequencing. *PLoS ONE* 8:e54848. doi: 10.1371/journal.pone.0054848

McCormack, J. E., Tsai, W. L., 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

Mora, C., Tittensor, D. P., Adl, S., Simpson, A. G., and Worm, B. (2011). How many species are there on earth and in the ocean? *PLoS Biol.* 9:e1001127. doi: 10.1371/journal.pbio.1001127

Moritz, C., and Cicero, C. (2004). DNA barcoding: promise and pitfalls. *PLoS Biol.* 2:e354. doi: 10.1371/journal.pbio.0020354

Newbold, T., Hudson, L. N., Hill, S. L. L., Contu, S., Lysenko, I., Senior, R. A., et al. (2015). Global effects of land use on local terrestrial biodiversity. *Nature* 520, 45–50. doi: 10.1038/nature14324

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

Piñol, J., Senar, M. A., and Symondson, W. O. C. (2018). The choice of universal primers and the characteristics of the species mixture determine when DNA metabarcoding can be quantitative. *Mol. Ecol.* 28, 407–419. doi: 10.1111/mec.14776

Porter, T. M., and Hajibabaei, M. (2018). Scaling up: a guide to high-throughput genomic approaches for biodiversity analysis. *Mol. Ecol.* 27, 313–338. doi: 10.1111/mec.14478

Quattrini, A. M., Faircloth, B. C., Dueñas, L. F., Bridge, T. C. L., Brugler, M. R., Calixto-Botía, I. F., et al. (2018). Universal target-enrichment baits for anthozoan (Cnidaria) phylogenomics: new approaches to long-standing problems. *Mol. Ecol. Resour.* 18, 281–295. doi: 10.1111/1755-0998.12736

Quince, C., Walker, A. W., Simpson, J. T., Loman, N. J., and Segata, N. (2017). Shotgun metagenomics, from sampling to analysis. *Nat. Biotechnol.* 35:833. doi: 10.1038/nbt.3935

Raposo do Amaral, F., Neves, L. G., Resende, M. F., Mobili, F., Miyaki, C. Y., Pellegrino, K. C., et al. (2015). ultraconserved elements sequencing as a low-cost source of complete mitochondrial genomes and microsatellite markers in non-model amniotes. *PLoS ONE* 10:e0138446. doi: 10.1371/journal.pone.0138446

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. (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

Reneker, J., Lyons, E., Conant, G. C., Pires, J. C., Freeling, M., Shyu, C.-R., et al. (2012). Long identical multispecies elements in plant and animal genomes. *Proc. Natl. Acad. Sci. U.S.A.* 109, E1183–E1191. doi: 10.1073/pnas.1121356109

Ritter, C. D., Häggqvist, S., Karlsson, D., Sääksjärvi, I. E., Muasya, A. M., Nilsson, R. H., et al. (2019). Biodiversity assessments in the 21st century: the potential of insect traps to complement environmental samples for estimating eukaryotic and prokaryotic diversity using high-throughput DNA metabarcoding. *Genome* 62, 147–159. doi: 10.1139/gen-2018-0096

Rubinoff, D., Cameron, S., and Will, K. (2006). A genomic perspective on the shortcomings of mitochondrial DNA for "Barcoding" identification. *J. Hered.* 97, 581–594. doi: 10.1093/jhered/esl036

Seebens, H., Blackburn, T. M., Dyer, E. E., Genovesi, P., Hulme, P. E., Jeschke, J. M., et al. (2018). Global rise in emerging alien species results from increased accessibility of new source pools. *Proc. Natl. Acad. Sci. U.S.A.* 115, E2264–E2273. doi: 10.1073/pnas.1719429115

Siepel, A., Bejerano, G., Pedersen, J. S., Hinrichs, A. S., Hou, M., Rosenbloom, K., et al. (2005). Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* 15, 1034–1050. doi: 10.1101/gr.3715005

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

Starrett, J., Derkarabetian, S., Hedin, M., Bryson, R. W., McCormack, J. E., and Faircloth, B. C. (2016). High phylogenetic utility of an ultraconserved element probe set designed for Arachnida. *Mol. Ecol. Resour.* 17, 812–823. doi: 10.1111/1755-0998.12621

Stat, M., Huggett, M. J., Bernasconi, R., DiBattista, J. D., Berry, T. E., Newman, S. J., et al. (2017). Ecosystem biomonitoring with eDNA: metabarcoding across the tree of life in a tropical marine environment. *Sci. Rep.* 7:12240. doi: 10.1038/s41598-017-12501-5

Stephen, S., Pheasant, M., Makunin, I. V., and Mattick, J. S. (2008). Large-scale appearance of ultraconserved elements in tetrapod genomes and slowdown of the molecular clock. *Mol. Biol. Evol.* 25, 402–408. doi: 10.1093/molbev/msm268

Sullins, D. S., Haukos, D. A., Craine, J. M., Lautenbach, J. M., Robinson, S. G., Lautenbach, J. D., et al. (2018). Identifying the diet of a declining prairie grouse using DNA metabarcoding. *Auk* 135, 583–608. doi: 10.1642/AUK-17-199.1

Taberlet, P., Coissac, E., Pompanon, F., Brochmann, C., and Willerslev, E. (2012). Towards next-generation biodiversity assessment using DNA metabarcoding: NEXT-GENERATION DNA METABARCODING. *Mol. Ecol.* 21, 2045–2050. doi: 10.1111/j.1365-294X.2012.05470.x

Taberlet, P., Coissac, E., Pompanon, F., Gielly, L., Miquel, C., Valentini, A., et al. (2007). Power and limitations of the chloroplast trnL (UAA) intron for plant DNA barcoding. *Nucleic Acids Res.* 35:e14. doi: 10.1093/nar/gkl938

Thomsen, P. F., and Willerslev, E. (2015). Environmental DNA – An emerging tool in conservation for monitoring past and present biodiversity. *Biol. Conserv.* 183, 4–18. doi: 10.1016/j.biocon.2014.11.019

Valentini, A., Pompanon, F., and Taberlet, P. (2009). DNA barcoding for ecologists. *Trends Ecol. Evol.* 24, 110–117. doi: 10.1016/j.tree.2008.09.011

Will, K. W., Mishler, B. D., and Wheeler, Q. D. (2005). The perils of DNA barcoding and the need for integrative taxonomy. *Syst. Biol.* 54, 844–851. doi: 10.1080/10635150500354878

Zarza, E., Connors, E. M., Maley, J. M., Tsai, W. L. E., Heimes, P., Kaplan, M., et al. (2018). Combining ultraconserved elements and mtDNA data to uncover lineage diversity in a Mexican highland frog (Sarcohyla; Hylidae). *PeerJ* 6:e6045. doi: 10.7717/peerj.6045

Zhang, G. K., Chain, F. J. J., Abbott, C. L., and Cristescu, M. E. (2018). Metabarcoding using multiplexed markers increases species detection in complex zooplankton communities. *Evol. Appl.* 11, 1901–1914. doi: 10.1111/eva.12694

Zinger, L., Bonin, A., Alsos, I. G., Bálint, M., Bik, H., Boyer, F., et al. (2019). DNA metabarcoding—Need for robust experimental designs to draw sound ecological conclusions. *Mol. Ecol.*, 28, 1857–1862. doi: 10.1111/mec.15060

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

frontiers in Ecology and Evolution

ORIGINAL RESEARCH
published: 19 December 2019
doi: 10.3389/fevo.2019.00495

The Challenge of DNA Barcoding Saproxylic Beetles in Natural History Collections—Exploring the Potential of Parallel Multiplex Sequencing With Illumina MiSeq

Lucas Sire[1], Delphine Gey[2], Régis Debruyne[2], Thierry Noblecourt[3], Fabien Soldati[3], Thomas Barnouin[3], Guilhem Parmain[4], Christophe Bouget[4], Carlos Lopez-Vaamonde[1,5] and Rodolphe Rougerie[6]*

[1] Institut de Recherche sur la Biologie de l'Insecte, UMR 7261, CNRS Université de Tours, Tours, France, [2] Service de Systématique Moléculaire, UMS 2700 2AD, Muséum national d'Histoire naturelle, Paris, France, [3] Office National des Forêts, Laboratoire National d'Entomologie Forestière, Quillan, France, [4] IRSTEA, Nogent-sur-Vernisson, France, [5] INRA, UR0633 Zoologie Forestière, Orléans, France, [6] Institut de Systématique, Evolution, Biodiversité (ISYEB), Muséum national d'Histoire naturelle, CNRS, Sorbonne Université, EPHE, Université des Antilles, Paris, France

OPEN ACCESS

Edited by:
Rodney L. Honeycutt,
Pepperdine University, United States

Reviewed by:
Anthony I. Cognato,
Michigan State University,
United States
Trevor Bringloe,
The University of Melbourne, Australia
Michael J. Raupach,
Bavarian State Collection of
Zoology, Germany

***Correspondence:**
Lucas Sire
lucas.sire@univ-tours.fr

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

Received: 01 July 2019
Accepted: 03 December 2019
Published: 19 December 2019

Citation:
Sire L, Gey D, Debruyne R,
Noblecourt T, Soldati F, Barnouin T,
Parmain G, Bouget C,
Lopez-Vaamonde C and Rougerie R
(2019) The Challenge of DNA
Barcoding Saproxylic Beetles in
Natural History Collections—Exploring
the Potential of Parallel Multiplex
Sequencing With Illumina MiSeq.
Front. Ecol. Evol. 7:495.
doi: 10.3389/fevo.2019.00495

Saproxylic beetles are important bioindicators of forest health but their enormous diversity makes their identification challenging. As an example, the French fauna of saproxylic beetles alone contains 2,663 species in 72 families. Recently, DNA barcoding was proposed as a promising tool for the identification and monitoring of saproxylic beetle species. However, the rate of DNA barcode recovery from specimens of natural history collections using standard Sanger sequencing protocols remains low and challenges the construction of reference libraries. In this study, we test the potential of high-throughput sequencing (HTS) technology to reduce this shortfall by increasing sequencing success rate and lowering processing cost per specimen. Using a dual-indexing strategy for library construction and sequencing on the Illumina MiSeq platform, we successfully sequenced the DNA barcodes of 286 dry-pinned saproxylic beetles out of 521 specimens aged from 1 to 17 years and sampled in natural history collections. Age at sequencing did not affect sequence recovery and the success rate (54.9%) of our approach is comparable to that obtained using Sanger sequencing technology in another study targeting beetle specimens from natural history collections, but the cost per specimen is significantly reduced when using HTS. Finally, we shortly discuss how the newly produced DNA barcodes contribute to the existing library and we highlight a few interesting cases in which the new sequences question current species boundaries.

Keywords: COI, coleoptera, high-throughput sequencing, degraded DNA, Sanger sequencing

INTRODUCTION

The assessment and monitoring of biodiversity are fundamental tasks for conservation management and ecosystem preservation. Both suffer heavily from their strong reliance on a too scarce taxonomic knowledge (Giangrande, 2003) and on the general absence of comprehensive, inexpensive, and user-friendly tools for species identification. This is especially critical for insects

(Green, 1998; Stork et al., 2015), which are massively impacted by environmental changes, with cascade effects on the functioning of ecosystems (Hallmann et al., 2017).

One recent methodological development that can reduce this shortfall in insect species identification is the use of a short and standardized DNA fragment, termed "DNA barcode" (Hebert et al., 2003a). This approach relies on the use of a 658 base pair (bp) fragment of the mitochondrial gene cytochrome *c* oxidase subunit 1 (COI) and on an online centralized database and workbench, the Barcode of Life Datasystems (BOLD) (Ratnasingham and Hebert, 2007, www.boldsystems.org) as a reference library ensuring the link between reference specimens identified by experts and these DNA barcode sequences. DNA barcoding is now a widely adopted tool for species delimitation and identification. As of today (10/09/2019) BOLD holds as many as 7480K DNA barcodes for 210K named species, and 645K BINs (Barcode Index Numbers, an automatic classification system of DNA barcode sequences that can be used as a proxy to species when records are unnamed; see Ratnasingham and Hebert, 2013). Such rapid—and ongoing—development has created a completely new and efficient access to taxonomic expertise for whoever can retrieve this short DNA snippet from specimens or their parts. Furthermore, it is now the ground for developing high-throughput approaches such as environmental sequencing or DNA metabarcoding (Deiner et al., 2017) that use new sequencing technologies to analyse tens to thousands of individuals and species simultaneously, thus opening new avenues in biodiversity assessments and monitoring (Yu et al., 2012; Ji et al., 2013).

Saproxylic beetles (i.e., "any beetle that depends, during some parts of its life cycle, upon wounded or decaying woody material from living, weakened, or dead trees") (Stokland et al., 2012) are of major importance in forest ecosystem functioning. Indeed, they are prime actors in the early process of wood decay (Stokland et al., 2012). Furthermore, their response to changes in environmental variables make them suitable bio-indicators for shaping conservation and economic managements as well as monitoring health of forest environments (Janssen et al., 2017). In France, saproxylic beetles are highly diversified but well-studied and 2,663 species from 72 families are currently recorded (Bouget et al., 2019). Yet, their identification requires a high level of expertise, which is scarce or possibly missing for many families. DNA barcode reference libraries using traditional Sanger sequencing have been developed for European Coleoptera (Pentinsaari et al., 2014; Hendrich et al., 2015; Rougerie et al., 2015), covering more than 2,100 European saproxylic beetle species so far. These libraries revealed the general consistency between morphologically characterized species and DNA barcode clusters, as in other insect orders, and thus supported the relevance of this genetic marker for delimitating and identifying beetle species.

Access to natural history collection is critical for the assembly of DNA barcode libraries, because it allows processing specimens of species that are very difficult to re-collect (e.g., rare or extinct species or populations), and access to specimens that have been authoritatively identified and/or type material that can facilitate stronger links between barcodes and species names (Hausmann et al., 2016). However, for many collection samples nothing is known about the way insects were collected, killed and preserved, which can in turn have a significant negative effect on the generation of a DNA barcode sequence (Prosser et al., 2015). Indeed, success rate of DNA barcoding of saproxylic beetles from natural history collections is reportedly low (61%, see Rougerie et al., 2015) in spite of the use of failure tracking technique, targeting shorter DNA fragments to improve the rate of PCR amplification success. This relatively low success of sequence recovery represents a major hurdle to the use of DNA barcoding for identifying beetles and challenges the very construction of reference libraries. Nevertheless, the majority of failed amplified samples were either old collection individuals or belonging to specific families (Rougerie et al., 2015; see also Pentinsaari et al., 2014). These observations are in accordance with studies showing that some taxa may be hard to sequence due to primer mismatches (Piñol et al., 2015) or when sampled in natural history collections (Van Houdt et al., 2010). High-throughput sequencing (HTS) has been shown to be an alternative to improve the DNA barcoding success of such taxa (Shokralla et al., 2014). HTS technologies have emerged and considerably developed over the past two decades and their sequencing power and quality has increased inversely to sequencing cost (Liu et al., 2012; van Dijk et al., 2014). Yet, one major issue when targeting the full-length DNA barcode is the relatively short reads produced by most of the HTS technologies. This usually requires the need to amplify multiple overlapping fragments and to use dual-indexing approaches in the lab to multiplex different samples or amplicons per samples, as well as bioinformatic expertise to separate samples and assemble the produced reads into a single consensus (Fadrosh et al., 2014; Bourlat et al., 2016; Leray et al., 2016). On the other hand, because sequencing of degraded DNA requires the amplification of shorter amplicons, it seems appropriate to adapt and use HTS to process collection specimens toward sequencing of DNA barcodes.

Nevertheless, while HTS technologies are widely used in environmental genomic approaches like metabarcoding with complex samples (Oliverio et al., 2018; Barsoum et al., 2019; Thomsen and Sigsgaard, 2019), their implementation in conventional DNA barcoding of single individuals still lags behind, despite evidence of their potential at multiple levels (Shokralla et al., 2014, 2015; Cruaud et al., 2017; Fagan-jeffries et al., 2018; Wang et al., 2018).

Overall, the reference library for French saproxylic beetle fauna is still largely incomplete, with 1,535 species barcoded out of 2,663 species (58%) (Rougerie et al., 2015). The application of HTS could potentially accelerate the pace of assembly of this library at reduced cost. Here we use a slightly modified version of the approach proposed by Shokralla et al. (2015) targeting two short amplicons on Illumina MiSeq sequencing to generate DNA barcodes for individual collection specimens of saproxylic beetles. Our main aim was to extend the taxonomic and geographical coverage of the French saproxylic beetle DNA barcode library, while testing the benefits of using HTS technology when processing specimens whose DNA is expected to be degraded.

MATERIALS AND METHODS

Specimen Sampling in Collections

We sampled dry-pinned specimens deposited at the national collection of forest insects held at the National Forest Office (ONF) in Quillan (Aude, France). Most samples lacked information on collecting methods and reagents used for preservation. We therefore selected specimens based only on their collection date, favoring those samples collected as recently as possible. Specimens belonged to species known to occur in France (http://saprox.mnhn.fr/) but not yet barcoded (Pentinsaari et al., 2014; Hendrich et al., 2015; Rougerie et al., 2015). We focused our sampling on species from the French Pyrenees where we are carrying out a metabarcoding analysis of forest biodiversity (CLIMTREE project). Tissue samples were placed in 96-well plates. For each individual, a midleg was sampled, except for *Dorcatoma* and *Stagetus* spp. for which an abdomen was taken after genitalia removal, due to the lack of significant diagnostic characters for taxonomic identification and the higher amount of tissue it provides. Sampling was done using sterilized forceps. Collecting data were compiled into a standard Darwin Code spreadsheet and vouchers were photographed using either a 14MP 1/2.3″ APTINA CMOS Sensor U3CMOS mounted on a stereomicroscope, or a Nikon D7200 with an AF-S DX NIKKOR 18-300MM F/3.5-5.6G ED VR Lens for the biggest individuals.

DNA Extraction and Illumina Library Preparation

DNA extraction of 521 individuals belonging to 343 species and 42 families sampled in six 96-well plates was carried out at the Service de Systématique Moléculaire (SSM) at the MNHN in Paris, using Macherey-Nagel NucleoSpin® 96 tissue kit following manufacturers protocol using either a semi-automated procedure with an Eppendorf Liquid Handling Workstation epMotion® 7075 VAC, or a manual approach through successive centrifugations.

To accommodate for the 658 bp length of the targeted DNA fragment and the limit in read length when using Illumina sequencing technology, we used internal primers to amplify two shorter amplicons here named B_R and C_F, of 325 and 418 bp in length, respectively; they encompass together the entire DNA barcode region with an 85 bp overlap (**Figure 1**). We carried out a dual indexing method similar to the one used in Shokralla et al. (2015) to permit de-multiplexing and assembly of the reads produced (**Figure 1**). Thus, 20 primer tags of 5 nucleotides were re-designed to remain unique after two potential nucleotide degenerations, containing all four nucleotides without more than two repetitions, and avoiding more than 3 identical successive nucleotides once added to the 5′ end of our primers. These primer tags were split in 2 sets of 10 each: AGTCT, ATTGC, ACGTC, ATGCG, AGATC, ATCTG, CATTG, CTAGG, CGGAT, CGTGA for forward primers and CTGTA, CGATT, TTGAC, TGGCA, TACAG, TGACG, TTCGA, TAGCC, TCGGA, TCTAG for reverse primers, respectively. Tagged-primers were synthetized in NGS grade with HPLC purification by Eurofins Genomics, 85560 Ebersberg, Germany.

The internal primers Ill_C_R (5′–GGIGGRTAIACIGTTCAICC–3′), and Ill_B_F (5′–CCIGAYATRGCITTYCCICG–3′) (Shokralla et al., 2015) were used in combination with Folmer et al. (1994) primers LCO1490 and HCO2198, respectively, to amplify the fragments B_R and C_F mentioned above independently. The use of Inosin nucleotide (*I*) allows a match with all four nucleotides more efficiently than with a four-fold degeneracy because of the reduction in concentration of each primer combination when using the latter option. Yet, due to *(I)* in our primers, we did not used a proof-reading polymerase to ensure avoiding synthesis bias (Knittel and Picard, 1993). PCR reactions were conducted separately in two plates (one for each amplicon) in 25 μL with 2.5 μL of 10X CoralLoad PCR buffer, 1 μL of 50X $MgCl_2$ (50 mM), 0.5 μL of dNTPs (6.6 mM), 1 μL of each primer (10 mM), 0.5 μL of DNA Taq Polymerase (5 U/μL) from Qiagen, 2 μL of DNA template and the final 17.5 μL in extra pure water. PCR started with initial denaturation at 95°C for 5 min, 35 cycles of 94°C for 40 s, 51°C for 1 min, 72°C for 30 s, and final elongation at 72°C for 5 min.

For each sample plate, the two independent plates of PCR products obtained, corresponding, respectively, to amplicons B_R and F_C, were pooled in 5 mL tubes before being processed through a second indexing step based on the protocol of Meyer and Kircher (2010). This started by a purification step of 400 μL of each pool of amplicons using NucleoMag 0.85X, then eluted in 50 μL TET buffer (0.1X), and followed by DNA quantification using Qubit® Broad Range. In contrast to Meyer and Kircher (2010), we performed a blunt-end repair using NEBNext End Repair Module before proceeding with a ligation step to attach Illumina adapters to our libraries and thus avoid an additional PCR step that may increase replication errors, especially when using a non-proofreading high-fidelity polymerase enzyme (Meyer et al., 2012; Leray et al., 2016; Chimeno et al., 2018). Approximately 500 ng of DNA were used with 5 μL of NEBNext Repair Reaction Buffer (10X) and 2.5 μL of NEBNext End Repair Enzyme Mix. Additional extra pure water was added to reach a 50 μL reaction volume, and the mix was incubated at 20°C for 30 min. A second purification step was carried out with NucleoMag 1X and an elution volume of 20 μL of TET buffer (0.1X). Adapter ligation was therefore performed in 40 μL by adding 10 μL extra pure water, 4 μL T' DNA ligase buffer (10X), 4 μL PEG-4000 (50%), 1 μL adapter mix (100 μM each), and T4 DNA ligase (5 U/μL) to the eluate, which was then incubated at 22°C for 30 min. A third purification with NucleoMag 1X was then performed in 20 μL of EBT buffer. To assess the success of the library preparation, we performed quantification using Qubit® High-Sensitivity kit and controlled products using migration on agarose gel of positive controls. The final PCR indexing enrichment was undertaken after different PCR trials to define the best number of cycles for each sample and starting DNA quantity. This final step was done in a 25 μL volume reaction, comprising 0.5 μL Qiagen Taq (5 U/μL), 2.5 μL of buffer Qiagen 10X, 0.2 μL of dNTPs (25 mM), 0.5 μL of IS4 primer (10 μM) and 50 ng of DNA template as well as 0.5 μL

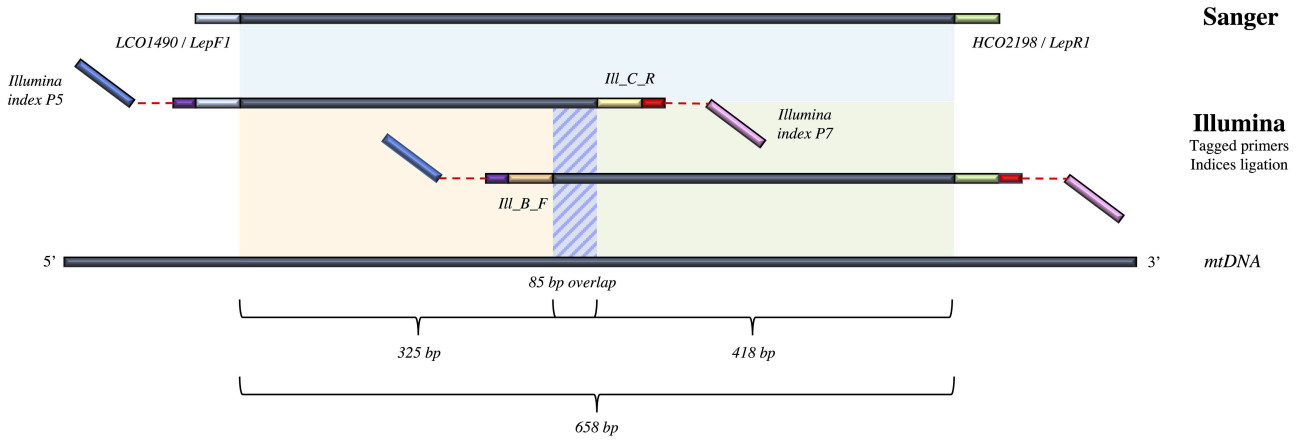


FIGURE 1 | Details of the different COI amplicons targeted in the present study and their associated lengths and overlaps, as well as the primers and sequencing technologies used for each fragment. For the Illumina MiSeq approach, the dual-tagging process and ligation of the Illumina indices are represented.

of indexing primer (10 μM) respective to each sample. PCR cycle was as follow: 94°C for 3 min, 7 cycles of 94°C for 30 s, 60°C for 30 s, and 72°C for 40 s, and final elongation at 72°C for 10 min. Final purification using NucleoMag 0.85X and eluted in 25 μL of EBT buffer was followed by quantification on Qubit® with High-Sensitivity well plate kit.

The six sample plates analyzed for the present study were processed along with 35 other plates from other projects and while our first indexing procedure (using dual tagged-primers) aims at demultiplexing reads per sample within each plate, the second step (by Illumina indices ligation) allows for demultiplexing reads by plate (Bourlat et al., 2016). The concentrations of the libraries corresponding to each plate were homogenized before pooling to obtain a fair balance of sequencing reads between the plates processed and according to their contents. Altogether, the six plates analyzed represented 5.6% in concentration of our pooled library, which was sequenced using a 600 cycles v3 kit (2 × 300 bp, paired-end sequencing) on an Illumina MiSeq at the CIRAD-AGAP sequencing platform in Montpellier, France.

Sanger Sequencing

We tried to amplify all 521 samples targeting the full-length DNA barcode for Sanger sequencing to compare with sequence quality of Illumina MiSeq reads. PCR amplifications were done in 20 μL with 2 μL of 10X CoralLoad PCR buffer, 2 μL of dNTP (6.6 mM), 0.6 μL of each primer (10 mM), 0.2 μL of DNA Taq Polymerase (5 U/μL) from Qiagen, 3 μL of DNA template and 12.2 μL of extra pure water. A primer cocktail named C_LepFol (Hernández-Triana et al., 2014) containing Folmer primers (Folmer et al., 1994) LCO1490 (5′–GGTCAACAAATCATAAAGATATTGG–3′)/HCO2198 (5′–TAAACTTCAGGGTGACCAAAAAATCA–3′) and primers LepF1 (5′–ATTCAACCAATCATAAAGATATTGG–3′)/LepR1 (5′–TAAACTTCTGGATGTCCAAAAAATCA–3′) (Hebert et al., 2004) was used to target and amplify a 658 bp part of the mitochondrial gene cytochrome oxidase subunit 1 (COI) (**Figure 1**). PCR conditions were 94°C during 5 min, followed by 35 cycles of 94°C during 30 s, 54°C for 40 s, and 72°C for 1 min, with a final 10 min extension at 72°C. PCR products were deposited on 2% agarose gel and only successfully amplified DNA templates where sent for Sanger sequencing on ABI 3730XL sequencer at Eurofins MWG Operon sequencing facilities (Ebersberg, Germany).

Demultiplexing and Sequence Analyses

Demultiplexing was done using customized workflows in Geneious V11.0.4 (Kearse et al., 2012). Reads were separated by primer tags with a maximum of one mismatch and a minimum of 2 reads per tag. Primers were trimmed and reads were aligned together with MUSCLE 3.8.425 using eight iterations. The two amplicons B_R and C_F were merged together by *De Novo Assembly* with four maximum ambiguities and two base pairs gap sizes over the 85 bp overlapping region, and the consensus was then saved in separate folders mirroring wells of sample plates for further curation of the sequences. To do so, we blasted each consensus against all barcode records on BOLD and NCBI. Prior morphological identification established by experts in the collection was used to control the blast results to species or to genus level, depending on the availability of DNA barcodes for closely related species. In case of multiple plausible consensus, the potential presence of identical sequences was checked in other samples from the same plate with particular focus on adjacent wells to assess for potential widespread cross-contaminations. In these cases, we also excluded potential pseudogenes by searching for STOP codons or indels, and we investigated possible chimeric sequences (from tag-jumping or incorrect amplicon assembly) through independent identification of both B_R and C_F fragments. The identification was also critically revised by experts through reexamination of voucher specimens, considering the different potential molecular identifications and taking into account existing synonymy, biogeography of sister taxa as well as intra- and interspecific genetic distances to

establish the genuine consensus. When discrimination of this genuine sequence was impossible, sequences were discarded.

Sanger electropherograms for both directions and fragments were assembled to form contigs using Geneious V11.0.4 (Kearse et al., 2012), then aligned and visually checked for quality and noise to resolve some of the ambiguities. For each sample, we ensured no pseudogene presence similarly than with HTS sequences, and we checked for potential cross-contamination by blasting sequences on BOLD to test similarity with conspecific and congeneric existing records. Low quality electropherograms (potentially due to low DNA concentration, DNA degradation or contaminants) were discarded.

Sequence analyses across individuals were performed with analytical tools integrated in BOLD's workbench (Ratnasingham and Hebert, 2007) using BOLD aligner and Kimura-2 Parameter (K2P) (Kimura, 1980) correction method to compute genetic distances and Neighbor Joining (NJ) trees (Saitou and Nei, 1987). The complete workflow of the study is pictured in **Figure 2**. To compare DNA barcodes produced with Illumina and with Sanger, we built a NJ tree combining the consensus sequences recovered from both technologies using Geneious V11.0.4 (Kearse et al., 2012) following Tamura-Nei genetic distance model (Tamura and Nei, 1993), with 1,000 non-parametric bootstrap replications (**Supplementary Material 2**).

Specimens were grouped in different categories according to their age at sequencing to test its effect on sequencing success. Ratio of sequencing success (successfully sequenced individuals divided by the total number of individuals sequenced) was plotted against age at sequencing for the following age categories: 1 year ($N = 86$), 2 years ($N = 61$), 3 years ($N = 39$), 4 years ($N = 41$), 5 years ($N = 61$), 6 years ($N = 49$), 7 years ($N = 72$), 8 years ($N = 60$), 9 years ($N = 30$), 10 years, and more ($N = 22$). We used R V3.6.1 (R Core Team, 2017) to run a non-parametric Spearman correlation analysis as outcome variables do not follow a normal distribution (cor.test, method = "spearman").

Genetic Analyses

To analyse the genetic distances between our newly generated sequences, we mined data of matching species or genus from existing French, German, and Finnish reference libraries (Pentinsaari et al., 2014; Hendrich et al., 2015; Rougerie et al., 2015). This joined dataset of 1,920 sequences and 490 species is available at dx.doi.org/10.5883/DS-COLSAPRO. The taxonomic tree for the total dataset is provided in **Supplementary Material 3**.

RESULTS

DNA Barcoding of Saproxylic Beetles With MiSeq

The HTS library we constructed for our 521 sampled individuals representing 343 different species in 39 different families produced an average of 173,664 paired-end reads per pooled plates (sd = 50 083; min. = 97 706 reads; max. = 248 324 reads) with a sequencing depth of around 450X per sample. We recovered 286 partial or complete DNA barcodes (i.e., 54.9% of all samples) representing 193 species (56.3% of all species analyzed). The consensus sequences produced were of high-quality with very few ambiguous base-calls (<1% *N*, except one sequence with <2% *N*). Sequence length varied with the amplification success of both or either one of the two fragments amplified: we recovered 147 full length DNA barcodes (658 bp), as well as 140 and 19 partial DNA barcodes from the C_F (418 bp) and B_R (325 bp) fragments, respectively.

All records (including failed samples) are publicly available in project PSFOR on BOLD, and all sequenced individuals can be found in dataset dx.doi.org/10.5883/DS-NEWCOLEO and in the Table in **Supplementary Material 1**.

Using the C_LepFol primers targeting the full-length DNA barcode, 170 (32.63%) samples produced visible PCR products on agarose gels and were sent for sequencing with Sanger technology. Eventually, 115 specimens (22.1% of the 521 samples) yielded long and high-quality sequences (mean length = 655.5; sd = 12.7; <1% *N*), of which 104 (90%) had also been successfully sequenced using the HTS approach. Overall, the quality of Sanger sequences is higher with only 3 ambiguous bases over 115 sequences (0.026 *N* per sequence), while consensus sequences from Illumina MiSeq reads include a total of 61 ambiguities distributed among the 286 DNA barcodes recovered (0.213 *N* per sequence) (**Table 1**). However, a neighbor-joining analysis (**Supplementary Material 2**) showed a near perfect match between DNA barcode sequences obtained using both sequencing technologies for the same individuals.

For the 286 sequences recovered, the correlation test indicates no significant effect of specimen age on the sequencing ratio success ($S = 264$, $p = 0.07312$, rho = −0.6); age of specimens at time of sequencing seems not to influence sequencing success.

Reliability and By-Products of Illumina MiSeq DNA Barcodes

Multiple contigs were often retrieved from Illumina sequencing. Across all the samples processed, the mean number of contigs per sample after demultiplexing in Geneious was 9, ranging from no sequence recovery to a maximum of 196 contigs for one sample. High discrepancies in sequence number were also observed varying from 1 read per contig to thousands, but generally tended to be in low proportion for by-products compared to the barcode of interest. These different contigs were either lower quality reads clustered apart, chimeras, contaminations, potential heteroplasmy, or bacterial sequences. Our samples being degraded and collected within working collections, where specimens are handled on a regular basis, many co-amplified human DNA or other contaminants from fresh organisms processed in the lab at the same time were found. We took care that no other insect experiment was being conducted during our wet-lab processing to avoid potential misleading contamination. Overall, we identified what we considered to be the genuine consensus sequence by first looking into the ones with the greater number of reads and by blasting these against BOLD or NCBI. In addition to the recovered DNA barcodes, we also recovered consensus

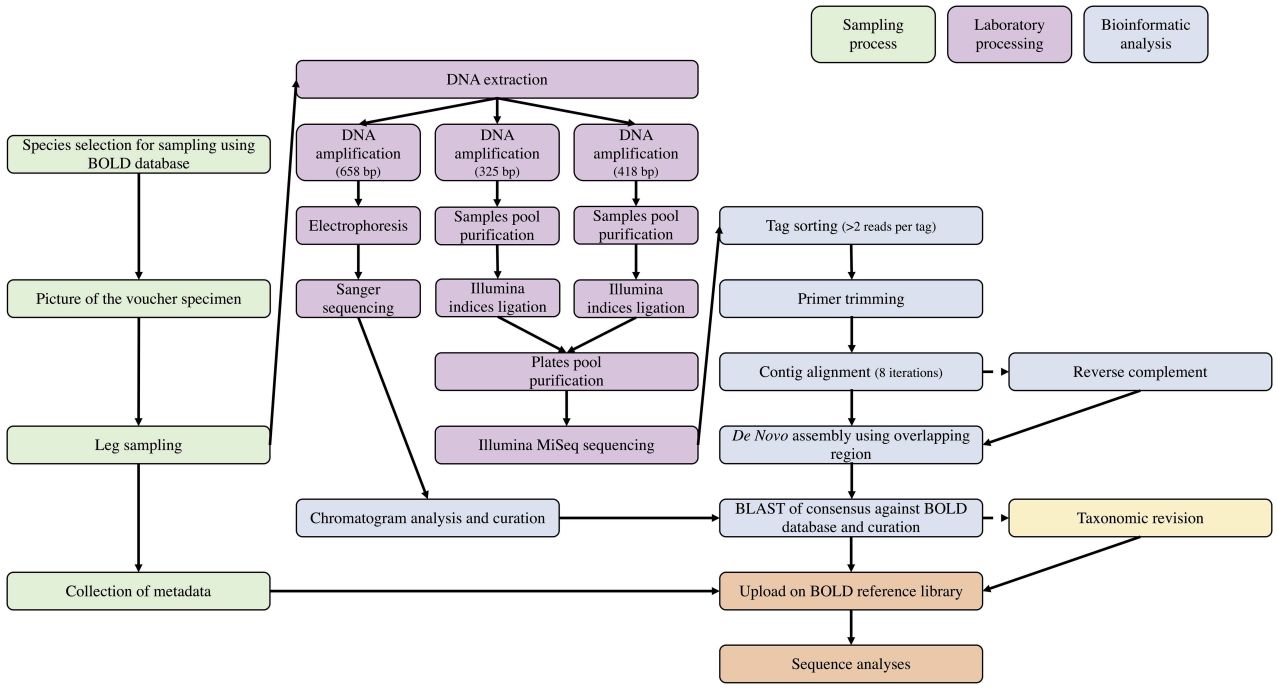


FIGURE 2 | Workflow used in this study for sampling preparation, laboratory processing, and bio-informatic analyses.

TABLE 1 | Summary of the sequencing results for the two sequencing technologies.

	Sequence recovery on total sampling (521)	Number of species recovered on the total sampling (343)	Unique sequence to sequencing method	Unique species to sequencing method	Average sequence length in bp (sd)	Total number of ambiguities (number per sequence)
Sanger	115 (22.1%)	79 (23%)	11	6	655.5 (12.7)	3 (0.026)
Illumina MiSeq	286 (54.9%)	193 (56.3%)	182	120	534.8 (128.3)	61 (0.213)

sequences of by-products potentially interesting for other studies. Indeed, out of 286 specimens successfully yielding a DNA barcode, we identified sequences of *Rickettsia* sp. and *Wolbachia* sp. in 11 and 5 samples, respectively (with one specimen showing co-occurrence of both; see Table in **Supplementary Material 4**).

Genetic Distance Analyses

Overall, we produced 297 new DNA barcodes, of which 180 are full-length (658 bp). From this, 286 were sequenced using our HTS approach, 104 with both HTS and Sanger sequencing, and 11 only with Sanger sequencing after amplification of the full-length DNA barcode fragment. These DNA barcodes represent 199 different species (58% of the species processed), of which 103 are new additions to the reference library for French saproxylic beetles; these new sequences also represent 82 new BINS for BOLD.

The genetic analysis of these 297 newly generated barcodes along with the 1,623 sequences mined from BOLD shows that the means of within-species and within-genus distances are 1.11 and 13.62%, respectively. Within species, genetic distance ranges from 0 to 18.70% whereas we observed 0% to >30.41% within genus (**Table 2**). The frequency distributions of within-species and within-genus distances we observed are consistent with previous results reported in beetles (Hebert et al., 2003b), showing a clear discontinuity in these distributions that form a "barcode gap" near 2% and showing an overall interspecific divergence comprised between 8 and 32% within genera. Our results also highlighted the overall reliability of shorter DNA barcodes to discriminate genera and species (Hajibabaei et al., 2006; Zuccon et al., 2012; Lanner et al., 2019).

As a preliminary investigation of our results of genetic distance analyses, we sought for possible conflicts between current taxonomic identification and DNA barcoding by applying an arbitrary 2% threshold (Hebert et al., 2003b) to highlight cases, in newly sequenced species, of high intra-specific or low inter-specific divergence. Overall, 14 species displayed a maximum intra-specific distance >2% (**Table 3**) and 18 species had a minimum inter-specific distance <2% (**Table 4**).

TABLE 2 | Summary of the genetic distances calculated for sequences with length >400 bp on BOLD with Kimura-2 Parameter and BOLD Aligner for the 297 newly sequenced individuals within the 1,920 sequences of the complete DNA barcode dataset combining our newly generated sequences and preexisting conspecific and congeneric records.

Scale	*n*	Taxa	Comparisons	Min dist (%)	Mean dist (%)	Max dist (%)	*SE* dist (%)
Within species	1,546	257	7,298	0.00	1.11	18.70	0.00
Within genus	1,626	79	42,579	0.00	13.62	30.41	0.00
Within family	1,564	16	78,953	7.90	21.43	34.70	0.00

TABLE 3 | List of newly sequenced species revealing a maximum intra-specific distance >2% using Kimura-2 Parameter, with *n* being the number of individuals (sequence length >400 bp, Bold Aligner alignment).

Family	Species	*n*	Max intra-spe. dist (%)
Cerambycidae	*Stenurella sennii**	2	7.11
Cerylonidae	*Cerylon histeroides**	18	2.34
Curculionidae	*Dryocoetes villosus**	15	15.96
Curculionidae	*Hylastes batnensis**	4	11.09
Elateridae	*Cardiophorus biguttatus**	2	7.46
Elateridae	*Melanotus castanipes*	14	8.08
Elateridae	*Melanotus villosus*	11	8.84
Melyridae	*Psilothrix viridicoeruleus**	3	2.66
Œdemeridae	*Anogcodes seladonius**	5	12.40
Scarabaeidae	*Protaetia cuprea*	20	2.34
Tenebrionidae	*Crypticus quisquilius**	15	4.01
Tenebrionidae	*Isomira murina**	5	15.72
Tenebrionidae	*Tenebrio molitor*	5	5.49
Tenebrionidae	*Tenebrio obscurus**	3	9.15

New cases revealed by this study are highlighted with "".*

TABLE 4 | List of newly sequenced species with a minimum inter-specific distance <2% using Kimura-2 Parameter, with (*n*) being the number of individuals (sequence length >400, Bold Aligner alignment).

Family	Species pairs (number of sequences)	Min inter-spe. dist (%)
Alexiidae	*Sphaerosoma quercus (1)/S. piliferum (4)**	1.47
Bostrichidae	*Sinoxylon perforans (1)/S. muricatum (1)**	0.15
Buprestidae	*Chrysobothris solieri (1)/C. igniventris (1)**	1.55
Cerambycidae	*Stenurella sennii (2)/S. melanura (16)**	0.24
Cerylonidae	*Cerylon impressum (1)/C. ferrugineum (12) /C. histeroides (25)*	0.82
Cleridae	*Opilo cf. domesticus (1)/O. barbarus (1)**	0
Curculionidae	*Kissophagus novaki (2)/kissophagus hederae (7)**	0
Curculionidae	*Pityogenes calcaratus (2)/P. bidentatus (18)**	0.73
Elateridae	*Melanotus villosus (11)/M. castanipes (14)/M. rufipes (11)*	0
Histeridae	*Gnathoncus rotundatus (1)/G. buyssoni (2)**	1.47
Melyridae	*Dasytes caeruleus (1)/Dasytes cyaneus (10)**	0.49
Œdemeridae	*Anogcodes seladonius (5)/A. fulvicollis (2)**	0.77
Ptinidae	*Dorcatoma dresdensis (2)/D. falli (4)**	0
Ptinidae	*Ernobius fulvus (1)/E. gallicus (1)**	0.49
Scarabaeidae	*Protaetia metallica (1)/Protaetia cuprea (20)*	0.49
Tenebrionidae	*Allecula suberina (1)/A. rhenana (1)**	0.24
Tenebrionidae	*Corticeus vanmeeri (1)/C. suturalis (2)**	1.01
Tenebrionidae	*Isomira hypocrita (1)/I. murina (5)/I. semiflava (12)**	0.49

New cases revealed by this study are highlighted with "".*

DISCUSSION

HTS Sequencing of DNA Barcodes From Collection Specimens of Saproxylic Beetles

Our recovery of DNA barcode sequences with Illumina MiSeq is relatively low (55%) though comparable to that reported in Rougerie et al. (2015) using Sanger sequencing and a similar PCR strategy including failure tracking with internal primers (61%). Other studies showed higher sequencing success but used fresh specimens collected specifically for DNA barcoding (Hendrich et al., 2015: 67%; Pentinsaari et al., 2014: 90%).

Sequencing results could vary with preservation, collection methods and age, as well as taxonomically biased primer amplification (Elbrecht and Leese, 2015). Although information about collecting methods was missing for most of our samples, these are known to result mostly from the use of traps that are not adequate for the preservation of DNA. These stay in place in the field for weeks and use non-toxic chemicals such as monopropylene glycol or soap to prevent evaporation and ensure the preservation of specimens. Our analyses show that specimen age at sequencing has no effect on sequence recovery, as opposed to the results reported in collections of Lepidoptera (Hebert et al., 2013) where age appears to be the main determinant of sequencing success. Here, although age certainly remains important, confounding factors linked to collecting and preservation methods might also strongly affect the success of our amplification attempts, despite the use of internal primers. Although we cannot directly measure from our results the possible difference in sequencing success rate using HTS and Sanger technology, as we did not attempt to amplify and sequence the B_F anc C_R fragments with Sanger, there seem to be no significant difference between the two approaches. In particular, we did not observe the increase in success that we had expected considering the high sensitivity of Illumina sequencing and the ability to handle co-amplifications when analyzing HTS reads, whereas these jeopardize the use

of the electropherograms produced with Sanger. Nonetheless, it is interesting that HTS produces multiple products, even in low abundance, that can permit detection and documentation of potential heteroplasmy, pseudogenes or, as exemplified in our dataset, of *Rickettsia* and *Wolbachia* infections (Shokralla et al., 2014; Lanner et al., 2019). Our sequencing depth and read quality did not allow us to have enough information to confirm the occurrence of heteroplasmy in our samples. Nevertheless, these non-targeted co-amplifications allowed us to confirm the presence of endosymbionts in some species (see Table in **Supplementary Material 4**) and suspect potential heteroplasmy.

Applicability and Laboratory Costs

Our recovery rate with HTS is not higher than Rougerie et al. (2015) but the costs are lower. Indeed, our current cost per sample of the Illumina approach we used here—in the molecular facilities at MNHN, from DNA extraction to sequencing, excluding labor—is 4 € per sample, of which we estimate sequencing cost to represent 0.5 € per sample. In comparison, the current cost of bidirectional sequencing using Sanger on a 96-well plate is 4.5 € per sample, meaning that the cost per sample would be 8 € if targeting a single amplicon, or 12.5 € if targeting two shorter, overlapping amplicons as was the case here when processing degraded DNA from collection specimens.

Here, we used a dual-tagging approach instead of a twin-tagging approach as it is advantageous in terms of costs (10-fold less in primers' synthesis costs) but can artificially increase the number of chimeras by tag-jumping during sequencing (Schnell et al., 2015), hence reducing the success of true barcode sequence recovery and increasing the time needed to demultiplex reads. However, both technological developments (e.g., all-in-one library preparation kits) and development of user-friendly bioinformatics tools (Blankenberg et al., 2010; Kearse et al., 2012; Dufresne, 2017) are expected to streamline this process in the future, thus empowering the potential for high-throughput, fast and affordable sequencing of DNA barcodes (Porter and Hajibabaei, 2018). Whereas, the sequencing cost of our approach itself will remain constant while increasing the number of samples processed, the overall cost of the Illumina library preparation could significantly be reduced by optimizing the cost of its multiple steps (e.g., home-made protocols and reagents instead of commercial kits, reduction of PCR volumes through the shift to 384-well plates, automation of purification steps, etc.) (Shokralla et al., 2015; Meier et al., 2016; Wang et al., 2018). Furthermore, this methodology can be applied to various taxa, from both newly collected samples and collection specimens, and allows processing of a large number of samples for a reduced cost.

Quality of HTS Sequences

From a sequence quality point of view, Sanger sequencing is still considered the gold-standard. Hebert et al. (2018) recently emphasized the high-throughput potential of the Sequel sequencing platform from Pacific Biosciences that can generate tens of thousands of full-length DNA barcodes per run from freshly collected samples with low levels of sequencing errors. They showed that sequences resulting from Sequel where largely identical to the ones retrieved with Sanger. Here, we show similar results with our Illumina MiSeq approach (**Supplementary Material 2**). Our recovered DNA barcodes were sometimes shorter than the standard DNA barcode (658 bp length), yet were still consistently useful for species discrimination (Hajibabaei et al., 2006; Lanner et al., 2019). One pitfall of sequencing DNA barcodes with Illumina MiSeq is dealing with multiple amplifications and the possibly resulting ambiguities in assembled consensus sequences. Yet, even though the overall quality of our Illumina produced sequences seems lower than with Sanger sequencing, the quality of each sequence independently remains similar and high (with $<1\%$ N) for all but one sequence. Furthermore, it has recently been shown by Lanner et al. (2019) that read quality from Illumina MiSeq sequencing was in fact equivalent to Sanger, and that drops in quality were mostly due to contamination and co-amplification, detectable with Illumina but less with Sanger. This is consistent with our results where we sometimes had very low number of different reads, artificially increasing the number of ambiguities. We explain it in two ways: first, we sampled specimens in daily-handled collections and captured with unknown killing and preservative reagents, making them more prone to both DNA degradation and environmental contaminations. Second, the use of dual-tagging approach can potentially increase the number of contigs by tag-jumping (Schnell et al., 2015) and therefore reduce the sequencing depth available for the true sequence of our samples. Both issues can blur the genuine signal of consensus sequences, resulting in a higher frequency of ambiguities.

DNA Barcode Reference Libraries of Saproxylic Beetles and Integrative Taxonomy

French Fauna of saproxylic beetles is already relatively well-known and described (Bouget et al., 2019). Overall, our results support previous findings that intra-/interspecific genetic distances derived from DNA barcode analyses do fit species defined on the basis of morphological expertise in most cases. However, we still have identified 14 cases of deep splits (species with high intraspecific divergence) and 18 species pairs that share DNA barcodes (see **Tables 3**, **4**, respectively). As erroneous identifications and synonymies can explain discrepancies between DNA barcoding results and proposed taxonomic names (Mutanen et al., 2016), we reviewed potential synonymies and TN, FS, TB, and GP re-validated together taxonomical identifications from original vouchered specimens for each conflicting result to correct potential errors. The cases mentioned in **Tables 3**, **4** are the result of this integrative dialogue that helped reducing and understand observed discrepancies.

In cases of high intra-specific divergence (**Table 3**), our data reinforce the taxonomic uncertainty already highlighted by Rougerie et al. (2015) in the *Melanotus villosus*/*M. villosus* var. *aspericollis* pair, where the morphological "variety" *aspericollis* consistently and greatly (ca. 5%) differ genetically from *M. villosus*. These results suggest a potential need for revising

the status of *M. villosus* var. *aspericollis* as a distinct species. Regarding *Cardiophorus biguttatus*, this taxon is known to be highly polymorphic and the observed genetic divergences may match different recognized "varieties" of the species that could also deserve distinct specific status. This divergence may also represent geographical structure among populations as the only two specimens sequenced so far come from different areas in France [Pyrénées-Orientales (66) and Var (83) administrative departments] with potential geographical barriers and thus low gene flow between populations. Cases where geographical structure might be driving intraspecific variability may occur within other species, as in the *Psilothrix viridicoeruleus* cluster, or within the newly sequenced group of *Tenebrio obscurus*, as one specimen comes from Romania, another from the Provence region in France and the last one from Corsica island. A case of high intra-specific divergence involving island context is also reported within the *Dryocoetes villosus* complex. Interestingly, we can see that the three newly sequenced individuals from Sainte Marguerite Island in France are highly divergent (over 15%) from their continental counterparts that themselves display low divergence among them (maximum divergence of 0.93%) and are represented by 12 individuals from Germany, Finland and France. Overall, we reveal here several new cases of high genetic divergences within species that may result from incomplete lineage sorting, phylogeographical structure, or represent cases of overlooked cryptic species. Thus, further sampling and analyses are desirable to shed light on these deep split cases. Presence of *Wolbachia* is also known to affect reproductive success and mitochondrial inheritance within the host. Even though *Wolbachia* infection seems not to affect DNA barcoding identification in insects in general (Smith et al., 2012), its potentiality to inflate mitochondrial divergence across populations should be kept in mind (Smith and Fisher, 2009). Nevertheless, our primers were not designed for this purpose, preventing us to shed further light on potential infections.

With respect to cases of low interspecific divergence (**Table 4**), the higher number of reported cases is actually an artifact of discrepancies in the curation of taxon names in databases. For instance, the absence of divergence between the two specimens of *Dorcatoma dresdensis* from France and the four specimens of *Dorcatoma falli* from Germany results from misidentification of the later specimens. Indeed, *D. falli* is a North American species absent in Europe. Careful examination of the available pictures of voucher specimens of the German *D. falli*, confirmed that they indeed are misidentified individuals of *D. dresdensis*. The species complex *Isomira murina*, *I. thoracica*, *I. hypocrita*, and *I. semiflava* (see **Supplementary Material 3**) is another example where different species apparently share similar or highly similar DNA barcodes. Further investigation revealed that *I. semiflava* is in fact a recognized synonym of *I. murina,* but also that the German *I. murina* has been erroneously identified. After verification of the specimen habitus from the voucher picture (BOLD sample ID: GBOL02228), it actually appears to be *I. thoracica*, therefore explaining the high intraspecific variability among *I. murina*. Nonetheless, we could not explain the low divergence between *I. murina* and *I. hypocrita*, two species that are quite distinct both morphologically and geographically, and further studies must be undertaken to understand this result. In the complex of *Protaetia cuprea*—a well-studied European flower beetle taxon—*P. metallica* is a morphologically, geographically and biologically recognized distinct species (Tauzin, 2015). Here, we found however a low genetic divergence between *P. metallica* and other representatives of *P. cuprea* (different subspecies were sampled in this study: *P. cuprea cuprea, P. cuprea olivacea, P cuprea bourgini*), which is consistent with what was previously highlighted in Rougerie et al. (2015) and more recently in Vondráček et al. (2018) from both COI and CytB markers. The later authors actually questioned the specific status of *P. metallica*. These results may suggest a recent origin of these taxa, or ongoing hybridization and introgression, although experimental crossing attempts in captivity failed, suggesting the later to be unlikely (Tauzin, 2015).

Overall, our study expands the current coverage of the DNA barcode reference library for European saproxylic beetles by adding 297 newly sequenced records representing 199 species in 31 families, of which 103 species (82 new BINs) are new additions to the Barcode of Life Datasystems, 26 of which represent genera yet unrepresented in the libraries.

This generated DNA barcode dataset of well-curated and identified collection specimens will be helpful for fast and reliable taxonomical identification for potential mass-trapping and broad biomonitoring studies using genetic approaches. Saproxylic beetles are of major interest with respect to forest health concerns and the need for identification at species level is of great importance to link functional traits and ecological patterns (Gossner et al., 2013).

In total, adding these new sequences to the PASSIFOR dataset (Rougerie et al., 2015) (656 barcodes of 410 species), DNA barcodes reference library now covers 22.4% (598 species out of 2,663 species) of the French fauna of saproxylic beetles (Bouget et al., 2019). When considering records available in BOLD from other European countries, only 1,128 species remain to be barcoded. We created a checklist in BOLD that can be used both for taxonomical curation and tracking of the completeness of the reference library for French saproxylic beetle's fauna. Presently, the completeness of the DNA barcode reference library for the French saproxylic beetle fauna is of 57.6%.

CONCLUSION

Our results emphasize the interest and potential of using HTS technologies—here Illumina MiSeq—as a fast, reliable, and affordable approach to barcode collection specimens that may be challenging or costly to process. The Illumina MiSeq approach used here, despite a relatively low sequencing success, allowed to recover good quality sequences from collection specimens at a reasonable cost.

By adding new sequences of specimens from southern Europe, our study also helps to better assess the intra- and interspecific variability of saproxylic beetles. It also promotes collaboration between specialists to gather enough specimens for sequencing at reasonable costs, and integrative taxonomy to resolve taxonomic uncertainties, correct wet-lab errors and curate public DNA barcode reference libraries. With ongoing

development of amplicon-assembly pipelines as well as long-reads HTS, associated to plummeting sequencing costs, we expect further development of HTS for DNA barcoding and for the sequencing of complete organelle genome. This will accelerate the assembly of DNA barcode reference libraries and reinforce studies relying on DNA-based species identification or delimitation (Tang et al., 2019).

DATA AVAILABILITY STATEMENT

All datasets generated for this study are included in the article/**Supplementary Material**.

AUTHOR CONTRIBUTIONS

RR, CL-V, CB, LS, DG, and RD designed the presented experiment. CB and GP listed the targeted organisms. LS carried out the sampling in natural history collections with the help of TN, TB, FS, GP, and RR. LS, DG, and RR carried out the wet-lab experiments. TN, TB, FS, and GP helped in taxonomic identifications. LS carried out the analyses and wrote the first draft. All authors provided critical feedbacks on the manuscript.

FUNDING

Experiments were funded both by ANR—Belmont Forum to CLIMTREE project: ANR-15-MASC-0002 (PI: CL-V) and by ANR to project SPHINX: anr-16-ce02-0011-05 (PI: RR).

ACKNOWLEDGMENTS

We would like to thank Elisabeth A. Herniou for helpful comments on the manuscript. We are thankful to the three reviewers and to the editor Prof. Rodney L. Honeycutt for their helpful comments on the manuscript. Laboratory work was carried out at the Service de Systématique Moléculaire (SSM), part of the Service Unit Acquisition et Analyse de Données pour l'Histoire Naturelle (2AD) (UMS2700) at the Muséum national d'Histoire naturelle in Paris.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fevo.2019.00495/full#supplementary-material

REFERENCES

Barsoum, N., Bruce, C., Forster, J., Ji, Y., and Yu, D. W. (2019). The devil is in the detail: metabarcoding of arthropods provides a sensitive measure of biodiversity response to forest stand composition compared with surrogate measures of biodiversity. *Ecol. Indic.* 101, 313–323. doi: 10.1016/j.ecolind.2019.01.023

Blankenberg, D., Von Kuster, G., Coraor, N., Ananda, G., Lazarus, R., Mangan, M., et al. (2010). "Galaxy: a web-based genome analysis tool for experimentalists," in *Current Protocols in Molecular Biology,* eds F. M. Ausubel, R. Brent, R. E. Kingston, D. D. Moore, J. G. Seidman, J. A. Smith, and K. Struhl (Hoboken, NJ: John Wiley & Sons, Inc), 19.10.1–19.10.21. doi: 10.1002/0471142727.mb1910s89

Bouget, C., Brustel, H., Noblecourt, T., and Zagatti, P. (2019). *Les Coléoptères Saproxyliques de France: Catalogue Écologique Illustré*. Paris: Muséum National D'Histoire Naturelle, 744. (Patrimoines naturels; 79).

Bourlat, S. J., Haenel, Q., Finnman, J., and Leray, M. (2016). "Preparation of amplicon libraries for metabarcoding of Marine eukrayotes using illumina miseq: the dual-PCR method," in *Methods in Molecular Biology*, Vol. 1452, ed S. J. Bourlat (New York, NY: Springer Science+Business Media), 197–207. doi: 10.1007/978-1-4939-3774-5_13

Chimeno, C., Morinière, J., Podhorna, J., Hardulak, L., Hausmann, A., Reckel, F., et al. (2018). DNA barcoding in forensic entomology - establishing a DNA reference library of potentially forensic relevant arthropod species. *J. Forensic Sci.* 64, 593–601. doi: 10.1111/1556-4029.13869

Cruaud, P., Rasplus, J. Y., Rodriguez, L. J., and Cruaud, A. (2017). High-throughput sequencing of multiple amplicons for barcoding and integrative taxonomy. *Sci. Rep.* 7:41948. doi: 10.1038/srep41948

Deiner, K., Bik, H. M., Mächler, E., Seymour, M., Lacoursière-Roussel, A., Altermatt, F., et al. (2017). Environmental DNA metabarcoding: transforming how we survey animal and plant communities. *Mol. Ecol.* 26, 5872–5895. doi: 10.1111/mec.14350

Dufresne, Y. (2017). *Double Tagged Amplicon Demultiplexing, GitHub Repository*. Available online at: https://github.com/yoann-dufresne/DoubleTagDemultiplexer

Elbrecht, V., and Leese, F. (2015). Can DNA-Based ecosystem assessments quantify species abundance? Testing primer bias and biomass—sequence relationships with an innovative metabarcoding protocol. *PLoS ONE* 10:e0130324. doi: 10.1371/journal.pone.0130324

Fadrosh, D. W., Ma, B., Gajer, P., Sengamalay, N., Ott, S., Brotman, R. M., et al. (2014). An improved dual-indexing approach for multiplexed 16S rRNA gene sequencing on the illumina MiSeq platform. *Microbiome* 2:6. doi: 10.1186/2049-2618-2-6

Fagan-jeffries, E. P., Cooper, S. J. B., Bertozzi, T., Bradford, T. M., and Austin, A. D. (2018). DNA barcoding of microgastrine parasitoid wasps (Hymenoptera: Braconidae) using high-throughput methods more than doubles the number of species known for Australia. *Mol. Ecol. Resour.* 18, 1132–1143. doi: 10.1111/1755-0998.12904

Folmer, O., Black, M., Hoeh, W., Lutz, R., and Vrijenhoek, R. (1994). DNA primers for amplification of mitochondrial cytochrome c oxidase subunit I from diverse metazoan invertebrates. *Mol. Mar. Biol. Biotechnol.* 3, 294–299.

Giangrande, A. (2003). Biodiversity, conservation, and the "Taxonomic impediment". *Aquat. Conserv.* 13, 451–459. doi: 10.1002/aqc.584

Gossner, M. M., Lachat, T., Brunet, J., Isacsson, G., Bouget, C., Brustel, H., et al. (2013). Current near-to-nature forest management effects on functional trait composition of saproxylic beetles in beech forests: functional diversity of beetles. *Conserv. Biol.* 27, 605–614. doi: 10.1111/cobi.12023

Green, S. V. (1998). The taxonomic impediment in orthopteran research and conservation. *J. Insect Conserv.* 2, 151–159. doi: 10.1023/A:1009633811789

Hajibabaei, M., Smith, M. A., Janzen, D. H., Rodriguez, J. J., Whitfield, J. B., and Hebert, P. D. N. (2006). A minimalist barcode can identify a specimen whose DNA is degraded. *Mol. Ecol. Notes* 6, 959–964. doi: 10.1111/j.1471-8286.2006.01470.x

Hallmann, C. A., Sorg, M., Jongejans, E., Siepel, H., Hofland, N., Schwan, H., et al. (2017). More than 75 percent decline over 27 years in total flying insect biomass in protected areas. *PLoS ONE* 12:e0185809. doi: 10.1371/journal.pone.0185809

Hausmann, A., Miller, S. E., Holloway, J. D., deWaard, J. R., Pollock, D., Prosser, S. W. J., et al. (2016). Calibrating the taxonomy of a megadiverse insect family: 3000 DNA barcodes from geometrid type specimens (Lepidoptera, Geometridae). *Genome* 59, 671–684. doi: 10.1139/gen-2015-0197

Hebert, P. D. N., Braukmann, T. W. A., Prosser, S. W. J., Ratnasingham, S., deWaard, J. R., Ivanova, N. V., et al. (2018). A sequel to sanger: amplicon sequencing that scales. *BMC Genomics* 19:219. doi: 10.1186/s12864-018-4611-3

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

Hebert, P. D. N., deWaard, J. R., Zakharov, E. V., Prosser, S. W. J., Sones, J. E., McKeown, J. T. A., et al. (2013). A DNA 'barcode blitz': rapid digitization and sequencing of a natural history collection. *PLoS ONE.* 8:e68535. doi: 10.1371/journal.pone.0068535

Hebert, P. D. N., Penton, E. H., Burns, J. M., Janzen, D. H., and Hallwachs, W. (2004). Ten species in one: DNA barcoding reveals cryptic species in the neotropical skipper butterfly astraptes fulgerator. *Proc. Natl. Acad. Sci. U.S.A.* 101, 14812–14817. doi: 10.1073/pnas.0406166101

Hebert, P. D. N., Ratnasingham, S., and deWaard, J. R. (2003b). Barcoding animal life: cytochrome c oxidase subunit 1 divergences among closely related species. *Proc. Biol. Sci.* 270(Suppl. 1), S96–S99. doi: 10.1098/rsbl.2003.0025

Hendrich, L., Moriniere, J., Haszprunar, G., Hebert, P. D. N., Hausmann, A., Köhler, F., et al. (2015). A comprehensive DNA barcode database for Central European beetles with a focus on Germany: adding more than 3500 identified species to BOLD. *Mol. Ecol. Resour.* 15, 795–818. doi: 10.1111/1755-0998.12354

Hernández-Triana, L. M., Prosser, S. W., Rodríguez-Perez, M. A., Chaverri, L. G., Hebert, P. D. N., and Gregory, T. R. (2014). Recovery of DNA barcodes from blackfly museum specimens (Diptera: Simuliidae) using primer sets that target a variety of sequence lengths. *Mol. Ecol. Resour.* 14, 508–518. doi: 10.1111/1755-0998.12208

Janssen, P., Fuhr, M., Cateau, E., Nusillard, B., and Bouget, C. (2017). Forest continuity acts congruently with stand maturity in structuring the functional composition of saproxylic beetles. *Biol. Conserv.* 205, 1–10. doi: 10.1016/j.biocon.2016.11.021

Ji, Y., Ashton, L., Pedley, S. M., Edwards, D. P., Tang, Y., Nakamura, A., et al. (2013). Reliable, verifiable and efficient monitoring of biodiversity via metabarcoding. *Ecol. Lett.* 16, 1245–1257. doi: 10.1111/ele.12162

Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., et al. (2012). Geneious basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* 28, 1647–1649. doi: 10.1093/bioinformatics/bts199

Kimura, M. (1980). A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J. Mol. Evol.* 16, 111–120. doi: 10.1007/BF01731581

Knittel, T., and Picard, D. (1993). PCR with degenerate primers 9 containing deoxyinosine fails with PFU DNA polymerase. *Genome Res.* 2, 346–347. doi: 10.1101/gr.2.4.346

Lanner, J., Curto, M., Pachinger, B., Neumüller, U., and Harald Meimberg, H. (2019). Illumina midi-barcodes: quality proof and applications. *Mitochondrial DNA A DNA Mapp. Seq. Anal.* 30, 490–499. doi: 10.1080/24701394.2018.1551386

Leray, M., Haenel, Q., and Bourlat, S. J. (2016). "Preparation of amplicon libraries for metabarcoding of marine eukaryotes using illumina miseq: the adapter ligation method," in *Marine Genomics*, ed S. J. Bourlat (New York, NY: Springer), 209–218. doi: 10.1007/978-1-4939-3774-5_14

Liu, L., Li, Y., Li, S., Hu, N., He, Y., Pong, R., et al. (2012). Comparison of next-generation sequencing systems. *J. Biomed. Biotechnol.* 2012, 1–11. doi: 10.1155/2012/251364

Meier, R., Wong, W., Srivathsan, A., and Foo, M. (2016). $1 DNA barcodes for reconstructing complex phenomes and finding rare species in specimen-rich samples. *Cladistics* 32, 100–110. doi: 10.1111/cla.12115

Meyer, M., and Kircher, M. (2010). Illumina sequencing library preparation for highly multiplexed target capture and sequencing. *Cold Spring Harb. Protoc.* 5, 1–10. doi: 10.1101/pdb.prot5448

Meyer, M., Kircher, M., Gansauge, M.-T., Li, H., Racimo, F., Mallick, S., et al. (2012). A high-coverage genome sequence from an archaic Denisovan individual. *Science* 338, 222–226. doi: 10.1126/science.1224344

Mutanen, M., Kivelä, S. M., Vos, R. A., Doorenweerd, C., Ratnasingham, S., Hausmann, A., et al. (2016). Species-level para- and polyphyly in DNA barcode gene trees: strong operational bias in European Lepidoptera. *Syst. Biol.* 65, 1024–1040. doi: 10.1093/sysbio/syw044

Oliverio, A. M., Gan, H., Wickings, K., and Fierer, N. (2018). A DNA metabarcoding approach to characterize soil arthropod communities. *Soil Biol. Biochem.* 125, 37–43. doi: 10.1016/j.soilbio.2018.06.026

Pentinsaari, M., Hebert, P. D. N., and Mutanen, M. (2014). Barcoding beetles: a regional survey of 1872 species reveals high identification success and unusually deep interspecific divergences. *PLoS ONE* 9:e108651. doi: 10.1371/journal.pone.0108651

Piñol, J., Mir, G., Gomez-Polo, P., and Agustí, N. (2015). Universal and blocking primer mismatches limit the use of high-throughput dna sequencing for the quantitative metabarcoding of arthropods. *Mol. Ecol. Resour.* 15, 819–830. doi: 10.1111/1755-0998.12355

Porter, T. M., and Hajibabaei, M. (2018). Automated high throughput animal co1 metabarcode classification. *Sci. Rep.* 8, 1–10. doi: 10.1038/s41598-018-22505-4

Prosser, S. W. J., deWaard, J. R., Miller, S. E., and Hebert, P. D. N. (2015). DNA barcodes from century-old type specimens using next-generation sequencing. *Mol. Ecol. Resour.* 16, 487–497. doi: 10.1111/1755-0998.12474

R Core Team (2017). *R: A Language and Environment for Statistical Computing (version 3.4.3)*. Vienna: R Foundation for Statistical Computing. Available online at: https://www.R-project.org/

Ratnasingham, S., and Hebert, P. D. N. (2007). BARCODING, BOLD: the barcode of life data system. *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

Rougerie, R., Lopez-Vaamonde, C., Barnouin, T., Delnatte, J., Moulin, N., Noblecourt, T., et al. (2015). PASSIFOR: a reference library of DNA barcodes for french saproxylic beetles (Insecta, Coleoptera). *Biodivers. Data J.* 3:e4078. doi: 10.3897/BDJ.3.e4078

Saitou, N., and Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* 4, 406–425.

Schnell, I. B., Bohmann, K., and Gilbert, M. T. P. (2015). Tag jumps illuminated - reducing sequence-to-sample misidentifications in metabarcoding studies. *Mol. Ecol. Resour.* 15, 1289–1303. doi: 10.1111/1755-0998.12402

Shokralla, S., Gibson, J. F., Nikbakht, H., Janzen, D. H., Hallwachs, W., and Hajibabaei, M. (2014). Next-generation DNA barcoding: using next-generation sequencing to enhance and accelerate dna barcode capture from single specimens. *Mol. Ecol. Resour.* 14, 892–901. doi: 10.1111/1755-0998.12236

Shokralla, S., Porter, T. M., Gibson, J. F., Dobosz, R., Janzen, D. H., Hallwachs, W., et al. (2015). Massively parallel multiplex DNA sequencing for specimen identification using an illumina miseq platform. *Sci. Rep.* 5:9687. doi: 10.1038/srep09687

Smith, M. A., Bertrand, C., Crosby, K., Eveleigh, E. S., Fernandez-Triana, J., Fisher, B. L., et al. (2012). Wolbachia and DNA barcoding insects: patterns, potential, and problems. *PLoS ONE* 7:e36514. doi: 10.1371/journal.pone.0036514

Smith, M. A., and Fisher, B. L. (2009). Invasions, DNA barcodes, and rapid biodiversity assessment using ants of Mauritius. *Front. Zool.* 6:31. doi: 10.1186/1742-9994-6-31

Stokland, J. N., Siitonen, J., and Jonsson, B.-G. (2012). *Biodiversity in Dead Wood*. Cambridge; Edimburgh: Cambridge University Press. doi: 10.1017/CBO9781139025843

Stork, N. E., McBroom, J., Gely, C., and Hamilton, A. J. (2015). New approaches narrow global species estimates for beetles, insects, and terrestrial arthropods. *Proc. Natl. Acad. Sci. U.S.A.* 112, 7519–7523. doi: 10.1073/pnas.1502408112

Tamura, K., and Nei, M. (1993). Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in human and chimpanzees. *Mol Biol Evol* 10, 512–526.

Tang, P., Zhu, J.-C., Zheng, B.-Y., Wei, S.-J., Sharkey, M., Chen, X.-X., et al. (2019). Mitochondrial phylogenomics of the hymenoptera. *Mol. Phylogenet. Evol* 131, 8–18. doi: 10.1016/j.ympev.2018.10.040

Tauzin, P.-H. (2015). Chorologie du complexe spécifique Protaetia (Potosia) Cuprea Fabricius, 1775 En France (Coleoptera, Cetoniinae, Cetoniini). *Lambillionea* 115, 99–174.

Thomsen, P. F., and Sigsgaard, E. E. (2019). Environmental DNA metabarcoding of wild flowers reveals diverse communities of terrestrial arthropods. *Ecol. Evol.* 9, 1665–1679. doi: 10.1002/ece3.4809

van Dijk, E. L., Auger, H., Jaszczyszyn, Y., and Claude Thermes, C. (2014). Ten years of next-generation sequencing technology. *Trends Genet.* 30, 418–426. doi: 10.1016/j.tig.2014.07.001

Van Houdt, J. K. J., Breman, F. C., Virgilio, M., and De Meyer, M. (2010). Recovering full DNA barcodes from natural history collections of tephritid fruitflies (Tephritidae, Diptera) using mini barcodes. *Mol. Ecol. Resour.* 10, 459–465. doi: 10.1111/j.1755-0998.2009.02800.x

Vondráček, D., Fuchsová A., Ahrens, D., Král, D., and Šípek P. (2018). Phylogeography and DNA-based species delimitation provide insight into the taxonomy of the polymorphic rose chafer protaetia (Potosia) cuprea species complex (Coleoptera: Scarabaeidae: Cetoniinae) in the Western Palearctic. *PLoS ONE* 13: e0192349. doi: 10.1371/journal.pone.0192349

Wang, W. Y., Srivathsan, A., Foo, M., Yamane, S. K., and Meier, R. (2018). Sorting specimen-rich invertebrate samples with cost-effective ngs barcodes: validating a reverse workflow for specimen processing. *Mol. Ecol. Resour.* 18, 490–501. doi: 10.1111/1755-0998.12751

Yu, D. W., Ji, Y., Emerson, B. C., Wang, X., Ye, C., Yang, C., et al. (2012). Biodiversity soup: metabarcoding of arthropods for rapid biodiversity assessment and biomonitoring: biodiversity soup. *Methods Ecol. Evol.* 3, 613–623. doi: 10.1111/j.2041-210X.2012.00198.x

Zuccon, D., Brisset, J., Corbari, L., Puillandre, N., Utge, J., and Samadi, S. (2012). An optimised protocol for barcoding museum collections of decapod crustaceans: a case-study for a 10-40-years-old collection. *Invertebr. Syst.* 26, 592–600. doi: 10.1071/IS12027

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.

frontiers in Ecology and Evolution

ORIGINAL RESEARCH
published: 21 February 2020
doi: 10.3389/fevo.2020.00028

PCR Cloning Combined With DNA Barcoding Enables Partial Identification of Fish Species in a Mixed-Species Product

*Anthony J. Silva[1], Michael Kawalek[2], Donna M. Williams-Hill[2] and Rosalee S. Hellberg[1]**

[1] One University Drive, Food Science Program, Schmid College of Science and Technology, Chapman University, Orange, CA, United States, [2] Pacific Southwest Food and Feed Laboratory, Office of Regulatory Affairs, Office of Regulatory Science, U.S. Food and Drug Administration, Irvine, CA, United States

DNA barcoding is a valuable tool for regulatory identification of fish species; however, it does not perform well when multiple species are present within the same food product. Therefore, the objective of this study was to examine the use of PCR cloning to identify fish in a mixed-species product that cannot be identified with standard DNA barcoding. A total of 15 fish ball mixtures were prepared with known amounts of Nile tilapia (*Oreochromis niloticus*), Pacific cod (*Gadus macrocephalus*), and walleye pollock (*Gadus chalcogrammus*). Three subsamples from each fish ball underwent DNA extraction, full DNA barcoding (655 bp), and mini-barcoding (226 bp) of the cytochrome *c* oxidase subunit 1 (CO1) gene. Subsamples that did not pass sequencing according to regulatory standards were further analyzed with PCR cloning. All fish balls made of just one species tested positive for that species (i.e., tilapia, cod, or pollock) with both full and mini-barcoding. However, only tilapia was detected in fish balls containing multiple species when tested with standard barcoding techniques, reflecting an inaccurate representation of the fish mixture and suggesting species bias. PCR cloning allowed for identification of Pacific cod in 86% of the mixed-species fish balls tested with full-barcode cloning and 100% of the mixed-species fish ball tested with mini-barcode cloning. However, PCR cloning did not enable the identification of walleye pollock. Standard full barcoding produced more high quality sequences compared to mini-barcoding yet failed to accurately detect all species present in the tested fish mixtures. Overall, the results of this study show that PCR cloning may be an effective method to identify certain fish in mixed-species products when standard DNA barcoding fails. However, additional research is needed to overcome the species bias observed in this study.

Keywords: DNA barcoding, fish mixtures, PCR cloning, species bias, species identification

OPEN ACCESS

Edited by:
David S. Thaler, Universität Basel, Switzerland

Reviewed by:
Mark Stoeckle, The Rockefeller University, United States
Andreanna J. Welch, Durham University, United Kingdom

***Correspondence:**
Rosalee S. Hellberg
hellberg@chapman.edu

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

Received: *19 September 2019*
Accepted: *31 January 2020*
Published: *21 February 2020*

Citation:
Silva AJ, Kawalek M, Williams-Hill DM and Hellberg RS (2020) PCR Cloning Combined With DNA Barcoding Enables Partial Identification of Fish Species in a Mixed-Species Product. Front. Ecol. Evol. 8:28. doi: 10.3389/fevo.2020.00028

INTRODUCTION

Food fraud, including species substitution and mislabeling, is a concern within the seafood industry (Pardo et al., 2016). The United States is a major importer of fish and fish-based products, with annual imports valued at United States $20.5 billion in 2016 (FAO, 2018). The vulnerability of fish-based products to fraud is high due to fluctuations in pricing, quality, supply, and demand. Species substitution and mislabeling is largely motivated through the economic gain that

results from substitution of an inexpensive fish for a premium fish (Khaksar et al., 2015). However, species substitution can have serious consequences, including exposure to toxins and allergens, infringement of religious practices, and financial loss (Armani et al., 2015).

DNA barcoding is typically used by the U.S. Food and Drug Administration (FDA) to identify fish species in food for regulatory purposes (Handy et al., 2011). In DNA full-barcoding, a ∼650 base-pair (bp) region of the cytochrome *c* oxidase subunit 1 (CO1) gene is sequenced and compared to reference sequences to enable species identification. While full barcoding has been shown to work well with raw or minimally processed single-species products, challenges have arisen in the identification of more processed products. One means of addressing these challenges has been the development of DNA mini-barcodes that target shorter regions (∼100–300 bp) of CO1 (Shokralla et al., 2015). DNA mini-barcodes have been found to perform well for species identification in a variety of processed products (Shokralla et al., 2015; Pollack et al., 2018). However, both full and mini DNA barcoding utilize Sanger sequencing and, therefore, often fail to identify species when two or more species are mixed in the same sample (Carvalho et al., 2017b). This is because the presence of multiple species in the same sample can lead to the generation of multiple, overlapping peaks on the resulting sequencing chromatogram, making it unreadable.

Some seafood products, such as fish balls, fish cakes and surimi, are made with a range of fish species and can readily be adulterated due to the lack of morphological identifiers (Galal-Khallaf et al., 2016; Carvalho et al., 2017a). For example, a previous study involving 22 processed cod products (including fish cakes) purchased in Brazil found that 41% of samples were mislabeled and 31% of samples consisted of two or more species (Carvalho et al., 2017b). Mixed fish products, such as fish cakes and fish balls, are consumed worldwide in regions such as Asia, Brazil, and Scandinavia. A wide variety of species are commonly used for production of mixed fish products, typically ranging from 2 to 3 fish species per mixture, including Pacific cod (*Gadus microcephalus*), walleye pollock (*Gadus chalcogrammus*), Pacific whiting (*Merluccius productus*), and tilapia (*Oreochromis* spp.) (Morrissey and Guenneugues, 2000; Ninan et al., 2010; Carvalho et al., 2017b). Cod is the most highly valued of these species, while the latter three are relatively inexpensive and sometimes mislabeled as more expensive fish (Stiles et al., 2013; NOAA, 2019).

PCR cloning has previously been used in combination with DNA barcoding for species identification in mixed-species fish products (Galal-Khallaf et al., 2016). This technique involves the use of an *E. coli*-based cloning vector to isolate DNA amplicons from different species in the same sample (Rondon et al., 2000; Galal-Khallaf et al., 2016). The resulting amplicons can then be sequenced separately and identified using DNA barcoding techniques. PCR cloning in combination with mini-barcoding (127 bp) of the CO1 gene was previously reported to identify species in 100% (29 out of 29) of commercial surimi products tested from China, Singapore, and India (Galal-Khallaf et al., 2016). This method enabled identification of an average of 2.3, 1.6, and 1.0 species per product from Singapore, China, and India, respectively. Common species identified in this study included Sutchi catfish (*Pangasianodon hypophthalmus*), yellowbelly threadfin bream (*Nemipterus bathybius*), and fringescale sardinella (*Sardinella fimbriata*). PCR cloning has been used previously for the identification of species in other applications involving mixed samples, such as detection of animal species in pet food (Donne-Gousse et al., 2005; Teletchea et al., 2005), identification of plant species in honey (Bruni et al., 2015), and analysis of fish species in the fecal material of predators (Deagle et al., 2005; Murray et al., 2011).

Although various DNA barcoding techniques have been established for species identification, no definitive research has been done on the ability of PCR cloning combined with DNA barcoding to identify specific fish in a mixed-species sample using known amounts of each species. Therefore, the objective of this study was to examine the use of PCR cloning combined with DNA barcoding to identify fish in a mixed-species product (i.e., fish balls) containing known amounts of each species. This is the only study to date that has assessed the use of these methods to identify species in known mixtures of fish with varying composition. In accordance with typical fish species testing procedures, all products were first tested with standard DNA barcoding. To simulate regulatory testing, all samples that passed sequencing with standard DNA barcoding were not additionally tested. Products that failed to produce a species identification underwent PCR cloning. This method was tested using both mini-barcoding and full barcoding in order to determine which barcoding technique is most appropriate for this application.

MATERIALS AND METHODS

Sample Collection and Preparation

Fifteen fish ball samples were prepared containing specific proportions of Nile tilapia, Pacific cod, and walleye pollock (**Table 1**). Filets corresponding to each species were purchased from local grocery stores in Orange County, CA, United States. Prior to use in this study, the filets were authenticated with DNA barcoding (described below) and then stored at −20°C until authentication was complete. Fish balls were prepared using an adapted recipe from China Sichuan Food[1]. The authenticated filets from the three species of fish were used to prepare 100-g mixtures at the proportions specified in **Table 1**. Each fish mixture was homogenized with 10 g ice and 10 ml deionized water in a sterile 12-speed Oster blender (Fort Lauderdale, FL, United States) for 2 min at speed 2. Next, 0.3 g of salt and 0.4 g of sugar were added and the mixture was blended for 1–2 min at speed 5. Then, an additional 8 g of ice and 3 ml deionized water were added and mixed for 2 min at speed 11. This step was repeated and blended at speed 4. Finally, 0.4 g of cornstarch and 5 ml deionized water was added to the mixture and blended for 2 min at speed 8. The mixture was then rolled into a 100-g fish ball and heated in 80°C deionized water for 1–2 min. After heating, the fish ball was cooled, placed in an individually labeled Ziploc

[1]https://www.chinasichuanfood.com/how-to-make-fish-balls/

freezer bag (Racine, WI, United States), and stored at −80°C until further analysis.

DNA Extraction

Three subsamples from each fish ball underwent DNA extraction using the DNeasy Blood and Tissue Kit (Qiagen, Valencia, CA, United States), Spin-Column protocol, with modifications. The amount of starting tissue was increased to 100 mg. The fish tissue was mixed with 500 µl Buffer ATL and 55.6 µl proteinase K in a 2-ml microcentrifuge tube and then incubated at 56°C for 2 h at 300 rpm using a Thermomixer C (Eppendorf, Hamburg, Germany). Next, equal parts (556 µl) Buffer AL and 95% ethanol were added to the sample tubes and the tubes were vortexed. A portion (177 µl) of each sample was transferred to a DNeasy Mini spin column in a 2 ml collection tube. Samples were centrifuged (8000 × *g*) for 1 min and the columns were transferred to new collection tubes. The subsequent wash and elution steps were performed as described in Handy et al. (2011). The extracted DNA was stored at −80°C until PCR and DNA sequencing. A reagent negative blank control was included for each set of DNA extractions.

PCR and DNA Sequencing

All DNA extracts underwent PCR and DNA sequencing using both full (655 bp) and mini-barcoding (226 bp) of the CO1 gene. PCR primers (**Table 2**) were synthesized by Integrated DNA Technologies (Coralville, IA, United States) and a Master Cycler Nexus Gradient Thermal Cycler (Eppendorf) was used to perform PCR. PCR amplification for the SH-E mini-barcode was carried out as described in Pollack et al. (2018) with 16 µL of molecular-grade water, 2.5 µL 10X buffer, 2.5 µL $MgCl_2$ (50 nM), 0.5 µL dNTPs (10 mM), 0.5 µL platinum Taq, 0.5 µL of 10 µM forward primer cocktail, 0.5 µL of 10 µM reverse primer, and 2.0 µL of template DNA (Pollack et al., 2018). The cycling conditions for fish mini-barcoding were: 95°C for 5 min; 35 cycles of 94°C for 40 s, 46°C for one min, and 72°C for 30 s; and a final extension step at 72°C for 5 min (Pollack et al., 2018). PCR for the fish full-barcode was carried out as described in Handy et al. (2011) using 6.25 µL 10% trehalose, 2 µL of molecular-grade water, 1.25 µL 10X PCR Buffer, 0.625 µL of $MgCl_2$ (50 mM), 0.062 µL dNTPs (10 mM), 0.060 µL Platinium Taq (5U/µl), 0.125 µL of 10 µM forward primer, 0.125 µL of 10 µM reverse primer, and 1.0 µL of template DNA. The cycling conditions for fish full barcoding were: 94°C for 2 min; 35 cycles of 94°C for 30 s, 55°C for 40 s, and 72°C for 1 min; and a final extension step at 72°C for 10 min (Handy et al., 2011). PCR product confirmation for full and mini-barcodes was carried out with 2% agarose E-Gels (Invitrogen, Carlsbad, CA, United States) run on an E-Gel iBase (Invitrogen) for 15 min (Pollack et al., 2018). The results were visualized using a FOTO/Analyst Express (Fotodyne, Hartland, WI, United States) and Transilluminator (Fisher Scientific, Waltham, MA, United States) combined with FOTO/Analyst PCImage (version 5.0.0.0, FOTODYNE). Samples with a PCR band correlating to the target region length were considered successfully amplified and prepared for DNA sequencing. PCR products were cleaned using ExoSAP-IT (Affymetrix, Santa Clara, CA, United States) following the manufacturer's instructions. Next, bi-directional cycle sequencing was carried out using the M13 primers as described in Handy et al. (2011). Sequencing purification was performed using a Performa DTR V3 96-well short plate (Edge Bio, Gaithersburg, MD, United States). Samples underwent sequencing using a 3500xl Genetic Analyzer (Thermo Fisher Scientific, Waltham, MA, United States) with POP-7 polymer (Thermo Fisher Scientific).

TABLE 1 | Sequencing results for fish ball subsamples tested with standard DNA barcoding techniques (no PCR cloning).

Fish ball sample no.	% Tilapia/cod/pollock (wt/wt/wt)	Full barcoding		Mini barcoding	
		No. of subsamples with acceptable sequences[a]	Top species match	No. of subsamples with acceptable sequences[a]	Top species match
1	98/1/1	2/3	Nile tilapia	3/3	Nile tilapia
2	1/98/1	2/3	Nile tilapia	0/3	N/A
3	1/1/98	2/3	Nile tilapia	1/3	Nile tilapia
4	90/5/5	3/3	Nile tilapia	2/3	Nile tilapia
5	5/90/5	3/3	Nile tilapia	0/3	N/A
6	5/5/90	3/3	Nile tilapia	3/3	Nile tilapia
7	80/10/10	3/3	Nile tilapia	3/3	Nile tilapia
8	10/80/10	1/3	Nile tilapia	0/3	N/A
9	10/10/80	2/3	Nile tilapia	0/3	N/A
10	50/25/25	1/3	Nile tilapia	0/3	N/A
11	25/50/25	3/3	Nile tilapia	2/3	Nile tilapia
12	25/25/50	2/3	Nile tilapia	0/3	N/A
13	100/0/0	3/3	Nile tilapia	3/3	Nile tilapia
14	0/100/0	3/3	Pacific cod	3/3	Pacific cod
15	0/0/100	3/3	Walleye pollock	3/3	Walleye Pollock

The fish balls were prepared with varying proportions of Nile tilapia (O. niloticus), Pacific cod (G. macrocephalus), and walleye pollock (T. chalcogramma), and three subsamples were tested for each fish ball. [a]Based on quality control parameters described in Handy et al. (2011) for full barcodes and Pollack et al. (2018) for mini-barcodes.

Sequence Analysis

Raw sequence data was assembled and edited using Geneious v.5.4.7 (Biomatters Ldt., Auckland, New Zealand) following steps described in Handy et al. (2011). Full barcodes were only considered acceptable if they met the following quality control (QC) parameters: bi-directional sequences with ≥500 bp and <2% ambiguities or single reads with ≥500 bp and ≥98% high-quality bases (Handy et al., 2011). Mini-barcodes were analyzed using QC parameters described in Pollack et al. (2018), which call for bi-directional sequences that are ≥171 bp and have <2% ambiguities or single reads that are ≥171 bp and have ≥98% high-quality bases. Samples that did not produce an assembled sequence underwent repeat DNA extraction, PCR, and sequencing. PCR amplicons from samples with assembled sequences that did not meet QC parameters were used for PCR cloning, due to the assumption that QC failure was due to the presence of a species mixture. Sequences that passed QC were identified to the species level using the Barcode of Life Database (BOLD) Animal Identification Request Engine[2], Species Level Barcode Records. The top species match in BOLD with >98% genetic identity to the query sequence was recorded as the identified species. All sequences obtained in this project were deposited in BOLD (Project Code: AJS). Sequences from each fish species analyzed in this project were uploaded to GenBank (Accession IDs: MN879772, MN879773, and MN879774).

PCR Cloning

Samples with assembled sequences that did not pass QC sequencing parameters were further analyzed through PCR cloning using the Qiagen PCR cloning Kit (Qiagen). Each PCR product (2 μl) was ligated to the commercially prepared Qiagen pDrive A/U cloning vector (1 μl) with 2x buffer (5 μl) and nuclease free water (2 μl) for 2 h at 4°C. Next, the ligations were transformed into *E. coli* competent cells with the addition of 2 μl of ligation-reaction mixture to QIAGEN EZ Competent Cells (Qiagen). This mixture was incubated on ice for 5 min, heated at 42°C for 30 s, and then incubated on ice for 2 min. Next, 250 μl of SOC broth was vortexed in each tube and 100 μl of the sample was plated on Luria Bertani agar containing ampicillin, X-Gal, and Isopropyl B-D-1 thiogalactopyranoside (IPTG). The plates were incubated at 37°C overnight. Next, white colonies bearing PCR strand inserts were transferred to fresh Trypticase Soy Broth with 0.6% Yeast Extract (TSBYE) broth. A plasmid mini-prep was performed on 10 independent plasmid clones for each sample, which served as the template for DNA sequencing. Prior to sequencing, each plasmid clone underwent a restriction digest that included 10 μl plasmid, 2.0 μl 10X buffer, 0.5 μl *Eco*RI, and 7.5 μl molecular grade H_2O incubated in a 37°C water bath for 2 h. The digested plasmids were then mixed with loading dye (5 μl) and 10 μl was pipetted to the appropriate wells of a 2% agarose E-gel to confirm that PCR inserts were still present. If individual cloned isolates did not have PCR inserts, additional clones were selected for a total of 10 PCR bearing clones. Plates with additional white colonies were stored at 4°C in case additional clones needed to be selected for analysis. Ten individual plasmid templates were DNA sequenced in the forward and reverse direction using T7 and SP6 primers, respectively. The raw sequences were analyzed, and top species matches were identified as described above in the "Sequence analysis" section.

[2]http://www.boldsystems.org/

In silico Primer Analysis

Based on the results of DNA barcoding, the full and mini-barcode primers were examined *in silico* for their potential to preferentially amplify Nile tilapia over the other two species tested in this study. All available COI gene sequences from complete mitochondrial genomes were downloaded from GenBank for Nile tilapia (Accession IDs:NC_013663, GU238433, GU370126, GU477624-GU477628), Pacific cod (Accession IDs: AP017650, KY296294, NC_036931) and walleye pollock (NC_004449, MH018252, AB094061, and AB182300-AB182308). The sequences were aligned in Geneious using MUSCLE. The nucleotides in the primer-binding regions of each sequence were examined for mismatches with the primer sequences shown in **Table 2**.

RESULTS AND DISCUSSION

Standard Full Barcoding Without PCR Cloning

As shown in **Table 1**, standard full-barcoding identified Nile tilapia in all 12 of the mixed-species fish ball samples and correctly identified each of the three fish species in the single-species fish ball samples. However, walleye pollock and Pacific cod were not identified in any of the mixed-species fish balls. The average length of the full barcodes that passed quality control was 650 bp, with a range of 558–655 bp (**Table 3**). The sequence quality and percent ambiguities averaged 77.9 and 0.49%, respectively. The overall percent of mixed-species

TABLE 2 | Primer sets used in this study.

Primer set	Primer name	Primer sequence (5′-3′)[a]	Barcode length	References
Fish full barcode	FISHCO1LBC_ts	CACGACGTTGTAAAACGACTCAACYAATCAYAAAGATATYGGCAC	655 bp	Handy et al., 2011
	FISHCO1HBC_ts	GGATAACAATTTCACACAGGACTTCYGGGTGRCCRAARAATCA		
Fish mini-barcode (Mini_SH-E)	Mini_SH-E_F	CACGACGTTGTAAAACGACACYAAICAYAAAGAYATIGGCAC	226 bp	Shokralla et al., 2015
	Mini_SH-E_R	GGATAACAATTTCACACAGGCTTATRTTRTTTATICGIGGRAAIGC		

[a]*Shaded portions of primer sequences indicate M13 tail.*

TABLE 3 | Sequencing results of methods assessed for sequencing parameters and quality control.

Method	No. of acceptable sequences obtained/total[a]	Sequence length (bp)		Sequence quality (% HQ)		Sequence ambiguities (%)	
		Average ± StDev	Range	Average ± StDev	Range	Average ± StDev	Range
Standard full barcoding	27/36	650 ± 27.2	558–655	77.9 ± 25.4	45.5–99.7	0.49 ± 0.60	0.0–1.9
Full barcoding + PCR cloning	55/90	628 ± 41.2	547–655	95.0 ± 10.5	43.6–100	0.12 ± 0.18	0.0–1.9
Standard mini-barcoding	14/36	224 ± 2.70	216–226	95.9 ± 4.18	83.2–99.1	0.23 ± 0.52	0.0–1.8
Mini-barcoding + PCR cloning	111/220	225 ± 0.24	225–226	99.7 ± 0.85	93.4–100	0.01 ± 0.09	0.0–1.0

[a] *36 subsamples underwent both full and mini barcoding. Any subsamples that failed standard barcoding underwent PCR cloning with 10 clones sequenced per subsample.*

subsamples with a species identification (tilapia) was 75.0% (27 of 36), ranging from 33.3% (1 of 3 subsamples) to 100% (3 of 3 subsamples) for individual fish balls. Similarly, Galal-Khallaf et al. (2016) reported a relatively low sequencing rate (45%) for surimi-based mixed fish products. This low rate may be attributed to multiple species producing peaks in a chromatogram during sequencing (Galimberti et al., 2013). In comparison, Pollack et al. (2018) reported a full barcoding sequencing rate of 90% for single-species fish products processed in a variety of ways. The variation in sequencing rates among subsamples in the current study may be due to the possibility of slight variations in the fish ball matrix combined with the use of only 100 mg of sample for DNA extraction.

The consistent detection of only Nile tilapia in all of the mixed-species samples suggests the occurrence of species bias. Bias for a particular species could be due to various factors, including primer bias, mitochondrial copy number differences, and/or genome duplications or insertions of the COI gene. The full barcode primers used in the current study were able to detect Pacific cod and walleye pollock in the single-species fish balls (sample nos. 14–15) and have previously demonstrated the ability to detect these species in single-species processed fish products (Di Pinto et al., 2013; Pollack et al., 2018). Given that this primer set is known to be effective in identifying these fish species, the inability to identify them in mixed-species fish balls suggests the possibility of preferential primer binding to Nile tilapia. Primer bias has not previously been reported with these specific primers; however, it has been reported for DNA barcoding of mixed-fish products using NGS techniques with the cytochrome *b* gene, in which an overrepresentation of skipjack tuna was identified (Kappel et al., 2017). Primer bias has also been reported to be a problem in other studies involving DNA barcoding, such as DNA metabarcoding research involving macroinvertebrates (Elbrecht and Leese, 2015; Deiner et al., 2017; Elbrecht and Leese, 2017).

Due to the possibility of primer bias in the current study, an *in silico* analysis was carried out to examine the potential for the full and mini-barcode primers to preferentially amplify Nile tilapia. The results of the analysis showed very few mismatches when comparing the primer sequences to each species and there was no apparent explanation for the bias observed in this study (**Figure 1**). When the results for all four primers were combined, Nile tilapia showed the greatest number of mismatches (n = 3) with the primer sequences, followed by walleye pollock (n = 1 or 2, depending on haplotype), and Pacific cod (n = 1). Furthermore, none of the mismatches observed with the primers occurred within the first 5 nucleotides of the 3′ end. These results suggest that the bias observed for Nile tilapia may have been due to biological factors, such as differences in mitochondrial copy number or insertions/duplications of the COI gene (Brown, 2008). Analytical bias for a particular species can lead to a misinterpretation of the actual composition of a mixed-species product and could be a concern for regulators and consumers. Future research should be carried out to investigate the likelihood of species bias in mixed products across a wider range of commercial fish species.

PCR Cloning Combined With Full Barcoding

The nine fish ball subsamples that did not generate acceptable sequences with standard full barcoding were partially identified through PCR cloning and DNA sequencing (**Table 4**). Out of the 90 clones sequenced, 55 (61%) had sequences that passed quality control parameters according to Handy et al. (2011). These sequences had an average full-barcode length of 639 bp, average sequence quality of 95.0%, and average ambiguities of 0.12% (**Table 3**). The percentage of clones from each subsample with acceptable sequences ranged from 40% (4 of 10) to 100% (10 of 10). The subsamples with the highest percentage of a single fish (e.g., 98/1/1%) had the highest average sequencing rate, at 70%. Subsamples with 80% of a single fish (e.g., 10/80/10%) had an average sequencing rate of 67% and subsamples in which no fish was present at >50% (e.g., 50/25/25%) had the lowest average sequencing rate, at 43% (**Table 4**).

As shown in **Table 4**, Nile tilapia was identified in all nine subsamples tested with PCR cloning, Pacific cod was identified in six of the subsamples, and walleye pollock was not identified in any of the subsamples. All species-level identifications showed high genetic similarity (≥99.6%) to sequences in BOLD. Overall, the combination of results from standard full barcoding and PCR cloning combined with full barcoding enabled identification of Nile tilapia in all 12 mixed-species fish balls and identification of Pacific cod in 6 of 12 mixed-species fish balls.

Interestingly, there was no correlation between the percentage of each fish in a mixture and the percentage of identifications for that species among the ten clones sequenced. For example, subsample 8-B contained 80% Pacific cod, 10% walleye pollock., and 10% Nile tilapia; however, the sequencing results showed Nile tilapia identifications for 80% of the 10 clones, and Pacific cod identifications for 20% of the clones. This discrepancy is likely a continued effect of the species bias observed with standard DNA barcoding combined with the low number of

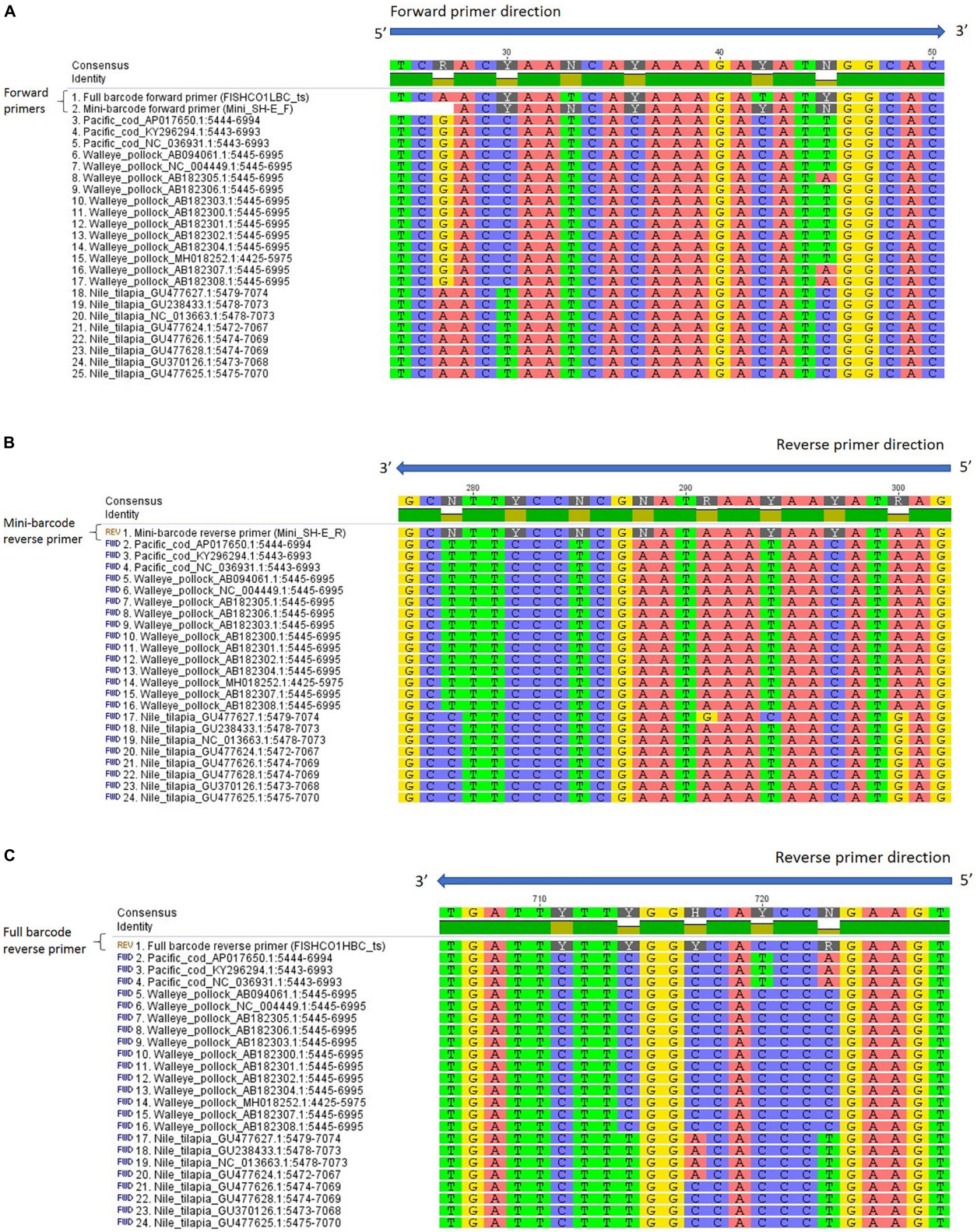


FIGURE 1 | Reference sequences from GenBank aligned with **(A)** mini SH-E and full-barcode forward primers, **(B)** mini SH-E reverse primer (reverse complement), and **(C)** full barcode reverse primer (reverse complement). The sequences were aligned in Geneious using MUSCLE and are shown in the 5′ to 3′ orientation. Primers are shown without the M13 tails and inosine is represented by an "N" in the primer sequence.

TABLE 4 | Sequencing results for fish ball subsamples that underwent PCR cloning and full DNA barcoding after failing standard full barcoding.

Fish ball subsample no.	% Tilapia/cod/pollock (wt/wt/wt)	No. of clones with acceptable sequences[a]	No. of clones identified as each species		
			Nile tilapia (*O. niloticus*)	Pacific cod (*G. macrocephalus*)	Walleye pollock (*T. chalcogramma*)
1-A	98/1/1	9/10	8	1	0
2-A	1/98/1	7/10	7	0	0
3-A	1/1/98	5/10	4	1	0
8-A	10/80/10	4/10	4	0	0
8-B	10/80/10	10/10	8	2	0
9-C	10/10/80	6/10	1	5	0
10-A	50/25/25	5/10	1	4	0
10-B	50/25/25	4/10	4	0	0
12-B	25/25/50	4/10	2	2	0

Ten clones were sequenced for each PCR product that failed standard barcoding. [a]Based on quality control parameters described in Handy et al. (2011).

clones sequenced per subsample. Although PCR cloning enables detection of individual species within mixtures, it is still reliant on the initial PCR amplification step to capture the amplicons representing each species. Bias for Nile tilapia over Pacific cod and walleye pollock during the initial PCR amplification step likely led to a greater number of Nile tilapia amplicons available for the subsequent cloning procedure. While it is possible that sequencing a higher number of clones may result in a more accurate representation of the species present, the matter of bias would also need to be reconciled.

Standard Mini-Barcoding Without PCR Cloning

Standard mini-barcoding identified Nile tilapia in 6 of the 12 mixed-species fish ball samples and correctly identified each of the three fish species in the single-species fish ball samples (**Table 1**). Similar to the results of full barcoding, mini-barcoding did not allow for identification of walleye pollock or Pacific cod in any of the mixed-species fish balls. This is likely due to the species bias described above. The mini-barcodes that passed quality control had an average sequence length of 224 bp, average sequence quality of 95.9% and average ambiguities of 0.23% (**Table 3**). The overall percent of mixed-species subsamples with a species identification (tilapia) was 38.9% (14 of 36), which is lower than that obtained for full barcoding (75.0%). The identification of a fewer number of samples with mini-barcoding as compared to full barcoding may actually be advantageous when working with mixed-species products. This is because sequencing failure is an indication that there may be more than one species in the product, among other things. A sample that fails to be identified with standard barcoding techniques could be flagged for additional analysis while it is likely that additional testing would not be carried out on a sample with a single species identified. This is concerning for the fish product testing sector due to the potential for misinterpretation of results. For example, in this study, 75% of full barcoding subsamples and 38.9% of mini barcoding subsamples produced high quality sequences and were incorrectly identified as 100% tilapia. The misidentification of species composition in a fish product could lead to serious issues, such as non-detection of fish associated with health risks, unwarranted fines for improper labeling, and inaccurate market data regarding the types of fish that are harvested and consumed. In order to enable proper identification of species composition, additional research should be carried out to determine the most appropriate technique for the analysis of mixed-species fish samples.

PCR Cloning Combined With Mini-Barcoding

Among the 22 mini-barcode subsamples that did not pass traditional sequencing, 21 were partially identified with PCR cloning and DNA sequencing (**Table 5**). Out of the 220 clones tested, 111 (50.5%) passed quality control parameters according to Pollack et al. (2018). These sequences had an average mini-barcode length of 225 bp, average sequence quality of 99.7%, and average ambiguities of 0.01% (**Table 3**). Interestingly, the subsamples in which all three species of fish were present at $\geq$25% (e.g., 50/25/25%) had the highest average sequencing rate (77%) and the subsamples with fish at levels as low as 1% (e.g., 98/1/1%) had the lowest average sequencing rate (36%). Similar to the results for PCR cloning of full barcodes, both Pacific cod and Nile tilapia were identified in the mixed-species subsamples. Nile tilapia was detected in the highest number of subsamples ($n = 18$), while Pacific cod was detected in 16 subsamples (**Table 5**). Both species showed high genetic similarity (99.1–100%) to sequences in BOLD. However, consistent with the full barcode cloning results of this study, walleye pollock was not identified in any of the mixed-species subsamples. Overall, the combination of standard mini-barcoding and PCR cloning combined with mini-barcoding enabled identification of Nile tilapia in all 12 mixed-species fish balls and identification of Pacific cod in 9 of 12 (75%) of mixed-species fish balls.

The percent of clones that passed for full barcode cloning was higher (61%) compared to mini barcode cloning (50.5%). There was no correlation between the percentage of each fish in a mixture and the percentage of identifications for that species among the ten clones sequenced. For example, mixture 10-C, which consisted of 50% walleye pollock, 25% Nile tilapia, and 25% Pacific cod, was indicated by sequencing to be 78% Nile tilapia, 22% Pacific cod, and 0% walleye pollock.

TABLE 5 | Sequencing results for fish ball subsamples that underwent PCR cloning and mini-barcoding after failing standard mini-barcoding.

Fish ball subsample no.	% Tilapia/cod/pollock (wt/wt/wt)	No. of clones with acceptable sequences[a]	No. of clones identified as each species		
			Nile tilapia (*O. niloticus*)	Pacific cod (*G. microcephalus*)	*Walleye pollock* (*T. chalcogramma*)
2-A	1/98/1	3/10	0	3	0
2-B	1/98/1	0/10	0	0	0
2-C	1/98/1	6/10	6	0	0
3-B	1/1/98	4/10	3	1	0
3-C	1/1/98	5/10	5	0	0
4-C	90/5/5	5/10	4	1	0
5-A	5/90/5	1/10	1	0	0
5-B	5/90/5	4/10	1	3	0
5-C	5/90/5	7/10	2	5	0
8-A	10/80/10	3/10	3	0	0
8-B	10/80/10	5/10	3	2	0
8-C	10/80/10	1/10	0	1	0
9-A	10/10/80	3/10	0	3	0
9-B	10/10/80	3/10	2	1	0
9-C	10/10/80	7/10	7	0	0
10-A	50/25/25	8/10	5	3	0
10-B	50/25/25	7/10	4	3	0
10-C	50/25/25	9/10	6	3	0
11-B	25/50/25	9/10	7	2	0
12-A	25/25/50	7/10	5	2	0
12-B	25/25/50	8/10	2	6	0
12-C	25/25/50	6/10	0	6	0

Ten clones were sequenced for each PCR product that failed standard barcoding. [a]Based on quality control parameters described in Pollack et al. (2018).

CONCLUSION

Overall, this study revealed the ability of PCR cloning combined with DNA barcoding to identify multiple fish in a mixed-species sample; however, this technique was unable to identify all fish species present. While only one species (Nile tilapia) was identified in mixed-species fish balls using standard DNA barcoding techniques, PCR cloning of the DNA barcode enabled the identification of a second species (Pacific cod) in 50% of fish balls tested with the full barcode and 75% of fish balls tested with the mini-barcode. However, none of the techniques was able to identify the presence of walleye pollock in any of the fish balls. Furthermore, PCR cloning was unable to identify the composition of specific ratios of each fish in the mixture. While the results of this study suggest the occurrence of species bias, additional research is needed to investigate this further. Additional research is also needed to determine whether alternative primer sets would improve detection rates for fish species using the techniques described in this study. The results from this study indicate a concern with the use of standard DNA barcoding for the analysis of mixed-species samples, as the identification of only one of the species within the mixture could be misleading. Therefore, the feasibility of using additional techniques such as PCR cloning or next-generation sequencing for the routine analysis of mixed-species samples should be explored further, including an assessment of the costs and labor involved.

DATA AVAILABILITY STATEMENT

The datasets generated for this study can be found in the BOLD Systems; Project Code AS; http://www.boldsystems.org/.

AUTHOR CONTRIBUTIONS

DW-H, MK, RH, and AS designed the study. AS and MK performed the laboratory work. AS, RH, and MK conducted analysis of the data and prepared the manuscript.

FUNDING

Materials for this study were provided by the FDA and Chapman University. This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

ACKNOWLEDGMENTS

The authors would like to thank Rachel Isaacs for editorial assistance. Materials for this study were provided by the FDA and Chapman University.

REFERENCES

Armani, A., Guardone, L., La Castellana, R., Gianfaldoni, D., Guidi, A., and Castigliego, L. (2015). DNA barcoding reveals commercial and health issues sold on the Italian market in ethnic seafood. *Food Control* 55, 206–214. doi: 10.1016/j.foodcont.2015.02.030

Brown, K. H. (2008). Fish mitochondrial genomics: sequence, inheritance and functional variation. *J. Fish Biol.* 72, 355–374. doi: 10.1111/j.1095-8649.2007.01690.x

Bruni, I., Galimberti, A., Caridi, L., Scaccabarozzi, D., Mattia, F. D., Casiraghi, M., et al. (2015). A DNA barcoding approach to identify plant species in multiflower honey. *Food Chem.* 170, 308–315. doi: 10.1016/j.foodchem.2014.08.060

Carvalho, D. C., Guedes, D., Gloria Trindade, M. D., Sartori Coelho, R. M., and de Lima Araujo, P. H. (2017a). Nationwide Brazilian governmental forensic programme reveals seafood mislabelling trends and rates using DNA barcoding. *Fish. Res.* 191, 30–35. doi: 10.1016/j.fishres.2017.02.021

Carvalho, D. C., Palhares, R. M., Drummond, M. G., and Gadanho, M. (2017b). Food metagenomics: next generation sequencing identifies species mixtures and mislabeling within highly processed cod products. *Food Control* 80, 183–186. doi: 10.1016/j.foodcont.2017.04.049

Deagle, B. E., Tollit, D. J., Jarman, S. N., Hindell, M. A., Trites, A. W., and Gales, N. J. (2005). Molecular scatology as a tool to study diet: analysis of prey DNA in scats from captive Steller sea lions. *Mol. Ecol.* 14, 1831–1842. doi: 10.1111/j.1365-294x.2005.02531.x

Deiner, K., Bik, H. M., Machler, E., Seymour, M., Lacoursiere-Roussel, A., Altermatt, F., et al. (2017). Environmental DNA metabarcoding: transforming how we survey animal and plant communities. *Mol. Ecol.* 26, 5872–5895. doi: 10.1111/mec.14350

Di Pinto, A., Di Pinto, P., Terio, V., Bozzo, G., Bonerba, E., Ceci, E., et al. (2013). DNA barcoding for detecting market substitution in salted cod fillets and battered cod chunks. *Food Chem.* 141, 1757–1762. doi: 10.1016/j.foodchem.2013.05.093

Donne-Gousse, C., Laudet, V., and Hanni, C. (2005). *Method for Determining the Existence of Animal or Vegetable Mixtures in Organic Substances*. Alexandria, VA: U.S. Geological Survey.

Elbrecht, V., and Leese, F. (2015). Can DNA-based ecosystem assessments quantify species abundance? testing primer bias and biomass-sequence relationships with an innovative metabarcoding protocol. *PLoS One* 10:e0130324. doi: 10.1371/journal.pone.0130324

Elbrecht, V., and Leese, F. (2017). Validation and development of COI metabarcoding primers for freshwater macroinvertebrate bioassessment. *Front. Environ. Sci.* 5:11. doi: 10.3389/fenvs.2017.00011

FAO (2018). *The State of World Fisheries and Aquaculture 2018. Contributing to Food Security and Nutrition For All*. Rome: Food and Agriculture Organization of the United Nations.

Galal-Khallaf, A., Ardura, A., Borrell, Y. J., and Garcia-Vazquez, E. (2016). Towards more sustainable surimi? PCR-cloning approach for DNA barcoding reveals the use of species of low trophic level and aquaculture in Asian surimi. *Food Control* 61, 62–69. doi: 10.1016/j.foodcont.2015.09.027

Galimberti, A., De Mattia, F., Losa, A., Bruni, I., Federici, S., Casiraghi, M., et al. (2013). DNA barcoding as a new tool for food traceability. *Food Res. Int.* 50, 55–63. doi: 10.1016/j.foodres.2012.09.036

Handy, S. M., Deeds, J. R., Ivanova, N. V., Hebert, P. D. N., Hanner, R. H., Ormos, A., et al. (2011). A single-laboratory validated method for the generation of DNA barcodes for the identification of fish for regulatory compliance. *J. Aoac Int.* 94, 201–210.

Kappel, K., Haase, I., Kaeppel, C., Sotelo, C. G., and Schroeder, U. (2017). Species identification in mixed tuna samples with next-generation sequencing targeting two short cytochrome b gene fragments. *Food Chem.* 234, 212–219. doi: 10.1016/j.foodchem.2017.04.178

Khaksar, R., Carlson, T., Schaffner, D. W., Ghorashi, M., Best, D., Jandhyala, S., et al. (2015). Unmasking seafood mislabeling in US markets: DNA barcoding as a unique technology for food authentication and quality control. *Food Control* 56, 71–76. doi: 10.1016/j.foodcont.2015.03.007

Morrissey, M. T., and Guenneugues, P. (2000). *Surimi and Surimi Seafood*. Boca Raton, FL: CRC Press.

Murray, D. C., Bunce, M., Cannell, B. L., Oliver, R., Houston, J., White, N. E., et al. (2011). DNA-based faecal dietary analysis: a comparison of qPCR and high throughput sequencing approaches. *PLoS One* 6:e25776. doi: 10.1371/journal.pone.0025776

Ninan, G., Bindu, J., and Joseph, J. (2010). Frozen storage studies of value-added mince-based products from Tilapia (Oreochromis Mossambicus, Peters 1852). *J. Food Process. Preserv.* 34, 255–271. doi: 10.1111/j.1745-4549.2009.00379.x

NOAA (2019). *Imports and Exports of Fishery Products Annual Summary, 2018*. Silver Spring, MA: NOAA.

Pardo, M. A., Jimenez, E., and Perez-Villarreal, B. (2016). Misdescription incidents in seafood sector. *Food Control* 62, 277–283. doi: 10.1016/j.foodcont.2015.10.048

Pollack, S. J., Kawalek, M. D., Williams-Hill, D. M., and Hellberg, R. S. (2018). Evaluation of DNA barcoding methodologies for the identification of fish species in cooked products. *Food Control* 84, 297–304. doi: 10.1016/j.foodcont.2017.08.013

Rondon, M. R., August, P. R., Bettermann, A. D., Brady, S. F., Grossman, T. H., Liles, M. R., et al. (2000). Cloning the soil metagenome: a strategy for accessing the genetic and functional diversity of uncultured microorganisms. *Appl. Environ. Microbiol.* 66, 2541–2547. doi: 10.1128/aem.66.6.2541-2547.2000

Shokralla, S., Hellberg, R. S., Handy, S. M., King, I., and Hajibabaei, M. (2015). A DNA mini-barcoding system for authentication of processed fish products. *Sci. Rep.* 5:15894. doi: 10.1038/srep15894

Stiles, M., Kagan, A., Lahr, H., Pullekines, E., and Walsh, A. (2013). *Seafood Sticker Shock: Why You May Be Paying Too Much for Your Fish*. Accessed at: www.oceana.org/costofseafoodfraud (accessed March 1, 2019).

Teletchea, F., Maudet, C., and Hanni, C. (2005). Food and forensic molecular identification: update and challenges. *Trend Biotechnol.* 23, 359–366. doi: 10.1016/j.tibtech.2005.05.006

Conflict of Interest: The authors declare that the research was conducted in the absence of any commerical or financial relationships that could be construed as a potential conflict of interest. The views in this publication represent those of the authors themselves and do not represent the views of the U.S. Food and Drug Administration. The inclusion of specific trade names or technologies does not imply endorsement by the U.S. Food and Drug Administration nor is criticism implied of similar commercial technologies not mentioned within.

frontiers in Ecology and Evolution

ORIGINAL RESEARCH
published: 26 February 2020
doi: 10.3389/fevo.2020.00027

The Essential Role of Taxonomic Expertise in the Creation of DNA Databases for the Identification and Delimitation of Southeast Asian Ambrosia Beetle Species (Curculionidae: Scolytinae: Xyleborini)

Anthony I. Cognato[1], Gina Sari[2], Sarah M. Smith[1], Roger A. Beaver[3], You Li[4], Jiri Hulcr[4], Bjarte H. Jordal[5], Hisashi Kajimura[6], Ching-Shan Lin[7], Thai Hong Pham[8], Sudhir Singh[9] and Wisut Sittichaya[10]*

[1] Department of Entomology, Michigan State University, East Lansing, MI, United States, [2] Indonesia Agricultural Quarantine Agency, Jakarta, Indonesia, [3] Retired, Chiang Mai, Thailand, [4] School of Forest Resources and Conservation, University of Florida, Gainesville, FL, United States, [5] University Museum of Bergen, University of Bergen, Bergen, Norway, [6] Graduate School of Bioagricultural Sciences, Nagoya University, Nagoya, Japan, [7] Department of Entomology, National Taiwan University, Taipei, Taiwan, [8] Vietnam National Museum of Nature and Graduate School of Science and Technology, Vietnam Academy of Science and Technology, Hanoi, Vietnam, [9] Forest Entomology Division, Forest Research Institute, New Forest, Dehradun, Uttarakhand, India, [10] Department of Pest Management, Faculty of Natural Resources, Prince of Songkla University, Hat Yai, Thailand

OPEN ACCESS

Edited by:
David S. Thaler,
Universität Basel, Switzerland

Reviewed by:
Michael J. Raupach,
Bavarian State Collection of Zoology, Germany
Quentin Wheeler,
SUNY College of Environmental Science and Forestry, United States

***Correspondence:**
Anthony I. Cognato
cognato@msu.edu

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

Received: 04 October 2019
Accepted: 31 January 2020
Published: 26 February 2020

Citation:
Cognato AI, Sari G, Smith SM, Beaver RA, Li Y, Hulcr J, Jordal BH, Kajimura H, Lin C-S, Pham TH, Singh S and Sittichaya W (2020) The Essential Role of Taxonomic Expertise in the Creation of DNA Databases for the Identification and Delimitation of Southeast Asian Ambrosia Beetle Species (Curculionidae: Scolytinae: Xyleborini). Front. Ecol. Evol. 8:27. doi: 10.3389/fevo.2020.00027

DNA holds great potential for species identification and efforts to create a DNA database of all animals and plants currently contains >7.5 million sequences representing ~300,000 species. This promise of a universally applicable identification tool suggests that morphologically based tools and taxonomists will soon not have utility. Here we demonstrate that DNA-based identification is not reliable without the contributions of taxonomic experts. We use ambrosia beetles (Xyleborini), which are known for great diversity as well as global invasions and damage, as a test case. Recent xyleborine introductions have caused major economic and ecological losses, thus timely species identifications of new invaders are necessary. This need is hampered by a paucity of identification tools and a fauna that is only moderately documented. To help alleviate deficiencies in their identification, we created COI and CAD DNA barcode databases (490 and 429 specimens), representing over half of the known fauna of Southeast Asia (165/316 species). Taxonomic experts identified species based on original descriptions and type specimens. Tree, distance, and iterative methods were used to assess the identification and delimitation of species. High intra- and interspecific COI distances were observed for congeneric species and attributed to the beetle's inbreeding system. Neither of the two markers provided 100% identification success but with the neighbor-joining tree-based method, 80% of species were identified by both genes. As for species delimitation, an obvious barcode gap between intra- and interspecific differences was not observed. Correspondence between distance-based groups and morphology-based

species was poor. In a demonstration of iterative taxonomy, we constructed parsimony-based phylogenies using COI and CAD sequences for two genera. Although not all clades were resolved or supported, we provided better explanations for species boundaries in light of morphological and DNA sequence differences. Confident species identifications demonstrated <3% COI and <1% CAD difference and recognition of new species became more probable when there was >10–12% COI and/or >2–3% CAD. Involvement of taxonomic experts from the start of this project was essential for the creation of a stable foundation for the DNA identification of xyleborine species. In general, their role in DNA barcoding cannot be underestimated and is further discussed.

Keywords: CAD, COI, DNA barcoding, species delimitation, species identification

INTRODUCTION

Xyleborine ambrosia beetles (Curculionidae: Scolytinae) occur throughout the world's forests with most of the diversity in the moist tropics where they comprise the majority of the scolytine diversity (Browne, 1961; Wood and Bright, 1992; Hulcr et al., 2015). These beetles exhibit two conspicuous life history traits: they cultivate symbiotic fungi for food in tunnels that they bore into recently dead trees (and their parts), and they are haplodiploid and highly inbred with female-skewed sex ratios averaging 13:1 (Kirkendall, 1993; Cooperband et al., 2016; Castro et al., 2019). These traits have allowed these beetles to colonize the world and gave them their infamous reputation as potential invasive species (Jordal et al., 2001; Gohli et al., 2016; Brockerhoff and Liebhold, 2017). One female with her fungal food stored in specialized body cavities (mycangia) can start a new population after establishing a fungal garden and laying an unfertilized egg which develops into a haploid son. After developing into an adult, the male mates with his mother to produce diploid eggs which develop into females. The adult females mate with their brother and then emerge from the natal nest to beget other families (Kirkendall, 1993; Wood, 2007). These traits allowed multiple lineage radiations on both remote Pacific islands as well as continents from at least 15 million years ago (Cognato et al., 2011, 2018; Jordal and Cognato, 2012). As a result, Xyleborini are the largest and most diverse scolytine tribe, representing nearly 20% of all described species (Hulcr et al., 2015; Gohli et al., 2017). Global trade and the use of wood products as ballast and crating have contributed to an accelerated rate of dispersal of these beetles in many parts of the world (Haack and Rabaglia, 2013; Cognato et al., 2015; Gohli et al., 2017; Meurisse et al., 2018). The first recored introduction of a xyleborine species in the US dates to 1817, but the rate of introduction accelerated with 17 new out of the total 28 exotic species in just the last 30 years (Haack and Rabaglia, 2013; Smith and Cognato, 2015; Gomez et al., 2018; Hoebeke et al., 2018). A subset of these species has also been introduced into Europe in the last two decades (Kirkendall and Faccoli, 2010; Dodelin, 2018).

Most introduced species have an apparently benign effect on the environment because most non-native species occupy woody debris unused by the meager native Holarctic xyleborine fauna (Wood, 1982; Knížek, 2011; Hulcr et al., 2017). However recent findings suggest that the native wood decay fungus community may be displaced by a non-native fungus proliferated by a non-native xyleborine (Hulcr et al., 2018). In addition, three species, *Euwallacea fornicatus*, *Euwallacea perbrevis*, and *Xyleborus glabratus*, and their associated fungi have caused economic and ecological destruction to US orchards and natural forests. These species threaten the multi-million dollar avocado industry and have already altered the ecology of natural landscapes with the loss of millions of red bay trees (Eskalen et al., 2012; Boland, 2016; Carrillo et al., 2016; Hughes et al., 2017).

The introduction of exotic xyleborines presents a serious threat to native forests and much time and funding has been invested at the national level in the US and Europe to detect non-native beetles (Kirkendall and Faccoli, 2010; Rabaglia et al., 2019). The faunas of Europe and America north of Mexico are well-known but a taxonomic impediment concerning tropical xyleborines challenges these efforts by hindering the identification of unknown specimens. Few species identification keys exist for the faunas of the New and Old World tropics where xyleborines are most speciose and this limited knowledge of their diversity hampers the ability to identify species (Kirkendall and Faccoli, 2010). It is estimated that only 75% of the Southeast Asian and 25% of the South American faunas have been described so far (Wood and Bright, 1992; Hulcr et al., 2015; Smith et al., 2017). Even with taxonomic tools, the small and subtle morphological differences that define many xyleborine species make it difficult for non-experts to accurately identify species (Cognato et al., 2015; Gomez et al., 2018; Hoebeke et al., 2018; Smith et al., 2019). Identification of immature stages to species or genus presents the greatest challenge even for experts. This taxonomic impediment could be remedied in part by the creation of a DNA database based on expertly identified specimens, as with other wood-boring beetles (Wu et al., 2017).

At the start of molecular systematics, the use of molecules, especially DNA, for the identification of species was recognized (e.g., Nanney, 1982; Sperling and Hickey, 1994; Foster et al., 2004). The franchise of "DNA barcoding" popularized the use of a ∼700 nucleotide section of mitochondrial cytochrome *c* oxidase subunit I gene (COI), amplified and sequenced with universal primers (Folmer et al., 1994), to identify most animal species (Hebert et al., 2003a). This rapid proliferation of sequences and application to most taxa demonstrated that many species could

be distinguished from related species with obvious reproductive barriers (Hebert et al., 2003b; Sperling, 2003). However, the best use of these data to identify species i.e., tree-based and DNA sequence match identification, was debated (Meier et al., 2006; Taylor and Harris, 2012). Although DNA barcoding was initially invisioned for species identification, diagnosticians readily suggested, and sometimes declared new species for non-monophyletic species recovered in neighbor-joining trees and those that transgressed the 2% barcoding gap (e.g., Hebert et al., 2003b; Barrett and Hebert, 2005; Zahiri et al., 2017). Thus, DNA barcoders trespassed into the field of taxonomy. Delimitation of species based solely on phenetic measures and disgaurd of basic taxonomic prinicples caused much controversy and response from the systematics community (e.g., Will and Rubinoff, 2004; DeSalle et al., 2005; Ebach and Holdrege, 2005; Prendini, 2005; Will et al., 2005; Brower, 2006; Cognato, 2006). Major objections included taxonomy based on one DNA locus, the use of a standardized barcoding gap, neighbor-joining analysis, and the absence of taxonomic expertise in the delimitation of species (see Prendini, 2005 for review). However, approached scientifically with deposition of vouchers, adequate sample size, and the phylogenetic/systematic framework, DNA barcoding data can identify species and contribute to the discovery of new taxa (e.g., Schindel and Miller, 2005; Packer et al., 2009b; Adamski and Miller, 2015; Taft and Cognato, 2017; Gibbs, 2018; DeSalle and Goldstein, 2019).

Issues with the implementation of DNA barcoding still exist for certain taxa (Taylor and Harris, 2012). The universal COI PCR primers fail to amplify DNA for some groups of taxa or particular species within groups (e.g., Hebert et al., 2004; Ward et al., 2005; Smith and Cognato, 2014). This has led to modifications of the original PCR primers to capture the barcoding region, to the use of different primer pairs to capture a partial barcoding region, or to the abandonment of the barcoding region (e.g., Jordal and Kambestad, 2014; Smith and Cognato, 2014). However, nearly all barcoding projects use the fragment as designated by Hebert et al. (2003a). Different evolutionary rates within some highly divergent or conserved taxa hamper identification because of non-uniform nucleotide differences and challenge the use of a standard barcoding gap to distinguish species (Hebert et al., 2003b; Cognato, 2006). In addition, taxa with non-sexual or inbreeding mating may defy standard species concepts as they do based on morphology. Issues with heteroplasmy and pseudogenes (numts) can also decrease the accuracy of identification with the use of the COI barcoding region and mitochondrial DNA in general (Song et al., 2008; Magnacca and Brown, 2010; Moulton et al., 2010; Jordal and Kambestad, 2014). The adoption of different genes for identification can help to alleviate these COI barcoding region issues for some taxa (e.g., Foster et al., 2013).

Taxonomic experts have been underutilized in developing DNA barcodes. Among DNA barcoding studies, either taxonomists are ignored (e.g., Lait and Hebert, 2018), mentioned only as identifiers (e.g., Kekkonen and Hebert, 2014), or called upon to interpret the taxonomic implications of the resulting data in subsequent studies (e.g., Barrett and Hebert, 2005). The exclusion of taxonomists or explicit taxonomic methodology for DNA barcoding studies can yield suspect conclusions or irreproducible results (e.g., Hebert et al., 2004; Chang et al., 2014). Also the discovered "new species" only add to the taxonomic impediment if the species are not formally described (e.g., Brower, 2010; Pinheiro et al., 2019). Incorporation of taxonomists from the start of a DNA barcoding project would alleviate many of the mentioned issues, as observed in the more informative barcoding studies (e.g., Trewick, 2008; Packer et al., 2009a).

Although there are potential issues and limitations of DNA barcoding using COI, preliminary data suggest the feasibility of identification and delimiting xyleborine species (Dole et al., 2010; Cognato et al., 2011, 2015, 2019; Jordal and Kambestad, 2014; Stouthamer et al., 2017; Gomez et al., 2018). Studies of a few closely related species of different genera demonstrated: (1) The universal or scolytine specific barcoding COI primers produce PCR products and DNA sequences for most species; (2) non-monophyletic species; (3) high intraspecific nucleotide difference ($>$ 10% as compared to 2–3% for outbreeding scolytines) (4) the use of nuclear genes as alternative diagnostic loci; and (5) the results of a few studies identified new species (Gomez et al., 2018; Cognato et al., 2019). In addition, there are currently overlapping generations of scolytine taxonomists that can identify specimens to species and can interpret the DNA results in reference to these identifications.

In this study, we develop a DNA identification foundation for 165 species of 316 Southeast (SE) Asian xyleborines representing more than half the known species. The goal is to create a DNA barcode resource in conjunction with the historically most comprehensive taxonomic revision of the fauna (Smith et al., in preparation), intended to serve as a model taxonomic product where DNA barcodes and morphological systematics are iteratively used and in mutual support. Another intent is to integrate fundamental biosystematics with direct application: species of this fauna are the most often intercepted wood borers at US ports-of-entry, therefore diagnosticians need a dataset of authenticated DNA sequences as an identification tool (Haack and Rabaglia, 2013). Anticipating the issue of high COI nucleotide divergence we tested the species identification potential of an alternative locus—in this case CAD. Although, any other gene locus could potentially provide species diagnostic DNA sequences such as, 28S rDNA, preliminary rDNA data suggested that the species level nucleotide variation of this locus was not consistant among scolytine taxa (Jordal and Kambestad, 2014; Cognato et al., 2019). We compare tree-based and DNA match methods for the identification of species and demonstrate the use of DNA barcodes for the discovery of species. We demonstrate that the use of COI and CAD can help the identification and delimitation of xyleborine species and discuss the role of the taxonomist in the creation of a DNA barcoding database.

MATERIALS AND METHODS

Specimens

Specimens were collected from various localities in SE Asia via excision of the beetles infesting wood or from ethanol baited

flight interception traps. A total of 508 individuals representing 33 genera and 258 species with more than half from SE Asia (165) were included in this study (**Supplementary Table 1**). Specifically, 490 and 429 individuals were included in the COI and CAD datasets, respectively. The head and pronotum were removed and placed in a 1.5 ml microfuge vial for the extraction of DNA. DNA extraction followed using Qiagen tissue extraction kit and protocol (Qiagen Ltd., Hilden, Germany). Pinned vouchers were deposited at the A.J. Cook Arthropod Research Collection, Michigan State University. Specimens were identified to species based on comparison to type specimens and published descriptions by SMS, RAB, and AIC. We consider these morphologically-based identifications null hypotheses of species limits.

DNA Amplification and Sequencing

DNA sequences of partial Cytochrome *c* Oxidase subunit I (COI) mtDNA were generated with primers LCO1490: 5′-GGTCAACAAATCATAAAGATATTGG-3′ and HCO2198: 5′-TAAACTCAGGGTGACCAAAAAATCA-3′ (Folmer et al., 1994). Each 25 ml PCR reaction contained 4.5 ml template DNA; 2.5 ml buffer; 1 ml $MgCl_2$; 0.5 ml dNTPs; 0.75 ml each primer; 0.25 ml of hot star taq and the reactions were subjected to the PCR thermal protocol listed in Hebert et al. (2003a). When PCR failed, a primer pair designed for scolytines was used (1495b: 5′-AACAAATCATAAAGATATTGGRAC-3′ and rev750: 5′-GAAATTATNCCAATTCCTGG-3′; Smith and Cognato, 2014). PCR amplification protocol consisted of 15 min denaturation at 95°C and 38 cycles at 94°; 50°C each for 30 s and 72°C 45 s, followed by a 5 min extension at 72°C.

Sequences of the nuclear protein coding gene (CAD) gene were generated with forward primers CADforB2 5′ GARAARGTNGCNCCNAGTATGGC-3′ (Jordal et al., 2011) or CADfor4 5′ TGGAARGARGTBGARTACGARGTGGTYCG-3′ (Danforth et al., 2006) and the reverse primer apCADrev1mod 5′ GCCATYRCTCBCCTACRCTYTTCAT-3′ (Danforth et al., 2006). Each 25 ml PCR reaction contained 4.5 ml template DNA; 2.5 ml buffer; 1 ml $MgCl_2$; 0.5 ml dNTPs; 0.75 ml each primer; 0.25 ml of hot star taq. PCR amplification protocol consisted of 15 min denaturation at 95°C and 35 cycles at 94°C for 30 s; 55°C for 30 s and 72°C for 1 min, followed by a 5 min extension at 72°C.

PCR products were electrophoresed and visualized on 1.5% TAE agarose gel stained with ethidium bromide. PCR products were purified of excess primers and unincorporated nucleotides using ExoSAP-IT™ following the manufacturer's protocol (Thermo Fisher Scientific). Sequencing of the purified PCR products occurred at the Research Technology Support Facility at Michigan State University using BigDye Terminator v.1.1 (Applied Biosystems, Foster City, CA, USA) cycle sequencing kit and visualized on an ABI 3730 or 3700 (Applied Biosystems). The DNA sequences were compiled and inspected with Sequencer 4.7 (Gene Codes, Ann Arbor, MI, USA). Sequences were assessed for potential pseudogenes following the recommendations of Jordal and Kambestad (2014). Consensus sequences derived from the forward and reverse sequences were used in subsequent analyses and deposited in Genbank (**Supplementary Table 1**).

Taxon Identification

For the tree-based method, COI and CAD sequences were assembled in separate NEXUS files using the software PAUP version 4.0a (build 161) (Swofford, 2002). Previously published sequences were also included from studies in which we could verify the species status of vouchers (Cognato et al., 2011, 2015, 2019). These specimens provided a global context as many of these species occurred outside the study area. Neighbor-joining trees were calculated using uncorrected "*p*"–distances. We used "*p*"–distance instead of Jukes-Cantor (Jukes and Cantor, 1969) or Kumura-2 (Kimura, 1980) models of nucleotide substitution because these models do not affect the interspecific distance among closely related species thus not benefiting the identification of species (Srivathsan and Meier, 2012). The number of monophyletic species and genera were recorded.

DNA sequence match methods rely only on DNA sequence similarity without reliance on the clustering of sequences in a neighbor-joining tree (Saitou and Nei, 1987). This is advantageous because it avoids the pitfalls of neighbor-joining analysis (DeSalle and Goldstein, 2019) and includes percent sequence difference criterion in species identification. Using the TaxonDNA software (Meier et al., 2006), we calculated DNA sequence match for the COI and CAD sequences and recorded the number of successful, ambiguous, and misidentifications of species. We varied the analyses by including best match, best close match, and all species barcode criteria. Best match criterion is the most relaxed given the query sequence needs to match only one sequence without regard to percent similarity. For the best close match criterion, the query sequence needs to match a threshold percent similarity observed in 95% conspecifics. The chosen threshold percent similarities for the genes were traditional barcodes gaps (2 and 3%) and approximate barcode gaps based on the empirical data. The species barcode criterion is similar to the best close match method but the query sequence needs to match all conspecific sequences as top matches.

Taxon Discovery

We used Automatic Barcode Gap Discovery (ABGD) and TaxaDNA to identify COI and CAD barcode gaps among species (Meier et al., 2006; Puillandre et al., 2012). Although other means for assessing barcode gaps exist (such as, Ratnasingham and Hebert, 2013), these methods provide assessment of multiple gap values and models of nucleotide evolution. We used TaxaDNA software to cluster sequences based on the barcode gaps and determined the number of violations of the predetermined taxonomy based on morphology and comparison to type specimens. Different barcode gap values were assessed with ABGD software (http://wwwabi.snv.jussieu.fr/public/abgd/abgdweb.html, accessed 9 August 2019) where Pmin = 0.001, Pmax = 0.1, Steps = 10, and the relative gap width (X) = 1.0 for both genes. Preferred groups of sequences were selected based on an intermediate value of P after an initial steep decline in number of estimated groups (Puillandre et al., 2012).

For two genera we provide examples of the application of iterative taxonomy (Yeates et al., 2011) to deliberate species limits. Based on monophyletic genera as found in the CAD NJ-tree, we created NEXUS files for the species of *Ambrosiophilus*

and *Euwallacea* which included COI and CAD sequences for the corresponding species. For these data sets, we performed maximum parsimony analyses using a heuristic search with 100 random stepwise additions. Non-parametric bootstrap (Felsenstein, 1985) values were calculated for all generic datasets with 500 pseudoreplicates using a heuristic search with simple stepwise additions. These results were discussed in reference to morphological characters typically used to diagnose species (Hulcr et al., 2007).

RESULTS

PCR and Sequencing

The PCR primer pairs do not reliably amplified the target locus for COI and CAD. The COI primers 1495b and rev750 and CAD primers ApCADfor4 and CADrev1mod amplified the target loci most often. The combination of COI and CAD primer pairs had 88 and 72% success rates, respectively. COI sequences showed no double peaks in chromatograms, however double peaks were observed in some in CAD chromatograms, which we attributed to allelic variation. These nucleotide positions were labeled with an appropriate ambiguity code.

Taxon Identification

In the tree based identification method, monophyly of genera, and species was not found for all taxa in COI and CAD neighbor-joining trees (**Table 1** and **Supplementary Figures 1**, **2**). However, of the ∼65% of species that were represented >1 sequence, 80% of the species were identified for both genes. CAD neighbor-joining tree resolved more monophyletic genera (17) as compared to the COI neighbor-joining tree (7) which is expected given the observed high COI nucleotide substitution rate (see below). Fifty percent of the COI sequences were successfully clustered with the same species, while 14% did not and 35% had an ambiguous placement. Fifty-two percent of the CAD sequences were successfully clustered with the same species, while 11% did not and 39% had an ambiguous placement.

The DNA sequence match identification did not perform as well as the tree based identification (**Table 2**). For both genes, best match of sequences performed the worst with 54–60% successful identifications but also with 35–40% misidentifications. For COI, the all-species barcode criterion was the most stringent and only 25 and 34% of identifications were successful at 3 and 9% thresholds respectfully. For COI, the best close match performed the best at 9% threshold with 55% successful identification as compared to 42% successful identification at a 3% threshold. The number of ambiguous and misidentified sequences was below 3%. At 2 and 3% thresholds for CAD, success with the best close match and all species barcode criteria was similar to COI however, ambiguous, and misidentification of sequence ranged from 4 to 49%.

Average interspecific difference for congenerics ranged 9.3–16.3% for COI and 0.86–10% for CAD (**Table 1**). Most genera with <13.6% COI interspecific difference were monophyletic while the association between interspecific difference and monophyly was not consistent. Intraspecific differences averaged 8.34% (most <10%) and 1.26% (most <2%) for COI and CAD, respectively (**Figure 1**).

Taxon Discovery

Barcode gaps between interspecific and intraspecific differences for COI and CAD were not distinct (**Figures 2**, **3**). These differences greatly overlapped between 12 and 17% for COI, and 1 and 3% for CAD (**Figures 2**, **3**). TaxonDNA analyses found minimum of DNA cluster threshold violations at 9% for COI and 3% for CAD, respectively (**Table 3**). The ABGD analyses did not find any gaps in the distribution of sequence differences for both genes. The correspondence between ABGD groups and taxonomically recognized species was poor. The species were divided into 394 and 251 groups for COI ($P = 0.00278$) and CAD ($P = 0.0017$), respectively which consisted of mostly over split species while in other cases different species were grouped together.

Iterative Taxonomy

Parsimony analysis found one most parsimonious tree for 11 individuals of *Ambrosiophilus* which was represented by five *A. osumiensis* specimens (**Figure 4**). The clade containing all *A. osumiensis* individuals and two internal clades had bootstrap values above 95%. All other clades had lower bootstrap values (<70%). Percentage COI and CAD sequence difference among the *A. osumiensis* individuals range from 3.5–7.5 and 1.2–2.7%. Compared to its sister species *A. subnepotulus*, the COI sequence difference ranged 12.9–15.8% (*A. subnepotulus* CAD was missing from the dataset). Total interspecific COI and CAD sequence differences ranged 15.3–20.2 and 3.6–7.9%, respectfully. Considerable morphological differences occur among the clades of *A. osumiensis.* Such variation occurs in the shape of the pronotum; in the minute structure of the elytral declivity and pronotal disc; interstriae width; strial puncture size; number and size of tubercles on declivital interstriae 2; antennal club type (Hulcr et al., 2007); amount of elytral vestiture; and body size, with individuals differing up to 0.9 mm in length (0.5 mm or less is typical, Smith, unpublished).

Parsimony analysis found 2,475 most parsimonious trees for 57 individuals of *Euwallacea* (**Figure 5**). Twenty species were included with seven species represented by more than one individual. Only *Euwallacea fornicatus*, *Euwallacea interjectus*, *Euwallacea velatus*, and *Euwallacea wallacei* were monophyletic. The COI and CAD sequence difference among the *Euwallacea fornicatus* individuals ranged 1.4–3.2 & 0.0–0.7% and between the sister-species *E. kuroshio*, 9.8–10.9 & 1.2–1.7%. *Euwallacea interjectus* was subdivided by two internal clades (A, B, & C) with bootstrap values above 95. Overall percentage COI and CAD sequence difference among the *E. interjectus* individuals ranged from 0.3–15.7 to 0.4–2.6%, respectfully. However, within clades A, B, and C 2.0, 2.7, and 0.3–3.7% COI sequence differences were observed, respectively. For CAD sequence differences only one comparison was observed for clade A (0.5%) while a range sequence differences (0.2–0.8%) was observed for clade C. *Euwallacea andamanensis, Euwallacea funereus, Euwallacea similis,* and *Euwallacea semirudis,* were not monophyletic and

TABLE 1 | Tree based identification: monophyly for xyleborine species and genera found in the neighbor-joining analyses.

									% Mean interspecific		Genus	
	#Of seq/spp.		Successful seq./spp.		Failed seq./spp.		Ambiguous		Divergence		Monophyletic	
	COI	CAD	COI	CAD	COI	CAD	COI	CAD	COI	CAD	COI	CAD
Amasa	18/12	11/10	8/2	0/0	0/0	2/1	10	9	13.6	3.3	NO	YES
Ambrosiodmus	20/7	20/7	11/3	15/3	6/1	2/1	3	3	14.5	2.8	NO	YES
Ambrosiophilus	11/6	8/5	5/1	5/2	2/1	0/0	4	3	15.4	4.3	NO	NO
Ancipitis	1/1	1/1	NA	NA	NA	NA	1	1	NA	NA	NA	NA
Anisandrus	22/16	20/15	11/4	9/4	0/0	0/0	11	11	14.6	3.7	NO	NO
Arixyleborus	14/10	13/11	6/2	3/1	0/0	0/0	8	10	15.6	4.2	NO	YES
Beaverium	7/6	7/6	2/1	2/1	0/0	0/0	5	5	15.1	2.6	NO	YES
Cnestus	17/8	13/6	12/3	9/2	0/0	0/0	5	4	13.7	8.1	NO	YES
Coptoborus	2/2	2/2	NA	NA	NA	NA	2	2	9.7	1.9	YES	YES
Coptodryas	9/7	8/6	0/0	3/1	3/1	0/0	6	5	14.4	7.7	NO	NO
Cryptoxyleborus	2/2	3/3	NA	NA	NA	NA	2	3	14.1	10	NO	NO
Cyclorhipidion	33/20	30/20	17/4	12/3	0/0	0/0	16	18	15.3	5.5	NO	NO
Debus	13/9	11/8	2/1	6/3	5/2	0/0	6	5	15.9	3.1	NO	YES
Diuncus	13/8	8/5	6/2	5/2	2/1	0/0	5	3	15.1	4.9	NO	YES
Eccoptopterus	9/6	7/4	0/0	0/0	4/1	4/1	5	3	13	4.1	YES	YES
Euwallacea	52/19	45/17	4/1	28/5	36/6	8/2	12	9	15.5	4.4	NO	NO
Hadrodemius	7/2	6/2	6/1	5/1	0/0	0/0	1	1	12.4	4.5	YES	YES
Heteroborips	2/2	2/2	NA	NA	NA	NA	2	2	10	3	YES	YES
Immanus	1/1	2/2	NA	NA	NA	NA	1	2	NA	5	NO	NO
Leptoxyleborus	3/1	3/1	3/1	3/1	0/0	0/0	0	0	NA	NA	NA	YES
Microperus	31/17	27/17	22/9	13/4	2/1	3/2	7	11	15.1	3.7	NO	NO
Planiculus	17/5	14/4	6/1	12/2	8/1	0/0	3	2	14.2	2.8	NO	YES
Sampsonius	1/1	1/1	NA	NA	NA	NA	1	1	NA	NA	NA	NA
Schedlia	1/1	1/1	NA	NA	NA	NA	1	1	NA	NA	NA	NA
Stictodex	2/2	2/1	NA	2/1	NA	0/0	2	0	9.7	NA	NA	NA
Streptocranus	3/3	3/3	NA	NA	NA	NA	3	3	9.3	1.7	YES	YES
Taurodemus	1/1	1/1	NA	NA	NA	NA	1	1	NA	NA	NA	NA
Theoborus	2/2	2/2	NA	NA	NA	NA	2	2	12.9	5	NO	NO
Truncaudum	7/2	7/2	6/1	6/1	0/0	0/0	1	1	13.6	0.86	YES	YES
Webbia	5/5	5/5	NA	NA	NA	NA	5	5	13.6	3	YES	YES
Xyleborinus	36/15	29/14	24/5	21/6	3/1	0/0	9	8	14.5	3.4	NO	YES
Xyleborus	56/28	58/28	37/9	32/6	0/0	6/2	19	20	16.3	3.8	NO	NO
Xylosandrus	72/24	58/22	60/12	30/7	0/0	16/3	12	12	14.8	5.4	NO	NO
Total	490/251	429/233	248/63	222/56	71/16	47/13	171	166			7/27	17/17

intraspecific COI and CAD sequence difference ranged 1.0–16.8 & 0.0–4.4%. Many *Euwallacea* species demonstrated little to no morphological variation in characters typically used to diagnose xyleborine species, particularly in the sculpturing of the elytral declivity.

DISCUSSION

This study is the first to describe the application of COI and CAD DNA sequences for the identification and delimitation of xyleborine ambrosia beetles based on the largest sampling of species, to date, representing nearly all genera. The most striking observation is the prevalent high amount of COI intraspecific and interspecific pairwise differences which also was observed in earlier studies of limited xyleborine species (**Figures 1–3**) (Dole et al., 2010; Cognato et al., 2011). There are many reasons for high intraspecific COI sequence differences including unrecognized putative cryptic species, poorly defined species boundaries, effects of *Wolbachia* infection, and pseudogenes (Funk and Omland, 2003; Rubinoff et al., 2006). Most of the morphologically defined species for all genera exhibit 10–12% difference; thus, we contend that this observation is not the result of rampant cryptic speciation that one would expect given a 2% standard sequence divergence between species as promoted by the Barcode initiative (Hebert et al., 2003b; Ashfaq

TABLE 2 | Identification success with various similarity thresholds using DNA match method.

COI	Success	Ambiguous	Misidentification	No match closer than 9%
Best match	289 (59%)	31 (6.3%)	169 (34.6%)	N/A
3%				
Best close match	206 (42.1%)	0	5 (1.0%)	278 (56.9%)
All species barcode	123 (25.2%)	87 (17.8%)	1 (0.2%)	278 (56.9%)
9%				
Best close match	269 (55%)	0	15 (3.1%)	205 (41.92)
All species	167 (34.2%)	116 (23.7%)	1 (0.2%)	205 (41.92)
CAD				
Best match	230 (53.6%)	28 (6.5%)	171 (39.9%)	N/A
2%				
Best close match	222 (51.7%)	19 (4.4%)	98 (22.8%)	90 (21.0%)
All species barcode	154 (35.9%)	168 (39.2%)	17 (4.0%)	90 (21.0%)
3%				
Best close match	226 (52.7%)	25 (5.8%)	137 (31.9%)	41 (9.6%)
All species barcode	157 (36.6%)	211 (49.2%)	20 (4.7%)	41 (9.6%)

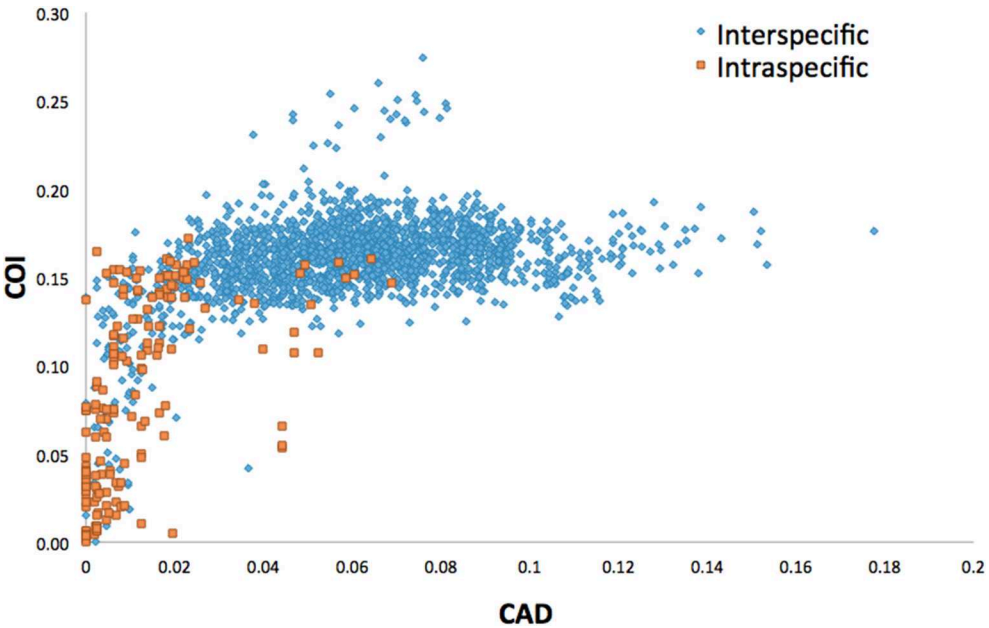


FIGURE 1 | Pairwise uncorrected "*p*" intra- and interspecific distances for COI and CAD.

and Hebert, 2016). Cryptic species are evident at intraspecific differences ∼13%, such as in the *E. fornicatus* species complex and other lineages (Gomez et al., 2018; Cognato et al., 2019; Smith et al., in preparation). Our sequence data shows no evidence of *Wolbachia* or pseudogenes. The uncommon haplodiploid mating system of Xyleborini may provide the best explanation for the high intraspecific COI sequence differences. The diploid female/haploid male sex-ratio is skewed on average 13:1 and ranges from 2:1 to 83:1 (French and Roeper, 1975; Beaver and Browne, 1979; Kirkendall, 1993; Cooperband et al., 2016; Castro et al., 2019). A female has an apparent greater chance of reproducing compared to diploid-diploid species because if unmated she lays a haploid egg which produces a male. She mates with her son to produce diploid daughters. Thereby, a single COI nucleotide mutation can be amplified to population levels in a short time (e.g., 13 female offspring each for 12 generations = $\sim 23 \times 10^{12}$ in a year, assuming all live and reproduce). Similarly high levels of intra- and interspecific COI differences have been observed among inbreeding bark-feeding scolytines with female skewed sex ratios (Kambestad et al., 2017). In comparison, the CAD intraspecific nucleotide differences were less for most pairwise intraspecific and interspecific comparisons at $< 2\%$ and as most as 10%, respectively (**Figure 3**). It is unknown if these sequence differences are unexpectedly

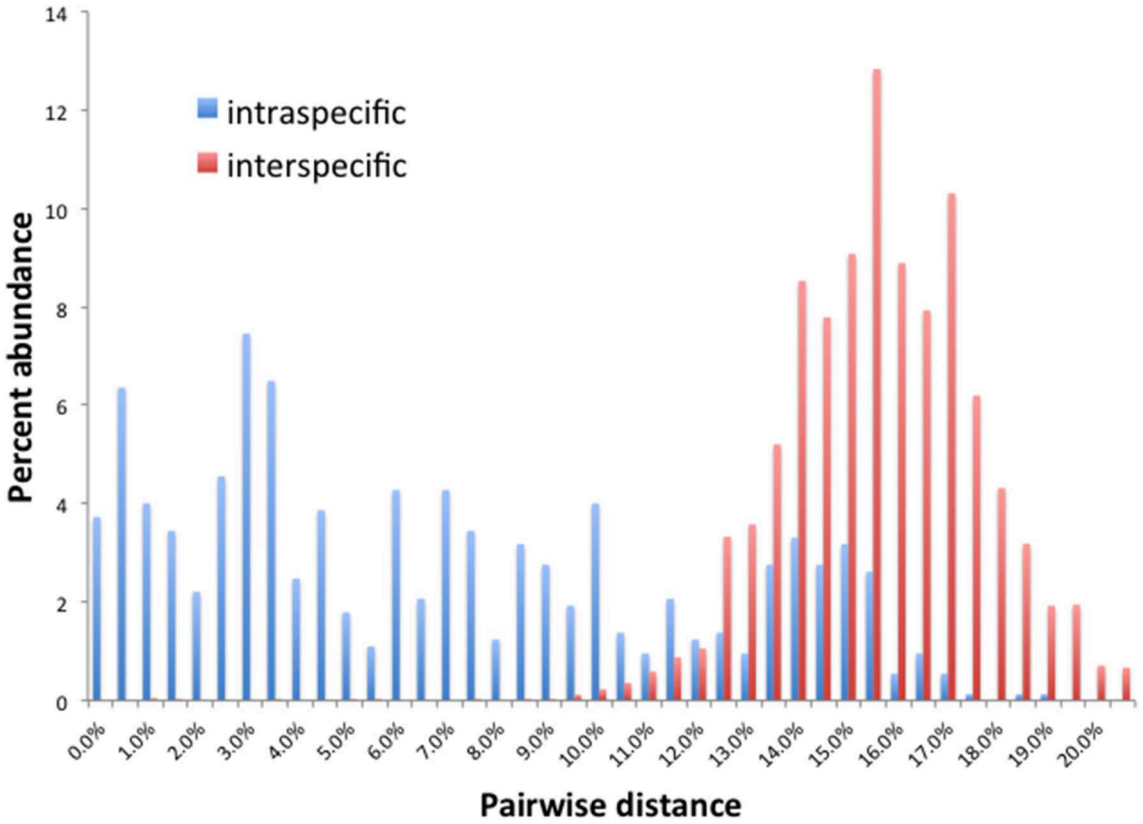


FIGURE 2 | Frequency of uncorrected "*p*" intra- and interspecific COI distances.

high like the COI differences because a comparable dataset of pairwise intraspecific values does not exist for diploid-diploid scolytine groups. However, these values may be as expected for single copy nuclear genes given that xyleborines may experience uncommon interfamilial matings which could maintain a minimal amount of gene flow within a species (Storer et al., 2017).

No one method clearly identified or delimited species. A barcode gap was not evident between intra- and interspecific COI and CAD sequences differences no matter the method used. While TaxonDNA analyses found DNA cluster thresholds (9% for COI and 3% for CAD) near or within the observed overlap of intra- and interspecific differences (**Figures 2**, **3**), ABGD split most species into multiple groups. The tree-based NJ analysis performed better where monophyly and an approximate percentage DNA sequence difference helped to recognize species boundaries. Even better was the iterative approach highlighted for two genera where monophyly was rigorously tested in a parsimony framework and association between the clades and diagnostic morphological characters were examined by taxonomic experts (**Figures 4**, **5**).

These authoritative DNA databases of >400 sequences of COI and CAD are stable foundations for the improved systematics of SE Asian xyleborine ambrosia beetles. However, they currently have limitations in the identification and delimitation of species as is the case for most other DNA identification databases (e.g., Ekrem et al., 2007). Correct determinations are limited to the included 161 of 316 known SE Asian species. Identifications will improve with time as the database grows with the addition of the missing known species. However, the addition of undescribed species is also expected as under-collected regions are sampled. Approximately 30% of the SE Asian xyleborine fauna remains undiscovered, so far (Smith, Beaver, Cognato, pers. observation). In addition, this study exposes taxonomic issues concerning polyphyly of some species and monophyletic species with variable morphology (see discussion below). Both situations suggest that further data is needed to test species limits. Delimitation and description of new and problematic species will be necessary in order to continue the accuracy of this identification database.

This study highlights three taxonomic scenarios that are expected as this database grows. (1) *Ambrosiophilus osumiensis* exemplifies the scenario of a monophyletic species with variable morphology (**Figure 4**). Differences in the number, position, and size of tubercles of the elytral declivity have been used to delimit xyleborine species (Hulcr et al., 2007; Wood, 2007). However some species were suspected as geographically variants of the same species (Hulcr and Cognato, 2013) and

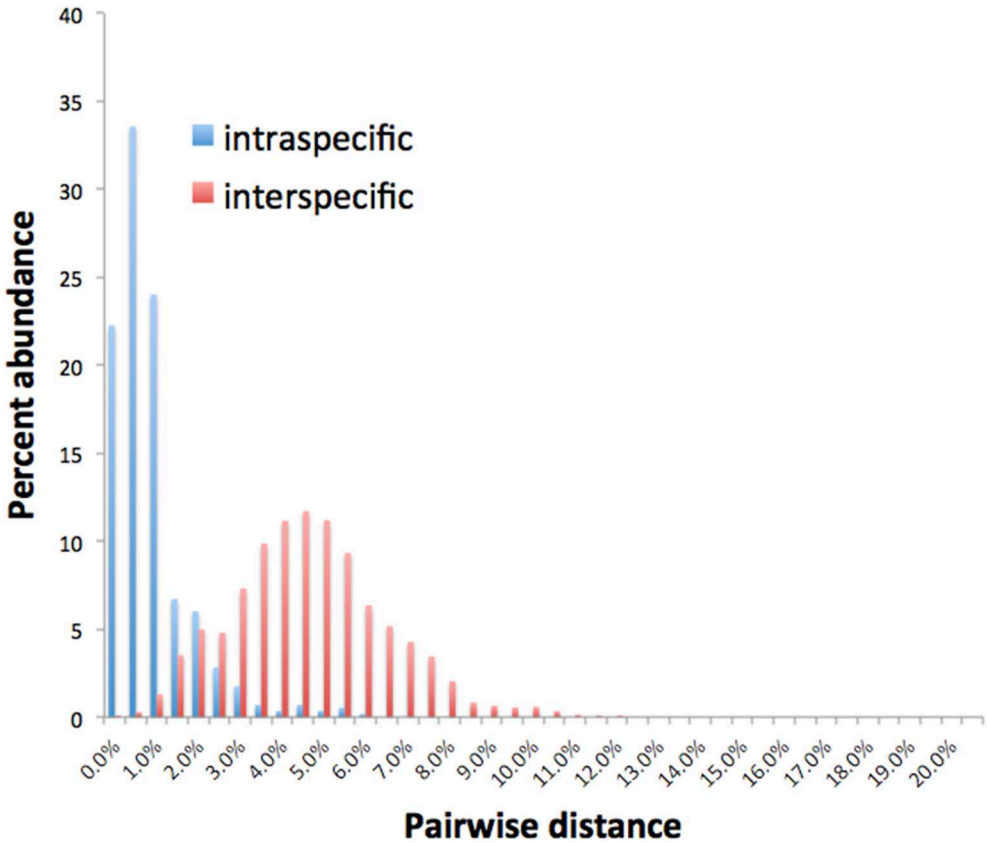


FIGURE 3 | Frequency of uncorrected "*p*" intra- and interspecific CAD distances.

TABLE 3 | DNA clusters based on a pair-wise distance thresholds for interspecific sequence comparisons.

Percent threshold Pairwise distance	#of DNA Profiles	Profiles with Threshold violations	Maximum Pairwise distance	Profiles compatible with Traditional species	Profiles with Only one species	Maximum # species per Profile
COI						
3	354	12 (3.38%)	4.91%	197 (56%)	351	2
6	315	10 (3.17%)	8.33%	203 (64%)	208	2
9	286	7 (2.44%)	10.41%	216 (76%)	277	2
12	114	4 (3.5%)	22.32%	83 (72%)	104	144
15	1	1 (100%)	24.70%	0	0	252
CAD						
1	227	16 (7.04%)	5.21%	157 (69%)	204	13
2	133	16 (12.03%)	13.29%	92 (70%)	109	63
3	62	4 (6.45%)	13.29%	44 (71%)	49	161
4	23	4 (17.39%)	13.46%	12 (52%)	15	206
5	6	2 (33.33%)	15.99%	4 (67%)	4	232

only recently the validity of some suspect species has been investigated in a phylogenetic context (e.g., Cognato et al., 2015; Gomez et al., 2018). The morphological variation illustrated for *Ambrosiophilus osumiensis* (**Figure 4**) was previously presumed diagnostic for three species (*Ambrosiophilus metanepolulus*, *Ambrosiophilus nodulosus*, *Ambrosiophilus osumiensis*) but given that only a maximum 7.5% COI and 2.7% CAD difference occurs among individuals, they are now considered one species (Smith et al., unpublished). Potentially these *A. osumiensis* morphotypes could represent valid species. Investigation of pre and/or post mating barriers in a phylogenetic context of a more widely sampled *A. osumiensis* individuals would aid in

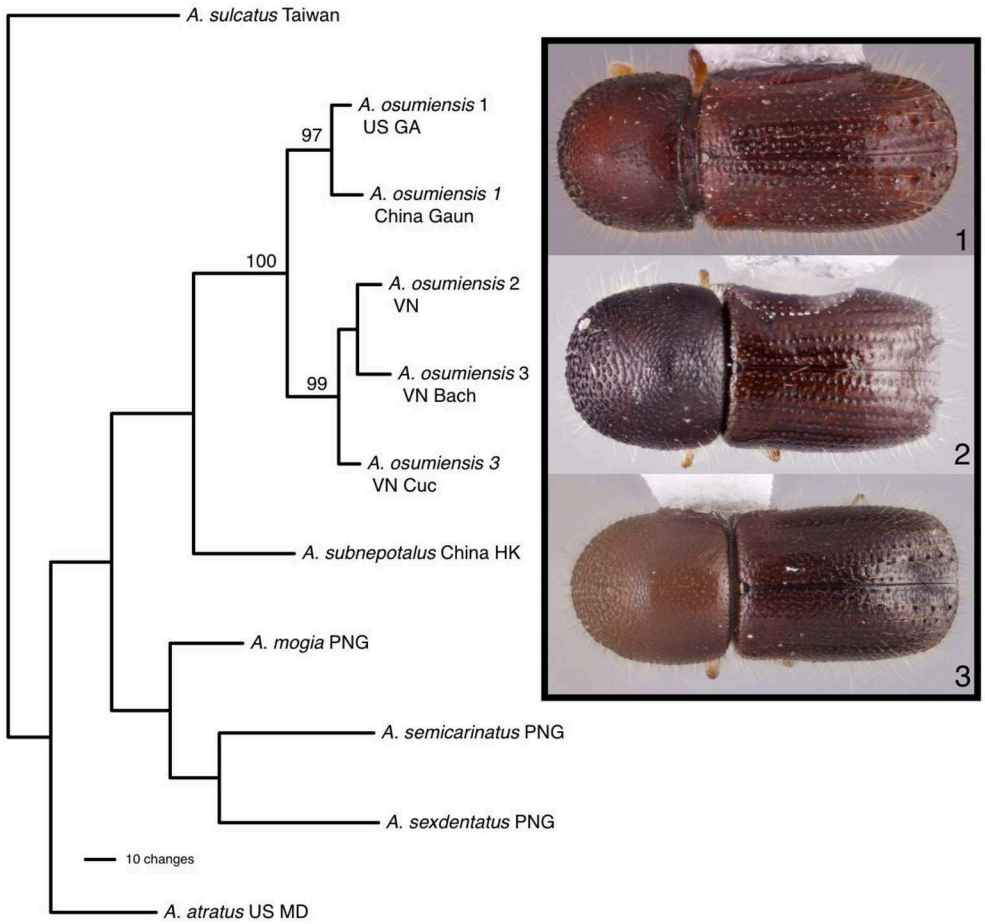


FIGURE 4 | The one most parsimonious tree found for *Ambrosiophilus* species based on COI and CAD data. Numbers indicate bootstrap values. Habitus images depict morphological variation observed among tubercles of the elytra.

discerning the species validity of the *A. osumiensis* morphotypes (as in Cooperband et al., 2015). (2) *Euwallacea* exemplifies a situation where little to no morphological difference occurs among polyphyletic species or monophyletic species in which subclades exhibit >10–12% COI and 2–3% CAD difference (**Figure 5**). The *E. fornicatus* species complex has recently received much taxonomic attention given their pest status and that different lineages impart various levels of economic damage. Although qualitative diagnostic characters were not observed, consistent quantitative characters, and morphometric analysis were congruent with lineages that demonstrated >10% COI difference compared to each other (Stouthamer et al., 2017; Gomez et al., 2018; Smith et al., 2019). In addition, potential pre- and post-mating reproductive barriers and fidelity with different symbiotic fungal strains support the validity of the recognized species (Kasson et al., 2013; Cooperband et al., 2015, 2017). Cryptic species may riddle *Euwallacea* given the > 12% COI difference observed in species like *E. interjectus* and polyphyly of others (**Figure 5**). Their species status will remain unknown until detailed morphometric and biological analyses can be conducted. (3) A recently published study on *Xyleborus glabratus* demonstrates an ideal situation where monophyly, molecular difference, and morphological variation coalesce to support the recognition of new species (Cognato et al., 2019). Upon discovery in the field SMS and AIC initially hypothesized that the included specimens were *X. glabratus* but upon inspection in the laboratory species level morphological diagnostic characters of the elytral declivity were noted. These characters associated with monophyletic groups and >14% COI and >1.5% CAD differences (**Figure 6**). Two species were described and much morphological difference within *X. glabratus* was documented. A lineage of *X. glabratus* with 9% COI difference was not described as a species because of the lack of morphological diagnostic characters. This study and others (Cognato and Sun, 2007; Kambestad et al., 2017; Gomez et al., 2018) are examples of the decision process for the recognition of scolytine species in the context of morphology and molecular phylogenies.

Based on the presented DNA databases and the case studies, we recommend the following conservative guidelines for the identification and delimitation of xyleborine taxa. (1) Confident identifications demonstrate <3 and <1%

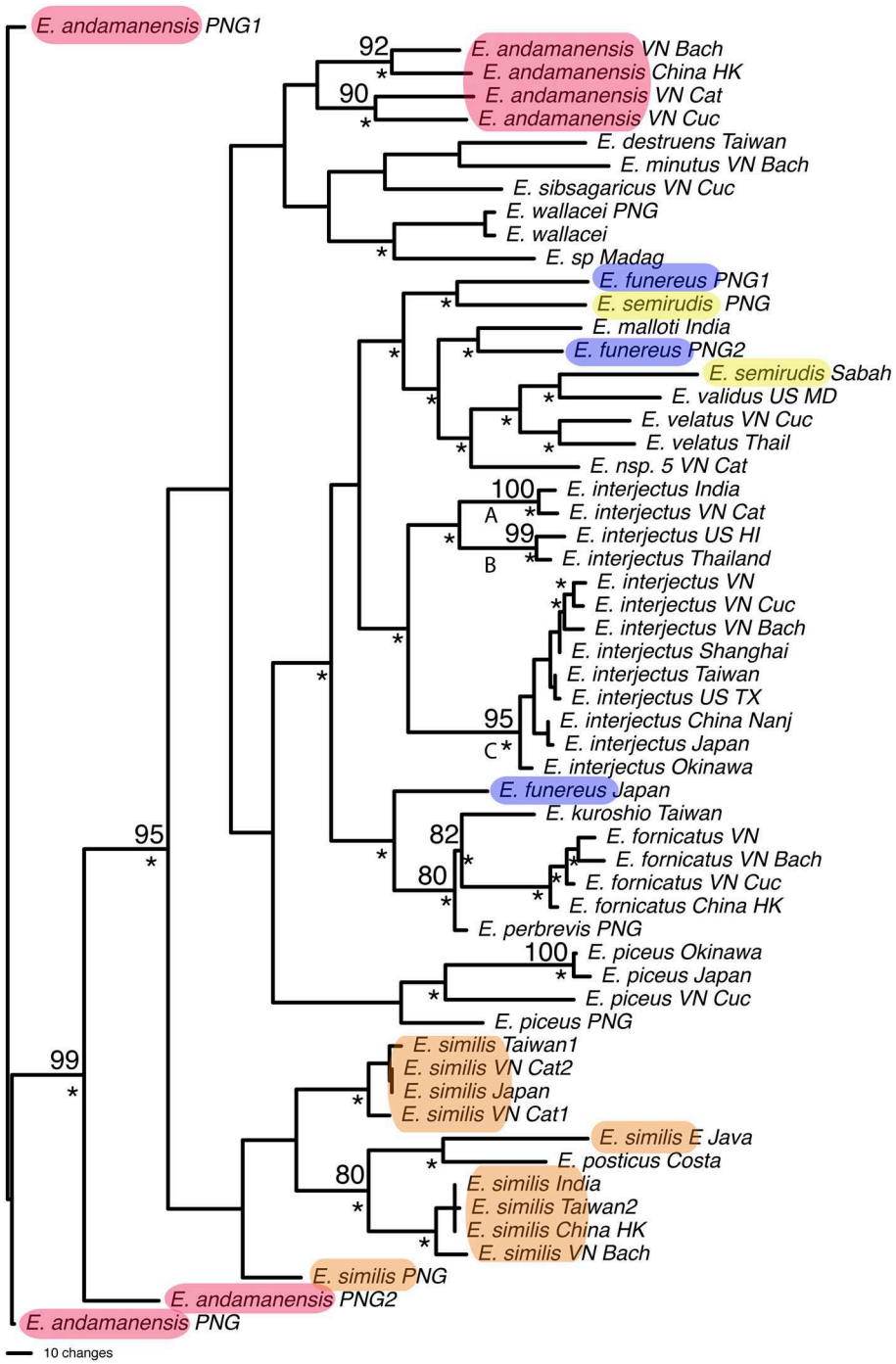


FIGURE 5 | One of 2,475 most parsimonious trees found for *Euwallacea* species based on COI and CAD data. Numbers indicate bootstrap values. *Clade found in the strict consensus of most parsimonious trees. Highlighted species are not monophyletic.

pairwise uncorrected "*p*" COI and/or CAD difference between an unknown and a named barcode DNA sequence. (2) Delimitation of new species becomes more probable when pairwise uncorrected "*p*" COI and/or CAD differences increase beyond 10–12 and 2–3%, respectively. These values are most useful for the naïve diagnostician or when specimens lack additional morphological diagnostic characters (such as, larvae). Indeed, there are cases were species can be identified with confidence when pairwise difference exceeds these pairwise percentages, for example, *X. glabratus* (Cognato et al., 2019) or when species fall near (or below) expected interspecific pairwise percentages. These cases

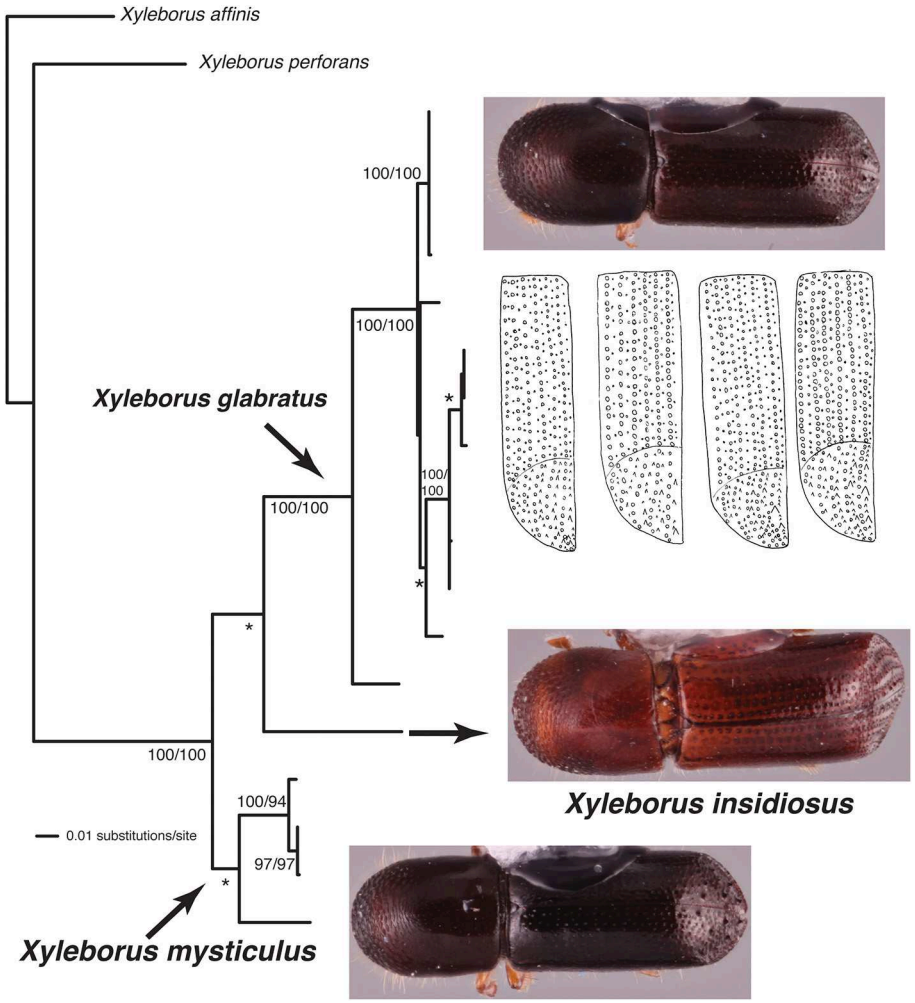


FIGURE 6 | Phylogeny of *Xyleborus glabratus* redrawn from Cognato et al. (2019). Habitus images depict intra- and interspecific morphological variation observed among granules of the elytra. Numbers at nodes indicate parsimony/likelihood bootstrap values and *bootstrap values < 70%.

will usually be evident with a sample size that includes a representative genotypic variation for the species. When in doubt, a taxonomic expert should review these cases using systematic methodology.

The taxonomic experts for this study (SMS, RAB, and AIC) have ∼75 years of combined experience in the identification and delimitation of scolytine species using both morphological and phylogenetic inference. Their initial morphologically-based (null) species hypotheses (i.e., identifications) were informed by this experience, the study of type specimens, and original species descriptions of all SE Asian species. Yet for several species, for example *A. osumiensis*, they reassessed the morphology-based identifications based on the COI/CAD phylogeny. In some cases this resulted in a broader morphological-based species concept and in other cases, the delimitation of new species (as in, Cognato et al., 2019). This iterative process [similar to reciprocal illumination (Hennig, 1966)] treats species as hypotheses of evolutionary lineages, which are tested with morphological, phylogenetic, and/or molecular evidence (Hey, 2006; Yeates et al., 2011). At this stage most of the species included in this study have diagnostic morphological characters, are monophyletic, and/or demonstrate >10 and >2% sequence difference for COI and CAD. The inclusion of more specimens and DNA sequence of different genes in subsequent phylogenetic studies will test species limits and likely improve the delimitation of xyleborine species especially for the highlighted problematic species (e.g., Cognato et al., 2020).

Involvement of taxonomic experts during the process of creating a DNA database for species identification is critical for a solid taxonomic foundation. Without their initial identification, followed by tests of and deliberation of species boundaries, the database would be incomplete and misleading; that is, DNA barcodes identified only to higher taxa or that are misidentified to species. For example, in the BoLD public database ∼10% of the ∼7000 Scolytinae specimens are not identified to species

and ∼6% are only identified to subfamily (http://v4.boldsystems.org, accessed 5 September 2019). These values are relatively good given that less than half of the sequences in Genbank (including BoLD data) are named to species (Page, 2016). The accuracy of species identifications in BoLD is difficult to assess because either vouchers are not imaged or the image quality does not allow for species identification. Also the specimen identifiers are not indicated and if the identifier is named then their taxonomic experience is unrecorded. The citations of the authoritative reference(s) used to make species identifications are mostly lacking. Although the BoLD system allows for the revision of identifications, the above missing information hampers peer-review of species names associated with DNA barcodes. Peer- review of taxonomic identifications is critical to the scientific process inherent in species identification. For example, relying on only a 2–3% percent sequence divergence standard for estimating species diversity, Ashfaq and Hebert (2016) suggested an unexpectedly high estimate of cryptic arthropod pest species. This estimate ignored the accuracy of the species determinations, limited sample size of COI haplotypes, and the biology of the pest. In one case, *Xylosandrus crassiusculus*, our data clearly shows that it is a highly variable (i.e., COI haplotypes) monophyletic species and not three potential cryptic species (Ashfaq and Hebert, 2016). Taking these steps to improve species identification and verification of species in current global databases will improve accuracy of the DNA barcodes (Wu et al., 2017) and applications to biodiversity assessment or the testing of ecological hypotheses (e.g., Caesar et al., 2006; Cognato and Caesar, 2006; Miller et al., 2016).

The initial DNA barcoding movement predicted an end to traditional taxonomy (Hebert et al., 2003a; Sperling, 2003; Smith, 2005; Will et al., 2005; Brower, 2006) and along with a call for DNA taxonomy, the taxonomist's role in these enterprises was uncertain (Tautz et al., 2003; Blaxter, 2004). In 16 years, DNA barcoding publications have proliferated and millions of DNA barcodes have been generated (Taylor and Harris, 2012; DeSalle and Goldstein, 2019). Despite this overwhelming zeal for barcoding, taxonomists remained relevant and advocates of DNA barcoding have welcomed more interaction with taxonomists (e.g., Miller, 2007; Packer et al., 2009a; Miller et al., 2016; Zahiri et al., 2017). For example, DNA barcoding funding helped stop a decline in traditional taxonomy in Canada but productivity had not returned to pre-decline levels of 1980 (Packer et al., 2009a). As already acknowledged, thousands of taxonomists are needed to describe newly collected morphological distinct species as well as species discovered as the result of DNA barcoding (Wheeler et al., 2012). Although taxonomists' involvement in DNA barcoding studies is essential for a reliable identification system and improved understanding of biodiversity, the monetary support future taxonomists is uncertain. For example, the recent $180 million global investment in DNA barcoding aims to discover two million new species; however, the number of traditional taxonomists employed to help with this endeavor is not apparent (BioScan, https://ibol.org/, accessed 16 September 2019; Pennisi, 2019). One would hope that as with past funding of DNA barcoding, this initiative will have a positive impact on training taxonomists and taxonomic publications (Packer et al., 2009a). If funding has not been allocated for taxonomists, then $180 million will only result in a backlog of "DNA barcode species" that will need further study and potentially formal description (Pinheiro et al., 2019).

Creation of a DNA database for species identification is not trivial. It relies on authoritatively identified specimens for use in the generation of DNA barcodes. Misidentified specimens result in a misleading DNA identification tool. For this reason, taxonomists should be part of barcoding ventures from beginning to end so to establish null hypotheses of species boundaries and to interpret non-monophyletic species and/or lineages with unexpected high sequence differences deemed as "DNA barcode species." The taxonomist could then quickly address these "DNA barcode species" by comparison of morphology or inclusion in a rigorously reconstructed multi-gene phylogeny so to test the "DNA barcode species" and to describe validated species. This study exemplifies this approach. Through an iterative process we tested our initial morphologically based species identifications with DNA barcodes (sequences from COI and CAD in this case) and then re-examined our identifications with additional specimens, morphological characters, and additional genes. Some "DNA barcode species" were validated and some were synonymized with known species. We will not contribute to the taxonomic impediment because this DNA barcode project occurred within the context of a traditional taxonomic review of the SE Asian xyleborine fauna and descriptions of new species will soon be published (Smith et al., in preparation). We believe that DNA barcodes are best delivered as an outcome of taxonomic reviews, revisions, or monographs. Indeed one could approach the discovery and description of new species with the DNA barcodes first followed by morphological and phylogenetic study (Puillandre et al., 2012; Kekkonen and Hebert, 2014; Miller et al., 2016; DeSalle and Goldstein, 2019), especially in cases where a taxonomic expert does not exist for the higher taxon. But it could take years for an expert to test the validity of the "DNA barcode species" if she is not vested in the initial project (Fontaine et al., 2012). Thus, it is prudent to include the taxonomic expert throughout a DNA barcoding project because (1) the resulting DNA barcodes will be tied to authoritatively identified species which increases the scientific value of future biodiversity research, (2) new species will be described faster (e.g., $<$ 4 years for species discovered in this study), and (3) other taxonomic tools and information may be produced (e.g., illustrated morphological keys and distribution maps). If a taxonomist for a particular taxon does not exist, then the barcoding project should take the opportunity to train an expert for the orphaned taxon through the employment of existing taxonomists as mentors of the new generation (as in, Rodman and Cody, 2003). By adopting a modern systematic approach, one that analyses all available data in phylogenetic context so to improve taxonomy (Will et al., 2005; Yeates et al., 2011; DeSalle and Goldstein, 2019), the barcoding initiative could make a more meaningful impact on our understanding of biodiversity.

DATA AVAILABILITY STATEMENT

The data generated for this study were deposited in GenBank (**Supplementary Table 1**) and NEXUS files can be found at www.canr.msu.edu/hisl/.

AUTHOR CONTRIBUTIONS

AC conceived the study and wrote the initial manuscript. AC, SMS, YL, JH, HK, C-SL, TP, SS, and WS collected or provided access to specimens. SMS, RB, and AC identified specimens. GS and AC generated and analyzed sequence data. AC, JH, BJ, and SMS revised drafts of the manuscript. AC and JH funded various aspects of this research. All authors read, commented on, and approved the final version of the manuscript.

FUNDING

This research was supported by USDA-APHIS Cooperative Agreement Award (16-8130-0666-CA), NSF (DEB-1256663), NSF-PEET (DEB-0328920), USDA Forest Service Early Detection Rapid Response program cooperative agreement (11-DG-11420004-257) and the Ernst Mayr Travel Grant in Animal Systematics (Harvard University) to AC. In addition grants from the JSPS KAKENHI #17H03831 to HK, USDA-APHIS Farm Bill section 10007, the USDA Forest Service, the National Science Foundation and the Florida Forest Service to JH supported various aspects of this study.

ACKNOWLEDGMENTS

We thank the curators of the following institutions who allowed access to or the loan of type specimens from the collections in their care: Dirk Ahrens (ZFMK), Max Barclay (NHML), Lutz Behne (SDEI), Johannes Bergsten (SMNH), James Boone (BPBM), Michel Brancucci and Isabelle Zürcher-Pfander (NHMB), Lourdes Chamorro (NMNH), Chris Grinter (CASC), Jiri Hájek (MNHP), Matthias Hartmann (NKME), Martin Husemann and Thure Dalsgaard (UHZM), Pol Limbourg (IRSNB), Katrina Menard (OMNH), Otto Merkl (HNHM), Hélène Perrin (MNHN), Roberto Poggi and Maria Tavano (MCG), Harald Schillhammer (NHMW), Kyle Schnepp (FSCA), Wichai Srisuka (QSBG), David Szymroszczyk (MIZ), Joachim Willers (MFNB), Hiraku Yoshitake (NIAES).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fevo.2020.00027/full#supplementary-material

Supplementary Figure 1 | Neighbor-joining tree based on pairwise uncorrected "*p*" distances for COI for 490 Xyleborini specimens.

Supplementary Figure 2 | Neighbor-joining tree based on pairwise uncorrected "*p*" distances for CAD for 429 Xyleborini specimens.

Supplementary Table 1 | Xyleborini specimens included in this study.

REFERENCES

Adamski, D., and Miller, S. E. (2015). Two new yellow-banded sister species of *Syntomaula* Meyrick (Lepidoptera: Gelechioidea: Cosmopterigidae) from Papua New Guinea associated with Rubiaceae. *J. Lepidopter. Soc.* 69, 307–316. doi: 10.18473/lepi.69i4.a6

Ashfaq, M., and Hebert, P. D. N. (2016). DNA barcodes for bio-surveillance: regulated and economically important arthropod plant pests. *Genome* 59, 933–945. doi: 10.1139/gen-2016-0024

Barrett, R. D. H., and Hebert, P. D. N. (2005). Identifying spiders through DNA barcodes. *Can. J. Zool.* 83, 481–491. doi: 10.1139/z05-024

Beaver, R. A., and Browne, F. G. (1979). The scolytidae and platypodidae (Coleoptera) of Penang, Malaysia. *Orient. Insects* 12, 575–624. doi: 10.1080/00305316.1978.10432538

Blaxter, M. L. (2004). The promise of DNA taxonomy. *Philos. Trans. R. Soc. B.* 359, 669–679. doi: 10.1098/rstb.2003.1447

Boland, J. M. (2016). The impact of an invasive ambrosia beetle on the riparian habitats of the Tijuana River Valley, California. *Peer J.* 4:e2141. doi: 10.7717/peerj.2141

Brockerhoff, E. G., and Liebhold, A. M. (2017). Ecology of forest insect invasions. *Biol. Invasions.* 19, 1–19. doi: 10.1007/s10530-017-1514-1

Brower, A. V. Z. (2006). Problems with DNA barcodes for species delimitation: 'Ten species' of *Astraptes fulgerator* reassessed (Lepidoptera: Hesperiidae). *System. Biodivers.* 4, 127–132 doi: 10.1017/S147720000500191X

Brower, A. V. Z. (2010). Alleviating the taxonomic impediment of DNA barcoding and setting a bad precedent: names for ten species of '*Astraptes fulgerator*' (Lepidoptera: Hesperiidae: Eudaminae) with DNA-based diagnoses. *System. Biodivers.* 8, 485–491. doi: 10.1080/14772000.2010.534512

Browne, F. G. (1961). The biology of Malayan Scolytidae and Platypodidae. *Malayan Forest Records* 22, 1–255.

Caesar, R. M., Sörensson, M., and Cognato, A. I. (2006). Integrating DNA data and traditional taxonomy to streamline biodiversity assessment: an example from edaphic beetles in the Klamath ecoregion, California, USA. *Diversity Distribut.* 12, 483–489. doi: 10.1111/j.1366-9516.2006.00237.x

Carrillo, D., Cruz, L., Kendra, P., Narvaez, T., Montgomery, W., Monterroso, A., et al. (2016). Distribution, pest status and fungal associates of *Euwallacea* nr. *fornicatus* in Florida avocado groves. *Insects* 7:55. doi: 10.3390/insects7040055

Castro, J., Smith, S. M., Cognato, A. I., Lanfranco, D., Martinez, M., and Guachambala, M. (2019). Life cycle and development of *Coptoborus ochromactonus* Smith and Cognato (Coleoptera: Curculionidae: Scolytinae). *J. Econ. Entomol.* 112, 729–735. doi: 10.1093/jee/toy403

Chang, H., Liu, Q., Hao, D., Liu, Y., An, Y., and Yang, X. (2014). DNA barcodes and molecular diagnostics for distinguishing introduced *Xyleborus* (Coleoptera: Scolytinae) species in China. *Mitochondr. DNA* 25, 63–69. doi: 10.3109/19401736.2013.779260

Cognato, A. I. (2006). Standard percent DNA sequence difference does not predict species boundaries. *J. Econ. Entomol.* 99, 1037–1045. doi: 10.1093/jee/99.4.1037

Cognato, A. I., and Caesar, R. M. (2006). Will DNA barcoding advance efforts to conserve biodiversity more efficiently than traditional taxonomic methods?: Introduction. *Front. Ecol. Environ.* 4, 268–269. doi: 10.1890/1540-9295(2006)004[0268:WDBAET]2.0.CO;2

Cognato, A. I., Hoebeke, E. R., Kajimura, H., and Smith, S. M. (2015). History of the exotic ambrosia beetles *Euwallacea interjectus* and *Euwallacea validus* (Coleoptera: Curculionidae: Xyleborini) in the United States. *J. Econ. Entomol.* 108, 1129–1135. doi: 10.1093/jee/tov073

Cognato, A. I., Hulcr, J., Dole, S. A., and Jordal, B. H. (2011). Phylogeny of haplo–diploid, fungus-growing ambrosia beetles (Curculionidae: Scolytinae: Xyleborini) inferred from molecular and morphological data. *Zool. Scr.* 4, 174–186. doi: 10.1111/j.1463-6409.2010.00466.x

Cognato, A. I., Jordal, B. H., and Rubinoff, D. (2018). Ancient "Wanderlust" leads to diversification of endemic hawaiian *Xyleborus* species (Coleoptera: Curculionidae: Scolytinae). *Insect Syst. Diver.* 2:1. doi: 10.1093/isd/ixy005

Cognato, A. I., Smith, S.M., and Beaver, R. A. (2020). Two new genera of Oriental xyleborine ambrosia beetles (Coleoptera, Curculionidae: Scolytinae). *Zootaxa* 4722, 540–554. doi: 10.11646/zootaxa.4722.6.2

Cognato, A. I., Smith, S. M., Li, Y., Pham, T. H., and Hulcr, J. (2019). Genetic variability among native *Xyleborus glabratus* Eichhoff populations native to Southeast Asia and the description of two related species. *J. Econ. Entomol.* 112, 1274–1284. doi: 10.1093/jee/toz026

Cognato, A. I., and Sun, J. H. (2007). DNA based cladograms augment the discovery of a new Ips species from China (Coleoptera: Curculionidae: Scolytinae). *Cladistics* 23, 539–551. doi: 10.1111/j.1096-0031.2007.00159.x

Cooperband, M., Cossé, A., Stouthamer, R., Carrillo, D., and Jones, T. (2017). Attractants of ambrosia beetles in the *Euwallacea fornicatus* species complex. Pheromones and other semiochemicals in integrated production. *IOBC-WPRS Bull.* 126, 89–92. doi: 10.7287/peerj.preprints.3175

Cooperband, M. F., Stouthamer, R., Carillo, D., Eskalen, A., Thibault, T., Cossé, A. A., et al. (2016). Biology of two members of the *Euwallacea fornicatus* species complex (Coleoptera: Curculionidae: Scolytinae), recently invasive in the U.S.A., reared on an ambrosia beetle artificial diet. *Agric. For. Entomol.* 18, 223–237. doi: 10.1111/afe.12155

Cooperband, M. F., Stouthamer, R., Carrillo, D., and Cossé, A. (2015). "A crossing study to evaluate invasive *Euwallacea* near *fornicatus* populations in California and Florida," in *26th USDA Interagency Research Forum on Invasive Species,* Vol. 26, K. A. McManus and K. W. Gottschalk (Washington, DC: US Forest Service), 62.

Danforth, B. N., Fang, J., and Sipes, S. D. (2006). Analysis of family level relationships in bees (Hymenoptera: Apiformes) using 28S and two previously unexplored nuclear genes: CAD and RNA polymerase II. *Mol. Phylogenet. Evol.* 39, 358–372. doi: 10.1016/j.ympev.2005.09.022

DeSalle, R., Egan, M. G., and Siddall, M. (2005). The unholy trinity: taxonomy, species delimitation and DNA barcoding. *Philos. Trans. R. Soc. B.* 360, 1905–1916. doi: 10.1098/rstb.2005.1722

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

Dodelin, B. (2018). *Espèce Invasive Nouveau Pour la Faune de France Scolyte Cyclorhipidion fukiense Installé en Europe*. Available online at: https://entomodata.wordpress.com/2018/04/24/cyclorhipidion-fukiense-installe-en-europe/

Dole, S. A., Jordal, B. H., and Cognato, A. I. (2010). Polyphyly of *Xylosandrus* reitter inferred from nuclear and mitochondrial genes (Coleoptera: Curculionidae: Scolytinae). *Mol. Phylogenet. Evol.* 54, 773–782. doi: 10.1016/j.ympev.2009.11.011

Ebach, M. C., and Holdrege, C. (2005). More taxonomy, not DNA barcoding. *Bioscience* 55, 822–823. doi: 10.1641/0006-3568(2005)055[0823:MTNDB]2.0.CO;2

Ekrem, T., Willassen, E., and Stur, E. (2007). A comprehensive DNA sequence library is essential for identification with DNA barcodes. *Mol. Phylogenet. Evol.* 43, 530–542. doi: 10.1016/j.ympev.2006.11.021

Eskalen, A., Gonzalez, A., Wang, D. H., Twizeyimana, M., Mayorquin, J. S., and Lynch, S. C. (2012). First report of a *Fusarium* sp. and its vector tea shot hole borer (Euwallacea fornicatus) causing Fusarium dieback on avocado in California. *Plant Dis.* 96:1070. doi: 10.1094/PDIS-03-12-0276-PDN

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

Folmer, O., Black, M., Hoeh, W., Lutz, R., and Vrijenhoek, R. (1994). DNA primers for amplification of mitochondrial cytochrome c oxidase subunit I from diverse metazoan invertebrates. *Mol. Mar. Biol. Biotechnol.* 3, 294–299.

Fontaine, B., Perrard, A., and Bouchet, P. (2012). 21 years of shelf life between discovery and description of new species. *Curr. Biol.* 22:R944. doi: 10.1016/j.cub.2012.10.029

Foster, B. T., Cognato, A. I., and Gold, R. E. (2004). DNA-based identification of the eastern subterranean termite, *Reticulitermes flavipes* (Isoptera: Rhinotermitidae). *J. Econ. Entomol.* 97, 95–101. doi: 10.1093/jee/97.1.95

Foster, P. G., Bergo, E. S., Bourke, B. P., Oliveira, T. M. P., Nagaki, S. S., Sant'Ana, D. C., et al. (2013). Phylogenetic analysis and DNA-based species confirmation in *Anopheles* (Nyssorhynchus). *PLoS ONE* 8:e54063. doi: 10.1371/journal.pone.0054063

French, J. R., and Roeper, R. A. (1975). Studies on the biology of the ambrosia beetle *Xyleborus dispar* (F) (Coleoptera: Scolytidae). *Zeitschr. Angew. Entomol.* 78, 241–247. doi: 10.1111/j.1439-0418.1975.tb04178.x

Funk, D. J., and Omland, K. E. (2003). Species-level paraphyly and polyphyly: Frequency, causes, and consequences, with insights from animal mitochondrial DNA. *Annu. Rev. Ecol. Evol. Syst.* 34, 397–423. doi: 10.1146/annurev.ecolsys.34.011802.132421

Gibbs, J. (2018). DNA barcoding a nightmare taxon: assessing barcode index numbers and barcode gaps for sweat bees. *Genome* 61, 21–31. doi: 10.1139/gen-2017-0096

Gohli, J., Kirkendall, L. R., Smith, S. M., Cognato, A. I., Hulcr, J., and Jordal, B. H. (2017). Biological factors contributing to bark and ambrosia beetle species diversification. *Evolution* 71, 1258–1272. doi: 10.1111/evo.13219

Gohli, J., Selvarajah, T., Kirkendall, L. R., and Jordal, B. H. (2016). Globally distributed *Xyleborus* species reveal recurrent intercontinental dispersal in a landscape of ancient worldwide distributions. *BMC Evol. Biol.* 16:37. doi: 10.1186/s12862-016-0610-7

Gomez, D. F., Skelton, J., Steininger, M. S., Stouthamer, R., Rugman-Jones, P., Sittichaya, W., et al. (2018). Species delineation within the *Euwallacea fornicatus* complex revealed by morphometric and phylogenetic analyses. *Insect Syst. Diver.* 2, 1–11. doi: 10.1093/isd/ixy018

Haack, R. A., and Rabaglia, R. J. (2013). "Exotic bark and ambrosia beetles in the USA: potential and current invaders" in *Potential Invasive Pests of Agricultural Crop Species*, ed J. Pena (Wallingford: CAB International), 48–74. doi: 10.1079/9781845938291.0048

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

Hebert, P. D. N., Penton, E. H., Burns, J. M., Janzen, D. H., and Hallwachs, W. (2004). Ten species in one: DNA barcoding reveals cryptic species in the neotropical skipper butterfly *Astraptes fulgerator*. *Proc. Natl. Acad. Sci. U.S.A.* 101, 14812–14817. doi: 10.1073/pnas.0406166101

Hebert, P. D. N., Ratnasingham, S., and deWaard, J. R. (2003b). Barcoding animal life: cytochrome c oxidase subunit 1 divergences among closely related species. *Proc. R. Soc. London Ser. B.* 270, S96–S99. doi: 10.1098/rsbl.2003.0025

Hennig, W. (1966). *Phylogenetic Systematics*. Translated by D. D. Davis and R. Zangerl. Urbana: University of Illinois Press.

Hey, J. (2006). On the failure of modern species concepts. *Trends Ecol. Evol.* 21, 447–450. doi: 10.1016/j.tree.2006.05.011

Hoebeke, E. R., Rabaglia, R. J., KníŽek, M., and Weaver, J. S. (2018). First records of *Cyclorhipidion fukiense* (Eggers) (Coleoptera: Curculionidae: Scolytinae: Xyleborini), an ambrosia beetle native to Asia, in North America. *Zootaxa* 4394, 243–250. doi: 10.11646/zootaxa.4394.2.7

Hughes, M. A., Riggins, J. J., Koch, F. H., Cognato, A. I., Anderson, C., Formby, J. P., et al. (2017). No rest for the laurels: symbiotic invaders causes unprecedented damage to southern USA forests. *Biol. Invasions* 19, 2143–2157. doi: 10.1007/s10530-017-1427-z

Hulcr, J., Atkinson, T. H., Cognato, A. I., Jordal, B. H., and McKenna, D. D. (2015). "Morphology, taxonomy and phylogenetics of bark beetles" in *Bark Beetles: Biology and Ecology of Native and Invasive Species,* eds F. E. Vega and R. W. Hofstetter (San Diego, CA: Elsevier Inc.), 41–84. doi: 10.1016/B978-0-12-417156-5.00002-2

Hulcr, J., Black, A., Prior, K., Chen, C.-Y., and Li, H. F. (2017). Studies of ambrosia beetles (Coleoptera: Curculionidae) in their native ranges help predict invasion Impact. *Fla. Entomol.* 100, 257–261. doi: 10.1653/024.100.0219

Hulcr, J., and Cognato, A. I. (2013). *Xyleborini of New Guinea: A Taxonomic Monograph*. Thomas Say Publications in Entomology. Maryland: Entomological Society of America.

Hulcr, J., Dole, S. A., Beaver, R. A., and Cognato, A. I. (2007). Cladistic review of generic taxonomic characters in Xyleborini (Coleoptera: Curculionidae: Scolytinae). *Syst. Entomol.* 32, 568–584. doi: 10.1111/j.1365-3113.2007.00386.x

Hulcr, J., Skelton, J., Johnson, A. J., Li, Y., and Jusino, M. A. (2018). Invasion of an inconspicuous ambrosia beetle and fungus may alter wood decay in Southeastern North America. *PeerJ. Preprints* 6:e27334v1. doi: 10.7287/peerj.preprints.27334v1

Jordal, B. H., Beaver, R. A., and Kirkendall, L. R. (2001). Breaking taboos in the tropics: incest promotes colonization by wood-boring beetles. *Glob. Ecol. Biogeogr.* 10, 345–357. doi: 10.1046/j.1466-822X.2001.00242.x

Jordal, B. H., and Cognato, A. I. (2012). Molecular phylogeny of bark and ambrosia beetles reveals multiple origins of fungus farming during periods of global warming. *BMC Evol. Biol.* 12:133. doi: 10.1186/1471-2148-12-133

Jordal, B. H., and Kambestad, M. (2014). DNA barcoding of bark and ambrosia beetles reveals excessive NUMTs and consistent east-west divergence across Palearctic forests. *Mol. Ecol. Resour.* 14, 7–17. doi: 10.1111/1755-0998.12150

Jordal, B. J., Sequeira, A., and Cognato, A. I. (2011). The age and phylogeny of wood boring weevils and the origin of subsociality. *Mol. Phylogenet. Evol.* 59, 708–724. doi: 10.1016/j.ympev.2011.03.016

Jukes, T. H., and Cantor, C. R. (1969). "Evolution of protein molecules" in *Mammalian Protein Metabolism,* ed H. N. Munro (New York, NY: Academic Press), 21–132. doi: 10.1016/B978-1-4832-3211-9.50009-7

Kambestad, M., Kirkendall, L. R., Knutsen, I. L., and Jordal, B. H. (2017). Cryptic and pseudo-cryptic diversity in the world's most common bark beetle—*Hypothenemus eruditus*. *Organ. Divers. Evol.* 17, 633–652. doi: 10.1007/s13127-017-0334-6

Kasson, M. T., O'Donnell, K., Rooney, A. P., Sink, S., Ploetz, R. C., Ploetz, J. N., et al. (2013). An inordinate fondness for *Fusarium*: phylogenetic diversity of fusaria cultivated by ambrosia beetles in the genus *Euwallacea* on avocado and other plant hosts. *Fungal Genet. Biol.* 56, 147–157. doi: 10.1016/j.fgb.2013.04.004

Kekkonen, M., and Hebert, P. D. N. (2014). DNA barcode-based delineation of putative species: efficient start for taxonomic workflows. *Mol. Ecol. Resour.* 14, 706–715. doi: 10.1111/1755-0998.12233

Kimura, M. (1980). A simple method for estimating evolutionary rate of base substitutions through comparative studies of nucleotide sequences. *J. Mole. Evol.* 16, 111–120. doi: 10.1007/BF01731581

Kirkendall, L. R. (1993). "Ecology and evolution of biased sex ratios in bark and ambrosia beetles" in *Evolution and Diversity of Sex Ratios in Insects and Mites*, eds. D. L. Wrensch and M. A. Ebbert (New York, NY: Chapman and Hall), 235–345. doi: 10.1007/978-1-4684-1402-8_8

Kirkendall, L. R., and Faccoli, M. (2010). Bark beetles and pinhole borers (Curculionidae, Scolytinae, Platypodinae) alien to Europe. *Zookeys* 56, 227–251. doi: 10.3897/zookeys.56.529

Knížek, M. (2011). "Scolytinae" in *Catalogue of Palaearctic Coleoptera, Vol. 7. Curculionoidea I*, eds I. Löbl and A. Smetana (Stenstrup, Apollo Books), 204–251.

Lait, L. A., and Hebert, P. D. N. (2018). Phylogeographic structure in three North American tent caterpillar species (Lepidoptera: Lasiocampidae): *Malacosoma americana*, *M. californica, and M. disstria. PeerJ*. 6:e4479. doi: 10.7717/peerj.4479

Magnacca, K. N., and Brown, M. J. F. (2010). Mitochondrial heteroplasmy and DNA barcoding in Hawaiian *Hylaeus* (*Nesoprosopis*) bees (Hymenoptera: Colletidae). *BMC Evol. Biol.* 10:174. doi: 10.1186/1471-2148-10-174

Meier, R., Shiyang, K., Vaidya, G., and Ng, P. K. L. (2006). DNA barcoding and taxonomy in Diptera: a tale of high intraspecific variability and low identification success. *Syst. Biol.* 55, 715–728. doi: 10.1080/10635150600969864

Meurisse, N., Rassati, D., Hurley, B. P., Brockerhoff, E. G., and Haack, R. A. (2018). Common pathways by which non-native forest insects move internationally and domestically. *J. Pest Sci.* (2004) 92, 13–27. doi: 10.1007/s10340-018-0990-0

Miller, S. E. (2007). DNA barcoding and the renaissance of taxonomy. *Proc. Natl Acad. Sci. U.S.A.* 104, 4775–4776. doi: 10.1073/pnas.0700466104

Miller, S. E., Hausmann, A., Hallwachs, W., and Janzen, D. H. (2016). Advancing taxonomy and bioinventories with DNA barcodes. *Phil. Trans. R. Soc. B.* 371:20150339. doi: 10.1098/rstb.2015.0339

Moulton, M. J., Song, H., and Whiting, M. F. (2010). Assessing the effects of primer specificity on eliminating numt coamplification in DNA barcoding: a case study from Orthoptera (Arthropoda: Insecta). *Mol. Ecol. Resour.* 10, 615–627. doi: 10.1111/j.1755-0998.2009.02823.x

Nanney, D. L. (1982). Genes and phenes in tetrahymena. *Bioscience* 32, 783–788. doi: 10.2307/1308971

Packer, L., Gibbs, J., Sheffield, C. S., and Hanner, R. (2009b). DNA barcoding and the mediocrity of morphology. *Mol. Ecol. Resour.* 9, 42–50. doi: 10.1111/j.1755-0998.2009.02631.x

Packer, L., Grixti, J. C., Roughley, R. E., and Hanner, R. (2009a). The status of taxonomy in Canada and the impact of DNA barcoding. *Can. J. Zool.* 87, 1097–1110. doi: 10.1139/Z09-100

Page, R. D. M. (2016). DNA barcoding and taxonomy: dark taxa and dark texts. *Phil. Trans. R. Soc. B.* 371:20150334. doi: 10.1098/rstb.2015.0334

Pennisi, E. (2019). $180 million DNA 'barcode' project aims to discover 2 million new species. *Science* doi: 10.1126/science.aay2877. [Epub ahead of print].

Pinheiro, H. T., Moreau, C. S., Daly, M., and Rocha, L. A. (2019). Will DNA barcoding meet taxonomic needs? *Science* 365, 873–874. doi: 10.1126/science.aay7174

Prendini, L. (2005). Comment on "Identifying spiders through DNA barcodes". *Can. J. Zool.* 83, 498–504. doi: 10.1139/z05-025

Puillandre, N., Lambert, A., Brouillet, S., and Achaz, G. (2012). ABGD, automatic Barcode Gap discovery for primary species delimitation. *Mol. Ecol.* 21, 1864–1877. doi: 10.1111/j.1365-294X.2011.05239.x

Rabaglia, R. J., Cognato, A. I., Hoebeke, E. R., Johnson, C. W., LaBonte, J. R., Carter, M. E., et al. (2019). Early detection and rapid response a 10-year summary of the USDA Forest Service program of surveillance for non-native bark and ambrosia beetles. *Am. Entomol.* 65, 29–42. doi: 10.1093/ae/tmz015

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

Rodman, J. E., and Cody, J. H. (2003). The taxonomic impediment overcome: NSF's Partnerships for Enhancing Expertise in Taxonomy (PEET) as a model. *Syst. Biol.* 52, 428–435. doi: 10.1080/10635150309326

Rubinoff, D., Cameron, S., and Will, K. (2006). A genomic perspective on the shortcomings of mitochondrial DNA for "barcoding" identification. *J. Hered.* 97, 581–594. doi: 10.1093/jhered/esl036

Saitou, N., and Nei, M. (1987). The neighbor-joining method a new method for reconstructing phylogenetic trees. *Mole. Biol. Evol.* 4, 406–425.

Schindel, D. E., and Miller, S. E. (2005). DNA barcoding a useful tool for taxonomists. *Nature* 435:17. doi: 10.1038/435017b

Smith, S. M., and Cognato, A. I. (2014). A taxonomic monograph of Nearctic *Scolytus* Geoffroy (Coleoptera, Curculionidae, Scolytinae). *Zookeys* 450, 1–182. doi: 10.3897/zookeys.450.7452

Smith, S. M., and Cognato, A. I. (2015). Ambrosiophilus peregrinus n. sp., an exotic ambrosia beetle discovered in Georgia, U.S.A. (Coleoptera: Curculionidae: Scolytinae). *Coleopter. Bull.* 69, 213–220. doi: 10.1649/0010-065X-69.2.213

Smith, S. M., Gomez, D. F., Beaver, R. A., Hulcr, J., and Cognato, A. I. (2019). Reassessment of the species in the *Euwallacea fornicatus* (Coleoptera: Curculionidae: Scolytinae) complex after the rediscovery of the 'lost' type specimen. *Insects* 10:261. doi: 10.3390/insects10090261

Smith, S. M., Petrov, A. V., and Cognato, A. I. (2017). Beetles (Coleoptera) of Peru: a survey of the Families. Curculionidae: Scolytinae. *Coleopter. Bull.* 71, 77–94. doi: 10.1649/0010-065X-71.1.77

Smith, V. S. (2005). DNA Barcoding: Perspectives from a "Partnerships for Enhancing Expertise in Taxonomy" (PEET) Debate. *Syst. Biol.* 54, 841–844. doi: 10.1080/10635150500354894

Song, H., Buhay, J. E., Whiting, M. F., and Crandall, K. A. (2008). Many species in one: DNA barcoding overestimates the number of species when nuclear mitochondrial pseudogenes are coamplified. *Proc. Natl. Acad. Sci. U.S.A.* 105, 13486–13491. doi: 10.1073/pnas.0803076105

Sperling, F. A. (2003). DNA barcoding: deus ex machina. *Newslett. Biol. Survey Canada* (Terrestrial Arthropods) 22, 50–53.

Sperling, F. A., and Hickey, D. A. (1994). Mitochondrial DNA sequence variation in the spruce budworm species complex (Choristoneura: Lepidoptera). *Mol. Biol. Evol.* 11, 656–665.

Srivathsan, A., and Meier, R. (2012). On the inappropriate use of Kimura-2-Parameter (K2P) divergences in the DNA – barcoding literature. *Cladistics* 28, 190–194. doi: 10.1111/j.1096-0031.2011.00370.x

Storer, C., Peyton, A., McDaniel, S., Jordal, B., and Hulcr, J. (2017). Cryptic genetic variation in an inbreeding and cosmopolitan pest, *Xylosandrus crassiusculus*, revealed using ddRADseq. *Ecol. Evol.* 7, 10974–10986. doi: 10.1002/ece3.3625

Stouthamer, R., Rugman-Jones, P., Thu, P. Q., Eskalen, A., Thibault, T., Hulcr, J., et al. (2017). Tracing the origin of a cryptic invader: phylogeography of the *Euwallacea fornicatus* (Coleoptera: Curculionidae: Scolytinae) species complex. *Agr. Forest Entomol.* 19, 366–375. doi: 10.1111/afe.12215

Swofford, D. L. (2002). *PAUP*: Phylogenetic Analysis Using Parsimony (*and Other Methods), 4.0b10*. Sunderland, MA: Sinauer Associates.

Taft, W. H., and Cognato, A. I. (2017). Recognition of a new species of *Carmenta* from New Mexico supported by morphology and mitochondrial cytochrome oxidase I data (Lepidoptera: Sesiidae: Sesiinae: Synanthedonini). *Zootaxa* 4337, 436–444. doi: 10.11646/zootaxa.4337.3.8

Tautz, D., Arctander, P., Minelli, A., Thomas, R. H., and Vogler, A. P. (2003). A plea for DNA taxonomy. *Trends Ecol. Evol.* 18, 70–74. doi: 10.1016/S0169-5347(02)00041-1

Taylor, H. R., and Harris, W. E. (2012). An emergent science on the brink of irrelevance: a review of the past 8 years of DNA barcoding. *Mol. Ecol. Resour.* 12, 377–388. doi: 10.1111/j.1755-0998.2012.03119.x

Trewick, S. A. (2008). DNA Barcoding is not enough: mismatch of taxonomy and genealogy in New Zealand grasshoppers (Orthoptera: Acrididae). *Cladistics* 24, 240–254. doi: 10.1111/j.1096-0031.2007.00174.x

Ward, R. D., Zemlak, T. S., Innes, B. H., Last, P. R., and Hebert, P. D. N. (2005). DNA barcoding Australia's fish species. *Phil. Trans. R. Soc. B.* 360, 1847–1857. doi: 10.1098/rstb.2005.1716

Wheeler, Q. D., Knapp, S., Stevenson, D. W., Stevenson, J., Blum, S. D., Boom, B. M., et al. (2012). Mapping the biosphere: exploring species to understand the origin, organization and sustainability of biodiversity. *Syst. Biodivers.* 10, 1–20. doi: 10.1080/14772000.2012.665095

Will, K. W., Mishler, B. D., and Wheeler, Q. D. (2005). The perils of DNA barcoding and the need for integrative taxonomy. *Syst. Biol.* 54, 844–851. doi: 10.1080/10635150500354878

Will, K. W., and Rubinoff, D. (2004). Myth of the molecule: DNA barcodes for species cannot replace morphology for identification and classification. *Cladistics* 20, 47–55. doi: 10.1111/j.1096-0031.2003.00008.x

Wood, S. L. (1982). The bark and ambrosia beetles of North and Central America (Coleoptera: Scolytidae), a taxonomic monograph. *Great Basin Naturalist Memoirs* 8, 1–1359.

Wood, S. L. (2007). *Bark and Ambrosia Beetles of South America (Coleoptera, Scolytidae)*. M. L. Bean Life Science Museum. Provo, UT: Brigham Young University.

Wood, S. L., and Bright, D. E. (1992). A catalog of Scolytidae and Platypodidae (Coleoptera), part 2: taxonomic index. *Great Basin Naturalist Memoirs* 13, 1–1533.

Wu, Y., Trepanowski, N. F., Molongoski, J. J., Reagel, P. F., Lingafelter, S. W., Nadel, H., Myers, S. W., et al. (2017). Identification of wood-boring beetles (Cerambycidae and Buprestidae) intercepted in trade-associated solid wood packaging material using DNA barcoding and morphology. *Sci. Rep.* 7:40316. doi: 10.1038/srep40316

Yeates, D. H., Seago, A., Nelson, L., Cameron, S. L., Joseph, L., and Trueman, J. W. H. (2011). Integrative taxonomy, or iterative taxonomy? *Syst. Entomol.* 36, 209–217. doi: 10.1111/j.1365-3113.2010.00558.x

Zahiri, R., Lafontaine, J. D., Schmidt, B. C., deWaard, J. R., Zakharov, E. V., and Hebert, P. D. N. (2017). Probing planetary biodiversity with DNA barcodes: the Noctuoidea of North America. *PLoS ONE* 12:e0178548. doi: 10.1371/journal.pone.0178548

Disclaimer: The authors are solely responsible for the writing of this paper.

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.

frontiers in Ecology and Evolution

ORIGINAL RESEARCH
published: 28 February 2020
doi: 10.3389/fevo.2020.00044

The Cycad Genus *Cycas* May Have Diversified From Indochina and Occupied Its Current Ranges Through Vicariance and Dispersal Events

Ledile T. Mankga[1], Kowiyou Yessoufou[2], Thendo Mugwena[2] and Munyaradzi Chitakira[3]*

[1] Department of Life and Consumer Sciences, University of South Africa, Pretoria, South Africa, [2] Department of Geography, Environmental Management and Energy Studies, University of Johannesburg, Johannesburg, South Africa, [3] Department of Environmental Sciences, University of South Africa, Pretoria, South Africa

Biogeographically, cycads were once widely distributed but the extant cycads are restricted to tropical and subtropical regions. They originated ∼ 300 Ma and re-diversified recently around 12 Ma, with the genus *Cycas* being the most rapidly diversified and largely distributed lineage. However, the forces that shaped the diversification and biogeography of the genus remain to be fully understood. Here, we first retrieved and used DNA sequences from GenBank (nuclear: PHYP, RPB1, HZP, AC3, F3H, SAMS, and GTP; chloroplasts: plant barcode *trn*H-*psb*A, *trn*L-*trn*F, *trn*S-*trn*G, and *psb*M-*trn*D) to assemble a complete dated phylogeny of *Cycas*. Then, we employed the Bayesian Binary Method to reconstruct the historical biogeography of the extant *Cycas* and finally, using the Bayesian approach for diversification analysis, we explored the evolutionary events that might have shaped the rapid diversification and large distribution of *Cycas* across the Pacific Islands. Our analysis pointed to Indo-China as the origin of the genus, which may have dispersed firstly across the Pacific Islands during the late Miocene aided by multiple excursions of sea levels and the development of a key innovation, i.e., a spongy endocarp particularly in the seeds of the subsection Rumphiae. The colonization of South China, which was thought to be the origin of the genus, may have occurred more recently aided by both dispersal and vicariance events. However, no significant shifts in the evolutionary events (speciation, extinction, mass extinction) that shaped the diversity of the genus were observed. Overall, our study re-clarifies the historical biogeography and the evolutionary forces that shaped the current diversity of the genus *Cycas*.

Keywords: cycads, DNA barcode, evolutionary diversification, historical biogeography, late Miocene, sea-level excursions

OPEN ACCESS

Edited by:
Rodney L. Honeycutt, Pepperdine University, United States

Reviewed by:
Michael Wink, Heidelberg University, Germany
Fabien L. Condamine, University of Gothenburg, Sweden

****Correspondence:***
Kowiyou Yessoufou
kowiyouy@uj.ac.za

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

Received: *04 November 2019*
Accepted: *11 February 2020*
Published: *28 February 2020*

Citation:
Mankga LT, Yessoufou K, Mugwena T and Chitakira M (2020) The Cycad Genus Cycas May Have Diversified From Indochina and Occupied Its Current Ranges Through Vicariance and Dispersal Events. Front. Ecol. Evol. 8:44. doi: 10.3389/fevo.2020.00044

INTRODUCTION

Cycads are dioecious and entomophilous plants that developed palm-like habit with stout trunks and large evergreen pinnate leaves (Jones, 2002). They share some characteristics with the ferns (e.g., spermatozoa with flagella) and angiosperms (e.g., seed productions; Guan, 1996; Norstog and Nicholls, 1997). The dispersal of cycad seeds is limited to 2–7 km, and is mostly mediated

by rodents, small fruit-eating bats (Yang and Meerow, 1996) and long dispersal via the sea (Keppel et al., 2009). Cycads represent the oldest lineage of plants, originating ∼ 300 million years ago (Ma) in the mid-Permian (Hendricks, 1987; Gao and Thomas, 1989; Calonje et al., 2017) and reaching their greatest diversity in the Jurassic era (Jones, 2002; Nagalingum et al., 2011). Geographically, cycads are restricted to tropical and subtropical or warm temperate regions with predominantly summer rainfalls (Jones, 2002). In total, 10 genera diversified within the cycad group, with the genus *Cycas* being the largest of all (Osborne et al., 2012; Calonje et al., 2017).

Specifically, *Cycas* is the only genus in the family Cycadaceae, an early diverging lineage to the cycad phylogenetic tree (Stevenson, 1992; Nagalingum et al., 2011). This genus is comprised of six Sections, including *Asiorientales, Panzhihuaenses, Wadeanae, Strongyloides, Indosinenses,* and *Cycas* (Hill, 2004). The genus *Cycas* is the most rapidly diversified clade in the cycad group with ∼ 112 species (Yessoufou et al., 2017). Fossil evidence points to Asia as the origin of the genus (Hill, 1995; see also Xiao and Möller, 2015). From Asia, the genus *Cycas* is further distributed southward to Australia, eastern Africa and the Pacific Islands (Hill, 2004).

In Asia, the genus is distributed across the Red River Fault between South China and the Indochina block, with Red River potentially constituting a geographical barrier for gene flow (Xiao and Möller, 2015). If this barrier was effective, we would expect to detect the signature of vicariance events in the evolutionary history of the genus *Cycas* (Keppel et al., 2008; Xiao and Möller, 2015). Then, the widespread distribution of *Cycas* from Asia to Africa, Australia and across the Pacific regions might have been mediated through long distance dispersal events across the ocean. However, the sample analyzed in a recent study that tested this hypothesis (Xiao and Möller, 2015) was taxonomically limited (only 31 species out of 112), although they included representatives of all six Sections of the genus in their analysis. Even in Keppel et al.'s (2008) study, only the Subsection Rumphiae of the Section *Cycas* was analyzed. As such, inferences on the evolutionary and ecological processes that shaped the biogeography of *Cycas* may require further investigations. In addition, in their recent analysis of the diversification rate comparison across the cycad tree of life, Yessoufou et al. (2017) revealed a diversification rate heterogeneity across the tree with the genus *Cycas* identified as the most rapidly diversifying clade, and they suggested that this rapid diversification might have mediated their widest geographic distribution. Unfortunately, they did not go further to elucidate the patterns of diversification events within this clade.

In the present study, our aim is to provide a refined understanding of the evolutionary and ecological processes that shaped the biogeography of the genus *Cycas*. Specifically, we assembled the most comprehensive dated phylogeny of the genus, which was then used to elucidate its historical biogeography as well as the ecological forces that mediated the observed diversity patterns.

MATERIALS AND METHODS

A Complete List of *Cycas* Species Used to Reconstruct a Dated *Cycas* Phylogeny

The full list of *Cycas* species is still a matter of debate. However, a recent study analyzed a large dataset of informative markers (DNA data) to estimate the total cycad diversity to 116 (100 accepted, 7 subspecies and 9 controversial species; Liu et al., 2018). To assemble a complete phylogeny for the 116 *Cycas* species, we retrieved from GenBank/EBI (accessed October 2018; Liu et al., 2018) DNA sequences of seven nuclear regions (PHYP, RPB1, HZP, AC3, F3H, SAMS, and GTP) and four plastid regions (including a complementary plant DNA barcode *trn*H-*psb*A as well as *psb*M-*trn*D, *trn*L-*trn*F and *trn*S-*trn*G) of *Cycas* species. The molecular matrix is available as **Supplementary Material** (DNA matrix; available at https://doi.org/10.5061/dryad.1g1jwstrn, Mankga et al., 2020b); accession numbers as well as the species names are presented in **Supplementary Table S1**. The dated phylogeny was assembled for 135 species including outgroups (*Bowenia* Hook.ex Hook.f., *Ceratozamia* Brongn, *Dioon* Lindl., *Encephalartos* Lehm., *Lepidozamia* Regel, *Macrozamia* Miq., *Microcycas calocoma* (Miq.) A. DC., *Stangeria eriopus* (Kunze) Baill., *Zamia* L., *Ginkgo biloba* L.) following the traditional Bayesian approach implemented in the BEAST program (Rambaut and Drummond, 2007).

The following steps were followed for the BEAST analysis. Firstly, an XML file using BEAUti (Drummond and Rambaut, 2007) was generated. Secondly, the best model GTR + I + Γ (based on Akaike information criterion evaluated using MODELTEST; Nylander, 2004) was selected as well as the birth-death process prior with uncorrelated relaxed lognormal model for rate variation among branches, following Condamine et al.'s (2015) recommendations. To calibrate the *Cycas* tree, uniform priors with minimum and maximum age estimates for nodes calibration were selected as the normal priors bias the node age estimates (Schenk, 2016). The following uniform calibration points were used following Condamine et al. (2015) for cycad group: *Cycas* SG (15.8–257.2 Ma), Cycads SG (273.9–364.9 Ma), *Dioon* SG (107–207.9 Ma), *Bowenia* SG (88.7–174.3 Ma), *Lepidozamia* SG (33.9–55 Ma), *Ceratozamia* SG (19.2–84.9 Ma), *Zamia* SG (14.6–57 Ma), *Encephalartos* SG (97.7–192.5 Ma). Lastly, MCMC was run for 100 million with trees sampled every 10,000 generations. At the end of the process of dated tree reconstruction, the ESS values ranged from 200 to 901 for the age estimates; the first 2,000 trees were burnt and the remaining 8,000 trees were combined using TREEANNOTATOR (Rambaut and Drummond, 2007) to generate a maximum clade credibility (MCC) tree. The node support on this MCC tree is interpreted as follows: not supported (PP $<$ 0.50), supported (PP = 0.60) and strongly supported (PP $>$ 0.60). In addition, the bootstrap node supports on the phylogeny were assessed using PAUP v40b10 (Swofford, 2002) approach. These node supports were assessed by reconstructing the Maximum Parsimony (MP) tree based on the heuristic search with 1000 random sequences additions keeping 10 trees. The bootstrap values

were interpreted as follows: BS > 70% indicates strong support and BS < 70% indicate weak support (Hillis and Bull, 1993; Wilcox et al., 2002).

Ancestral Area Reconstruction States: Historical Biogeography of *Cycas*

To reconstruct the historical biogeography of the genus *Cycas*, we grouped all species into three categories based on their current geographic distribution (Osborne et al., 2012) and following Xiao and Möller (2015). The category (A) includes species from South China, Taiwan-Ryukyu Archipelago, and Palawan islands (we refer henceforth to category A simply as South China). The category (B) includes species from Indochina, and (C) include Islands of Southeast Asia plus the Malay Peninsula, the Indian subcontinent, East Africa and North Australia.

We used Bayesian Binary Model (BBM) analysis implemented in RASP to reconstruct the possible ancestral ranges of the genus *Cycas* on the phylogenetic trees. In this analysis, the frequencies of an ancestral range at a node in ancestral reconstructions are averaged over all trees generated by RASP in Bayesian analysis (Yan et al., 2010). To account for uncertainties in phylogeny, we used 20,000 trees from MCMC output generated with BBM model. The MCMC chains were run simultaneously for 5,000,000 generations. The state was sampled every 1000 generations. Fixed JC + G (Jukes-Cantor + Gamma) were used with null root distribution and the maximum number of areas for this analysis was kept as 3.

Diversification Analysis

All the diversification analyses were run using R library TESS (Höhna et al., 2015). Firstly, we identified the branching model that fits the diversification of the genus *Cycas* and then compared the number of taxa of the *Cycas* tree to the posterior-predictive distribution of 1000 simulated trees under a constant-rate birth-death model. The constant-rate birth-death model was parameterized by drawing rate parameters from the joint posterior densities inferred from the phylogenetic tree. This parameterized model was used to simulate 1000 phylogenies, which were then used to calculate the expected number of taxa. If the actual number of taxa falls near the center of the posterior-prediction distribution, then the model can be used to simulate the *Cycas* trees, indicating that it provides a good absolute fit and the diversification rates of *Cycas* are constant over time. Conversely, if the summary statistics fell outside the 95% credible interval of the posterior-predictive distributions, then the constant rate birth-death model is not suitable to predict the simulated trees and the diversification has significantly changed over time (Höhna et al., 2015).

In addition, we plotted the posterior-predictive distribution of the lineage accumulation curves (LTT plots for simulated trees) and compared the predictive distribution to the LTT plot of the observed tree. If the observed LTT plot falls within the simulated LTT plots, then the diversification rate of the genus *Cycas* has been constant over time and if not, this means that the diversification has experienced some evolutionary shifts.

Finally, the evolutionary models that explain the diversification patterns depicted by the observed LTT plot were identified. The models tested include a constant-rate birth-death model and three rate-variation models. The rate-variation models include a birth-death model with an exponentially decreasing speciation rate, a birth-death model with piecewise-constant rates (i.e., rates of speciation and extinction change over time but the diversification rate remains constant; Höhna et al., 2015) and a birth-death model of evolution punctuated by a mass-extinction event. Using Bayes Factors (BF; Baele et al., 2013), a pairwise comparison of these models was done to select the best model. For two models M_0 and M_1, BF values were interpreted following Jeffreys (1961). Specifically, $BF(M_0,M_1) < 1$ means the model M_1 is supported; $1 < BF(M_0,M_1) < 3.2$ suggests that M_0 is barely worth-mentioning; $3.2 < BF(M_0,M_1) < 10$ indicates a substantial support for M_0, $10 < BF(M_0,M_1) < 100$ is indicative of a strong support for M_0, and $BF(M_0,M_1) > 100$ is interpreted as decisive support for M_0 (Jeffreys, 1961).

The Compound Poisson Process Mass-Extinction Times (CoMET) Analysis

To investigate whether the genus *Cycas* has experienced some mass extinctions events (if so, when?), the CoMET [Compound Poisson Process (CPP) on Mass Extinction Time)] approach was employed (May et al., 2016). This approach has the advantage to fit not only all possible birth-death models to the data at hand but also to specifically model mass extinction events. The CoMET approach treats the number of speciation-rate shifts, extinction-rate shifts, mass-extinction events as well as the parameters associated with these events as random variables, and then estimates their joint posterior distribution. For this analysis, hyperpriors were set both empirically and *a priori*.

RESULTS

Phylogenetic Tree of *Cycas*

The combined DNA data consisting of seven nuclear genes (PHYP, RPB1, HZP, AC3, F3H, SAMS, and GTP) and four chloroplasts (*trn*H-*psb*A, *trn*L-*trn*F, *trn*S-*trn*G, and *psb*M-*trn*D) includes 10788 characters, 3947 potential parsimony informative sites and 3584 constant characters (**Supplementary Table S2**). The missing data is less than 5% (**Supplementary Table S1**).

The phylogenetic tree reconstructed is, in general, well supported. Among all the nodes whose support values are reported on **Figure 1**, 77% of them have PP $\geq$ 0.80, whereas 59% of these nodes have BP > 70% (**Figure 1**). Further, the ESS values ranged from 200 to 901 for the age estimates, suggesting convergence between posterior distributions and the MCMC estimates. The dated tree suggests that the genus *Cycas* may have diverged around 12 Ma (95% HPD, 10.4 – 14.7; **Figure 1**). Even though the origin of the genus dates back to 12 Ma, most *Cycas* diversification was initiated in the Pleistocene and reached its peak in the Holocene (**Figure 1**).

In addition, of all the six sections of the genus (*Cycas*, *Wadeae, Asiorientales, Stangerioides, Panzhihuaenses* and *Indosinenses*), the section *cycas* is the largest (67 species out of 116 species), is

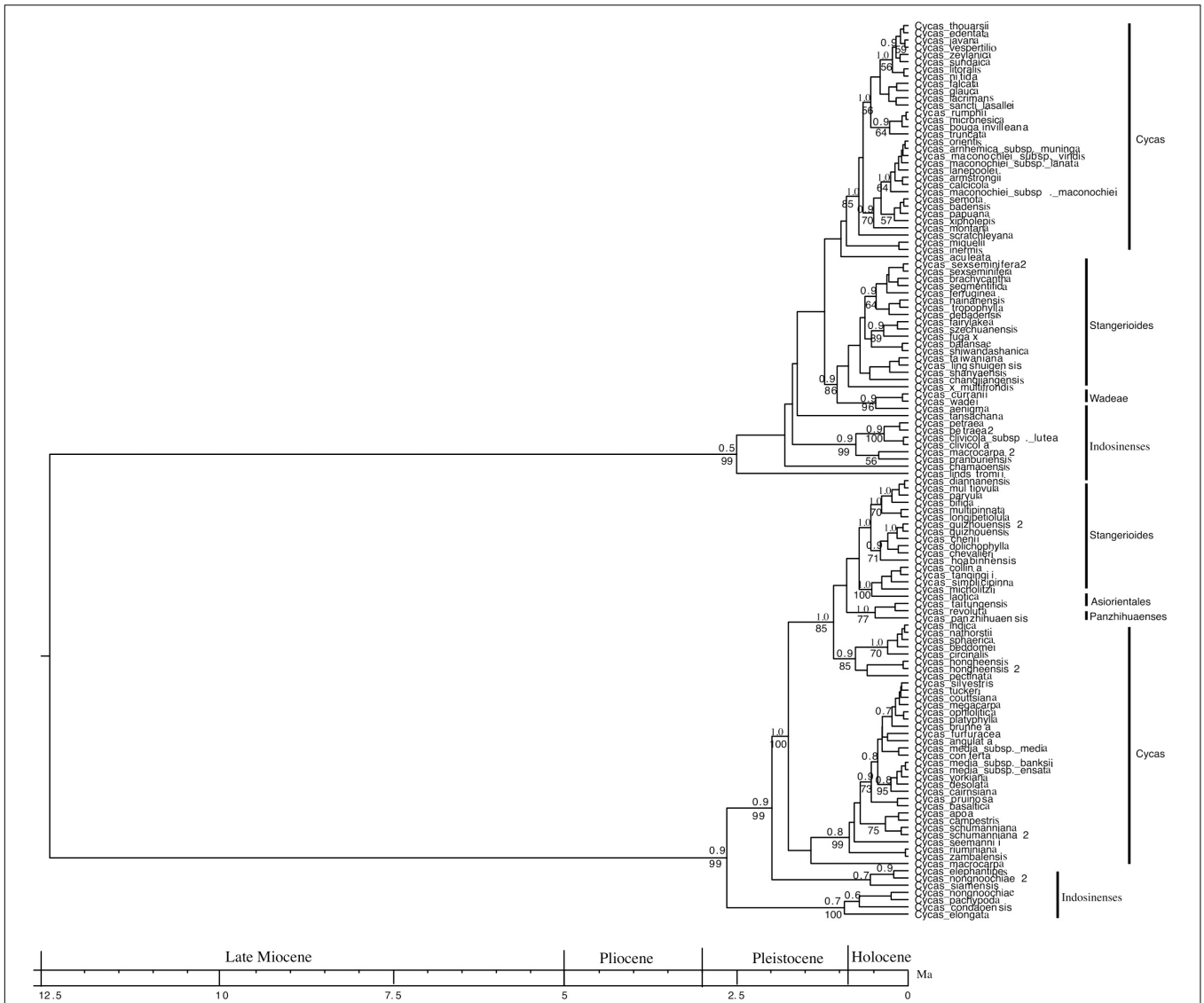


FIGURE 1 | A complete dated phylogenetic tree of the genus *Cycas* from combined seven nuclear genes and three chloroplasts based on Bayesian Inference. The numbers above the branches represent Bayesian Posterior Probability (PP) and numbers below the branches represent the bootstrap value (BP).

polyphyletic and radiated ∼2 Ma (95% HPD, 1.09 – 2.6; **Figure 1**). The sections *Stangerioides* and *Indosinenses* are not monophyletic and most species in these sections radiated ∼1 Ma (95% HPD, 0.61–1.90; 0.17–2.4 and 1.09–3.03, respectively). However, *Panzhihuaenses* and *Asiorientales* sections are monophyletic with a strong support (PP/BP = 1.0/77). Finally, the section *Wadeae*, consisting of two species that are monophyletic with a strong support (PP/BP = 0.9/96), is the most recently radiated section (95% HPD, 0.01–0.37; **Figure 1**). All the six sections form two clearly defined major clades which both diverged in the Pleistocene (**Figure 1**). Each of these two clades is subtended by a long stem branch (phylogenetic fuse) connecting each clade to the origin of the genus. The early diverging clade (age = 2.75 Ma) is made up of the sections *Indosinenses*, *Cycas*, *Panzhihuaenses*, *Asiorientales and Stangerioides*. The sections *Indosinenses*, *Wadeae*, *Stangerioides* and *Cycas* (age = 2.5 Ma) form the second major clade (**Figure 1**).

Historical Biogeography

Our analysis points to Indochina (∼99%) as the origin of the genus *Cycas*, which dated back to around 12 Ma (node I, **Figure 2**). Around 2 Ma, the genus diverged from Indochina to the Islands of Southeast Asia, including the Malay Peninsula, the Indian subcontinent, East Africa and North Australia where the diversification was mostly mediated through vicariance (**Figures 2**, **3**), but the origin is uncertain (node II; probability <50%). Around the same time period, the genus further diversified within Indochina (nodes III, probability 99%), and colonized South China around ∼ 1.5 Ma (node IV, probability 90%) aided by vicariance (**Figures 2**, **3**).

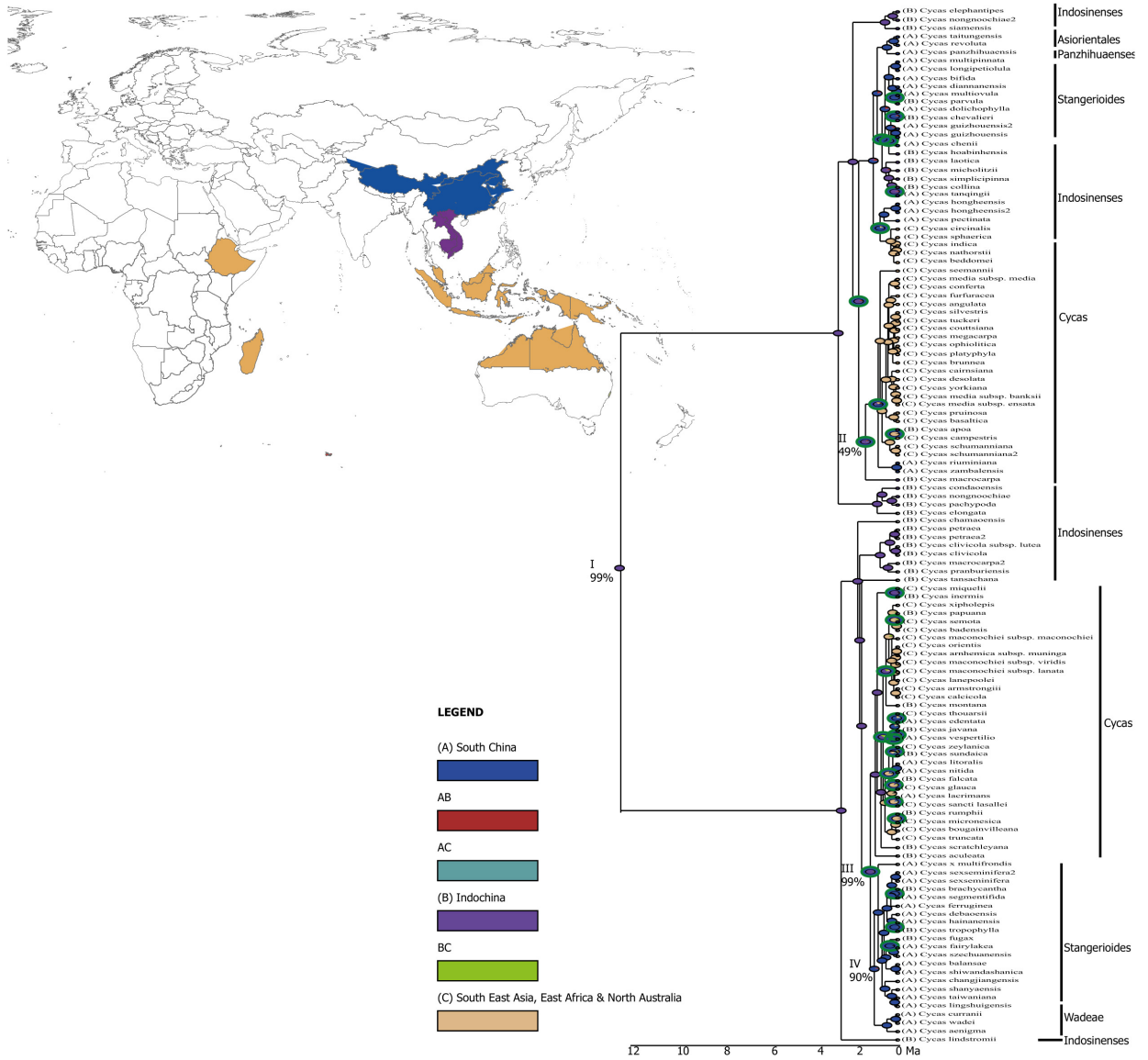


FIGURE 2 | A graphical output from RASP showing the results of ancestral reconstruction area from BBM (Bayesian Binary Method) analysis. Pie charts at each node show probabilities of alternative ancestral ranges. The green circles around the node represent vicariance events and the blue circles represent dispersal events. Key major nodes representing historical origin of different taxonomic sections are noted I-IV and the probability of the origin at these nodes are also indicated (%).

Diversification Analysis

Most of the diversification events occurred in the last two million years (Pleistocene; **Figures 1**, **4A,B**). These diversification events may have followed a constant diversification model as revealed in the following findings. The actual number of taxa (116) falls within the 95% credible interval of its posterior predictive distributions (**Figure 5**; left panel). This means that the constant-rate birth-death model used to reconstruct the posterior predictive distributions provides a good absolute fit to the evolutionary diversification of the genus *Cycas*. In addition, our LTT-plot does not depart significantly from those of the simulated trees under a constant-rate birth-death model (**Figure 5**; right panel). This is an additional support for the constant diversification over time. Finally, when testing alternative models using Bayes Factors to select the best diversification model, we found that a constant birth-death model is once more strongly supported (BF = 72.40; **Supplementary Table S3**).

The COMET Results

We tested several diversification events that might shape the biogeographical patterns. The diversification hyperpriors were specified *a priori* and empirically. Only the results of *a priori* hyperpriors are reported below as these are similar to those of the empirically set priors. The analysis indicates a general trend of increased speciation rate within the window of 2 to

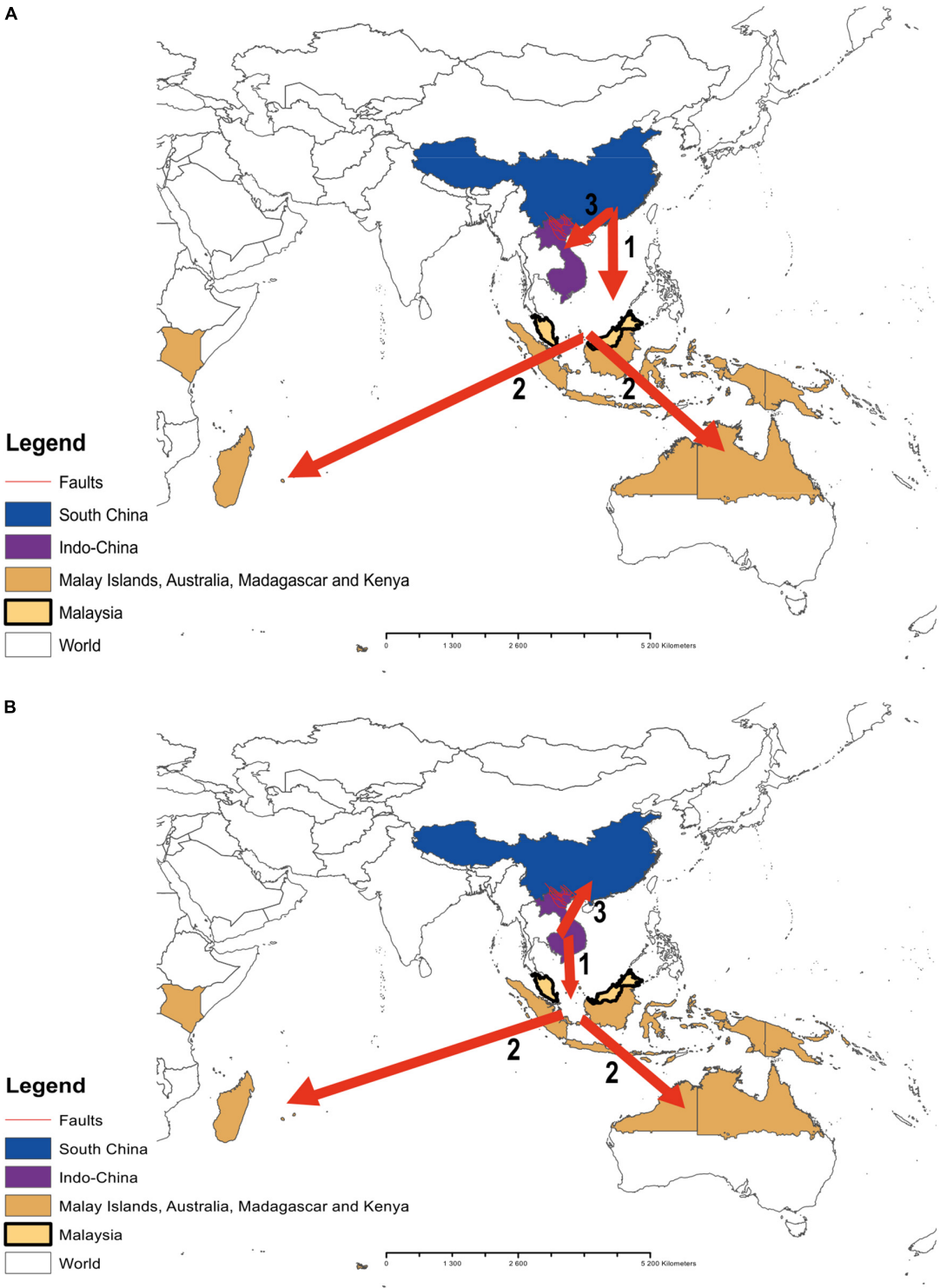


FIGURE 3 | A representation of the potential colonization routes of the genus *Cycas* across the Pacific regions. Numbers 1, 2, and 3 indicate the probable sequences of colonization routes. **(A)** route reconstruction by Xiao and Möller (2015); **(B)** our route reconstruction.

4 species per million years (Myr; **Figure 6A**). These multiple speciation events did not correspond to any significant or dramatic shift (**Figure 6B**). Although the extinction rate remains roughly constant at 4 species Myr^{-1} from 12 to ~4 Ma, it decreases gradually during the last period of diversification (4-0 Ma) to 3 species Myr^{-1} (**Figure 6C**). Again, none of

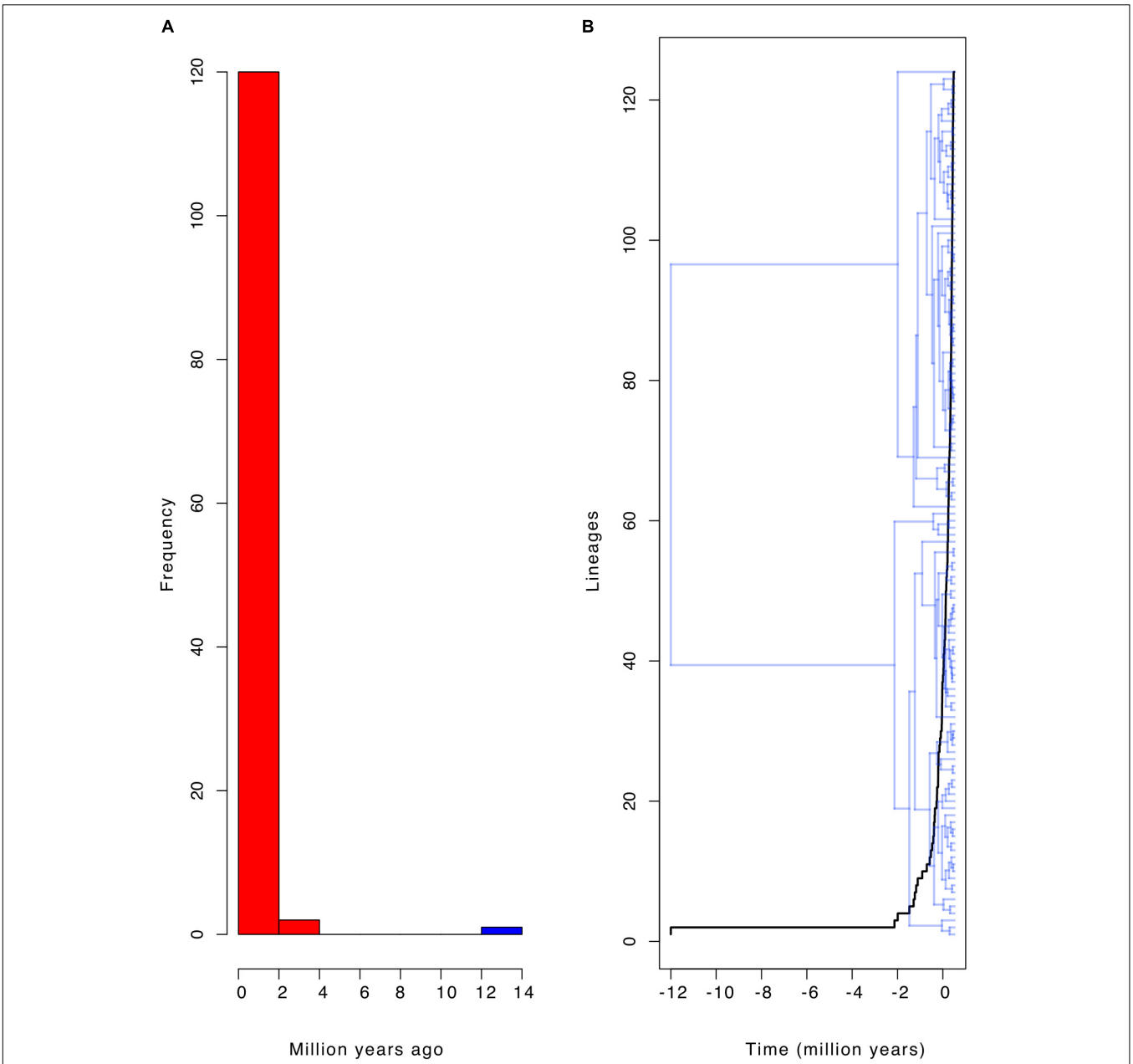


FIGURE 4 | The diversification frequency of the genus *Cycas*. **(A)** Histogram showing the frequency of branching time of the phylogeny of *Cycas*; red color indicates earlier branching events, which shows that most branching events occurred in the last two million years; **(B)** Lineage-through-time plot matching the branching time on the phylogeny of *Cycas*.

these extinction events was significant, and there was no evidence of any significant shift in mass extinction events ($2\ln BF < 6$; **Figures 6D–F**).

DISCUSSION

Phylogenetic Tree of *Cycas*

In comparison with the phylogeny reported in Liu et al. (2018), our phylogeny is similar in terms of the topology and the node support. This is not surprising because we used their DNA sequences. Three of the six Sections of the genus are polyphyletic (*Cycas, Stangerioides, Indosinense*) and the remaining sections are monophyletic as previously reported (Xiao and Möller, 2015; Liu et al., 2018). There are a few points worth highlighting. In our phylogeny, the species *Cycas macrocarpa* and *Cycas pranburiensis* are nested within the section Indosinenses, but they were included in the section *Cycas* in previous studies (Hill and Yang, 1999; Liu et al., 2018). Our finding is likely due to the following reason. The sections *Cycas* and *Indosinenses* have

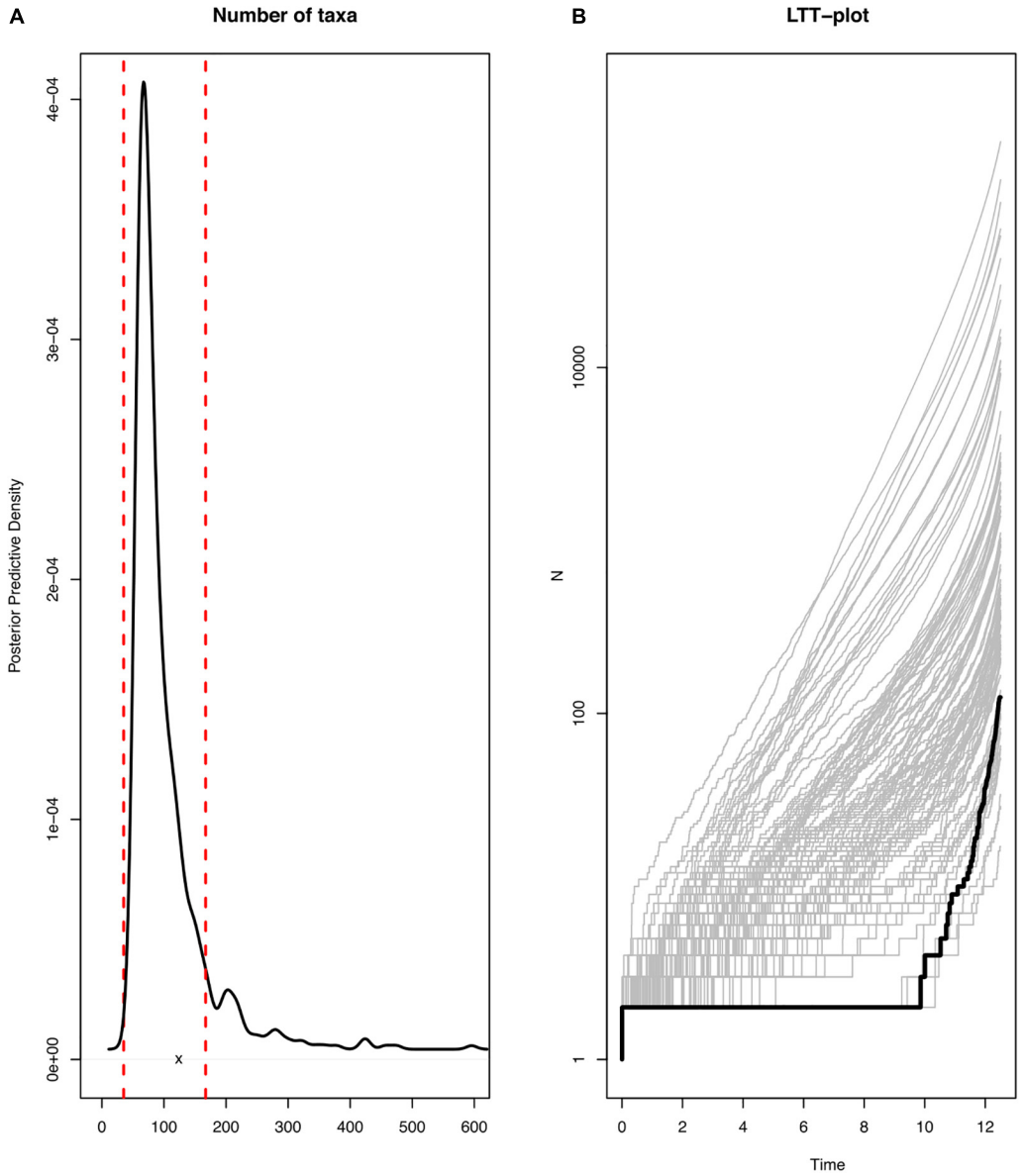


FIGURE 5 | Evidence of constant-rate diversification of the genus *Cycas*. **(A)** actual number of species in the phylogeny (x) in comparison to the simulated number of species (posterior predictive distribution); **(B)** actual Lineage Through Time (LTT) plot in comparison to the simulated LLT plots.

overlapping distribution pattern in Southeast China and India that might have caused a gene flow within the two sections (Yang and Meerow, 1996), making it difficult to distinguish species of these two sections on a phylogeny.

Diversification and Historical Biogeography

The Dispersal-Extinction-Cladogenesis (DEC) model provides alternative option to BBM for historical biogeographic analysis as it takes into account, unlike the BBM, the adjacent configuration of the areas through time (Ree and Smith, 2008; Beeravolu and Condamine, 2016). However, we reported only the results of BBM based on the following reasons. First, Xiao and Möller (2015) conducted a similar study on the same genus but with limited sampling size; in their study they used BBM, and for us to be able to compare our findings with theirs, we used the same BBM model. Our findings are indeed different from theirs. Second, we did run the DEC analysis, but the results point to an uncertain origin for the genus (60% of uncertainty) as opposed to the finding of the BBM (10% of uncertainty). In addition to the differences in sampling size between both studies (ours and that of Xiao and Möller, 2015), it is important to highlight the influence of priors (e.g., Yule vs. birth-death) on age estimates or divergence times. This has been showed in recent studies. For example, Condamine et al. (2015) demonstrated striking

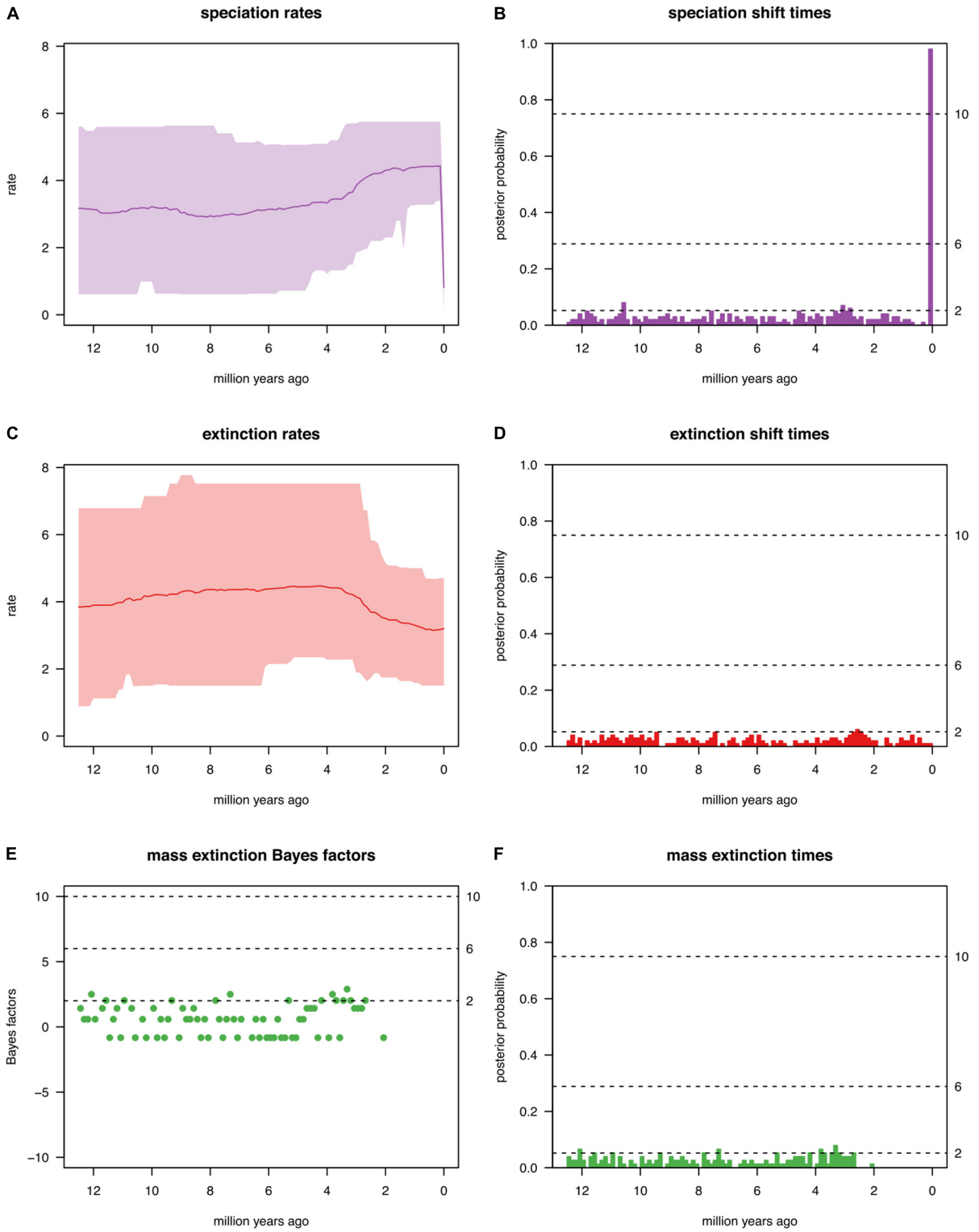


FIGURE 6 | Summary of all evolutionary events reported in this study by fitting the CoMET model. Results reported are for the diversification hyperpriors specified *a priori.* **(A)** Speciation rate; **(B)** Speciation shift times; **(C)** Extinction rates; **(D)** Extinction shift times; **(E)** Mass extinction Bayes factors; **(F)** Mass extinction times.

differences in node age estimates between the Yule versus the birth-death priors employed on the same dataset to assemble a phylogenetic tree of cycads (see also Couvreur et al., 2010). This is an additional but essential justification for the present study to re-investigate the biogeography of the genus *Cycas*.

Indeed, the biogeography of the genus *Cycas* has been investigated in recent studies (e.g., Keppel et al., 2008; Xiao and Möller, 2015). In their study, Xiao and Möller (2015) indicated, with high level of confidence (∼94%), that South China is the origin of the genus. Our analysis, instead, pointed to Indo-China as the origin of the genus, which dating back to ∼ 12 Ma (evolutionary age of the genus). They also indicated that Indo-China was the first geographic region to be colonized by the genus through vicariance and dispersal from South China (with a relatively low level of confidence of 46%) with a series of late dispersal events across the Malay archipelagos through to Australia and East Africa. In our study, the colonization routes are different. Specifically, we found that the colonization route might actually have started first from Indo-China (ancestral area B; **Figure 2**) to the ancestral area (C) (Malay islands southward to Australia and westward to Madagascar, East-Africa) and last from Indo-China (B) to South China (A).

Indeed, the historical biogeography of the Pacific Island's flora has always been a matter of debate (e.g., see Keppel et al., 2009). Our study adds to this debate specifically with regard to the origin and the ecological forces that might have driven the distribution of the genus *Cycas* across the island. The differences between our findings and those of Xiao and Möller (2015) could be linked to the differences in the sampling size between both studies. Although they included representatives of the major sections of the genus into their analysis, only 31 species were analyzed whilst ours includes the complete sample (116 species) of the genus. In addition, our analysis further contradicts theirs in term of the sequences of the colonization events. As opposed to Xiao and Möller (2015), we found that the colonization of South China occurred actually not at an early stage but at the last, after the rest of the genus distribution ranges across the Pacific Islands has been colonized. However, our study agrees with Xiao and Möller (2015) concerning the ecological processes (dispersal and vicariance) that might have mediated the colonization. On this aspect, the Red River Fault may have played an important role, which may include the role of a geographic barrier between Indo-China and South China (Xiao and Möller, 2015; Zheng et al., 2016). This barrier may account for the delay of the colonization of South China in comparison to the early colonization of the Malay archipelagos and the distribution ranges of the genus previously reported (Xiao and Möller, 2015).

In this early colonization of the Malay archipelagos, Malaysia might have played the role of a source area from which the genus might have dispersed westward to East Africa and eastward into the Pacific Islands (centre-periphery hypothesis, Brown, 1984; Hampe and Petit, 2005; Kawecki, 2008; Gaston, 2009). The centre-periphery hypothesis provides an explanation to the biogeographical distribution of species from their centre of origin to their peripheral ranges. The hypothesis predicts that populations are more isolated and less abundant toward the periphery of their distribution (Sexton et al., 2009). Although we did not explicitly test this hypothesis in this study, early studies reported an overall decrease in taxonomic diversity of various plant groups from Malesia eastward in the Pacific region (Corner, 1963; van Balgooy, 1969; Woodroffe, 1987). Even this report holds for the genus *Cycas* as, for example, most *Cycas* species in the subsection Rumphiae are centred in or around Malesia (Hill, 1996a,b; Keppel et al., 2008).

The debate on the colonization process of the Pacific Islands (Keast and Miller, 1996; Ebach and Tangey, 2006) revolves around vicariance and long-distance dispersal events (see Keppel et al., 2009). The vicariance biogeography (Nelson and Platnick, 1981) was initially believed to be the major force structuring the flora of the Pacific regions (Whitmore, 1973; Ladiges et al., 2003; Heads, 2006, 2008; Ladiges and Cantrill, 2007). However, the long distance dispersal process has also been central in the early debate (Darwin, 1859; Guppy, 1906; Ridley, 1930; Mayr, 1954; Carlquist, 1967). Interestingly, mounting evidence, including molecular data, supports the long distance dispersal scenario (Turner et al., 2001; Price and Clague, 2002; Winkworth et al., 2002, 2005; Perrie and Brownsey, 2007). For the genus *Cycas*, the long distance dispersal is more likely the main event through which the entire geographic ranges of *Cycas* has been colonized (Keppel et al., 2008; Xiao and Möller, 2015). There are various scenarios for this mode of dispersal event, including the hitch-hiking, stepping-stones, and long distance dispersal scenarios (Keppel et al., 2009) mediated through a floatation-facilitating layer in the seeds of *Cycas* (Xiao and Möller, 2015; Zheng et al., 2016).

To further elucidate this historical biogeographic process, we first tested for the diversification model explaining the diversification history of the genus. All the tests point to an overall constant diversification history over time. Such constant diversification has recently been reported for another cycad genus, the African-endemic genus *Encephalartos* (Yessoufou et al., 2014; Mankga et al., 2020b). This suggests that cycads in general, not only diverged globally at the similar period (Nagalingum et al., 2011), but their diversification may have, perhaps, followed a similar overall pattern of constant-rate diversification. We explored several evolutionary events that might shape the diversification of *Cycas*, including speciation, extinction, and mass extinctions. Around 12 Ma, we found an initial speciation rate that is very similar to the overall speciation rate reported for gymnosperm in general (Crisp and Cook, 2011). However, the overall speciation corresponds to the late Miocene (Tortonian-Messinian), a period characterized in the Pacific regions by frequent sea level excursions (e.g., eight sea level excursions; Aharon et al., 1993). These multiple frequent rises and falls of sea level would likely promoted long-distance dispersal of *Cycas* seeds across the Pacific Islands through to Australia, Madagascar and East Africa. For example, species in the subsection Rumphiae developed seeds with spongy layer inside the sclerotesta (de Laubenfels and Adema, 1998); the "spongy" characteristic of the seeds facilitates the floatation of the seeds, potentially promoting a long *trans*-oceanic dispersal across the Pacific Islands (de Laubenfels and Adema, 1998; Xiao and Möller, 2015; Zheng et al., 2017).

Cycads have a fascinating evolutionary history (Mankga et al., 2020b) starting around 300 Ma (Hendricks, 1987), and

the extant cycads re-diversified around 12-2 Ma (Nagalingum et al., 2011). They share morphological characteristics of ferns and angiosperms (Norstog and Nicholls, 1997; Brenner et al., 2003), and these characteristics make them a unique taxonomic and evolutionary group. In this group, the genus *Cycas* has recently been identified as the most rapidly diversified and widely distributed clade (Yessoufou et al., 2017). Here we build upon this knowledge to reconstruct the historical biogeography and the evolutionary events that might shape the rapid diversification and wide distribution across the Pacific Islands. Our analysis indicated that Indo-China may have been the origin of the genus (but see Xiao and Möller, 2015), and that the pacific island may have been first colonized through dispersal before the genus reaches South China. This dispersal may have been facilitated by multiple excursions of sea level and the development of a key innovation, a spongy endocarp. Our study therefore clarifies the historical biogeography and the evolutionary events that shaped the current diversity of the genus.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found in the **Supplementary Material**.

AUTHOR CONTRIBUTIONS

KY designed the project and wrote the manuscript with editorial works from all other authors. LM collected the data. LM and KY analyzed the data. MC did the editorial works. TM did the GIS mapping.

FUNDING

We acknowledge the South Africa's National Research Foundation (NRF) for funding (Research Development Grants for Y-Rated Researchers Grant No: 112113) to Dr. KY. Ms. LM acknowledges the staff grants support from the University of South Africa.

ACKNOWLEDGMENTS

We acknowledge the two reviewers including the editor whose comments improve greatly an initial draft of this manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fevo.2020.00044/full#supplementary-material

TABLE S1 | The species names and GenBank accession numbers of sequences used in the data analyses. All the sequences *trn*H*psb*A, trnLF, trnS-trnG,psbM-trnD, RPB1, HZP, AC3, F3H, SAMS, PHYP and GTP were retrieved from GenBank. "—" indicates DNA sequences that are not available.

TABLE S2 | Summary statistics of the aligned DNA matrix.

TABLE S3 | The Bayes factor (BF) values calculated for each birth-death model tested for the phylogeny of the genus *Cycas*. ConstBD, constant-rate birth-death model; DecrBD, continuously variable-rate birth-death model; EpisodicBD, episodically variable-rate birth-death model; MassExtinctionBD, explicit mass-extinction birth-death model.

REFERENCES

Aharon, P., Goldstein, S. L., Wheeler, C. W., and Jacobson, G. (1993). Sea level events in the South Pacific linked with the messinian salinity crisis. *Geology* 21, 771–775.

Baele, G., Lemey, P., and Vansteelandt, S. (2013). Make the most of your samples: bayes factor estimators for high-dimensional models of sequence evolution. *BMC Bioinform.* 14:85. doi: 10.1186/1471-2105-14-85

Beeravolu, C. R., and Condamine, F. L. (2016). An extended maximum likelihood inference of geographic range evolution by dispersal, local extinction and cladogenesis. *BioRxiv* [Preprint]. doi: 10.1101/038695

Brenner, E. D., Stevenson, D. W., and Twigg, R. W. (2003). Cycads: evolutionary innovations and the role of plant-derived neurotoxins. *Trends Plant Sci.* 8, 446–452. doi: 10.1016/s1360-1385(03)00190-0

Brown, J. H. (1984). On the relationship between abundance and distribution of species. *Am. Natl.* 124, 255–279. doi: 10.1086/284267

Calonje, M., Stevenson, D. W., and Stanberg, L. (2017). *The World List of Cycads, Online edition [Internet]. 2013-2017*. Available at: http://www.cycadlist.org (accessed January, 2017).

Carlquist, S. (1967). The biota of long-distance dispersal V. plant dispersal to Pacific Islands. *Bull. Torrey Bot. Club* 94, 129–162.

Condamine, F. L., Nagalingum, N. S., Marshall, C. R., and Morlon, H. (2015). Origin and diversification of living cycads: a cautionary tale on the impact of the branching prior in Bayesian molecular dating. *BMC Evol. Biol.* 15:65. doi: 10.1186/s12862-015-0347-8

Corner, E. J. H. (1963). "Ficus in the pacific region," in *Pacific Basin Biogeography – A Symposium*, ed. J. L. Gressitcpesnm, (Hawaii: Bishop Museum Press), 233–245.

Couvreur, T. L. P., Franzke, A., Al-Shehbaz, I. A., Bakker, F., Koch, M., and Mummenhoff, K. (2010). Molecular phylogenetics, temporal diversification and principles of evolution in the mustard family (Brassicaceae). *Mol. Biol. Evol.* 27, 55–71. doi: 10.1093/molbev/msp202

Crisp, M. D., and Cook, L. G. (2011). Cenozoic extinctions account for the low diversity of extant gymnosperms compared with angiosperm. *New Phytol.* 192, 997–1009. doi: 10.1111/j.1469-8137.2011.03862.x

Darwin, C. (1859). *On the Origin of Species by Means of Natural Selection or the Preservation of Favored Races in the Struggle for Life*. London: John Murray. doi: 10.1111/j.1469-8137.2011.03862.x

de Laubenfels, D. J., and Adema, F. A. (1998). Taxonomic revision of the genera Cycas and Epicycas gen. nov. (*Cycadaceae*). *Blumea* 43, 351–400.

Drummond, A. J., and Rambaut, A. (2007). BEAST: bayesian evolutionary analysis by sampling trees. *BMC Evol. Biol.* 7:214. doi: 10.1186/1471-2148-7-214

Ebach, M. C., and Tangey, R. S. (2006). *Biogeography in a Changing World*. Boca Raton, FL: CRC Press. doi: 10.1186/1471-2148-7-214

Gao, Z. F., and Thomas, B. A. (1989). A review of fossil cycad megasporophylls, with new evidence of crossozamia pommel and its associated leaves from the lower Permian of Taiyuan, China. *Rev. Palaeobot. Palynol.* 60, 205–223. doi: 10.1016/0034-6667(89)90044-4

Gaston, K. J. (2009). Geographic range limits: achieving synthesis. *Proc. R. Soc. B Biol. Sci.* 276, 1395–1406. doi: 10.1098/rspb.2008.1480

Guan, Z. T. (1996). *Cycads of China*. Chengdu: Sichuan Science and Technology Press. doi: 10.1098/rspb.2008.1480

Guppy, H. B. (1906). *Observations of a Naturalist in the Pacific between 1896 and 1899. Volume II: Plant Dispersal*. London: Macmillan and Co. doi: 10.1098/rspb.2008.1480

Hampe, A., and Petit, R. J. (2005). Conserving biodiversity under climate change: the rear edge matters. *Ecol. Lett.* 8, 461–467. doi: 10.1111/j.1461-0248.2005.00739.x

Heads, M. (2006). Seed plants of Fiji: an ecological analysis. *Biol. J. Linnean Soc.* 89, 407–431. doi: 10.1111/j.1095-8312.2006.00682.x

Heads, M. (2008). Panbiogeography of New Caledonia, southwest Pacific: basal angiosperms on basement terranes, ultramafic endemics inherited from volcanic island arcs and old taxa endemic to young islands. *J. Biogeogr.* 35, 2153–2175. doi: 10.1111/j.1365-2699.2008.01977.x

Hendricks, J. G. (1987). The gondwanan cycas. *Encephalartos* 10, 24–25.

Hill, K. D. (1995). "Infrageneric relationships, phylogeny and biogeography of the genus Cycas (Cycadaceae)," in *CYCAD 93, The 3rd International Conference on Cycad Biology, Proceedings*, ed. P. Vorstercpesnm, (Stellenbosch: Cycad Society), 139–162.

Hill, K. D. (1996a). "Cycas, an evolutionary perspective. Biology and conservation of cycads," in *Proceedings of the Fourth International Conference on Cycad Biology*, Panzhihua.

Hill, K. D. (1996b). "Cycads in the Pacific. The origin and evolution of Pacific island biotas," in *New Guinea to Eastern Polynesia: Patterns and Processes*, eds A. Keastcpesnm, and S. E. Millercpesnm, (Amsterdam: SPB Academic Publishing), 267–274.

Hill, K. D. (2004). "Character evolution, species recognition and classification concepts in the Cycadaceae," in *Cycad Classification, Concepts and Recommendations*, eds T. Walterscpesnm, and R. Osbornecpesnm, (Wallingford: CABI Publishing), 23–44. doi: 10.1079/9780851997414.0023

Hill, K. D., and Yang, S. L. (1999). The genus cycas (Cycadaceae) in Thailand. *Brittonia* 51, 48–73. doi: 10.1016/j.ympev.2018.05.019

Hillis, D. M., and Bull, J. J. (1993). An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis. *Syst. Biol.* 42, 182–192. doi: 10.1093/sysbio/42.2.182

Höhna, S., May, M. R., and Moore, B. R. (2015). TESS: an R package for efficiently simulating phylogenetic trees and performing Bayesian inference of lineage diversification rates. *Bioinform* 32, 789–791. doi: 10.1093/bioinformatics/btv651

Jeffreys, H. (1961). *The Theory of Probability*. Oxford: Oxford University Press. doi: 10.1093/bioinformatics/btv651

Jones, D. L. (2002). *Cycads of the World-Ancient Plant in Today's Landscape*. Washington, DC: Smithsonian Institution Press. doi: 10.1093/bioinformatics/btv651

Kawecki, T. J. (2008). Adaptation to marginal habitats. *Ann. Rev. Ecol. Evol. Syst.* 39, 321–342. doi: 10.1146/annurev.ecolsys.38.091206.095622

Keast, A., and Miller, S. E. (1996). *The Origin and Evolution of Pacific Island Biotas, New Guinea to Eastern Polynesia: Patterns and Processes*. Amsterdam: SPB Academic Publishing. doi: 10.1146/annurev.ecolsys.38.091206.095622

Keppel, G., Hodgskiss, P. D., and Plunkett, G. M. (2008). Cycads in the insular Southwest Pacific: dispersal or vicariance? *J. Biogeogr.* 35, 1004–1015. doi: 10.1111/j.1365-2699.2007.01869.x

Keppel, G., Lowe, A. J., and Possingham, H. P. (2009). Changing perspective on the biogeographical of the tropical South Pacific: influences of dispersal, vicariance and extinction. *J. Biogeogr.* 36, 1035–1054. doi: 10.1111/j.1365-2699.2009.02095.x

Ladiges, P. Y., and Cantrill, D. (2007). New Caledonia–Australia connections: biogeographic patterns and geology. *Aust. Syst. Bot.* 20, 383–389.

Ladiges, P. Y., Udovicic, F., and Nelson, G. (2003). Australian biogeographical connections and the phylogeny of large genera in the plant family Myrtaceae. *J. Biogeogr.* 30, 989–998. doi: 10.1046/j.1365-2699.2003.00881.x

Liu, J., Zhang, S., Nagalingum, N. S., Chiang, Y., Lindstrom, A., and Gong, X. (2018). Phylogeny of the gymnosperms genus Cycas L. (Cycadaceae) as inferred from plastid and nuclear loci based on a large-scale sampling: evolutionary relationships and taxonomical implications. *Mol. Phylogenet. Evol.* 127, 87–97. doi: 10.1016/j.ympev.2018.05.019

Mankga, L. T., Yessoufou, K., and Chitakira, M. (2020a). On the origin and diversification history of the African genus *Encephalartos*. *South Afr. J. Bot.* 130, 231–239. doi: 10.1016/j.sajb.2019.12.007

Mankga, L. T., Yessoufou, K., Mugwena, T., and Chitakira, M. (2020b). The genus cycas may have diversified from Indochina and occupied its current ranges through vicariance and dispersal events. *Dryad, Dataset*. https://doi.org/10.5061/dryad.1g1jwstrn

May, M. R., Höhna, S., and Moore, B. R. (2016). A Bayesian approach for detecting mass-extinction events when rates of lineage diversification vary. *Syst. Biol.* 7, 947–959. doi: 10.1111/2041-210x.12563

Mayr, E. (1954). "Change of genetic environment and evolution," in *Evolution as a Process*, eds J. Huxley, A. C. Hardy, and E. B. Fordcpesnm, (London: Allen and Unwin), 157–180.

Nagalingum, N. S., Marshall, C. R., Quental, T. B., Rai, H. S., Little, D. P., and Mathews, S. (2011). Recent synchronous radiation of a living fossil. *Science* 334, 796–799. doi: 10.1126/science.1209926

Nelson, G. and Platnick, N. I. (1981). Recent synchronous radiation of a living fossil. *Systematics and Biogeography: Cladistics and Vicariance*. New York, NY: Columbia University Press. doi: 10.1126/science.1209926

Norstog, K. J., and Nicholls, T. J. (1997). "The fossil Cycadophytes," in *The Biology of the Cycads*, eds K. J. Norstogcpesnm, and T. J. Nicholllscpesnm, (Ithaca, NY: Cornell University Press), 169–201.

Nylander, J. A. A. (2004). *Modeltest v2. Program Distributed by the Author*. Sweden: Evolutionary Biology Centre.

Osborne, R., Calonje, M. A., Hill, K. D., Stanberg, L., and Stevenson, D. W. (2012). The world list of Cycads. *Mem. N.Y. Bot. Gard.* 106, 480–510.

Perrie, L. R., and Brownsey, P. (2007). Molecular evidence for long-distance dispersal in the New Zealand pteridophyte flora. *J. Biogeogr.* 34, 2028–2038. doi: 10.1111/j.1365-2699.2007.01748.x

Price, J. P., and Clague, D. A. (2002). How old is the Hawaiian biota? Geology and phylogeny suggest recent divergence. *Proc. R. Soc. B Biol. Sci.* 269, 2429–2435. doi: 10.1098/rspb.2002.2175

Rambaut, A., and Drummond, A. J. (2007). *Tree Annotator (Version 1.5.4)*. Available at http://beast.bio.ed.ac.uk (accessed January, 2017).

Ree, R. H., and Smith, S. A. (2008). Maximum likelihood inference of geographic range evolution by dispersal, local extinction, and cladogenesis. *Syst. Biol.* 57, 4–14. doi: 10.1080/10635150701883881

Ridley, N. H. (1930). *Dispersal of Plants Throughout the World*. Ashford: L. Reeve and Co. doi: 10.1080/10635150701883881

Schenk, J. J. (2016). Consequences of secondary calibrations on divergence time estimates. *PLoS One* 11:e0148228. doi: 10.1371/journal.pone.0148228

Sexton, J. P., McIntyre, P. J., Angert, A. L., and Rice, K. J. (2009). Evolution and ecology of species range limits. *Ann. Rev. Ecol. Evol. Syst.* 4, 415–436. doi: 10.1146/annurev.ecolsys.110308.120317

Stevenson, D. W. (1992). A formal classification of the extant cycads. *Brittonia* 44, 220–223.

Swofford, D. L. (2002). *PAUP: Phylogenetic Analysis Using Parsimony (and Other methods), Version 4.0b10*. Sunderland, MA: Sinauer Associates.

Turner, H., Hovenkamp, P. H., and van Welzen, P. C. (2001). Biogeography of Southeast Asia and the West Pacific. *J. Biogeogr.* 28, 217–230. doi: 10.1046/j.1365-2699.2001.00526.x

van Balgooy, M. M. J. (1969). A study on the diversity of island floras. *Biodiv. Evol. Biogeogr. Plants* 17, 139–178.

Whitmore, T. C. (1973). Plate tectonics and some aspects of Pacific plant geography. *New Phytol.* 72, 1185–1190. doi: 10.1111/j.1469-8137.1973.tb02095.x

Wilcox, T. P., Zwicki, D. J., Heath, T. A., and Hillis, D. M. (2002). Phylogenetic relationships of the dwarf boas and a comparison of Bayesian and bootstrap measures of phylogenetic support. *Mol. Phylogenet. Evol.* 25, 361–371. doi: 10.1016/s1055-7903(02)00244-0

Winkworth, R. C., Wagstaff, S. J., Glenny, D., and Lockhart, P. J. (2002). Plant dispersal N.E.W.S. from New Zealand. *Trends Ecol. Evol.* 17, 514–520.

Winkworth, R. C., Wagstaff, S. J., Glenny, D., and Lockhart, P. J. (2005). Evolution of the New Zealand mountain flora: origins diversification and dispersal. *Org. Diver. Evol.* 5, 237–247. doi: 10.1016/j.ode.2004.12.001

Woodroffe, C. D. (1987). Pacific Island mangroves: distribution and environmental settings. *Pac. Sci.* 41, 166–185.

Xiao, L. Q., and Möller, M. (2015). Nuclear ribosomal ITS functional paralogs resolve the phylogenetic relationships of a late-Miocene radiation cycad *Cycas* (Cycadaceae). *PLoS One* 10:e0117971. doi: 10.1371/journal.pone.0117971

Yan, Y., Harris, A. J., and Xingjin, H. (2010). S-DIVA (Statistical Dispersal-Vicariance Analysis): a tool for inferring biogeographic histories. *Mol. Phylogenet. Evol.* 56, 848–850. doi: 10.1016/j.ympev.2010.04.011

Yang, S. L., and Meerow, A. W. (1996). The *Cycas pectinata* (Cycadaceae) complex: genetic structure and gene flow. *Int. J. Plant Sci.* 157, 468–483. doi: 10.1086/297364

Yessoufou, K., Bamigboye, S. O., Daru, B. H., and Van der Bank, M. (2014). Evidence of constant diversification punctuated by a mass extinction in the African cycads. *Ecol. Evol.* 4, 50–58. doi: 10.1002/ece3.880

Yessoufou, K., Daru, B. H., Tafirei, R., Elansary, H. O., and Rampedi, I. (2017). Integrating biogeography, threat and evolutionary data to explore extinction crisis in the taxonomic group of cycads. *Ecol. Evol.* 7, 2735–2746. doi: 10.1002/ece3.2660

Zheng, Y., Liu, J., Feng, X., and Gong, X. (2017). The distribution, diversity, and conservation status of Cycas in China. *Ecol. Evol.* 7, 3212–3224. doi: 10.1002/ece3.2910

Zheng, Y., Liu, J., and Gong, X. (2016). Tectonic and climatic impacts on the biota within the red river fault, evidence from phylogeography of *Cycas dolichophylla* (Cycadacaeae). *Sci. Rep.* 6:33540. doi: 10.1038/srep33540

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.

frontiers in Ecology and Evolution

BRIEF RESEARCH REPORT
published: 28 February 2020
doi: 10.3389/fevo.2020.00009

Early Alert of Biological Risk in a Coastal Lagoon Through eDNA Metabarcoding

*Marcos Suarez-Menendez[1], Serge Planes[2,3], Eva Garcia-Vazquez[1] and Alba Ardura[1]**

[1] Department of Functional Biology, University of Oviedo, Oviedo, Spain, [2] USR 3278-CRIOBE-CNRS-EPHE, Laboratoire d'excellence "CORAIL," Universite de Perpignan-CBETM, Perpignan, France, [3] Laboratoire d'Excellence "CORAIL," Centre de Recherche Insulaire et Observatoire de l'Environnement (CRIOBE), Papetoai, French Polynesia

Mediterranean coastal lagoons are environmentally important ecosystems whose conservation has been challenged due to anthropogenic impacts that promoted the expansion of non-indigenous and, sometimes, invasive species. Therefore, it is crucial to inventory biodiversity in these areas for the development of strategies of conservation and management. Classical methods used for biodiversity surveys and detection of non-native species may be unsuccessful for the detection and identification of species in early development stages such as cryptic, microscopic, elusive, and new coming species at low population density. The development of metabarcoding techniques in the last decade offers new opportunities for reliable biodiversity surveillance and facilitates early detection of nuisance species. The objective of this study was to analyze the species occurring in the protected coastal lagoon Canet-Saint Nazaire using a simple sampling protocol based on water samples and environmental DNA (eDNA) metabarcoding with a single barcode (cytochrome c oxidase subunit I [COI] gene). Two invasive species (*Polydora cornuta* and *Acartia tonsa*), two polychaete bioindicators of pollution (*Hediste diversicolor* and *Capitella capitata*), and one alga that produces harmful algal blooms were detected from only 6 L of water, indicating environmental degradation in the lagoon despite its protected status. These results demonstrate the importance of COI as single barcode together with eDNA as an ecological early warning system and suggest the need for environmental restoration in this lagoon.

Keywords: COI "barcode," next generation sequencing, Canet lagoon, Nature 2000 areas, conservation

OPEN ACCESS

Edited by:
David S. Thaler, Universität Basel, Switzerland

Reviewed by:
Alejandro Trujillo Gonzalez, University of Canberra, Australia
Rebecca Barr Simmons, University of North Dakota, United States

***Correspondence:**
Alba Ardura
arduraalba@uniovi.es

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

Received: *03 October 2019*
Accepted: *15 January 2020*
Published: *28 February 2020*

Citation:
Suarez-Menendez M, Planes S, Garcia-Vazquez E and Ardura A (2020) Early Alert of Biological Risk in a Coastal Lagoon Through eDNA Metabarcoding. Front. Ecol. Evol. 8:9. doi: 10.3389/fevo.2020.00009

INTRODUCTION

Coastal lagoons occupy 13% of the coastal area worldwide (Barnes, 1980) and are among the marine habitats showing the highest biological productivity, by providing diverse habitat types for many species, nursery areas, and feeding grounds for marine and estuarine fishes (Pérez-Ruzafa et al., 2011). They are distinctive ecosystems because they are shallow coastal water bodies separated from the ocean by a barrier and connected intermittently to the sea (Kjerfve, 1994). They also support important fisheries and allow for intensive aquaculture exploitation (Cataudella et al., 2015). Despite their environmental and economic importance, and protected status, lagoons suffer from several threats derived from human activities such as the effects of climate change, pollution, eutrophication, and the introduction of non-indigenous species (NIS) (Reizopoulou et al., 1996;

Chapman, 2012). Human activities, such as the increases in maritime traffic and the opening of the Suez Canal, have facilitated NIS introductions in the Mediterranean Sea (Katsanevakis et al., 2014). In coastal lagoons, NIS settlement is facilitated by their naturally stressful conditions, pollution, loss of native species, and intense shipping in the numerous nearby harbors (Ruiz et al., 2000; Crooks et al., 2011). NIS can become invasive and affect both native species and economic activities in the area (Galil, 2007).

Together with NIS, eutrophication plays an important role in the degradation of these estuarine systems. Inefficient wastewater management and intense agricultural activities in the catchment area of the lagoons can increase the concentration of nutrients (Carlier et al., 2008). The consequent eutrophic state facilitates the uncontrolled growth of organisms such as dinoflagellates, diatoms, and cyanobacteria that produce harmful algal blooms (HABs, also referred as red tides) (Anderson et al., 2002). The increasing algal bloom episodes is not only favored by eutrophication but also by climate change and the introduction of new strains by ballast water (Hallegraeff, 1993; Moore et al., 2008). HABs have several effects on the environment. The mere growth and following decay of organic material cause anoxic conditions, leading to the death of aquatic life. Some HABs also produce toxins that affect both marine life and human health even at low densities (Sellner et al., 2003). In humans, these toxins can cause different types of poisoning (diarrheal, neurotoxic, and paralytic). The most frequent are diarrheic shellfish poisoning (DSP), mainly due to toxic strains of *Prorocentrum* spp. and *Dinophysis* spp. (Yasumoto et al., 1980; Bravo et al., 2001), and paralytic shellfish poisoning (PSP) predominantly linked to toxic strains of *Alexandrium* spp. (Anderson et al., 2012).

The correct management and conservation of coastal areas requires efficient assessment of their biodiversity and detection and identification of species that may be of environmental concern, such as pollution indicators and NIS. For this purpose, species surveys are carried out periodically, generally consisting of manual sampling and visual identification of sampled specimens by taxonomical experts. Such a classical morphological analysis requires high sampling effort and might be inefficient in the detection of some species, for example those that are in early development stages, cryptic species, or microorganisms (Ficetola et al., 2008). For NIS, their early detection is crucial to prevent their establishment and dispersion because eradication or control are more efficient when the species are at low density soon after introduction (Gozlan et al., 2010). However, low population densities require greater survey efforts for detection, and many newly established populations may go unnoticed (Blanchet, 2012). Classical approaches for phytoplankton identification are often considered time consuming while requiring expertise in taxonomical identification. Identifying dinoflagellates represent a greater challenge due to high morphological similarities and a lack of unique characters between different species (Lin et al., 2009). DNA metabarcoding is a rapidly evolving method for assessing biodiversity that exceeds the limitations of traditional methods. It combines the use of environmental DNA (eDNA) and next-generation sequencing (NGS) allowing the detection of species from single cells in an environmental sample (e.g., soil or water) (Taberlet et al., 2012). Although eDNA metabarcoding was found to be less sensitive than a targeted monitoring approach for generating detailed specific distribution data (Furlan et al., 2015; Ardura et al., 2016; Bylemand et al., 2019), it has the ability to provide baseline information on biodiversity patterns (e.g., cryptic species and novel incursions of invasive species) (Blackman et al., 2017), capturing a great part of the aquatic community diversity (Zaiko et al., 2015). Borrell et al. (2017) have proposed the use of metabarcoding as an early alert method for the detection of invasive species in ports. They used two metabarcodes (cytochrome oxidase I and 18S rRNA genes) on 3 L of water from each sampling point. For the need of easy methods for alert of environmental disturbances, here, we will test a similar but simplified approach for the exploration of nuisance species, including invasive species and HABs.

We apply metabarcoding (eDNA) for the detection of aquatic nuisance organisms in the coastal lagoon Canet-Saint Nazaire, situated in the French Mediterranean coast. Previous studies developed in this lagoon using a rapid assessment of invertebrate species and barcoding (Ardura and Planes, 2017) suggested a high level of degradation and vulnerability of the lagoon. However, in that study, small species like HABs and organisms in early development stages could not be recognized. In the present investigation, eDNA was extracted from water samples from the canal that connects the lagoon to the open sea, and a region of the cytochrome c oxidase subunit I (COI) gene was used as a single metabarcode for species identification, since it has been used as a barcode of high-resolution power (Hebert et al., 2003a) and reference databases for aquatic organisms are more complete for this gene than other barcodes (Weigand et al., 2019). The principal aim of this study was focused on the detection of species that may reveal a risk for the conservation status and environmental health of the lagoon.

MATERIALS AND METHODS

Study Area

Sampling was conducted in the canal that connects the Canet-Saint Nazaire Lagoon to the open sea (**Figure 1**). This lagoon is a special protection area (SPA) for birds within the Natura 2000 network, and it is one of the many coastal lagoons spanning the French Mediterranean coastline. It is a semiclosed system, with its principal water sources coming from two rivers and streams and, to a lesser extent, a canal connected to the sea. This, along with seasonal variations in temperature, drought periods, and sudden and intense rainfalls, causes important changes in its depth and salinity that varies from 13.2 to 35.6 during the year (Vouvé et al., 2014). Besides this, the lagoon suffers from intense pollution, eutrophication, and sediment filling due to human activities in its catchment area (agriculture, water treatment, and tourism) and scarce water exchange with the sea (Carlier et al., 2008; Souchu et al., 2010). Within 4 km of the lagoon, Canet port is located, which is a popular location for recreational boating with space for 977 vessels up to 24 m in length (Portbooker website, 2018). The presence of this harbor, along with the lagoon's brackish water, makes Canet-Saint Nazaire extremely

FIGURE 1 | Lagoon of Canet-Saint Nazaire (below) and port of Canet (above) indicating the sampling site (Google Maps, 2018).

vulnerable to introduction of non-native species (Paavola et al., 2005; Ardura and Planes, 2017).

Sampling

Water samples of the lagoon were taken in the proximity of the canal (42°39′25.19″N 3° 1′39.30″E). Six liters of surface water was collected in three sterile bottles of 2 L (three replicates) on November–December 2016. All the water from each bottle was vacuum filtered through several filters of different material and pore size to collect all DNA from the sample without clogging the filters. Bottles are sterile, and cross-contamination is not expected because the sampling is done in unique point. First, the whole volume of each sample was filtered through 10-μm nylon filters; then, each sample was divided in four and filtered through 0.8-μm polyethersulfone (PES) filters. Finally, the whole volume of each sample was pooled together and filtered through 0.2-μm PES filters. In total, 18 filters (3 replicates × 6 filters each one) were obtained, which were thereafter preserved with 96% ethanol until eDNA extraction.

DNA and Bioinformatics Analysis

DNA was extracted from the filters using PowerWater® DNA Isolation Kit (MOBIO Laboratories, EE. UU.) following the manufacturer's extraction protocol. In addition, ethanol was centrifuged, and the pellet, with precipitated DNA, was added in the extracting process. A negative control of pure water was added at this step to monitor contamination during the extraction process. In addition, bovine serum albumin (BSA) was added to the PCR reactions to increase PCR yields from low-purity templates and to avoid, as much as possible, the effect of inhibitors present in the water. Extracted DNA from all the filters was sent to Macrogen (Seoul, South Korea) were it was quantified and sequenced. The eDNA was quantified by a fluorescence-based method Victor 3 (Picogreen, Invitrogen). The modified universal COI primers mlCOIintF/jgHCO2198 (Leray et al., 2013) were used for PCR amplification of a fragment of ∼313 bp within the mitochondrial gene coding for the COI. The primers were modified to include Illumina sequencing adapters and sample-specific dual indices (i5 and i7) following the Illumina (2013) protocol in which the conditions for the amplicon PCR were changed for the ones described by Leray et al. (2013). The PCR reactions were undertaken by Macrogen, following the protocol Illumina (2013). After constructing the library, it was charged into the platform Illumina MiSeq v3 reagents that generates paired-end sequences (2 × 301). Adapters and indices were removed from the raw data along with reads <36 bp using Scythe and Buffalo software, respectively.

The trimmed data were received from Macrogen in Fastaq format and were further processed using the platform Qiime1 (Caporaso et al., 2010). All the sequence reads were assessed for quality by applying a Phred quality score threshold of 20 and were filtered by length (200 bp ≤ reads ≤400 bp) from the downstream analysis.

The paired-end reads from each sample were merged when they presented an overlapping region of at least 100 bp and <15% of differences within this region. This step is necessary to merge forward and reverse reads. Some errors often appear in the Illumina data as the merging is not always 100%, due to sequencing length and PCR errors. For this, one of the thresholds in the merging algorithm must be the percentage of coincidence within the merging region to avoid artifact sequences when merging them and to improve the following assignment against reference database.

For the taxonomic assignation, Basic Local Alignment Search Tool (BLAST) alignment was performed against NCBI database[1] of COI sequences (obtained in 09/2017) filtered for environmental reads and using as threshold criteria: maximum E value = 1×10^{-50} and minimum percent identity = 97.0, which is enough in most of the cases for species identification from COI barcode (Hebert et al., 2003b). Some reads may have multiple BLAST hits at a 100% identity to different species, or the best BLAST hit have no species-level reference available. In those cases, the operational taxonomic unit (OTU) was assigned to a genus level.

OTU tables, a list of OTUs obtained for each sample and the number of sequences assigned to them, were constructed, clustering reads with a 100% identity between them and maintaining all assigned sequences including singletons to retain maximum sensitivity for species detection. The removal of singletons is usually employed to eliminate false positives as proposed by Scott et al. (2018), in the context of species survival, NIS early detection, or marine biosecurity surveillance; a false negative is most costly than a false positive (von Ammon et al., 2018). Sequences of organisms without relevance for the study (e.g., human, insects, terrestrial plants, etc.) were removed from the dataset (**Figure 2**). Generally, chimera formation is not extensive, although some chimeras can be found. In addition, one sequence per OTU was BLASTed manually to verify the reliability of the pipeline parameters and discard the presence of chimeras (Nilsson et al., 2012). The taxonomic information from the remaining OTUs was checked against the World Register

[1]https://www.ncbi.nlm.nih.gov/

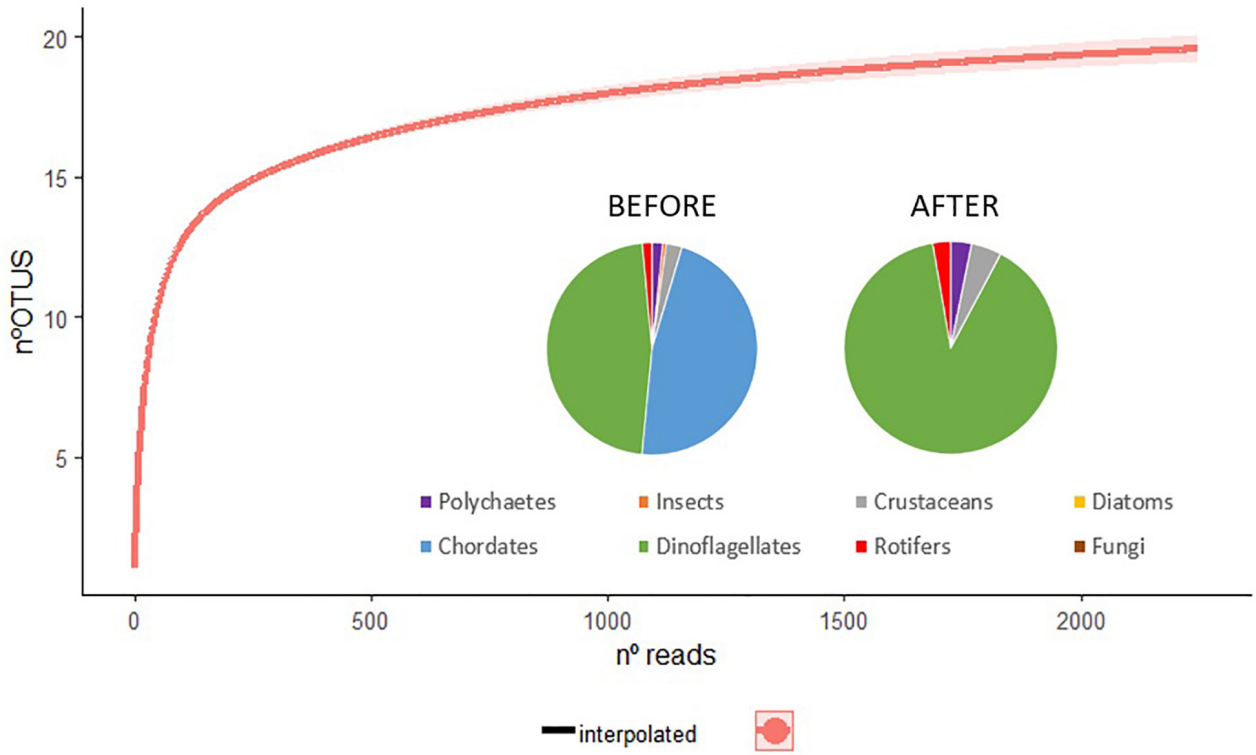


FIGURE 2 | Rarefaction curve plotting the number of reads by the number of operational taxonomic units (OTUs) for the sample analyzed using high-throughput sequencing. Interpolated line in red. Graphics with the proportion of different taxonomical groups detected in the lagoon before and after expert check (eliminating the reads assigned to terrestrial and exclusive marine species).

of Marine Species[2] and AlgaeBase[3]. The number of reads per OTU, as a proxy for species abundance, was used to generate OTU rarefaction curves using Vegan package in R software (Oksanen et al., 2013).

RESULTS

The minimum concentration of the extracted eDNA was 0.076 ng/μl. A total of 581,786 raw paired-end reads were sequenced from positive PCR amplicons (average read length, 301 bp). Following initial filtering, a total of 328,220 reads (56.4%) were retained for downstream analysis (average read length, 269 bp). After merging, and a second quality filtering, 239,842 sequences remained (41.2%), of which 98,282 matched a correct reference sequence from the database (16.9%; average read length, 359 bp) (**Table 1**).

The rarefaction curve of sequences indicated that the sample reached a plateau (**Figure 2**). This figure represents three 2-L superficial water samples, which were then grouped for the purpose of downstream analysis, hence the single plotted line. This indicates that, in terms of the number of sequences generated, our sequencing had enough depth to detect most of the possible species present within the sample.

After the application of expert check, eliminating the reads assigned to terrestrial and exclusive marine species, 51,624 sequences were left assigned to aquatic organisms at least at class level (**Table 1**). Of these reads, 10.5% were identified to a species level, while the majority was assigned to a genus or higher level. The presence of singletons may prevent reaching the asymptote in the curve of accumulation of species. In this case, the rare species (singletons and doublets) represent 11.1% (3 of the 27 OTU obtained), which did not impede to reach the asymptote, indicating that, in a number of sequences generated, the sample was adequately sized (**Figure 2**).

Regarding the biota profile captured from NGS after expert check, the assigned sequences were clustered in 27 different OTUs of which 17 belonged to the Class Dinophyceae (89.5% of the reads), while the rest was divided into four different classes by sequences number order: crustaceans, polychaetes, diatoms, and rotifers (**Figure 2**).

Considering only the sequences assigned at least at a genus level (28,029 in total), 25 taxa were identified, 10 of them at a species level. The genus with more reads was *Protoperidinium* (19.4%), followed by *Scrippsiella* (19.1%), and *Levanderina* (16.9%) (**Table 1**), pointing out at a clear dominance of Dinophyceae in the water samples analyzed not only in the number of taxa but also in abundance of reads. The following taxon in abundance of reads was the acorn barnacle *Perforatus perforatus* (8.3% of reads), then the whip mudworm *Polydora*

[2]http://www.marinespecies.org/
[3]http://www.algaebase.org/

TABLE 1 | Sequences matched a correct reference sequence from the GenBank database per filter, after merging and quality filtering.

GenBank ID	Phylum	Class	Species	0.2 μ mesh size	0.8 μ mesh size	10 μ mesh size
KP254366.1	Annelida	Polychaeta	*Capitella capitata*	0	0	3
KR916843.1	Annelida	Polychaeta	*Hediste diversicolor*	0	0	4
AB636160.1	Annelida	Polychaeta	*Polydora cornuta*	0	0	1,619
KM578774.1	Arthropoda	Insecta	*Attagenus gobicola*	0	0	2
KJ962220.1	Arthropoda	Insecta	*Dermestes lardarius*	7	0	289
HQ824544.1	Arthropoda	Insecta	*Dichotomius nisus*	1	0	1
AF253029.1	Arthropoda	Insecta	*Ochlerotatus detritus*	0	0	125
KM578800.1	Arthropoda	Insecta	*Trogoderma yunnaeunsis*	0	0	107
KC287363.1	Arthropoda	Maxillopoda	*Acartia tonsa*	0	0	1
KP136566.1	Arthropoda	Maxillopoda	*Cyclopoida* environmental sample	0	0	2
KF297550.1	Arthropoda	Maxillopoda	*Perforatus perforatus*	0	2,324	0
DQ059772.1	Arthropoda	Maxillopoda	*Sacculina carcini*	0	0	46
FJ590523.1	Ascomycota	Dothideomycetes	*Cladosporium bruhnei*	28	0	0
GQ844253.1	Bacillariophyta	Bacillariophyceae	*Cylindrotheca fusiformis*	0	0	1
HF563534.1	Bacillariophyta	Bacillariophyceae	*Haslea crucigera*	0	0	9
FJ519930.1	Chordata	Chondrichthyes	*Galeocerdo cuvier*	0	0	9
AP008737.1	Chordata	Mammalia	*Homo sapiens*	20,333	0	25,756
GQ501113.1	Dinoflagellata	Dinophyceae	*Akashiwo sanguinea*	0	0	670
GQ501119.1	Dinoflagellata	Dinophyceae	*Alexandrium affine*	0	0	84
GQ501128.1	Dinoflagellata	Dinophyceae	*Alexandrium catenella*	0	0	210
GQ501141.1	Dinoflagellata	Dinophyceae	*Alexandrium lusitanicum*	0	0	1
GQ501157.1	Dinoflagellata	Dinophyceae	*Alexandrium minutum*	0	0	667
GQ501159.1	Dinoflagellata	Dinophyceae	*Alexandrium ostenfeldii*	0	0	64
GQ501142.1	Dinoflagellata	Dinophyceae	*Alexandrium* sp. RFS-2009a	0	0	1
GQ914937.1	Dinoflagellata	Dinophyceae	*Azadinium obesum*	0	0	1,104
KJ503235.1	Dinoflagellata	Dinophyceae	*Dinophysis acuminata*	0	0	1,046
GQ501853.1	Dinoflagellata	Dinophyceae	*Dinophysis* sp. PL9-13	0	0	76
GQ501849.1	Dinoflagellata	Dinophyceae	*Dinophysis* sp. PL9-3	0	0	7
GQ501848.1	Dinoflagellata	Dinophyceae	*Dinophysis* sp. PL9-4	0	0	1
GQ501217.1	Dinoflagellata	Dinophyceae	*Gonyaulax* sp. AC551	0	0	3
GQ501250.1	Dinoflagellata	Dinophyceae	*Karenia brevis*	0	0	9
GQ501256.1	Dinoflagellata	Dinophyceae	*Karlodinium veneficum*	0	0	125
GQ501233.1	Dinoflagellata	Dinophyceae	*Lepidodinium chlorophorum*	0	0	95
GQ501243.1	Dinoflagellata	Dinophyceae	*Levanderina fissa*	0	0	4,751
GQ501269.1	Dinoflagellata	Dinophyceae	*Peridinium inconspicuum*	0	0	51
GQ502055.1	Dinoflagellata	Dinophyceae	*Peridinium* sp. ES18-106	0	0	1
GQ502043.1	Dinoflagellata	Dinophyceae	*Peridinium* sp. ES18-111	0	0	5
GQ502054.1	Dinoflagellata	Dinophyceae	*Peridinium* sp. ES18-113	0	0	42
GQ502044.1	Dinoflagellata	Dinophyceae	*Peridinium* sp. ES18-128	0	0	1
GQ502074.1	Dinoflagellata	Dinophyceae	*Prorocentrum micans*	0	0	1,137
GQ502077.1	Dinoflagellata	Dinophyceae	*Prorocentrum* sp. ES11-87	0	0	1
GQ502071.1	Dinoflagellata	Dinophyceae	*Prorocentrum* sp. ES18-118	0	0	6
GQ502090.1	Dinoflagellata	Dinophyceae	*Prorocentrum* sp. ES5-50	0	0	1
GQ501301.1	Dinoflagellata	Dinophyceae	*Protoceratium reticulatum*	0	0	14
GQ502107.1	Dinoflagellata	Dinophyceae	*Protoperidinium* cf. *depressum*	0	0	5,448
GQ501312.1	Dinoflagellata	Dinophyceae	*Scrippsiella* cf. *precaria*	0	0	540
GQ501317.1	Dinoflagellata	Dinophyceae	*Scrippsiella precaria*	0	0	2
GQ501320.1	Dinoflagellata	Dinophyceae	*Scrippsiella* sp. CS-297	0	0	4,620
GQ501314.1	Dinoflagellata	Dinophyceae	*Scrippsiella* sp. DINOB785-08	0	0	44
GQ501322.1	Dinoflagellata	Dinophyceae	*Scrippsiella sweeneyae*	0	0	7
GQ501325.1	Dinoflagellata	Dinophyceae	*Scrippsiella trochoidea*	0	0	144
GQ501400.1	Dinoflagellata	Dinophyceae	*Thoracosphaera heimii*	0	0	1,215
GQ501756.1	Dinoflagellata	Dinophyceae	Uncultured *dinoflagellate*	913	0	22,512
GQ501539.1	Dinoflagellata	Dinophyceae	Uncultured *Peridinium*	0	0	168

(Continued)

TABLE 1 | Continued

GenBank ID	Phylum	Class	Species	0.2 μ mesh size	0.8 μ mesh size	10 μ mesh size
GQ501810.1	Dinoflagellata	Dinophyceae	Uncultured *Prorocentrum*	0	0	6
GQ501436.1	Dinoflagellata	Dinophyceae	Uncultured *Thecadinium*	0	0	421
JN809389.1	Rotifera	Monogononta	*Testudinella clypeata*	0	0	1,402
			Total per filter	21,282	2,324	74,676
			Total			98,282

cornuta (5.8%). Other two polychaetes, *Capitella capitata* and *Hediste diversicolor*, and another barnacle (*Sacculina carcini*) were represented by less reads: 3, 4, and 46, respectively. One rotifer accounting for 5% of the reads was identified: *Testudinella clypeata*. Diatoms were represented by only two species, *Haslea crucigera* and *Cylindrotheca fusiformis*, being together <0.04% of the reads. The copepod *Acartia tonsa* was also found at very low abundance in number of reads, which is only one read (**Table 1**).

One half of the species detected from metabarcoding in the analyzed lagoon can be considered nuisance species or indicators of bad environmental quality[4] (**Table 2**). The three polychaetes are bioindicators of pollution. The polychaete *P. cornuta* and the copepod *A. tonsa* are considered invasive species in the Mediterranean Sea. Finally, the dinoflagellate *Akashiwo sanguinea* forms algal blooms (**Table 2**).

Moreover, seven of the dinoflagellate genera are known to contain numerous species that are considered HABs because they produce toxins and/or form algal blooms, being cataloged in the HAB list of Intergovernmental Oceanographic Commission (IOC)-Unesco (Moestrup et al., 2009). According to the Harmful Algal Events Database (HAEDAT)[4], the genera *Alexandrium* (13 species) and *Dynophysis* (10 species) found in our study (**Table 1**) are responsible for most algal blooms in the area. The other five genera of dinoflagellates found in our study and listed as HABs were *Karenia* (10 species), *Prorocentrum* (12 species), *Karlodinium* (6 species), *Protoceratium* (1 species), and *Azadinium* (3 species) (**Table 1**).

DISCUSSION

The present study, based on a single metabarcode and 6 L of surface water from only one sampling point from one lagoon of 600 ha, did reveal bioindicators of three different environmental or biosecurity threats in Canet-Saint Nazaire lagoon: pollution (three bioindicator species), biological invasions (two species), and harmful algal blooms (one species). Despite extreme methodological simplicity, the resource-efficient methodology employed in this study was capable of obtaining important environmental information. Although it is known that DNA remains in the environment for prolonged periods especially in cold conditions and that individuals may not be present at the time of sampling, these results confirm the utility of COI as an eDNA metabarcoding tool for early alert of biological risks as proposed by Borrell et al. (2017) for biological invasions in ports, expanding the applications to pollution assessments and HABs.

The results indicate a poor ecological status for the lagoon Canet-Saint Nazaire. The three polychaete species were found to have been previously used as pollution indicators due to their ability to survive in highly polluted waters, with wide ranges of temperature and salinity, hypoxia, and eutrophication states (Surugiu, 2005; Dean, 2008; Maranho et al., 2014). *P. cornuta* is also considered an invasive species in the Mediterranean due to its ability to change the composition and abundance of native species, replacing native species like *C. capitata* (Cinar et al., 2005; Çinar et al., 2012; Zenetos et al., 2012). It was first discovered in the Mediterranean from polluted sediments in Valencia's harbor (Spain) (Tena et al., 1991); however, it now has a wider distribution. Its larval phase can survive transport in ship ballast water, and adults may form dense settlements through fouling of ship hulls facilitating its spread (Radashevsky and Selifonova, 2013). On the other hand, the invasive *F. enigmaticus* that was found in a previous survey in this lagoon (Ardura and Planes, 2017) could have facilitated the introduction and establishment of *P. cornuta* in the lagoon as reported in other locations (Read and Gordon, 1991; Heiman et al., 2008).

Acartia tonsa deserves additional attention, since it is a neritic copepod commonly found in the western Atlantic, the Pacific, and the Indian Ocean, especially in estuarine and coastal waters. It was first described in the Mediterranean in 1985 (Gaudy and Viñas, 1985) and has been previously reported in the Gulf of Lion in Berre Lagoon (Cervetto et al., 1999). It is considered an invasive species due to its ability to replace native species of copepods and its propensity to spread to new areas through ballast water (David et al., 2007).

On the other hand, despite having a single sampling point with only 6 L of water and forming only a limited description of the diversity of species in the whole ecosystem; the data set shows a great diversity of dinoflagellates in the sampling point analyzed. Most reads (89.5%) and OTUs (60.7%) detected were assigned to the class Dinophycea. This could be explained by the

[4]http://haedat.iode.org/ accessed on August 2019

TABLE 2 | Assessment of environmental threats from the species found in Canet-Saint Nazaire lagoon.

Species	Environmental threat	References
Polydora cornuta	Invasive species, pollution bioindicator	Dean, 2008; Zenetos et al., 2012
Hediste diversicolor	Pollution bioindicator	Dean, 2008
Capitella capitata	Pollution bioindicator	Dean, 2008
Acartia tonsa	Invasive species	David et al., 2007
Akashiwo sanguinea	Harmful algal blooms	Jones et al., 2017

sampling protocol, since the samples were taken from the lagoon surface where a greater number of these organism are expected. In any case, this is a signal of the lagoon's poor ecological state and high eutrophication and nitrogen and phosphorus levels that facilitate the proliferation of these organisms (Carlier et al., 2008; Heisler et al., 2008). Moreover, several dinoflagellates can produce resistance cysts, allowing them to survive during harsher periods in the lagoon (Sellner et al., 2003). However, the lack of variability in the COI region used for the species identification does not allow differences between species or strains (native or not) from the same genus to be resolved, making it impossible to detect species that can cause harm to the ecosystem. Several species of the genera detected (e.g., *Alexandrium*, *Dinophysis*, *Karenia*, or *Prorocentrum*) are known to cause HABs that could harm the lagoon ecosystem (Sellner et al., 2003). *A. sanguinea* (only species in the genus *Akashiwo*) blooms have been reported to produce powerful surfactant-like proteins that can coat bird plumage and collapse their feathers, causing loss of waterproofing and thermal insulation and even mass mortality (Jones et al., 2017). This could have serious effects in the lagoon because it is located along one of Europe's principal migration routes, and since March 2006, it has been classified as a SPA (Ardura and Planes, 2017). Further analysis with more specific identification tools would be necessary to assess the risks to the lagoon, but the data obtained in this study should be taken as warning signal. In Thau Lagoon, 100 km away from Canet, a toxic strain from *Alexandrium catenella* originating in the West Pacific was probably introduced from ballast water released in Séte's harbor that is connected to the lagoon (Lilly et al., 2002).

The current study has some limitations, likely because the sampling was limited to only one point and 6 L of water taken from the surface; thus, benthic organisms are unlikely detected. One limitation could be a relatively modest number of reads recovered (0.5 million reads). The presence of inhibitors due to the polluted state of the lagoon cannot be discarded as an additional explanation for this, since pollution can inhibit the PCR amplification of eDNA (Jane et al., 2015). The lack of detection of some invasive macroscopic species detected by Ardura and Planes (2017) in the lagoon (the bivalves *Abra* sp. and *Cerastoderma glaucum* and the tubeworm *Ficopomatus enigmaticus*) could be explained by the different sampling methods employed in the two studies. The species *C. glaucum* occurs in the bottom and is very scarce in the lagoon (in the previous survey, more sampling points were analyzed and only one individual for was obtained). *Abra* sp. and *F. enigmaticus*, which were abundant in the lagoon, were not seen where the water samples were taken in the present study; thus, it is possible that their DNA was simply absent from our water samples. Additional explanations could be very degraded DNA (Barnes et al., 2014) and/or lack of universality of the primers that may not anneal equally well in all taxa (Wilcox et al., 2013). In Ardura and Planes (2017), two different barcodes were amplified, e.g., COI (Geller et al., 2013) and 16S rRNA (Palumbi, 1996), but high-quality sequences for *F. enigmaticus* and *Abra* sp. were obtained only for the 16S rRNA marker, while in the present study, only COI primers were employed.

The number of detected OTUs in metabarcoding samples tends to be higher than in the visual and barcoding analysis, and effectively, the number of species detected here was much higher than in the previous work by Ardura and Planes (2017). False positives in metabarcoding due to incorrect assignations during the bioinformatic analysis (Ficetola et al., 2015) could be reasonably discarded since all the OTUs detected in this study were BLASTed manually to avoid incorrect assignations. On the other hand, it is possible to detect species that are not alive in the sampling area but whose DNA could come from the sea or from neighboring rivers (Pochon et al., 2017). If it was the case, they would still be around and be able to reach the lagoon eventually. In general, and despite possible errors and the obvious limitations of the simplified sampling scheme, the metabarcoding technique used in this study allowed the successful completion of a general assessment of the lagoon's environmental state from a list of nuisance and indicator species. Despite this, the data should be considered qualitative rather than empirical. Although eDNA concentration and number of sequences yielded from NGS are positively correlated with biomass or population density, the accurate conversion of genome abundances to cell numbers and estimates of absolute abundance are still imprecise (Kelly et al., 2014; Bonk et al., 2018). A greater number of reads assigned to a specific OTU do not necessarily imply a greater abundance (Thomas et al., 2016).

This study provides new results from a new case study that support the applicability of the COI barcode as a metabarcoding tool using eDNA (e.g., Thomsen and Willerslev, 2015; Zaiko et al., 2015; Borrell et al., 2017; Pochon et al., 2017; Forster et al., 2019), showing in this particular case its usefulness to detect HAB-causing organisms. However, although it is useful to assign some organisms at species level, such as the genus Akashiwo, in most cases, this assignment is not possible because this small fragment, 300 bp, is not informative enough.

The simple sampling scheme in this study, with only one-point sample and three replicates, provides a proof of concept for the use of single samples as an approximation of ecological status. It is important to remark that the study trials a simplified sampling method to remark the value that small sample numbers can have on remote locations for future studies: even one sample with three replicates is informative. However, the limitations associated to collect small sample numbers should not be forgotten, bringing attention to considerations that must be had when analyzing NGS data collected from one sample. In these cases, the data must be analyzed as proof of concept. Further research with additional samples (at different depths and situational periods) would be necessary to build a more complete survey of the study area. Moreover, larger COI fragment or different molecular markers and the enlargement and improvement of barcode databases (Ardura, 2018) are also necessary and will provide a better representation of the ecosystem's biodiversity.

On the other hand, new methodological tools are been developed to improve biological surveys, such as the use of environmental RNA (eRNA) (Pochon et al., 2017). Compared to eDNA, eRNA degrades more rapidly in

the marine environment (typically hours to days, Thomsen et al., 2012; Sassoubre et al., 2016) and is therefore considered a better proxy for detecting living biota. However, susceptibility of RNA makes it difficult to work with. The collection of RNA samples requires dedicated sampling protocols, more careful preservation, and storage. There is also additional processing time and costs associated with isolation and reverse transcription of RNA (Laroche et al., 2016), making it more expensive and challenging and thus a less attractive molecule to work with (Zaiko et al., 2018).

As a final remark, the presence of invasive species, bioindicators of contamination, and organisms that can cause damage to the ecosystem in this single sampling point suggests a high level of degradation and vulnerability of the lagoon. This is a protected area under the Natura 2000 network, and the results may question the effectiveness of the current conservation programs, indicating that COI and metabarcoding can play an important role in monitoring the progress of conservation efforts.

DATA AVAILABILITY STATEMENT

The Fastaq files are available in GenBank, reference (PRJNA549294).

AUTHOR CONTRIBUTIONS

AA: conception and design. SP: sampling. MS-M and AA: laboratory analysis. MS-M: NGS analysis. EG-V and AA: statistical analysis. AA and EG-V: writing the article. Finally, all authors revised the manuscript.

FUNDING

This study has been supported by the Spanish Ministry of Economy and Competitiveness Grant CGL2016-79209-R and the project FUO-201-17 supported by the Centre National de la Recherche Scientifique: "eDNA in Mediterranean lagoons." This is a contribution from the Marine Observatory of Asturias (OMA), Asturias University Institute of Biotechnology (IUBA), and the Spanish Research Group of Excellence ARENA. AA holds a Juan de la Cierva-reincorporation Grant from the Spanish Ministry.

ACKNOWLEDGMENTS

The authors thank Martin Desmalades for helping with the sampling tasks.

REFERENCES

Anderson, D. M., Alpermann, T. J., Cembella, A. D., Collos, Y., Masseret, E., and Montresor, M. (2012). The globally distributed genus Alexandrium: multifaceted roles in marine ecosystems and impacts on human health. *Harmful Algae* 14, 10–35. doi: 10.1016/j.hal.2011.10.012

Anderson, D. M., Glibert, P. M., and Burkholder, J. M. (2002). Harmful algal blooms and eutrophication: nutrient sources, composition, and consequences. *Estuaries* 25, 704–726. doi: 10.1007/bf02804901

Ardura, A. (2018). Species-specific markers for early detection of marine invertebrate invaders through eDNA methods: gaps and priorities in GenBank as database example. *J. Nat. Conserv.* 47, 51–57. doi: 10.1016/j.jnc.2018.11.005

Ardura, A., and Planes, S. (2017). Rapid assessment of non-indigenous species in the era of the eDNA barcoding: a mediterranean case study. *Estuar.Coast. Shelf Sci.* 188, 81–87. doi: 10.1016/j.ecss.2017.02.004

Ardura, A., Zaiko, A., Borrell, Y., Samuiloviene, A., and Garcia-Vazquez, E. (2016). Novel tools for early detection of a global aquatic invasive, the zebra mussel Dreissena polymorpha. *Aquat. Conserv.* 27, 165–176. doi: 10.1002/aqc.2655

Barnes, M. A., Turner, C. R., Jerde, C. L., Renshaw, M. A., Chadderton, W. L., and Lodge, D. M. (2014). Environmental conditions influence eDNA persistence in aquatic systems. *Environ. Sci. Technol.* 48, 1819–1827. doi: 10.1021/es404734p

Barnes, R. S. K. (1980). *Coastal lagoons*, Vol. 1. Cambridge: CUP Archive.

Blackman, R. C., Constable, D., Hahn, C., Sheard, A., and Durkota, J. (2017). Detection of a new non-native freshwater species by DNA metabarcoding of environmental samples – first record of Gammarus fossarum in the UK. *Aquat. Invasion.* 12, 177–189. doi: 10.3391/ai.2017.12.2.06

Blanchet, S. (2012). The use of molecular tools in invasion biology: an emphasis on freshwater ecosystems. *Fish. Manag Ecol.* 19, 120–132. doi: 10.1111/j.1365-2400.2011.00832.x

Bonk, F., Popp, D., Harms, H., and Centler, F. (2018). PCR-based quantification of taxa-specific abundances in microbial communities: quantifying and avoiding common pitfalls. *J. Microbiol. Methods* 153, 139–147. doi: 10.1016/j.mimet.2018.09.015

Borrell, Y. J., Miralles, L., Do Huu, H., Mohammed-Geba, K., and Garcia-Vazquez, E. (2017). DNA in a bottle—Rapid metabarcoding survey for early alerts of invasive species in ports. *PLoS One* 12:e0183347. doi: 10.1371/journal.pone.0183347

Bravo, I., Fernández, M. L., Ramilo, I., and Martı nez, A. (2001). Toxin composition of the toxic dinoflagellate Prorocentrum lima isolated from different locations along the Galician coast (NW Spain). *Toxicon* 39, 1537–1545. doi: 10.1016/s0041-0101(01)00126-x

Bylemand, J., Gleeson, D. M., Duncan, R. P., Hardy, C. M., and Furlan, E. M. (2019). A performance evaluation of targeted eDNA and eDNA metabarcoding analyses for freshwater fishes. *Environ. DNA* 1, 402–414. doi: 10.1002/edn3.41

Caporaso, J. G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F. D., Costello, E. K., et al. (2010). QIIME allows analysis of high-throughput community sequencing data. *Nat. Methods* 7, 335–336. doi: 10.1038/nmeth.f.303

Carlier, A., Riera, P., Amouroux, J. M., Bodiou, J. Y., Desmalades, M., and Grémare, A. (2008). Food web structure of two Mediterranean lagoons under varying degree of eutrophication. *J. Sea Res.* 60, 264–275. doi: 10.1016/j.seares.2008.10.006

Cataudella, S., Crosetti, D., and Massa, F. (2015). "Mediterranean coastal lagoons: sustainable management and interactions among aquaculture, capture fisheries and the environment," in *General Fisheries Commission for the Mediterranean*. Rome: FAO.

Cervetto, G., Gaudy, R., and Pagano, M. (1999). Influence of salinity on the distribution of Acartia tonsa (Copepoda, Calanoida). *J. Exp. Mar. Biol Ecol.* 239, 33–45. doi: 10.1016/s0022-0981(99)00023-4

Chapman, P. M. (2012). Management of coastal lagoons under climate change. Estuarine, Coastal and Shelf Science, 110, 32-35. g data.*Nature Methods* 7:335.

Cinar, M. E., Ergen, Z., Dagli, E., and Petersen, M. E. (2005). Alien species of spionid polychaetes (Streblospio gynobranchiata and Polydora cornuta) in Izmir Bay, eastern Mediterranean. *J. Mar. Biol. Assoc. UK.* 85, 821–827. doi: 10.1017/s0025315405011768

Çinar, M. E., Katagan, T., Öztürk, B., Bakir, K., and Dagli, E. (2012). Spatio-temporal distributions of zoobenthos in soft substratum of Izmir Bay (Aegean Sea, eastern Mediterranean), with special emphasis on alien species and ecological quality status. *J. Mar. Biol. Assoc. UK* 92, 1457–1477. doi: 10.1017/s0025315412000264

Crooks, J. A., Chang, A. L., and Ruiz, G. M. (2011). Aquatic pollution increases the relative success of invasive species. *Biol. Invasions* 13, 165–176. doi: 10.1007/s10530-010-9799-3

David, M., Gollasch, S., Cabrini, M., Perkoviè, M., Bošnjak, D., and Virgilio, D. (2007). Results from the first ballast water sampling study in the Mediterranean Sea–the Port of Koper study. *Mar. Pollut. Bull.* 54, 53–65. doi: 10.1016/j.marpolbul.2006.08.041

Dean, H. K. (2008). The use of polychaetes (Annelida) as indicator species of marine pollution: a review. *Rev. Biol. Trop.* 56, 11–38.

Ficetola, G. F., Miaud, C., Pompanon, F., and Taberlet, P. (2008). Species detection using environmental DNA from water samples. *Biol. Lett.* 4, 423–425. doi: 10.1098/rsbl.2008.0118

Ficetola, G. F., Pansu, J., Bonin, A., Coissac, E., Giguet—Covex, C., De Barba, M., et al. (2015). Replication levels, false presences and the estimation of the presence/absence from eDNA metabarcoding data. *Mol. Ecol. Resour.* 15, 543–556. doi: 10.1111/1755-0998.12338

Forster, D., Filker, S., Kochems, R., Breiner, H.-W., Cordier, T., Pawlowski, J., et al. (2019). A comparison of different ciliate metabarcode genes as bioindicators for environmental impact assessments of salmon aquaculture. *Eukaryot. Microbiol.* 66, 294–308. doi: 10.1111/jeu.12670

Furlan, E. M., Gleeson, D., Hardy, C. M., and Duncan, R. I. P. (2015). A framework for estimating the sensitivity of eDNA surveys. *Mol. Ecol. Resour.* 16, 641–654. doi: 10.1111/1755-0998.12483

Galil, B. S. (2007). Loss or gain? Invasive aliens and biodiversity in the Mediterranean Sea. *Mar. Pollut. Bull.* 55, 314–322. doi: 10.1016/j.marpolbul.2006.11.008

Gaudy, R., and Viñas, M. D. (1985). Première signalisation en Méditerranée du copépode pélagique Acartia tonsa. *Rapp. Comm. Int. Mer Médit.* 29, 227–229.

Geller, J., Meyer, C., Parker, M., and Hawk, H. (2013). Redesign of PCR primers for mitochondrial cytochrome c oxidase subunit I for marine invertebrates and application in all—taxa biotic surveys. *Mol. Ecol. Resour.* 13, 851–861. doi: 10.1111/1755-0998.12138

Gozlan, R. E., Britton, J. R., Cowx, I., and Copp, G. H. (2010). Current knowledge on non—native freshwater fish introductions. *J Fish Biol.* 76, 751–786. doi: 10.1111/j.1095-8649.2010.02566.x

Hallegraeff, G. M. (1993). A review of harmful algal blooms and their apparent global increase. *Phycologia* 32, 79–99. doi: 10.1186/1476-069X-7-S2-S4

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

Hebert, P. D., Ratnasingham, S., and de Waard, J. R. (2003b). Barcoding animal life: cytochrome c oxidase subunit 1 divergences among closely related species. *Proc. R. Soc. Lon. B Biol. Sci.* 270(Suppl. 1), S96–S99.

Heiman, K. W., Vidargas, N., and Micheli, F. (2008). Non-native habitat as home for non-native species: comparison of communities associated with invasive tubeworm and native oyster reefs. *Aquati. Biol.* 2, 47–56. doi: 10.3354/ab00034

Heisler, J., Glibert, P. M., Burkholder, J. M., Anderson, D. M., Cochlan, W., Dennison, W. C., et al. (2008). Eutrophication and harmful algal blooms: a scientific consensus. *Harmful Algae* 8, 3–13. doi: 10.1016/j.hal.2008.08.006

Illumina, (2013). "16S metagenomic sequencing library preparation," in *Preparing 16S Ribosomal RNA Gene Amplicons for the Illumina MiSeq System*, (Illumina, CA: Illumina), 1–28.

Jane, S. F., Wilcox, T. M., McKelvey, K. S., Young, M. K., Schwartz, M. K., Lowe, W. H., et al. (2015). Distance, flow and PCR inhibition: eDNA dynamics in two headwater streams. *Mol. Ecol. Resour.* 15, 216–227. doi: 10.1111/1755-0998.12285

Jones, T., Parrish, J. K., Punt, A. E., Trainer, V. L., Kudela, R., Lang, J., et al. (2017). Mass mortality of marine birds in the Northeast Pacific caused by Akashiwo sanguinea. *Mar. Ecol. Progr. Ser.* 579, 111–127. doi: 10.3354/meps12253

Katsanevakis, S., Coll, M., Piroddi, C., Steenbeek, J., Ben Rais Lasram, F., Zenetos, A., et al. (2014). Invading the mediterranean sea: biodiversity patterns shaped by human activities. *Front. Mar. Sci.* 1:32. doi: 10.3389/fmars.2014.00032

Kelly, R. P., Port, J. A., Yamahara, K. M., and Crowder, L. B. (2014). Using environmental DNA to census marine fishes in a large mesocosm. *PLoS One* 9:e86175. doi: 10.1371/journal.pone.0086175.t001

Kjerfve, B. (ed.) (1994). *Coastal lagoon processes*, Vol. 60. Berlin: Elsevier.

Laroche, O., Wood, S. A., Tremblay, L. A., Ellis, J. I., Lejzerowicz, F., Pawlowski, J., et al. (2016). First evaluation of foraminiferal metabarcoding for monitoring environmental impact from an offshore oil drilling site. *Mar. Environ. Res.* 120, 225–235. doi: 10.1016/j.marenvres.2016.08.009

Leray, M., Yang, J. Y., Meyer, C. P., Mills, S. C., Agudelo, N., Ranwez, V., et al. (2013). A new versatile primer set targeting a short fragment of the mitochondrial COI region for metabarcoding metazoan diversity: application for characterizing coral reef fish gut contents. *Front. Zool.* 10:34. doi: 10.1186/1742-9994-10-34

Lilly, E. L., Kulis, D. M., Gentien, P., and Anderson, D. M. (2002). Paralytic shellfish poisoning toxins in France linked to a human-introduced strain of Alexandrium catenella from the western Pacific: evidence from DNA and toxin analysis. *J. Plankton Res.* 24, 443–452. doi: 10.1093/plankt/24.5.443

Lin, S., Zhang, H., Hou, Y., Zhuang, Y., and Miranda, L. (2009). High-level diversity of dinoflagellates in the natural environment, revealed by assessment of mitochondrial cox1 and cob genes for dinoflagellate DNA barcoding. *Appl. Environ. Microbiol.* 75, 1279–1290. doi: 10.1128/AEM.01578-08

Maranho, L. A., Baena-Nogueras, R. M., Lara-Martín, P. A., DelValls, T. A., and Martín-Díaz, M. L. (2014). Bioavailability, oxidative stress, neurotoxicity and genotoxicity of pharmaceuticals bound to marine sediments. The use of the polychaete Hediste diversicolor as bioindicator species. *Environ. Res.* 134, 353–365. doi: 10.1016/j.envres.2014.08.014

Moestrup, Ø, Akselmann, R., Fraga, S., Hoppenrath, M., Iwataki, M., Komárek, J., et al. (eds) (2009). *IOC-UNESCO Taxonomic Reference List of Harmful Micro Algae*. Avaliable at: http://www.marinespecies.org/hab (accessed December 16, 2018).

Moore, S. K., Trainer, V. L., Mantua, N. J., Parker, M. S., Laws, E. A., Backer, L. C., et al. (2008). Impacts of climate variability and future climate change on harmful algal blooms and human health. *Environ. Health* 7 (Suppl. 2):S4.

Nilsson, R. H., Tedersoo, L., Abarenkov, K., Ryberg, M., Kristiansson, E., Hartmann, M., et al. (2012). Five simple guidelines for establishing basic authenticity and reliability of newly generated fungal ITS sequences. *MycoKeys* 4, 37–63. doi: 10.3897/mycokeys.4.3606

Oksanen, J., Blanchet, F., Kindt, R., Legendre, P., Minchin, R., O'Hara, R., et al. (2013). *Vegan: Community Ecology Package. R Package Version. 2.0–10*.

Paavola, M., Olenin, S., and Leppäkoski, E. (2005). Are invasive species most successful in habitats of low native species richness across European brackish water seas? *Estuar.Coast. Shelf Sci.* 64, 738–750. doi: 10.1016/j.ecss.2005.03.021

Palumbi, S. R. (1996). "Nucleic acids II: the polymerase chain reaction," in *Molecular Systematics*, eds D. M. Hillis, C. Moritz, and B. K. Mable (Sunderland: Sinauer Associates, Inc.), 205–247.

Pérez-Ruzafa, A., Marcos, C., and Pérez-Ruzafa, I. M. (2011). Mediterranean coastal lagoons in an ecosystem and aquatic resources management context. *Phys,. Chem. Earth Parts A/B/C* 36, 160–166. doi: 10.1016/j.pce.2010.04.013

Pochon, X., Zaiko, A., Fletcher, L. M., Laroche, O., and Wood, S. A. (2017). Wanted dead or alive? Using metabarcoding of environmental DNA and RNA to distinguish living assemblages for biosecurity applications. *PLoS One* 12:e0187636. doi: 10.1371/journal.pone.0187636

Radashevsky, V. I., and Selifonova, Z. P. (2013). Records of Polydora cornuta and Streblospio gynobranchiata (Annelida, Spionidae) from the Black Sea. *Mediterr. Mar. Sci.* 14, 261–269.

Read, G. B., and Gordon, D. P. (1991). Adventive occurrence of the fouling serpulid Ficopomatus enigmaticus (Polychaeta) in New Zealand. *N. Z. J Mar. Freshwa. Res.* 25, 269–273. doi: 10.1080/00288330.1991.9516478

Reizopoulou, S., Thessalou-Legaki, M., and Nicolaidou, A. (1996). Assessment of distubance in Mediterranean lagoons: an evaluation of methods. *Mar. Biol.* 125, 189–197. doi: 10.1007/bf00350773

Ruiz, G. M., Fofonoff, P. W., Carlton, J. T., Wonham, M. J., and Hines, A. H. (2000). Invasion of coastal marine communities in North America: apparent patterns, processes, and biases. *Annu. Rev. Ecol. Syst.* 31, 481–531. doi: 10.1146/annurev.ecolsys.31.1.481

Sassoubre, L. M., Yamahara, K. M., Gardner, L. D., Block, B. A., and Boehm, A. B. (2016). Quantification of environmental DNA (eDNA) shedding and decay rates for three marine fish. *Environ. Sci. Technol.* 50, 10456–10464. doi: 10.1021/acs.est.6b03114

Scott, R., Zhan, A., Brown, E. A., Chain, F. J. J., Cristescu, M. E., Gras, R., et al. (2018). Optimization and performance testing of a sequence processing pipeline applied to detection of nonindigenous species. *Evol. Appl.* 11, 891–905. doi: 10.1111/eva.12604

Sellner, K. G., Doucette, G. J., and Kirkpatrick, G. J. (2003). Harmful algal blooms: causes, impacts and detection. *J. Ind. Microbiol. Biotechnol.* 30, 383–406. doi: 10.1007/s10295-003-0074-9

Souchu, P., Bec, B., Smith, V. H., Laugier, T., Fiandrino, A., Benau, L., et al. (2010). Patterns in nutrient limitation and chlorophyll a along an anthropogenic eutrophication gradient in French Mediterranean coastal lagoons. *Can. J. Fish Aquat. Sci.* 67, 743–753. doi: 10.1139/f10-018

Surugiu, V. (2005). The use of polychaetes as indicators of eutrophication and organic enrichment of coastal waters: a study case–Romanian Black Sea Coast. Analele Ştiinţifice ale Universităţii "Al.I. Cuza". Iaşi, s. *Biol. Anim.* 51, 55–62.

Taberlet, P., Coissac, E., Pompanon, F., Brochmann, C., and Willerslev, E. (2012). Towards next−generation biodiversity assessment using DNA metabarcoding. *Mol. Ecol.* 21, 2045–2050. doi: 10.1111/j.1365-294X.2012.05470.x

Tena, J., Capaccioni-Azzati, R., Porras, R., and Torres-Gavilá, F. J. (1991). Cuatro especies de poliquetos nuevas para las costas mediterráneas españolas en los sedimentos del antepuerto de Valencia. *Miscel· Zool.* 15, 29–41.

Thomas, A. C., Deagle, B. E., Eveson, J. P., Harsch, C. H., and Trites, A. W. (2016). Quantitative DNA metabarcoding: improved estimates of species proportional biomass using correction factors derived from control material. *Mol. Ecol. Resour.* 16, 714–726. doi: 10.1111/1755-0998.12490

Thomsen, P. F., and Willerslev, E. (2015). Environmental DNA – An emerging tool in conservation for monitoring past and present biodiversity. *Biol. Conserv.* 183, 4–18. doi: 10.1016/j.biocon.2014.11.019

Thomsen, P. T., Kielgast, J., Iversen, L., Møller, P. R., Rasmussen, M., and Willerslev, E. (2012). Detection of a diverse marine fish fauna using environmental DNA from seawater samples. *PLoS One* 7:e41732. doi: 10.1371/journal.pone.0041732

von Ammon, U., Wood, S. A., Larocge, O., Zaiko, A., Tait, L., Lavery, S., et al. (2018). Combining morpho-taxonomy and metabarcoding enhances the detection of non-indigenous marine pests in biofouling communities. *Sci. Rep.* 8, 16290. doi: 10.1038/s41598-018-34541-1

Vouvé, F., Buscail, R., Aubert, D., Labadie, P., Chevreuil, M., Canal, C., et al. (2014). Bages-Sigean and Canet-St Nazaire lagoons (France): physico-chemical characteristics and contaminant concentrations (Cu, Cd, PCBs and PBDEs) as environmental quality of water and sediment. *Environ. Sci. Pollut. Res.* 21, 3005–3020. doi: 10.1007/s11356-013-2229-1

Weigand, A., Beermann, A. J., Èiampor, F., Costa, F. O., Csabai, Z., Duarte, S., et al. (2019). DNA barcode reference libraries for the monitoring of aquatic biota in Europe: Gap-analysis and recommendations for future work. *Sci. Total Environ.* 678, 499–524. doi: 10.1016/j.scitotenv.2019.04.247

Wilcox, T. M., McKelvey, K. S., Young, M. K., Jane, S. F., Lowe, W. H., Whiteley, A. R., et al. (2013). Robust detection of rare species using environmental DNA: the importance of primer specificity. *PLoS One* 8:e59520. doi: 10.1371/journal.pone.0059520

Yasumoto, T., Oshima, Y., Fukuyo, Y., Oguri, H., Igarashi, T., and Fujita, N. (1980). Identification of Dinophysis fortii as the causative organism of diarrhetic shellfish poisoning. *Bull. Japan. Soc. Sci. Fish.* 46, 1405–1411. doi: 10.2331/suisan.46.1405

Zaiko, A., Pochon, X., Garcia-Vazquez, E., Olenin, S., and Wood, S. A. (2018). Advantages and limitations of environmental DNA/RNA tools for marine biosecurity: management and surveillance of non-indigenous species. *Front. Mar. Sci.* 5:322. doi: 10.3389/fmars.2018.00322

Zaiko, A., Samuiloviene, A., Ardura, A., and Garcia-Vazquez, E. (2015). Metabarcoding approach for nonindigenous species surveillance in marine coastal waters. *Mar. Pollut. Bull.* 100, 53–59. doi: 10.1016/j.marpolbul.2015.09.030

Zenetos, A., Gofas, S., Morri, C., Rosso, A., Violanti, D., García Raso, J. E., et al. (2012). Alien species in the mediterranean Sea by 2012," in a contribution to the application of european union's marine strategy framework directive (MSFD), Part2. introduction trends and pathways. *Mediterr. Mar. Sci.* 13, 328–352. doi: 10.12681/mms.327

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.

frontiers in Ecology and Evolution

REVIEW
published: 06 April 2020
doi: 10.3389/fevo.2020.00057

Sixteen Years of DNA Barcoding in China: What Has Been Done? What Can Be Done?

Cai-qing Yang†, Qing Lv† and Ai-bing Zhang‡*

College of Life Sciences, Capital Normal University, Beijing, China

OPEN ACCESS

Edited by:
Rodney L. Honeycutt, Pepperdine University, United States

Reviewed by:
Zhi Chao, Southern Medical University, China
Karen Leanne Bell, The University of Western Australia, Australia
Xiaohui Pang, Chinese Academy of Medical Sciences, China
Paul Z. Goldstein, United States Department of Agriculture, United States

***Correspondence:**
Ai-bing Zhang
zhangab2008@cnu.edu.cn; zhangab2008@mail.cnu.edu.cn

†These authors have contributed equally to this work

‡ORCID:
Ai-bing Zhang
orcid.org/0000-0003-3450-5421

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

Received: *12 November 2019*
Accepted: *26 February 2020*
Published: *06 April 2020*

Citation:
Yang C, Lv Q and Zhang A (2020) Sixteen Years of DNA Barcoding in China: What Has Been Done? What Can Be Done? Front. Ecol. Evol. 8:57. doi: 10.3389/fevo.2020.00057

Over the past 16 years, more than half (59.68%) of research papers in China on DNA barcoding have been published in Chinese rather than English. Using the records in the BOLD (Barcode of Life Data) system, we found Chinese scientists have contributed nearly 120,000 DNA barcodes for more than 16,000 species as of September 2019, with barcoded species distributed throughout China. Based on 2,624 articles and 494 dissertations published during the last 16 years, we reviewed the basic statistics of these studies as well as the type of articles contributed by Chinese scientists, the preference of taxonomic groups, the characteristic of barcoding studies in China, the current limitations, and potential future directions as well. We found that most barcode data pertain primarily to plants and animals. Most work in China has focused on verification of the authenticity of species used in traditional Chinese medicine, while other applications have paid more attention to food safety, inspection and quarantine, and the control of pests and invasive species. In methodology and technology, a number of new DNA barcoding methods have been developed by Chinese scientists. However, there are several significant limitations to research into DNA barcoding in China in general, such as the lack of leadership in pioneering international projects, the absence of an open bioinformatics infrastructure, and the fact that some Chinese journals do not clearly require data transparency and availability for DNA barcodes, impeding the further development of barcode libraries and research in China. In the future, Chinese scientists should build authoritative online libraries, while aiming for theoretical innovations for both concepts and methodology of DNA barcoding.

Keywords: DNA barcode, sequence assignment, COI, ITS, matK, BOLD

INTRODUCTION

Since the inception of DNA barcoding in 2003 (Hebert et al., 2003a,b), it has become widely used as a taxonomic tool (DeSalle and Goldstein, 2019). It is especially useful for species identification when accurate morphological information and taxonomic expertise are limiting factors (Ahrens et al., 2007; Valentini et al., 2009). With additional development and

methodologies, barcoding is becoming increasingly useful outside of taxonomy (Hebert and Gregory, 2005), and it is becoming more popular in ecological (e.g., ecological interactions and food webs) studies, biodiversity surveys (Hajibabaei et al., 2007; Joly et al., 2013), conservation biology, biosecurity, and medicine and pharmacology (Pečnikar and Buzan, 2014).

In China, some taxonomists, such as those who work on plants, are deeply involved in the study of barcoding, providing many significant contributions to the international community of DNA barcoding. For example, 62 researchers from 19 research institutes and universities across the country have formed the "China Plant BOL (Barcode of Life) Group" to conduct in-depth research on the DNA barcoding of seed plants. Based on the barcode combinations recommended by the Consortium for Barcode of Life (CBOL), they proposed that ITS/ITS2 should be incorporated into the core barcode for seed plants after conducting a large number of tests on four DNA barcode candidate fragments of 6,286 specimens (China-Plant-BOL-Group et al., 2011). Their research not only solved the problem of low resolution using only *rbcL* + *matK* but also represented another step forward toward standardizing the routine use of DNA sequence data (Hollingsworth, 2011). Besides DNA barcoding of plants, other Chinese scientists have applied different DNA barcodes in their own taxonomic groups (Cheng et al., 2011). However, a systematic review on DNA barcoding research in China is lacking, especially in an international context, given the relative inaccessibility of this language to those who cannot read Chinese.

To this end, we systematically searched for articles published by Chinese scientists in both domestic and international journals from 2003 to August 2019 and summarized the contributions of Chinese scientists in DNA barcoding research in terms of their publications and data outputs. We have also pointed out severe limitations and potential future directions for barcoding research in China.

LITERATURE SEARCHING AND MANUAL DATA MINING

According to the ecological theory of species–area relations (Arrhenius, 1921; Gleason, 1922), countries with large land areas theoretically possess higher biodiversity. With a land area of more than 9.63 million square kilometers, China is the third largest country in the world. In the context of DNA barcoding, the Chinese scientific community is responsible for documenting an immense wealth of biodiversity and corresponding barcode sequences. To gauge the amount of barcode data generated and shared by China, the current number of records and related species were retrieved with the keyword "China" (incl. Taiwan) from the BOLD system (The Barcode of Life Data[1]; Ratnasingham and Hebert, 2007), which is one of the world's most authoritative online barcode databases. The coordinates of those records were also downloaded to visualize their geographic distribution at the same time. Barcode data from the five other largest countries (excluding China), Russia, Canada, America, Brazil, and Australia were also downloaded for comparison.

To determine the proportion of the publications on DNA barcoding from Chinese scientists worldwide, a preliminary retrieval from the Web of Science (WOS[1]) database with the phrase "DNA barcode*" (the asterisk was used to enable the return of results containing the words "barcode," "barcodes" or "barcoding") as the keyword was implemented. To make the results more general, we searched for publications where the keyword appeared throughout the full text of articles (with "topic" field tag in WOS) rather than just in title, which is slightly different from previous reviews (Taylor and Harris, 2012; DeSalle and Goldstein, 2019).

A final database was then assembled. To review the DNA barcoding studies contributed by Chinese scientists during the last 16 years, a comprehensive literature search was conducted from not only WOS[1] but also China National Knowledge Infrastructure (CNKI[2]) for articles published during the period between January 2003 and August 2019. The latter database was generally ignored in previous studies by western researchers due to language issues. We searched for "DNA barcode*" in the full text of the paper, with Chinese institutions/universities as the first research institute (**Supplementary Data Sheet S1**). Because of the partial overlap between these two online databases, we manually removed the duplicative records for subsequent analyses. Then, to summarize the problems and potential directions of DNA barcoding research in China in the future, information of each publication was listed, covering taxonomic groups, article types, journals, barcode selections, and research institutions.

[1]http://isiknowledge.com

[2]https://www.cnki.net/

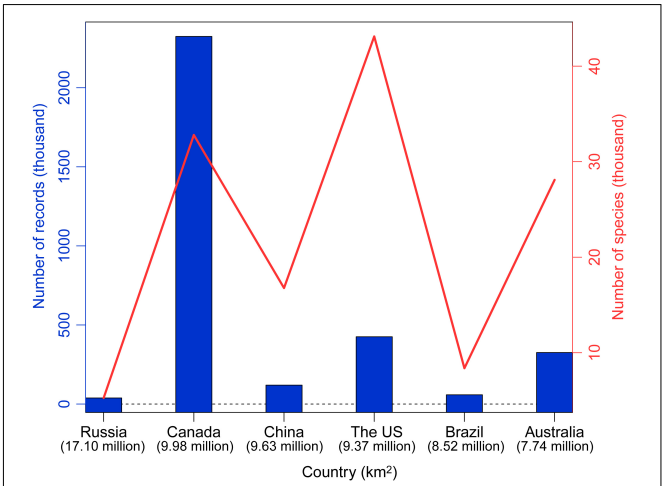


FIGURE 1 | Numbers of barcoding records contributed by scientists from the six largest countries in the world. Countries are given on the *X* axis, and the *Y* axis indicates the number of barcode records (blue bar) and number of species (red solid line). Country area is given under the name.

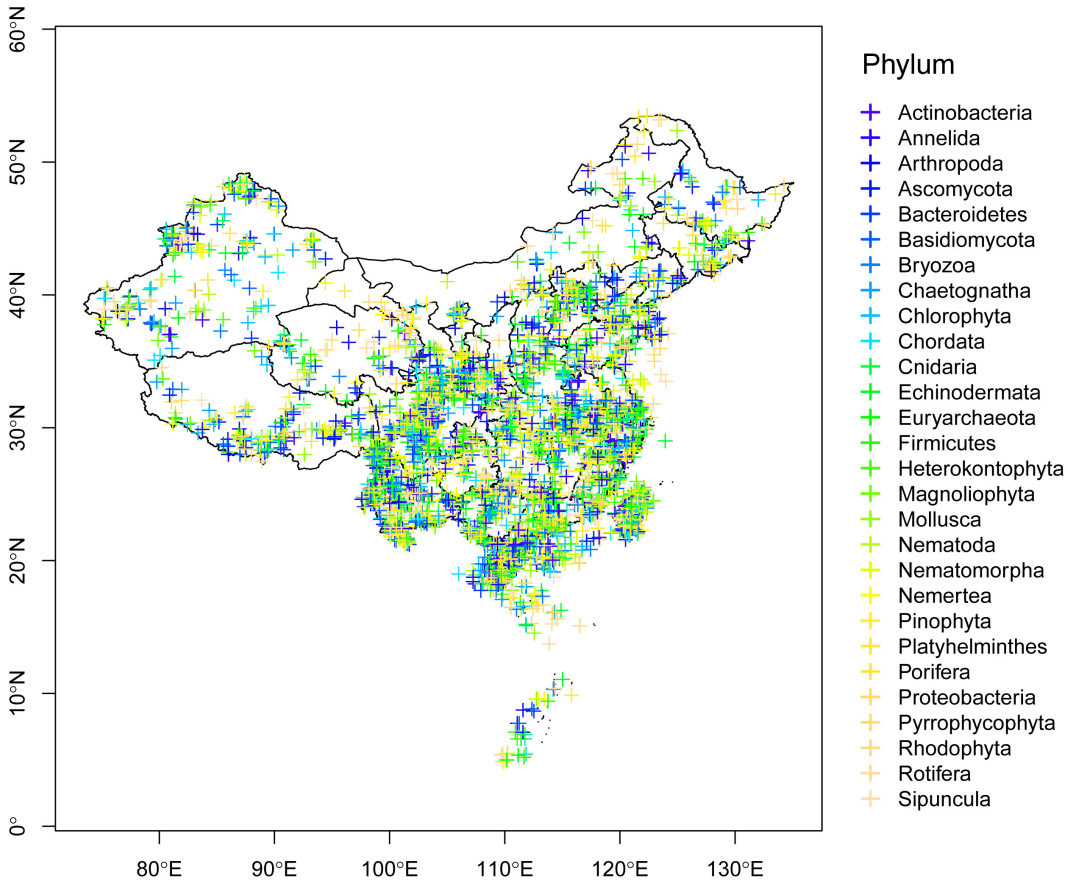


FIGURE 2 | Geographical distribution of barcodes in the BOLD system contributed by Chinese scientists, with samples collected in China. Phyla are denoted by different colors.

DNA BARCODING AND ITS CURRENT SITUATION IN CHINA

As the third largest country in the world, China possesses one of 25 global biodiversity hotspots (Myers et al., 2000). Based on a survey of the global DNA barcoding library BOLD system (The Barcode of Life Database[3]; Ratnasingham and Hebert, 2007), Chinese scientists have contributed 119,745 DNA barcodes belonging to 16,772 species as of September 2019 (**Figure 1**). Of the six largest countries examined, the only countries that have contributed more barcodes are Australia, Canada, and the United States. Geographically, studies have taken place throughout much of China, although fewer have been conducted in the northwest and northeast (**Figure 2**).

According to the data from WOS[1], 1,993 articles from China (incl. Taiwan) that were published between 2003 and 2018 include the phrase "DNA barcode*" in their "topic" field tag. Following a review on barcoding published in a domestic journal in 2004, Chinese scientists started publishing their DNA barcoding research in international journals in 2006, and the number of articles began to increase in 2009. By the end of 2018, the total number of publications on DNA barcoding contributed from Chinese researchers reached 20.06% of the total number of DNA barcoding papers published throughout the world (**Figure 3**), indicating that China has become one of the major countries dedicated to research on DNA barcoding. However, this is only part of China's contribution to DNA barcoding because more than half (59.68%) of their publications occur in internal Chinese journals (most are not databased in WOS).

PUBLICATIONS CONTRIBUTED BY CHINESE SCIENTISTS

More Empirical but Fewer Methodological Studies

In this study, all 2,624 articles were classified into four categories: Category 1 – basic studies, where one or more DNA barcodes are established for specific taxonomic groups; Category 2 – practical studies, where DNA barcodes are used to identify species or other ecologically related research; Category 3 – methodological studies, where new algorithms or methods of species identification are developed, computer programs are

[3]http://www.boldsystems.org

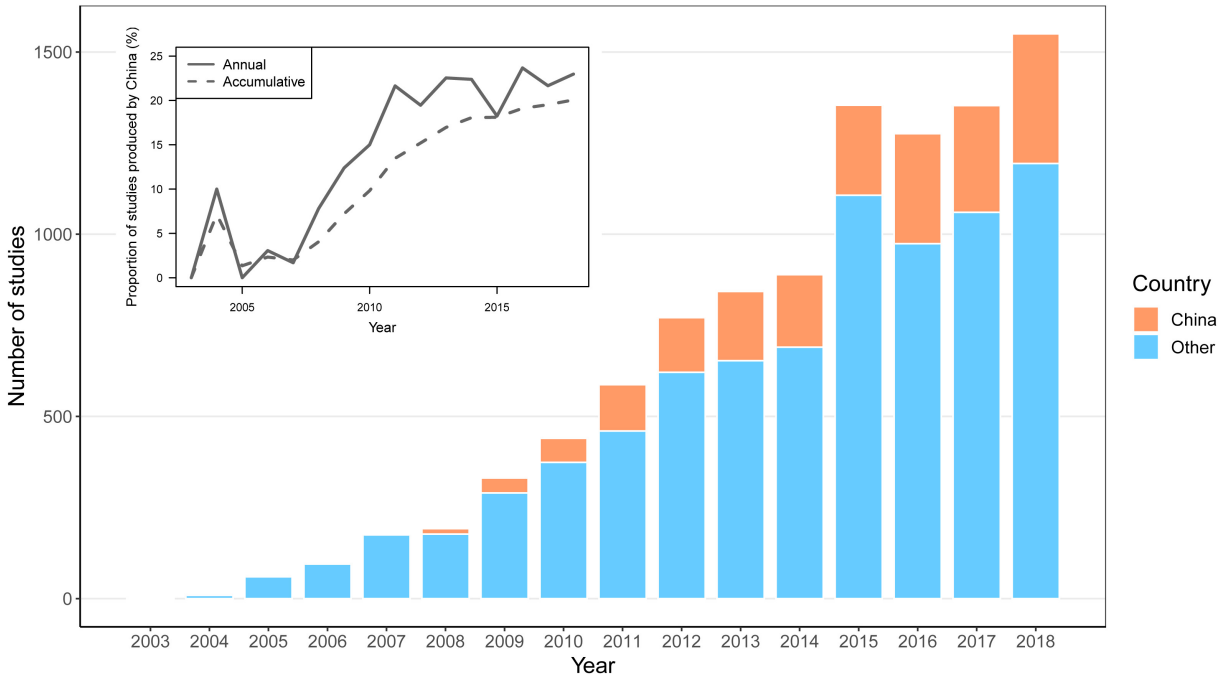


FIGURE 3 | Numbers of articles published on DNA barcoding by Chinese scientists and researchers from all other countries between 2003 and 2018. The *X* axis gives the year of publication, and the *Y* axis indicates the number of studies. Orange bars represent the number of articles published by Chinese scientists, while blue bars indicate the number of articles published by other researchers from all other countries. The small figure in the upper left corner represents the proportion of the number of articles from Chinese studies to the total number of articles, where the solid line represents the proportion of articles contributed by Chinese scientists to all others each year, and the dotted line represents the proportion of accumulated articles to all others.

established, or comparisons are made between different DNA barcoding approaches; and Category 4 – reviews that summarize recent advances in DNA barcoding, including those focusing on certain groups of taxa.

Based on the statistics derived from these different types of articles, we found that the number of articles pertaining to Category 1 showed an annual increase, and, by August 2019, they represented nearly half (47.14%) of the total number of articles published (**Figure 4**); Category 2 showed a similar trend, with 28.24% of articles, and this result implies that China has a huge demand for DNA barcoding technology, including demand from traditional Chinese medicine and social needs related to food safety, inspection and quarantine, pest control, and other applications (see below).

In contrast to Category 1 and 2, the number of articles pertaining to Category 3 has increased at a much slower rate (**Figure 4**). Methodological studies, accounting for 10.98% of the total studies, are considerably less common than practical studies (28.24%). Despite the small percentage of methodological studies, they may have comprehensive and profound effects on other DNA barcoding studies. Therefore, we have paid more attention to them here. In this category, internationally, one of the earliest algorithms for sequence assignment was the BLAST algorithm (Altschul et al., 1990), which relies on local similarity between sequences. However, the credibility of the assignment results can be questionable in DNA barcoding (Ross et al., 2008). Most tree-based methods, such as maximum parsimony (MP; Czelusniak et al., 1990), maximum-likelihood (ML; Felsenstein, 1981), and Bayesian approaches (Huelsenbeck and Ronquist, 2001; Munch et al., 2008), are probably more accurate, but they usually require long processing times and high-RAM (random access memory) when dealing with very large DNA datasets (Austerlitz et al., 2009) except neighbor joining (NJ; Saitou and Nei, 1987). Chinese scientists have used these approaches in their DNA barcoding studies. The last decade, however, has also witnessed significant progress in the methodology of DNA barcoding given many new approaches proposed by Chinese scientists (Zhang et al., 2008, 2012a,b, 2017; Yu et al., 2012; Liu et al., 2013, 2017; Jin et al., 2018; Shi et al., 2018). The main advances include both algorithm development and the optimization of sequencing strategies, as summarized below.

Artificial intelligence (AI) is used for industrialized applications in China, and Chinese scientists appear to be among the first to introduce AI into species identification algorithms (Zhang et al., 2008). The proposed method is used for identification of species with unknown barcodes based on referencing library trained back-propagation (BP) neural networks. The BP-based method appears to be superior to commonly used distance-based methods, particularly in cases involving incomplete lineage sorting (Zhang et al., 2008). Species identification algorithms for non-coding barcode sequences based on machine learning methods, such as DV-RBF and FJ-RBF, also performed well (Zhang et al., 2012a). The

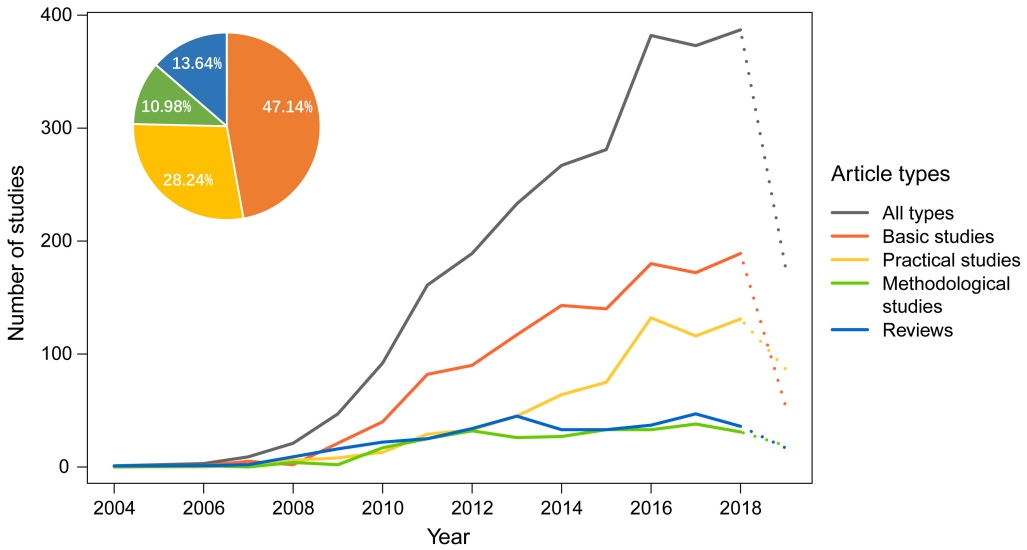


FIGURE 4 | The number of different types of articles on DNA barcoding published by Chinese scientists each year. The *X* axis represents the year of publication, and the *Y* axis represents the number of studies. A complete dataset for 2019 is not yet available and is shown as a dotted line. The pie chart at the top left represents the proportion of the total number of research articles of each type derived over 16 years.

problem of species membership can also be solved by linking it to fuzzy-set-theory (Zadeh, 1965), which efficacy has been demonstrated by its successful application to empirical datasets (Zhang et al., 2012b). Compared with other methods, the fuzzy-set-theory-based approach has great efficacy in reducing false-positive species identification when conspecifics of the query are absent from the reference database (Zhang et al., 2012b). In addition, Shi et al. (2018) combined the Hidden Markov Model (HMM; Eddy, 1998) algorithm with the fuzzy membership function and further improved the processing speed of this approach for exploring large datasets. Naturally, the expanding number of available methods begets a need for an integrated toolkit for DNA barcoding. BarcodingR is one of the most useful software packages that provides a comprehensive implementation of species identification methods with additional new functions in R (Zhang et al., 2017). With the great facility of this package for DNA barcoding research, the high performance of machine learning approaches has been successfully applied in studies, such as wood barcoding (He et al., 2019).

Aside from analysis algorithms, in the optimization of sequencing strategies, scientists are also developing more efficient means of obtaining accurate metadata. Yu et al. (2012) proposed protocols for the extraction of ecological, taxonomic, and phylogenetic information from bulk samples by combining mass trapping, mass-PCR amplification, pyrosequencing, and bioinformatics analysis. They demonstrated that metabarcoding allows for a broad and efficient estimate of biodiversity for the first time, which can facilitate assessment of the state of current ecosystems worldwide. One problem with barcodes derived from next-generation-sequencing (NGS) analyses is the shorter maximum read lengths (typically $<$ 150 bp) and consequent lost taxonomic information. To overcome this problem, Liu et al. (2013) presented a new Illumina-based pipeline (SOAPBarcode) that allows for the full-length recovery of COI barcodes from mixed samples. Their assemblage protocol involves the use of two libraries: the full-length library (insert size = 658 bp) and the shotgun library (insert size = 200 bp). This approach can deliver reliable and taxonomically informative metabarcoding outcomes for biodiversity-related research (Liu et al., 2013). Although the introduction and optimization of metabarcoding has applications for biodiversity studies, the most accurate approach for taxonomists is to obtain the complete barcode sequence by amplification from a single sample. Because Sanger sequencing is approaching its limits in terms of throughput and chemistry cost, Liu et al. (2017) developed an Illumina-based pipeline, HIFI-Barcode, to produce full-length COI barcodes from pooled PCR amplicons generated by individual specimens. The accuracy of barcode sequences generated by the new pipeline is comparable to sequences derived from the Sanger method and only requires about one-tenth of the current cost (Liu et al., 2017).

The ever-increasing number of DNA barcoding methods has led to many reviews on the subject. The number of reviews accounted for 13.64% of all articles. The first review of DNA barcoding was published in 2004 (Xiao et al., 2004), and it was the first to introduce Chinese scientists to the concept, basic principles, and potentials of DNA barcoding. The increase in the number of reviews came after 2010. Many papers summarized the application and methods of barcoding technology in different taxonomic groups (e.g., Cheng et al., 2011; Yao et al., 2013; Chen et al., 2014; Liang et al., 2015; Sun et al., 2016). Lately, some researchers have also reviewed DNA barcoding from the perspective of ecological communities, and they have proposed a "purpose-driven barcode" fit for multi-level applications (Pei et al., 2017).

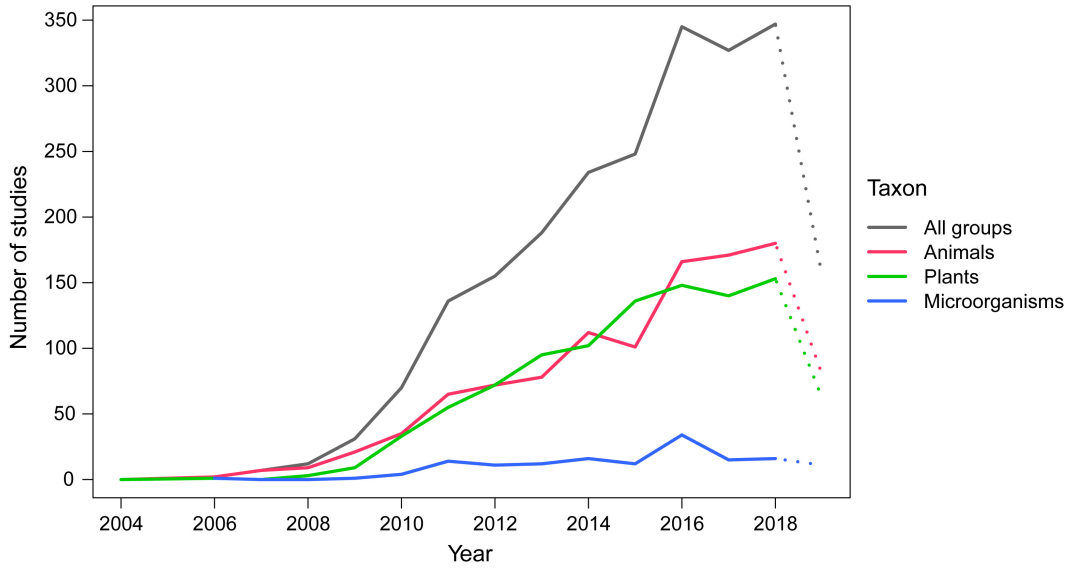


FIGURE 5 | The number of DNA barcoding studies by Chinese scientists published over different years on different taxonomic groups. The *X* axis represents the year of publication, and the *Y* axis represents the number of studies. A complete dataset for 2019 is not yet available and is shown as a dotted line. The count of articles on different taxonomic groups is not completely mutually exclusive – some articles involve two or more taxonomic groups, and these articles were used to estimate the statistics for each taxonomic group respectively. Therefore, the total number of articles is not a simple sum of the number of articles of the three groups.

The Vast Majority of Species Barcoded in China: Animals and Plants

As originally proposed in 2003, DNA barcoding largely focused on species of animals (Hebert et al., 2003a), thus indicating a taxonomic bias in that other groups were less studied (Taylor and Harris, 2012). This trend continues in China (**Figure 5**). As of August 2019, the total number of articles related to animal groups in China reached 1,104, nearly half (48.72%) of the total number of research papers (2,266, excluding review articles). Likewise, plant barcoding studies showed a trend of continuous and rapid increase similar to that of animal groups after 2009 (**Figure 5**). The rapid growth of DNA barcoding research on plant groups is probably related to Chinese traditional medicine culture (see below). At the same time, Chinese researchers have paid less attention to DNA barcoding of microorganisms. As of August 2019, only 147 research papers on other groups were published, and most are related to the classification and identification of fungi as well as viruses and pathogens (**Figure 5**).

Internal Publication Chinese Barcoding Research

In order to present the contributions made by Chinese scientists to the worldwide efforts focused on DNA barcoding, we compared the 20 journals where Chinese scientists published their research most frequently over the last 16 years. As shown in **Figure 6**, two thirds of the publications were in Chinese journals. More than half of these domestic journals pertained to traditional Chinese medicine, indicating the great need of DNA barcoding technology for medically related studies. These types of studies are more likely to be of use to Chinese researchers than a global audience, so Chinese journals may be the most appropriate. The English journals PLoS One, Zootaxa, Scientific Reports, Molecular Ecology Resources, ZooKeys, Mitochondrial DNA Part A, Systematics and Evolution, and Ecology and Evolution comprised 46.2% of the publications contributed by Chinese scientists (**Figure 6**) and a majority of what could be considered systematics, evolution, ecology, and biodiversity studies. One potential benefit from publishing the research in Chinese journals is that access to the research is locally available, thus enhancing more general use of barcode data. One drawback from publishing primarily in Chinese journals is that contributions made by scientists from China are inaccessible to scientists from other countries. Therefore, the contributions made by Chinese scientists are underappreciated, but data transparency is less acute in the ecology and evolution literature rather than that in medical or pharmaceutical publications.

DNA BARCODING-RELATED RESEARCH AREAS IN CHINA

Species Identification and Diversity

DNA barcoding was firstly proposed to simplify the taxonomic identification of species by providing an efficient and accurate method that did not require taxonomic expertise (Hebert and Gregory, 2005). Based on the prevalence of specific "keywords" in articles published by Chinese scientists, the current application of DNA barcoding in China is primarily for species identification (**Figure 7**). More recently, the application of DNA barcodes for species identification has matured, and researchers have turned from the exploration

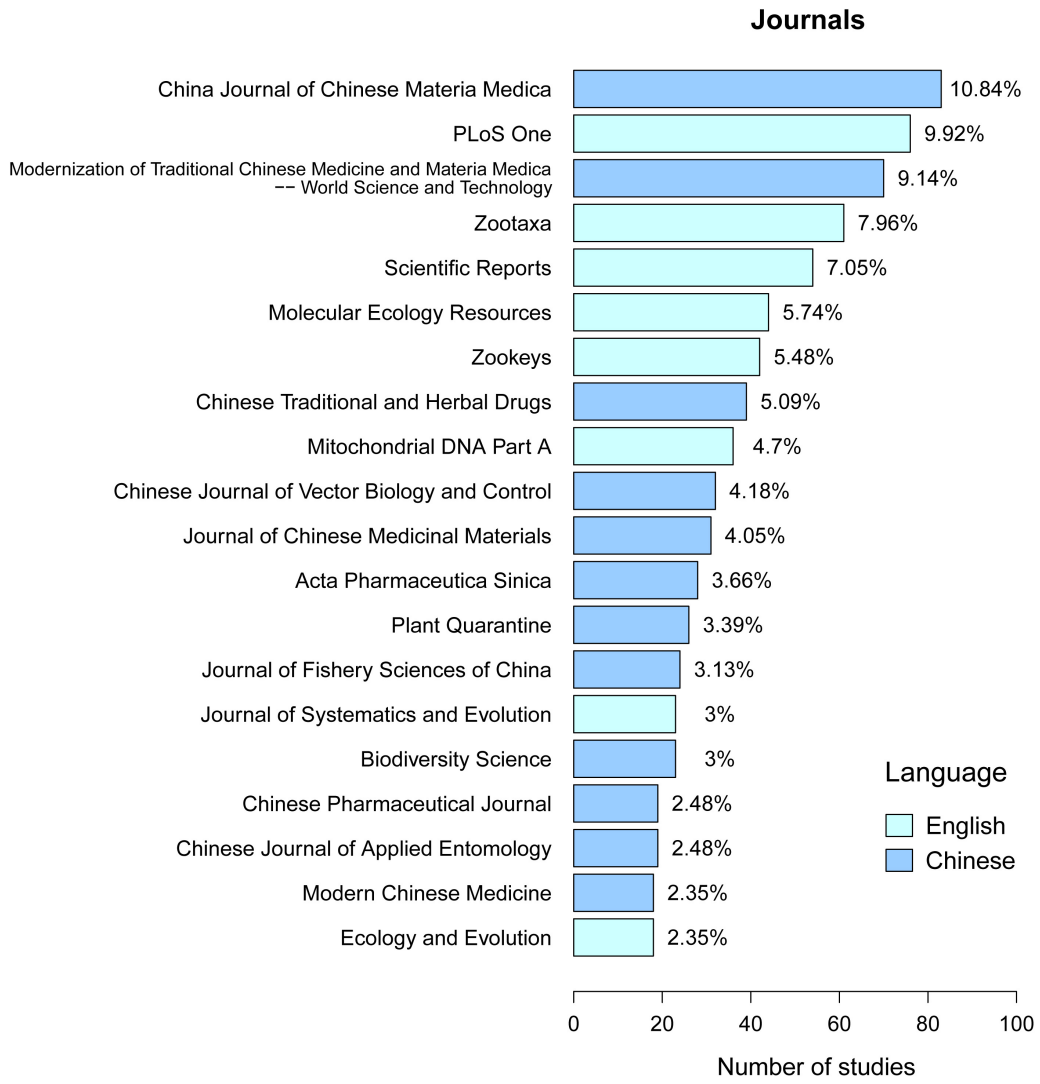


FIGURE 6 | Top 20 journals that published DNA barcoding studies contributed by Chinese scientists. The *X* axis represents the number of studies, and the *Y* axis represents journal names. Percentages represent the number of articles published by each journal as a percentage of the total number of the top 20 journals.

and verification of barcode technology to the applications and solutions of practical problems in the taxonomic groups they specialize in.

In addition to species identification, Chinese scientists are using DNA barcodes in phylogenetics (e.g., Feng et al., 2016; Liu et al., 2016; Chesters, 2017), the discovery of new or cryptic species (e.g., Liu et al., 2011a,b; Qin et al., 2018), and the evaluation of the levels of biodiversity (e.g., Chen et al., 2015; Chesters et al., 2015; Li et al., 2017; Sun et al., 2018; **Figure 7**).

Based on the statistics of keywords with the top 20 highest frequency in different literatures, 1.57% of the Chinese barcode research pertains to herbal medicine and 2.09% for identification of adulterants (identification of fraudulent products) in Chinese herbal medicine (**Figure 7**). Therefore, the emergence of DNA barcoding technology has indeed proven important for research on Chinese traditional medicine.

Standard DNA Barcodes for Plant Groups Need Further Exploration

Ideally, DNA barcodes should at least satisfy the following criteria: (1) specificity – the DNA fragment must be nearly identical in the same species but different between different species; (2) uniformity – the section must be standardized (the same section should be used in different taxonomic groups); and (3) robustness – the marker must have conservative primer binding sites that allow it to be amplified and sequenced from a large number of groups (Pečnikar and Buzan, 2014). Despite years of effort to find universal DNA barcodes for different taxonomic groups, people have to admit that searching for a universal barcode for all species is utopian. The top five most commonly used barcodes by Chinese scientists for their own taxonomic groups are listed in **Figure 8**. It was found that COI was used in nearly all studies involving the barcoding of animal

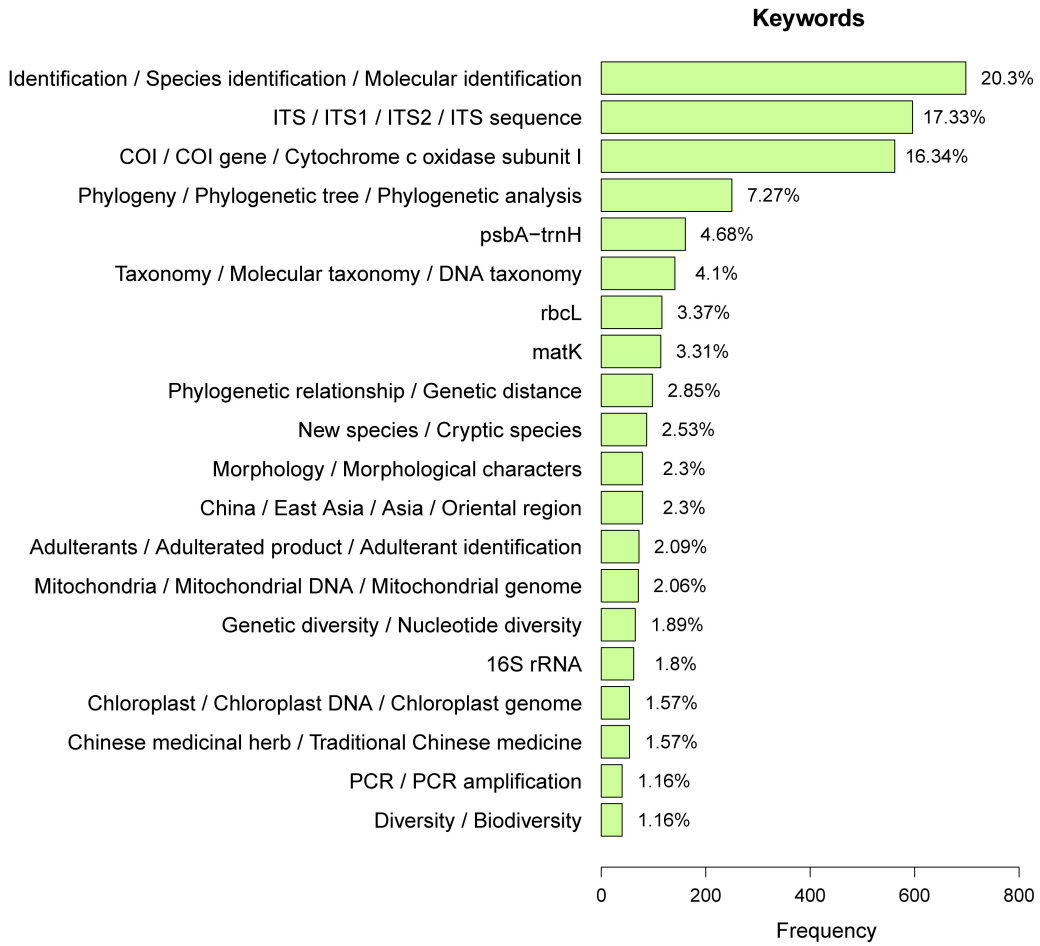


FIGURE 7 | Frequency of key words in DNA barcoding studies published by Chinese researchers. The *X* axis represents the frequency of various keywords in the search results of all articles, and the *Y* axis represents similar categories of keywords. Percentages represent the ratio of the frequency of each keyword to the total number of terms in the top 20. There are no terms such as "DNA barcoding," "DNA barcode," "DNA barcodes," and "barcoding" presented here because they are obviously the most frequent ones.

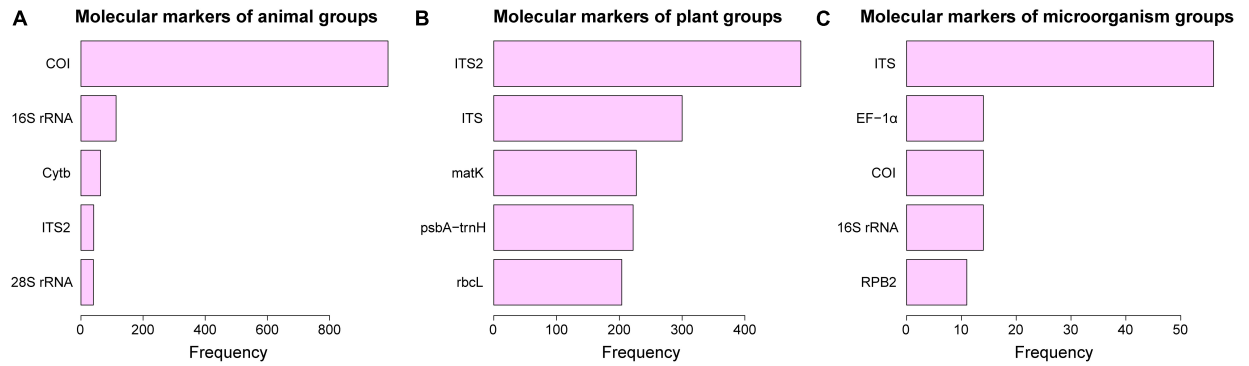


FIGURE 8 | Frequency of DNA barcodes articles pertaining to animals, plants, and microorganisms. **(A)** animal, **(B)** plant, and **(C)** microorganism groups. The *X* axis represents the frequency of each molecular marker used in the article, and the *Y* axis represents the name of the gene used as a molecular marker.

groups in China (**Figure 8A**), indicating that the COI region has been consistently important for the general use of DNA barcodes of animal groups due to the fact that COI barcodes perform excellently in most animal groups (e.g., Hebert et al., 2003a,b; Rougerie et al., 2009; Steinke et al., 2009). Although other markers, such as 16S rRNA, *Cytb*, ITS2, etc., have also

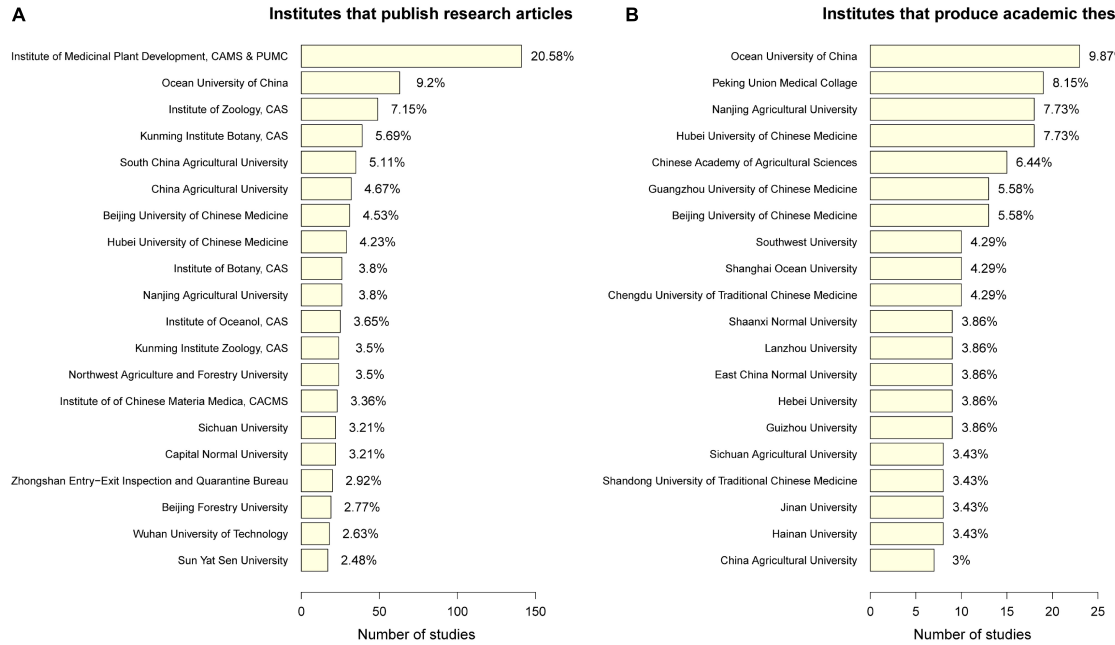


FIGURE 9 | Top 20 Chinese institutions that published **(A)** research articles and **(B)** academic dissertations in China. The *X* axis represents the number of studies, and the *Y* axis represents the name of the institution. The abbreviations are as follows: CAMS & PUMC – Chinese Academy of Medical Sciences & Peking Union Medical College; CAS – Chinese Academy of Sciences; and CACMS – China Academy of Chinese Medical Sciences.

been used in some studies of animal groups, they were co-analyzed with COI in most cases (e.g., Li et al., 2010; Jin et al., 2018; Huang et al., 2019). Similarly, ITS genes are the most commonly used molecular markers (**Figure 8C**) in studies that focus on microorganisms, while other genes are used relatively infrequently and are generally used as auxiliary barcodes.

However, in plant groups, the most frequently used molecular markers are not as obvious (**Figure 8B**). ITS2 and ITS are the most widely used markers in **Figure 8B**, which were proposed as novel barcodes for medicinal plants by Chen et al. (2010) and were suggested to be incorporated into the core barcode for seed plants by China-Plant-BOL-Group et al. (2011). *MatK*, *psbA-trnH*, and *rbcL* are high-frequency candidate barcodes for plants as well, which may be related to the joint use of multiple plant barcodes in most studies (e.g., Yang et al., 2012; Jin et al., 2014; Gong et al., 2016; Bao et al., 2018).

In fact, a large part of the studies on plant barcodes in China are carried out on Chinese medicinal herbs, and the barcodes selected for these studies are often different. For example, Li et al. (2014) identified the herbal medicinal materials from *Aristolochia* using the *matK*, *rbcL*, *psbA-trnH*, and *trnL-trnF* DNA regions. Guo et al. (2017) identified the herbal materials from *Cynanchum* using the ITS2 barcode; Gong et al. (2018) constructed a DNA barcode reference library for "Nan Yao" (crude drugs mainly produced in or imported through tropical and subtropical China, especially the Lingnan region, i.e., the territories south of the Nanling Mountains) using ITS2; and Jiao et al. (2018) identified the medicinal Polygonati Rhizoma (a traditional medicinal and edible product with *Polygonatum* polysaccharides, saponins, phenols, and flavonoids) efficiently and accurately using ITS2 and *psbA-trnH* sequences. This shows that the selection of molecular markers for plant groups in China still relies heavily on the combination of multiple markers.

PRIMARY RESEARCH INSTITUTIONS ON DNA BARCODING

The Institute of Chinese Medicine Science and Marine Biology: Dominant Institutions Focusing on DNA Barcoding Research in China

As shown in **Figure 9A**, the top five Chinese institutions with the largest number of articles published on DNA barcodes include (in order of the most to fewest publications) the Institute of Medicinal Plant Development (Chinese Academy of Medical Sciences & Peking Union Medical College), the Ocean University of China, the Institute of Zoology (Chinese Academy of Sciences), the Kunming Institute of Botany (Chinese Academy of Sciences), and the South China Agricultural University. Research at these institutions mainly focuses on traditional Chinese medicine, marine organisms, and other animals and plants.

Comparatively, **Figure 9B** lists the top 20 universities or research institutions that have contributed the highest proportion of 494 dissertations related to DNA barcoding. The Ocean University of China has produced the most master's and doctoral dissertations, followed by Peking Union Medical Collage, Nanjing Agricultural University, Hubei University Chinese Medicine, and the Chinese Academy of Agricultural Sciences.

These dissertations focused primarily on marine organisms and traditional Chinese medicine. Together, these figures reveal which institutions have pioneered barcoding research in China.

CONCLUSION AND FUTURE PERSPECTIVES

Based on our analyses, the contribution over the last 16 years by Chinese scientists to research using DNA barcoding is underappreciated, primarily because of the bias in that over half of relevant articles were published in the Chinese rather than the international literature. In terms of the overall number of barcode entries, Chinese scientists have produced a considerable amount of information on plants and animals (**Figure 9**), and the amount of data is close to that produced by several other leading countries in the field (**Figure 1**). Yet some DNA barcode sequences are not totally publicly available due to Chinese journals not clearly requiring data transparency and accessibility for DNA barcodes. In addition to barcode information on a variety of species, Chinese scientists are involved in the development of new barcoding methods as well as the analysis of barcode data from a large amount of sequencing information.

During the inception of barcoding, research in China was less well developed than the rest of the world, but it has, since 2009, witnessed a rapid growth (**Figure 3**). This growth of DNA barcoding in China is continuously expanding from medicinal plants to including other plants and animals, but the primary focus is still on medically and economically important species in need of identification. Additionally, the application of barcode technology is expanding, with studies related to phylogenetics, population genetics, and biodiversity becoming more common.

There are several potential research directions for Chinese scientists:

(1). Developing integrated evolutionary and/or ecological projects implementing DNA barcoding. We must admit that most current barcoding studies in China represent follow-up research and lack conceptual originality. The main important concepts and initiatives of DNA barcoding were not proposed by Chinese scientists in general (Pei et al., 2017). Studies with barcode data that appear in western journals where data transparency is required are often concerned with solving important ecological and evolutionary problems. However, China has the funding for – and satisfies the conditions of – the development of comprehensive research projects and promotion of theoretical innovation. In China, there is still a lot of unsurveyed biodiversity, from rainforests to deserts, where both taxonomists and evolutionary biologists could conduct investigations via DNA barcoding. This technology may also be applied to studies on macroevolution, interactions and food webs, environmental monitoring (Valentini et al., 2009; Garlapati et al., 2019). To maximize the value of DNA barcoding data, the people who collect it must collaborate with ecologists and evolutionary biologists (Joly et al., 2013; Cristescu, 2014) to expand the usefulness of barcode data. In the process, Chinese scientists have the opportunity to come up with their own new ideas and approaches to barcoding by developing integrated evolutionary and/or ecological projects implementing DNA barcoding.

(2). Proposing new approaches and *de novo* assigning algorithms for NGS related DNA barcoding. The concept of metabarcoding (Taberlet et al., 2012) has greatly expanded the potential scope of applications of DNA barcoding in recent years. A few scientists from China have published important papers on metabarcoding (e.g., Yu et al., 2012; Zhou et al., 2013; Liu et al., 2017; Lang et al., 2019), showing great potential in this field. As DNA barcoding technology matures, we think Chinese scientists should make more contributions in metabarcoding. Currently, fewer methodological studies are optimizing sequencing procedures or proposing new assignment algorithms to better address the challenges of the big data era (Coissac et al., 2016). The need for biodiversity-related research also poses new challenges for barcode bioinformatics analysis (Taberlet et al., 2012; Wang et al., 2019). For example, neither PCR-based nor PCR-free metabarcoding protocol allows the accurate estimation of species abundance (Braukmann et al., 2019), several barriers are still exist in metabarcoding when solving quantitative ecological issues. As each method has its shortcomings in certain contexts (Paz and Crawford, 2012), no perfect DNA barcoding method has been proposed for all cases (Li et al., 2013). The direction of multi-gene, multi-method, and multi-discipline combinations will become a primary focus in the future (Yang et al., 2018), and that is why there is so much space for the development of methodological advances, given the high demand for biodiversity research in China.

(3). Constructing a national-level DNA barcoding reference library. This has also been suggested by some other scientists (Pei et al., 2017). Although there are a few local barcoding libraries constructed for specific taxa (e.g., Hou et al., 2017; Gong et al., 2018; Liu et al., 2018), few leading and international DNA barcoding libraries have been created or have been hosted by Chinese scientists. Chen et al. (2014) established and continually maintain an online DNA barcoding database for herbal materials[4] with 78,847 barcode records belonging to 23,262 species, which shows the possibility of constructing national-level DNA barcode sequence libraries in China. Based on such efforts to build a foundation for barcoding, China can achieve far more toward documenting its immense biodiversity (Xu et al., 2015; Pei et al., 2017).

(4). Integrating into global research by making their DNA barcode data available to global barcoding research

[4]http://www.tcmbarcode.cn

communities. Some Chinese journals do not clearly require authors to submit their DNA barcodes to a publicly available database (e.g., submission to GenBank), rendering these DNA barcodes invisible to the broader scientific community, impeding DNA barcoding research both globally and in China. Together with help from the global scientist community, Chinese scientists must further their efforts to close the gap with their international counterparts, especially in data standardization and disclosure. With the efforts made by the biodiversity committee of the Chinese Academy of Sciences since 2013, GBIF (Global Biodiversity Information Facility) has made a Chinese portal[5]. If a Chinese edition of GenBank can be established, as proposed in (3), and be accessible to the researchers all over the world, submitting the data (including but not only DNA barcodes) to the library should be equivalent to submitting to GenBank. Chinese and overseas researchers are to be encouraged to submit data to both of them simultaneously before publishing their works. Currently, the National Genomics Data Center[6] may be the most appropriate candidate for a Chinese DNA barcode repository.

AUTHOR CONTRIBUTIONS

AZ designed the study. CY performed the research. CY and QL analyzed the data. CY and AZ wrote the first draft of the manuscript. All authors contributed substantially to revisions.

[5] http://www.gbifchina.org/

[6] http://bigd.big.ac.cn/

FUNDING

This work was supported by China National Funds for Distinguished Young Scientists (Grant Number 31425023), the Natural Science Foundation of China (Grant Number 31772501), Support Project of High-level Teachers in Beijing Municipal Universities (Grant Number IDHT20180518), and Academy for Multidisciplinary Studies, Capital Normal University.

ACKNOWLEDGMENTS

The authors are grateful to Dr. Michael C. Orr and the editor, Dr. RH, whose comments have considerably improved the manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fevo.2020.00057/full#supplementary-material

DATA SHEET S1 | Summary of articles with Chinese institutions/universities as the first research institute during the period between 2003 and August 2019.

REFERENCES

Ahrens, D., Monaghan, M. T., and Vogler, A. P. (2007). DNA-based taxonomy for associating adults and larvae in multi-species assemblages of chafers (Coleoptera: Scarabaeidae). *Mol. Phylogenet. Evol.* 44, 436–449. doi: 10.1016/j.ympev.2007.02.024

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.* 215, 403–410.

Arrhenius, O. (1921). Species and area. *J. Ecol.* 9, 95–99. doi: 10.2307/2255763

Austerlitz, F., David, O., Schaeffer, B., Bleakley, K., Olteanu, M., Leblois, R., et al. (2009). DNA barcode analysis: a comparison of phylogenetic and statistical classification methods. *BMC Bioinformatics* 10:S10. doi: 10.1186/1471-2105-10-S14-S10

Bao, W., Li, D., and Li, X. (2018). DNA barcoding of *Actinidia* (Actinidiaceae) using internal transcribed spacer, *matK*, *rbcL* and *trnH-psbA*, and its taxonomic implication. *N. Z. J. Bot* 56, 360–371. doi: 10.1080/0028825X.2018.1491009

Braukmann, T. W., Ivanova, N. V., Prosser, S. W., Elvrecht, V., Steinke, D., Ratnasingham, S., et al. (2019). Metabarcoding a diverse arthropod mock community. *Mol. Ecol. Resour.* 19, 711–727. doi: 10.1111/1755-0998.13008

Chen, S., Pang, X., Song, J., Shi, L., Yao, H., Han, J., et al. (2014). A renaissance in herbal medicine identification: from morphology to DNA. *Biotechnol. Adv.* 32, 1237–1244. doi: 10.1016/j.biotechadv.2014.07.004

Chen, S., Yao, H., Han, J., Liu, C., Song, J., Shi, L., et al. (2010). Validation of the ITS2 region as a novel DNA barcode for identifying medicinal plant species. *PLoS One* 5:e8613. doi: 10.1371/journal.pone.0008613

Chen, W., Ma, X., Shen, Y., Mao, Y., and He, S. (2015). The fish diversity in the upper reaches of the Salween River, Nujiang River, revealed by DNA barcoding. *Sci. Rep.* 5:17437. doi: 10.1038/srep17437

Cheng, X., Wang, A., Gu, Z., Wang, Y., Zhan, X., and Shi, Y. (2011). Current progress of DNA barcoding. *Genom. Appl. Biol.* 30, 748–758. doi: 10.3969/gab.030.000748

Chesters, D. (2017). Construction of a species-level tree of life for the insects and utility in taxonomic profiling. *Syst. Biol.* 66, 426–439. doi: 10.1093/sysbio/syw099

Chesters, D., Zheng, W. M., and Zhu, C. D. (2015). A DNA barcoding system integrating multigene sequence data. *Methods Ecol. Evol.* 6, 930–937. doi: 10.1111/2041-210X.12366

China Plant Bol Group, Li, D. Z., Gao, L. M., Li, H. T., Wang, H., Ge, X. J., et al. (2011). Comparative analysis of a large dataset indicates that internal transcribed spacer (ITS) should be incorporated into the core barcode for seed plants. *Proc. Natl. Acad. Sci. U.S.A.* 108, 19641–19646. doi: 10.1073/pnas.1104551108

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

Cristescu, M. E. (2014). From barcoding single individuals to metabarcoding biological communities: towards an integrative approach to the study of global biodiversity. *Trends Ecol. Evol.* 29, 566–571. doi: 10.1016/j.tree.2014.08.001

Czelusniak, J., Goodman, M., Moncrief, N. D., and Kehoe, S. M. (1990). Maximum parsimony approach to construction of evolutionary trees from aligned homologous sequences. *Methods Enzymol.* 183, 601–615. doi: 10.1016/0076-6879(90)83039-C

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

Eddy, S. R. (1998). Profile hidden Markov models. *Bioinformatics* 14, 755–763. doi: 10.1093/bioinformatics/14.9.755

Felsenstein, J. (1981). Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol.* 17, 368–376. doi: 10.1007/BF01734359

Feng, S., Jiang, M., Shi, Y., Jiao, K., Shen, C., Lu, J., et al. (2016). Application of the ribosomal DNA ITS2 region of *Physalis* (Solanaceae): DNA barcoding and phylogenetic study. *Front. Plant Sci.* 7:1047. doi: 10.3389/fpls.2016.01047

Garlapati, D., Charankumar, B., Ramu, K., Madeswaran, P., and Murthy, M. V. R. (2019). A review on the applications and recent advances in environmental

DNA (eDNA) metagenomics. *Rev. Environ. Sci. Biotechnol.* 28, 389–411. doi: 10.1007/s11157-019-09501-4

Gleason, H. A. (1922). On the relation between species and area. *Ecology* 3, 158–162. doi: 10.2307/1929150

Gong, L., Qiu, X. H., Huang, J., Xu, W., Bai, J. Q., Zhang, J., et al. (2018). Constructing a DNA barcode reference library for southern herbs in China: a resource for authentication of southern Chinese medicine. *PLoS One* 13:e0201240. doi: 10.1371/journal.pone.0201240

Gong, W., Liu, Y., Chen, J., Hong, Y., and Kong, H. H. (2016). DNA barcodes identify Chinese medicinal plants and detect geographical patterns of *Sinosenecio* (Asteraceae). *J. Syst. Evol.* 54, 83–91. doi: 10.1111/jse.12166

Guo, M., Ren, L., and Pang, X. (2017). Inspecting the true identity of herbal materials from *Cynanchum* using ITS2 barcode. *Front. Plant Sci.* 8:1945. doi: 10.3389/fpls.2017.01945

Hajibabaei, M., Singer, G. A., Hebert, P. D., and Hickey, D. A. (2007). DNA barcoding: how it complements taxonomy, molecular phylogenetics and population genetics. *Trends Genet.* 23, 167–172. doi: 10.1016/j.tig.2007.02.001

He, T., Jiao, L., Wiedenhoeft, A. C., and Yin, Y. (2019). Machine learning approaches outperform distance- and tree-based methods for DNA barcoding of *Pterocarpus* wood. *Planta* 249, 1617–1625. doi: 10.1007/s00425-019-03116-3

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

Hebert, P. D., and Gregory, T. R. (2005). The promise of DNA barcoding for taxonomy. *Syst. Biol.* 54, 852–859. doi: 10.1080/10635150500354886

Hebert, P. D., Ratnasingham, S., and deWaard, J. R. (2003b). Barcoding animal life: cytochrome c oxidase subunit 1 divergences among closely related species. *Proc. R. Soc. Lond. Ser. B Biol. Sci.* 270, S96–S99.

Hollingsworth, P. M. (2011). Refining the DNA barcode for land plants. *Proc. Natl. Acad. Sci. U.S.A.* 108, 19451–19452. doi: 10.1073/pnas.1116812108

Hou, G., Chen, W. T., Lu, H. S., Cheng, F., and Xie, S. G. (2017). Developing a DNA barcode library for perciform fishes in the South China Sea: species identification, accuracy and cryptic diversity. *Mol. Ecol. Resour.* 18, 137–146. doi: 10.1111/1755-0998.12718

Huang, X. C., Su, J. H., Ouyang, J. X., Ouyang, S., Zhou, C. H., and Wu, X. P. (2019). Towards a global phylogeny of freshwater mussels (Bivalvia: Unionida): species delimitation of Chinese taxa, mitochondrial phylogenomics, and diversification patterns. *Mol. Phylogenet. Evol.* 130, 45–59. doi: 10.1016/j.ympev.2018.09.019

Huelsenbeck, J. P., and Ronquist, F. (2001). MrBayes: bayesian inference of phylogenetic trees. *Bioinformatics* 17, 754–755. doi: 10.1093/bioinformatics/17.8.754

Jiao, J., Huang, W., Bai, Z., Liu, F., Ma, C., and Liang, Z. (2018). DNA barcoding for the efficient and accurate identification of medicinal polygonati rhizoma in China. *PLoS One* 13:e0201015. doi: 10.1371/journal.pone.0201015

Jin, Q., Hu, X. M., Han, H. L., Chen, F., Cai, W. J., Ruan, Q. Q., et al. (2018). A two-step DNA barcoding approach for delimiting moth species: moths of Dongling Mountain (Beijing, China) as a case study. *Sci. Rep.* 8:14256. doi: 10.1038/s41598-018-32123-9

Jin, W. T., Jin, X. H., Schuiteman, A., Li, D. Z., Xiang, X. G., Huang, W. C., et al. (2014). Molecular systematics of subtribe Orchidinae and Asian taxa of Habenariinae (Orchideae, Orchidaceae) based on plastid *matK*, *rbcL* and nuclear ITS. *Mol. Phylogenet. Evol.* 77, 41–53. doi: 10.1016/j.ympev.2014.04.004

Joly, S., Davies, T. J., Archambault, A., Bruneau, A., Derry, A., Kembel, S. W., et al. (2013). Ecology in the age of DNA barcoding: the resource, the promise and the challenges ahead. *Mol. Ecol. Resour.* 14, 221–232. doi: 10.1111/1755-0998.12173

Lang, D. D., Tang, M., Hu, J. H., and Zhou, X. (2019). Genome-skimming provides accurate quantification for pollen mixtures. *Mol. Ecol. Resour.* 19, 1433–1446. doi: 10.1111/1755-0998.13061

Li, G. Q., Xue, X. F., Zhang, K. J., and Hong, X. Y. (2010). Identification and molecular phylogeny of agriculturally important spider mites (Acari: Tetranychidae) based on mitochondrial and nuclear ribosomal DNA sequences, with an emphasis on *Tetranychus*. *Zootaxa* 2647, 1–15. doi: 10.11646/zootaxa.2647.1.1

Li, J., Liu, X., Guo, H., Yue, B., and Li, J. (2013). Progress of analytic methods in animal DNA barcoding. *Sichuan J. Zool.* 32, 950–954. doi: 10.3969/j.issn.1000-7083.2013.06.030

Li, L., Josef, B. A., Liu, B., Zheng, S., Huang, L., and Chen, S. (2017). Three-dimensional evaluation on ecotypic diversity of traditional Chinese medicine: a case study of *Artemisia annua* L. *Front. Plant Sci.* 8:1225. doi: 10.3389/fpls.2017.01225

Li, M., Au, K. Y., Lam, H., Chen, L., But, P. P., and Shaw, P. C. (2014). Molecular identification and cytotoxicity study of herbal medicinal materials that are confused by *Aristolochia* herbs. *Food Chem.* 147, 332–339. doi: 10.1016/j.foodchem.2013.09.146

Liang, F., Dai, Y., Yue, L., Li, F., and Liu, X. (2015). DNA barcoding and taxonomic review of the barklouse genus *Stenopsocus* (Psocoptera: Stenopsocidae) from Taiwan. *Zootaxa* 4057, 191–209. doi: 10.11646/zootaxa.4057.2.2

Liu, J., Li, Q., Kong, L., and Zheng, X. (2011a). Cryptic diversity in the pen shell *Atrina pectinata* (Bivalvia: Pinnidae): high divergence and hybridization revealed by molecular and morphological data. *Mol. Ecol.* 20, 4332–4345. doi: 10.1111/j.1365-294X.2011.05275.x

Liu, J., Möller, M., Gao, L. M., Zhang, D. Q., and Li, D. Z. (2011b). DNA barcoding for the discrimination of Eurasian yews (*Taxus* L., Taxaceae) and the discovery of cryptic species. *Mol. Ecol. Resour.* 11, 89–100. doi: 10.1111/j.1755-0998.2010.02907.x

Liu, J. X., Wei, M. J., Li, G., Cheng, S. H., Zhao, C. Y., Borjigidai, A., et al., (2018). Construction of ITS2 barcode database of *Scutellariae radix* and establishment of DNA barcode identification method for its seeds. *Chin. J. Exper. Tradi. Med. Form.* 24, 37–45. doi: 10.13422/j.cnki.syfjx.20180906

Liu, S., Li, Y., Lu, J., Su, X., Tang, M., Zhang, R., et al. (2013). SOAPBarcode: revealing arthropod biodiversity through assembly of Illumina shotgun sequences of PCR amplicons. *Methods Ecol. Evol.* 4, 1142–1150. doi: 10.1111/2041-210X.12120

Liu, S., Yang, C., Zhou, C., and Zhou, X. (2017). Filling reference gaps via assembling DNA barcodes using high-throughput sequencing—moving toward barcoding the world. *Gigascience* 6, 1–8. doi: 10.1093/gigascience/gix104

Liu, X., Liang, M., Etienne, R. S., Gilbert, G. S., and Yu, S. (2016). Phylogenetic congruence between subtropical trees and their associated fungi. *Ecol. Evol.* 6, 8412–8422. doi: 10.1002/ece3.2503

Munch, K., Boomsma, W., Huelsenbeck, J. P., Willerslev, E., and Nielsen, R. (2008). Statistical assignment of DNA sequences using Bayesian phylogenetics. *Syst. Biol.* 57, 750–757. doi: 10.1080/10635150802422316

Myers, N., Mittermeier, R. A., Mittermeier, C. G., da Fonseca, G. A. B., and Kent, J. (2000). Biodiversity hotspots for conservation priorities. *Nature* 403, 853–858. doi: 10.1093/acrefore/9780199389414.013.95

Paz, A., and Crawford, A. J. (2012). Molecular-based rapid inventories of sympatric diversity: a comparison of DNA barcode clustering methods applied to geography-based vs clade-based sampling of amphibians. *J. Biosci.* 37, 887–896. doi: 10.1007/s12038-012-9255-x

Pečnikar, F. Z., and Buzan, E. V. (2014). 20 years since the introduction of DNA barcoding: from theory to application. *J. Appl. Genet.* 55, 43–52. doi: 10.1007/s13353-013-0180-y

Pei, N., Chen, B., and Kress, W. J. (2017). Advances of community-level plant DNA barcoding in China. *Front. Plant Sci.* 8:225. doi: 10.3389/fpls.2017.00225

Qin, Y. G., Zhou, Q. S., Yu, F., Wang, X. B., Wei, J. F., Zhu, C. D., et al. (2018). Host specificity of parasitoids (Encyrtidae) toward armored scale insects (Diaspididae): untangling the effect of cryptic species on quantitative food webs. *Ecol. Evol.* 8, 7879–7893. doi: 10.1002/ece3.4344

Ratnasingham, S., and Hebert, P. D. N. (2007). BOLD: the barcode of life data system (www.barcodinglife.org). *Mol. Ecol. Notes* 7, 355–364. doi: 10.1111/j.1471-8286.2007.01678.x

Ross, H. A., Murugan, S., and Li, W. L. S. (2008). Testing the reliability of genetic methods of species identification via simulation. *Syst. Biol.* 57, 216–230. doi: 10.1080/10635150802032990

Rougerie, R., Decaëns, T., Deharveng, L., Porco, D., James, S. W., Chang, C. H., et al. (2009). DNA barcodes for soil animal taxonomy. *Pesqui. Agropecu. Bras.* 44, 789–801. doi: 10.1590/S0100-204X2009000800002

Saitou, N., and Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* 4, 406–425. doi: 10.1093/oxfordjournals.molbev.a040454

Shi, Z. Y., Yang, C. Q., Hao, M. D., Wang, X. Y., Ward, R. D., and Zhang, A. B. (2018). FuzzyID2: a software package for large data set species identification via barcoding and metabarcoding using hidden Markov models and fuzzy set methods. *Mol. Ecol. Resour.* 18, 666–675. doi: 10.1111/1755-0998.12738

Steinke, D., Zemlak, T. S., and Hebert, P. D. (2009). Barcoding nemo: DNA-based identifications for the ornamental fish trade. *PLoS One* 4:e6300. doi: 10.1371/journal.pone.0006300

Sun, W., Li, J. J., Xiong, C., Zhao, B., and Chen, S. L. (2016). The potential power of Bar-HRM technology in herbal medicine identification. *Front. Plant Sci.* 7:367. doi: 10.3389/fpls.2016.00367

Sun, X., Bedos, A., and Deharveng, L. (2018). Unusually low genetic divergence at COI barcode locus between two species of intertidal *Thalassaphorura* (Collembola: Onychiuridae). *PeerJ* 6:e5021. doi: 10.7717/peerj.5021

Taberlet, P., Coissac, E., Pompanon, F., Brochmann, C., and Willerslev, E. (2012). Towards next-generation biodiversity assessment using DNA metabarcoding. *Mol. Ecol.* 21, 2045–2050. doi: 10.1111/j.1365-294x.2012.05470.x

Taylor, H. R., and Harris, W. E. (2012). An emergent science on the brink of irrelevance: a review of the past 8 years of DNA barcoding. *Mol. Ecol. Resour.* 12, 377–388. doi: 10.1111/j.1755-0998.2012.03119.x

Valentini, A., Pompanon, F., and Taberlet, P. (2009). DNA barcoding for ecologists. *Trends Ecol. Evol.* 24, 110–117. doi: 10.1016/j.tree.2008.09.011

Wang, X., Hua, F., Wang, L., Wilcove, D. S., and Yu, D. W. (2019). The biodiversity benefit of native forests and mixed-species plantations over monoculture plantations. *Divers. Distrib.* 25, 1721–1735. doi: 10.1111/ddi.12972

Xiao, J. H., Xiao, H., and Huang, D. W. (2004). DNA barcoding: new approach of biological taxonomy. *Acta Zool. Sin.* 50, 852–855. doi: 10.3969/j.issn.1674-5507.2004.05.023

Xu, C., Dong, W., Shi, S., Cheng, T., Li, C., Liu, Y., et al. (2015). Accelerating plant DNA barcode reference library construction using herbarium specimens: improved experimental techniques. *Mol. Ecol. Resour.* 15, 1366–1374. doi: 10.1111/1755-0998.12413

Yang, H. Q., Dong, Y. R., Gu, Z. J., Liang, N., and Yang, J. B. (2012). A preliminary assessment of *matK*, *rbcL* and *trnH-psbA* as DNA barcodes for *Calamus* (Arecaceae) species in China with a note on ITS. *Ann. Bot. Fenn.* 49, 319–330. doi: 10.5735/085.049.0603

Yang, Q. Q., Liu, S. W., and Yu, X. P. (2018). Research progress on DNA barcoding analysis methods. *Chin. J. Appl. Ecol.* 29, 1006–1014. doi: 10.13287/j.1001-9332.201803.032

Yao, X. N., Liu, Y., Xue, K., Feng, X. X., Zhang, S. X., Ma, X., et al. (2013). Review of domestic research progress on animal taxonomy DNA barcoding. *J. Agric. Sci. Technol.* 15, 99–106.

Yu, D. W., Ji, Y., Emerson, B. C., Wang, X., Ye, C., Yang, C., et al. (2012). Biodiversity soup: metabarcoding of arthropods for rapid biodiversity assessment and biomonitoring. *Methods Ecol. Evol.* 3, 613–623. doi: 10.1111/j.2041-210X.2012.00198.x

Zadeh, L. A. (1965). Fuzzy sets. *Inf. Control* 8, 338–353. doi: 10.1016/S0019-9958(65)90241-X

Zhang, A. B., Feng, J., Ward, R. D., Wan, P., Gao, Q., Wu, J., et al. (2012a). A new method for species identification via protein-coding and non-coding DNA barcodes by combining machine learning with bioinformatic methods. *PLoS One* 7:e30986. doi: 10.1371/journal.pone.0030986

Zhang, A. B., Muster, C., Liang, H. B., Zhu, C. D., Crozier, R., Wan, P., et al. (2012b). A fuzzy-set-theory-based approach to analyse species membership in DNA barcoding. *Mol. Ecol.* 21, 1848–1863. doi: 10.1111/j.1365-294X.2011.05235.x

Zhang, A. B., Hao, M. D., Yang, C. Q., and Shi, Z. Y. (2017). BarcodingR: an integrated R package for species identification using DNA barcodes. *Methods Ecol. Evol.* 8, 627–634. doi: 10.1111/2041-210X.12682

Zhang, A. B., Sikes, D. S., Muster, C., and Li, S. Q. (2008). Inferring species membership using DNA sequences with back-propagation neural networks. *Syst. Biol.* 57, 202–215. doi: 10.1080/10635150802032982

Zhou, X., Li, Y. Y., Liu, S. L., Yang, Q., Su, X., Zhou, L. L., et al. (2013). Ultra-deep sequencing enables high-fidelity recovery of biodiversity for bulk arthropod samples without PCR amplification. *Gigascience* 2:4. doi: 10.1186/2047-217X-2-4

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.

frontiers in Ecology and Evolution

ORIGINAL RESEARCH
published: 19 June 2020
doi: 10.3389/fevo.2020.00179

DNA mtCOI Barcodes for Maritime Biosecurity: A Proof of Concept in French Polynesia Ports

Eva Garcia-Vazquez[1], Alba Ardura[1] and Serge Planes[2,3]*

[1] Department of Functional Biology, University of Oviedo, Oviedo, Spain, [2] USR 3278 CNRS – EPHE, Centre de Recherche Insulaire et Observatoire de l'Environnement, Moorea, French Polynesia, [3] Laboratoire d'Excellence CORAIL, Centre de Recherche Insulaire et Observatoire de l'Environnement, Moorea, French Polynesia

DNA barcodes have been proposed for diverse applications as markers for species identification. One application that is not fully explored yet is their use for assessing the species biodiversity and presence of invasive alien species (IAS) in maritime biosecurity. The phylogeographical signals of the mitochondrial COI (mtCOI) gene have been sometimes used to infer the number of introductions and the origin of biological invasions. Here, we employed mtCOI barcodes of mollusks and acorn barnacles (N = 751) from ports of French Polynesia to infer the effect of port size, maritime traffic, and degree of openness in the risk of biological invasions. With 17.2% of non-indigenous species (NIS) recorded here, significant differences in diversity were found among docks and between long-time docked ships and their closest piers. A higher proportion of NIS was found from sheltered compared to open ports regardless of their size and traffic. Less frequent wave washing, a lower effect of currents, and partial isolation in sheltered ports could explain the difference. The results suggest that port biota surveys should focus first on ports sheltered from the open sea and emphasize the value of mtCOI barcodes for the early detection of potential invasive species and for prioritizing surveillance efforts.

Keywords: biosecurity, French Polynesia, mtCOI barcode, barcode applications, maritime ports

OPEN ACCESS

Edited by:
David S. Thaler, Biozentrum, Universität Basel, Switzerland

Reviewed by:
Yonas Isaak Tekle, Spelman College, United States
Karen Frances Armstrong, Lincoln University, New Zealand

****Correspondence:***
Eva Garcia-Vazquez
egv@uniovi.es

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

Received: *03 October 2019*
Accepted: *19 May 2020*
Published: *19 June 2020*

Citation:
Garcia-Vazquez E, Ardura A and Planes S (2020) DNA mtCOI Barcodes for Maritime Biosecurity: A Proof of Concept in French Polynesia Ports. Front. Ecol. Evol. 8:179. doi: 10.3389/fevo.2020.00179

INTRODUCTION

More than 90% of global trade goods are transported by ship[1]. This means the maritime ports convey most of the world trade traffic together with the organisms attached to the ships or transported in ballast water (Molnar et al., 2008). Merchandise imports are indeed significant in the introduction of biological invasions (Hulme, 2009). Ports are the hubs of marine invasions (e.g., Seebens et al., 2013; Bellard et al., 2016), and the factors that enhance their risk of biopollution should be identified as soon as possible. Among these, human population size explains biological invasions better than any other factor (Pyšek et al., 2010), so the size of port cities could increase biopollution risks. Ports located in estuaries—typically of low salinity—may have a higher risk of some biological pollutants, for example, Ponto-Caspian species (Paiva et al., 2018). Empty niches, suitable environmental conditions, and availability of vectors might be the most effective predictor for the invasibility of brackish water areas and estuaries (e.g., Paavola et al., 2005; Pejovic et al., 2016).

[1]www.imo.org

Early detection is the best tool to avoid establishment of new invasions (Gozlan et al., 2010; Blanchet, 2012). However, this is not always possible, not least because the need for accurate identification of species, as an essential component of this for biosecurity and conservation management strategies (Bax et al., 2001), is not able to be achieved. For example, traditional methods that rely on visual identification of specimens have been criticized for their poor ability to identify juvenile life stages that may be critically important in the establishment and spread of invasive populations. Also, there can be limited taxonomic resolution in many taxa, where morphologically cryptic species are difficult to distinguish, possibly confusing the distinction of exotic and native species (Caesar et al., 2006). Besides this, the samples need to be collected using specific manual sampling devices for different taxa (e.g., nets, electrofishing, filtering large water volumes, sediment cores, and SCUBA diving) and then sorted and individually taxonomically identified under the microscope in most cases. This limits how many samples and replicates can be collected and analyzed (Zaiko et al., 2018).

Zaiko et al. (2018) highlight the need to employ robust DNA-based tools, such as genetic barcoding, in aquatic biosecurity studies. Biosecurity not only prevents the arrival and establishing of new Invasive alien species (IAS), but it's also for the management and analysis of existing pests, where, studying their entry retrospectively, we may have information that could help to prevent similar situations occurring again. DNA barcoding has been cited as a reliable, cheap, rapid, and accurate tool for non-indigenous species (NIS) identification and monitoring (Cross et al., 2010; Briski et al., 2011; Ardura et al., 2015a; Ardura and Planes, 2017). DNA-based tools, together with rapid assessment sampling, allow species identification at any life stage based on DNA extraction from a single individual, facilitating the early detection of new arriving species before an introduced population becomes fully established in a new habitat (Armstrong and Ball, 2005; Chown et al., 2008; Briski et al., 2011; Zhan and MacIsaac, 2015).

The use of mitochondrial COI (mtCOI) DNA barcodes (Hebert et al., 2003) for ascertaining the identity of species present in marine surveys is especially important for guaranteeing biosecurity in maritime ports (Madden et al., 2019). mtCOI also has a relatively low intraspecific variability, making it useful for species identification through DNA barcoding (Meyer and Paulay, 2005). In addition, its strong phylogeographical signal in some invertebrates make this region useful for various purposes related to biosecurity beyond exotic species gene detection. For example it has been used to trace the invasion paths of green crab *Carcinus maenas* in Australian shores (e.g., Burden et al., 2014), to infer the occurrence of multiple invasion hits of the pygmy mussel *Xenostrobus securis* in the Bay of Biscay (Devloo-Delva et al., 2016), and to identify geographic donor regions (Miralles et al., 2018). Conveniently it also has a substantial database with more than 3,000,000 sequences of species and populations from around the world[2].

In this study, we employ mtCOI barcodes to ascertain NIS of Mollusca (mollusks) and Arthropoda: Crustacea (acorn barnacles) present in French Polynesia ports of different size that are connected by frequented or unfrequented maritime routes. We have chosen these taxonomic groups because they contain numerous highly invasive species that travel attached to hulls and also in ballast water (e.g., Molnar et al., 2008). We have considered port size, fresh water, sheltering level, human population nearby, and number of maritime routes (as international vs. local traffic) as key features that contribute to the arrival and establishment of marine NIS. The initial expectation was that big ports in a region have more NIS than small local ports, assuming homogeneity of the other factors considered in our study.

MATERIALS AND METHODS

Study Locations

On the island of Moorea (French Polynesia), coastal NIS can be attributed to maritime traffic (e.g., Ardura et al., 2015a) with marine protection areas as a moderator (Ardura et al., 2016a). For this reason, we have targeted only samples taken within or close to ports located beyond protected areas (**Figure 1**). Three pairs of ports were considered, taking into account the distance between them (located at <30 km of coastline between each other) and connectivity (directly connected by regular lines). Therefore, the small ports of Afareitou and Vai'eane in the south part of the island, Papetoai and Pao-Pao in the north part, and the international Papeete harbor (in Tahiti) and Port Vai'are (Moorea island) that are connected by ferry with two companies operating several times a day all year round were analyzed.

Port Features

The total length of the docks and piers was taken as a proxy of port size. It was estimated using the "distance measurement tool" in Google Maps (©2018 Google) with the maximum zooming possible. In addition to port size, the following features were considered: exposure to open sea (scored as sheltered, 0; semi-exposed, 1; open, 2), brackish water (vicinity of fresh water discharges as a proxy), and size of surrounding human population taken from national institute of statistics and economic studies[3]. These factors have been reported to be associated with marine biological invasions in other studies (e.g., Paavola et al., 2005; Molnar et al., 2008; Hulme, 2009; Pyšek et al., 2010).

Sampling

Mollusks were targeted for the regional study, and sessile crustaceans (acorn barnacles) were also considered for comparing docks and ships. For the non-native status, NIS are those species not listed or reported as a native to French Polynesia according to current inventories of Moorea fauna and the native distribution of each species (World Register of Marine Species, www.marinespecies.org; Encyclopedia of

[2]https://www.ncbi.nlm.nih.gov/nuccore/?term=COI

[3]https://www.insee.fr/fr/statistiques/3294362?sommaire=2122700

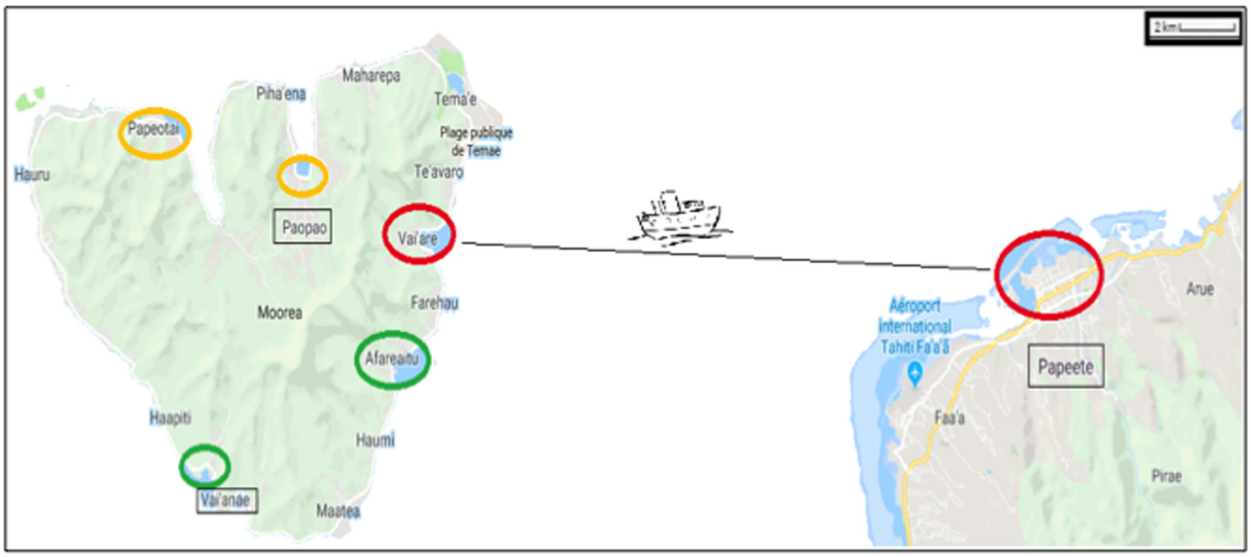


FIGURE 1 | Map showing the regions studied within the Windward Islands (French Polynesia) at 17° 31–40′S/149° 25–50′W. The ports are marked with circles proportional to their size and colored according to their proportion of non-indigenous species as red > orange > green. Ships from Vai'are and Papeete ports (small and big red circle, respectively) were also sampled.

Life[4]). Invasive alien species status (IAS) are species listed in the globally invasive species database of the International Union for Nature Conservation[5]. Invasive alien species are highly invasive in several regions of the world and, thus, pose a real biosecurity risk to Moorea. For statistical analysis, they were considered with other NIS regardless of their invasive status.

The method of sampling employed in docks of small ports was described in Ardura et al. (2015a). Briefly, sampling was carried out picking (at random within species) mollusk and acorn barnacle individuals from the intertidal range (upper to lower), which is quite short in Moorea (maximum tidal range of 0.40 m), between August 26 and September 10, 2011 (**Figure 1**). An effort was made to obtain representative samples, proportional to the abundance of each species. The methodology described in Miralles et al. (2016) sampling from rectangles of approximately 200 m^2 was followed in the larger ports: three rectangles in Papeete, two in Vai'are.

Mollusks and acorn barnacles were sampled from three ships docked in Papeete and one in Vai'are ports as well as from the closest dock. Fouling biota from three quadrats of approximately 30 cm × 30 cm was scratched with a spatula and then the mollusks and acorn barnacles were sorted, identified de visu with the help of taxonomic guides and voucher specimens from the collection of the CRIOBE in Moorea (French Polynesia) to species level when it was possible. A part of tissue (digestive tract was avoided to prevent possible contamination with gut content) was excised and stored in absolute ethanol (100%) for further DNA extraction and genetic identification from barcodes.

DNA Barcoding

DNA barcoding was carried out to ascertain de visu taxonomic identification as described in Ardura et al. (2015a). Total DNA was extracted from a small piece of tissue following the standard protocol described by Estoup et al. (1996), employing Chelex® resin (Bio-Rad Laboratories). The E.Z.N.A Mollusk DNA kit (IOMEGA, bio-tek) was used for the species with high content of mucopolysaccharides in muscle tissues, following the manufacturer's instructions. In both cases, the tubes were stored at 4°C for immediate DNA analysis, and aliquots were frozen at −20°C for long-term preservation. A fragment within the mitochondrial Cytochrome oxidase I gene (COI) was PCR amplified and sequenced using Geller et al. (2013) primers. Some individuals were double-checked with a second marker, the 16S rRNA gene with the primers described by Palumbi (1996), to confirm the species when identification using COI was not sufficiently accurate (99% match, at least 450 nucleotides coverage). For species identification, the sequences were compared with international databases BOLD system for COI[6] and the program BLAST within NCBI for 16S rRNA gene sequences[7].

Statistics

Distribution normality of the different variables analyzed in the port data set was checked first, employing Shapiro–Whilk tests. Parametric or non-parametric tests were employed accordingly for further analysis.

In the exploratory analysis of the port data, pairwise correlations between habitat and community variables were

[4]http://www.eol.org, accessed October 2019.
[5]http://www.iucngisd.org/, accessed October 2019.
[6]http://www.boldsystems.org/
[7]http://www.ncbi.nlm.nih.gov/

done. Individual rarefaction curves for each port, and sample rarefaction curves for big and small ports of each region were constructed with the aim of estimating if the mollusk diversity was sufficiently represented in samples. Diversity of samples was estimated using the Shannon–Weaver index. Differences in diversity between samples were then estimated using permutations tests (n = 9 999). Two-sample paired tests (pairs of ports located in the same area but different in size) were performed to compare means (t tests) of big versus small ports for the percentage of NIS. For estimating the contribution of different independent variables to the variation of a dependent variable, a multiple linear regression model was applied.

All statistics were conducted using PAST version 3.8 (Hammer et al., 2001).

RESULTS

Total Diversity Identified by DNA

The port characteristics are in **Supplementary Table 1**, including the proportion of NIS. The ports considered were very different in size, exposure to open sea, fresh water proximity, and surrounding human population (**Supplementary Table 1**). The number of individuals for each species found in each port and ship are presented in **Supplementary Table 2**. A total of 751 mollusks were identified to species from the French Polynesia ports analyzed (excluding ships) and another 30 specimens from ships of Papeete and Vai'are ports. In addition to the mollusks, more than 100 acorn barnacles were found from Vai'are port (>50 attached on a ship, from which 50 were analyzed). The majority of species were able to be identified using the COI gene. Only two species required the additional 16S marker for their genetical identification: *Nerita plicata* and *Pinctada maculata*. COI and 16S sequences have been deposited in GenBank[8] with the accession numbers KT149303 and KT149305 for 16S and KT149306, KT149308, KT149314-6, KT149319-23, KT290130, MH197042-4, and KJ663817-KJ663819 for COI.

Rarefaction curves for the mollusks (**Supplementary Figure 1**) generally reached a plateau, suggesting that the mollusk communities were representative, i.e., that no significant change to the species represented would occur with further sampling. In total, 155 NIS individuals (i.e., individuals of a species whose native distribution does not include the studied region) were found (17.34% of the samples; see **Table 1**). We found *Drupa albolabris* from the Philippines in Papeete (Tahiti Island) and Vai'are (Moorea Island) as well as the gastropods *Nerita tessellata* (Atlantic Ocean), *Littoraria glabrata*, and *Semiricinula tissoti* (Indian Ocean) and the invasive oyster of the Indian Ocean *Saccostrea cucullata* in Moorea Island. The Caribbean *Dendostrea frons* oyster was found in Papeete. To our knowledge, the last Polynesian mollusk inventory was published in 2009 (Tröndlé and Boutet, 2009). None of these species was described in that inventory. Therefore, from our knowledge, and taking into account the distribution described in the World Register of Marine Species (WORMS)[9], none of the species listed here were previously recorded in this area. In addition to mollusks, the highly invasive West Pacific *Amphibalanus amphitrites* and West Atlantic *Chthamalus proteus* acorn barnacles were found from Vai'are port in Moorea Island. These two species were reported by Ardura et al. (2016a). For their relative abundance, most of the mollusk NIS were scarce except a few with >5% frequency: *Saccrostrea cuculata* (9.4%) and *S. tissoti* (6.3%) in Pao-Pao, *L. glabrata* in Papetoai (6.3%), *N. tessellata* in Vai'eane (8.1%). All the acorn barnacles analyzed, in contrast, were NIS.

Analyzing all the ports together for the proportion of NIS, significant negative correlation was found only with exposure (i.e., how open is the entrance of the harbor): r = -0.82, 4 d.f., P = 0.041, the more exposed ports having a lower proportion of NIS than the more sheltered. Port size was positively correlated with the species richness (r = 0.83, 4 d.f., P = 0.040), and both port size and human population were significant in a multiple regression model (F = 27.9, df1 = 3, df2 = 2, P = 0.011) for explaining the species richness although the human population had a negative coefficient (**Table 2**). Fresh water was not significantly associated with any biotic measure.

At a subregional scale, the three pairs of ports considered (located at < 30 km of coastline between each other, or directly connected by regular lines as in the case of Papeete and Vai'are) coincided in a significantly higher proportion of NIS individuals occurring in the *smaller* port of the pair regardless of its degree of exposure (two-sample paired t test for differences between means with t = 5.500, P = 0.031 for a mean difference of 0.074 and 95% conf. 0.016–0.132) (**Figure 2**). Noteworthy, significant negative correlations were found between native biodiversity estimated from Shannon index and both%NIS individuals (tau = -0.6, 4 d.f., P = 0.038) and%NIS (tau = -0.867, 4 d.f., P = 0.014) in these Polynesian ports; this essentially means the higher the biodiversity, the lower the proportion of NIS and NIS individuals.

Small-Scale: Docks Versus Ships

In the small-scale study with four ships, the most evident result was the difference between the mollusks and acorn barnacles attached on ships and those fouling on the very close docks (**Supplementary Table 3**). The total number of individuals was greater on the docks than on the ships. This was with the exception of the Vai'are Marina sites, where the numbers were similar and influenced heavily by the large numbers of barnacles at both. Second, the species occupying the two types of substrate diverged remarkably in all cases. For example, the native *Littoraria* species and *Siphonaria normalis* were prevalent on the docks, and the native *P. maculata* was prevalent on the ships. Of the IAS, both barnacle species were found at Vai'are, but *C. proteus* was only on the dock while *A. amphithrite* only on the ship. Moreover, the species richness and diversity estimates were obviously different, being much greater in the docks than on the ships, where only a few species occurred (**Table 3**). The differences in diversity were statistically significant between Papeete Douane (custom) dock and both ship 1 and ship 2

[8]https://www.ncbi.nlm.nih.gov/genbank/

[9]http://www.marinespecies.org/index.php

TABLE 1 | Non-indigenous species found in the docks examined.

		Docks					Ships		
Species	**Native distribution**	**PA**	**PP**	**VN**	**VR**	**PT**	**WPT**	**DPT**	**VR**
Drupa albolabris	Philippines	0	0	0	4	1	0	0	0
Littoraria glabrata	Indian Ocean	6	0	0	3	0	0	0	0
Nerita tessellata	W Atlantic	0	0	6	0	0	0	0	0
Saccostrea cucullata	**Indian Ocean**	**0**	**9**	**0**	**0**	**0**	**0**	**0**	**0**
Semiricinula tissoti	Indo-West Pacific	0	6	0	0	0	0	0	0
Dendostrea frons	Caribbean	0	0	0	0	0	1	3	0
Amphibalanus amphithrite*	**West Pacific**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**50**
Chthamalus proteus*	**West Atlantic**	**0**	**0**	**0**	**50**	**0**	**0**	**0**	**0**

*Results are presented as number of individuals of a NIS found in each dock or ship (gray shade highlights where these occur). Acorn barnacle species (Crustaceans, Sessilia) are marked with *. W and D, are respectively for West and Douane inside Papeete port. Invasive Alien Species (IAS) are highlighted in bold. PA, PP, VN, VR, and PT are Papetoai, Pao-Pao, Vai'eane, Vai'are and Papeete, respectively.*

TABLE 2 | Multiple linear regression model with species richness as dependent variable.

Variable	Coefficient	SE	*t*	*P*
Constant Species richness	0.342	0.043	7.91	0.004
vs Port size	0.777	0.168	4.62	0.019
vs Human population	−3.258	0.837	3.89	0.03
vs Freshwater	−0.855	1.55	0.55	0.64

SE, standard error; t, significance test and P, p-value.

(permutation tests for Shannon–Weaver diversity indices with P value = 0.001 in the two cases) and between Vai'are Marine dock and the ship sampled nearby there (P = 0.0013); there was no significant difference between the Papeete West dock and the ship. For the difference in diversity between the Papeete West dock and the ship therein, it was not statistically significant (P = 0.385 in the permutation test).

DISCUSSION

The results of this study confirmed that employing mtCOI as a genetic barcode is very useful for a rapid assessment of invertebrate species biodiversity when the individuals can be taken easily from the environment in general and from docks and ships in particular as here. It is a fast, cheap, and easy technique to inventory biodiversity and detect NIS as well as allowing morphological identifications to be confirmed in the case of cryptic or difficult to classify species (Lara et al., 2010; Williams et al., 2012). Previous substantial study to define Moorea's biodiversity through barcodes was essential for the development of this new study about NIS presence in ports and nearby because all the species (native and NIS) found during the sampling were already in GenBank (Ardura et al., 2015a), making the use of the mtCOI marker very effective.

The use of the mtCOI gene as a main DNA barcode demonstrated the occurrence of NIS in ports of French Polynesia at a rate exceeding 17% of the total biodiversity. This is a level of biopollution higher than that found in more populated areas; for example, in some ports of Bay of Biscay with 9% of NIS and less than 15,000 inhabitants (Miralles et al., 2016). This should be considered a call for attention because Polynesian-Micronesian islands have been identified as a biodiversity hotspot and highly vulnerable to biological invasions if the current rates of global change persist over time (Bellard et al., 2014).

Another significant result for biosecurity was a higher frequency of NIS found in the smaller and sheltered ports. This was opposite to the expectation of more NIS in large ports anticipated due to major traffic associated with them. This could be related to the higher species richness found in larger ports in this study, where, following Stachowicz et al. (1999), biotic resistance could be involved; essentially, niches would be filled in large Polynesian ports with high native diversity, and new arrivals would have lesser opportunities to settle down. Despite the limited number of ports examined in this proof of concept, a negative correlation between native biodiversity and NIS (low native biodiversity, high proportion of NIS) suggests that biotic resistance is occurring as shown in ports of other regions (Miralles et al., 2016). In a previous study, this effect was not significant (Ardura et al., 2015a), probably because the ecosystems analyzed in that study were too heterogeneous (ports, protected areas, others), and here, we have considered only ports. As other authors point out (Shea and Chesson, 2002), biotic resistance may act at short or medium spatial scales, and its effect is likely diluted when ecologically distant sites are analyzed jointly. Further investigations could focus on protected Polynesian areas, including remote islands, where limited anthropogenic uses would favor native biodiversity (Ardura et al., 2016b) and less maritime traffic, which likely reduces the opportunities of new arrivals. Biotic resistance would accordingly be expected to be much higher there.

Port exposure was salient in our study for explaining the proportion of NIS over other factors of recognized effect on marine invasions, such as human population size (Pyšek et al., 2010) and freshwater discharge (Paavola et al., 2005). Port exposure was negatively correlated with NIS. This could be explained from the presence of waves that may disturb NIS settlement in open ports washing propagules away (unsettled larvae, young adults detaching from hulls or recently deposited

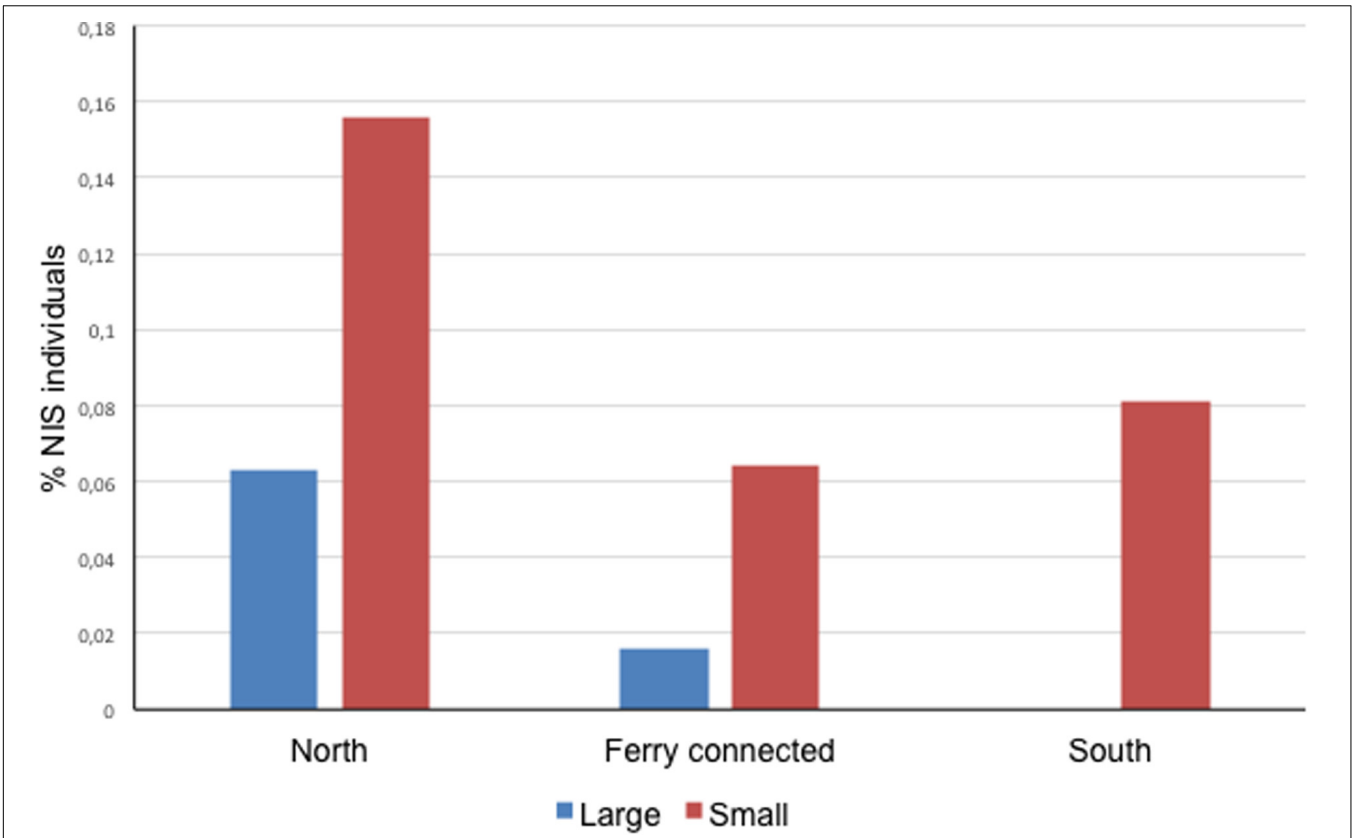


FIGURE 2 | Proportion of NIS individuals in the ports sampled in this study organized by zones. Each zone contains a pair of ports, one bigger (large) and one smaller (small) (Zero NIS recorded in large port in the South area).

on the rocks, etc.). Another spatial factor that influenced the level of NIS in our study was the location of the ports in the north or south coasts of the island with NIS being higher in the north. The explanation in this case is likely to be higher maritime traffic in the north than in the south because big cruise ships anchor in the profound Cook's and Opunohu bays (next to Pao-Pao and Papetoai ports, respectively; see **Figure 1**).

Significant differences were found between ship hulls and their closely associated docks. The most evident result was the difference between the mollusks and acorn barnacles attached on ships and those fouling on the very close docks (**Table 3**). Ship hulls only had bivalves attached although gastropods were much more abundant than bivalve species on the rocks nearby; similarly, acorn barnacles *A. amphitrites* were found on a ship, and *C. proteus* was sampled from docks. These differences could be a matter of time because biofouling species diversity depends on the time a vessel remains in a recipient region (Hopkins and Forrest, 2008). However, here, the ships had been docked in the same place for a minimum of 3 months, a time that would be enough for their organisms to move to the closest rocks. Another explanation for the difference in species between ships and docks could be differential preferences of sessile animals for substrates and materials. As an example in mollusks, Rech et al. (2018) find many more bivalves than gastropods on artificial substrates; our results were clearly in concordance with this. Regarding acorn barnacles, there is genetic evidence of multiple introductions from different regions of the two species mentioned above (Ardura et al., 2016a), suggesting they are colonizing the island via maritime traffic. Why only one was attached to a ship? The explanation could be the same: acorn barnacles exhibit different substrate preferences depending on the species; for

TABLE 3 | Diversity of ships and nearby docks in the studied ports.

Location	Species richness	Diversity	Number of NIS
PW-Dock	5	1.079	0
PW-Ship	2	0.759	1
PD-Dock	6	1.238*	0
PD-Ship 1	2	0.199*	1
PD-Ship 2	2	0.234*	1
VM-Dock	7	1.085#	1
VM-Ship	1	0.097#	1
Total docks	11	1.575	1
Total ships	4	1.181	2

*PT and VR are for Papeete and Vai'are ports; W, D and I for West, Douane and Marina docks. Species richness = total number of species. Diversity, Shannon–Weaver diversity index estimated from 99999 permutations. NIS, non-indigenous species. *Differences in diversity statistically significant between P-D dock and both P-D ship-1 and P-D ship-2 and #between Vai'are Marine dock and the ship sampled nearby there.*

example, on Swedish shores *Amphibalanus* species were found preferentially on wood although other genera (e.g., *Balanus*) preferred harder substrates (Garcia-Vazquez et al., 2018). Our results of *Amphibalanus* attached to a ship hull and *Chthamalus* to dock rocks would be consistent with those findings in Sweden and also with other studies that have found big differences between ship hulls and harbor fouling organisms (e.g., Sylvester et al., 2011). In some ways, ships could be considered partially isolated habitats carrying their own fouling biota that, in some circumstances, may provide the pattern for them to move to surrounding habitats. The mixed origins of the NIS found in our study—none of them imported for aquaculture and the majority from the Indian Ocean and other regions of the Pacific, but also one Caribbean and another two from the West Atlantic—would confirm that the main vector of NIS in Polynesia is maritime traffic (Ardura et al., 2015a).

The study illustrates that local conditions can influence the nature of resident NIS and, presumably, help prioritize surveillance efforts. The risk of biological invasions is especially important in islands because they depend on maritime trade (Hulme, 2009), but not all areas seem to be equally vulnerable to them. From our results, the areas around the ports in the Windward Islands should be periodically monitored and samples from the different species mtCOI barcoded. A higher surveillance of the beaches nearby Vai'are and especially Pao-Pao, corresponding to sheltered ports and the second one being located in the north of the island, would be recommended because they have already a quite high frequency of NIS. As seen in this proof of concept, barcoding has considerable potential in port biosecurity, which emphasizes its use for early detection of potential invasive species.

Even with these positive results, this methodology implies a sampling, requiring a large effort of human resources and many specialists systematically sampling all ecosystems. The effort is greater in habitats with difficult physical accession that have to be reached from the sea and/or diving. In addition, some species cannot be detected when they are at a low density, in their first development stages, or have high mobility (not sessile species). In these cases, the use of environmental DNA and metabarcoding is useful to detect non-target species by traces of their DNA in the water (Ardura et al., 2015b; Zaiko et al., 2015, 2018). However, some drawbacks must be taken into account because environmental DNA techniques involve higher costs and substantial analytical effort (bioinformatics) to ensure efficient exploration of the sequence data obtained from multispecies communities (Blanchet, 2012). Therefore, the best methodology should be assessed for each particular study, depending on economic and material resources and the data available and necessary in each case (Ardura and Planes, 2017).

Finally, it is important to highlight that, although these molecular methods can answer some questions about biosecurity questions, a complete marine biosecurity program should integrate complementary scientific approaches, including traditional surveys, mathematical modeling, risk assessment frameworks, citizen collaboration, and molecular techniques.

DATA AVAILABILITY STATEMENT

The datasets generated for this study can be found in the GenBank, KT149305, KT149315-6, KT149319-20, KT290130, MH197042-4, and KJ663817–KJ663819.

AUTHOR CONTRIBUTIONS

AA performed laboratory analysis and collaborated in writing. SP collaborated in the study design and writing. EG-V designed the study, sampled and wrote the article draft.

FUNDING

This study was funded from the Spanish Ministry of Economy and Competitiveness under the Grant MINECO CGL2016-79209-R and the Government of Asturias Principality under the Grant IDI-2018-000201. This is a contribution from the Marine Observatory of Asturias (OMA), Asturias University Institute of Biotechnology, and the Spanish Research Group of Excellence ARENA. AA holds a postdoctoral Juan de la Cierva fellowship and received an IRCP-CRIOBE Grant (2018).

ACKNOWLEDGMENTS

We are grateful to two reviewers of Frontiers for their constructive comments on the manuscript that helped us so much to improving it.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fevo.2020.00179/full#supplementary-material

REFERENCES

Ardura, A., and Planes, S. (2017). Rapid assessment of non-indigenous species in the era of the eDNA barcoding: a mediterranean case study. *Est. Coast. Shelf Sci.* 188, 81–87. doi: 10.1016/j.ecss.2017.02.004

Ardura, A., Planes, S., and Garcia-Vazquez, E. (2015a). Aliens in paradise. boat density and exotic coastal mollusks in moorea Island (French Polynesia). *Mar. Environ. Res.* 112, 56–63. doi: 10.1016/j.marenvres.2015.08.007

Ardura, A., Zaiko, A., Martinez, J. L., Aurelija Samuiloviene, A., Borrell, Y., and Garcia-Vazquez, E. (2015b). Environmental DNA evidence of transfer of North Sea molluscs across tropical waters through ballast water. *J. Mollusc. Stud.* 81, 495–501. doi: 10.1093/mollus/eyv022

Ardura, A., Juanes, F., Planes, S., and Garcia-Vazquez, E. (2016a). Rate of biological invasions is lower in coastal marine protected areas. *Sci. Rep.* 6:33013. doi: 10.1038/srep33013

Ardura, A., Planes, S., and Garcia-Vazquez, E. (2016b). Phylogenetic analysis for detection of multiple fouling events: a pilot study of barnacles at Moorea Island (French Polynesia). *Crustaceana* 89, 863–875. doi: 10.1163/15685403-00003554

Armstrong, K., and Ball, S. (2005). DNA barcodes for biosecurity: invasive species identification. *Phil. Trans. R. Soc. B Biol. Sci.* 360, 1813–1823. doi: 10.1098/rstb.2005.1713

Bax, N., Carlton. J. T., Mathews-Amos, A., Haedrich, R. L., Howarth, F. G., Purcell, J. E., et al. (2001). The control of biological invasions in the world's oceans. *Conserv. Biol.* 15, 1234–1246. doi: 10.1111/j.1523-1739.2001.99487.x

Bellard, C., Leclerc, C., Leroy, B., Bakkenes, M., Veloz, S., Thuiller, W., et al. (2014). Vulnerability of biodiversity hotspots to global change. *Global Ecol. Biogeogr.* 23, 1376–1386. doi: 10.1111/geb.12228

Bellard, C., Leroy, B., Thuiller, W., Rysman, J. F., and Courchamp, F. (2016). Major drivers of invasion risks throughout the world. *Ecosphere* 7, 1–14.

Blanchet, S. (2012). The use of molecular tools in invasion biology: an emphasis on freshwater ecosystems. *Fisheries Manag. Ecol.* 19, 120–132. doi: 10.1111/j.1365-2400.2011.00832.x

Briski, E., Cristescu, M. E., Bailey, S. A., and MacIsaac, H. J. (2011). Use of DNA barcoding to detect invertebrate invasive species from diapausing eggs. *Biol. Inv.* 13, 1325–1340. doi: 10.1007/s10530-010-9892-7

Burden, C. T., Stow, A. J., Hoggard, S. J., Coleman, M. A., and Bishop, M. J. (2014). Genetic structure of Carcinus maenas in southeast Australia. *Mar. Ecol. Progr. Ser.* 500, 139–147. doi: 10.3354/meps10704

Caesar, R. M., Sorensson, M., and Cognato, A. I. (2006). Integrating DNA data and traditional taxonomy to streamline biodiversity assessment: an example from edaphic beetles in the Klamath ecoregion. California USA. *Divers. Distrib.* 12, 483–489. doi: 10.1111/j.1366-9516.2006.00237.x

Chown, S. L., Sinclair, B. J., and van Vuuren, B. J. (2008). DNA barcoding and the documentation of alien species establishment on sub-Antarctic Marion Island. *Polar Biol.* 31, 651–655. doi: 10.1007/s00300-007-0402-z

Cross, H. B., Lowe, A. J., and Gurgel, F. D. (2010). "DNA barcoding of invasive species," in *Fifty Years of Invasion Ecology: The Legacy of Charles Elton*, ed. D. Richardson (Oxford: Blackwells).

Devloo-Delva, F., Miralles, L., Ardura, A., Borrell, Y. J., Pejovic, I., Tsartsianidou, V., et al. (2016). Detection and characterisation of the biopollutant Xenostrobus securis (Lamarck 1819) Asturian population from DNA Barcoding and eBarcoding. *Mar. Pol. Bull.* 105, 23–29. doi: 10.1016/j.marpolbul.2016.03.008

Estoup, A., Largiader, C. R., Perrot, E., and Chourrout, D. (1996). Rapid one-tube DNA extraction for reliable PCR detection of fish polymorphic markers and transgenes. *Mol. Mar. Biol. Biotech.* 5, 295–298.

Garcia-Vazquez, E., Cani, A., Diem, A., Ferreira, C., Geldhof, R., Marquez, L., et al. (2018). Leave no traces – Beached marine litter shelters both invasive and native species. *Mar. Pol. Bull.* 131, 314–322. doi: 10.1016/j.marpolbul.2018.04.037

Geller, J., Meyer, C., Parker, M., and Hawk, H. (2013). Redesign of PCR primers for mitocondrial cytochrome c oxidase subunit I for marine invertebrates and application in all-taxa biotic surveys. *Mol. Ecol. Res.* 13, 851–861. doi: 10.1111/1755-0998.12138

Gozlan, R. E., Britton, J. R., Cowx I., and Copp, G. H. (2010). Current knowledge on non-native freshwater fish introductions. *J. Fish Biol.* 76, 751–786. doi: 10.1111/j.1095-8649.2010.02566.x

Hammer, Ø, Harper, D. A. T., and Ryan, P. D. (2001). PAST: paleontological statistics software package for education and data analysis. *Palaeontol. Electronica* 4, 1–9.

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

Hopkins, G. A., and Forrest, B. M. (2008). Management options for vessel hull fouling: an overview of risks posed by in-water cleaning. *ICES J. Mar. Sci.* 65, 811–815. doi: 10.1093/icesjms/fsn026

Hulme, P. E. (2009). Trade, transport and trouble: managing invasive species pathways in an era of globalization. *J. App. Ecol.* 46, 10–18. doi: 10.1111/j.1365-2664.2008.01600.x

Lara, A., Ponce, J. L., Rodriguez, R., Casane, D., Côté, G., Bernatchez, L., et al. (2010). DNA barcoding of Cuban freshwater fishes: evidence for cryptic species and taxonomic conflicts. *Mol. Ecol. Res.* 10, 421–430. doi: 10.1111/j.1755-0998.2009.02785.x

Madden, M., Young, R. G., Brown, J. W., Miller, S. E., Frewin, A. J., and Hanner, R. H. (2019). Using DNA barcoding to improve invasive pest identification at U.S. ports-of-entry. *PloS One* 14:e0222291. doi: 10.1371/journal.pone.0222291

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

Miralles, L., Ardura, A., Arias, A., Borrell, Y. J., Clusa, L., Dopico, E., et al. (2016). Barcodes of marine invertebrates from north Iberian ports: native diversity and resistance to biological invasions. *Mar. Pol. Bull.* 112, 183–188. doi: 10.1016/j.marpolbul.2016.08.022

Miralles, L., Ardura, A., Clusa, L., and Garcia-Vazquez, E. (2018). DNA barcodes of antipode marine invertebrates in bay of biscay and gulf of lion ports suggest new biofouling challenges. *Sci. Rep.* 8:16214.

Molnar, J. L., Gamboa, R. L., Revenga, C., and Spalding, M. D. (2008). Assessing the global threat of invasive species to marine biodiversity. *Frontiers Ecol. Environ.* 6:485–492. doi: 10.1890/070064

Paavola, M., Olenin, S., and Leppakoski, E. (2005). Are invasive species most successful in habitats of low native species richness across European brackish water seas? *Est. Coast. Shelf Sci.* 64, 738–750. doi: 10.1016/j.ecss.2005.03.021

Paiva, F., Barco, A., Chen, Y., Mirzajani, A., Chan, F. T., Lauringson, V., et al. (2018). Is salinity an obstacle for biological invasions? *Global Change Biol.* 24, 2708–2720. doi: 10.1111/gcb.14049

Palumbi, S. R. (1996). "Nucleic acids II: the polymerase chain reaction," in *Molecular Systematics*, eds D. M. Hillis, C. Moritz, and B. K. Mable (Sunderland, MA: Sinauer Associates, Inc), 205–247.

Pejovic, I., Ardura, A., Miralles, L., Arias, A., Borrell, Y. J., and Garcia-Vazquez, E. (2016). DNA barcoding for assessment of exotic molluscs associated with maritime ports in northern Iberia. *Mar. Biol. Res.* 12, 168–176. doi: 10.1080/17451000.2015.1112016

Pyšek, P., Jarošík, V., Hulme, P. E., Kühn, I., Wild, J., Arianoutsou, M., et al. (2010). Disentangling the role of environmental and human pressures on biological invasions across Europe. *Proc. Natl. Acad. Sci. U.S.A.* 107, 12157–12162.

Rech, S., Borrell, Y. J., and Garcia-Vazquez, E. (2018). Anthropogenic marine litter composition in coastal areas may be a predictor of potentially invasive rafting fauna. *PLoS One* 13:e0191859. doi: 10.1371/journal.pone.0191859

Seebens, H., Gastner, M. T., and Blasius, B. (2013). The risk of marine bioinvasion caused by global shipping. *Ecol. Lett.* 16, 782–790. doi: 10.1111/ele.12111

Shea, K., and Chesson, P. (2002). Community ecology theory as a framework for biological invasions. *Trends Ecol. Evol.* 17, 170–176. doi: 10.1016/s0169-5347(02)02495-3

Stachowicz, J. J., Whitlatch, R. B., and Osman, R. W. (1999). Species diversity and invasion resistance in a marine ecosystem. *Science* 286, 1577–1579.

Sylvester, F., Kalaci, O., Leung, B., Lacoursière-Roussel, A., Clarke Murray, C., Choi, F., et al. (2011). Hull fouling as an invasion vector: can simple models explain a complex problem? *J. Appl. Ecol.* 48, 415–423. doi: 10.1111/j.1365-2664.2011.01957.x

Tröndlé, J., and Boutet, M. (2009). Inventory of Marine Molluscs of French Polynesia. *Atoll Res. Bull.* 570, 1–90. doi: 10.5479/si.00775630.570.1

Williams, P. H., Mark, B. J. F., Jamrs, C. C., Jiandong, A., Murat, A. A., Lincoln, B. R., et al. (2012). Unveiling cryptic species of the bumblebee subgenus Bombus s. str. worldwide with COI barcodes (*Hymenoptera*: *Apidae*). *Syst. Biodivers.* 10, 21–56. doi: 10.1080/14772000.2012.664574

Zaiko, A., Pochon, X., Garcia-Vazquez, E., Olenin, S., and Wood, S. (2018). Advantages and limitations of environmental DNA/RNA tools for marine biosecurity: management and surveillance of non-indigenous species. *Front. Mar. Sci.* 5:322. doi: 10.3389/fmars.2018.00322

Zaiko, A., Samuiloviene, A., Ardura, A., and Garcia-Vazquez, E. (2015). Metabarcoding approach for nonindigenous species surveillance in marine coastal waters. *Mar. Poll. Bull.* 100, 53–59. doi: 10.1016/j.marpolbul.2015.09.030

Zhan, A., and MacIsaac, H. J. (2015). Rare biosphere exploration using high-throughput sequencing: research progress and perspectives. *Conserv. Genet.* 16, 513–522. doi: 10.1007/s10592-014-0678-9

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.

frontiers in Ecology and Evolution

REVIEW
published: 11 August 2020
doi: 10.3389/fevo.2020.00248

Putting COI Metabarcoding in Context: The Utility of Exact Sequence Variants (ESVs) in Biodiversity Analysis

*Teresita M. Porter and Mehrdad Hajibabaei**

Centre for Biodiversity Genomics and Department of Integrative Biology, University of Guelph, Guelph, ON, Canada

DNA barcoding and metabarcoding are techniques that focus on signature genomic regions that in theory provide species level resolution, but in practice this is not always possible. We place animal-focused COI metabarcoding in context with respect to the use of marker gene sequencing in microbial and fungal ecology. We focus on three specific aspects of metabarcodes: (1) the process of metabarcode sequence clustering, (2) how metabarcode cluster types affect the results of biodiversity analyses, and (3) the current state of reference sequence databases used for metabarcode identification. Using examples from the arthropod COI metabarcode literature, we show that exact sequence variants (ESVs) detect more unique taxa than operational taxonomic units (OTUs) but with similar patterns in taxonomic resolution. We also show that the difference between ordinations based on ESVs or OTUs recover similar groupings. We compile a list of reference sequence databases useful for multi-marker metabarcoding and present a list of reference sequence databases specifically formatted for use with a naive Bayesian classifier for rigorous metabarcode taxonomic assignments. Sophisticated tools and reference databases are available for analyzing COI sequences, and these compare favorably with those available for other metabarcode markers such as the ribosomal RNA genes used to target microbes and fungi.

Keywords: exact sequence variant, amplicon sequence variant, operational taxonomic unit, DNA barcode, mini-barcode, metabarcode, taxonomic assignment

OPEN ACCESS

Edited by:
David S. Thaler, Biozentrum, Universität Basel, Switzerland

Reviewed by:
Andrew Dopheide, Manaaki Whenua Landcare Research, New Zealand
Helena Kristiina Wirta, University of Helsinki, Finland

****Correspondence:***
Mehrdad Hajibabaei
mhajibab@uoguelph.ca

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

Received: *13 December 2019*
Accepted: *07 July 2020*
Published: *11 August 2020*

Citation:
Porter TM and Hajibabaei M (2020) Putting COI Metabarcoding in Context: The Utility of Exact Sequence Variants (ESVs) in Biodiversity Analysis. Front. Ecol. Evol. 8:248. doi: 10.3389/fevo.2020.00248

BACKGROUND

The objective of DNA barcoding is to permit specimen identification to the species rank. Part of the DNA barcoding process involves building a high-quality reference database containing geographic, morphological, and taxonomic information that is submitted along with a high-quality reference sequence providing species-level resolution (Hebert et al., 2003). DNA barcodes can then be used to help identify unknown specimens when compared to a reference sequence database. Cytochrome c oxidase subunit I (COI) mitochondrial DNA (mtDNA) barcodes for animal species are about 650 bp, the length supported by Sanger sequencing, but modern barcoding has been able to scale up by using newer sequencing technology (**Box 1**). In practice, however, only a proportion of DNA barcode records themselves represent fully-identified specimens at the species rank

BOX 1 | Scaling up DNA barcoding.

Though DNA barcodes can be generated for a few samples at a time to help fill out a reference dataset for a particular study, the process can also be scaled up tremendously where researchers have access to automation, liquid handling machines, and high throughput sequencing technology (Hebert et al., 2003, 2018; Hajibabaei et al., 2005). Initially, DNA barcodes were generated in batches using Sanger sequencing. Later, protocols were adapted for high throughput sequencing using an Illumina MiSeq platform where multiple overlapping mini-barcode regions were targeted and then assembled into full length barcodes (Shokralla et al., 2015b). More recently, scalability has been increased and overall cost per sequence decreased by using asymmetric unique molecular identifier (UMI) tagging to track individual samples with single molecule real time (SMRT) technology on the PacBio SEQUEL system (Hebert et al., 2018). This new system ramps up the throughput from 96-sample batches using Sanger sequencing up to 10,000 samples per SEQUEL run. For example, the International Barcodes of Life (iBOL) consortium has released more than 2.6 million DNA barcode sequences from 500,000 species as a part of the BARCODE 500K project (available from https://www.boldsystems.org). Most recently, the current BIOSCAN project is expected to generate DNA barcode sequences for more than 2 million species (Hobern and Hebert, 2019; Hobern, 2020).

(Porter and Hajibabaei, 2018b). Some issues that hamper rapid taxonomic identification include dwindling taxonomic expertise (Ebach et al., 2011); hyperdiversity in certain taxa such as insects, microbes, and fungi (Lozupone and Knight, 2007; Blackwell, 2011; Basset et al., 2012; Tedersoo et al., 2014); and lack of morphological characters at certain life stages such as immature insect larva or asexual fungal cultures. Even specimens with degraded DNA, however, such as food products or archival specimens, have been successfully sequenced using mini-barcodes (**Box 2**). The commonality of these challenges across multiple fields of study, from microbes to animals, has driven the development of DNA-based methods to detect and identify organisms.

The fields of microbial ecology and animal biodiversity each came up with their own solution to a shared problem: How do you consistently label sequences from specimens that cannot be identified to the species rank? In mycology, internal transcribed spacer region of ribosomal DNA (ITS rDNA) sequences are clustered into species hypotheses (SHs) that are given a numeric identifier and can be used as a common label for sequences that cannot be identified to the species rank (Koljalg et al., 2013). In the field of COI barcoding, the barcode index number (BIN) serves a similar purpose (Ratnasingham and Hebert, 2013). Specialized databases such as BOLD for COI mtDNA and UNITE for ITS rDNA barcodes house reference sequences and their corresponding BINs or SHs that attempt to approximate species units (Ratnasingham and Hebert, 2007, 2013; Koljalg et al., 2013). In the future, it is possible that BINs could be adapted to include high quality metabarcode (environmental) sequences lacking a physical specimen in the way that fungal species hypotheses (SHs) currently do (Kõljalg et al., 2019; Nilsson et al., 2019).

To transition from sampling individuals (DNA barcoding) to whole communities (DNA metabarcoding) requires the use of "culture-free" and "capture-free" approaches based on targeting environmental DNA (**Box 3**). DNA metabarcoding is a technique similar to the culture independent marker gene sequencing routinely used in the microbial and fungal ecology literature. The term *DNA metabarcoding*, however, also implies species-level taxonomic assignment (Taberlet et al., 2012b). Species level resolution of metabarcodes, however, may not be possible if there are gaps in the reference sequence database, the chosen marker lacks species-level resolution (Hajibabaei et al., 2011; Hajibabaei, 2012), or if the metabarcode sequences are too short to provide enough variable characters for a confident assignment (Porter and Hajibabaei, 2018a). In the microbial literature, it is accepted that 16S rRNA gene sequences may only provide genus level taxonomic assignments (Wang et al., 2007). Popular bioinformatic pipelines used in the microbial ecology and microbiome literature, such as QIIME, produce rank-flexible taxonomic assignments (Caporaso et al., 2010). In the DNA barcoding and metabarcoding literature, this type of rank flexible taxonomic assignment was specifically termed "metasystematics" (Hajibabaei, 2012).

From microbes to macrofauna, DNA metabarcoding can be conducted without having to isolate or identify individuals using morphological characters and leverages the sequence and taxonomic information contained in reference databases built from DNA barcodes (Hajibabaei et al., 2011; Taberlet et al., 2012b; Yu et al., 2012). Often, metabarcodes range from about 200–400 bp to correspond to the length supported by current high throughput sequencing platforms such as the Illumina MiSeq (Hajibabaei et al., 2011; Taberlet et al., 2012b). For some applications, such as with ancient DNA, even shorter regions may be targeted (D'Costa et al., 2011). In this paper, we focus on how metabarcodes are generated, analyzed, and identified. We ask three questions: (1) Why do we cluster metabarcode reads? (2) Does metabarcode cluster type affect the results of biodiversity analyses? (3) What resources are available for metabarcode identification?

WHY DO WE CLUSTER METABARCODE READS?

If the DNA metabarcode sequences themselves provide the finest level of resolution, why do many metabarcode bioinformatic pipelines include a clustering step (**Box 4**)? First, clustering metabarcode sequences allows users to reduce the size of the data files and facilitate downstream processing. Second, sequence clustering may absorb artifactual sequences caused by PCR or sequencing error. This clustering step was needed because the early methods of denoising were computationally intensive and difficult to implement on large datasets (Reeder and Knight, 2009). Current denoising methods are incorporated into several existing programs and pipelines such as DADA2, USEARCH, VSEARCH, and Deblur (Callahan et al., 2016; Edgar, 2016; Rognes et al., 2016; Amir et al., 2017; Nearing et al., 2018). Reads may be clustered to approximate species units. In the field of microbial ecology, it was shown that if a phylogenetic species definition requires at least 70% or greater DNA similarity, this

BOX 2 | Mini-barcodes for difficult samples.

Mini-barcodes can be thought of as partial DNA barcodes where very short regions about 100–200 bp in length are generated from individual specimens (Hajibabaei et al., 2006). These minimalist barcodes are ideal for identifying very old or poorly preserved specimens or highly processed material (e.g., food products) where DNA is very degraded and longer barcode sequences are difficult to amplify (Hajibabaei et al., 2006; Shokralla et al., 2015a). In the original study that describes a minimalist barcode, a dataset of over 200 Australian fish species and four species-rich lepidopteran genera show that 109–218 bp regions of COI mtDNA had sufficient variation to allow for identification (Hajibabaei et al., 2006).

Mini-barcodes, and even metabarcodes, can also be generated from sample preservative such as ethanol (Hajibabaei et al., 2012; Erdozain et al., 2019). In one of the first studies describing this non-destructive technique, DNA was isolated from mescal, a liquor containing the larva of the Agave butterfly, and a sequence from the family that includes the Agave butterfly was successfully recovered (Shokralla et al., 2010). The optimization of non-destructive DNA barcoding to identify single specimens and entire communities from sample preservative continues (Shokralla et al., 2010; Hajibabaei et al., 2012; Erdozain et al., 2019; Marquina et al., 2019; Gauthier et al., 2020; Zenker et al., 2020).

BOX 3 | Environmental DNA.

Environmental DNA (eDNA) refers to DNA that can be extracted from environmental samples, without having to isolate individual organisms (Taberlet et al., 2012a). In the microbial and fungal literature, "culture-free" methods were used to extract eDNA directly from, for example, soil or water without having to isolate, culture, and identify individual strains (Pace et al., 1986; Handelsman, 2004). The term "bulk" was used to refer to a bulk environmental sample such as soil or water. The advantage of "culture-free" methods was the avoidance of known culture-bias such as in the "great plate count anomaly" described from microbial studies (Staley and Konopka, 1985). More recently in animal-focused studies, "capture-free" methods using eDNA have been adopted to facilitate the detection of organisms in the environment (Darling, 2019). In animal-focused studies, eDNA methods allow for the detection of organisms that are difficult to catch using traditional methods, especially if they are rare.

The term "extracellular DNA" should not be confused with eDNA as we use the term here. In some of the modern eDNA literature, extracellular DNA has been targeted to improve the chances of recovering enough DNA to detect non-microbial organisms such as plants and invertebrates from soil or water. Extracellular DNA can adsorb to sand, clay, silt, or organic compounds such as humic acids. It has been shown that extracellular DNA is more resistant to DNase digestion and adsorbed DNA may persist longer than free-DNA in the environment (Romanowski et al., 1991; Nielsen et al., 2007; Pietramellara et al., 2009). It has also been suggested that focusing metabarcoding on extracellular DNA allows for more efficient detection of non-microbial organisms compared with using methods that extract both intra- and extra-cellular DNA from environmental samples that are dominated by microbial DNA (Taberlet et al., 2012c). In the eDNA literature, water samples are filtered to isolate the extracellular DNA used to indirectly monitor fish and other aquatic animals using metabarcoding or species-specific qPCR (Hänfling et al., 2016; Hernandez et al., 2020). The focus on extracellular DNA for animal-focused metabarcoding can be contrasted with that in the microbial soil ecology literature where DNA adsorbed to particles has been termed "relic DNA." Such relic DNA has been considered problematic as it may obscure estimates of microbial diversity (Carini et al., 2016).

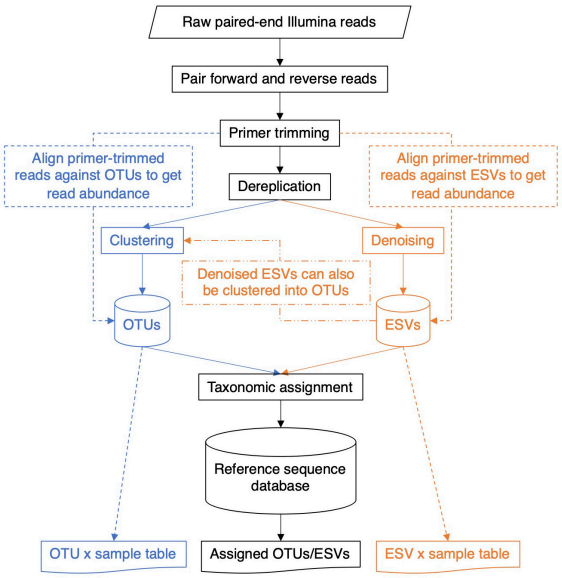
In eDNA studies, a further distinction is also often made between environmental DNA comprised of degraded extracellular DNA or DNA from mixed community samples (Deiner et al., 2017). Such mixed community samples are sometimes referred to as "bulk" tissue samples that are comprised of whole organisms such as those collected from traps or nets (Taberlet et al., 2012b; Yu et al., 2012; Creer et al., 2016). For example, the arthropods collected from a Malaise trap or kick-net sample can be homogenized together, whole community DNA can be extracted, then one or more primer sets are used for metabarcoding (Hajibabaei et al., 2011; Gibson et al., 2014; Barsoum et al., 2019).

The terminology used in microbial versus animal metabarcoding studies needs to be understood from the history of the field and context in terms of the targeted organisms to avoid misunderstandings.

corresponds to ∼97% sequence similarity in the 16S rRNA gene region (Stackebrandt and Goebel, 1994). A recent study, however, suggests that 99–100% thresholds may be more appropriate (Edgar, 2018b). In current fungal ecology, 97–99% cutoffs for the ITS rDNA are sometimes used to approximate species units (Koljalg et al., 2013). In COI metabarcoding studies, a variety of sequence similarity cutoffs have been used ranging from 95–100% to maximize genetic diversity recovered while controlling for the effect of sequence errors, resulting in species-like groupings (Elbrecht et al., 2017; Braukmann et al., 2019; Tapolczai et al., 2019). In many cases, a 97% sequence similarity cutoff is used because existing bioinformatic pipelines were originally developed to process microbial rRNA gene sequences, and this threshold is often a default value. In all cases, use of a single sequence similarity threshold, such as 97% OTUs, may not reproduce species units across all taxa defined by traditional species concepts or across the variety of markers used for metabarcoding today.

The reasons for clustering metabarcodes may vary, but the result are two types of metabarcode clusters, operational taxonomic units (OTUs) or exact sequence variants (ESVs). OTUs, or molecular OTUs (mOTUs), represent a cloud of similar sequences whose composition may vary depending on the order of the sequences being clustered, making them difficult to reproduce and compare across studies (He et al., 2015). Any single OTU is usually represented by a single sequence, such as the centroid, and the remaining sequences in the OTU are disregarded in further analyses obscuring the underlying nucleotide variation within any single OTU. On the other hand, exact sequences variants (ESVs), also known as amplicon sequence variants (ASVs) (Callahan et al., 2017), zero-radius OTUs (Edgar, 2016), or simply error-corrected OTUs defined by 100% sequence identity, each represent sequence variation down to single-nucleotide resolution. To ensure high quality ESVs, steps need to be taken to remove artifactual sequences such as putative chimeras, sequences with predicted errors, and contaminants (Callahan et al., 2016; Edgar, 2016). We make the case here that ESVs are appropriate for analyzing metabarcodes from any taxon, from microbes to arthropods, using any marker from rRNA genes to COI. The advantages of using ESVs includes improved taxonomic resolution down to single nucleotides as well as improved reproducibility and comparability across studies that use the same marker (Callahan et al., 2017). In theory, ESVs are comparable to haplotypes used commonly in population genetics and phylogeography (Callahan et al., 2017) and are already starting to be treated as such in the COI metabarcoding

BOX 4 | A general bioinformatic pipeline for metabarcode clusters based on operational taxc



DNA metabarcodes are often generated using paired-end Illumina sequencing. Forward and reverse reads are paired, then the ends of the sequence matching the primers are removed. In some pipelines, primers are trimmed first, then forward and reverse reads are paired. Each of these steps may require the user to set a minimum Phred quality score cutoff as well as a cutoff for the number of mismatches tolerated.

At this point, sequence files belonging to each sample are often pooled together for a "global" analysis. Dereplication involves obtaining just the unique sequences from the set. The number of reads matching each unique sequence is retained as this information is needed for both the clustering and denoising methods described below. The output is usually sorted by decreasing read abundance, but other sort orders are possible. Because many clustering methods are "greedy" to improve computation time, changing the input order of the sequences can change the composition of the resulting OTUs.

The operational taxonomic unit (OTU) clustering part of the pipeline is shown in blue. An identity threshold is chosen, for example, 0.97, to cluster sequences with at least 97% sequence similarity. Steps to remove putative chimeric sequences and rare sequences that may contain sequence errors will also be conducted at this step. In pipelines run in USEARCH or VSEARCH, each OTU is represented by a single centroid sequence in any future analyses (Edgar, 2013; Rognes et al., 2016). To create an OTU × sample table containing read numbers, primer-trimmed paired sequences can be aligned to each OTU centroid sequence in the database. This step may require numerous parameters to be chosen such as the identity threshold, for example, 0.97, to retain sequences with at least 97% sequence similarity to an OTU centroid sequence.

The exact sequence variant (ESV) denoising pipeline is shown in orange. In USEARCH or VSEARCH, the UNOISE3 algorithm performs denoising (Edgar, 2016) by clustering identical sequences together, similar to using an identity threshold of 1.0 to cluster sequences that have 100% sequence similarity. During this process, sequences with predicted sequence errors, putative PhiX carry-over from Illumina sequencing, putative chimeric sequences, and rare sequences are removed. Each denoised ESV is represented by a single sequence in any future analysis. To create an ESV x sample table containing read numbers, primer-trimmed paired sequences can be aligned to each unique ESV sequence in the database. This step may require numerous parameters to be chosen such as the identity threshold of 1.0 to retain sequences with at least 100% sequence similarity to a denoised ESV sequence.

Several metabarcode denoising programs have been compared and the USEARCH UNOISE3 algorithm was shown to be the fastest and DADA2 was found to generate the greatest number of ESVs (Callahan et al., 2016; Nearing et al., 2018). USEARCH is proprietary software with a free 32-bit version available and DADA2 is open source software. VSEARCH is another useful open source software program that allows you to use as much memory as your system supports to facilitate large analyses, and it can also run the UNOISE3 algorithm.

Metabarcode identification can be performed a number of ways using similarity-, phylogeny-, or composition-based methods (Porter and Hajibabaei, 2018c). One most popular method for high-throughput identification of large batches of COI metabarcodes is to perform BLAST comparisons against the GenBank nucleotide or other custom databases. We have developed the COI classifier v4 that uses a method initially developed to taxonomically assign rRNA gene sequences. This naive Bayesian classifier was trained on a curated set of COI sequences from BOLD and GenBank to make rapid, accurate taxonomic assignments (Altschul et al., 1997; Wang et al., 2007; Porter and Hajibabaei, 2018a). Recently, a python package called BOLDigger has been developed to help automate batch query submissions to the BOLD identification engine and can be used to identify COI, ITS, rbcL, and matK sequences (Buchner and Leese, 2020). For each of these methods, there are trade-offs in terms of ease of use, speed, and rigor. Users should carefully consider the output: Similarity-based methods provide a measure of how similar a query sequence is to a target sequence whereas taxonomic assignment methods provide a statistical measure of confidence for a taxonomic placement at each rank. Each of these approaches relies on comparing unknown metabarcode sequences against a reference sequence database of known sequences. The quality, coverage, and availability of these reference sequences can be quite varied for COI and other popular metabarcode markers and is discussed below (also see **Table 1**).

literature (Elbrecht et al., 2018). In terms of reproducibility and comparability, it is relatively straightforward to align new reads using a 100% sequence similarity threshold to an ESV reference database. It is more complicated to align new reads to an OTU reference database because an arbitrary similarity threshold needs to be chosen or to regenerate OTUs from scratch since greedy algorithms are affected by sequence input order and may not generate OTUs with the same composition as before (He et al., 2015). For studies that require species estimates, fungal ITS or animal COI ESVs can be aligned to ITS SHs or COI BINs using a meaningful threshold for sequence similarity, say 97% sequence similarity. In the fungal literature, ESVs and OTUs were both shown to recover similar ecological patterns (Glassman and Martiny, 2018). In this paper, we show how the analysis of COI metabarcode clusters based on ESVs and OTUs affects biodiversity analyses (see next section).

After choosing whether metabarcode clusters will be based on OTUs or ESVs, it will be necessary to decide on which approach to take for taxonomically assigning or identifying the clusters. For assessing biodiversity, there is no need to limit analyses to only the portion of the dataset confidently identified to species. Instead, we recommend that metabarcode clusters are annotated to the most specific taxonomic rank possible. For example, the taxonomic lineage "Cellular Organisms; Eukaryota; Metazoa; Arthropoda; Arachnida; Araneae; Amarobiidae; Amarobius; *Amarobius borealis*; F230R_Otu231" represents an OTU identified to the species rank, *Amarobius borealis*; and the taxonomic lineage "Cellular Organisms; Eukaryota; Metazoa; Arthropoda; Insecta; Diptera; F230R_Otu1794" represents an OTU identified to the order rank. Using a taxonomic assignment method such as the COI Classifier v4, instead of a similarity-based method, can help to delimit the finest level of

resolution that can be made with confidence (**Table 1**; Porter and Hajibabaei, 2018a). Filtering for bootstrap support values that exceed cutoff values can also help reduce the rate of false positive taxonomic assignments (Porter and Hajibabaei, 2018a). This may be an important consideration in cases where the cost of making a false-positive assignment is high, such as where falsely detecting an invasive species could be a cause for alarm. Methods that use a naive Bayesian classifier such as the RDP classifier, phylogenetic-based taxonomic assignment such as SAP, Bayesian multinomial regression such as PROTAX, or non-Bayesian k-mer based methods such as SINTAX each produce measures of confidence for taxonomic assignments for each rank (Wang et al., 2007; Munch et al., 2008; Huson et al., 2016; Somervuo et al., 2016). Some methods even take into consideration species that exist but may not have a reference sequence, new species, and mislabeled sequences (Somervuo et al., 2016, 2017).

HOW DOES CLUSTER METHOD CHOICE AFFECT DIVERSITY ANALYSES?

For biodiversity analyses, the choice between using ESVs or OTUs can affect recovered alpha diversity/richness (Hajibabaei et al., 2019). We reanalyzed the data from a study that used COI metabarcoding to assess invertebrates directly from forest soils and directly compared the data reanalyzed two ways: using denoised ESVs and using denoised ESVs clustered into OTUs with 97% sequence similarity (Porter et al., 2019; **Box 4**). Taxonomic assignments were made using a naive Bayesian classifier trained using a COI reference set (Wang et al., 2007; Porter and Hajibabaei, 2018a). Using this method, we were able to filter for taxonomic assignments to ensure 95% accuracy

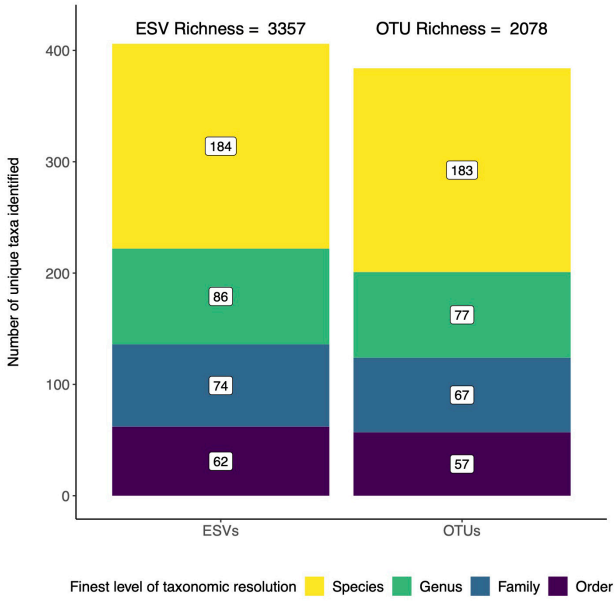


FIGURE 1 | ESVs detect more unique taxa than OTUs, but both reveal similar patterns in taxonomic resolution. Data is from a study that assessed arthropod diversity using COI metabarcoding of forest soil (Porter et al., 2019). The data was analyzed twice, first using denoised exact sequence variants (ESVs) and second using denoised ESVs that were clustered into operational taxonomic units (OTUs) based on 97% sequence similarity.

TABLE 1 | Taxonomic assignment using the COI classifier produces a measure of statistical support at each rank.

ESV	Rank	Taxon	COI Classifier v4 bootstrap support*
F230R_Otu231	Root	Cellular organisms	1.0
	Superkingdom	Eukaryota	1.0
	Kingdom	Metazoa	1.0
	Phylum	Arthropoda	1.0
	Class	Arachnida	1.0
	Order	Araneae	1.0
	Family	Amaurobiidae	1.0
	Genus	Amaurobius	1.0
	Species	*Amaurobius borealis*	1.0**
F230R_Otu1794	Root	Cellular organisms	1.0
	Superkingdom	Eukaryota	1.0
	Kingdom	Metazoa	1.0
	Phylum	Arthropoda	1.0
	Class	Insecta	1.0
	Order	Diptera	0.94**
	Family	Hybotidae	0.16
	Genus	Crossopalpus	0.13
	Species	*Crossopalpu nigritellus*	0.13

**Bootstrap support ranges from 0 to 1. These values can be filtered using appropriate cutoff values that vary according to taxonomic rank and query sequence length to ensure 95 or 99% accuracy. Assumes that the query sequence is in the reference sequence database. **Indicates the finest resolution for the taxonomic assignment to ensure 99% correct assignments for a COI metabarcode ~200 bp in length.*

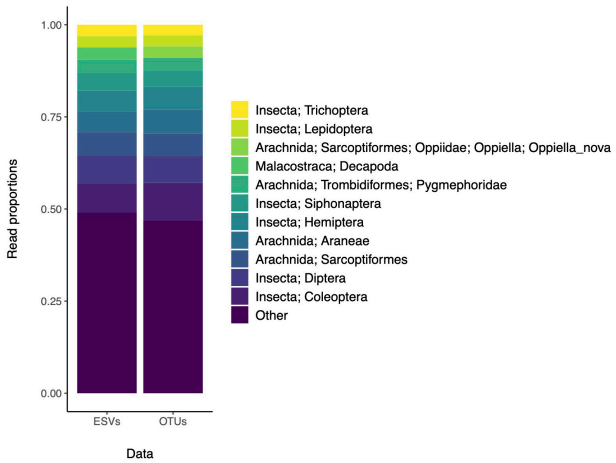


FIGURE 2 | The most abundant recovered taxa based on ESVs or OTUs are similar. Data is from a study that assessed arthropod diversity using COI metabarcoding of forest soil (Porter et al., 2019). The top 10 most abundant taxa are color coded according to the legend. Remaining clusters are binned into the "Other" category.

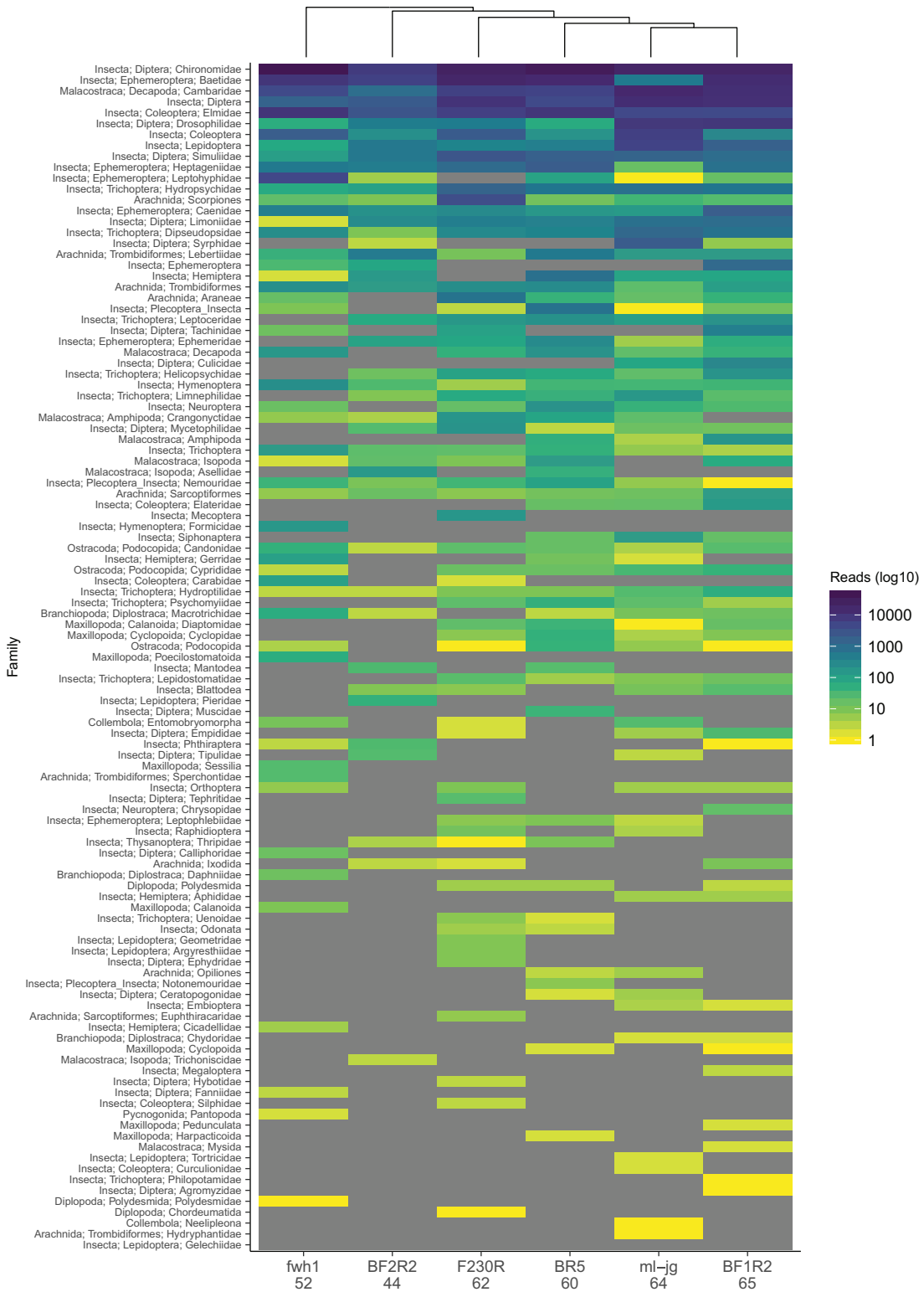


FIGURE 3 | Observed community composition can vary based on choice of COI amplicon. Data is from a study that assessed arthropod diversity using COI metabarcoding of 6 amplicons from freshwater kick net samples (Hajibabaei et al., 2019). Number of unique taxa detected from each primer set is indicated below the COI amplicon name on the *x*-axis. Results are based on ESVs whose taxonomic assignments have been summarized to the family rank where possible on the *y*-axis and ordered by decreasing read number (see legend). A UPGMA dendrogram is shown above the heatmap, indicating which amplicons recover communities that are most similar to each other. Fields in "gray" indicate that zero reads were detected.

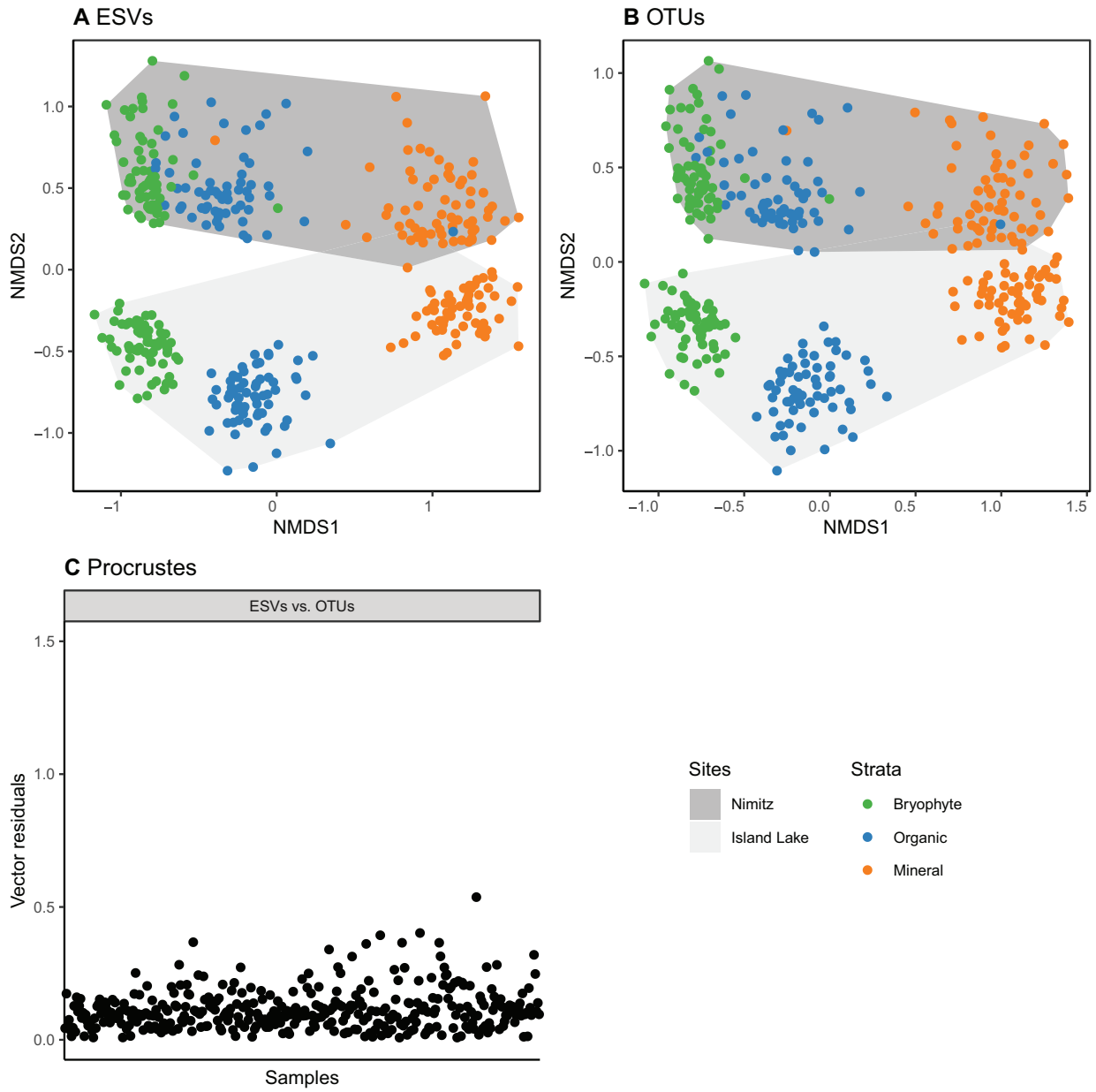


FIGURE 4 | Sample sites and soil strata are similarly distinguished using either ESV or OTU clusters. Data is from a study that assessed arthropod diversity using COI metabarcoding from forest soil (Porter et al., 2019). Parts **(A,B)** show that non-metric multidimensional scaling plots based on binary Bray Curtis (Sorensen) dissimilarities. A Procrustes analysis (least squares orthogonal mapping) was used in part to assess differences in the ordinations based on the analysis of ESVs and OTUs. The vector residuals plotted in **(C)** show the differences between the original ESV ordination and the OTU ordination. Smaller residuals indicate smaller differences between the ordinations.

at the species rank and 99% accuracy at all other ranks. As expected, we detected greater number of unique ESVs (3,357) than OTUs (2,078) (**Figure 1**). We also, however, found a similar distribution in taxonomic assignment resolution with almost half the clusters being identified to species, and just over half resolved to more inclusive ranks from genus to order. In the original study, analyzing the data with ESVs or OTUs did not make a difference to our final conclusions, so the final data was presented using ESVs.

We also assessed whether community composition patterns were affected by the use of ESVs or OTUs (**Figure 2**). The top 10 most abundant taxa are found in similar proportions whether the data are analyzed according to ESVs or OTUs. Again, in the original study, the analysis of ESVs or OTUs showed similar patterns and the final results were shown using ESVs. The taxonomic resolution of these results are typical of most studies, where many sequence clusters cannot yet be assigned to the species rank with confidence, and

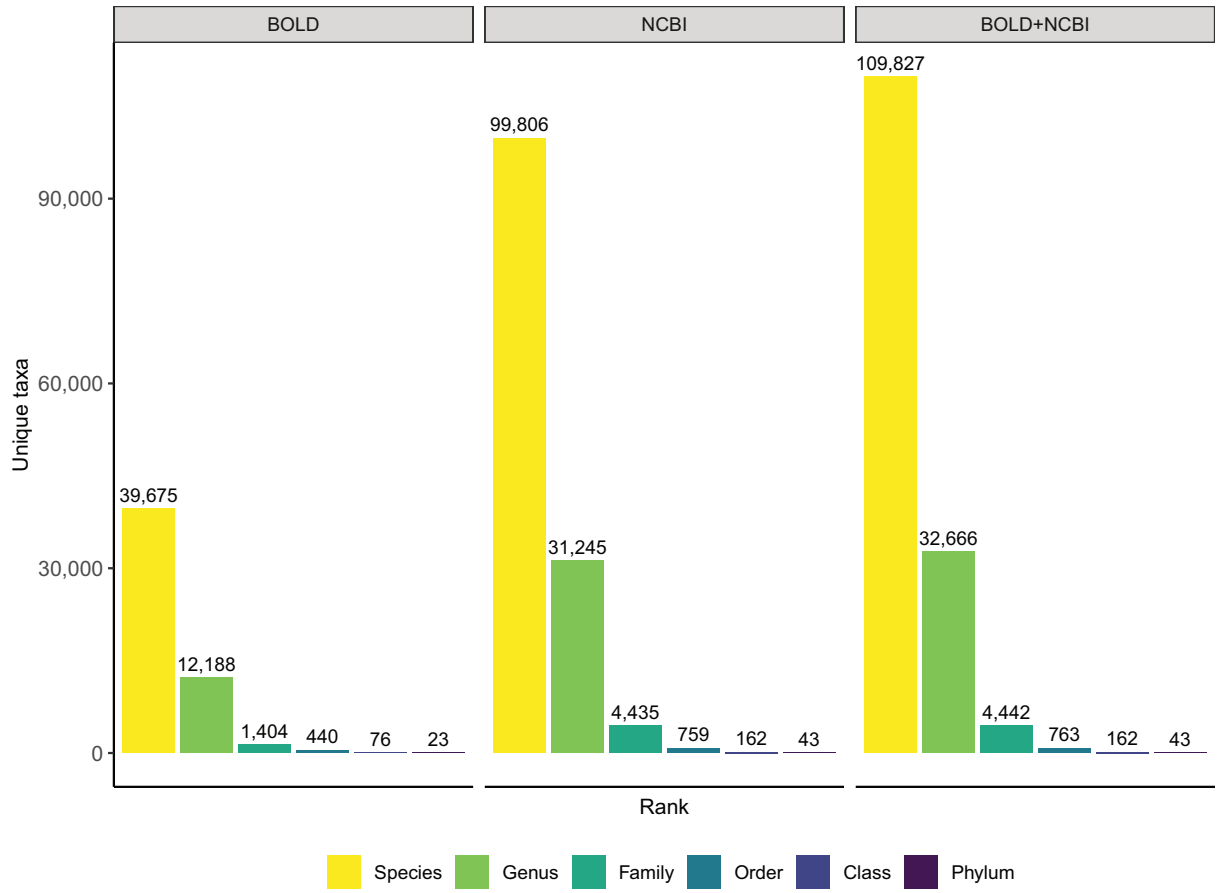


FIGURE 5 | Merging COI sequences from the BOLD and NCBI nucleotide database improves taxonomic coverage. Comparison of high-quality eukaryote COI reference sequences at a variety of ranks. Data is based on the BOLD data releases (no more than 3 Ns in the barcode sequence and at least 500 bp long), the NCBI nucleotide database accessed April 2019 (no Ns, at least 500 bp long, human and bacterial contaminants removed), and the combined database released in the CO1 Classifier v4 that merges data from BOLD and GenBank.

indicate where to target additional barcoding efforts. This is especially important for geographic locations that are poorly sampled, where diversity is high, and where the reference database is incomplete.

Both richness and community composition can be assessed based on metabarcoding data generated using a single primer set, but how would these results be affected if the primers were found to be biased in some way? Some of the early microbiome literature used only a single primer set to produce single amplicon datasets, and this has facilitated large scale studies and brought a measure of standardization to the field (Gilbert et al., 2014; Thompson et al., 2017). There are many examples, however, showing the effect of primer bias for a variety of commonly used metabarcoding primers (Hong et al., 2009; Bellemain et al., 2010; Clarke et al., 2014; Gibson et al., 2014; Elbrecht et al., 2019; Hajibabaei et al., 2019). There is also difficulty in designing "universal" COI primers to capture broad swaths of phylogenetic diversity and a switch to a multi-marker approach has been proposed for assessing animal diversity (Deagle et al., 2014). In the microbiome literature, there has been a shift to the use of PCR-free metagenomic methods to both avoid PCR-bias as well as to aid in quantitative assessments (Nayfach and Pollard, 2016). PCR-free methods have also been proposed to study terrestrial arthropod biodiversity, but these approaches are not often used due to cost and technical challenges for application in large scale studies (Zhou et al., 2013; Shokralla et al., 2016). For now, the most cost-effective approach to capture a wide array of phylogenetic diversity using COI metabarcoding is to use multiple primers sets.

To look at the effect of primer bias, we reanalyzed the data from a study that used 6 different COI metabarcode amplicons to sample arthropods from freshwater kick net samples (Hajibabaei et al., 2019). This study includes two COI amplicons that we have routinely used in our own work to survey freshwater macroinvertebrates, BR5 (B/ArR5) and F230R (LCO1490/230_R) (Folmer et al., 1994; Hajibabaei et al., 2012; Gibson et al., 2014, 2015); a primer set designed for marine taxa but has been shown to perform well for detecting arthropods in other environments, ml-jg (mlCOIintF/jgHCO2198) (Geller et al., 2013; Leray et al., 2013); as well as a few other recently published primer sets that look promising for macroinvertebrate

biomonitoring, BF1R2 (BF2/BR2), BF2R2 (BF2/BR2) (Elbrecht and Leese, 2017), and fwh1 (fwhF1/fwhR1) (Vamos et al., 2017). Taxonomic assignments were carried out as described above using the naive Bayesian classifier and summarized to the family rank where possible. Read number was normalized across each amplicon using rarefaction to account for differences in library size. We compared the results for each COI amplicon and found similarities among taxa represented by the greatest number of reads and many differences among taxa represented by fewer reads (**Figure 3**). Binary data was also used to create a Jaccard dissimilarity matrix to generate the UPGMA dendrogram clustering the COI amplicons. Community dissimilarities across amplicons ranged from 32 to 56%, with the community detected by ml-jg and BF1R2 being the most similar. The number of unique taxa detected by each amplicon ranged from 44 to 65, with the highest number of unique taxa detected by BF1R2. The

TABLE 2 | Reference sequence databases useful for taxonomically assigning metabarcodes.

Database	Content/Markers (Taxa)	Number of reference sequences	Website	References
International Nucleotide Sequence Database Collaboration (INSDC)	Repository for raw sequence data, alignments/assemblies/ annotations, sample/experimental metadata available through the NCBI, ENA, DDJB *	216,531,829 in GenBank [April 2020]	http://www.insdc.org/	Cochrane et al., 2016
Barcode of Life Data System v4 (BOLD)	COI (mostly), rbcL, matK, ITS (eukaryotes)	Available for searching: 7,389,954 COI (public and private BOLD + INSD); 2,027,132 COI (public BOLD + INSD) Available for download: 2,869,168 in data release packages	https://www.boldsystems.org/	Ratnasingham and Hebert, 2007
SILVA release 138	16S + 18S SSU, 23S + 28S LSU (bacteria, archaea, eukaryotes)	510,984–9,469,656 SSU; TBD LSU**	https://www.arb-silva.de/	Pruesse et al., 2007; Yilmaz et al., 2014
Greengenes 13.5	16S (bacteria, archaea)	1,262,986	https://greengenes.secondgenome.com/	DeSantis et al., 2006; McDonald et al., 2012
Genome Taxonomy Database (GTDB release 89)	120 proteins and 16S SSU (bacteria, archaea)	145,904 genomes; 284,051 SSU	https://gtdb.ecogenomic.org/	Parks et al., 2020
Ribosomal Database Project (RDP) release 11	16S SSU (bacteria + archaea), 28S LSU (Fungi)***	3,196,041 (bacteria) + 160,767 (archaea) SSU; 125,525 (fungi) LSU	https://rdp.cme.msu.edu/	Cole et al., 2014
The All-Species Living Tree Project (LTP) 132 (SSU) + 123 (LSU)	16S + 23S type strains (bacteria, archaea)	13,903 SSU; 1,614 LSU	https://www.arb-silva.de/projects/living-tree	Yilmaz et al., 2014
The Protist Ribosomal Reference Database (PR^2) v4.12.0	16S, 18S (protists plus metazoans, land plants, macrosporic fungi, and eukaryotic organelle outgroups)	6,010 16S; 177,934 18S	https://pr2-database.org/	Guillou et al., 2012
ITS2 database V	ITS2 (eukaryotes)	711,172	http://its2.bioapps.biozentrum.uni-wuerzburg.de/	Ankenbrand et al., 2015
UNITE v8.2	ITS fungi/eukaryotes (UNITE + INSD)	714,329 fungi; 1,796,591 eukaryotes	https://unite.ut.ee/	Kõljalg et al., 2019
PLANiTS	ITS (plants)	104,584 ITS1; 101,584 ITS2; 104,342 ITS	https://github.com/apallavicini/PLANiTS	Banchi et al., 2020
R-Syst:Diatom v7	18S, 28S, ITS, rbcL, COI (diatoms)	2,647 18S; 315 28S; 293 COI; 83 ITS2; 3,504 rbcL	https://www6.inrae.fr/r-syst_eng/Databases/R-Syst-diatom	Rimet et al., 2019
MitoFish	Mitochondrial genomes (fish)	2,853 genomes	http://mitofish.aori.u-tokyo.ac.jp/	Sato et al., 2018
rbcL Bell	rbcL (plants)	87,352	https://figshare.com/collections/rbcL_reference_library/3466311/1	Bell et al., 2017

**National Centre for Biotechnology Information (NCBI), European Nucleotide Archive (ENA), DNA Data Bank of Japan (DDBJ). **To be determined (TBD), LSU has not been released yet. ***A fungal ITS classifier is also provided.*

total number of unique taxa detected by all 6 COI amplicons was 109. It is clear from our example that taxa represented by the greatest number of reads tend to be similar across amplicons, but combining the results from multiple amplicons improves the overall recovery of the greatest diversity of taxa. In the original study, we showed that using at least two COI amplicons from this set of six could detect most species, genera, and families. Previous work has used *in silico* PCR using ITS primers to detect fungi (Bellemain et al., 2010) and mock community studies in bacterial (Brooks et al., 2015) and terrestrial arthropod communities (Elbrecht et al., 2019) to demonstrate the effect of PCR bias. Here we show the effect of primer bias on a real community with realistic complexity and template background.

We have shown that alpha diversity, richness, is sensitive both to choice of metabarcode cluster type and primer choice, but what does this mean for beta diversity? For arthropods sampled using COI metabarcoding from freshwater or soil samples, beta diversity assessments have been shown to be robust to both variations in primer choice and sampling method (Hajibabaei et al., 2019; Porter et al., 2019). Does this hold true for differences in clustering strategy and resolution of the matrix? In our research we have found that beta diversity estimates are robust to the use of either ESVs or OTUs (**Figure 4**). The difference between ordinations based on ESVs and OTUs is minimal, and the site and soil layer groupings are visually distinct using either sequence cluster type. In the original study, clustering patterns observed from NMDS plots and permutational analysis of variance (PERMANOVA) results were not affected by the analysis of ESVs or OTUs. As a result, we prefer the use of ESVs over OTUs to improve reproducibility, facilitate comparisons across studies, and permit within-species analyses.

HOW CAN WE LEVERAGE TAXONOMIC COVERAGE ACROSS REFERENCE DATABASES?

The composition, quality, and completeness of reference sequences databases determines our ability to identify unknown specimens using DNA barcodes and metabarcodes. BOLD has become the canonical COI reference sequence database, with official DNA barcode sequences available for download through data releases available from https://www.boldsystems.org/index.php/datarelease. The BOLD system also contains sequences mined from GenBank as well as private data that is available for comparison when using the BOLD identification engine (Ratnasingham and Hebert, 2007). Recently an R package was released that facilitates mining BOLD data; however, it can still be challenging to retrieve large amounts of data at one time, for example, the entire reference database of all arthropoda (Chamberlain, 2019). The NCBI nucleotide database, GenBank, has accumulated over 2.5 million COI sequences

TABLE 3 | Curated reference sequence databases specifically formatted to work with the RDP naive Bayesian classifier.

Marker	Name version (year)	Target taxa	Number of reference sequences	Availability	References*
SSU (16S)	16S trainsetNo16 (2016)	Prokaryotes	13,212	https://sourceforge.net/projects/rdp-classifier/	Wang et al., 2007
SSU (18S)	18S classifier v4** (2020)	Eukaryotes	42,301	https://github.com/terrimporter/18SClassifier	Pruesse et al., 2007
SSU	SSU Diatom Classifier v1.0 (2020)	Diatoms	2,962	https://github.com/terrimporter/SSUdiatomClassifier	Rimet et al., 2019
LSU	Fungi LSU trainsetNo11 (2014)	Fungi	11,442	https://sourceforge.net/projects/rdp-classifier/	Liu et al., 2012
ITS	Fungalits UNITE 07042014 (2014)	Fungi	145,019	https://sourceforge.net/projects/rdp-classifier/	Abarenkov et al., 2010
ITS	Fungalits Warcup v2 (2016)	Fungi	17,878	https://sourceforge.net/projects/rdp-classifier/	Deshpande et al., 2016
rbcL	rbcL Classifier v1 (2020)	Eukaryotes	164,454	https://github.com/terrimporter/rbcLClassifier	Benson et al., 2012
rbcL	rbcL Diatom Classifier v1.0 (2020)	Diatoms	3,504	https://github.com/terrimporter/rbcLdiatomClassifier	Rimet et al., 2019; Maitland et al., 2020
COI	CO1 Classifier v4 (2019)	Eukaryotes	1,221,528	https://github.com/terrimporter/CO1Classifier	Porter and Hajibabaei, 2018a
12S	12S fish Classifier v1.0 (2020)	Fish	2,853	https://github.com/terrimporter/12SfishClassifier	Iwasaki et al., 2013

**References for the database where sequences were obtained and/or for the trained naive Bayesian classifier if available. **Based on SILVA 138 SSURef Nr99.*

since the advent of the DNA barcoding initiative in 2003 (Hebert et al., 2003; Benson et al., 2012; Porter and Hajibabaei, 2018b). Since BOLD has a policy of depositing DNA barcodes in GenBank, much of the public BOLD data is also available through GenBank. Neither BOLD nor GenBank, however, is entirely complete, and each database provides complementary taxonomic coverage as has been shown for Canadian freshwater invertebrates (Curry et al., 2018). Combining these databases would improve both sequence and taxonomic coverage. Making the merged reference data available in plain text formats would make it relatively straightforward to reformat so they can be used as the basis for alternative taxonomic assignment tools such as those that provide rank-flexible statistical measures of confidence. For example, the BOLD_NCBI_Merger script provides a means to combine data from BOLD and the NCBI nucleotide database for use with MEGAN (Huson et al., 2016; Macher et al., 2017). Our own approach has been to update the underlying reference sequence database used by the COI classifier v4 to combine data from BOLD and GenBank, and it is available from https://github.com/terrimporter/CO1Classifier (Wang et al., 2007; Porter and Hajibabaei, 2018a). We demonstrate the improved taxonomic composition when COI reference sequences from the BOLD data releases are combined with COI sequences mined from GenBank (**Figure 5**). The combined reference set is available as a FASTA file as are the trained files needed to use these reference sets with the naive Bayesian classifier.

We have mainly focused on using a single marker, such as COI for animal metabarcoding, but the field has progressed such that investigators are now using multi-marker approaches (Drummond et al., 2015) to conduct food web studies or comprehensive biodiversity monitoring across phylogenetically diverse taxa. As such, we should be aware of tools available for analyzing other widely used metabarcoding markers (**Table 2**). The largest source for reference sequence information is through the International Nucleotide Sequence Database Collaboration (INSDC) comprised of the NCBI (GenBank, Short Read Archive), EMBL-EBI, and DDJB. In North America, most users are familiar with GenBank, a repository for marker gene sequences (also see European Nucleotide Archive and DDJB), and the Short Read Archive (SRA) where raw metabarcode reads are stored. For COI barcodes, public data in BOLD is automatically transferred to GenBank, and additional barcode sequences are retrieved from GenBank to complement the BOLD database. Multi-marker or genome projects focused on particular taxonomic groups are also valuable sources of reference sequence information. For example, DNA barcodes found to be most useful for diatom identification includes 18S, 28S rDNA, internal transcribed spacer 2 (ITS2), rbcL cpDNA, and COI mtDNA and are available through the Diat.barcode library (Chaumeil et al., 2018; Rimet et al., 2019). Additionally, though COI DNA barcodes are readily available for fish identification (Becker et al., 2011; Weigand et al., 2019), 12S mtDNA has a history of use for vertebrate detection (Kitano et al., 2007; Sato et al., 2018). Throughout the course of our own work, we have mined existing databases and created our own curated reference sets reformatted to work with a naive Bayesian classifier to make rank-flexible taxonomic assignments with a statistical measure of confidence (**Table 3**). Each of these curated datasets are also available as FASTA files. These resources show how the field of eukaryote metabarcoding is diversifying to use multiple markers and support a variety of taxonomic assignment methods.

Choosing a database for any given DNA barcode or marker often comes down to one's preferred species concept, database coverage, as well as the availability and ease-of-use of related tools. The NCBI database is the primary source of raw sequence data for most of the databases listed in **Table 2**. What makes each of the rRNA gene databases unique, however, is that they filter the data using their own quality control standards, and they follow their own taxonomic roadmap (Balvočiūtė and Huson, 2017). For example, a phylogenetic species concept is often preferred in microbial ecology where taxa are challenging to study and describe using traditional methods and undescribed environmental diversity is exceedingly high. In this case, both Greengenes and SILVA assume that trees based on available SSU sequences reflect evolutionary relatedness, and any taxonomic inconsistencies are resolved to make classification consistent with phylogeny. The RDP, however, follows Bergey's classification system (Cole et al., 2014). When the goal is to identify unknown environmental sequences from metabarcode sequences, the so-called "dark taxa," the microbial and fungal communities have come up with their own methods. For prokaryotes, the GTDB includes metagenome assembled genomes (MAGs) represented in their database (Chaumeil et al., 2019). The RDP, SILVA, and Greengenes databases each contain many environmental sequences for comparison, but the taxonomic assignment can be based on different criteria using an algorithm (RDP) or phylogenetic placement and manual curation (SILVA, Greengenes). For fungi, the UNITE database has made a concerted effort to incorporate fungal dark taxa in their SHs and have introduced Taxon Hypotheses (THs) to allow for the communication of SHs using different classification schemes at the same time (Nilsson et al., 2019). If a fungal or animal study requires species estimates, then using a database that attempts to approximate species using fungal SHs or animal COI BINs may be preferred. For studies where few taxa can be confidently identified, using a large database that includes environmental sequences will provide the most coverage, and using a method that provides a statistical measure of confidence can allow the user to adjust for the recovery of false negatives or false positives according to the study aims (Edgar, 2018a).

CONCLUDING REMARKS

Over the last 15 years the use of standardized DNA-based biodiversity markers such as DNA barcodes has become a routine practice in various scientific and socioeconomic endeavors. A much wider spatiotemporal biodiversity perspective is now achievable through bulk analysis of metabarcodes. Our ability to fully identify metabarcodes from particularly diverse taxonomic groups or samples may be currently limited, but

with continued DNA barcoding efforts these databases are expected to become more representative over time. Insufficiently identified sequence clusters, those not confidently identified to the species rank, can still be used for biodiversity analyses including richness assessment, community composition, and beta diversity assessments. For improved reproducibility, comparison across studies, and nucleotide-level resolution, we encourage the use of ESV level analyses. For studies that require species estimates, we suggest aligning ESVs to fungal ITS SHs or animal COI BINs which both attempt to approximate species units. If representative BIN sequences were made available in an easily parsed format, this would allow taxonomic assignments to be made using tools outside the BOLD system built-in barcode identification engine and would allow inclusion in metabarcode bioinformatic pipelines that are already widely used for analyzing large metabarcode datasets. COI metabarcoding offers a sophisticated toolset and reference databases suitable for large scale studies; as such, it is now firmly established as a marker for animals in molecular ecological and biodiversity studies.

DATA AVAILABILITY STATEMENT

Code used to generate figures is available on GitHub from https://github.com/terrimporter/PorterHajibabaei2020_ESVs_vs_OTUs.

AUTHOR CONTRIBUTIONS

MH conceived of the idea. TP and MH wrote and edited the manuscript. TP generated the figures. Both authors contributed to the article and approved the submitted version.

ACKNOWLEDGMENTS

We would like to thank the Hajibabaei group for their support and collaboration. This study is funded by the Government of Canada through Genome Canada and Ontario Genomics.

REFERENCES

Abarenkov, K., Nilsson, R. H., Larsson, K.-H., Alexander, I. J., Eberhardt, U., Erland, S., et al. (2010). The UNITE database for molecular identification of fungi – recent updates and future perspectives. *New Phytol.* 186, 281–285. doi: 10.1111/j.1469-8137.2009.03160.x

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., et al. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25, 3389–3402. doi: 10.1093/nar/25.17.3389

Amir, A., McDonald, D., Navas-Molina, J. A., Kopylova, E., Morton, J. T., Zech Xu, Z., et al. (2017). Deblur rapidly resolves single-nucleotide community sequence patterns. *mSystems* 2:e00191-16. doi: 10.1128/mSystems.00191-16

Ankenbrand, M. J., Keller, A., Wolf, M., Schultz, J., and Förster, F. (2015). ITS2 database V: twice as much. *Mol. Biol. Evol.* 32, 3030–3032. doi: 10.1093/molbev/msv174

Balvočiūtė, M., and Huson, D. H. (2017). SILVA, RDP, Greengenes, NCBI and OTT — how do these taxonomies compare? *BMC Genomics* 18:114. doi: 10.1186/s12864-017-3501-4

Banchi, E., Ametrano, C. G., Greco, S., Stanković, D., Muggia, L., and Pallavicini, A. (2020). PLANiTS: a curated sequence reference dataset for plant ITS DNA metabarcoding. *Database* 2020:baz155. doi: 10.1093/database/baz155

Barsoum, N., Bruce, C., Forster, J., Ji, Y.-Q., and Yu, D. W. (2019). The devil is in the detail: metabarcoding of arthropods provides a sensitive measure of biodiversity response to forest stand composition compared with surrogate measures of biodiversity. *Ecol. Indic.* 101, 313–323. doi: 10.1016/j.ecolind.2019.01.023

Basset, Y., Cizek, L., Cuénoud, P., Didham, R. K., Guilhaumon, F., Missa, O., et al. (2012). Arthropod diversity in a tropical forest. *Science* 338, 1481–1484. doi: 10.1126/science.1226727

Becker, S., Hanner, R., and Steinke, D. (2011). Five years of FISH-BOL: brief status report. *Mitochondrial DNA* 22, 3–9. doi: 10.3109/19401736.2010.535528

Bell, K. L., Loeffler, V. M., and Brosi, B. J. (2017). An *rbcL* reference library to aid in the identification of plant species mixtures by DNA metabarcoding. *Appl. Plant Sci.* 5:1600110. doi: 10.3732/apps.1600110

Bellemain, E., Carlsen, T., Brochmann, C., Coissac, E., Taberlet, P., and Kauserud, H. (2010). ITS as an environmental DNA barcode for fungi: an in silico approach reveals potential PCR biases. *BMC Microbiol.* 10:189. doi: 10.1186/1471-2180-10-189

Benson, D. A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., et al. (2012). GenBank. *Nucleic Acids Res.* 41, D36–D42. doi: 10.1093/nar/gks1195

Blackwell, M. (2011). The Fungi: 1, 2, 3 …5.1 million species? *Am. J. Bot.* 98, 426–438. doi: 10.3732/ajb.1000298

Braukmann, T. W. A., Ivanova, N. V., Prosser, S. W. J., Elbrecht, V., Steinke, D., Ratnasingham, S., et al. (2019). Metabarcoding a diverse arthropod mock community. *Mol. Ecol. Resour.* 19, 711–727. doi: 10.1111/1755-0998.13008

Brooks, J. P., Edwards, D. J., Harwich, M. D., Rivera, M. C., Fettweis, J. M., Serrano, M. G., et al. (2015). The truth about metagenomics: quantifying and counteracting bias in 16S rRNA studies. *BMC Microbiol.* 15:66. doi: 10.1186/s12866-015-0351-6

Buchner, D., and Leese, F. (2020). BOLDigger – a Python package to identify and organise sequences with the Barcode of Life Data systems. *Metabarcoding Metagenomics* 4:e53535. doi: 10.3897/mbmg.4.53535

Callahan, B. J., McMurdie, P. J., and Holmes, S. P. (2017). Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. *ISME J.* 11, 2639–2643. doi: 10.1038/ismej.2017.119

Callahan, B. J., McMurdie, P. J., Rosen, M. J., Han, A. W., Johnson, A. J. A., and Holmes, S. P. (2016). DADA2: high-resolution sample inference from Illumina amplicon data. *Nat. Methods* 13, 581–583. doi: 10.1038/nmeth.3869

Caporaso, J. G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F. D., Costello, E. K., et al. (2010). QIIME allows analysis of highthroughput community sequencing data. *Nat. Methods* 7, 335–336. doi: 10.1038/nmeth.f.303

Carini, P., Marsden, P. J., Leff, J. W., Morgan, E. E., Strickland, M. S., and Fierer, N. (2016). *Relic DNA is Abundant in Soil and Obscures Estimates of Soil Microbial Diversity*. Available online at: http://biorxiv.org/lookup/doi/10.1101/043372 (accessed July 6, 2016).

Chamberlain, S. (2019). *bold: Interface to Bold Systems API*. Available online at: https://CRAN.R-project.org/package=bold (accessed July 23, 2020).

Chaumeil, P., Fischer-Le Saux, M., Frigerio, J.-M., Grenier, E., Rimet, F., Streito, J.-C., et al. (2018). *R-Syst: A Network Providing Curated Molecular Databases and Data Analysis Tools for Taxonomy and Systematics (Prokaryotes and Eucaryotes)*. Available online at: https://doi.org/10.15454/OEDAUS.

Chaumeil, P.-A., Mussig, A. J., Hugenholtz, P., and Parks, D. H. (2019). GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. *Bioinformatics* btz848. doi: 10.1093/bioinformatics/btz848 [Epub ahead of print].

Clarke, L. J., Soubrier, J., Weyrich, L. S., and Cooper, A. (2014). Environmental metabarcodes for insects: *in silico* PCR reveals potential for taxonomic bias. *Mol. Ecol. Resour.* 14, 1160–1170. doi: 10.1111/1755-0998.12265

Cochrane, G., Karsch-Mizrachi, I., Takagi, T., and Sequence Database Collaboration, I. N. (2016). The international nucleotide sequence database collaboration. *Nucleic Acids Res.* 44, D48–D50. doi: 10.1093/nar/gkv1323

Cole, J. R., Wang, Q., Fish, J. A., Chai, B., McGarrell, D. M., Sun, Y., et al. (2014). Ribosomal database project: data and tools for high throughput rRNA analysis. *Nucleic Acids Res.* 42, D633–D642. doi: 10.1093/nar/gkt1244

Creer, S., Deiner, K., Frey, S., Porazinska, D., Taberlet, P., Thomas, W. K., et al. (2016). The ecologist's field guide to sequence-based identification of biodiversity. *Methods Ecol. Evol.* 7, 1008–1018. doi: 10.1111/2041-210X.12574

Curry, C. J., Gibson, J. F., Shokralla, S., Hajibabaei, M., and Baird, D. J. (2018). Identifying North American freshwater invertebrates using DNA barcodes: are existing COI sequence libraries fit for purpose? *Freshw. Sci.* 37, 178–189. doi: 10.1086/696613

Darling, J. A. (2019). How to learn to stop worrying and love environmental DNA monitoring. *Aquat. Ecosyst. Health Manage.* 22, 440–451. doi: 10.1080/14634988.2019.1682912

D'Costa, V. M., King, C. E., Kalan, L., Morar, M., Sung, W. W. L., Schwarz, C., et al. (2011). Antibiotic resistance is ancient. *Nature* 477, 457–461. doi: 10.1038/nature10388

Deagle, B. E., Jarman, S. N., Coissac, E., Pompanon, F., and Taberlet, P. (2014). DNA metabarcoding and the cytochrome c oxidase subunit I marker: not a perfect match. *Biol. Lett.* 10:20140562. doi: 10.1098/rsbl.2014.0562

Deiner, K., Bik, H. M., Mächler, E., Seymour, M., Lacoursière-Roussel, A., Altermatt, F., et al. (2017). Environmental DNA metabarcoding: transforming how we survey animal and plant communities. *Mol. Ecol.* 26, 5872–5895. doi: 10.1111/mec.14350

DeSantis, T. Z., Hugenholtz, P., Larsen, N., Rojas, M., Brodie, E. L., Keller, K., et al. (2006). Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Appl. Environ. Microbiol.* 72, 5069–5072. doi: 10.1128/AEM.03006-05

Deshpande, V., Wang, Q., Greenfield, P., Charleston, M., Porras-Alfaro, A., Kuske, C. R., et al. (2016). Fungal identification using a Bayesian classifier and the Warcup training set of internal transcribed spacer sequences. *Mycologia* 108, 1–5. doi: 10.3852/14-293

Drummond, A. J., Newcomb, R. D., Buckley, T. R., Xie, D., Dopheide, A., Potter, B. C., et al. (2015). Evaluating a multigene environmental DNA approach for biodiversity assessment. *GigaScience* 4:46. doi: 10.1186/s13742-015-0086-1

Ebach, M. C., Valdecasas, A. G., and Wheeler, Q. D. (2011). Impediments to taxonomy and users of taxonomy: accessibility and impact evaluation. *Cladistics* 27, 550–557. doi: 10.1111/j.1096-0031.2011.00348.x

Edgar, R. C. (2013). UPARSE: highly accurate OTU sequences from microbial amplicon reads. *Nat. Methods* 10, 996–998. doi: 10.1038/nmeth.2604

Edgar, R. C. (2016). UNOISE2: improved error-correction for Illumina 16S and ITS amplicon sequencing. *bioRxiv* [Preprint]. doi: 10.1101/081257

Edgar, R. C. (2018a). Accuracy of taxonomy prediction for 16S rRNA and fungal ITS sequences. *PeerJ* 6:e4652. doi: 10.7717/peerj.4652

Edgar, R. C. (2018b). Updating the 97% identity threshold for 16S ribosomal RNA OTUs. *Bioinformatics.* 34, 2371–2375. doi: 10.1093/bioinformatics/bty113

Elbrecht, V., Braukmann, T. W. A., Ivanova, N. V., Prosser, S. W. J., Hajibabaei, M., Wright, M., et al. (2019). Validation of COI metabarcoding primers for terrestrial arthropods. *PeerJ* 7:e7745. doi: 10.7717/peerj.7745

Elbrecht, V., and Leese, F. (2017). Validation and development of COI metabarcoding primers for freshwater macroinvertebrate bioassessment. *Front. Environ. Sci.* 5:11. doi: 10.3389/fenvs.2017.00011

Elbrecht, V., Vamos, E. E., Meissner, K., Aroviita, J., and Leese, F. (2017). Assessing strengths and weaknesses of DNA metabarcoding-based macroinvertebrate identification for routine stream monitoring. *Methods Ecol. Evol.* 8, 1265–1275. doi: 10.1111/2041-210X.12789

Elbrecht, V., Vamos, E. E., Steinke, D., and Leese, F. (2018). Estimating intraspecific genetic diversity from community DNA metabarcoding data. *PeerJ* 6:e4644. doi: 10.7717/peerj.4644

Erdozain, M., Thompson, D. G., Porter, T. M., Kidd, K. A., Kreutzweiser, D. P., Sibley, P. K., et al. (2019). Metabarcoding of storage ethanol vs. conventional morphometric identification in relation to the use of stream macroinvertebrates as Ecol. Indic. in forest management. *Ecol. Indic.* 101, 173–184. doi: 10.1016/j.ecolind.2019.01.014

Folmer, O., Black, M., Hoeh, W., Lutz, R., and Vrijenhoek, R. (1994). DNA primers for amplification of mitochondrial cytochrome c oxidase subunit I from diverse metazoan invertebrates. *Mol. Mar. Biol. Biotechnol.* 3, 294–299.

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

Geller, J., Meyer, C., Parker, M., and Hawk, H. (2013). Redesign of PCR primers for mitochondrial cytochrome c oxidase subunit I for marine invertebrates and application in all-taxa biotic surveys. *Mol. Ecol. Resour.* 13, 851–861. doi: 10.1111/1755-0998.12138

Gibson, J., Shokralla, S., Curry, C., Baird, D. J., Monk, W. A., King, I., et al. (2015). Large-scale biomonitoring of remote and threatened ecosystems via high-throughput sequencing. *PLoS One* 10:e0138432. doi: 10.1371/journal.pone.0138432

Gibson, J., Shokralla, S., Porter, T. M., King, I., van Konynenburg, S., Janzen, D. H., et al. (2014). Simultaneous assessment of the macrobiome and microbiome in a bulk sample of tropical arthropods through DNA metasystematics. *Proc. Natl. Acad. Sci. U.S.A.* 111, 8007–8012. doi: 10.1073/pnas.1406468111

Gilbert, J. A., Jansson, J. K., and Knight, R. (2014). The Earth Microbiome project: successes and aspirations. *BMC Biol.* 12:69. doi: 10.1186/s12915-014-0069-1

Glassman, S. I., and Martiny, J. B. (2018). Ecological patterns are robust to use of exact sequence variants versus operational taxonomic units. *mSphere* 3:e00148-18. doi: 10.1101/283283

Guillou, L., Bachar, D., Audic, S., Bass, D., Berney, C., Bittner, L., et al. (2012). The Protist Ribosomal Reference database (PR2): a catalog of unicellular eukaryote Small Sub-Unit rRNA sequences with curated taxonomy. *Nucleic Acids Res.* 41, D597–D604. doi: 10.1093/nar/gks1160

Hajibabaei, M. (2012). The golden age of DNA metasystematics. *Trends Genet.* 28, 535–537. doi: 10.1016/j.tig.2012.08.001

Hajibabaei, M., deWaard, J. R., Ivanova, N. V., Ratnasingham, S., Dooh, R. T., Kirk, S. L., et al. (2005). Critical factors for assembling a high volume of DNA barcodes. *Philos. Trans. R. Soc. B* 360, 1959–1967. doi: 10.1098/rstb.2005.1727

Hajibabaei, M., Porter, T. M., Wright, M., and Rudar, J. (2019). COI metabarcoding primer choice affects richness and recovery of indicator taxa in freshwater systems. *PLoS One* 14:e0220953. doi: 10.1371/journal.pone.0220953

Hajibabaei, M., Shokralla, S., Zhou, X., Singer, G. A. C., and Baird, D. J. (2011). Environmental barcoding: a next-generation sequencing approach for biomonitoring applications using river benthos. *PLoS One* 6:e17497. doi: 10.1371/journal.pone.0017497

Hajibabaei, M., Smith, M. A., Janzen, D. H., Rodriguez, J. J., Whitfield, J. B., and Hebert, P. D. N. (2006). A minimalist barcode can identify a specimen whose DNA is degraded: BARCODING. *Mol. Ecol. Notes* 6, 959–964. doi: 10.1111/j.1471-8286.2006.01470.x

Hajibabaei, M., Spall, J. L., Shokralla, S., and van Konynenburg, S. (2012). Assessing biodiversity of a freshwater benthic macroinvertebrate community through non-destructive environmental barcoding of DNA from preservative ethanol. *BMC Ecol.* 12:28. doi: 10.1186/1472-6785-12-28

Handelsman, J. (2004). Metagenomics: application of genomics to uncultured microorganisms. *Microbiol. Mol. Biol. Rev.* 68, 669–685. doi: 10.1128/mmbr.68.4.669-685.2004

Hänfling, B., Lawson Handley, L., Read, D. S., Hahn, C., Li, J., Nichols, P., et al. (2016). Environmental DNA metabarcoding of lake fish communities reflects long-term data from established survey methods. *Mol. Ecol.* 25, 3101–3119. doi: 10.1111/mec.13660

He, Y., Caporaso, J. G., Jiang, X.-T., Sheng, H.-F., Huse, S. M., Rideout, J. R., et al. (2015). Stability of operational taxonomic units: an important but neglected property for analyzing microbial diversity. *Microbiome* 3:20. doi: 10.1186/s40168-015-0081-x

Hebert, P. D. N., Braukmann, T. W. A., Prosser, S. W. J., Ratnasingham, S., deWaard, J. R., Ivanova, N. V., et al. (2018). A Sequel to Sanger: amplicon sequencing that scales. *BMC Genomics* 19:219. doi: 10.1186/s12864-018-4611-3

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

Hernandez, C., Bougas, B., Perreault-Payette, A., Simard, A., Côté, G., and Bernatchez, L. (2020). 60 specific eDNA qPCR assays to detect invasive, threatened, and exploited freshwater vertebrates and invertebrates in Eastern Canada. *Environ. DNA.* 2, 373–386. doi: 10.1002/edn3.89

Hobern, D., and Hebert, P. (2019). BIOSCAN - revealing eukaryote diversity, dynamics, and interactions. *Biodivers. Inf. Sci. Stand.* 3:e37333. doi: 10.3897/biss.3.37333

Hobern, D. G. (2020). BIOSCAN: DNA Barcoding to accelerate taxonomy and biogeography for conservation and sustainability. *Genome*. doi: 10.1139/gen-2020-0009 [Epub ahead of print].

Hong, S., Bunge, J., Leslin, C., Jeon, S., and Epstein, S. S. (2009). Polymerase chain reaction primers miss half of rRNA microbial diversity. *ISME J.* 3, 1365–1373. doi: 10.1038/ismej.2009.89

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

Iwasaki, W., Fukunaga, T., Isagozawa, R., Yamada, K., Maeda, Y., Satoh, T. P., et al. (2013). MitoFish and MitoAnnotator: a mitochondrial genome database of fish with an accurate and automatic annotation pipeline. *Mol. Biol. Evol.* 30, 2531–2540. doi: 10.1093/molbev/mst141

Kitano, T., Umetsu, K., Tian, W., and Osawa, M. (2007). Two universal primer sets for species identification among vertebrates. *Int. J. Legal Med.* 121, 423–427. doi: 10.1007/s00414-006-0113-y

Kõljalg, U., Abarenkov, K., Nilsson, R. H., Larsson, K.-H., and Taylor, A. F. S. (2019). The UNITE database for molecular identification and for communicating fungal species. *Biodivers. Inf. Sci. Stand.* 3:e37402. doi: 10.3897/biss.3.37402

Koljalg, U., Nilsson, R. H., Abarenkov, K., Tedersoo, L., Taylor, A. F. S., Bahram, M., et al. (2013). Towards a unified paradigm for sequence-based identification of fungi. *Mol. Ecol.* 22, 5271–5277.

Leray, M., Yang, J. Y., Meyer, C. P., Mills, S. C., Agudelo, N., Ranwez, V., et al. (2013). A new versatile primer set targeting a short fragment of the mitochondrial COI region for metabarcoding metazoan diversity: application for characterizing coral reef fish gut contents. *Front. Zool.* 10:34. doi: 10.1186/1742-9994-10-34

Liu, K.-L., Porras-Alfaro, A., Kuske, C. R., Eichorst, S. A., and Xie, G. (2012). Accurate, rapid taxonomic classification of fungal large-subunit rRNA Genes. *Appl. Environ. Microbiol.* 78, 1523–1533. doi: 10.1128/AEM.06826-11

Lozupone, C. A., and Knight, R. (2007). Global patterns in bacterial diversity. *Proc. Natl. Acad. Sci. U.S.A.* 104, 11436–11440. doi: 10.1073/pnas.0611525104

Macher, J.-N., Macher, T.-H., and Leese, F. (2017). Combining NCBI and BOLD databases for OTU assignment in metabarcoding and metagenomic datasets: the BOLD_NCBI _Merger. *Metabarcoding Metagenomics* 1:e22262. doi: 10.3897/mbmg.1.22262

Maitland, V. C., Robinson, C. V., Porter, T. M., and Hajibabaei, M. (2020). Freshwater diatom biomonitoring through benthic kick-net metabarcoding. *bioRxiv* [Preprint]. doi: 10.1101/2020.05.25.115089

Marquina, D., Esparza-Salas, R., Roslin, T., and Ronquist, F. (2019). Establishing arthropod community composition using metabarcoding: surprising inconsistencies between soil samples and preservative ethanol and homogenate from Malaise trap catches. *Mol. Ecol. Resour.* 19, 1516–1530. doi: 10.1111/1755-0998.13071

McDonald, D., Price, M. N., Goodrich, J., Nawrocki, E. P., DeSantis, T. Z., Probst, A., et al. (2012). An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. *ISME J.* 6, 610–618. doi: 10.1038/ismej.2011.139

Munch, K., Boomsma, W., Huelsenbeck, J., Willerslev, E., and Nielsen, R. (2008). Statistical assignment of DNA sequences using bayesian phylogenetics. *Syst. Biol.* 57, 750–757. doi: 10.1080/10635150802422316

Nayfach, S., and Pollard, K. S. (2016). Toward accurate and quantitative comparative metagenomics. *Cell* 166, 1103–1116. doi: 10.1016/j.cell.2016.08.007

Nearing, J. T., Douglas, G. M., Comeau, A. M., and Langille, M. G. I. (2018). Denoising the Denoisers: an independent evaluation of microbiome sequence error-correction approaches. *PeerJ* 6:e5364. doi: 10.7717/peerj.5364

Nielsen, K. M., Johnsen, P. J., Bensasson, D., and Daffonchio, D. (2007). Release and persistence of extracellular DNA in the environment. *Environ. Biosaf. Res.* 6, 37–53. doi: 10.1051/ebr:2007031

Nilsson, R. H., Larsson, K.-H., Taylor, A. F. S., 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. *Nucleic Acids Res.* 47, D259–D264. doi: 10.1093/nar/gky1022

Pace, N. R., Stahl, D. A., Lane, D. J., and Olsen, G. J. (1986). "The analysis of natural microbial populations by ribosomal RNA sequences," in *Advances in Microbial Ecology Advances in Microbial Ecology*, ed. K. C. Marshall (Boston, MA: Springer), 1–55. doi: 10.1007/978-1-4757-0611-6_1

Parks, D. H., Chuvochina, M., Chaumeil, P.-A., Rinke, C., Mussig, A. J., and Hugenholtz, P. (2020). A complete domain-to-species taxonomy for Bacteria and Archaea. *Nat. Biotechnol.* [Epub ahead of print].

Pietramellara, G., Ascher, J., Borgogni, F., Ceccherini, M. T., Guerri, G., and Nannipieri, P. (2009). Extracellular DNA in soil and sediment: fate and ecological relevance. *Biol. Fertil. Soils* 45, 219–235. doi: 10.1007/s00374-008-0345-8

Porter, T. M., and Hajibabaei, M. (2018a). Automated high throughput animal CO1 metabarcode classification. *Sci. Rep.* 8:4226.

Porter, T. M., and Hajibabaei, M. (2018b). Over 2.5 million COI sequences in GenBank and growing. *PLoS One* 13:e0200177. doi: 10.1101/353904

Porter, T. M., and Hajibabaei, M. (2018c). Scaling up: a guide to high-throughput genomic approaches for biodiversity analysis. *Mol. Ecol.* 27, 313–338. doi: 10.1111/mec.14478

Porter, T. M., Morris, D. M., Basiliko, N., Hajibabaei, M., Doucet, D., Bowman, S., et al. (2019). Variations in terrestrial arthropod DNA metabarcoding methods recovers robust beta diversity but variable richness and site indicators based on exact sequence variants. *Sci. Rep.* 9:18218.

Pruesse, E., Quast, C., Knittel, K., Fuchs, B. M., Ludwig, W., Peplies, J., et al. (2007). SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Res.* 35, 7188–7196. doi: 10.1093/nar/gkm864

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

Reeder, J., and Knight, R. (2009). The 'rare biosphere': a reality check. *Nat. Methods* 6, 636–637. doi: 10.1038/nmeth0909-636

Rimet, F., Gusev, E., Kahlert, M., Kelly, M. G., Kulikovskiy, M., Maltsev, Y., et al. (2019). Diat.barcode, an open-access curated barcode library for diatoms. *Sci. Rep.* 9:15116. doi: 10.1038/s41598-019-51500-6

Rognes, T., Flouri, T., Nichols, B., Quince, C., and Mahé, F. (2016). VSEARCH: a versatile open source tool for metagenomics. *PeerJ* 4:e2584. doi: 10.7717/peerj.2584

Romanowski, G., Lorenz, M. G., and Wackernagel, W. (1991). Adsorption of plasmid DNA to mineral surfaces and protection against DNase I. *Appl. Environ. Microbiol.* 57, 1057–1061. doi: 10.1128/AEM.57.4.1057-1061.1991

Sato, Y., Miya, M., Fukunaga, T., Sado, T., and Iwasaki, W. (2018). MitoFish and MiFish Pipeline: a mitochondrial genome database of fish with an analysis pipeline for environmental DNA Metabarcoding. *Mol. Biol. Evol.* 35, 1553–1555. doi: 10.1093/molbev/msy074

Shokralla, S., Gibson, J., King, I., Baird, D., Janzen, D., Hallwachs, W., et al. (2016). Environmental DNA barcode sequence capture: targeted, PCR-free sequence capture for biodiversity analysis from bulk environmental samples. *bioRxiv* [Preprint]. doi: 10.1101/087437

Shokralla, S., Hellberg, R. S., Handy, S. M., King, I., and Hajibabaei, M. (2015a). A DNA mini-barcoding system for authentication of processed fish products. *Sci. Rep.* 5:15894. doi: 10.1038/srep15894

Shokralla, S., Porter, T. M., Gibson, J. F., Dobosz, R., Janzen, D. H., Hallwachs, W., et al. (2015b). Massively parallel multiplex DNA sequencing for specimen identification using an Illumina MiSeq platform. *Sci. Rep.* 5:9687. doi: 10.1038/srep09687

Shokralla, S., Singer, G., and Hajibabaei, M. (2010). Direct PCR amplification and sequencing of specimens' DNA from preservative ethanol. *Biotechniques* 48, 233–234. doi: 10.2144/000113362

Somervuo, P., Koskela, S., Pennanen, J., Henrik Nilsson, R., and Ovaskainen, O. (2016). Unbiased probabilistic taxonomic classification for DNA barcoding. *Bioinformatics* 32, 2920–2927. doi: 10.1093/bioinformatics/btw346

Somervuo, P., Yu, D. W., Xu, C. C. Y., Ji, Y., Hultman, J., Wirta, H., et al. (2017). Quantifying uncertainty of taxonomic placement in DNA barcoding and metabarcoding. *Methods Ecol. Evol.* 8, 398–407. doi: 10.1111/2041-210X.12721

Stackebrandt, E., and Goebel, B. M. (1994). Taxonomic note: a place for DNA-DNA Reassociation and 16S rRNA sequence analysis in the present species definition in bacteriology. *Int. J. Syst. Evol. Microbiol.* 44, 846–849. doi: 10.1099/00207713-44-4-846

Staley, J. T., and Konopka, A. (1985). Measurement of in situ activities of nonphotosynthetic microorganisms in aquatic and terrestrial habitats. *Annu. Rev. Microbiol.* 39, 321–346.

Taberlet, P., Coissac, E., Hajibabaei, M., and Rieseberg, L. H. (2012a). Environmental DNA. *Mol. Ecol.* 21, 1789–1793.

Taberlet, P., Coissac, E., Pompanon, F., Brochmann, C., and Willerslev, E. (2012b). Towards next-generation biodiversity assessment using DNA metabarcoding. *Mol. Ecol.* 21, 2045–2050.

Taberlet, P., Prud'Homme, S. M., Campione, E., Roy, J., Miquel, C., Shehzad, W., et al. (2012c). Soil sampling and isolation of extracellular DNA from large amount of starting material suitable for metabarcoding studies: EXTRACTION OF EXTRACELLULAR DNA FROM SOIL. *Mol. Ecol.* 21, 1816–1820. doi: 10.1111/j.1365-294X.2011.05317.x

Tapolczai, K., Vasselon, V., Bouchez, A., Stenger-Kovács, C., Padisák, J., and Rimet, F. (2019). The impact of OTU sequence similarity threshold on diatom-based bioassessment: a case study of the rivers of Mayotte (France, Indian Ocean). *Ecol. Evol.* 9, 166–179. doi: 10.1002/ece3.4701

Tedersoo, L., Bahram, M., Polme, S., Koljalg, U., Yorou, N. S., Wijesundera, R., et al. (2014). Global diversity and geography of soil fungi. *Science* 346:1256688. doi: 10.1126/science.1256688

Thompson, L. R., Sanders, J. G., McDonald, D., Amir, A., Ladau, J., Locey, K. J., et al. (2017). A communal catalogue reveals Earth's multiscale microbial diversity. *Nature* 551, 457–463. doi: 10.1038/nature24621

Vamos, E., Elbrecht, V., and Leese, F. (2017). Short COI markers for freshwater macroinvertebrate metabarcoding. *Metabarcoding Metagenomics* 1:e14625. doi: 10.3897/mbmg.1.14625

Wang, Q., Garrity, G. M., Tiedje, J. M., and Cole, J. R. (2007). Naive bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl. Environ. Microbiol.* 73, 5261–5267. doi: 10.1128/AEM.00062-07

Weigand, H., Beermann, A. J., Čiampor, F., Costa, F. O., Csabai, Z., Duarte, S., et al. (2019). DNA barcode reference libraries for the monitoring of aquatic biota in Europe: Gap-analysis and recommendations for future work. *Sci. Total Environ.* 678, 499–524. doi: 10.1101/576553

Yilmaz, P., Parfrey, L. W., Yarza, P., Gerken, J., Pruesse, E., Quast, C., et al. (2014). The SILVA and "All-species Living Tree Project (LTP)" taxonomic frameworks. *Nucleic Acids Res.* 42, D643–D648. doi: 10.1093/nar/gkt1209

Yu, D. W., Ji, Y., Emerson, B. C., Wang, X., Ye, C., Yang, C., et al. (2012). Biodiversity soup: metabarcoding of arthropods for rapid biodiversity assessment and biomonitoring: biodiversity soup. *Methods Ecol. Evol.* 3, 613–623. doi: 10.1111/j.2041-210X.2012.00198.x

Zenker, M. M., Specht, A., and Fonseca, V. G. (2020). Assessing insect biodiversity with automatic light traps in Brazil: pearls and pitfalls of metabarcoding samples in preservative ethanol. *Ecol. Evol.* 10, 2352–2366. doi: 10.1002/ece3.6042

Zhou, X., Li, Y., Liu, S., Yang, Q., Su, X., Zhou, L., et al. (2013). Ultra-deep sequencing enables high-fidelity recovery of biodiversity for bulk arthropod samples without PCR amplification. *Gigascience* 2:4. doi: 10.1186/2047-217X-2-4

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.

frontiers in Ecology and Evolution

PERSPECTIVE
published: 19 April 2021
doi: 10.3389/fevo.2021.565649

Is Global Microbial Biodiversity Increasing, Decreasing, or Staying the Same?

David S. Thaler[1,2]*

[1] Department Biozentrum, University of Basel, Basel, Switzerland, [2] Program for the Human Environment, Rockefeller University, New York, NY, United States

Animal and plant biodiversity is decreasing. In contrast, the global direction and the pace of change in microbial, including viral, biodiversity is unknown. Important niches for microbial diversity occur in highly specific associations with plants and animals, and these niches are lost as hosts become extinct. The taxonomic diversity of human gut bacteria is reported to be decreasing. On the other hand, SARS-CoV-2 variation is increasing. Where microbes are concerned, Darwin's "tangled bank" of interdependent organisms may be composed mostly of other microbes. There is the likelihood that as some classes of microbes become extinct, others evolve and diversify. A better handle on all processes that affect microbial biodiversity and their net balance is needed. Lack of insight into the dynamics of evolution of microbial biodiversity is arguably the single most profound and consequential unknown with regard to human knowledge of the biosphere. If some or all parts of microbial diversity are relentlessly increasing, then survey approaches may be too slow to ever catch up. New approaches, including single-molecule or single-cell sequencing in populations, as well as focused attention on modulators and vectors of vertical and horizontal evolution may offer more direct insights into some aspects of the pace of microbial evolution.

Keywords: biodiversity, microbial diversity, extinction rate, generation of diversity, speciation, bottleneck, DNA bar coding, 16S/18S ribosomal RNA gene analysis

OPEN ACCESS

Edited by:
Mariana Mateos,
Texas A&M University, United States

Reviewed by:
Frederick M. Cohan,
Wesleyan University, United States
Dmitry Yurievich Sherbakov,
Limnological Institute (RAS), Russia

***Correspondence:**
David S. Thaler
davidsthaler@gmail.com;
david.thaler@unibas.ch

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

Received: *25 May 2020*
Accepted: *16 March 2021*
Published: *19 April 2021*

Citation:
Thaler DS (2021) Is Global Microbial Biodiversity Increasing, Decreasing, or Staying the Same? Front. Ecol. Evol. 9:565649. doi: 10.3389/fevo.2021.565649

A PROFOUND IGNORANCE

Animal and plant biodiversity on earth is decreasing. Many important features of this decrease are unclear, including ways in which the pace, i.e., the rate of decrease, is comparable to great extinctions defined by paleontology and how the current decrease is distributed among different phylogenetic domains and ecosystems (Di Marco et al., 2019; Trisos et al., 2020). However, the overall trajectory is clearly downward (Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services, 2020). The decrease in global biodiversity of "macrobes," i.e., eukaryotic multicellular differentiated organisms (EMDOs), commonly known as plants and animals, is a key issue of the Anthropocene.

In contrast to what we know of the world of plants and animals, we have no idea whether global microbial diversity is increasing, decreasing, or staying the same. For the purpose of this discussion microbial biodiversity includes eubacteria, archaea, protists, single-celled fungi, and viruses of all forms, bacteriophages, archaeaphages, and viruses of eukaryotes, including viruses of animals and plants. It is a blind spot– almost a scandal– that the question of the global trends of microbial

biodiversity seems never to have been raised. We raise it here. Once the question is asked, are there direct ways to address it? The intent here is to review literature in search of additional ways to address pace and direction. If there are ways to short-circuit survey approaches, what might they be? Can the enormous amount of data on microbial genomes and metagenomics be examined with rate of change in mind? What new approaches might shed light on the question? The approaches proposed in this article are by no means final and are certainly not protocols to solve the problem. The purpose here is to frame the rate of change of microbial biodiversity as an interesting and important question on which consequential progress is possible.

Consider the graphs below in which the Y axis represents biological diversity and the X axis is time **Figure 1**). The trajectory of biodiversity at any point in time is the first derivative (defined in calculus as the tangent of the curve at that point) of total biodiversity versus time. Our goal is to directly measure the straight line that is tangent to the curve at the present time).

Preliminaries to Frame the Question

Three preliminary issues require consideration: (a) What is meant by microbial biodiversity? (b) By what metrics is biodiversity in microbial realms comparable to the biodiversity of EMDOs (i.e., animals and plants?) (c) What baseline knowledge of microbial biodiversity is necessary in order to analyze how that diversity changes over time? I propose the following point of view: (a) Microbial biodiversity is the distribution of individuals in sequence space. (b) Microbial distribution in sequence space is similar enough to EMDO distribution that meaningful comparisons are possible. (c) Global directions of microbial evolution need not depend on catalogs of species and phylogenies. In some cases, direct measurement of the derivative at a single point need not depend on knowledge of the shape of the curve or the equation for the entire line. These interrelated issues are expanded below. The answers proposed below are meant to initiate discussion of the problem, not to prescribe specific approaches.

What Is Diversity in Microbiological Realms and How Is It Related to Animal Biodiversity?

Carl Woese pioneered the use of small subunit (SSU) RNA as a tool for species identification and phylogenetic analysis across the entirety of the living world (with the notable exception of viruses) (Woese and Fox, 1977; Woese et al., 1990). Woese's insight of universal sequence with sufficient variation also applies to mitochondrial cytochrome oxidase I (mtCOI) DNA barcoding in the animal kingdom. The key recognized by Woese is the need for sequences similar enough across the groups of interest that they can be aligned and compared. Small subunit RNA works marvelously well to divide life into large groups. Subsequent work has moved beyond genes encoding SSU RNA, using other genes and gene families (Zhu et al., 2019; Williams et al., 2020). Whole genomes have also been used for parsing bacteria and are able to distinguish clusters of individuals (species or strains) separated by horizontal gene transfer (HGT) rather than only by point mutations in shared genes (Rodriguez et al., 2018; Murray et al., 2020). The relative roles of vertical and horizontal evolution are of great interest (Woese, 2004; Frazao et al., 2019).

DNA barcodes and their relationship to animal species invite comparison to measures of microbial biodiversity. DNA barcoding by mtCOI has been more extensively discussed elsewhere (Stoeckle and Thaler, 2018). The clustering pattern of macroscopic life was elegantly articulated by Dobzhansky in his foundational book *Genetics and the Origin of Species* (Dobzhansky, 1937), page 4:

> *If we assemble as many individuals living at a given time as we can, we notice that the observed variation does not form a single probability distribution or any other kind of continuous distribution. Instead, a multitude of separate, discrete, distributions are found. In other words, the living world is not a single array of individuals in which any two variants are connected by unbroken series of intergrades, but an array of more or less distinctly separate arrays, intermediates between which are absent or at least rare. Each array is a cluster of individuals, usually possessing some common characteristics and gravitating to a definite modal point in their variation. Therefore the biological classification is simultaneously a man-made system of pigeonholes devised for the pragmatic purpose of recording observations in a convenient manner and an acknowledgement of the fact of organic discontinuity.*

DNA barcodes constitute a single metric by which the "feeling that it must be right" can be given a single quantitative meaning across the entire animal kingdom. Important findings have emerged from analysis of several million COI DNA barcodes. In groups throughout the animal kingdom, DNA barcode clusters largely correspond to what experts in each group have determined to be species. The extent of variance within clusters is similar and small (0.0% to 0.5% with most ∼0.2%) as determined through average pairwise difference (APD) within species from widely different groups including birds, mammals, fish, and insects. In most cases the APD separating nearest neighbor clusters is 1% to 2%. A key controversy regarding sequence clustering is the extent to which variance within clusters is neutral. For mitochondrial DNA barcodes most variation changes synonymous codons. The conclusion is not certain but the preponderance of evidence is that synonymous codon variation in the mitochondrial genome is neutral (Stoeckle and Thaler, 2018). If so, then synonymous variation is a passive passenger and also an indicator of population processes such that the accumulated variation is a function of population size and time (Kimura, 1986).

One takeaway is that the clustering pattern seen in macroscopic life is in a general way also a property of microscopic forms of life (Tibayrenc et al., 2015). Mechanisms necessarily differ because in many microbes (and fewer macrobes) the reproductive life cycle is not coupled to genetic exchange.

One avenue for further work will be the critical comparison of variation within SSUs (genes encoding small subunit RNA, 16S in eubacteria and archaea; 18S from the cytoplasmic ribosomes of eukaryotes) with the variance of mitochondrial COI DNA barcodes. In this way we could learn whether the amount of variation accumulated in extant species of EMDOs corresponds to that found in groups of microbes. The same APD approach should be applicable to mtCOI DNA barcodes and SSU analysis.

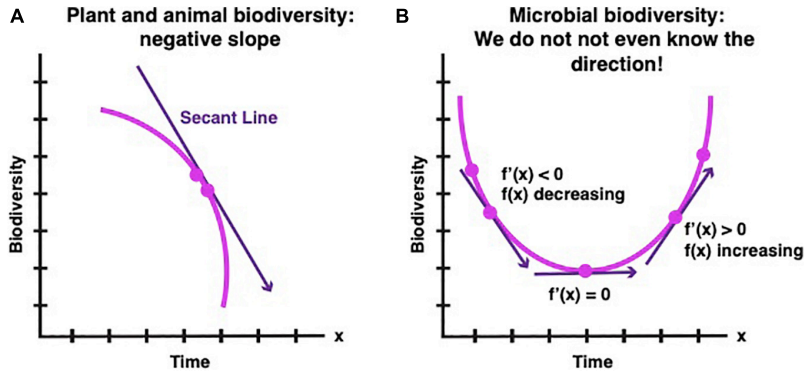


FIGURE 1 | (A) In pre-calculus, the secant is a straight line connecting two points on a curve. The present-day approach to measure changes in biodiversity is to count the number of species at different times and determine the slope of the secant. In the case of animal and plant biodiversity the slope of the secant is negative. **(B)** The breakthrough of calculus allows defining and finding the slope at a single point. A breakthrough analogous to calculus would be tremendously helpful to determining the trajectory of microbial biodiversity because the secant approach is in many cases difficult or seemingly impossible to implement.

A related area of important uncertainty in both cases (SSU RNA and mtCOI) is which variants are selectively neutral and which are subject to selection. Patterns of clustering of variation among individuals in extant populations of microbes may prove meaningfully comparable with analogous measures in plants and animals.

THE DERIVATIVE OF HORIZONTAL GENE TRANSFER

Microbial diversity is generated through both vertical and horizontal mechanisms. Evolution involves both the generation of diversity and selection among variants. However, these two processes are not always neatly separated (Thaler, 1994). HGT and mutation in vertical lineages are each generated through the action of enzymes encoded by genes that are themselves subject to evolution. The presence of genes involved in HGT, their activity and allelic state are indicative of the rates by which combinatorial microbial variants are generated.

There are several types of horizontal gene transfer mediated by bacteriophage through either specialized (Morse et al., 1956) or generalized (Zinder and Lederberg, 1952) transduction, conjugation (Lederberg and Tatum, 1946) (Lederberg et al., 1952; Cavalli et al., 1953), DNA-mediated transformation (Avery et al., 1944), or cell fusion (Gratia and Thiry, 2003; Gratia, 2005). New DNA entering a cell may either replicate independently or integrate into the host chromosome, or both. Integration into the host chromosome may involve illegitimate (Scwacha and Kleckner, 1994), site specific (Campbell, 1965) or homologous (Clark and Margulies, 1965) recombination. The degree of sequence similarity required to support homologous recombination is modulated by the mismatch repair and SOS systems which themselves are composed of genes and genetic networks subject to mutation and other genetic and physiological changes (Rayssiguier et al., 1989; Thaler, 1994; Magnasco and Thaler, 1996; Moxon and Thaler, 1997; Field et al., 1999). Sequence studies have inferred many of these processes independently (Cohan, 2017, 2019). Thus we know from experimental work that horizontal gene transfer can happen, with some detail as to molecular mechanisms, and from sequencing studies that these processes occur frequently in nature. Approaches need to be developed that directly assay mechanisms that mediate horizontal gene transfer.

Important niches for microbial diversity occur in highly specific associations with EMDOs, such as the gut microbiota in animals and nitrogen-fixing nodules on the roots of legumes. These specialized microbial communities probably cease to exist along with the extinction of their associated animals and plants. On the other hand, much of microbial life is understood only in the context of other microbes. For microbes, Darwin's "tangled bank" of "elaborately constructed forms, so different from each other, and so dependent upon each other" may consist mostly of other microbes.

A "species" may be considered to be a cluster in nucleotide sequence space. The number of different animal and plant species, although unknown, is the subject of reasonable estimation (Mora et al., 2011). The number of 16S eubacterial and archaeal sequence clusters has also been estimated (Louca et al., 2019) albeit with the caveat that estimates based on 16S necessarily underestimate total sequence diversity (Shevchenko et al., 2019). The difficulty in estimating sequence diversity of viruses inclusive of bacteriophage and archaeaphage lies partly in the plausibility of strains appearing and disappearing more rapidly than a comprehensive survey at any one moment could catch (Hadfield et al., 2018). What seems most likely is that the biosphere's sequence diversity at present is probably dominated by microbes, including viruses. A possible qualification to the statement of microbial dominance of biospheric sequence space is if somatic immune-system diversity is counted (Lin et al., 2020; Roskin et al., 2020) and turns out to be sufficient to shift the accounting of total diversity in favor of EMDOs. A different type of possible exception concerns heritable biological variation that is not based in polynucleotide sequence. DNA sequence could, in principle, account for all hereditary information. However, the fact that extant cells come only from other cells leaves open the

possibility that DNA sequence alone is not sufficient for biological continuity (Thaler, 2009).

ESTABLISHING A BASELINE OF MICROBIAL BIODIVERSITY

The classical way to measure the rate of change in biodiversity is to measure diversity at different times and divide the difference by the amount of time that has passed. For example, if the number of species decreases by half over the course of a year, the rate of loss is 0.5/yr. This classical approach of counting species and seeing how the counts change with time can be applied in specialized contexts, such as monitoring changes in microbial diversity within the intestinal microbiome (Magro et al., 2019). But there are two related problems making this approach difficult and perhaps impossible to generalize.

The first problem is that the extent of current microbial biodiversity is unknown (Dance, 2020), possibly with a large fraction in hard-to-access, rare, or extreme environments (Sogin et al., 2006; Amaral-Zettler et al., 2010). The deep hot biosphere may contain the majority of our planet's microbial biodiversity (Magnabosco et al., 2019). The second problem is a possible "chicken and egg" paradox that would preclude direct measurement of the rate of microbial biodiversification through a series of independent timepoints: Suppose that the rate of generation of new microbial diversity is very fast. If so, a baseline sequence library may never be finished because new diversity is generated more rapidly than it can be measured. It might require 20 years before there is sufficient understanding of the deep biosphere and other hard-to-access environments to have an adequate baseline library. However, if microbial–including viral–evolution is as rapid as it might be, establishing a complete baseline inventory may prove impossible. These two problems motivate the pursuit of approaches with potential for directly measuring the first derivative of microbial biodiversity, in order to gauge instantaneous directions and rates of change.

MICROBIAL DISTRIBUTION IN SEQUENCE SPACE

Species, in common practice of mitochondrial barcode sequence analysis, are usually assigned on the basis of their closeness to a consensus sequence. Counts taken at successive times determine if consensus sequences are missing and if new ones have appeared. Census counts of a species are enumerated as the number of individuals that are "close enough" to the consensus sequence. Microbiome analysis often enumerates not "species" but the diversity of "operational taxonomic units" (OTUs) in the context of SSU-encoding genes (usually 16S). OTUs are clusters of sequences separated by more than 2% (for fungi) or more than 3% (for eubacteria and archaea) from the nearest cluster.

"Quasispecies" are clouds in a sequence space formed by populations possessing mutable genomes (Eigen, 1993; Bull et al., 2005; Domingo and Perales, 2019; Lu et al., 2020). Modeling suggests that favored quasispecies allow many neutral mutations. Selectively favored mutations migrate the quasispecies cloud or establish a new one. Consider the case when a founder clone seeds a new quasispecies. The distribution of the quasispecies in sequence space is a function of the number of generations elapsed since clonal founding as well as the mutation rate and spectrum. The greater the number of generations and the higher the mutation rate, the larger the quasispecies sequence cloud will be. The size of a quasispecies cloud can be accessed through the positively correlated statistic APD among individuals within the cloud (Dridi et al., 2015; de Azevedo et al., 2017). Consider the case of a population that starts from a single sequence, a founder and its clonal descendants. A larger APD and cloud are functions of the number of generations and the mutation rate per generation. The quasispecies concept is related to the molecular clock, neutral evolution, and Luria-Delbrück mutation.

Zuckerkandl and Pauling in 1965 hypothesized a molecular clock based on the rate of amino acid substitutions in hemoglobin compared with fossil evidence giving an independent measure of time of divergence (Zuckerkandl and Pauling, 1965). Kimura in 1968 proposed that most sequence changes are selectively neutral or nearly neutral and the accumulation of variation in a population follows from the mutation rate, the number of generations, and the chance-driven loss or gain of variants (Kimura, 1968). The Luria-Delbrück interpretation of mutant clone sizes in a growing bacterial population implicitly includes the assumption that the accumulating variants are neutral during an exponential growth phase before selection (Luria and Delbrück, 1943; Stewart et al., 1990).

Phylogeography of mitochondrial genome coalescence, in combination with archeology and paleontology dates the origin of the modern human mitochondrial sequence in the range of 150,000 to 200,000 years ago (Mellars, 2006). Most mitochondrial sequence variation appears to be neutral (Richards et al., 2000; Tishkoff and Williams, 2002; Forster, 2004; Kivisild et al., 2006), but for a different view see Mishmar et al. (2003). The APD of mitochondrial sequences within animal species tends toward the same low value and most variation is probably neutral in divergent phyla such as insects, birds, mammals and, fish (Stoeckle and Thaler, 2018). Humans are an average animal species when analyzed in this way; the human APD is 0.1% and the majority of animal species are in the range of 0.1% to 0.2% (Thaler and Stoeckle, 2016). Outgrowth from a clonal sequence in a similar way over a similar time frame (for different species this varies from a range of ca. "last week" to half a million years ago, with a median at 100,000 to 200,000 years ago) is a plausible way to account for the similarity of mitochondrial sequence clustering within different species across the animal kingdom (Stoeckle and Thaler, 2018). The variance within clusters measured by APD is a prototype of an independent way to obtain the derivative of biodiversity–the pace of change in the sense of population replacement. Successive clonal replacement is very much the manner of evolution in the microbial realm (Tibayrenc et al., 2015; Baym et al., 2016). Baym and colleagues have produced a superb YouTube video demonstrating in an experimental setting the evolution of bacteria to survive in increasingly high concentrations of an antibiotic (Link in **Figure 2** caption). It

seems reasonable that a method akin to that used to estimate the age since expansion from a clonal sequence of mitochondria within an extant population, i.e., the variance of individuals around a consensus sequence as measured by APD, can also be used to estimate the time or the number of generations that have elapsed since a clonal origin of cellular microbial and viral populations.

In the eubacterial and archael realms of microbiome metagenomic sequence many analyses use OTUs rather than the word "species." OTUs are clusters separated by 2% or 3% in sequence space. The lack of further definitive knowledge follows in some cases because the organisms harboring them have never been cultured. Another part of the ignorance is simply that there are so many different sequence clusters. A related aspect is the semi-ignorance of sequence classification in the absence of experiment. A sequence may be somewhat similar to one in another organism that has been cultured, or to an enzyme for which a reaction has been characterized. There may be a temptation to overestimate the certainty when a new sequence is assigned to an old role. A less presumptuous way to state the level of certainty is that it is a generator of hypotheses whose degree of certainty and testability differs by specific case. OTUs are often based on SSUs (small subunit RNA, 16S in eubacteria and archaea) although with longer reads and

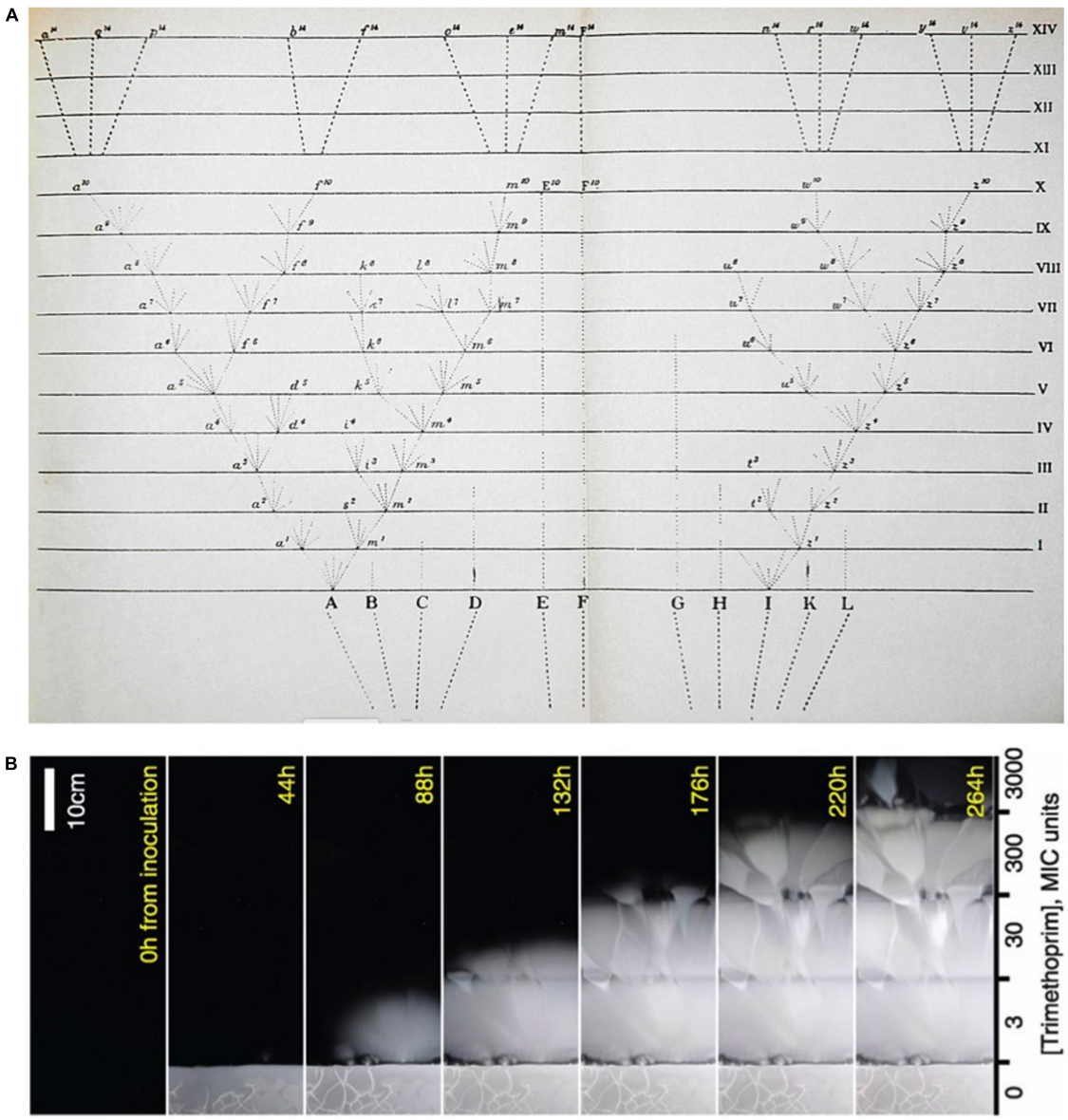


FIGURE 2 | (A) Darwin's "Tree of Life" diagram from his *Origin of Species* (Darwin, 1860) juxtaposed with **(B)** a time-lapse study of bacterial growth across a step-wise concentration gradient of antibiotic (Baym et al., 2016). A video of this marvellous experiment may be accessed here: https://www.youtube.com/watch?v=plVk4NVlUh8.

better total assembly, more extensive chromosomal contiguous segments are available and horizontal evolution can more easily be taken into account (Nguyen et al., 2016; Palmer et al., 2019). A saving grace of eubacterial, archaea, and eukaryotic cellular sequencing is that there is an overall tree of life on which to organize sequences (Woese et al., 1990; York, 2020). Such an overarching organizing principle is not available for viruses of either prokaryotes or eukaryotes.

The sequences of viruses of eukaryotes and prokaryotes (bacteriophage, archaeaphage) are in some ways easier and in others harder to characterize. Easier, because viral genomes are smaller and tend to be less potentially confusing than others, making it straightforward to assemble full genomes even from short read lengths (e.g., ∼200 bp from Illumina). There are differences in sample preparation to optimize for viruses, and they may have been missed in some studies. Phylogenetically, there are lineages and groups of related viruses, but the trail of relatedness "runs cold" far short of the all-encompassing tree of cellular life (Mavrich and Hatfull, 2017; Low et al., 2019). It seems likely that groups of viruses evolved independently, i.e., viruses of bats did not evolve from viruses of bacteria. Instead, it seems likely that viruses have branched off from cellular life many independent times and may continue to do so. There is recombination creating functional combinatorial diversity among viruses but most of this is between viruses that were related to begin with (Botstein, 1980; Brown et al., 2016).

Combining Quasispecies and Phylogeny for a "Stars and Galaxies" View of Life

Consider the distribution of life in sequence space as similar to that of stars in the universe. In this analogy, individuals are stars and species are galaxies. Galaxies correspond to quasispecies in the Eigen formulation. The sequence space outside quasispecies is unused and, at least in the neighborhood of quasispecies-galaxies, largely neutral with regard to selection. Higher taxa might approximate galaxy clusters or other cosmological structures in this planetarium-inspired representation. The enumeration of higher taxa in eukaryotes is proposed as a way to estimate the number of species and characterize their phylogenetic distribution (Mora et al., 2011). It will be of interest to learn if this approach can be usefully extended to microbial life, prokaryotic and eukaryotic.

It is instructive to compare the famous tree diagram from Darwin's *Origin of Species* (Darwin, 1860) with a more recent time series of bacterial evolution upward against a gradient of increasing concentrations of an antibiotic (Baym et al., 2016). Darwin's stick figure diagram of the origin of new species by descent is essentially unchanged in contemporary phylogenetic diagrams and their conceptualization (**Figure 2A**). When compared with a time-lapse visualization of bacteria evolving against an antibiotic gradient (**Figure 2B**), what we see are not stick figures but fans of selected clones followed by expansion, apparent neutrality being manifest in the symmetry of each fan's outgrowth. The accumulation of neutral diversity in the outgrowth is invisible to the eye, but certainty of its occurrence follows from the very property that Darwin precisely articulated in the final paragraph of his *Origin of Species* as inheritance with variability. Evidence of accumulated invisible diversity during clonal outgrowth is revealed when the expanding edge of growth reaches a higher concentration of antibiotic. At the point where a more resistant clone has occurred, it is selected at the border and the outgrowth scenario iterates.

NGS approaches to microbial metagenomics where each molecule is sequenced separately are amenable to direct measurement of variation in microbial and viral populations, thereby giving precise instantiation to Eigen's quasispecies. Most sequence variation is likely to be neutral, as Kimura first proposed. A series of sequences across a horizontal line in the bacterial experiment would presumably manifest as a cloud in the manner of Eigen's quasispecies. Virus outbreak sequences such as HIV and SARS-CoV-2, could also be interpreted as quasispecies as well as the more usual approach as phylogenetic trees, which are optimal for epidemiological tracing (Hadfield et al., 2018).

A gedanken experiment follows. Consider two sets of metagenomic microbiome sequences, i.e., every bit of DNA and RNA sequence that can be obtained from an environmental or medical sample. Map each metagenome in sequence space. They will form a universe of galaxies. Compare the two universes that came from applying similar sequencing and analysis methods to different samples. Suppose that the galaxies coincide. The maps are congruent with respect to the center of all galaxies. However, suppose in one case the galaxies are bigger and more diffuse. What would this mean? The more diffuse galaxies imply more neutral accumulation, which implies that the population has existed longer without going through conditions that enforce sequence uniformity. The compactness of galaxies would be a measure of how recently populations have been through conditions that enforce sequence uniformity. There will be further information if the variance within galaxies that are OTUs within a sample tend to be similar and whether they tend to be similar among different samples. This would be the microbial and viral analog of the analysis that led to the conclusion that the extant population of most animal species is within an order of magnitude similar to humans in terms of age or number of generations since their mitochondria passed though conditions that led to sequence uniformity.

Sequence uniformity in a population can come about *via* different mechanisms including: (a) Clonal bottlenecks (b) Selective Sweeps or (c) Sorting. These mechanisms sometimes overlap and they cannot be distinguished within the sequence-clustering context discussed here [A more extensive discussion can be found in the section "**Conditions that favor clonal uniformity are frequent in biology"** of Stoeckle and Thaler (2018)].

TWO IMPERFECT APPROACHES TO AN INTERESTING AND IMPORTANT QUESTION

Learning the direction and pace of microbial change in biodiversity is predicted to be a key parameter for better understanding all evolution and better thinking about human futures. Both of the approaches proposed here are indirect and

imperfect. However, they correlate with the first derivative of microbial biodiversity as it changes over time. The first approach follows from the reasoning of Kimura and Luria and Delbrück to be the integral of the mutation rate and the number of generations since a population originated from a single sequence. This is a measurement of vertical evolution. The second is to inventory elements of horizontal gene transfer, e.g., the origin of transfer for promiscuous conjugative plasmids and transducing bacteriophage that harbor eubacterial or archaeal sequences in their capsids.

A key shortcoming of the proposed approaches, because they try to extract all their information from only a single time point, is that they are insensitive to the loss or gain of an entire group. If an entire group has become extinct, then it is by definition impossible to assay the variation within it. Conversely, when a group is within the sample, it is impossible to say whether or not it was not in the sampled ecosystem previously, e.g., a year ago. This insensitivity to large-scale changes appears (at this moment) to be an insurmountable weakness of any method based on sampling only a single moment in time. In situations where it is possible to take multiple measurements at different times, it will be worthwhile to inquire if and under what specific circumstances measurements indicative of the rate of change within existing populations correlate with the loss or gain of entire groups. The hygiene hypothesis and its intellectual descendant, the disappearing microbiota hypothesis, assert that over years and generations the taxonomic diversity of human microbiomes has been on a downward trajectory that negatively impacts health (Finlay et al., 2021).

These questions, framed here in the context of microbial diversity, are familiar concepts and controversies regarding macroscopic plants, animals, and the relationships between population and evolutionary processes.

INCREASING MICROBIAL BIODIVERSITY AND DECREASING ANIMAL AND PLANT BIODIVERSITY?

Microbes are sometimes thought of as living fossils (Schopf et al., 2015). Some "big history" views imply that microbial evolution must be finished or insignificant because eukaryotes culminating in humans with culture and science now cut the edges of evolution (Chardin, 1959; Marchetti, 1983). Concepts of hierarchical evolution may be biologically inaccurate and lead to flawed interpretations and predictions.

A complication to global thinking is that the derivative of microbial biodiversity is not the same everywhere. On the contrary, trends in microbial biodiversity are situation and context specific. Human gut bacterial diversity is reported to be taxonomically decreasing (Finlay et al., 2021). On the other hand, the diversity of SARS-CoV-2 is increasing, although understanding the phylogeny and significance of variants is difficult (Morel et al., 2020; Mascola et al., 2021). Approaches for analyzing microbial biodiversity are context specific, and new efforts will be needed to synthesize them.

Consider a scenario in which global microbial biodiversity increases while plant and animal diversity decreases (a quite plausible condition for the world in which we live). Does it matter? Might there be consequences for possible human futures? Most variation in evolution is neutral and has no selective consequence. Nevertheless, variation is the raw material from which new evolutionary possibilities are made. Microbial novelty might alter macroscopic ecosystems. It is worth considering how this change might occur. New human pathogens might arise. Prediction is not perfect, but possibilities merit thought. The cyberneticist Ashby framed a context between competitors each of which is considered to be or to possess an array of variety (Ashby, 1952). Each element of variety in one competitor's array is countered by an element in the other competitor's array. The advantage in Ashby's game goes to the competitor with the most variety. As Ashby famously put it, "Only variety can destroy/absorb variety." Serious thought should be given to consideration of whether Ashby's metaphor aptly describes human interactions with the microbial world.

If Ashby's game theory is apt, then it would benefit our side of the competition, the EMDO team, to identify and enhance our own relevant variety. The metaphor encourages the proposal that human needs might be well taken care of while leaving much of our biosphere in a wild, and biologically more complex, state (Waggoner et al., 1996; Wilson, 2016).

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.

AUTHOR CONTRIBUTIONS

DT conceived and wrote the manuscript.

FUNDING

This work was supported by the Program for the Human Environment, Rockefeller University.

ACKNOWLEDGMENTS

Thanks to Jesse Ausubel, Fiona Doetsch, Wandrille Duchemin, Jackie Faherty, Geoffrey Fucile, Mark Stoeckle, and Iddo Wernick for discussion and encouragement. The submitted manuscript has been significantly improved through insightful anonymous as well as editorial review and through language polishing by Dale Langford. I believe Socrates described "profound ignorance" as a dangerous lack of awareness of what one does not know. This essay is dedicated to my three sisters: Susanna Drogsvold, Joan Dobbie, and Ellen Beeler. I am a guest associate editor for a special research topics section in Frontiers of Ecology and Evolution: DNA Barcodes: Controversies, Mechanisms and Future Applications.

REFERENCES

Amaral-Zettler, L., Artigas, L., Baross, J., Bharathi, P., Boetius, A., Chandramohan, D., et al. (2010). "A global census of marine microbes," in *Life in the World's Oceans: Diversity, Distribution, and Abundance*, ed. A. D. McIntyre (Hoboken, NJ: Wiley Blackwell Publishing Ltd), 221–245.

Ashby, W. R. (1952). *Design for a Brain; the Origin of Adaptive Behavior*. New York, NY: Wiley.

Avery, O., Macleod, C., and McCarty, M. (1944). Studies on the chemical nature of the substance inducing transformation of pneumococcal types. *J. Expt. Med.* 79, 137–158.

Baym, M., Lieberman, T. D., Kelsic, E. D., Chait, R., Gross, R., Yelin, I., et al. (2016). Spatiotemporal microbial evolution on antibiotic landscapes. *Science* 353, 1147–1151. doi: 10.1126/science.aag0822

Botstein, D. (1980). A theory of modular evolution for bacteriophages. *Ann. N. Y. Acad. Sci.* 354, 484–490. doi: 10.1111/j.1749-6632.1980.tb27987.x

Brown, P. A., Touzain, F., Briand, F. X., Gouilh, A. M., Courtillon, C., Allee, C., et al. (2016). First complete genome sequence of European turkey coronavirus suggests complex recombination history related with US turkey and guinea fowl coronaviruses. *J. Gen. Virol.* 97, 110–120. doi: 10.1099/jgv.0.000338

Bull, J. J., Meyers, L. A., and Lachmann, M. (2005). Quasispecies made simple. *PLoS Comput. Biol.* 1:e61. doi: 10.1371/journal.pcbi.0010061

Campbell, A. (1965). The steric effect in lysogenization by bacteriophage lambda. I. Lysogenization of a partially diploid strain of *Escherichia coli* K-12. *Virology* 27, 329–339. doi: 10.1016/0042-6822(65)90112-1

Cavalli, L. L., Lederberg, J., and Lederberg, E. M. (1953). An infective factor controlling sex compatibility in Bacterium coli. *J. Gen. Microbiol.* 8, 89–103. doi: 10.1099/00221287-8-1-89

Chardin, P. T. D. (1959). *The Phenomenon of Man*. New York, NY: Harper Perennial.

Clark, A. J., and Margulies, A. D. (1965). Isolation and characterization of recombination-deficient mutants of *Escherichia coli* K-12. *Proc. Natl. Acad. Sci. U.S.A.* 53, 451–459.

Cohan, F. M. (2017). Transmission in the origins of bacterial diversity, from ecotypes to phyla. *Microbiol. Spectr.* 5:MTB-0014-2016. doi: 10.1128/microbiolspec.MTBP-0014-2016

Cohan, F. M. (2019). Systematics: the cohesive nature of bacterial species taxa. *Curr. Biol.* 29, R169–R172. doi: 10.1016/j.cub.2019.01.033

Dance, A. (2020). The search for microbial dark matter. *Nature* 582, 301–303. doi: 10.1038/d41586-020-01684-z

Darwin, C. (1860). *On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life*. New York, NY: D. Appleton and Company.

de Azevedo, S. S. D., Caetano, D. G., Cortes, F. H., Teixeira, S. L. M., Dos Santos Silva, K., Hoagland, B., et al. (2017). Highly divergent patterns of genetic diversity and evolution in proviral quasispecies from HIV controllers. *Retrovirology* 14:29. doi: 10.1186/s12977-017-0354-5

Di Marco, M., Ferrier, S., Harwood, T. D., Hoskins, A. J., and Watson, J. E. M. (2019). Wilderness areas halve the extinction risk of terrestrial biodiversity. *Nature* 573, 582–585. doi: 10.1038/s41586-019-1567-7

Dobzhansky, T. (1937). *Genetics and the Origin of Species*, 3rd Edn. New York, NY: Columbia University Press.

Domingo, E., and Perales, C. (2019). Viral quasispecies. *PLoS Genet.* 15:e1008271. doi: 10.1371/journal.pgen.1008271

Dridi, M., Rosseel, T., Orton, R., Johnson, P., Lecollinet, S., Muylkens, B., et al. (2015). Next-generation sequencing shows West Nile virus quasispecies diversification after a single passage in a carrion crow (*Corvus corone*) in vivo infection model. *J. Gen. Virol.* 96, 2999–3009. doi: 10.1099/jgv.0.000231

Eigen, M. (1993). Viral quasispecies. *Sci. Am.* 269, 42–49.

Field, D., Magnasco, M. O., Moxon, E. R., Metzgar, D., Tanaka, M. M., Wills, C., et al. (1999). Contingency loci, mutator alleles, and their interactions. Synergistic strategies for microbial evolution and adaptation in pathogenesis. *Ann. N. Y. Acad. Sci.* 870, 378–382.

Finlay, B. B., Amato, K. R., Azad, M., Blaser, M. J., Bosch, T. C. G., Chu, H., et al. (2021). The hygiene hypothesis, the COVID pandemic, and consequences for the human microbiome. *Proc. Natl. Acad. Sci. U.S.A.* 118:e2010217118. doi: 10.1073/pnas.2010217118

Forster, P. (2004). Ice Ages and the mitochondrial DNA chronology of human dispersals: a review. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 359, 255–264. doi: 10.1098/rstb.2003.1394

Frazao, N., Sousa, A., Lassig, M., and Gordo, I. (2019). Horizontal gene transfer overrides mutation in *Escherichia coli* colonizing the mammalian gut. *Proc. Natl. Acad. Sci. U.S.A.* 116, 17906–17915. doi: 10.1073/pnas.1906958116

Gratia, J. P. (2005). Noncomplementing diploidy resulting from spontaneous zygogenesis in *Escherichia coli*. *Microbiology* 151(Pt 9), 2947–2959.

Gratia, J. P., and Thiry, M. (2003). Spontaneous zygogenesis in *Escherichia coli*, a form of true sexuality in prokaryotes. *Microbiology* 149(Pt 9), 2571–2584.

Hadfield, J., Megill, C., Bell, S. M., Huddleston, J., Potter, B., Callender, C., et al. (2018). Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics* 34, 4121–4123. doi: 10.1093/bioinformatics/bty407

Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services (2020). *Assessing the State of Knowledge on Biodiversity*. Available online at: https://ipbes.net/assessing-knowledge [Accessed April 25 2020].

Kimura, M. (1968). Evolutionary rate at the molecular level. *Nature* 217, 624–626.

Kimura, M. (1986). DNA and the neutral theory. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 312, 343–354.

Kivisild, T., Shen, P., Wall, D. P., Do, B., Sung, R., Davis, K., et al. (2006). The role of selection in the evolution of human mitochondrial genomes. *Genetics* 172, 373–387. doi: 10.1534/genetics.105.043901

Lederberg, J., Cavalli, L. L., and Lederberg, E. M. (1952). Sex compatibility in *Escherichia Coli*. *Genetics* 37, 720–730.

Lederberg, J., and Tatum, E. L. (1946). Gene recombination in *Escherichia coli*. *Nature* 158:558. doi: 10.1038/158558a0

Lin, H., Peng, Y., Chen, X., Liang, Y., Tian, G., and Yang, J. (2020). T cell receptor repertoire sequencing. *Methods Mol. Biol.* 2204, 3–12. doi: 10.1007/978-1-0716-0904-0_1

Louca, S., Mazel, F., Doebeli, M., and Parfrey, L. W. (2019). A census-based estimate of Earth's bacterial and archaeal diversity. *PLoS Biol.* 17:e3000106. doi: 10.1371/journal.pbio.3000106

Low, S. J., Dzunkova, M., Chaumeil, P. A., Parks, D. H., and Hugenholtz, P. (2019). Evaluation of a concatenated protein phylogeny for classification of tailed double-stranded DNA viruses belonging to the order *Caudovirales*. *Nat. Microbiol.* 4, 1306–1315. doi: 10.1038/s41564-019-0448-z

Lu, I. N., Muller, C. P., and He, F. Q. (2020). Applying next-generation sequencing to unravel the mutational landscape in viral quasispecies. *Virus Res.* 283:197963. doi: 10.1016/j.virusres.2020.197963

Luria, S. E., and Delbrück, M. (1943). Mutations of bacteria from virus sensitivity to virus resistance. *Genetics* 28, 490–510.

Magnabosco, C., Biddle, J., Cockell, C., Jungbluth, S., and Twing, K. (2019). "Biogeography, ecology, and evolution of deep life," in *Deep Carbon: Past to Present*, eds B. Orcutt, I. Daniel, and R. Dasgupta (Cambridge: Cambridge University Press), 524–555.

Magnasco, M., and Thaler, D. S. (1996). Changing the pace of evolution. *Phys. Lett. A* 221, 287–292.

Magro, D. O., Santos, A., Guadagnini, D., de Godoy, F. M., Silva, S. H. M., Lemos, W. J. F., et al. (2019). Remission in Crohn's disease is accompanied by alterations in the gut microbiota and mucins production. *Sci. Rep.* 9:13263. doi: 10.1038/s41598-019-49893-5

Marchetti, C. (1983). On the role of science in the postindustrial society "Logos" - the empire builder. *Technol. Forecast. Soc. Change* 24, 197–206.

Mascola, J. R., Graham, B. S., and Fauci, A. S. (2021). SARS-CoV-2 viral variants-tackling a moving target. *JAMA* 11:2776542. doi: 10.1001/jama.2021.2088

Mavrich, T. N., and Hatfull, G. F. (2017). Bacteriophage evolution differs by host, lifestyle and genome. *Nat. Microbiol.* 2:17112. doi: 10.1038/nmicrobiol.2017.112

Mellars, P. (2006). Why did modern human populations disperse from Africa ca. 60,000 years ago? A new model. *Proc. Natl. Acad. Sci. U.S.A.* 103, 9381–9386. doi: 10.1073/pnas.0510792103

Mishmar, D., Ruiz-Pesini, E., Golik, P., Macaulay, V., Clark, A. G., Hosseini, S., et al. (2003). Natural selection shaped regional mtDNA variation in humans. *Proc. Natl. Acad. Sci. U.S.A.* 100, 171–176. doi: 10.1073/pnas.0136972100

Mora, C., Tittensor, D. P., Adl, S., Simpson, A. G., and Worm, B. (2011). How many species are there on Earth and in the ocean? *PLoS Biol.* 9:e1001127. doi: 10.1371/journal.pbio.1001127

Morel, B., Barbera, P., Czech, L., Bettisworth, B., Hubner, L., Lutteropp, S., et al. (2020). Phylogenetic analysis of SARS-CoV-2 data is difficult. *Mol Biol Evol.* 15:msaa314. doi: 10.1093/molbev/msaa314

Morse, M. L., Lederberg, E. M., and Lederberg, J. (1956). Transduction in *Escherichia Coli* K-12. *Genetics* 41, 142–156.

Moxon, E. R., and Thaler, D. S. (1997). The tinkerer's evolving toolbox. *Nature* 387, 659–662.

Murray, A. E., Freudenstein, J., Gribaldo, S., Hatzenpichler, R., Hugenholtz, P., Kampfer, P., et al. (2020). Roadmap for naming uncultivated Archaea and Bacteria. *Nat. Microbiol.* 5, 987–994. doi: 10.1038/s41564-020-0733-x

Nguyen, N. P., Warnow, T., Pop, M., and White, B. (2016). A perspective on 16S rRNA operational taxonomic unit clustering using sequence similarity. *NPJ Biofilms Microbiomes* 2:16004. doi: 10.1038/npjbiofilms.2016.4

Palmer, M., Venter, S. N., Coetzee, M. P. A., and Steenkamp, E. T. (2019). Prokaryotic species are sui generis evolutionary units. *Syst. Appl. Microbiol.* 42, 145–158. doi: 10.1016/j.syapm.2018.10.002

Rayssiguier, C., Thaler, D. S., and Radman, M. (1989). The barrier to recombination between *Escherichia coli* and *Salmonella* typhimurium is disrupted in mismatch-repair mutants. *Nature* 342, 396–401. doi: 10.1038/342396a0

Richards, M., Macaulay, V., Hickey, E., Vega, E., Sykes, B., Guida, V., et al. (2000). Tracing European founder lineages in the Near Eastern mtDNA pool. *Am. J. Hum. Genet.* 67, 1251–1276.

Rodriguez, R. L., Gunturu, S., Harvey, W. T., Rossello-Mora, R., Tiedje, J. M., Cole, J. R., et al. (2018). The microbial genomes atlas (MiGA) webserver: taxonomic and gene diversity analysis of *Archaea* and *Bacteria* at the whole genome level. *Nucleic Acids Res.* 46, W282–W288. doi: 10.1093/nar/gky467

Roskin, K. M., Jackson, K. J. L., Lee, J. Y., Hoh, R. A., Joshi, S. A., Hwang, K. K., et al. (2020). Aberrant B cell repertoire selection associated with HIV neutralizing antibody breadth. *Nat. Immunol.* 21, 199–209. doi: 10.1038/s41590-019-0581-580

Schopf, J. W., Kudryavtsev, A. B., Walter, M. R., Van Kranendonk, M. J., Williford, K. H., Kozdon, R., et al. (2015). Sulfur-cycling fossil bacteria from the 1.8-Ga Duck Creek Formation provide promising evidence of evolution's null hypothesis. *Proc. Natl. Acad. Sci. U.S.A.* 112, 2087–2092. doi: 10.1073/pnas.1419241112

Scwacha, A., and Kleckner, N. (1994). Identification of joint molecules that form frequently between homologs but rarely between sister chromatids during yeast meiosis. *Cell* 76, 51–63.

Shevchenko, S. G., Radey, M., Tchesnokova, V., Kisiela, D., and Sokurenko, E. V. (2019). *Escherichia coli* clonobiome: assessing the strain diversity in feces and urine by deep amplicon sequencing. *Appl. Environ. Microbiol.* 85:e01866-19. doi: 10.1128/AEM.01866-19

Sogin, M. L., Morrison, H. G., Huber, J. A., Welch, D. M., Huse, S. M., Neal, P. R., et al. (2006). Microbial diversity in the deep sea and the underexplored "rare biosphere". *Proc. Natl. Acad. Sci. U.S.A.* 103, 12115–12120.

Stewart, F. M., Gordon, D. M., and Levin, B. R. (1990). Fluctuation analysis: the probability distribution of the number of mutants under different conditions. *Genetics* 124, 175–185.

Stoeckle, M., and Thaler, D. (2018). Why should mitochondria define species? *Hum. Evol.* 33, 1–30.

Thaler, D. S. (1994). The evolution of genetic intelligence. *Science* 264, 224–225.

Thaler, D. S. (2009). The cytoplasmic structure hypothesis for ribosome assembly, vertical inheritance, and phylogeny. *Bioessays* 31, 774–783. doi: 10.1002/bies.200800190

Thaler, D. S., and Stoeckle, M. Y. (2016). Bridging two scholarly islands enriches both: COI DNA barcodes for species identification versus human mitochondrial variation for the study of migrations and pathologies. *Ecol. Evol.* 6, 6824–6835. doi: 10.1002/ece3.2394

Tibayrenc, M., Avise, J. C., and Ayala, F. J. (2015). In the light of evolution IX: clonal reproduction: alternatives to sex. *Proc. Natl. Acad. Sci. U.S.A.* 112, 8824–8826. doi: 10.1073/pnas.1508087112

Tishkoff, S. A., and Williams, S. M. (2002). Genetic analysis of African populations: human evolution and complex disease. *Nat. Rev. Genet.* 3, 611–621. doi: 10.1038/nrg865

Trisos, C. H., Merow, C., and Pigot, A. L. (2020). The projected timing of abrupt ecological disruption from climate change. *Nature* 580, 496–501. doi: 10.1038/s41586-020-2189-9

Waggoner, P., Ausubel, J., and Wernick, I. (1996). Lightening the tread of population on the land: American examples. *Popul. Dev. Rev.* 22, 531–545.

Williams, T. A., Cox, C. J., Foster, P. G., Szollosi, G. J., and Embley, T. M. (2020). Phylogenomics provides robust support for a two-domains tree of life. *Nat. Ecol. Evol.* 4, 138–147. doi: 10.1038/s41559-019-1040-x

Wilson, E. O. (2016). *Half-Earth: Our Planet's Fight for Life*. New York, NY: Norton & Company.

Woese, C. R. (2004). A new biology for a new century. *Microbiol. Mol. Biol. Rev.* 68, 173–186.

Woese, C. R., and Fox, G. E. (1977). Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *Proc. Natl. Acad. Sci. U.S.A.* 74, 5088–5090.

Woese, C. R., Kandler, O., and Wheelis, M. L. (1990). Towards a natural system of organisms: proposal for the domains of archea, bacteria, and eukarya. *Proc. Natl. Acad. Sci. U.S.A.* 87, 4576–4579.

York, A. (2020). New data for the tree of life. *Nat. Rev. Microbiol.* 18:63. doi: 10.1038/s41579-019-0317-z

Zhu, Q., Mai, U., Pfeiffer, W., Janssen, S., Asnicar, F., Sanders, J. G., et al. (2019). Phylogenomics of 10,575 genomes reveals evolutionary proximity between domains Bacteria and Archaea. *Nat. Commun.* 10:5477. doi: 10.1038/s41467-019-13443-4

Zinder, N. D., and Lederberg, J. (1952). Genetic exchange in *Salmonella*. *J. Bacteriol.* 64, 679–699.

Zuckerkandl, E., and Pauling, L. (1965). Molecules as documents of evolutionary history. *J. Theor. Biol.* 8, 357–366.

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

Advantages of publishing in Frontiers

OPEN ACCESS
Articles are free to read for greatest visibility and readership

FAST PUBLICATION
Around 90 days from submission to decision

HIGH QUALITY PEER-REVIEW
Rigorous, collaborative, and constructive peer-review

TRANSPARENT PEER-REVIEW
Editors and reviewers acknowledged by name on published articles

Frontiers
Avenue du Tribunal-Fédéral 34
1005 Lausanne | Switzerland

Visit us: www.frontiersin.org
Contact us: frontiersin.org/about/contact

REPRODUCIBILITY OF RESEARCH
Support open data and methods to enhance research reproducibility

DIGITAL PUBLISHING
Articles designed for optimal readership across devices

FOLLOW US
@frontiersin

IMPACT METRICS
Advanced article metrics track visibility across digital media

EXTENSIVE PROMOTION
Marketing and promotion of impactful research

LOOP RESEARCH NETWORK
Our network increases your article's readership