ASSESSING BIODIVERSITY IN THE PHYLOGENOMIC ERA

EDITED BY: Michael G. Campana, Melissa T. R. Hawkins and Susana Caballero

PUBLISHED IN: Frontiers in Ecology and Evolution

frontiers Research Topics

frontiers

ISSN 1664-8714
ISBN 978-2-88974-124-3
DOI 10.3389/978-2-88974-124-3

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

ASSESSING BIODIVERSITY IN THE PHYLOGENOMIC ERA

Topic Editors:
Michael G. Campana, Smithsonian Conservation Biology Institute (SI), United States
Melissa T. R. Hawkins, Smithsonian Institution, United States
Susana Caballero, University of Los Andes, Colombia

Citation: Campana, M. G., Hawkins, M. T. R., Caballero, S., eds. (2022). Assessing Biodiversity in the Phylogenomic Era. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-88974-124-3

Table of Contents

frontiers in Ecology and Evolution

EDITORIAL
published: 03 December 2021
doi: 10.3389/fevo.2021.803188

Editorial: Assessing Biodiversity in the Phylogenomic Era

Michael G. Campana[1]*, *Melissa T. R. Hawkins*[2] *and Susana Caballero*[3]

[1] *Center for Conservation Genomics, Smithsonian National Zoological Park and Conservation Biology Institute, Smithsonian Institution, Washington, DC, United States,* [2] *Department of Vertebrate Zoology, Division of Mammals, National Museum of Natural History, Smithsonian Institution, Washington, DC, United States,* [3] *Laboratorio de Ecología Molecular de Vertebrados Acuáticos, Departamento de Ciencias Biológicas, Universidad de Los Andes, Bogotá, Colombia*

Keywords: biodiversity, phylogenomics, taxonomy, species delimitation, population structure, hybridization, conservation

Editorial on the Research Topic

Assessing Biodiversity in the Phylogenomic Era

Over the last 15 years, rapid advances in high-throughput sequencing technology and bioinformatics have permitted the generation of phylogenomic datasets across the tree-of-life. Ongoing phylogenomic analyses range in scope from in-depth analyses of single species in time and space to attempts to sequence the genomes of every species on earth (Earth BioGenome Project: Lewin et al., 2018). These projects are necessary and timely due to the ongoing mass extinction known as the "Anthropocene" (Ceballos et al., 2020). Assessments of extant and extinct biodiversity is required to understand the evolution of life on earth, guide environmental policies, and inform species conservation efforts.

In this Research Topic, we collect nine research articles using current phylogenomic techniques to (re)assess patterns of biodiversity across the tree-of-life. Articles range in content from new bioinformatic tools to combine disparate genomic datasets (Fountain et al.) to species delimitation (Leaché et al.), disentangling reticulate speciation (Grummer et al.; Nge et al.), population structure and demographic analyses (Timm et al.; Martínez-García et al.), and metagenomics and microbiomics (Kaufmann and Cassin-Sackett; Lozano Mojica and Caballero; O'Rourke et al.). The breath of subjects covered in this Research Topic illustrates the wide utility of phylogenomic methods for assessing biodiversity.

Due to the proliferation of phylogenomic techniques, one of the current challenges is the combination of datasets from disparate sequencing technologies ranging from traditional single gene Sanger sequencing to the multitude of different high-throughput approaches. Additionally, sequencing technologies are often developed for specific model organisms, necessitating their adaptation to non-model organisms. Fountain et al. illustrate that iScan microarrays can be adapted to non-model organisms for high-throughput single nucleotide polymorphism (SNP) discovery. They also develop a script to combine SNP datasets from multiple sequencing technologies to make the best use of novel and existing data. By using these methods, Fountain et al. are able to genotype great apes at a far higher resolution than previously possible.

The increased resolution of phylogenomic methods have revealed fine-scale population structure (e.g., Cortes-Rodriguez et al., 2019; Gallego-García et al., 2021) and possible cryptic species in previously designated species and populations (e.g., Jin et al., 2020). The taxonomic significance of this structure needs further evaluation to determine whether these populations are significantly reproductively isolated. Here, Leaché et al. use double digest Restriction-Associated sequencing (ddRADseq) and a reference-based taxonomy to delimit species in the Greater Short-horned Lizard species complex (*P. hernandesi*).

OPEN ACCESS

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

***Correspondence:**
Michael G. Campana
campanam@si.edu

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

Received: *27 October 2021*
Accepted: *31 October 2021*
Published: *03 December 2021*

Citation:
Campana MG, Hawkins MTR and Caballero S (2021) Editorial: Assessing Biodiversity in the Phylogenomic Era. Front. Ecol. Evol. 9:803188. doi: 10.3389/fevo.2021.803188

Additionally, phylogenomic analyses routinely detect population admixture and hybridization, which can both generate and reduce biodiversity (e.g., Lamichhaney et al., 2015; Grossen et al., 2016; Kearns et al., 2018; Lavretsky et al., 2019). Introgression can be a frequent source of cytonuclear discordance via plastid capture (e.g., Hawkins et al., 2016), necessitating the analysis of both nuclear and plastid sequences to reconstruct the species tree. Using hybridization enrichment, Nge et al. identified repeated introgression and chloroplast capture in the Australian endemic plant genus *Adenanthos*. Similarly, Grummer et al. identified repeated hybridization between four species of *Liolaemus* lizards in Argentina using ddRADseq and mitochondrial DNA. These hybridization events necessitate revisions to taxonomic and conservations units, especially in legal systems where species protection is predicated on taxonomic distinctiveness (e.g., Waples et al., 2018).

Beyond redefinition of taxonomic units, phylogenomic techniques have critical implications for practical population management. Lozano Mojica and Caballero analyze environmental DNA from Colombian water bodies to assess vertebrate species richness and revise species ranges in this critically understudied biodiversity hotspot. Both Timm et al. and Martínez-García et al. use phylogenomic methods to inform fisheries management for commercially important species that underwent recent severe population collapses. Timm et al. use ddRADseq to identify previously unknown population structure in the Gulf of Mexico pink shrimp (*Farfantepenaeus duorarum*) and largely confirm an existing model of larval transport (Criales et al., 2000). By comparing whole mitogenomes obtained from 48 archaeological cod (*Gadus morhua*) specimens to 496 recent samples, Martínez-García et al. show that cod mitogenomic diversity reflects past demographic history rather than recent and historical overfishing. Analysis of nuclear genomes and greater sample sizes may better resolve impacts of overexploitation by humans.

Moreover, phylogenomics has expanded beyond the host organisms—investigations now routinely include analyses of associated microbiomes, pathogens, and diets. Kaufmann and Cassin-Sackett investigate patterns of microbial succession in sedimentary DNA found in Black-tailed prairie dogs (*Cynomys ludovicianus*) burrows. Black-tailed prairie dogs have undergone significant die-offs due to outbreaks of sylvatic plague (*Yersinia pestis*). Microbial communities in the burrows reflect usage patterns of the prairie dogs and the deposition of corpses due to plague infection. Using fecal DNA, O'Rourke et al. analyze the diet and foraging ecology of the endangered Indiana bat (*Myotis sodalis*). Their analyses revealed that Indiana bats were generalist consumers that most frequently foraged within riparian habitats. Their results indicate that conservation of the Cypress Creek National Wildlife Refuge's riparian habitat is critical to the Indiana bat's conservation.

Phylogenomic techniques are rapidly displacing earlier single genetic marker analyses. As sequencing throughputs continue to increase and costs continue to drop, we anticipate that these methods will only become more important. The articles in this Research Topic provide a snapshot of the current state-of-the-art.

AUTHOR CONTRIBUTIONS

MC drafted the manuscript. All authors revised and approved the final manuscript.

FUNDING

The Smithsonian Institution supported MC and MH.

REFERENCES

Ceballos, G., Ehrlich, P. R., and Raven, P. H. (2020). Vertebrates on the brink as indicators of biological annihilation and the sixth mass extinction. *Proc. Natl. Acad. Sci. U.S.A.* 117, 13596–13602. doi: 10.1073/pnas.1922686117

Cortes-Rodriguez, N., Campana, M. G., Berry, L., Faegre, S., Derrickson, S. R., Ha, R. R., et al. (2019). Population genomics and structure of the critically endangered Mariana Crow (*Corvus kubaryi*). *Genes* 10:187. doi: 10.3390/genes10030187

Criales, M. M., Bello, M. J., and Yeung, C. (2000). Diversity and recruitment of penaeoid shrimps (Crustacea: Decapoda) at Bear Cut, Biscayne Bay, Florida, USA. *Bull. Mar. Sci.* 67, 773–788. Available online at: https://www.ingentaconnect.com/content/umrsmas/bullmar/2000/00000067/00000002/art00007

Gallego-García, N., Caballero, S., and Shaffer, H. B. (2021). Are genomic updates of well-studied species worth the investment for conservation? a case study of the Critically Endangered Magdalena river turtle. *J. Hered.* esab063. doi: 10.1093/jhered/esab063

Grossen, C., Seneviratne, S. S., Croll, D., and Irwin, D. E. (2016). Strong reproductive isolation and narrow genomic tracts of differentiation among three woodpecker species in secondary contact. *Mol. Ecol.* 25, 4247–4266. doi: 10.1111/mec.13751

Hawkins, M. T. R., Leonard, J. A., Helgen, K. M., McDonough, M. M., Rockwood, L. L., and Maldonado, J. E. (2016). Evolutionary history of endemic Sulawesi squirrels constructed from UCEs and mitogenomes sequenced from museum specimens. *BMC Evol. Biol.* 16:80. doi: 10.1186/s12862-016-0650-z

Jin, M., Zwick, A., Slipiński, A., and de Keyzer, R., Pang, H. (2020). Museomics reveals extensive cryptic diversity of Australian prionine longhorn beetles with implications for their classification and conservation. *Syst. Entomol.* 45, 745–770. doi: 10.1111/syen.12424

Kearns, A. M., Restani, M., Szabo, I., Schrøder-Nielsen, A., Kim, J. A., Richardson, H. M., et al. (2018). Genomic evidence of speciation reversal in ravens. *Nat. Commun.* 9:906. doi: 10.1038/s41467-018-03294-w

Lamichhaney, S., Berglund, J., Almén, M. S., Maqbool, K., Grabherr, M., Martinez-Barrio, A., et al. (2015). Evolution of Darwin's finches and their beaks revealed by genome sequencing. *Nature* 518, 371–375. doi: 10.1038/nature14181

Lavretsky, P., McInerney, N. R., Mohl, J. E., Brown, J. I., James, H. F., McCracken, K. G., et al. (2019). Assessing changes in genomic divergence following a century of human-mediated secondary contact among wild and captive-bred ducks. *Mol. Ecol.* 29, 578–595. doi: 10.1111/mec.15343

Lewin, H. A., Robinson, G. E., Kress, W. J., Baker, W. J., Coddington, J., Crandall, K. A., et al. (2018). Earth BioGenome Project: sequencing life for the future of life. *Proc. Natl. Acad. Sci. U.S.A.* 115, 4325–4333. doi: 10.1073/pnas.1720115115

Waples, R. S., Kays, R., Fredrickson, R. J., Pacifici, K., and Mills, L. S. (2018). Is the red wolf a listable unit under the US Endangered species act? *J. Hered.* 109, 585–597. doi: 10.1093/jhered/esy020

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

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

frontiers in Ecology and Evolution

ORIGINAL RESEARCH
published: 13 January 2021
doi: 10.3389/fevo.2020.616741

Reticulate Evolution, Ancient Chloroplast Haplotypes, and Rapid Radiation of the Australian Plant Genus *Adenanthos* (Proteaceae)

Francis J. Nge[1,2], Ed Biffin[1,2], Kevin R. Thiele[3,4] and Michelle Waycott[1,2]*

[1] School of Biological Sciences, Faculty of Science, The University of Adelaide, Adelaide, SA, Australia, [2] State Herbarium of South Australia, Adelaide, SA, Australia, [3] School of Biological Sciences, The University of Western Australia, Crawley, WA, Australia, [4] Western Australian Herbarium, Biodiversity and Conservation Science, Department of Biodiversity, Conservation and Attractions, Bentley Delivery Centre, Bentley, WA, Australia

Cytonuclear discordance, commonly detected in phylogenetic studies, is often attributed to hybridization and/or incomplete lineage sorting (ILS). New sequencing technologies and analytical approaches can provide new insights into the relative importance of these processes. Hybridization has previously been reported in the Australian endemic plant genus *Adenanthos* (Proteaceae). Like many Australian genera, *Adenanthos* is of relatively ancient origin, and provides an opportunity to examine long-term evolutionary consequences of gene flow between lineages. Using a hybrid capture approach, we assembled densely sampled low-copy nuclear and plastid DNA sequences for *Adenanthos*, inferred its evolutionary history, and used a Bayesian posterior predictive approach and coalescent simulations to assess relative contributions of hybridization and ILS to cytonuclear discordance. Our analyses indicate that strong incongruence detected between our plastid and nuclear phylogenies is not only the result of ILS, but also results from extensive ancient introgression as well as recent chloroplast capture and introgression between extant *Adenanthos* species. The deep reticulation was also detected from long-persisting chloroplast haplotypes shared between evolutionarily distant species. These haplotypes may have persisted for over 12 Ma in localized populations across southwest Western Australia, indicating that the region is not only an important area for old endemic lineages and accumulation of species, but is also characterized by persistence of high genetic diversity. Deep introgression in *Adenanthos* coincided with the rapid radiation of the genus during the Miocene, a time when many Australian temperate plant groups radiated in response to large-scale climatic change. This study suggests that ancient introgression may play an important role in the evolution of the Australian flora more broadly.

Keywords: *Adenanthos*, ancient hybridization, chloroplast capture, incongruence, introgression, Proteaceae, radiation, reticulate evolution

OPEN ACCESS

Edited by:
Melissa T. R. Hawkins, Smithsonian Institution, United States

Reviewed by:
Filipe Sousa, University of Algarve, Portugal
Leo Joseph, CSIRO Ecosystem Sciences, Australia

****Correspondence:***
Francis J. Nge
francis.nge@adelaide.edu.au

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

Received: *13 October 2020*
Accepted: *09 December 2020*
Published: *13 January 2021*

Citation:
Nge FJ, Biffin E, Thiele KR and Waycott M (2021) Reticulate Evolution, Ancient Chloroplast Haplotypes, and Rapid Radiation of the Australian Plant Genus Adenanthos (Proteaceae). Front. Ecol. Evol. 8:616741. doi: 10.3389/fevo.2020.616741

INTRODUCTION

Hybridization is important in the evolution of many plant groups (Arnold, 1992; Soltis and Soltis, 2009; Givnish, 2010). Examples of gene flow between species are common in plants across many different evolutionary and phylogenetic scales, from deep reticulate introgression events (Folk et al., 2017; García et al., 2017) to the evolutionary process of speciation by hybridization (Mallet, 2007;

Soltis and Soltis, 2009). Reticulation can be indicated by discordance between organellar (plastid and mitochondrial) and nuclear molecular datasets, due to the different modes of inheritance and evolution between the two genomes (Birky, 1995; Soltis and Kuzoff, 1995; Small et al., 2004). However, cytonuclear discordance may also result from incomplete lineage sorting (ILS) or poor resolution among sampled loci (Willyard et al., 2009; Gurushidze et al., 2010). Addressing the causes of cytonuclear incongruence is increasingly realistic using next-generation sequencing (NGS) approaches including targeted hybrid capture (Lemmon et al., 2012; Lemmon and Lemmon, 2013; Weitemier et al., 2014). These methods, which can generate sequences from multiple nuclear and organellar loci, allow rigorous exploration of causes of cytonuclear incongruence, including hybridization, using robustly supported phylogenies (Howarth and Baum, 2005; Vargas et al., 2017).

Along with the developments in sequencing technology, there has been significant progress in analytical approaches to untangling the influence of hybridization and ILS on cytonuclear discordance. While studies applying these approaches cannot rule out the presence of ILS, they can confidently separate the signals of hybridization from ILS (e.g., Joly et al., 2009). Several recent studies have made inferences of deep reticulation from multiple introgression events throughout the evolutionary history of their study groups (Folk et al., 2017; García et al., 2017). Introgression can play an important role in plant evolution and has been linked to rapid radiations in some of these groups (Seehausen, 2004). Most studies to date have focused on Northern Hemisphere plants (Francisco-Ortega et al., 1996; Barrier et al., 1999; Stankowski and Streisfeld, 2015). Different evolutionary drivers may have been involved in the Southern Hemisphere due to the older age of its biota (Hopper, 2009 and references therein). Many prominent lineages in the Australian contemporary flora are thought to have originated in the Cretaceous (Crisp et al., 2011; Lamont and He, 2012; Crisp and Cook, 2013) and show a radiation pulse in the mid-Cenozoic (25–10 Ma) (Crisp et al., 2004) in response to increased seasonality initiated at the end of the Eocene (*c.* 33 Ma) and subsequent aridification after the mid-Miocene (*c.* 14 Ma) (Macphail, 2007). However, no studies to date have explored the link between large scale climatic change, radiation, and hybridization in the early evolution of Australian plants. Similarly, while adaptive introgression has been shown to have spurred radiations of many groups in other regions of the world (Barrier et al., 1999; Seehausen, 2004; Givnish, 2010), a conclusive link between the two has not yet been demonstrated in Australia.

Natural hybridization has been documented in a number of Australian plants (Ashton and Sandiford, 1988; Griffin et al., 1988; Sedgley et al., 1992; Holman and Playford, 2000; Walker et al., 2009). In a few cases it is extensive (Leach and Whiffin, 1978; Potts and Reid, 1985; McIntosh et al., 2014). Hybridization has been suspected in the endemic Australian plant genus *Adenanthos* Labill. (Proteaceae), based on morphology alone (Nelson, 1977), and has subsequently been confirmed using molecular data (Walker et al., 2018). *Adenanthos* comprises 31 extant species, the majority of which (29 of 31 species) occur in southwest Western Australia (SWA) (**Figure 1**). Two species are disjunct from the rest of the genus across the Nullarbor Plain and are restricted to the southern peninsulas of South Australia. The genus consists of perennial shrubs or in some cases trees and are thought to be bird pollinated (Keighery, 1982; Collins and Rebelo, 1987). High outcrossing rates associated with bird pollination and also general self-incompatibility found in most members of Proteaceae suggests that population dynamics of *Adenanthos* might be fundamentally different to the majority of plants that are insect pollinated (Keighery, 1982; Goldingay and Carthew, 1998). However, detailed population genetic studies on *Adenanthos* are currently lacking. All species of *Adenanthos* and its close relatives (e.g., *Isopogon* Knight, *Petrophile* R.Br. ex Knight, *Leucadendron* L.) that have chromosome counts are $n = 13$ (Ramsay, 1963; Stace et al., 1998); ploidy variation within genera and clades is relatively uncommon in Proteaceae.

Adenanthos diverged from its sister-group (Leucadendrinae P. H. Weston and N. P. Barker) at the Eocene–Oligocene boundary (stem age *c.* 33.9 Ma) (Sauquet et al., 2009), thus providing an excellent case study to investigate potential reticulate patterns of evolution across deep timescales in the context of the Australian flora. Here, we use a NGS hybrid capture approach to infer nuclear and chloroplast phylogenies of *Adenanthos* to: (1) reconstruct its evolutionary and biogeographic history, and (2) assess for signs of hybridization and deep reticulate evolution within the genus.

MATERIALS AND METHODS

Sampling

We included 44 samples (30 of the 31 recognized species and 2 putative hybrids) covering all infrageneric sections and subsections within *Adenanthos* according to the most recent taxonomic revision (Nelson, 1977). We included a natural hybrid between *A. cuneatus* and *A. sericeus* (*A.* × *cunninghamii*) in our study. Half of the samples were collected in the field with fresh leaf tissue dried in silica gel (**Supplementary Table S1A**). The remaining samples were sourced from recently collected herbarium specimens (after 1960) lodged in PERTH and AD (**Supplementary Table S1B**).

DNA Extraction, Library Preparation, and Sequencing

Approximately 20 mg of silica dried leaf material per sample was used for DNA extractions, performed by Intertek Group plc using magnetic bead-based chemistry. We used a set of 30–100 single-copy nuclear and 13 plastid loci developed as phylogenetic markers for angiosperms (Waycott et al., in preparation) using the MYBaits target enrichment system (MYcroarray, Ann Arbor, Michigan) for sequence capture of the selected loci. In brief, genomic DNA (normalized to 1 ng/uL) was sheared using a Diagenode Bioruptor Pico sonicator for seven cycles to fragment lengths of *c.* 400–600 bp. DNA libraries were constructed using a JetSeq Flex DNA Library preparation kit (Bioline). To enable bioinformatics processing following hybrid capture, two 8 bp synthetic barcodes were annealed at each end of the DNA fragments. During the hybrid capture protocol, for each 96-well plate, the first barcode is replaced every 48 samples (i.e., two barcodes, one for each half of the plate), while the second

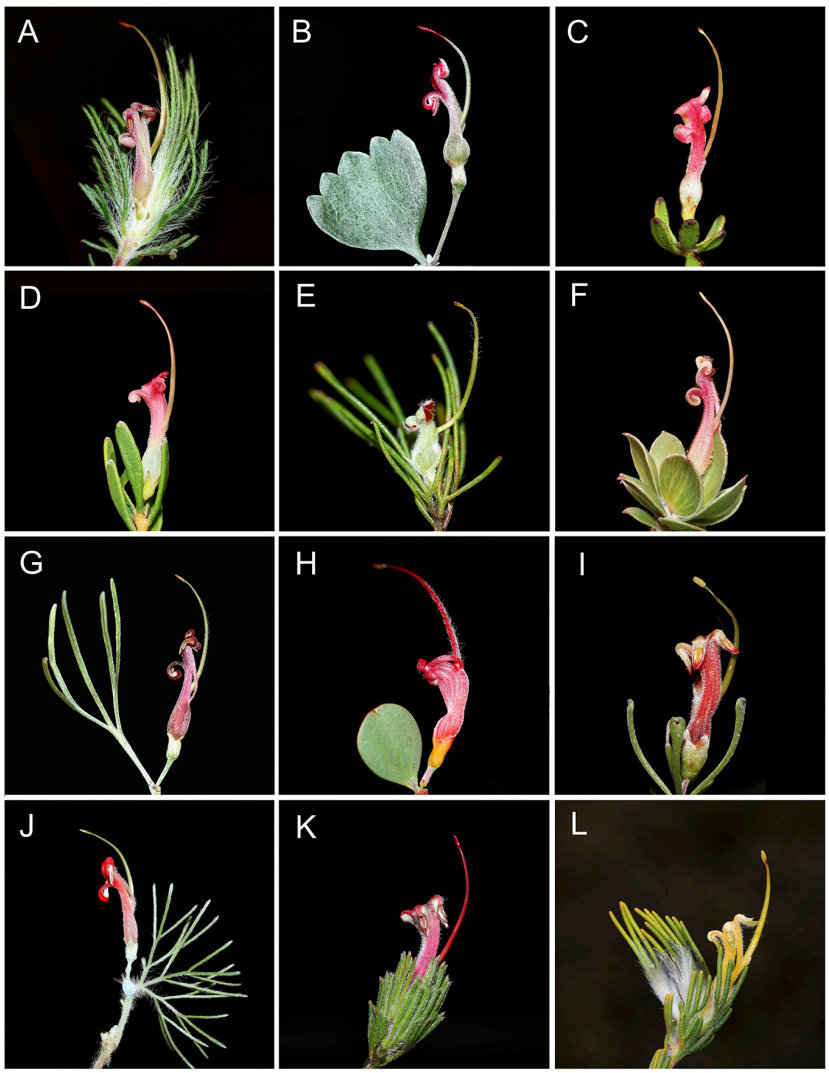

FIGURE 1 | Representative floral and leaf diversity of *Adenanthos*: **(A)** *Adenanthos cygnorum* subsp. *chaemaephyton* E.C.Nelson. **(B)** *A. stictus* A.S.George. **(C)** *A. glabrescens* subsp. *exasperatus* E.C.Nelson. **(D)** *A. glabrescens* E.C.Nelson subsp. *glabrescens*. **(E)** *A. linearis* Meisn. **(F)** *A. venosus* Meisn. **(G)** *A.* × *cunninghamii* Meisn. **(H)** *A. obovatus* Labill. **(I)** *A. forrestii* F. Muell. **(J)** *A. sericeus* [Labill. cultivated.] **(K)** *A. macropodianus* [E.C.Nelson.] **(L)** *A. terminalis* [R.Br.] Photos: F. J. Nge.

barcode is unique to each sample of each half-plate (i.e., 48 different barcodes). This ensured that each sample has a unique combination of the two barcodes for downstream identification. Libraries were pooled in equimolar concentrations and sent for Illumina paired-end sequencing (2 × 150) on a lane of a HiSeqX Ten at the Garvan Institute for Medical Research in Sydney.

Sequence Assembly

High-throughput 150 bp paired-end reads were processed using CLC Genomics Workbench v7.5.1[1]. Following demultiplexing and quality trimming (Phred-score threshold of 20), we used *de novo* assembly of pooled *Adenanthos* samples to generate a set of reference contigs for each sample. In order to recover the targeted nuclear loci, the *de novo* assembly was converted to a BLAST database and reference genomic sequences of *Aquilegia coerulea* (downloaded from Phytozome v 12[2]) used as query sequences using an *E*-value ≤ 1E-20. The *de novo* contigs matching the *Aquilegia* genes were used as a mapping reference for each individual to generate a per sample assembly at each locus. From these, we extracted the majority rule consensus sequence inserting "Ns" when coverage was lower than 5.

The resultant mapping files were exported in BAM format and allele phasing was performed using SAMTools Phase with default parameters applied (Li et al., 2009). SAMTools calls heterozygous SNPs (single nucleotide polymorphisms) at one site and segregates the reads (which contain one or the other heterozygous SNP) into two new "phased" BAM files. Reads lacking the given SNP site (but in part overlapping the segregated

[1]https://www.qiagenbioinformatics.com

[2]https://phytozome.jgi.doe.gov

reads) are segregated randomly to either BAM file. The phased BAM files were then imported into Geneious v.1.11.5 (Kearse et al., 2012), and a majority-rule consensus extracted using a 65% cut-off, then aligned using the MUSCLE (Edgar, 2004) plugin with default parameters.

In the majority of samples, we also recovered the 18S–26S nuclear ribosomal internal transcribed spacer (ITS) region. Although not specifically targeted, nuclear ribosomal DNA has a high copy number and can be recovered as by-catch. We used an ITS reference sequence (*Isopogon sphaerocephalus,* GenBank accession number AF508820.1) as a query sequence for BLAST and generated a per sample assembly as outlined above.

Plastid (chloroplast) targets were recovered using the chloroplast genome sequence of *Macadamia integrifolia* (GenBank reference number 34480) as a mapping reference. Reads from each sample were mapped to the reference using default parameters with a length fraction of 0.7 and a similarity fraction of 0.9. Consensus sequences were extracted as above. Consensus sequences for each individual and locus were imported into Geneious v1.11.5 (Kearse et al., 2012) and aligned using the MUSCLE (Edgar, 2004) plugin with default parameters, then manually checked and adjusted. Samples with more than 70% missing data in both nuclear and plastid alignments were excluded from our final dataset.

Phylogenetic Analyses and Divergence Time Estimation

Maximum Likelihood (ML) analyses were implemented in RAxML v.8.2.10 (Stamatakis, 2014) for two datasets: (1) 35 nuclear contigs phased and unphased (44 taxa, 25,646 bp), and (2) concatenated chloroplast sequences (43 taxa, 34,218 bp), using the GTR + I + G substitution model and bootstrap support obtained with 1,000 standard boostrap replicates. Single-gene trees were also estimated for a subset of our unphased nuclear dataset (19 nuclear contigs) that excluded potential paralogs, with 100 standard bootstrap replicates each. We used BLAST searches against *de novo* assemblies to screen for potential paralogs, assuming that divergent and overlapping contigs recovered for a single target gene represent paralogy. To assess for phylogenetic congruence and signal among loci, well-supported clades (>75% bootstrap) in each nuclear gene tree were compared with all other gene tree topologies manually. Bayesian analyses were conducted in BEAST v.2.4.7 (Bouckaert et al., 2014) for our concatenated datasets to obtain age estimates for *Adenanthos* for each dataset using a range of fossil calibration regimes (see **Supplementary Tables S2–S4** for details). Available nuclear (ITS) and plastid (*matK, rbcL*) sequences for outgroups were sourced from GenBank (**Supplementary Table S5**). For these three gene regions, we used the fossil calibrations applied in the Proteaceae family-wide study of Sauquet et al. (2009) to obtain divergence estimates for *Adenanthos.* One fossil calibration point (*Cranwellipollis palisadus;* for stem of *Franklandia*) was available within subfamily Proteoideae, which includes *Adenanthos.* We also included five additional calibration points in other subfamilies within Proteaceae to increase the accuracy of these estimates, applying uniform calibration priors following recommended practice (Sauquet et al., 2012). Because NGS sequences were not available for the outgroups, we applied similar calibration regimes for our full NGS datasets and compared the divergence age estimates with those obtained from secondary calibrations derived from these estimates. Secondary calibrations included (i) the stem age of *Isopogon*, the sister genus of *Adenanthos* and Leucadendrinae (set as log-normal distribution, offset = 41.5 Ma, *SD* = 0.23), and (ii) *Adenanthos* and Leucadendrinae crown (set as log-normal distribution, offset = 16 Ma, *SD* = 0.23) obtained from the plastid (*matK*) BEAST run.

BEAUti v2.4.7 was used to create input files for BEAST. We used a GTR + I + G substitution model and a relaxed lognormal clock model. Three parallel BEAST runs were performed for each analysis with the number of Markov Chain Monte Carlo (MCMC) generations and sampling frequency dependent on the size of the dataset (**Supplementary Table S6**). The first 20% of runs were discarded as burn-in. Tracer v1.6.0 (Rambaut et al., 2015) was used to assess convergence of the posterior, which was determined when effective sample size (ESS) reached ≥200. Tree output files were combined using LogCombiner v2.4.7, summarized in TreeAnnotator v2.4.7, and visualized using FigTree v1.4.3 (Rambaut, 2012). Lineage-through-time plots were constructed from pruned nuclear and chloroplast BEAST trees where each species was represented with only one terminal, using the 'phytools' package (Revell, 2012) in R (R Core Team, 2016).

All RAxML and BEAST analyses were run on the CIPRES Science Gateway portal (Miller et al., 2010). Conflicts between the nuclear and chloroplast ML topologies were visualized using the tanglegram tool in Dendroscope v. 3.5.10 (Scornavacca et al., 2011; Huson and Scornavacca, 2012).

Hybridization Assessment

In order to distinguish nuclear and chloroplast topological discordance as a result of either hybridization or ILS, we simulated plastid gene trees using scaled nuclear trees to obtain estimates of ILS for the plastid dataset, following the approach of García et al. (2017) and Folk et al. (2017). The simulated trees were then compared with the actual plastid topology and hybridization events inferred by areas that are incongruent. We utilized ASTRAL v. 5.6.3 (Zhang et al., 2018) to obtain a species tree with coalescent branch lengths from the individual nuclear gene trees obtained through RAxML. ASTRAL utilizes unrooted gene trees to generate phylogenetic quartets, which is relevant for our dataset as our individual gene trees did not contain outgroups. This is beneficial as random rooting can mimic the coalescent process (Rosenfeld et al., 2012; Tian and Kubatko, 2014). Support was assessed through the final normalized quartet scores of the overall species trees and local posterior probabilities of each branch terminal as a measure of gene tree conflict. Branch lengths of the species tree were scaled by a factor of four to account for organellar inheritance, as maternal inheritance of the plastid genome is typical for flowering plants (Mogensen, 1996). We simulated 1,000 gene trees from the scaled ASTRAL species tree by applying a coalescent model using a python-based script

with DendroPy (Mirarab et al., 2014). The simulated gene trees were visualized in DensiTree 2 (Bouckaert and Heled, 2014).

We used JML v.1.3.1 (Joly, 2012) following the approach of Joly et al. (2009) to assess the relative contributions of hybridization and ILS to discordance. This method uses the posterior distribution of species trees, population size, and branch lengths estimated in *BEAST (Bouckaert et al., 2014) to simulate sequence data under coalescent scenarios with no migration. To achieve this the minimum pairwise distance between sequences of two extant species from the simulated dataset were compared with empirical data. Hybridization or introgression can be inferred when observed pairwise distances from empirical data are significantly smaller than the simulated dataset derived from JML analyses, rejecting ILS as the only cause for topological conflict between the datasets (Joly et al., 2009; Joly, 2012). A coalescent tree was inferred in *BEAST using a combined nuclear and chloroplast dataset. Genetic distances were calculated using JML with only the chloroplast dataset, as Joly et al.'s (2009) approach assumes that the markers used to estimate the genetic distances are non-recombinant. JML analyses were conducted on the complete dataset, and also on individual clades which were shown to be incongruent between our nuclear and chloroplast topologies separately (**Supplementary Table S7**). We tested the performance of JML to assess deeper introgression events, evident by particular subclades showing incongruence among nuclear and chloroplast datasets. Analysis of subsets of the total data were conducted where only a fraction of the sampled taxa were included, including only one representative from each pair of conflicting clades across the two topologies. For each analysis, 1,000 simulations were computed for the chloroplast pairwise distance comparisons.

Haplotype and Splitstree Networks

As bifurcating trees may not accurately represent reticulate events among closely related taxa, we used network analyses to better represent relationships and assess for conflict between the nuclear and chloroplast datasets. Haplotype networks were constructed for the chloroplast dataset using TCS 1.13 (Clement et al., 2000) in PopART v.1.7 (Leigh and Bryant, 2015), classified into different subregions in SWA according to the Interim Biogeographic Regionalization for Australia (IBRA7) bioregional classification scheme[3]. Distance-based Neighbour-Nets were created in SplitsTree v.4.14.4 (Huson and Bryant, 2006) for (i) nuclear, (ii) plastid, and (iii) combined datasets using uncorrected *p*-distances.

RESULTS

Our plastid alignment through reference mapping and BLAST for downstream analyses contained 13 curated contigs that were 34,218 bp in length and included 43 taxa. Total curated nuclear alignments had a length of 25,646 bp comprising 35 independent loci after potential paralogs were excluded, covering 44 sampled taxa. Of the 44 samples, 39 were present in both nuclear and plastid datasets after removal of taxa with poor quality sequences or missing data.

Phylogenetic Relationships and Conflict in Nuclear and Plastid Data

Incongruence is significant between the plastid and nuclear ML topologies, with the two datasets recovering a different number of clades and statistically well-supported conflicting relationships across species and clades (**Figure 2**) (for further details see **Supplementary Results**).

Both the concatenated ML and coalescent ASTRAL analyses gave largely congruent results for the nuclear dataset (**Figures 2**, **3**). Several clades are resolved with high support from the ML topology (Clade C: bootstrap BS = 92; Clade D: BS = 82; Clade E: BS = 90; Clade F: BS = 74), however, the backbone of the tree was unresolved (**Figure 2**). Similarly, the backbone of the plastid topology was largely unresolved.

Divergence Age Estimates and Radiation of *Adenanthos*

Age estimates for *Adenanthos* obtained from our NGS nuclear trees were older than those from the plastid dataset, despite employing similar fossil calibration constraints (**Supplementary Table S8**). The divergence time estimates from the combined ITS, *matK,* and *rbcL*, as well as ITS-only topologies, are closer to those obtained from our NGS plastid topology (**Supplementary Table S8**). These differences in divergence time estimates were consistent across all fossil calibration schemes, including those obtained from secondary calibration of NGS data only, when outgroups were excluded due to missing data (**Supplementary Table S8**).We focus on the divergence age estimates of the NGS plastid and ITS combined topologies here, as the older age estimates from the nuclear NGS data likely reflect the lack of available NGS data for outgroup taxa used for the calibration regimes as well as missing data from our NGS nuclear dataset.

The stem age of *Adenanthos* was estimated at 36.1 Ma (95% CI: 15.3–33.2 Ma) for our NGS plastid topology and 38.4 Ma (95% CI: 15.6–36.6) for the ITS, *matK* and *rbcL* combined topology, employing the Proteaceae-wide calibration scheme (**Figure 4A** and **Supplementary Table S8**). The crown age was estimated at 24 Ma (95% CI: 15.3–33.2 Ma) for the plastid topology and 25.1 Ma (95% CI: 24.5–41.0 Ma) for the combined ITS and plastid topologies, respectively, in the late-Oligocene–Miocene (**Figure 4A** and **Supplementary Table S1**). Radiations of clades in the mid-Miocene (15–20 Ma) was consistent in both the NGS plastid and ITS-plastid combined topologies (**Figure 4A** and **Supplementary Figure S1**).

Interestingly, both our chloroplast and nuclear trees (ITS and NGS) showed that the southeastern *Adenanthos* clade is strongly nested within other SWA clades, even despite the strong topological conflicts between the two datasets. The divergence of the southeastern Australian clade from one of the SWA subclades was estimated at 16.5 Ma (95% CI: 9.5–24.8 Ma) based on the plastid NGS topology (**Figure 4A**). In contrast, the crown age of the southeastern species was inferred to be significantly older

[3]http://www.environment.gov.au/topics/land/national-reserve-system/science-maps-and-data/australias-bioregions-ibra

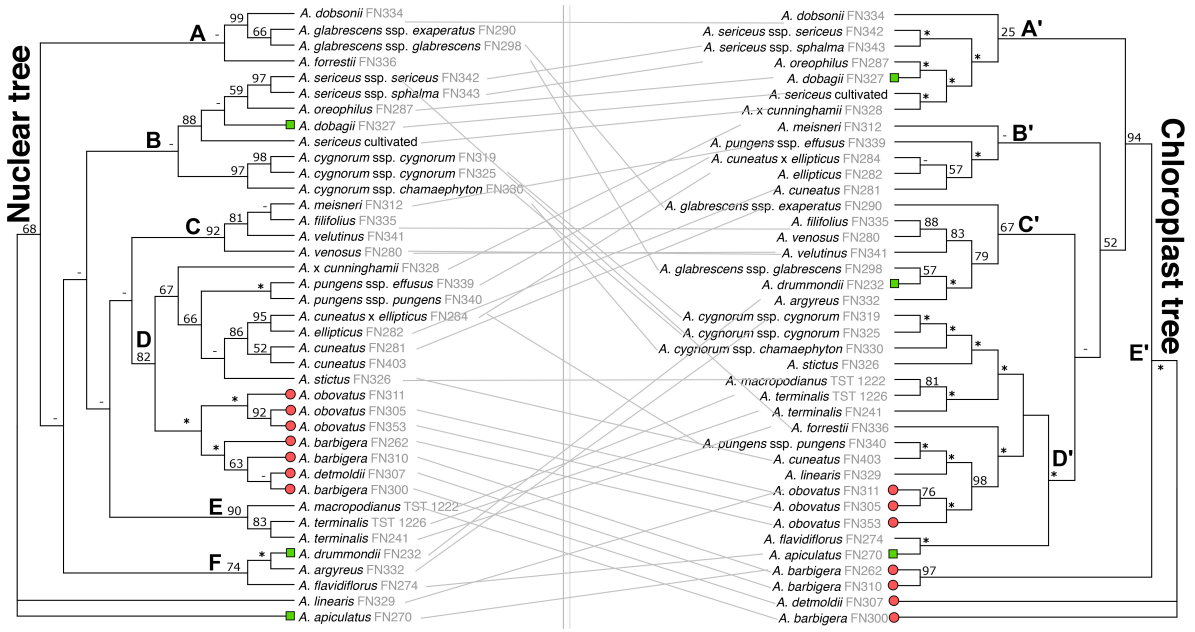

FIGURE 2 | Tanglegram comparing the phylogenetic placement of *Adenanthos* taxa connected by gray lines from maximum-likelihood concatenated RAxML nuclear (left) and plastid (right) trees. Bootstrap support (BS) is indicated above branches: very strongly supported clades (BS = 100%) are represented by "*", weakly supported branches with <50% BS are represented by "–". Letters **(A–F)** and **(A'–E')** refer to clades discussed in text. Terminal tips represent taxa from *A.* sect. *Eurylaema* (red circles), *A.* sect. *Adenanthos* subsect. *Anaclastos* (green squares), and *A.* subsect. *Adenanthos* (unlabeled).

based on the ITS topology, with divergence of *A. macropodianus* from *A. terminalis* estimated in the Miocene *c.* 8.4 Ma (95% CI: 2.8–14.7 Ma) compared with a Pleistocene divergence *c.* 1.3 Ma (95% CI: 0.29–2.9 Ma) inferred from the NGS plastid topology (**Figure 4A** and **Supplementary Figure S2**).

Hybridization Assessment

Reticulate evolution and introgression within *Adenanthos* were detected amongst our molecular datasets, indicated by the widespread discordance between the nuclear and plastid topologies. Evidence for hybridization was supported by both the gene tree simulations and JML approaches. The simulated plastid gene tree distribution derived from the nuclear ASTRAL species tree indicates that the discordance between the nuclear and plastid datasets is at least partly due to hybridization, as the simulated plastid topology did not match the actual plastid topology (**Figure 5**). The discordance between the simulated and actual plastid topology indicates that several of the observed plastid-nuclear discordances are almost never expected to occur under coalescence alone, indicating that they are unlikely to be caused by ILS alone.

Instances of ancient and putative recent hybridization were supported by our JML analyses (**Supplementary Table S9**). In the dataset with 25 taxa, only 9 out of 253 pairwise comparisons had non-significant values (p value > 0.1); that is, 96% (244/253) of the pairwise chloroplast distance comparisons are significantly smaller than expected in a scenario with only ILS (p value < 0.1) (**Supplementary Table S9A**). This indicates that the model cannot accurately predict the observed minimum distances and that a strict bifurcating species tree model is inadequate, due to the presence of hybridization. Subsequent analyses on the three subsets all detected signals of introgression between different clades within *Adenanthos* (**Supplementary Table S9B**).

Haplotype and Reticulate Networks

The majority of species have chloroplast haplotypes that were exclusive to each taxon, with unique haplotypes for sampled individuals within species (**Figure 6**). Only four instances where haplotypes were shared across multiple species were detected in our genus-wide haplotype network. Twenty haplotypes inferred by TCS were not present in analyzed individuals, and represent missing individuals or extinct lineages in the network. Interspecific geographic structure was evident in the network, with instances of shared haplotypes in species occurring in the same region (**Figure 6** and **Supplementary Figure S3**). In contrast, less infraspecific geographic patterning was noted with different populations of species exhibiting unique haplotypes across their geographic range, and in some instances scattered across the network. The two southeastern Australian species share a single haplotype, as do sympatric populations of *A. ellipticus* and *A. cuneatus* in the South Coast region of SWA. Southeastern Australia has less chloroplast haplotype diversity than SWA, containing only one haplotype shared between its two species. This haplotype is also less divergent than other haplotypes in SWA, being separated from its nearest extant haplotype by two extinct haplotype lineages. Potentially long-persisting haplotypes were detected across multiple species; in some cases the age of the chloroplast haplotype pre-dates the

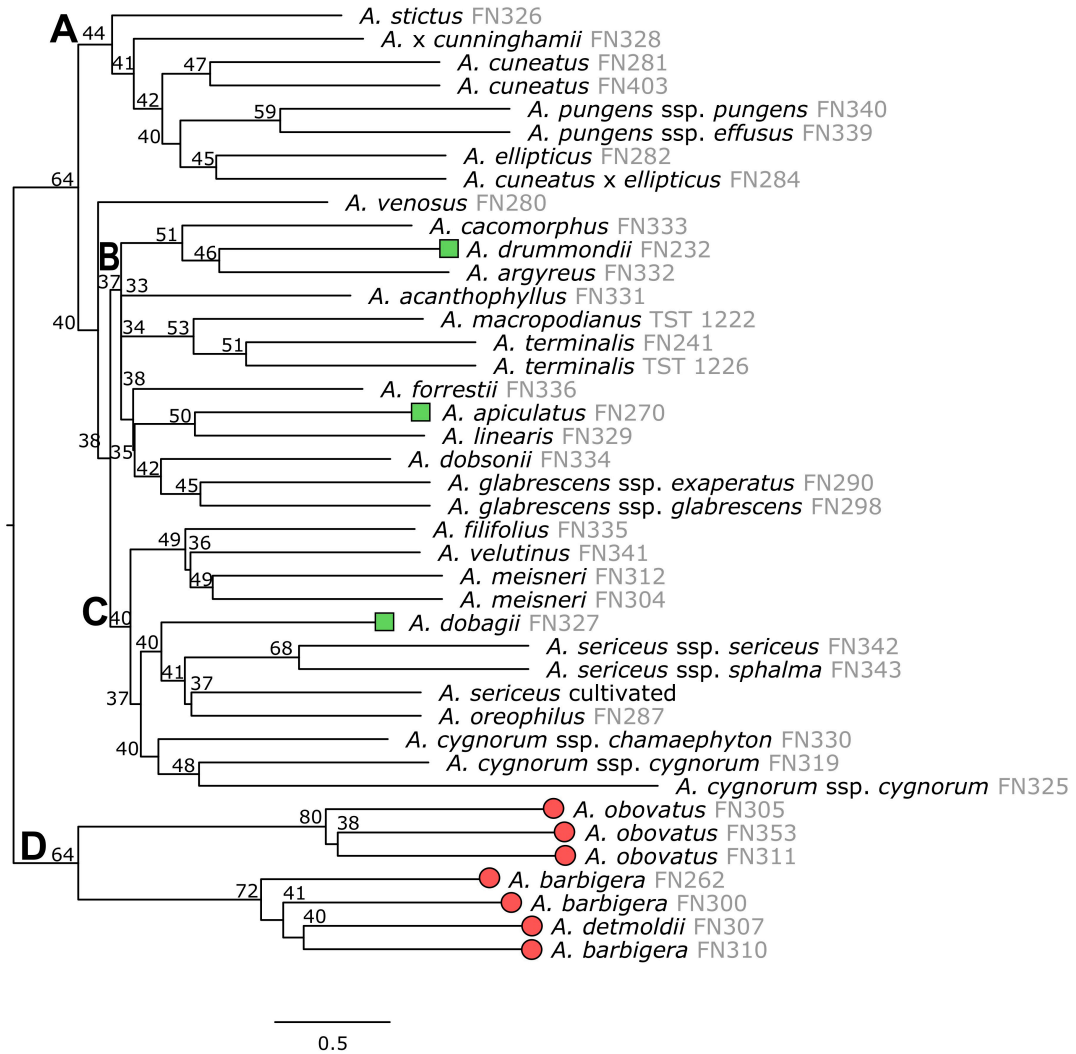

FIGURE 3 | ASTRAL species trees of *Adenanthos* inferred using the multispecies coalescent model and gene trees of nuclear loci. Support values for the ASTRAL tree represent the local quartet score of each branch. Letters **(A–D)** refer to clades discussed in text. Terminal tips represent taxa from *A.* sect. *Eurylaema* (red circles), *A.* sect. *Adenanthos* subsect. *Anaclastos* (green squares), and *A.* subsect. *Adenanthos* (unlabeled).

radiation of the lineage i.e., these haplotypes persisted in extant lineages from their most recent common ancestor (**Table 1**, **Figure 4**, and **Supplementary Figure S2**).

The splitstree networks suggest that reticulate relationships in both nuclear and plastid data sets are largely confined to the backbone (**Supplementary Figures S4–S6**). Both nuclear and plastid networks resolved distinct clades which were in conflict, resulting in the combined dataset having high levels of reticulate relationships throughout the network (**Supplementary Figure S6**).

DISCUSSION

We present the first well-resolved, densely sampled phylogeny of *Adenanthos*. Our results indicate that extensive hybridization is present throughout the evolution of this genus, including deep reticulation events coinciding with the radiation of the genus in the Miocene. A revised infrageneric classification is also warranted to better reflect evolutionary relationships within the genus.

Phylogenetic Incongruence and Reticulate Evolution

Extensive cytonuclear discordance in *Adenanthos* is best explained by multiple introgression events throughout its evolution that are particularly evident across deeper timescales. Several instances of recent introgression between closely related extant species were also detected. While we cannot rule out the presence of ILS as a factor contributing to the observed incongruence in our plastid versus nuclear data, our analyses

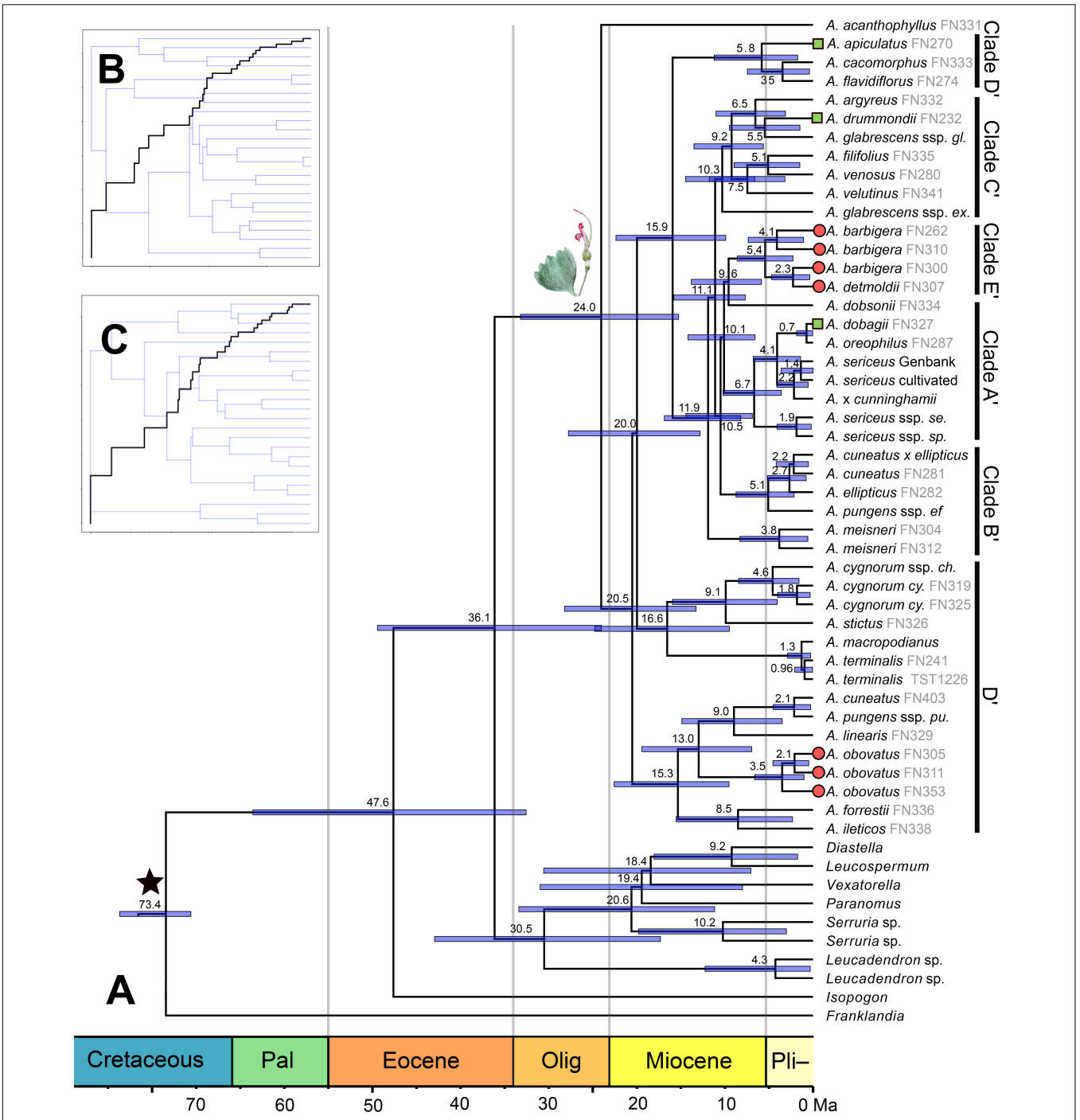

FIGURE 4 | (A) Calibrated BEAST chronogram of *Adenanthos* inferred from 13 chloroplast markers, with branch lengths scaled according to time (Ma). Non-Proteoideae outgroups used in the family-wide calibration scheme were pruned. Divergence times (Ma) are given at each node, with 95% confidence intervals indicated as blue bars. Clade labels denote the clades from the plastid ML RAxML tree. *Adenanthos stictus* illustration denotes the crown of the genus. Star denotes the fossil calibration used in Proteoideae (*Franklandia*). Subspecies are abbreviated by the first two letters of the epithet. **(B)** Lineage-through-time plot of *Adenanthos* based on the BEAST plastid topology. **(C)** Lineage-through-time plot of *Adenanthos* based on the BEAST ITS nuclear topology.

indicate that these conflicts cannot be solely due to ILS, as they are never expected to occur under the coalescent model alone.

At least four independent ancient introgression events in *Adenanthos* were detected in our plastid simulation and JML analyses. In theory, the JML method was developed for detecting hybrids between extant species pairs (Joly et al., 2009; Joly, 2012), and hence might not be optimal for detecting hybridization events that are ancestral in a clade (i.e., from non-extant

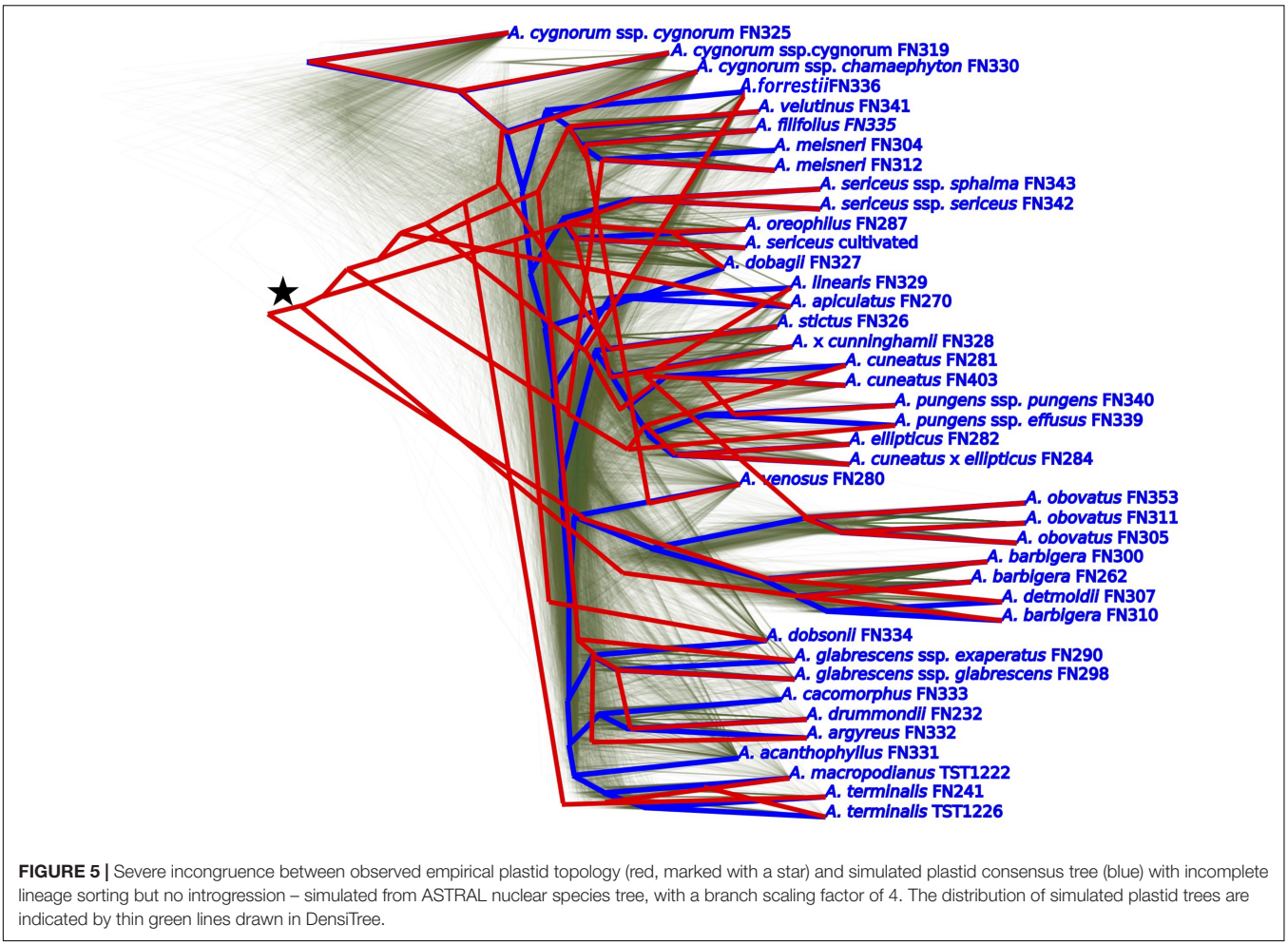

FIGURE 5 | Severe incongruence between observed empirical plastid topology (red, marked with a star) and simulated plastid consensus tree (blue) with incomplete lineage sorting but no introgression – simulated from ASTRAL nuclear species tree, with a branch scaling factor of 4. The distribution of simulated plastid trees are indicated by thin green lines drawn in DensiTree.

lineages) (García et al., 2017; Vargas et al., 2017). However, in our study, this method has been able to detect introgression events between clades, by including a representative subset from each clade in the analyses instead of including all taxa. Indeed, this approach has been recommended to increase statistical power for detecting introgression (Joly, 2012). The detection of ancestral hybridization through JML may also be dependent on the strength of the signal. In *Adenanthos,* hybridization is widespread and deep reticulation events were detectable using pairwise comparisons of extant taxa, whereas in other groups such as rain-lilies (Hippeastreae; Amaryllidaceae), the signal was insufficient or no longer present across sampled taxa (García et al., 2017). Nevertheless, the presence of introgression across clades in *Adenanthos* and rain-lilies was corroborated with other gene simulation analyses. We encourage the use of multiple methods for detecting hybridization events throughout the evolution of a study group. We also acknowledge potential limitations in detecting introgression through the use of nuclear data in simulating a plastid topology with only ILS, as nuclear genes may also show signals of introgression. This is not the case in our study, where the majority of species showed no signs of introgression in the nuclear topology except for *A.* × *cunninghamii.* Further studies with additional population sampling and methological advancements in distinguishing ILS from introgression should provide us with a greater understanding on this topic.

Different plant groups are prone to different degrees of hybridization (Whitney et al., 2010). While hybrids have been identified in some genera of Proteaceae (Lamont et al., 2003; Pharmawati et al., 2004; Milner et al., 2012; McIntosh et al., 2014; Mitchell and Holsinger, 2018), hybridization appears to be uncommon within the family as a whole. Apart from *Adenanthos,* only the eastern Australian genus *Lomatia* shows extensive signals of interspecific hybridization (McIntosh et al., 2014). To our knowledge, our study is the first to demonstrate deep reticulation events within a genus in Proteaceae. The biological mechanisms that maintain species boundaries in the face of such extensive past and present introgression in *Adenanthos* are currently unknown.

Radiation of *Adenanthos* and Long-Persisting Haplotypes

The unresolved backbone of *Adenanthos* found in both our NGS nuclear and chloroplast topologies is suggestive of a rapid radiation in the Oligocene-Miocene. Organismal groups that

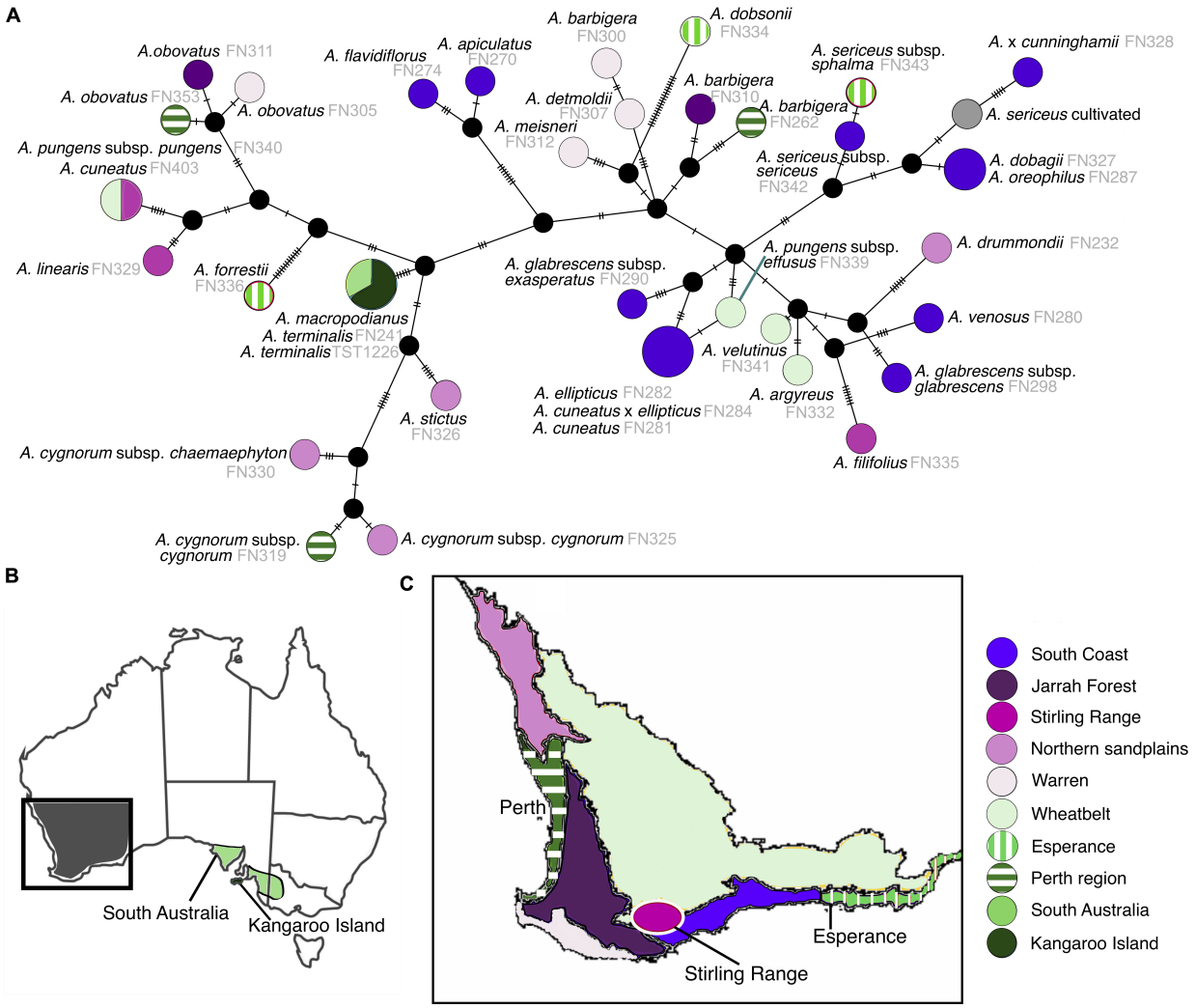

FIGURE 6 | (A) TCS network of *Adenanthos* chloroplast haplotypes colored according to subregions of southwestern and South Australia. Circle size represents number of taxa (1–3) sharing chloroplast haplotypes. Black dots represent missing, unsampled, or extinct intermediate haplotype lineages. **(B)** Map of Australia with box highlighting southwestern Australia and colored regions in South Australia indicate distribution of *Adenanthos terminalis* and *macropodianus*. **(C)** Map of southwestern Australia showing each of the subregions.

have undergone a rapid radiation appear to be particularly prone to reticulate evolution (Anderson and Stebbins, 1954; Seehausen, 2004, 2013; Mallet et al., 2007; Genner and Turner, 2011; Vargas et al., 2017). Species boundaries may be more porous during a radiation event (Dilley et al., 2000; Smith et al., 2008), potentially driving reticulate evolutionary patterns and incongruence between nuclear and plastid genealogies, as observed in *Adenanthos*.

We show that, in some cases, sister species or sampled individuals within species have highly divergent chloroplast haplotypes, each of which is most similar to haplotypes found in phylogenetically distant extant species according to the nuclear topology. For example, the haplotypes of *A. obovatus* and *A. barbigera*, which are sister species in the nuclear phylogeny, are separated by at least six extinct haplotypes across the backbone of the network. We interpret this as resulting from introgression between these species and extinct lineages. Introgression may have occurred at any time between the divergence of the sister pair at *c.* 11.6 Ma and the divergence of populations within *A. obovatus* and *A. barbigera* (*c.* 2 Ma) (**Supplementary Figure S2**: ITS BEAST topology). The older estimate of *c.* 11.6 Ma for these potentially long-persisting haplotypes is significantly older than ancient chloroplast haplotypes noted in other plants, for example, *c.* 4 Ma in Jakob and Blattner (2006).

Complex geographic patterning of *Adenanthos* chloroplast haplotypes is seen in many areas of southwestern Western Australia, where multiple distinct and highly divergent haplotypes are present in localized areas. The South Coast subregion contains the highest haplotype diversity across SWA, followed by the Esperance, Perth, and Northern Sandplains subregions, indicating that highly divergent, old chloroplast haplotypes have persisted in these areas (Byrne, 2008). Several

TABLE 1 | Divergence age estimates for long-persisting chloroplast haplotypes in *Adenanthos*, with nuclear divergence estimate obtained from the ITS topology.

Long-persisting haplotypes	Haplotype age (Ma)	ITS nuclear divergence (Ma)
A. obovatus stem	13.0(6.9 – 19.4)	11.6(4.2 – 18.75)
A. obovatus crown	3.5(1.0 – 6.6)	1.8(0 – 4.3)
A. barbigera stem	9.6(5.9 – 13.8)	11.6(4.2 – 18.75)
A. barbigera crown	5.4(2.3 – 8.6)	1.9(0.03 – 4.7)
A. forrestii	8.5(2.3 – 15.5)	2.2(0.01 – 6.0)
A. apiculatus	5.8(1.8 – 11.2)	2.1(0.04 – 5.3)
A. linearis	9.0(3.5 – 14.9)	2.1(0.04 – 5.3)
A. drummondii	5.5(1.5 – 9.5)	2.8(0.4 – 6.1)
A. argyreus	6.5(3.1 – 11.0)	2.8(0.4 – 6.1)
A. glabrescens subsp. *exasperatus*	10.3(6.6 – 14.5)	4.2(0.3 – 8.8)
A. dobsonii	9.6(5.9 – 13.8)	4.2(0.3 – 8.8)

The 95% credibility intervals are noted in brackets.

of these regions (Perth and Northern Sandplains) are also centres of floristic species richness (Hopper and Gioia, 2004) and phylogenetic diversity (Rosauer et al., 2009) in SWA, and hence are of high conservation value. Sniderman et al. (2013) and Nge et al. (2020) have hypothesized that relatively low extinction rates in SWA compared to other regions of Australia, due to climatic buffering over the course of multiple large-scale Eocene–Pleistocene climatic events, is one of the main drivers for these patterns.

The lower chloroplast haplotype diversity in southeastern Australia can be attributed to either higher local extinction rates compared with SWA, or a founder effect where a lineage was dispersed from SWA to southeastern Australia. Our divergence estimates based on the ITS topology for the disjunction of *Adenanthos* species across southern Australia postdates or coincides with the uplift of the Nullarbor Plain *c.* 14–13 Ma, which is a strong climatic and edaphic barrier for plant migration between the two southern temperate mesic regions. However, we caution against attributing this divergence solely to this vicariance event, as the 95% CI of this divergence event in *Adenanthos* (4.8–15.3 Ma) is too wide to discriminate between divergence as a direct result of the uplift of the Nullarbor Plain or post-uplift dispersal (see also Crisp and Cook, 2007). Further studies on the population genetics of the southeastern Australian species and additional sampling of outgroups with NGS nuclear data should provide us with more precise divergence age estimates of the clade in relation to its SWA sister groups. Not only does the southeastern Australian clade contain lower haplotype diversity, the clade is also relatively depauperate, containing only two species compared to more species-rich clades found in SWA. Future research on drivers of this disparity in species richness linking it with genetic mechanisms and results of this study would be especially promising.

The radiation of *Adenanthos* coincides with that inferred in many other Australian plant radiations during the Miocene (Crisp et al., 2004; Cardillo and Pratt, 2013; Puente-Lelièvre et al., 2013; Jabaily et al., 2014; Mast et al., 2015; Thornhill et al., 2019). Intensification of aridity and seasonality of rainfall across the continent at this time resulted in the retreat of mesic vegetation types and expansion of sclerophyllous and xeromorphic vegetation (Byrne et al., 2011; Crisp and Cook, 2013). These changes opened new niches, potentially allowing *Adenanthos* and other sclerophyllous groups to diversify. Hybridization between distinct lineages spurring adaptive radiations has been demonstrated for Hawaiian silverwords (Barrier et al., 1999) and African lake cichlids (Meier et al., 2017), with introgression providing genetic novelty leading to diversification into new niches. A recent review (Berner and Salzburger, 2015) has suggested that many adaptive radiations exhibit signals of hybridization, highlighting an important link between novel genetic variation derived from hybridization or introgression of adaptive genes and evolutionary radiations (Seehausen, 2004). Ours is the first study to assess these links in the context of the Australian flora; testing whether deep reticulate evolution is common or detectable in other Australian plant groups with similar radiations is an important next step. Explicitly testing for the adaptive function of introgressed genes in groups that experienced a rapid radiation and show signals of hybridization would further advance our understanding of the role that hybridization plays in the evolution of such groups (Suarez-Gonzalez et al., 2018).

Introgression between extant *Adenanthos* species in this study and others (Walker et al., 2018) show that reticulate evolution is ongoing. Porous species boundaries may still be evolutionarily advantageous in *Adenanthos,* resulting in intermediate phenotypes that can occupy different niches to the parental species (Givnish, 2010). This has been observed in *Adenanthos*, with hybrids observed to occupy intermediate or disturbed habitats (Nelson, 1977) and some taxa are known to be disturbance specialists (Groom and Lamont, 2015). Further studies are required to investigate whether ongoing gene flow across extant species plays an important role in the speciation of the genus (e.g., through homoploid hybrid speciation).

Recent Introgression and Hybridization Events

Several instances of recent introgression between extant *Adenanthos* species, while mainly isolated to closely related clade-specific lineages, were also detected in addition to strong incongruence predominantly shown across deeper scales across the backbone of the genus. These include potential chloroplast capture events where sympatric populations of one species share their plastid with another species with an overlapping geographic range (Rieseberg and Soltis, 1991). Examples include *A. macropodianus–A. terminalis, A. ellipticus–A. cuneatus,* and *A. dobagii–A. oreophilus* (**Figures 2**, **6**). The chloroplast capture event for the southeastern Australian species occurred relatively recently in the Pleistocene (*c.* 1.3 Ma, 95% CI: 0.3–2.9 Ma) compared with the species divergence of the pair (*A. macropodianus–A. terminalis*) estimated at 15.3 Ma (95% CI: 8.2–23.0 Ma) based on the nuclear topology. Signals of ancient chloroplast capture events were also evident for *A. stictus–A. cygnorum,* and

A. pungens/A.cuneatus–A. linearis, which have close geographic proximity and exhibit closely related chloroplast haplotypes but are phylogenetically distant in the nuclear phylogeny (**Figures 2**, **6**). In these cases, the chloroplast reflects more the geographic patterning of these taxa instead of species relationships. Other hybridization events between extant species include *Adenanthos* × *cunninghamii* which shows conflicting placements in the nuclear and plastid topologies (sister to *A. cuneatus* and *A. sericeus,* respectively), corroborating the study by Walker et al. (2018) that it is a hybrid between *A. cuneatus* and *A. sericeus*. The cultivated *A. sericeus* is sister to *A. dobagii, A. oreophilus* and *A. sericeus* in both our nuclear and plastid topologies. We hypothesize that it is likely of a hybrid origin, with a parent species (likely maternal based on the plastid topology) from the *A. sericeus* clade and another undetermined parental species. Indeed the *A. sericeus* cultivar might be of multiple hybrid origins resulting from repeated back-crossing events, as two unsampled (or extinct) plastid haplotypes link it with the wild *A. sericeus* samples, and one unsampled haplotype with *A. dobagii* and *A. oreophilus* (Rieseberg and Brunsfeld, 1992). The putative *A. cuneatus* × *ellipticus* hybrid is sister to *A. ellipticus* in both our nuclear and plastid topologies but shares the same chloroplast haplotype with sympatric *A. ellipticus* and *A. cuneatus* individuals. Further studies applying extensive population sampling to assess for introgression across populations for these putative hybrids are required to confirm their status as well as the identities of their parent species.

Species Tree and Infrageneric Classification of *Adenanthos*

Extensive hybridization, as detected by our analyses, best explains the observed incongruence between nuclear and chloroplast phylogenies of *Adenanthos*. The nDNA topology is largely congruent with taxonomic concepts for the genus derived from morphology, whereas the plastid topology contains strong signals of multiple introgression events. This finding coupled with little to no detectable signal of introgression from our nuclear data warrants further discussion. It is possible that selection for adaptive organellar introgression or prevention of nuclear introgression could explain our results (Bonnet et al., 2017). Based on the simulaton study of Bonnet et al. (2017), these scenarios are the main drivers for this pattern where there is little evidence for nuclear introgression despite strong discordance between nuclear and organellar genomes. Local selection for different chloroplast genomes have been linked to different environmental performance of these genomes (Sambatti et al., 2008; Sloan et al., 2017). In sunflowers, for example, local adaptation to drier or wetter parts of a species' range has contributed to multiple organellar introgression events across the genus (Sambatti et al., 2008; Lee-Yaw et al., 2019). It would be interesting to test in future studies whether the diverse chloroplast haplotypes and introgression events found within *Adenanthos* in SWA are the result of strong selection pressure for adaptation to local environmental conditions.

Many other studies have demonstrated strong cytonuclear discordance and separate evolutionary histories of plastid/mitochondrial vs. nuclear DNA (Soltis and Kuzoff, 1995; Yoo et al., 2002; Barrett et al., 2015) and some (e.g., Acosta and Premoli, 2010; Othman et al., 2010; Barrett et al., 2015; Folk et al., 2017; Vargas et al., 2017) have documented cases where plastid topologies are not reflective of species trees in comparison with nuclear data. Because the plastome is non-recombining and uniparentally inherited, a chloroplast lineage in one species can be replaced by an alien one following a single hybridization event (chloroplast capture), with the newly acquired chloroplast inherited by descendant lineages and persisting over long evolutionary timescales. This is expected to lead to plastome gene trees that are highly discordant with the species tree. Tree inference under the coalescent model using multiple nuclear loci is expected to provide a more accurate estimate of the species tree as compared to organellar genes that exhibit uniparental inheritance (Birky, 1995). For this reason, a cautious approach is needed when interpreting evolutionary signals between organellar and nuclear data. In particular, combining these datasets when they are in strong conflict will most likely compromise interpretations of evolutionary history.

While our nDNA topology is largely consistent with morphology in *Adenanthos*, nevertheless the infrageneric classification proposed by Nelson (1977) is partially inadequate. Nelson recognized two sections in *Adenanthos,* sect. *Eurylaema* and sect. *Adenanthos*, based on anther and style morphology, and two subsections within sect. *Adenanthos* based solely on perianth length. In our study, *A.* sect. *Eurylaema* was resolved as monophyletic in the nuclear topology but not the plastid topology, likely due to an ancient introgression event between *A. obovatus* and/or *A. barbigera* (sect. *Eurylaema*) with an extinct lineage from sect. *Adenanthos*. Both subsections of *A.* sect. *Adenanthos* were recovered as polyphyletic in both our plastid and nuclear datasets, from concatenated and coalescent analyses. We do not support the recognition of subsections within *A.* sect. *Adenanthos* and recommend they be merged.

CONCLUSION

Our study used complementary simulation approaches to detect introgression events across multiple scales within *Adenanthos,* and linked deep reticulate evolution to a rapid radiation in the Miocene coinciding with widespread aridification of the Australian continent. Dense sampling within *Adenanthos* allowed us to infer the extent and timing of introgression events within the genus. Reticulate signals were detected in a complex pattern of long-persisting haplotypes scattered across phylogenetically distant extant species. Some of these ancient chloroplast haplotypes are estimated to have diverged up to 12 Ma and may have persisted in southwestern Western Australia due to the relative stability of the landscape and buffering from major extinctions. Important open questions are the degree to which other Australian plant radiations show similar signals of reticulate

evolution, and the effects of hybridization and introgression on their diversification.

DATA AVAILABILITY STATEMENT

Datasets presented in this study can be found on the Dryad Digital Repository: doi: 10.5061/dryad.zw3r22876. Nge et al. (2020), data for the manuscript: reticulate evolution, ancient chloroplast haplotypes, and rapid radiation of the Australian plant genus - Adenanthos (*Proteaceae*), Dryad, Dataset: doi: 10.5061/dryad.zw3r22876.

AUTHOR CONTRIBUTIONS

FN, EB, KT, and MW conceived the ideas for this study. FN designed the study. FN, EB, and KT collected the data. FN and EB complied the data. FN analyzed the data. FN wrote the manuscript with contributions from EB, KT, and MW. All authors contributed to the article and approved the submitted version.

FUNDING

Funding for this project was supported by the South Australian Department of Environment, Water and Natural Resources (D0004335204). FN was supported by an Australian Government Research Training Program (RTP) scholarship.

ACKNOWLEDGMENTS

We thank Hans Lambers, Marion Cambridge, Roberta Dayrell, Graham Zemunik, Rosalie and Robert Lawrence, Anne Rick and members of the Wildflower Society of Western Australia for their assistance in the field, Greg Keighery for sharing his knowledge on the current phrase-named taxa as well as hybrids within the genus, Dan Duval for providing material from the South Australian Seed Conservation Centre, Korjent van Dijk for guidance and help in laboratory work, Ryan Folk and Nicolás García for insightful discussions and help with the DendroPy simulation script, and Oscar Vargas for helpful discussions and clarification on the technicality of the JML software. We acknowledge the Department of Biodiversity, Conservation and Attractions (Western Australia; permit no. SW 019140), and Department of Environment, Water and Natural Resources (South Australia; permit no. G25787-4) for their permission to collect leaf samples on land under their administration. We also thank Amanda Shade and Fernanda Veraldo for permission and help in obtaining fresh leaf material from *Adenanthos* specimens grown at the Kings Park Botanic Gardens in Perth. Lastly, we would also like to thank the curation teams from both the Western Australian (PERTH) and South Australian Herbarium (AD) for their assistance and for providing access to herbarium specimens.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fevo.2020.616741/full#supplementary-material

REFERENCES

Acosta, M. C., and Premoli, A. C. (2010). Evidence of chloroplast capture in Aouth American Nothofagus (subgenus Nothofagus, Nothofagaceae). *Mol. Phylogenet. Evol.* 54, 235–242.

Anderson, E., and Stebbins, G. L. Jr. (1954). Hybridization as an evolutionary stimulus. *Evolution* 8, 378–388.

Arnold, M. L. (1992). Natural hybridization as an evolutionary process. *Annu. Rev. Ecol. Syst.* 23, 237–261.

Ashton, D., and Sandiford, E. (1988). Natural hybridisation between *Eucalyptus regnans F. Muell.* and *E. macrorhyncha F. Muell.* in the Cathedral Range, Victoria. *Austral. J. Bot.* 36, 1–22.

Barrett, R. A., Bayly, M. J., Duretto, M. F., Forster, P. I., Ladiges, P. Y., and Cantrill, D. J. (2015). A chloroplast phylogeny of *Zieria* (Rutaceae) in Australia and New Caledonia shows widespread incongruence with species-level taxonomy. *Austral. Syst. Bot.* 27, 427–449.

Barrier, M., Baldwin, B. G., Robichaux, R. H., and Purugganan, M. D. (1999). Interspecific hybrid ancestry of a plant adaptive radiation: allopolyploidy of the Hawaiian silversword alliance (Asteraceae) inferred from floral homeotic gene duplications. *Mol. Biol. Evol.* 16, 1105–1113.

Berner, D., and Salzburger, W. (2015). The genomics of organismal diversification illuminated by adaptive radiations. *Trends Genet.* 31, 491–499.

Birky, C. W. (1995). Uniparental inheritance of mitochondrial and chloroplast genes: mechanisms and evolution. *Proc. Natl. Acad. Sci. U.S.A.* 92, 11331–11338.

Bonnet, T., Leblois, R., Rousset, F., and Crochet, P. A. (2017). A reassessment of explanations for discordant introgressions of mitochondrial and nuclear genomes. *Evolution* 71, 2140–2158.

Bouckaert, R., and Heled, J. (2014). DensiTree 2: seeing trees through the forest. *BioRxiv* [Preprint]. doi: 10.1101/012401v1

Bouckaert, R., Heled, J., Kühnert, D., Vaughan, T., Wu, C.-H., Xie, D., et al. (2014). BEAST 2: a software platform for Bayesian evolutionary analysis. *PLoS Comput. Biol.* 10:e1003537. doi: 10.1371/journal.pcbi.1003537

Byrne, M. (2008). Evidence for multiple refugia at different time scales during Pleistocene climatic oscillations in southern Australia inferred from phylogeography. *Quatern. Sci. Rev.* 27, 2576–2585.

Byrne, M., Steane, D. A., Joseph, L., Yeates, D. K., Jordan, G. J., Crayn, D., et al. (2011). Decline of a biome: evolution, contraction, fragmentation, extinction and invasion of the Australian mesic zone biota. *J. Biogeogr.* 38, 1635–1656.

Cardillo, M., and Pratt, R. (2013). Evolution of a hotspot genus: geographic variation in speciation and extinction rates in *Banksia* (Proteaceae). *BMC Evol. Biol.* 13:155. doi: 10.1186/1471-2148-13-155

Clement, M., Posada, D., and Crandall, K. A. (2000). TCS: a computer program to estimate gene genealogies. *Mol. Ecol.* 9, 1657–1659.

Collins, B. G., and Rebelo, T. (1987). Pollination biology of the Proteaceae in Australia and southern Africa. *Austral. J. Ecol.* 12, 387–421.

Crisp, M. D., Burrows, G. E., Cook, L. G., Thornhill, A. H., and Bowman, D. M. (2011). Flammable biomes dominated by eucalypts originated at the Cretaceous-Palaeogene boundary. *Nat. Commun.* 2, 1–8.

Crisp, M. D., Cook, L., and Steane, D. (2004). Radiation of the Australian flora: what can comparisons of molecular phylogenies across multiple taxa tell us about the evolution of diversity in present–day communities? *Philos. Trans. R. Soc. Lond. B* 359, 1551–1571.

Crisp, M. D., and Cook, L. G. (2007). A congruent molecular signature of vicariance across multiple plant lineages. *Mol. Phylogenet. Evol.* 43, 1106–1117.

Crisp, M. D., and Cook, L. G. (2013). How was the Australian flora assembled over the last 65 million years? A molecular phylogenetic perspective. *Annu. Rev. Ecol. Evol. Syst.* 44, 303–324.

Dilley, J. D., Wilson, P., and Mesler, M. R. (2000). The radiation of Calochortus: generalist flowers moving through a mosaic of potential pollinators. *Oikos* 89, 209–222.

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* 32, 1792–1797.

Folk, R. A., Mandel, J. R., and Freudenstein, J. V. (2017). Ancestral gene flow and parallel organellar genome capture result in extreme phylogenomic discord in a lineage of angiosperms. *Syst. Biol.* 66, 320–337.

Francisco-Ortega, J., Jansen, R. K., and Santos-Guerra, A. (1996). Chloroplast DNA evidence of colonization, adaptive radiation, and hybridization in the evolution of the *Macaronesian flora*. *Proc. Natl. Acad. Sci. U.S.A.* 93, 4085–4090.

García, N., Folk, R. A., Meerow, A. W., Chamala, S., Gitzendanner, M. A., de Oliveira, R. S., et al. (2017). Deep reticulation and incomplete lineage sorting obscure the diploid phylogeny of rain-lilies and allies (Amaryllidaceae tribe Hippeastreae). *Mol. Phylogenet. Evol.* 111, 231–247.

Genner, M. J., and Turner, G. F. (2011). Ancient hybridization and phenotypic novelty within Lake Malawi's cichlid fish radiation. *Mol. Biol. Evol.* 29, 195–206.

Givnish, T. J. (2010). Ecology of plant speciation. *Taxon* 59, 1326–1366.

Goldingay, R. L., and Carthew, S. M. (1998). Breeding and mating systems of Australian Proteaceae. *Austral. J. Bot.* 46, 421–437.

Griffin, A., Burgess, I., and Wolf, L. (1988). Patterns of natural and manipulated hybridisation in the genus *Eucalyptus* L'hérit. <i>1 a review. *Austral. J. Bot.* 36, 41–66.

Groom, P. K., and Lamont, B. (2015). *Plant Life of Southwestern Australia: Adaptations for Survival*. Berlin: Walter de Gruyter GmbH & Co KG.

Gurushidze, M., Fritsch, R. M., and Blattner, F. R. (2010). Species–level phylogeny of Allium subgenus Melanocrommyum: incomplete lineage sorting, hybridization and trnF gene duplication. *Taxon* 59, 829–840.

Holman, J., and Playford, J. (2000). Molecular and morphological variation in the Senna artemisioides complex. *Austral. J. Bot.* 48, 569–579.

Hopper, S. D. (2009). OCBIL theory: towards an integrated understanding of the evolution, ecology and conservation of biodiversity on old, climatically buffered, infertile landscapes. *Plant Soil* 322, 49–86.

Hopper, S. D., and Gioia, P. (2004). The Southwest Australian floristic region: evolution and conservation of a global hot spot of biodiversity. *Annu. Rev. Ecol. Evol. Syst.* 35, 623–650.

Howarth, D. G., and Baum, D. A. (2005). Genealogical evidence of homoploid hybrid speciation in an adaptive radiation of Scaevola (Goodeniaceae) in the Hawaiian Islands. *Evolution* 59, 948–961.

Huson, D. H., and Bryant, D. (2006). Application of phylogenetic networks in evolutionary studies. *Mol. Biol. Evol.* 23, 254–267.

Huson, D. H., and Scornavacca, C. (2012). Dendroscope 3: an interactive tool for rooted phylogenetic trees and networks. *Syst. Biol.* 61, 1061–1067.

Jabaily, R. S., Shepherd, K. A., Gardner, A. G., Gustafsson, M. H., Howarth, D. G., and Motley, T. J. (2014). Historical biogeography of the predominantly Australian plant family Goodeniaceae. *J. Biogeogr.* 41, 2057–2067.

Jakob, S. S., and Blattner, F. R. (2006). A chloroplast genealogy of Hordeum (Poaceae): long-term persisting haplotypes, incomplete lineage sorting, regional extinction, and the consequences for phylogenetic inference. *Mol. Biol. Evol.* 23, 1602–1612.

Joly, S. (2012). JML: testing hybridization from species trees. *Mol. Ecol. Resour.* 12, 179–184.

Joly, S., McLenachan, P. A., and Lockhart, P. J. (2009). A statistical approach for distinguishing hybridization and incomplete lineage sorting. *Am. Nat.* 174, E54–E70.

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.

Keighery, G. (1982). "Bird-pollinated plants in Western Australia," in *Pollination and Evolution. Royal Botanic Gardens Sydney*, eds J. A. Armstrong, J. W. Powell, and A. J. Richards (Cham: Springer), 77–89.

Lamont, B., He, T., Enright, N., Krauss, S., and Miller, B. (2003). Anthropogenic disturbance promotes hybridization between Banksia species by altering their biology. *J. Evol. Biol.* 16, 551–557.

Lamont, B. B., and He, T. (2012). Fire-adapted Gondwanan angiosperm floras evolved in the *Cretaceous*. *BMC Evol. Biol.* 12:223. doi: 10.1186/1471-2148-12-223

Leach, G., and Whiffin, T. (1978). Analysis of a hybrid swarm between *Acacia brachybotrya* and *A. calamifolia* (Leguminosae). *Bot. J. Linn. Soc.* 76, 53–69.

Lee-Yaw, J. A., Grassa, C. J., Joly, S., Andrew, R. L., and Rieseberg, L. H. (2019). An evaluation of alternative explanations for widespread cytonuclear discordance in annual sunflowers (Helianthus). *New Phytol.* 221, 515–526.

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

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

Lemmon, E. M., and Lemmon, A. R. (2013). High-throughput genomic data in systematics and phylogenetics. *Annu. Rev. Ecol. Evol. Syst.* 44, 99–121.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics* 25, 2078–2079.

Macphail, M. (2007). *Australian Palaeoclimates: Cretaceous to Tertiary a Review of Palaeobotanical and Related Evidence to the Year 2000*. Bentley: CRC LEME.

Mallet, J. (2007). Hybrid speciation. *Nature* 446:279.

Mallet, J., Beltrán, M., Neukirchen, W., and Linares, M. (2007). Natural hybridization in heliconiine butterflies: the species boundary as a continuum. *BMC Evol. Biol.* 7:28. doi: 10.1186/1471-2148-7-28

Mast, A. R., Olde, P. M., Makinson, R. O., Jones, E., Kubes, A., Miller, E. T., et al. (2015). Paraphyly changes understanding of timing and tempo of diversification in subtribe Hakeinae (Proteaceae), a giant Australian plant radiation. *Am. J. Bot.* 102, 1634–1646.

McIntosh, E. J., Rossetto, M., Weston, P. H., and Wardle, G. M. (2014). Maintenance of strong morphological differentiation despite ongoing natural hybridization between sympatric species of *Lomatia* (Proteaceae). *Ann. Bot.* 113, 861–872.

Meier, J. I., Marques, D. A., Mwaiko, S., Wagner, C. E., Excoffier, L., and Seehausen, O. (2017). Ancient hybridization fuels rapid cichlid fish adaptive radiations. *Nat. Commun.* 8, 1–11.

Miller, M. A., Pfeiffer, W., and Schwartz, T. (2010). "Creating the CIPRES Science Gateway for inference of large phylogenetic trees," in *Proceedings of the Gateway Computing Environments Workshop (GCE)*, New Orleans, LA, 1–8.

Milner, M. L., Rossetto, M., Crisp, M. D., and Weston, P. H. (2012). The impact of multiple biogeographic barriers and hybridization on species-level differentiation. *Am. J. Bot.* 99, 2045–2057.

Mirarab, S., Bayzid, M. S., Boussau, B., and Warnow, T. (2014). Statistical binning enables an accurate coalescent-based estimation of the avian tree. *Science* 346:1250463.

Mitchell, N., and Holsinger, K. (2018). Cryptic natural hybridization between two species of Protea. *S. Afr. J. Bot.* 118, 306–314.

Mogensen, H. L. (1996). The hows and whys of cytoplasmic inheritance in seed plants. *Am. J. Bot.* 83, 383–404.

Nelson, E. C. (1977). A taxonomic revision of the genus Adenanthos (Proteaceae). *Brunonia* 1, 303–406.

Nge, F. J., Biffin, E., Thiele, K. R., and Waycott, M. (2020). Extinction pulse at Eocene–Oligocene boundary drives diversification dynamics of the two Australian temperate floras. *Proc. R. Soc. B* 287:20192546.

Othman, R. N. A., Jordan, G. J., Worth, J. R., Steane, D. A., and Duretto, M. F. (2010). Phylogeny and infrageneric classification of Correa Andrews (Rutaceae) on the basis of nuclear and chloroplast DNA. *Plant Syst. Evol.* 288, 127–138.

Pharmawati, M., Yan, G., Sedgley, R., and Finnegan, P. (2004). Chloroplast DNA inheritance and variation in Leucadendron species (Proteaceae) as revealed by PCR-RFLP. *Theor. Appl. Genet.* 109, 1694–1701.

Potts, B., and Reid, J. (1985). Analysis of a hybrid swarm between *Eucalyptus risdonii* Hook. f. and *E. amygdalina* Labill. *Austral. J. Bot.* 33, 543–562.

Puente-Lelièvre, C., Harrington, M. G., Brown, E. A., Kuzmina, M., and Crayn, D. M. (2013). Cenozoic extinction and recolonization in the New Zealand flora: the case of the fleshy-fruited epacrids (Styphelieae, Styphelioideae, Ericaceae). *Mol. Phylogenet. Evol.* 66, 203–214.

R Core Team (2016). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Rambaut, A. (2012). *FigTree v1. 4*. Available online at: http://tree.bio.ed.ac.uk/software/figtree/ (accessed March 23, 2020).

Rambaut, A., Suchard, M. A., Xie, D., and Drummond, A. J. (2015). *Tracer v1. 6. 2014 MCMC Trace File Analyser*. Available online at: http://beast.bio.ed.ac.uk/Tracer (accessed March 23, 2020).

Ramsay, H. P. (1963). Chromosome numbers in the Proteaceae. *Austral. J. Bot.* 11, 1–20.

Revell, L. J. (2012). phytools: an R package for phylogenetic comparative biology (and other things). *Methods Ecol. Evol.* 3, 217–223.

Rieseberg, L. H., and Brunsfeld, S. J. (1992). *Molecular evidence and plant introgression. Molecular Systematics of Plants*. Cham: Springer, 151–176.

Rieseberg, L. H., and Soltis, D. (1991). Phylogenetic consequences of cytoplasmic gene flow in plants. *Evol. Trends Plants* 5, 65–84.

Rosauer, D., Laffan, S. W., Crisp, M. D., Donnellan, S. C., and Cook, L. G. (2009). Phylogenetic endemism: a new approach for identifying geographical concentrations of evolutionary history. *Mol. Ecol.* 18, 4061–4072.

Rosenfeld, J. A., Payne, A., and DeSalle, R. (2012). Random roots and lineage sorting. *Mol. Phylogenet. Evol.* 64, 12–20.

Sambatti, J. B., Ortiz-Barrientos, D., Baack, E. J., and Rieseberg, L. H. (2008). Ecological selection maintains cytonuclear incompatibilities in hybridizing sunflowers. *Ecol. Lett.* 11, 1082–1091.

Sauquet, H., Ho, S. Y., Gandolfo, M. A., Jordan, G. J., Wilf, P., Cantrill, D. J., et al. (2012). Testing the impact of calibration on molecular divergence times using a fossil-rich group: the case of Nothofagus (Fagales). *Syst. Biol.* 61, 289–313.

Sauquet, H., Weston, P. H., Anderson, C. L., Barker, N. P., Cantrill, D. J., Mast, A. R., et al. (2009). Contrasted patterns of hyperdiversification in Mediterranean hotspots. *Proc. Natl. Acad. Sci. U.S.A.* 106, 221–225.

Scornavacca, C., Zickmann, F., and Huson, D. H. (2011). Tanglegrams for rooted phylogenetic trees and networks. *Bioinformatics* 27, i248–i256.

Sedgley, M., Harbard, J., Smith, R., Wickneswari, R., and Griffin, A. (1992). Reproductive-biology and interspecific hybridization of *Acacia mangium* and *Acacia auriculiformis A. Cunn.* ex Benth (Leguminosae, Mimosoideae). *Austral. J. Bot.* 40, 37–48.

Seehausen, O. (2004). Hybridization and adaptive radiation. *Trends Ecol. Evol.* 19, 198–207.

Seehausen, O. (2013). Conditions when hybridization might predispose populations for adaptive radiation. *J. Evol. Biol.* 26, 279–281.

Sloan, D. B., Havird, J. C., and Sharbrough, J. (2017). The on-again, off-again relationship between mitochondrial genomes and species boundaries. *Mol. Ecol.* 26, 2212–2236.

Small, R. L., Cronn, R. C., and Wendel, J. F. (2004). Use of nuclear genes for phylogeny reconstruction in plants. *Austral. Syst. Bot.* 17, 145–170.

Smith, S. D., Hall, S. J., Izquierdo, P. R., and Baum, D. A. (2008). Comparative pollination biology of sympatric and allopatric andean iochroma (Solanaceae). *Ann. Missouri Bot. Garden* 95, 600–617.

Sniderman, J. K., Jordan, G. J., and Cowling, R. M. (2013). Fossil evidence for a hyperdiverse sclerophyll flora under a non–Mediterranean-type climate. *Proc. Natl. Acad. Sci. U.S.A.* 110, 3423–3428.

Soltis, D. E., and Kuzoff, R. K. (1995). Discordance between nuclear and chloroplast phylogenies in the Heuchera group (Saxifragaceae). *Evolution* 49, 727–742.

Soltis, P. S., and Soltis, D. E. (2009). The role of hybridization in plant speciation. *Annu. Rev. Plant Biol.* 60, 561–588.

Stace, H. M., Douglas, A. W., and Sampson, J. F. (1998). Did 'paleo-polyploidy' really occur in Proteaceae? *Austral. Syst. Bot.* 11, 613–629.

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

Stankowski, S., and Streisfeld, M. A. (2015). Introgressive hybridization facilitates adaptive divergence in a recent radiation of monkeyflowers. *Proc. R. Soc. B Biol. Sci.* 282:20151666.

Suarez-Gonzalez, A., Lexer, C., and Cronk, Q. C. (2018). Adaptive introgression: a plant perspective. *Biol. Lett.* 14:20170688.

Thornhill, A. H., Crisp, M. D., Külheim, C., Lam, K. E., Nelson, L. A., Yeates, D. K., et al. (2019). A dated molecular perspective of eucalypt taxonomy, evolution and diversification. *Austral. Syst. Bot.* 32, 29–48.

Tian, Y., and Kubatko, L. S. (2014). Gene tree rooting methods give distributions that mimic the coalescent process. *Mol. Phylogenet. Evol.* 70, 63–69.

Vargas, O. M., Ortiz, E. M., and Simpson, B. B. (2017). Conflicting phylogenomic signals reveal a pattern of reticulate evolution in a recent high-Andean diversification (Asteraceae: Astereae: Diplostephium). *New Phytol.* 214, 1736–1750.

Walker, E., Byrne, M., Macdonald, B., Nicolle, D., and McComb, J. (2009). Clonality and hybrid origin of the rare *Eucalyptus bennettiae* (Myrtaceae) in Western Australia. *Austral. J. Bot.* 57, 180–188.

Walker, E., McComb, J., and Byrne, M. (2018). Genetic and morphological evidence supports the hybrid status of *Adenanthos cunninghamii* (now Adenanthos × cunninghamii). *S. Afr. J. Bot.* 118, 299–305.

Weitemier, K., Straub, S. C., Cronn, R. C., Fishbein, M., Schmickl, R., McDonnell, A., et al. (2014). Hyb-Seq: Combining target enrichment and genome skimming for plant phylogenomics. *Appl. Plant Sci.* 2:1400042.

Whitney, K. D., Ahern, J. R., Campbell, L. G., Albert, L. P., and King, M. S. (2010). Patterns of hybridization in plants. *Perspect. Plant Ecol. Evol. Syst.* 12, 175–182.

Willyard, A., Cronn, R., and Liston, A. (2009). Reticulate evolution and incomplete lineage sorting among the ponderosa pines. *Mol. Phylogenet. Evol.* 52, 498–511.

Yoo, K. O., Lowry, P. P., and Wen, J. (2002). Discordance of chloroplast and nuclear ribosomal DNA data in Osmorhiza (Apiaceae). *Am. J. Bot.* 89, 966–971.

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

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: 25 January 2021
doi: 10.3389/fevo.2020.617948

Applications of eDNA Metabarcoding for Vertebrate Diversity Studies in Northern Colombian Water Bodies

Juan Diego Lozano Mojica and Susana Caballero*

Laboratorio de Ecología Molecular de Vertebrados Acuáticos (LEMVA), Departamento de Ciencias Biológicas, Universidad de Los Andes, Bogotá, Colombia

Environmental DNA metabarcoding is a tool with increasing use worldwide. The uses of such technology have been validated several times for diversity census, invasive species detection, and endangered/cryptic/elusive species detection and monitoring. With the help of this technology, water samples collected ($n = 37$) from several main river basins and other water bodies of the northern part of Colombia, including the Magdalena, Sinú, Atrato, and San Jorge river basins, were filtered and analyzed and processed using universal 12S primers for vertebrate fauna and NGS. Over 200 native taxa were detected, the majority of them being fish species but also including amphibia, reptiles, and several non-aquatic species of birds and mammals (around 78, 3, 2, 9, and 8%, respectively). Among the matches, vulnerable, and endangered species such as the catfish *Pseudoplatystoma magdaleniatum* and the Antillean manatee (*Trichechus manatus)* were detected. The manual revision of the data revealed some geographical incongruencies in classification. No invasive species were detected in the filters. This is, to our knowledge, the first time this technique is used in rivers of the country and this tool promises to bring advances in monitoring and conservation efforts, since its low cost and fast deployment allows for sampling in small periods of time, together with the fact that it can detect a wide range of species, allows for a new way of censing the vertebrate diversity in Colombia. Diversity analysis showed how the species identified using this method point to expected community structure although still much needs to be improved in rates of detection and genomic reference databases. This technique could be used in citizen science projects involving local communities in these regions.

Keywords: eDNA metabarcoding, vertebrates, fish communities, Colombia, Magdalena river, Atrato river

OPEN ACCESS

Edited by:
Joong-Ki Park,
Ewha Womans University,
South Korea

Reviewed by:
Allan D. McDevitt,
University of Salford, United Kingdom
Tony Dejean,
Spygen, France

****Correspondence:***
Juan Diego Lozano Mojica
jd.lozano@uniandes.edu.co

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

Received: *15 October 2020*
Accepted: *23 December 2020*
Published: *25 January 2021*

Citation:
Lozano Mojica JD and Caballero S (2021) Applications of eDNA Metabarcoding for Vertebrate Diversity Studies in Northern Colombian Water Bodies. Front. Ecol. Evol. 8:617948. doi: 10.3389/fevo.2020.617948

INTRODUCTION

The term environmental DNA (eDNA) has been used to make reference to the DNA collected from microbial organisms in sediments (Ogram et al., 1987). However with the development of better tools for sequencing and analyzing large amounts of information it was possible to adapt both the technique and the definition to all the DNA found in large environmental samples, both for micro and macroorganisms (Venter et al., 2004; Ficetola et al., 2008). Samples now may come from a wide variety of sources including water, soil, air and feces but most studies have focused on water samples (Drummond et al., 2015; Johnson et al., 2019; Sousa et al., 2019; Yates et al., 2019).

Although it existed well-before this millennium (Ogram et al., 1987), most of the development of this technique (environmental DNA analyses from water samples) occurred in the last 15 years and is already showing important results for species detection and diversity analysis (Ficetola et al., 2008; Jerde et al., 2011; Phalen et al., 2011; Hunter et al., 2015, 2018; Bakker et al., 2017; Castelblanco-Martínez et al., 2018; Tsuji et al., 2019; Yates et al., 2019). Most of these studies have been performed in Europe, Japan or North America (Myers et al., 2000; Arbeláez-Cortés, 2013; Habel et al., 2019). However, the most biodiverse areas in the planet are in developing countries (Myers et al., 2000) and little representation of these places is found among eDNA studies (Sales et al., 2019).

Studying the diversity of an area has always been troublesome, particularly when such areas are of difficult access. The Colombian biodiversity began to be studied with the royal botanical expedition of the New Granada in the late eighteenth century and have been occurring to this day. Increased knowledge has been available in later years by having higher access to previously unreachable locations (due to environmental conditions and safety concerns) and expanding the basis of biological knowledge through biodiversity inventories (Ayala López et al., 2018). While there is a high interest in reaching and studying all the regions of Colombia, keeping updated data from every corner of the country has been a less valued objective. Time, funding, and trained personnel are required in order for these tasks to be completed, and these factors are not as in developed countries. Basic abundance and distribution data remains relevant regardless of the place for reasons including protected areas research and evaluation of human impact on ecosystems evaluation (Pearce and Boyce, 2006; Leathwick et al., 2008; Bakker et al., 2017).

Environmental DNA metagenomics analysis has helped in the study of entire communities (Handley et al., 2019; Nichols and Marko, 2019), specific taxonomic groups (Ostberg et al., 2019), rare/cryptic species (Sakai et al., 2019), vulnerable species (Hunter et al., 2018), and also invasive species (Hunter et al., 2015; Robinson et al., 2019) making it an ideal tool to work on distribution censuses of many taxa. Presence/absence measures are now possible but abundance measures are still not entirely achievable since correct estimations of abundance based on eDNA are not precise enough currently, due to primer sensitivity to target DNA, seasonal variation of eDNA and environmental factors that diminish the correlation between eDNA and abundance (Bylemans et al., 2019; Yates et al., 2019).

For many regions of Colombia, eDNA metabarcoding may be a reliable source of initial information to improve existing biodiversity information by updating or completing it. The easiness with which this technique can be applied in a waterbody could help biologist, local governments, local communities, and NGOs to better understand the natural treasure found in these places. However, since there is only one previous study with this technique in Colombia [focused on tropical reef fish (Polanco Fernández et al., 2020)], much of the information will be hard to compare even with previously obtained data since there is not much genetic information available and databases with said information for comparisons may be incomplete. Other challenges include the physical and chemical properties of the water itself and the preservation methods used in order to obtain good results (Strickler et al., 2015; Sales et al., 2019; Tsuji et al., 2019).

With all of the above in mind, we present initial information on data collected of several water bodies from four river basins in the northern part of Colombia. The general objective was to collect the first diversity data using eDNA metabarcoding in rivers and water bodies from northern Colombia and to explore its opportunities to detect rare, endangered, invasive and cryptic species.

METHODS

Sampling Locations

Two field trips were made in 2019 to the Magdalena, San Jorge and Sinú river basins and to the Atrato river basin (from July 11th to July 20th and October 31st to November 4th, respectively). The chosen places consisted of water bodies and rivers from the four main river basins in northern Colombia-Caribbean region. Several locations required access via canoe or other type of aquatic transportation since all samples were collected from a boat. **Figure 1** presents sampling locations in three main river basins of northern Colombia. Additionally, saltwater samples were taken at Cispatá Bay, and a positive control was made at the lake in the Number 1 marine infantry mobility battalion, for known communities. **Figure 2** presents the four locations where sampling was made in the gulf of Urabá with samples from the Atrato river basin.

Sample Collection

At each sampling location, up to seven, one-liter (1 L) subsamples of water were pooled in a bucket covered with a sterile plastic bag. Each sample was taken from surface water or up to 1 m depth using a plastic bottle and sterile gloves avoiding the contact of skin with the water to avoid human DNA contamination. Each subsample was collected either 50–200 m upstream when in narrow water channels and rivers or in an area of ∼1 km around in a circular transect when in wider water bodies (i.e., swamps). The bottle and bucket were disinfected with 70% alcohol thoroughly (bleach or a more concentrated alcohol were not available at many places and their transport was not viable for many locations) to prevent cross contamination. After taking each sample, the plastic bag was changed for each sampling event to prevent the mixing of water in the bucket. Once all the subsamples were taken the process of filtration began using NatureMetrics eDNA collection kit. The water went through a 0.8 um pore size filter inside a plastic disk until it was clogged, point at which total filtered volume was measured and the kits preservative was added to the filters in order to avoid DNA degradation. Between one and four disks were taken per sampling event due to limited funding to purchase additional filters. Filters were stored in their respective envelopes and later after collection was ended, kept cool in Styrofoam fridges with ice packs until their shipping to NatureMetrics laboratory facilities in England for analysis.

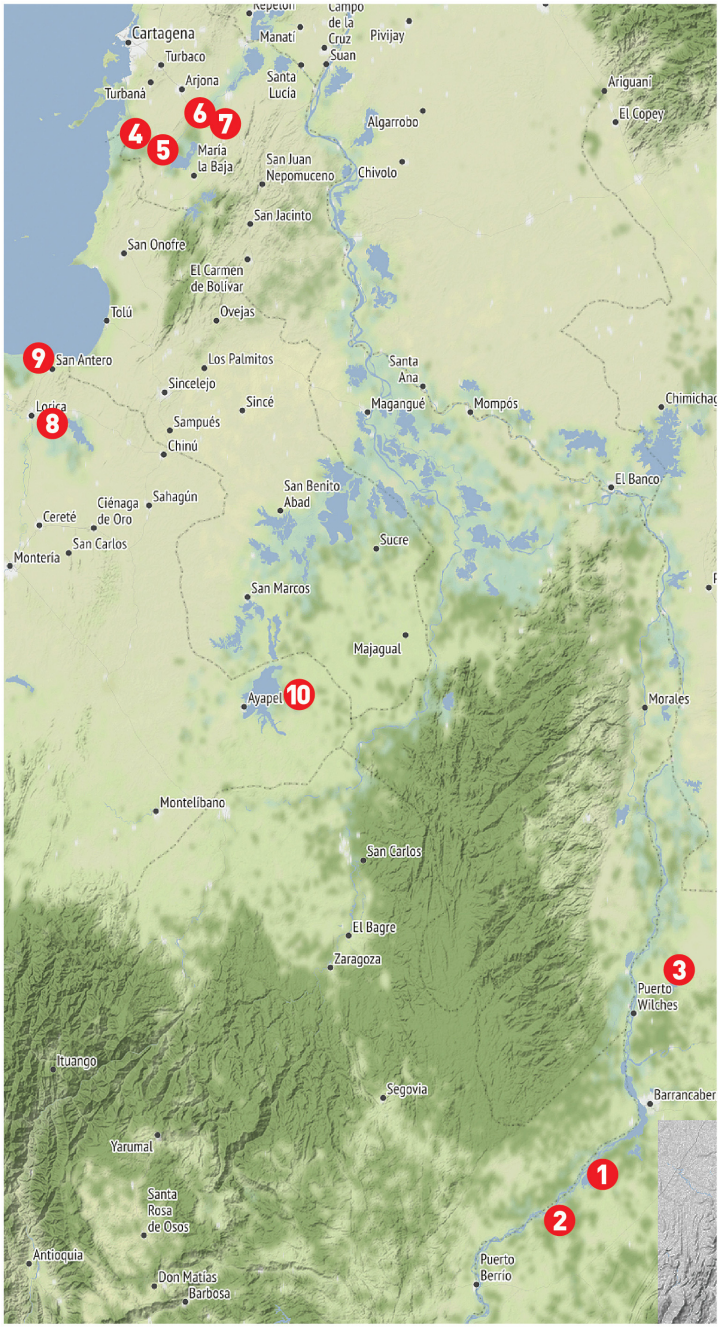

FIGURE 1 | Northern Colombia sampling places: Twenty-five (25) eDNA filters were collected across 10 locations in the central northern region of Colombia. The first three places belong to the middle Magdalena basin. The Chucuri swamp (1) and the San Juan River (2) used 3 filters while the Paredes swamp (3) was sampled with four filters. Samples 4 to 7 belong to the Canal del Dique region where the Magdalena river is deviated from its natural flow. Samples were taken directly in the canal (5) in two of the adjacent and connected swamps (4 and 6) and an artificial lake in the Nr 1 marine infantry mobility battalion (7) for a total of 6 filters between all these places. Sample 8 corresponds to the Lorica swamp (Sinú river basin), sample 9 to the Cispatá bay and sample 10 to the Ayapel swamp as part of the San Jorge river basin (3 filters each).

Sample Processing

Once the filters arrived in the laboratory, DNA was extracted and purified from each filter using DNeasy Blood and tissue kits (Quiagen). Twelve replicate PCRs for the hyper variable region of the 12S rRNA gene with vertebrate primers (Riaz et al., 2011) were run for each sample/filter. Positive controls were made alongside regular PCRs using mock communities of known non-native fish composition in order to verify sequence quality

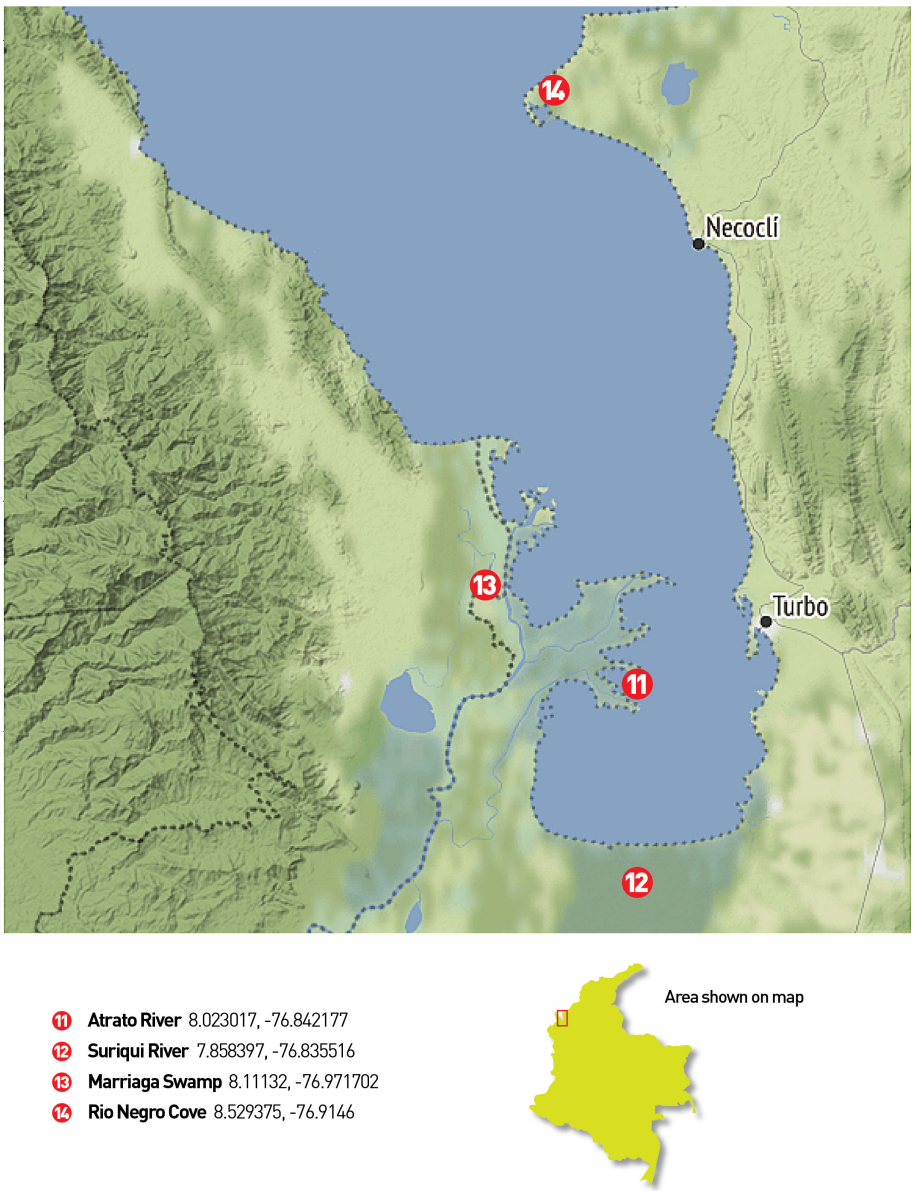

FIGURE 2 | Gulf of Urabá sampling places: 12 filters were collected in the Gulf of Urabá. The first three were taken on one of the Atrato river arms near its end (11), the next four were taken in the Suriquí river and its secondary channels (12), other four filters were used at the Marriaga swamp (13) and the last filter was used near the Rio Negro Cove in the northeastern part of the Gulf (14).

and also a negative control using only distilled water to detect cross contamination if present. Success of the amplifications was confirmed via gel electrophoresis. All amplicons were purified, and adapters were added before pooling all replicates and sequencing them using Illumina MiSeq at 12pM and a 10% PhiX spike in (Miseq V2 2x250 cartridges were used for this process) Sequences were processed using custom bioinformatic pipelines for quality filtering, denoising, and clustering at 99% similarity. Read pairs were merged with usearch v11 (Edgar, 2010) and only keeping pairs with at least 80% agreement in the overlapping region. Cutadapt 2.3 (Martin, 2011) was used to remove primers and short sequences. Quality filter was performed with usearch at an expected error rate of 0.001 and after that they were dereplicated. For the denoising step, unoise was used (Edgar, 2016) and also were clustered at 99%. OTUs were taxonomically assigned to species, genus, family order or class by searching for similarities with the NCBI nucleotide database (GenBank) and PROTAX. Species with matches of 99% or higher similarity and no ambiguity were retained, and genus level matches went through a similar process with matches at 95% similarity or higher. Cases were multiple species were possible, manual check of records of GBIF and IUCN were used to solve the ambiguity.

OTUs that were ≥99% similar and hat similar co-occurrence patterns were combined with LULU (Frøslev et al., 2017) and OTUs were relative abundance in the sample was lower than 0.05% or <10 reads (whichever was the higher) were omitted. Human and livestock sequences were also removed. A second run of taxonomical analysis was made in order to search specially for invasive species designated for the country according to current law (Ministerio De Ambiente Territorial Vivienda Y Desarrollo, 2008; Ministerio De Ambiente Vivienda Y Desarrollo, 2010).

Statistical Analysis

R studio (RStudio Team, 2020) (R Project for Statistical Computing, RRID:SCR_001905) version 3.6.0 was used to perform correlation tests among variables of sampling and results and to perform diversity analysis using the vegan package (Oksanen et al., 2019). Diversity indexes (Shannon-Wienner and Simpson) and statistical analysis were used to evaluate alpha diversity and beta diversity was evaluated using Bray-Curtis dissimilarity index. Since tetrapod detections were scarce and not present at each sample unlike fish, community analysis was performed on fish data only, at genus and species level to compare results between detected data with basic geographic corrections (genus level) and data with confirmed accuracy using available data for the sampling locations (species level).

RESULTS

Sample and Sequencing Quality and Identity

Thirty seven filters were collected at 15 different locations as seen in **Figures 1**, **2**. At each location up to 4 filters were collected. For the 25 samples belonging to the Magdalena, San Jorge and Sinú basins along with the samples from Cispatá bay and the artificial small lake containing a known community (sample 16), 2,695,309 sequences from northern Colombia and 620,828 aditional sequences from the gulf of Urabá were obtained and went through taxonomic assignment resulting in 169 taxa identified. Sixty one of the assigned taxa had a 99% or higher similarity with species reference data and therefore could be assigned up to the aforementioned level. Another 68 taxa could be identified up to the genus level and for the remaining 40, assignment was possible to either family or order (whichever was the lowest possible). Of the 169 taxa, 133 were identified as fish and this group was usually the most abundant taxa in each sample. The remaining 36, belonged to amphibians (4 taxa), birds (16 taxa), mammals (13 taxa), and reptiles (3). Sequencing depth was higher than 10,000 sequences with the exception of the data from samples 25–29 (**Table 1**).

For the remaining 12 samples taken from the Gulf of Urabá and the Atrato river basin (**Figure 2**), results showed 89 taxa detected in 620,828 high quality sequences. The distribution of taxa between main vertebrate groups and between distinct taxonomical categories followed a similar pattern to previous results. Seventy taxa belonged to fish, three to amphibians, six to birds, eight to mammals, and the remaining two were assigned to reptiles. Of these taxa, 38 could be assigned to species level and 29 more to genus level while the remaining 22 belonged to family (12) and order (10). For both sets of samples, human DNA contamination was present and ranged between 1 to 96.45%.

Community analyses were performed with detected genera of fish (**Figure 3A**) and also using only OTUs that could be identified to species and matched with previous reports for its presence to contrast the original obtained data against revised filtered information at the smallest taxonomic level possible (**Figure 3B**). If a detected species did not match any of the current information sources, geographical ranges were checked to decide if it was plausible that it was a new detection (these cases are elaborated further bellow in the discussions) or if it was a misidentification due to genetic similarity to other more plausible species. If this was the case, the detection was only considered up to the genus level. Environmental DNA analysis has been proved to be a reliable source of information for fish communities (Handley et al., 2019; Li et al., 2019; Sales et al., 2019), while other vertebrates detected in this study (i.e., tetrapods) still are mostly occasional detections and therefore are not included in the community analysis. Nonetheless genera and species of tetrapods detected for the sampling locations are also displayed (**Figures 4A,B**). Alpha diversity was calculated using Shannon and Simpsons indexes in vegan package (Oksanen et al., 2019) in order to present them based on eDNA. **Table 2** shows alpha diversity calculated for each of the 37 samples. After testing normality for the samples, beta diversity analysis was calculated using Bray-Curtis dissimilarity as seen in **Figure 5**. Diversity analysis showed some significant differences at the alpha level (**Figures 6A,B**). Significant differences were found in both diversity indexes between the Paredes swamp and three other locations: The Canal del Dique ($p = 0.027$), the Marriaga Swamp ($p = 0.029$) and the Suriqui river ($p = 0.029$). Bray Curtis dissimilarity pointed to the highest difference between saltwater and freshwater locations, leaving the Cispata bay (location 9) and the Rio Negro cove (location 14) in a separate branch to the remaining sampling locations, even if they were geographically closer (**Figures 1**, **2**). The Battalion sample (location 7) was also highly different to other locations and on the other extreme, the San Juan river and the Chucurí swamp were the most similar locations despite of the level of taxa used (**Figure 5**).

DISCUSSION

Many eDNA studies are coupled with traditional survey techniques since there are still some doubts regarding the usefulness and detection capacity of this technique, and to the fact that false negatives are possible (Pinfield et al., 2019). Still, eDNA as a cheap and efficient alternative for classic diversity census must be explored. Some studies are beginning to only work with filter information (Hunter et al., 2015; Bakker et al., 2017; Pinfield et al., 2019). In this study a small, yet relevant (since it's the one of the first times it is done) number of eDNA samples were taken in several water bodies of the northern Colombia. As expected, most of the results were from fish taxa (Jeunen et al., 2020). The other vertebrate groups showed also in smaller numbers.

Comparisons of the data generated in this study against available data for these sampling regions (Aguilera, 2006;

TABLE 1 | Water and DNA collection results: 37 samples of water were collected northern Colombia and filtered in order to extract the DNA and asses the quality of the sample and to correlate it with Taxa detected (**Figure 6**).

Sample	Total vol (ml)	Filtered vol (ml)	Detected taxa	DNA (ng/ul)	# of sequences	# of OTUS
1	6,000	260	6	2.44	23,642	22
2	6,000	170	4	5.74	110,510	23
3	6,000	176	4	10.6	43,603	21
4	6,000	342	10	6.86	86,896	37
5	6,000	460	8	10.8	70,203	33
6	6,000	372	7	10.8	75,793	28
7	6,000	413	0	5.64	49,978	13
8	6,000	482	6	8.22	23,993	24
9	6,000	454	3	5.66	25,499	13
10	2,400	337	2	5.3	77,223	16
11	6,000	952	10	6.42	50,820	25
12	6,000	1,520	12	6.48	49,958	35
13	6,000	233	5	0.578	20,595	24
14	6,000	362	8	3.72	33,834	26
15	6,000	1,261	10	8.52	78,455	37
16	6,000	517	7	3.22	28,812	20
17	6,000	274	18	2.06	43,912	53
18	6,000	444	11	2.96	83,032	32
19	6,000	475	6	3.66	91,230	23
20	6,000	1,980	23	20	35,105	57
21	6,000	1,382	18	20	51,714	31
22	7,000	996	6	20	83,350	13
23	6,000	406	4	4.42	37,845	24
24	6,000	335	5	10.2	68,133	26
25	6,000	259	6	2.68	7,728	20
26	6,000	397	8	0.476	5,750	20
27	6,000	342	9	1.07	9,017	23
28	6,000	124	2	0.412	5,577	12
29	6,000	660	9	20.6	3,932	17
30	6,000	438	14	18.2	96,372	40
31	6,000	507	9	12	86,822	35
32	6,000	643	3	27.8	21,884	16
33	6,000	508	5	41	40,456	19
34	6,000	408	9	94.6	74,041	26
35	6,000	183	6	5.92	90,229	18
36	6,000	362	7	86.8	103,026	24
37	6,000	619	6	55.4	83,722	21

Usually 6 L of water were collected however samples 10 and 22 has different values due to special circumstances presented at the moment of sampling.

Maldonado-Ocampo et al., 2006; Mojica et al., 2006; Ríos-Pulgarín et al., 2008; Mojica-Figueroa and Díaz-Olarte, 2016; Arango-Sánchez et al., 2019) showed some degree of correlation between available information from traditional sampling techniques and information obtained from eDNA (**Table 3**). At the genus level, around 60% of the recovered fish genera in the filters matched available information sources and a quarter of the species as well. It is worth mentioning that with the exception of the two swamps (Paredes and Ayapel), the information used to compare with the filters is not exactly of the designated area but rather the smallest range possible that includes the places sampled. In many cases detailed and updated diversity studies for these locations are missing, since long term field studies were not possible due to the internal conflict in the last decades and therefore it should not be seen as a negative result but rather the first on which to build further data obtained using this method. The initially high differences contrasts with studies comparing traditional sampling and eDNA filters, where the species recovered with eDNA were close to be the same amount (or even higher) that normal sampling methods found for groups like fishes, corals and soil eDNA (Drummond et al., 2015; Handley et al., 2019; Nichols and Marko, 2019). In most

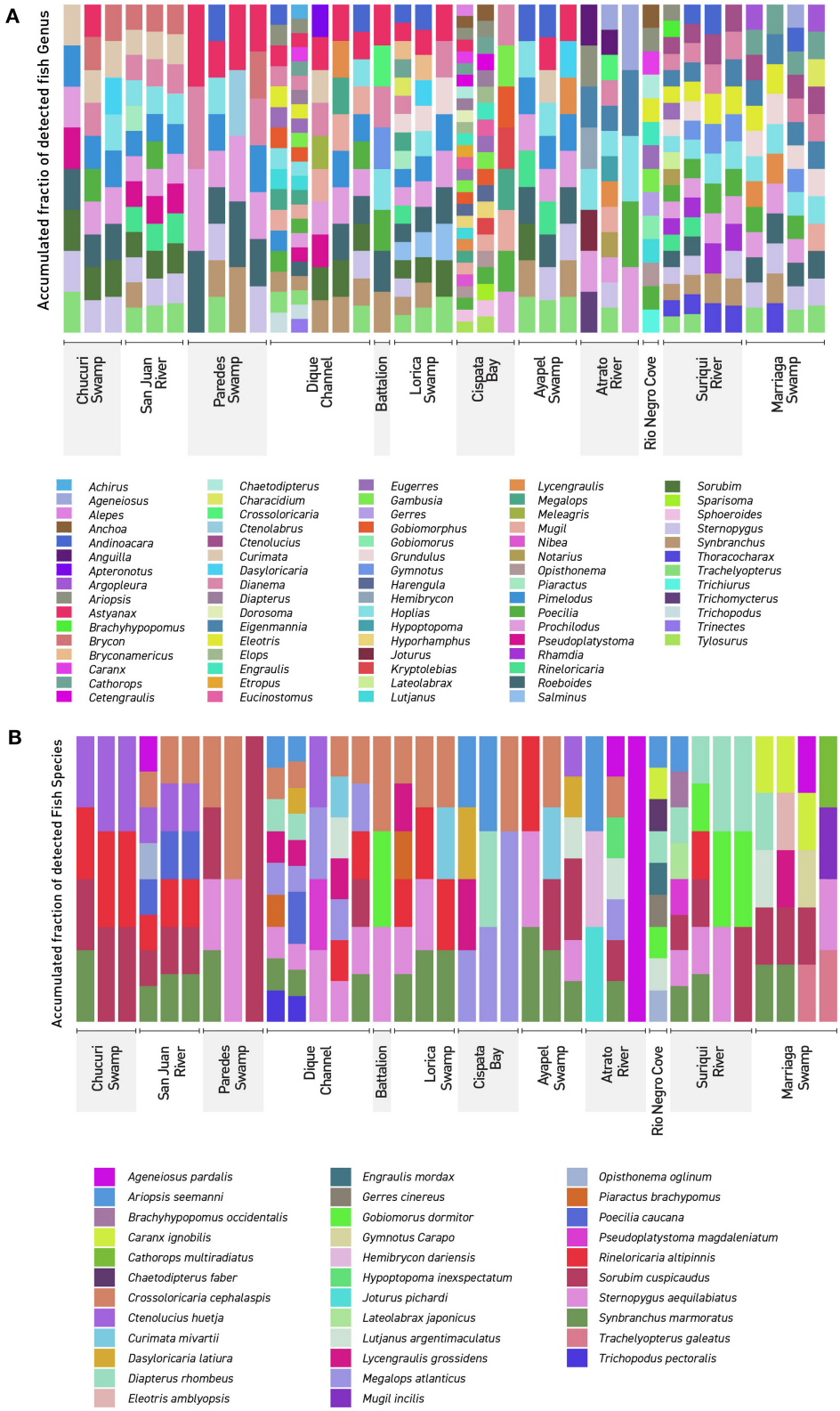

FIGURE 3 | Detected Northern Colombia fish communities: The figure shows every detected fish genera and species using eDNA metabarcoding. Colors don't represent similar lineages or taxa but rather are there to clearly differentiate. **(A)** Genera detected in the 37 filters used in this work. **(B)** Species detected in the 37 filters used in this work. Dique Cannel comprises sampling locations 4, 5, and 6.

FIGURE 4 | Detected Tetrapods in Northern Colombia: The figure shows every detected tetrapod **(A)** genera in the 37 filters used in this work. **(B)** Detected tetrapod species in the 37 filters collected in this work. The names showcased correspond to the sampling locations seen in (**Figures 1, 2**). Dique Channel comprises sampling locations 4, 5, and 6.

TABLE 2 | Alpha diversity indexes of Shannon and Simpson for each sampling location. Indexes are based on detected fish genera.

Samples	Water body	Shannon	Simpson
1–3	Chucuri Swamp	2.193 ± 0.111	0.887 ± 0.012
4–6	San Juan River	2.482 ± 0.083	0.916 ± 0.007
7–10	Paredes Swamp	1.784 ± 0.358	0.824 ± 0.061
11–15	Canal del Dique	2.569 ± 0.364	0.919 ± 0.026
17–19	Lorica Swamp	2.550 ± 0.346	0.918 ± 0.028
20–22	Cispata Bay	2.802 ± 0.648	0.929 ± 0.047
23–25	Ayapel Swamp	2.232 ± 0.060	0.892 ± 0.006
26–28	Atrato River	2.084 ± 0.477	0.866 ± 0.062
30–33	Suriqui River	2.677 ± 0.283	0.929 ± 0.019
34–37	Marriaga Swamp	2.521 ± 0.103	0.919 ± 0.008

A Kruskal–Wallis test for both indexes. Both cases showed significant differences ($p = 0.047$ for both cases) Samples 16 and 29 were omitted since one sample is not enough for statistical analysis.

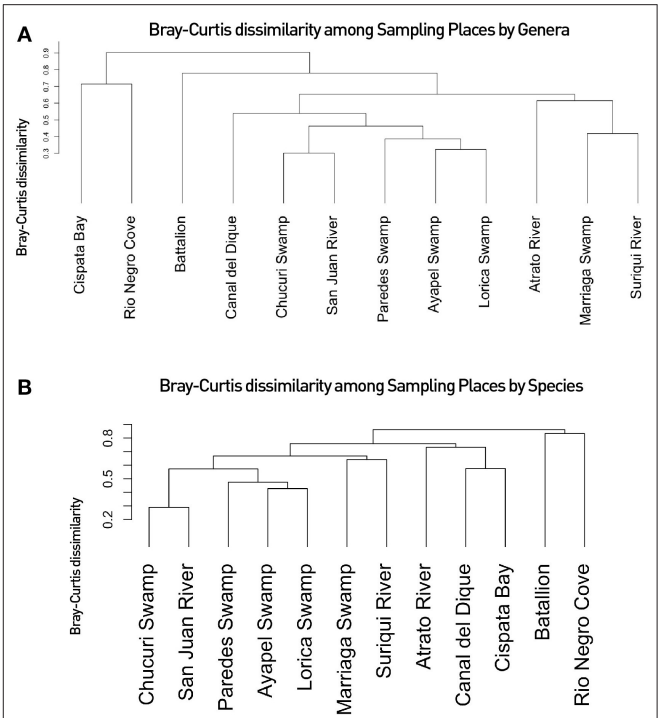

FIGURE 5 | Fish communities of Northern Colombia: The figure shows the beta diversity based on the Bray-Curties dissimilarity Index based on: **(A)** Confirmed detected fish genera for all 37 filters used in this work. **(B)** Confirmed detected fish species per filter. The names showcased correspond to the sampling locations seen in (**Figures 1, 2**). Confirmed detected genera/species indicate that the taxa has been detected both in eDNA filters used for this study and are registered in literature or may be based on habitat ranges. Dique Channel comprises sampling locations 4, 5, and 6.

of these studies, multiple gene primers were designed and tested and or the communities were much smaller in question like in Handley et al. (2019) where the fish community consisted of a total of 16 species where the only two undetected species were lampreys and later the authors explained that these were not detectable through the assay they were using.

Several reasons may explain this discrepancy between datasets. As mentioned before, the fact that current information is not specific for the studied areas in most cases, but instead covers larger areas along these basins. Other studies also have encountered problems to detect or assign sequences to species due to issues such as the aforementioned lack of genetic information but also others such as the current sequence and/or specimen being classified to other species. Also there may be a lack of enough genetic variation for the 12S region to separate species (Cilleros et al., 2019; Sales et al., 2021). The most usual solutions to this problems include the use of more than one primer set so that more species can be recovered in the case that some groups are either too genetically similar or do not work well with one primer set (Polanco Fernández et al., 2020; Sales et al., 2020b) or complementing it with other sampling techniques (Cilleros et al., 2019). These solutions however raise costs. Environmental DNA at the scale used in this study can be a useful initial tool for "snapshotting" communities and regions and once initial results are analyzed, further and deeper analysis can be done focusing on specific groups where the 12S primer fails to differentiate at a deeper more desired level, or coupling it with net fishing, electrofishing, toxicants, or trap cameras (Cilleros et al., 2019; Sales et al., 2020b).

The small volumes of filtered water could explain in part of the lack of detection. The total filtered volume varied between 124 and 1,980 ml (**Table 1**) with the mean being at 542 ml. **Figure 7** supports in part this idea, showing that there is a small but significant correlation between filtered volume and total species detected ($R = 0.43$, $p = 0.0081$) and also is in accordance with literature (Leduc et al., 2019). Other studies used vacuum pumps or peristaltic pumps instead of manual pumps or syringes like the one used here, since it would increase the amount of filtered water used (Hunter et al., 2015; Baker et al., 2018; Leduc et al., 2019; Wineland et al., 2019).

False negatives are also a possibility also and have occurred in other studies due to low amounts of target DNA in the water (Pinfield et al., 2019). While this could explain lack of detection for species that move long distances in rivers such as *Trichechus manatus*, for fish in particular is not highly feasible to explain the absence of many species. Besides, the 12S primers used in this study an also other sets have shown to be effective for use in fish (Bylemans et al., 2019; Li et al., 2019; Sales et al., 2019).

Upon further inspection of the data, particularly of species detected by filters but not found in other information sources, some geographical incongruences were detected. Some of the species showed for the Urabá region are distributed solely in the Pacific coast (such as *Engraulis mordax* or *Caranx ignobilis*) even though the whole sampling was made in the Caribbean coast or in rivers that eventually end in the Caribbean Sea. One possibility is that this confusion derives from sister species split after the Isthmus of Panama formed, allowing for allopatric speciation (Rocha et al., 2008; Aguilar et al., 2019) but this must also be treated carefully since as Rocha et al. (2008) points out, many of the speciation events for the genus *Haemulon* occurred after the closure of the Isthmus and so this could also be the case. Of the

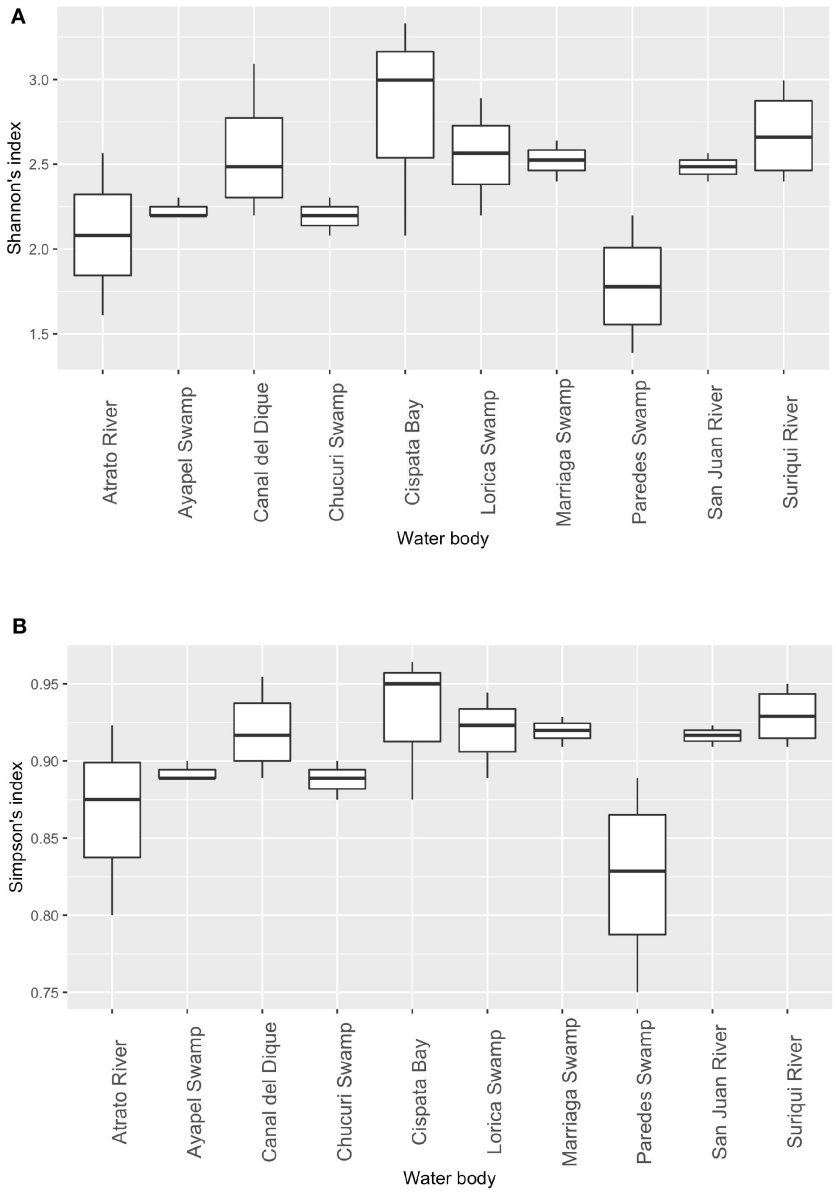

FIGURE 6 | Diversity indexes: Shannon's diversity index between sampling places. Significant differences between sampling locations based on Shannon's diversity index. **(A)** and Simpson's diversity index **(B)**. Statistical differences for both indexes were detected (Wicoxons Rank Sum test) at an alpha of 0.05 were found between Paredes swamp and the following: Canal del Dique, Marriaga Swamp, and Suriqui river ($p = 0.027, 0.029, 0.029$, respectively).

71 fully identified species, 36 did match with bibliography and 35 were out of their distribution range after a final search in GBIF database (GBIF.org, 2020).

Diversity analyses showed some promising results. In **Figure 5**, water bodies should group according to the basin they belong to. Results show that all basin samples were grouped in one clade separated from the saltwater samples and the Battalion sample. Inside the branch of the basins the Atrato samples were separated from the other basins. The Lorica swamp, the Ayapel Swamp and the Paredes swamp were together in another clade inside the basins clade. Certainly these places share many species leaving the possibility of similarity high in the charts (Aguilera, 2006; Ríos-Pulgarín et al., 2008; Lasso et al., 2011; Mojica-Figueroa and Díaz-Olarte, 2016). If based on species detection data, Bray's dissimilarity showed some different patterns (**Figure 5B**). The Chucurí swamp and the San Juan river are still together as well as the Ayapel, Paredes and Lorica swamps but now all the previously mentioned places are the sister branch to the Marriaga swamp and Suriquí river instead of the Canal del Dique, which now is in the same clade as the Atrato river

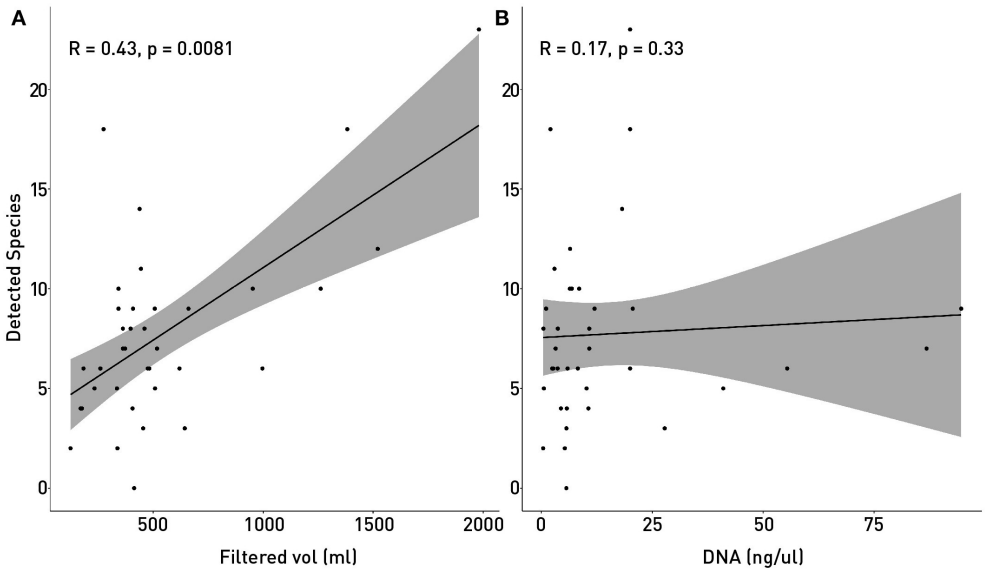

FIGURE 7 | Correlation of detected species with sampling and processing variables: The number of detected and identified species was tested with Spearmans correlation coefficient in order to determine if the small volume of filtered water **(A)** or the amount of recovered DNA **(B)**. In **(A)** *P*-value showed that the correlation between Filtered volume of water and the number of detected species is significant.

TABLE 3 | Comparison between filter obtained information and available information: 5 places had available information to compare with filter data although filter data had to be joined at times to make a better analysis since not every dataset vas specific for the sampled region in this work.

	Middle magdalena basin	Paredes swamp	Canal del dique	Ayapel swamp	Atrato river basin	Average
Known genus	78	24	26	36	66	–
Detected genus	19	12	30	15	38	–
Known Species	128	28	30	38	140	–
Detected species	16	4	16	8	27	–
Shared Genus	21	8	10	12	18	–
Shared Species	11	1	9	5	9	–
Detected confirmed species	69%	25%	56%	63%	37%	50 ± 26%
Detected confirmed genus	48%	66%	36%	85%	48%	56 ± 19%

Middle Magdalena comprises the filters 1 through 10 (Paredes swamp also showed separately since data for this location was available). Canal del Dique species include samples 11–15, Ayapel swamp samples are 23–25 and Atrato river basin used samples 26–37. Samples 16–22 were not used since no information from traditional monitoring on a desired scale was found to compare for comparisons. The known genus and species values were extracted from literature and the detected genus and species values were based on the taxa identified via eDNA metabarcoding from the water samples used in this study. Finally, the shared genus and species values represent the number of taxa of each kind which were found in both literature and filter data. The last two rows indicate the percentage of shcared taxa regarding the detected one to better illustrate the capacities of the eDNA metabarcoding process.

and the saltwater samples of the Cispatá bay. The Gulf of Urabá was also paired with the Batallion lake this time. **Figure 8** is a Venn diagram showing fish genera shared among the four basins (Atrato, Sinú, San Jorge and Magdalena) where it is seen that the Sinu and San Jorge river basins have no unique genera or genera that aren't shared with the Magdalena basin according to data available on GBIF (Herrera-Collazos et al., 2018) and therefore are grouped together with the Paredes swamp (the Lorica swamp and the Ayapel swamp, respectively, represent these basins) which supports their position in the dendrograms.

Many challenges still lay ahead related to obtaining consistent results using this technique. There are not many reference genomes or even gene sequences available for many of the species that inhabit the sampled waters. Projects such as the Earth BioGenome Project (Lewin et al., 2018) or Vertebrate Genomes Project are still only beginning their second phase of work focusing on higher taxa rather than on species leading many organisms still without a decent genomic frame to compare with and also most of the species in these projects are distributed in temperate areas rather than in tropical regions. Alpha diversity can greatly influence beta diversity analysis even if it shouldn't (Jost, 2007) and rare species can have a high impact in diversity assessments (Fontana et al., 2008).

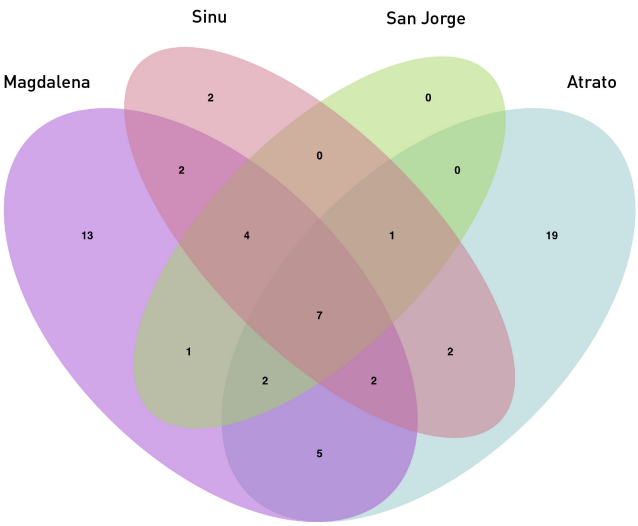

FIGURE 8 | Shared genera among Basins: The Venn diagram shows the shared fish genera between the four basins samples in this study. Magdalena Basin comprises samples 1 through 15 (San Juan River, Chucuri swamp, Paredes swamp, and the Canal del Dique), the Sinú basin is represented by samples 17–19 (Lorica swamp), the San Jorge basin is represented by samples 23–25 (Ayapel swamp) and the Atrato basin comprises samples 26–28 and 30–37 (Atrato river, Suriqui river, and Marriaga swamp).

Threatened and endangered species were detected in several places. The most relevant results include the detection of the endangered "Bagre rayado" *Pseudoplatystoma magdaleniatum* in samples belonging to the Chucuri swamp, San Juan river and Canal del Dique (1, 4–6, and 12 and 13) matching literature (Mojica et al., 2016) together with other six vulnerable fish species (*Curivata mivartii*, *Megalops atlanticus, Ageneiosus pardalis, Sorubim cuspicaudus, Mugil liza,* and *Mugil incilis*, the Antillean manatee, which is considered vulnerable (Self-Sullivan and Mignucci-Giannoni, 2008) and the endangered brown-headed spider monkey *Ateles fusciceps* from the Suriquí river (Samples 30–32) and the Marriaga swamp (Sample 37) (**Figure 4B**). The Antillean manatee *Trichechus manatus* was found in a total of six samples including the Battalion sample, designated as a positive control for *T. manatus*. Its presence was detected in samples 12, 14, 15, 16, 18, and 26 (**Figures 1**, **2**), respectively, belonging to the swamps around the Canal del Dique (an artificial deviation of the natural course of the Magdalena river (samples 12, 14, 15, 16), the Lorica swamp (Sinú basin) and one of the mouths of the Atrato river. While literature and local fishermen and boat drivers report the presence of the animal in all places where samples were taken, only these six spots captured DNA belonging to the species. On a side note, visual detection of the animal was made while collecting samples 22, 33, and 35 (Cispatá bay, Suriquí river, and Marriaga swamp), however none of these samples reported positive results, since most likely either the animals arrived recently to the area or in low numbers, resulting in non-significant amounts of DNA being shed into the water.

Some species detections were interesting (see Appendices 1, 2 in **Supplementary Material**) for complete list of species detected). For samples 26 and 27, taken in the Atrato river mouth, the American eel, *Anguilla rostrata* was detected. This species was not detected in the Gulf of Urabá even when its presence should have been detected based on their distribution range and known habitats in the Caribbean and in Colombia (Benchetrit and McCleave, 2015; Arango-Sánchez et al., 2019). Another interesting detection was a match for *Lateolabrax japonicus* (Japanese sea bass), one of three species from the genus *Lateolabrax*, all belonging to the western side of the western Pacific Ocean and all had their complete mitochondrial genome sequenced (Shan et al., 2016). No close relative(at least at the genus level) can be used to explain this match and the lateolabraciade family is placed as the sister branch of the acropomatidae family where perhaps a possible candidate for confusion may be found (Betancur et al., 2017).

Sample 16 was a particular case also since it was an "unofficial positive control." Upon arrival at the place, only the Antillean manatee (*Trichechus manatus*) was supposed to be at the place besides some common fish for the area: *Ctenolucius huetja*, *Synbrancus marmoratus* and *Gymnotus carapo* which is not listed for the area is likely to be *Gymnotus ardilai* based on registers (Mojica et al., 2006). The sample also showed positive results for the spectacled caiman (*Caiman crocodylus*), a turtle assigned as *Trachemys scripta* although most likely *Trachemys callirostris* (Galvis-Rizo et al., 2016) and for the largest rodent, the capybara (*Hydrochoerus hydrochoaeris*). The reason these results are particularly interesting, is because this is an enclosed artificial lake of the battalion. The two most likely explanations as to how the detections appeared are: (1) perhaps the most likely is that all three species live in nearby water bodies that occasionally feed the lake, and their DNA traveled with the current to the lake. This could help to better understand the flow of eDNA through current and how far can it travel if the position of the creatures in relation to the lake is more precisely determined. Studies support transportation of eDNA in short distances (Li et al., 2019; Wacker et al., 2019) and studying the transport of eDNA in small areas such as this could help to further develop this technique and its uses in open uncontrolled environments. The other possibility (2) is of course that these species recently were in the lake but were not seen, and it was thanks to eDNA that they could be detected.

Invasive vertebrate species for Colombia (Ministerio De Ambiente Territorial Vivienda Y Desarrollo, 2008; Ministerio De Ambiente Vivienda Y Desarrollo, 2010) were surprisingly not detected. Common invasive fishes such as the Nile Tilapia and the Mozambique Tilapia were not detected in the samples of this study, Cichlids were however detected although not identified (Appendix 2 in **Supplementary Material**). Additionally, in samples taken for another project in Colombia (Caballero, Personal communication) they have been also been identified. Tilapia species were initially introduced but rapidly expanded their range beyond planned and became invasive (Dirección de Recursos Naturales, 2017). It is unclear as to how they were not detected since they are reported for most of Colombia. Very low

numbers or highly degraded DNA are perhaps the only possible explanations since the detection of these fish species has been proven to be possible and yield good results (Keskin, 2014).

The detection of many not aquatic species was a surprise and not many studies of eDNA have included terrestrial species (Drummond et al., 2015; Ishige et al., 2017; Johnson et al., 2019) even with aquatic eDNA (Ushio et al., 2017; Williams et al., 2018; Seeber et al., 2019; Sales et al., 2020a,b). This study, however, presents evidence from very open sampling locations, unlike the ponds or waterholes with high eDNA concentrations mentioned by Ushio or Seeber who even went further into using DNA hybridization techniques in order to recover increased amounts of mammal eDNA. The fact that endangered species such as *Ateles fusciceps* or the southern tamandua (*Tamandua tetradactyla* only identified to genus and therefore not included in the main results, see Appendix 2 in **Supplementary Material**) shows that water samples could be used to monitor threatened or rare mammals. Coupled with habitat prediction computer programs it could help improve the determination of previously unknown habitat ranges for some species, like it has been made with the Yamato salamander in Japan (Sakai et al., 2019). Many of the most recognizable groups of terrestrial mammals were detected (see Appendices 1, 2 in **Supplementary Material**). However, as pointed in Seeber et al. (2019), rarer species may have lower representation in samples, due to low quality sequences than are filtered and eliminated and therefore not included in further analysis, or in such low amounts that is impossible to determine even family level, which may be the case for the order Chiroptera that appeared in very small quantities (see Appendix 3 in **Supplementary Material**) Both studies from Sales indicate that eDNA is very capable of detecting mammals, specially herbivores. Of these two studies one was performed in south America and identified 15 different mammal families including some bats to the species level. Primer selection in this study was a clear difference with both Sales studies were mammal primers were used unlike the universal vertebrate primers used here This would explain some of the differences in the identification to the species level. The Sales study performed in England showed confident data on the detection of at least three mammal species (water vole, filed vole and red deer) using just four water samples per location. While the number of samples might be close or equal for both studies, it has also been mentioned that conditions on tropical waters are different to those in the lakes and ponds of temperate regions, likely affecting the integrity of DNA. Fifteen bird species were identified in this study (Appendices 1, 2 in **Supplementary Material**). Bird eDNA showed frequently also and most likely derived from fecal matter (Bohmann et al., 2014) for species like *Ramphastos swainsoni* or *Ara ararauraana* that are not considered aquatic species. A migrant bird (*Catharus ustulatus*) was found among the data collected in the Atrato river (Sample 26). This suggests that the presence of migrant birds might be monitored via eDNA, however not much has been done to date to use eDNA in monitoring bird species. Studies focused on birds have not been published extensively, with the exception from of preliminary tests in small scale environments (Ushio et al., 2018) or by exploring other types of eDNA such as saliva in fruits or soil eDNA (Drummond et al., 2015; Monge et al., 2020). Since many species of migrant birds are attracted to waters, aquatic eDNA could be used in the future to monitor them as well.

CONCLUSIONS

As the whole country becomes easier to access, more detailed biodiversity sampling will be a possibility. The advantage of eDNA metabarcoding relies on its simplicity to deploy to the point that communities can work along scientists to generate valid results (Sakai et al., 2019). Communities were close to all sampling places and it has been a long time since the relevance of local communities in conservations efforts was noted (Wells and Brandon, 1993) and many successful examples exist such as The California environmental DNA "CALeDNA" program (Meyer et al., 2019) that already is working with a well-established network to allow both scientists and volunteers to provide samples from project associated or random places in the California state and could even enter the Earth BioGenome Project (Lewin et al., 2018). Environmental metabarcoding sampling in this work showed that there are still aspects to work on to improve the application of this technique, but the amount of information recovered from <3l of water per sampling place showed the great potential for this monitoring technique for to further biodiversity studies in Colombia.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found at: DRYAD Repository (https://datadryad.org/stash/share/Qtb2EFuCHoZdUSrMnKu6m2FeWJ01x1_mL0IO96mkauA).

ETHICS STATEMENT

Ethical review and approval was not required for the animal study because all information and data was obtained through the collection of water samples and environmental DNA in it. There was no contact with animals during the whole process of sampling. Since no animals were collected, handled or harmed, no ethical review was necessary.

AUTHOR CONTRIBUTIONS

JL: field work, sample processing, data analysis, statistical analysis, manuscript writing, and editing. SC: project conceptualization, field work, sample processing, manuscript writing, and editing.

FUNDING

Funding for this project was available from a private donation to Universidad de los Andes from Programa de Investigacion initiative from Facultad de Ciencias, Universidad de los Andes

[Project Name: Conservación del Manatí Antillano (Trichechus manatus) en Colombia y el Caribe uso de nuevas tecnologías como apoyo efectivo en procesos de recuperación de especies amenazadas].

ACKNOWLEDGMENTS

The authors would like to thank D. A. Quiroga, M. Luna, and C. Rosso for their help with field work, Fundación Omacha for logistics support in Lorica, Cispatá, and Canal del Dique sampling and the Cabildo Verde de Sabana de Torres organization, as well as to Don Mora, Seferino, and the people of the towns of Bocas del Carare and Marriaga for their hospitality during the sampling process.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fevo.2020.617948/full#supplementary-material

REFERENCES

Aguilar, C., Miller, M. J., Loaiza, J. R., González, R., Krahe, R., and De León, L. F. (2019). Tempo and mode of allopatric divergence in the weakly electric fish Sternopygus dariensis in the Isthmus of Panama. *Sci. Rep.* 9:18828. doi: 10.1038/s41598-019-55336-y

Aguilera, M. M. (2006). *El Canal Del Dique Y Su Subregion Una Economía Basada En La Riqueza Hidrica.* Colombia: Cartagena de Indias.

Arango-Sánchez, L. B., Correa-Herrera, T., and Correa-Rendón, J. D. (2019). Diversidad de peces en hábitats estuarinos delta del río atrato, golfo de urabá. *Bol. Cient. Mus. Hist. Nat. Univ. Caldas. 23,* 191–207. doi: 10.17151/bccm.2019.23.1.7

Arbeláez-Cortés, E. (2013). Knowledge of Colombian biodiversity: Published and indexed. *Biodivers. Conserv.* 22, 2875–2906. doi: 10.1007/s10531-013-0560-y

Ayala López, L., Murcia, L. M., and Barriga, J. (2018). "Expediciones científicas nacionales," in *Biodiversidad 2017*. Available online at: http://reporte.humboldt.org.co/biodiversidad/2017/cap1/104/index.html#seccion10 (accessed April 23, 2020).

Baker, S., Steel, D., Nieukirk, S., and Klink, H. (2018). Environmental DNA (eDNA) from the wake of the whales: droplet digital PCR for detection and species identification. *Front. Mar. Sci.* 1:133. doi: 10.3389/fmars.2018.00133

Bakker, J., Wangensteen, O. S., Chapman, D. D., Boussarie, G., Buddo, D., Guttridge, T. L., et al. (2017). Environmental DNA reveals tropical shark diversity in contrasting levels of anthropogenic impact. *Sci. Rep.* 7:16886. doi: 10.1038/s41598-017-17150-2

Benchetrit, J., and McCleave, J. D. (2015). Current and historical distribution of the American eel *Anguilla rostrata* in the countries and territories of the wider Caribbean. *ICES J. Mar. Sci.* 73, 122–134. doi: 10.1093/icesjms/fsv064

Betancur, R. R., Wiley, E. O., Arratia, G., Acero, A., Bailly, N., Miya, M., et al. (2017). Phylogenetic classification of bony fishes. *BMC Evol. Biol.* 17:162. doi: 10.1186/s12862-017-0958-3

Bohmann, K., Evans, A., Gilbert, M. T. P., Carvalho, G. R., Creer, S., Knapp, M., et al. (2014). Environmental DNA for wildlife biology and biodiversity monitoring. *Trends Ecol. Evol.* 29, 358–367. doi: 10.1016/j.tree.2014.04.003

Bylemans, 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

Castelblanco-Martínez, D. N., dos Reis, V., and de Thoisy, B. (2018). How to detect an elusive aquatic mammal in complex environments? A study of the endangered *Antillean manatee Trichechus manatus manatus* in French guiana. *Oryx* 52, 382–392. doi: 10.1017/S0030605316000922

Cilleros, K., Valentini, A., Allard, L., Dejean, T., Etienne, R., Grenouillet, G., et al. (2019). Unlocking biodiversity and conservation studies in high-diversity environments using environmental DNA (eDNA): a test with Guianese freshwater fishes. *Mol. Ecol. Resour.* 19, 27–46. doi: 10.1111/1755-0998.12900

Dirección de Recursos Naturales (2017). *Plan de Prevención, Control Y Manejo De la Tilapia Del Nilo (Oreochromis niloticus) En La Jurisdicción Car Cundinamarca. Plan De Prevención, Control Y Manejo De La Tilapia Del Nilo (Oreochromis niloticus) En La*. Available online at: https://www.car.gov.co/uploads/files/5b90332505307.pdf (accessed September 25, 2020).

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

Edgar, R. C. (2010). Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* 26, 2460–2461. doi: 10.1093/bioinformatics/btq461

Edgar, R. C. (2016). UNOISE2: improved error-correction for Illumina 16S and ITS amplicon sequencing. *BioRxiv [Preprint]* 081257. doi: 10.1101/081257

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

Fontana, G., Ugland, K. I., Gray, J. S., Willis, T. J., and Abbiati, M. (2008). Influence of rare species on beta diversity estimates in marine benthic assemblages. *J. Exp. Mar. Biol. Ecol.* 366, 104–108. doi: 10.1016/j.jembe.2008.07.014

Frøslev, T. G., Kjøller, R., Bruun, H. H., R., E., Brunbjerg, A. K., et al. (2017). Algorithm for post-clustering curation of DNA amplicon data yields reliable biodiversity estimates. *Nat. Commun.* 8:1188. doi: 10.1038/s41467-017-01312-x

Galvis-Rizo, C., Carvajal-Cogollo, J. E., Arredondo, J. C., Passos, P., López-Victoria, M., Velasco, J. A., et al. (2016). *Libro Rojo de Reptiles de Colombia 2015*. Bogotá. Available online at: http://repository.humboldt.org.co/handle/20.500.11761/9303 (accessed August 12, 2020).

GBIF.org (2020). *Free and Open Access to Biodiversity Data.* GBIF. Available online at: https://www.gbif.org/ (accessed August 8, 2020).

Habel, J. C., Rasche, L., Schneider, U. A., Engler, J. O., Schmid, E., Rödder, D., et al. (2019). Final countdown for biodiversity hotspots. *Conserv. Lett.* 12:e12668. doi: 10.1111/conl.12668

Handley, L. L., Read, D. S., Winfield, I. J., Kimbell, H., Johnson, H., Li, J., et al. (2019). Temporal and spatial variation in distribution of fish environmental DNA in England's largest lake. *Environ. DNA* 1, 26–39. doi: 10.1002/edn3.5

Herrera-Collazos, H.-C., Herrera,_R., G., DoNascimiento, C., and Maldonado-Ocampo, J. A. (2018). *Lista de Especies De Peces De Agua Dulce De Colombia/Checklist of the Freshwater Fishes of Colombia. v2.10.* Bogotá: Asociation Colombia Ictiólogos.

Hunter, M. E., Meigs-Friend, G., Ferrante, J. A., Takoukam Kamla, A., Dorazio, R. M., Keith-Diagne, L., et al. (2018). Surveys of environmental DNA (eDNA): a new approach to estimate occurrence in vulnerable manatee populations. *Endanger. Species Res.* 35, 101–111. doi: 10.3354/esr00880

Hunter, M. E., Oyler-McCance, S. J., Dorazio, R. M., Fike, J. A., Smith, B. J., Hunter, C. T., et al. (2015). Environmental DNA (eDNA) sampling improves occurrence and detection estimates of invasive burmese pythons. *PLoS ONE* 10:e0121655. doi: 10.1371/journal.pone.0121655

Ishige, T., Miya, M., Ushio, M., Sado, T., Ushioda, M., Maebashi, K., et al. (2017). Tropical-forest mammals as detected by environmental DNA at natural saltlicks in Borneo. *Biol. Conserv.* 210, 281–285. doi: 10.1016/j.biocon.2017.04.023

Jerde, C. L., Mahon, A. R., Chadderton, W. L., and Lodge, D. M. (2011). "Sight-unseen" detection of rare aquatic species using environmental DNA. *Conserv. Lett.* 4, 150–157. doi: 10.1111/j.1755-263X.2010.00158.x

Jeunen, G., Lamare, M. D., Knapp, M., Spencer, H. G., Taylor, H. R., Stat, M., et al. (2020). Water stratification in the marine biome restricts vertical environmental DNA (eDNA) signal dispersal. *Environ. DNA* 2, 99–111. doi: 10.1002/edn3.49

Johnson, M. D., Cox, R. D., and Barnes, M. A. (2019). Analyzing airborne environmental DNA: A comparison of extraction methods, primer type,

and trap type on the ability to detect airborne eDNA from terrestrial plant communities. *Environ. DNA* 1, 176–185. doi: 10.1002/edn3.19

Jost, L. (2007). Partitioning diversity into independent alpha and beta components. *Ecology* 88, 2427–2439. doi: 10.1890/06-1736.1

Keskin, E. (2014). Detection of invasive freshwater fish species using environmental DNA survey. *Biochem. Syst. Ecol.* 56, 68–74. doi: 10.1016/j.bse.2014.05.003

Lasso, C. A., de Paula Gutiérrez, F., Morales-Betancourt, M. A., Agudelo, E., Ramírez -Gil, H., and Ajiaco-Martínez, R. E. (2011). *Pesquerías Con- Tinentales De Colombia: Cuencas Del Magdalena-Cauca, Sinú, Canalete, Atrato, Orinoco, Amazonas y Vertiente del Pacífico.* Bogotá: Instituto de In- vestigación de los Recursos Biológicos Alexander von Humboldt.

Leathwick, J., Moilanen, A., Francis, M., Elith, J., Taylor, P., Julian, K., et al. (2008). Novel methods for the design and evaluation of marine protected areas in offshore waters. *Conserv. Lett.* 1, 91–102. doi: 10.1111/j.1755-263X.2008.00012.x

Leduc, N., Lacoursière-Roussel, A., Howland, K. L., Archambault, P., Sevellec, M., Normandeau, E., et al. (2019). Comparing eDNA metabarcoding and species collection for documenting Arctic metazoan biodiversity. *Environ. DNA* 1, 342–358. doi: 10.1002/edn3.35

Lewin, H. A., Robinson, G. E., Kress, W. J., Baker, W. J., Coddington, J., Crandall, K. A., et al. (2018). Earth BioGenome Project: Sequencing life for the future of life. *Proc. Natl. Acad. Sci. U.S.A.* 115, 4325–4333. doi: 10.1073/pnas.1720115115

Li, J., Lawson Handley, L. J., Harper, L. R., Brys, R., Watson, H. V., Di Muri, C., et al. (2019). Limited dispersion and quick degradation of environmental DNA in fish ponds inferred by metabarcoding. *Environ. DNA* 1, 238–250. doi: 10.1002/edn3.24

Maldonado-Ocampo, J. A., Antonio Villa-Navarro, F., Ortega-Lara, A., Prada-Pedreros, S., Jaramillo Villa, U., Claro, A., et al. (2006). *Peces del río Atrato, Zona Hidrogeografica Del Caribe, Colombia.* Available online at: www.fishbase.org (accessed October 29, 2019).

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. J.* 17, 10–12. doi: 10.14806/ej.17.1.200

Meyer, R. S., Curd, E. E., Schweizer, T., Gold, Z., Ramos, D. R., Shirazi, S., et al. (2019). The California Environmental DNA "CALeDNA" program. *bioRxiv*. 503383. doi: 10.1101/503383

Ministerio De Ambiente Vivienda Y Desarrollo (2010). *Resolución Número 0207 3 de Febrero de 2010.* Available online at: http://www.minambiente.gov.co/images/BosquesBiodiversidadyServiciosEcosistemicos/pdf/Politicas-de-conservacion-de-la-Biodiversidad/res_0207_030210.pdf (accessed April 20, 2020).

Ministerio De Ambiente and Territorial Vivienda Y Desarrollo (2008). *Resolución Número 0848 del 23 de Mayo de 2008.* Available online at: http://www.parquesnacionales.gov.co/portal/wp-content/uploads/2013/08/res_0848.pdf (accessed April 20, 2020).

Mojica, J., Valderrama, M., Jimenez-Segura, L., and Alonso, J. C. (2016). *Pseudoplatystoma magdaleniatum (Bagre rayado).* Available online at: https://www.iucnredlist.org/species/58439165/61474168#bibliography (accessed July 9, 2020).

Mojica, J. I., Galvis, G., Sánchez-Duarte, P., Castellanos, C., and Villa-Navarro, F. A. (2006). Peces del valle medio del río Magdalena, Colombia. *Biota Colomb.* 7, 23–38.

Mojica-Figueroa, B. H., and Díaz-Olarte, J. J. (2016). Comunidad de peces de la ciénaga de Paredes, Magdalena Medio, Santander (Colombia) y su asociación con variables espacio-temporales y ambientales. *Biota Colomb.* 16, 27–43. doi: 10.21068/c2016s01a02

Monge, O., Dumas, D., and Baus, I. (2020). Environmental DNA from avian residual saliva in fruits and its potential uses in population genetics. *Conserv. Genet. Resour.* 12, 131–139. doi: 10.1007/s12686-018-1074-4

Myers, N., Mittermeler, R. A., Mittermeler, C. G., Da Fonseca, G. A. B., and Kent, J. (2000). Biodiversity hotspots for conservation priorities. *Nature* 403, 853–858. doi: 10.1038/35002501

Nichols, P. K., and Marko, P. B. (2019). Rapid assessment of coral cover from environmental DNA in Hawai'i. *Environ. DNA* 1, 40–53. doi: 10.1002/edn3.8

Ogram, A., Sayler, G. S., and Barkay, T. (1987). The extraction and purification of microbial DNA from sediments. *J. Microbiol. Methods* 7, 57–66. doi: 10.1016/0167-7012(87)90025-X

Oksanen, J., Blanchet, F., Kindt, R., Legendre, P., Minchin, P., O'Hara, R., et al. (2019). *Vegan: Community Ecology Package.*

Ostberg, C. O., Chase, D. M., Hoy, M. S., Duda, J. J., Hayes, M. C., Jolley, J. C., et al. (2019). Evaluation of environmental DNA surveys for identifying occupancy and spatial distribution of Pacific Lamprey (*Entosphenus tridentatus*) and *Lampetra* spp. in a Washington coast watershed. *Environ. DNA* 1, 131–143. doi: 10.1002/edn3.15

Pearce, J. L., and Boyce, M. S. (2006). Modelling distribution and abundance with presence-only data. *J. Appl. Ecol.* 43, 405–412. doi: 10.1111/j.1365-2664.2005.01112.x

Phalen, D., Slapeta, J., King, J., and Rose, K. (2011). *Development and Validation of a Rapid Field Test to Detect the Chytrid Fungus Batrachochytrium dendrobatidis at a High Specificity and Sensitivity, for Use in Surveys to Determine the Distribution of Chytridiomycosis.* Taronga Conservation Society Australia. Available online at: http://www.environment.gov.au/system/files/resources/d3fb8c1a-1d58-4d7e-bcd0-f9fd63d67bdc/files/chytrid-fungus-field-test.pdf (accessed March 10, 2019).

Pinfield, R., Dillane, E., Runge, A. K. W., Evans, A., Mirimin, L., Niemann, J., et al. (2019). False-negative detections from environmental DNA collected in the presence of large numbers of killer whales (*Orcinus orca*). *Environ. DNA* 1, 316–328. doi: 10.1002/edn3.32

Polanco Fernández, A., Marques, V., Fopp, F., Juhel, J., Borrero-Pérez, G. H., Cheutin, M., et al. (2020). Comparing environmental DNA metabarcoding and underwater visual census to monitor tropical reef fishes. *Environ. DNA*. 1–15. doi: 10.1002/edn3.140

Riaz, T., Shezad, W., Viari, A., Pompanon, F., Taberlet, P., and Coissac, E. (2011). ecoPrimers: inference of new DNA barcode markers from whole genome sequence analysis. *Nucleic Acids Res.* 39:e145. doi: 10.1093/nar/gkr732

Ríos-Pulgarín, M. I., Jiménez-Segura, L. F., Palacio, J. A., and Ramírez-Restrepo, J. J. (2008). *Comunidad De Peces En La Ciénaga De Ayapel, Río Magdalena (Córdoba) Colombia: Cambios Espacio-Temporales En Su Asociación the Fish Community of the Ayapel Floodplain Lagoon, Magdalena River (Córdoba).* Colombia: Spacio-Temporal Changes in its Assemblage. Available at: https://www.researchgate.net/profile/Maria_Rios-Pulgarin/publication/305279431_COMUNIDAD_DE_PECES_EN_LA_CIENAGA_DE_AYAPEL_RIO_MAGDALENA_CORDOBA_COLOMBIA_CAMBIOS_ESPACIO-TEMPORALES_EN_SU_ASOCIACION_THE_FISH_COMMUNITY_OF_THE_AYAPEL_FLOODPLAIN_LAGOON_MAGDALENA_RIVER_CORDOBA_COLOMBI/links/57865a2108ae3949cf5558a4.pdf (accessed December 2, 2019).

Robinson, C. V., Garcia de Leaniz, C., Rolla, M., and Consuegra, S. (2019). Monitoring the eradication of the highly invasive topmouth gudgeon (*Pseudorasbora parva*) using a novel eDNA assay. *Environ. DNA* 1, 74–85. doi: 10.1002/edn3.12

Rocha, L. A., Lindeman, K. C., Rocha, C. R., and Lessios, H. A. (2008). Historical biogeography and speciation in the reef fish genus *Haemulon* (*Teleostei*: *Haemulidae*). *Mol. Phylogenet. Evol.* 48, 918–928. doi: 10.1016/j.ympev.2008.05.024

RStudio Team (2020). *RStudio: Integrated Development for R.* RStudio Team. Available online at: http://www.rstudio.com/.

Sakai, Y., Kusakabe, A., Tsuchida, K., Tsuzuku, Y., Okada, S., Kitamura, T., et al. (2019). Discovery of an unrecorded population of yamato salamander (*Hynobius vandenburghi*) by GIS and eDNA analysis. *Environ. DNA* 1, 281–289. doi: 10.1002/edn3.31

Sales, N. G., Kaizer, M. C., Coscia, I., Perkins, J. C., Highlands, A., Boubli, J. P., et al. (2020a). Assessing the potential of environmental DNA metabarcoding for monitoring neotropical mammals: a case study in the Amazon and Atlantic Forest, Brazil. *Mamm. Rev.* 50, 221–225. doi: 10.1101/750414

Sales, N. G., McKenzie, M. B., Drake, J., Harper, L. R., Browett, S. S., Coscia, I., et al. (2020b). Fishing for mammals: landscape-level monitoring of terrestrial and semi-aquatic communities using eDNA from riverine systems. *J. Appl. Ecol.* 57, 707–716. doi: 10.1111/1365-2664.13592

Sales, N. G., Wangensteen, O. S., Carvalho, D. C., Deiner, K., Præbel, K., Coscia, I., et al. (2021). Space-time dynamics in monitoring neotropical fish communities using eDNA metabarcoding. *Sci. Total Environ.* 754:142096. doi: 10.1016/j.scitotenv.2020.142096

Sales, N. G., Wangensteen, O. S., Carvalho, D. C., and Mariani, S. (2019). Influence of preservation methods, sample medium and sampling time on eDNA recovery in a neotropical river. *Environ. DNA* 1, 119–130. doi: 10.1002/edn3.14

Seeber, P. A., McEwen, G. K., Löber, U., Förster, D. W., East, M. L., Melzheimer, J., et al. (2019). Terrestrial mammal surveillance using hybridization capture of environmental DNA from African waterholes. *Mol. Ecol. Resour.* 19, 1486–1496. doi: 10.1111/1755-0998.13069

Self-Sullivan, C., and Mignucci-Giannoni, A. (2008). *Caribbean Manatee Trichechus Manatus ssp. Manatus*. Available online at: https://www.iucnredlist.org/species/22105/9359161 (accessed March 4, 2019).

Shan, B., Song, N., Han, Z., Wang, J., Gao, T., and Yokogawa, K. (2016). Complete mitochondrial genomes of three sea basses *Lateolabrax* (*Perciformes*, *Lateolabracidae*) species: genome description and phylogenetic considerations. *Biochem. Syst. Ecol.* 67, 44–52. doi: 10.1016/j.bse.2016.04.007

Sousa, L. L., Silva, S. M., and Xavier, R. (2019). DNA metabarcoding in diet studies: unveiling ecological aspects in aquatic and terrestrial ecosystems. *Environ. DNA* 1, 199–214. doi: 10.1002/edn3.27

Strickler, K. M., Fremier, A. K., and Goldberg, C. S. (2015). Quantifying effects of UV-B, temperature, and pH on eDNA degradation in aquatic microcosms. *Biol. Conserv.* 183, 85–92. doi: 10.1016/j.biocon.2014.11.038

Tsuji, S., Takahara, T., Doi, H., Shibata, N., and Yamanaka, H. (2019). The detection of aquatic macroorganisms using environmental DNA analysis—a review of methods for collection, extraction, and detection. *Environ. DNA* 1, 99–108. doi: 10.1002/edn3.21

Ushio, M., Fukuda, H., Inoue, T., Makoto, K., Kishida, O., Sato, K., et al. (2017). Environmental DNA enables detection of terrestrial mammals from forest pond water. *Mol. Ecol. Resour.* 17, e63–e75. doi: 10.1111/1755-0998.12690

Ushio, M., Murata, K., Sado, T., Nishiumi, I., Takeshita, M., Iwasaki, W., et al. (2018). Demonstration of the potential of environmental DNA as a tool for the detection of avian species. *Sci. Rep.* 8:4493. doi: 10.1038/s41598-018-22817-5

Venter, J. C., Remington, K., Heidelberg, J. F., Halpern, A. L., Rusch, D., Eisen, J. A., et al. (2004). Environmental genome shotgun sequencing of the sargasso sea. *Science* 304, 66–74. doi: 10.1126/science.1093857

Wacker, S., Fossøy, F., Larsen, B. M., Brandsegg, H., Sivertsgård, R., and Karlsson, S. (2019). Downstream transport and seasonal variation in freshwater pearl mussel (*Margaritifera margaritifera*) eDNA concentration. *Environ. DNA* 1, 64–73. doi: 10.1002/edn3.10

Wells, M. P., and Brandon, K. E. (1993). The principles and practice of buffer zones and local participation in biodiversity conservation. *Ambio* 22, 157–162.

Williams, K. E., Huyvaert, K. P., Vercauteren, K. C., Davis, A. J., and Piaggio, A. J. (2018). Detection and persistence of environmental DNA from an invasive, terrestrial mammal. *Ecol. Evol.* 8, 688–695. doi: 10.1002/ece3.3698

Wineland, S. M., Arrick, R. F., Welch, S. M., Pauley, T. K., Mosher, J. J., Apodaca, J. J., et al. (2019). Environmental DNA improves Eastern hellbender (*Cryptobranchus alleganiensis alleganiensis*) detection over conventional sampling methods. *Environ. DNA* 1, 86–96. doi: 10.1002/edn3.9

Yates, M. C., Fraser, D. J., and Derry, A. M. (2019). Meta-analysis supports further refinement of eDNA for monitoring aquatic species-specific abundance in nature. *Environ. DNA* 1, 5–13. doi: 10.1002/edn3.7

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: 16 February 2021
doi: 10.3389/fevo.2021.623655

Lord of the Diptera (and Moths and a Spider): Molecular Diet Analyses and Foraging Ecology of Indiana Bats in Illinois

Devon R. O'Rourke[1†], Matthew T. Mangan[2], Karen E. Mangan[2], Nicholas A. Bokulich[3], Matthew D. MacManes[1] and Jeffrey T. Foster[1†]*

[1] Department of Molecular, Cellular, and Biomedical Sciences, University of New Hampshire, Durham, NH, United States, [2] US Fish and Wildlife Service, Cypress Creek National Wildlife Refuge, Ullin, IL, United States, [3] Laboratory of Food Systems Biotechnology, Institute of Food, Nutrition, and Health, ETH Zürich, Zurich, Switzerland

OPEN ACCESS

Edited by:
Susana Caballero, University of Los Andes, Colombia

Reviewed by:
Bruce D. Patterson, Field Museum of Natural History, United States
Łukasz Kajtoch, Institute of Systematics and Evolution of Animals (PAN), Poland

***Correspondence:**
Devon R. O'Rourke
devon@outermostlab.com

†Present address:
Devon R. O'Rourke, Pathogen and Microbiome Institute, Northern Arizona University, Flagstaff, AZ, United States

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

Received: *30 October 2020*
Accepted: *06 January 2021*
Published: *16 February 2021*

Citation:
O'Rourke DR, Mangan MT, Mangan KE, Bokulich NA, MacManes MD and Foster JT (2021) Lord of the Diptera (and Moths and a Spider): Molecular Diet Analyses and Foraging Ecology of Indiana Bats in Illinois. Front. Ecol. Evol. 9:623655. doi: 10.3389/fevo.2021.623655

Effective management of endangered or threatened wildlife requires an understanding of how foraging habitats are used by those populations. Molecular diet analysis of fecal samples offers a cost-effective and non-invasive method to investigate how diets of wild populations vary with respect to spatial and temporal factors. For the federally endangered Indiana bat (*Myotis sodalis*), documenting its preferred food sources can provide critical information to promote effective conservation of this federally endangered species. Using cytochrome oxidase I amplicon sequence data from Indiana bat guano samples collected at two roosting areas in Cypress Creek National Wildlife Refuge, we found that dipteran taxa (i.e., flies) associated with riparian habitats were the most frequently detected taxon and represented the majority of the sequence diversity among the arthropods sampled. A select few arthropods from other taxa—especially spiders—are also likely important to Indiana bat diets in this refuge. A supervised learning analysis of diet components suggest only a small fraction of the frequently detected taxa are important contributors to spatial and temporal variation. Overall, these data depict the Indiana bat as a generalist consumer whose diet includes some prey items associated with particular seasonal or spatial components, along with other taxa repeatedly consumed throughout the entire foraging season. These molecular diet analyses suggest that protecting foraging resources specifically associated with the riparian habitat of Cypress Creek National Wildlife Refuge is essential to promote effective Indiana bat conservation.

Keywords: animal diets, metabarcoding, cytochrome oxidase, myotis sodalis, bat diet

INTRODUCTION

The Indiana bat, *Myotis sodalis*, has the dubious distinction of being the first North American bat listed under the Endangered Species Preservation Act (Udall, 1967). The historically broad distribution of Indiana bats once spanned much of the eastern United States (Thomson, 1982), however populations were dramatically reduced through decades of anthropogenic effects on habitat and required regional and national efforts to mitigate declines (Brady et al., 1983; O'Shea and Bogan, 2003; Lewis, 2007). Indiana bat populations appeared stable from the 1980s through

the early 2000s (Thogmartin et al., 2012; King, 2019), but the emergence of White-Nose Syndrome (WNS)—an infectious disease caused by a fungal pathogen (Lorch et al., 2011; Warnecke et al., 2012)—has decimated several bat species, resulting in near complete loss of some species at particular hibernation sites (Frick et al., 2010; Turner et al., 2011). WNS has been particularly devastating to Indiana bats in the Northeastern U.S. (Thogmartin et al., 2012; Jachowski et al., 2014; King, 2019), and populations are currently concentrated primarily in just four states; Kentucky, Missouri, Indiana, and Illinois populations constitute over 95% of all Indiana bats detected in winter 2019 (King, 2019).

Effective bat conservation requires protecting critical resources such as winter and summer habitats (Lewis, 2007; Johnson and King, 2018). Importantly, these summer habitat resources consist of both maternity colony sites as well as foraging areas. Understanding the particular foraging habitats used by bats from maternity colony roosts, for example, has led to refined strategies by policy-holders to engage with land managers (Johnson and King, 2018). However, Indiana bats occupy distinct territories within a landscape and often travel several kilometers between foraging habitats and roost sites (Garner and Gardner, 1992; Murray and Kurta, 2004). Thus, research that identifies preferences about roost site selection, for example Jachowski et al. (2016), provides essential information for guiding conservation practices, but does not fully convey the habitat needs of the species. Understanding food preferences may identify unique and additional required habitat in need of protection.

Radio-telemetry has identified foraging preferences of Indiana bats for forested areas in largely agricultural (Menzel et al., 2005; Womack et al., 2013) and urban (Sparks et al., 2005) landscapes. These studies highlight the growing importance of protecting the increasingly fragmented forested environments these bats use for both maternity colony roosts as well as foraging. Nevertheless, telemetry data may underestimate the home range used by Indiana bats (Womack et al., 2013) and discriminating which parts of the landscape are required habitat for the primary prey items is inherently challenging. For example, a bat may be infrequently detected over water, but aquatic prey items may be essential to the bat's diet. Alternatively, diet analysis can offer insights into the particular taxa consumed by the bat species, and thus further refine which habitats are essential for foraging, and therefore in greatest need of management and protection.

Previous studies using visual identification of arthropods in bat guano suggest that Indiana bats are frequent consumers of dipterans (flies), coleopterans (beetles), and lepidopterans (moths and butterflies) (Sparks et al., 2005; Tuttle et al., 2006), as well as trichopterans (caddisflies) in certain conditions (Murray and Kurta, 2004). However, such studies are limited by the number of samples analyzed and the specificity of the diet components identified: manual inspection requires substantial taxonomic expertise and time to classify arthropod contents. Further, even expert visual identification of arthropods in bat diets are typically limited to order or family-level specificity, and can fail to identify some prey completely—particularly soft bodied taxa (Kunz and Whitaker, 1983). The lack of precise taxonomic identification of food items makes it challenging to translate observations into detailed management strategies. Fortunately, adopting a molecular approach to diet analysis can provide the necessary taxonomic resolution to detail the breadth and specificity of Indiana bat foraging behaviors, and therefore give a more complete understanding of the habitat needs of the species. Furthermore, this workflow scales efficiently to hundreds or thousands of samples without requiring months or years of time invested, and can provide detailed information of arthropod diet composition regardless of the particular bat species. This allows for a comprehensive evaluation of diet and therefore foraging habitat requirements for many of the critically endangered bat species in North America. In the case of the Cypress Creek National Wildlife Refuge, this information can be used to inform the particular habitats in most need of protection.

Located in between the Ohio and Mississippi Rivers in Southern Illinois, Cypress Creek National Wildlife Refuge contains riparian bottomland hardwood forests—ideal summer roosting habitats (Cable et al., 2020). In addition, it is within 8 km of a large Indiana bat hibernaculum (Brown and Melius, 2014). However, concerns about habitat loss and limited roost availability served as an impetus to evaluate if artificial roost structures installed in the refuge would expand roosting use to areas that were otherwise not suitable for maternity colonies (Mangan and Mangan, 2017). Prior mist-netting and radio-telemetry surveys of the region indicated that bats occupied a particular stretch of riparian habitat surrounded by agricultural landscapes (Mangan and Mangan, 2019a). In fact, this radio-tracking led to confirmation of an Indiana bat occupying one of Egner roosts, which served as an impetus for conducting this diet work. These results indicated the area as suitable roosting habitat for bat maternity colonies, but it was unclear whether or not the same habitat was important for bat foraging.

DNA barcoding (or metabarcoding) provides a cost-effective method to rapidly generate datasets rich with taxonomic information (Valentini et al., 2009; Pompanon et al., 2012; Alberdi et al., 2018, 2019; Nielsen et al., 2018). Molecular diet analyses have been widely applied to a range of systems and organisms, although the methodology is not without challenges and biases (Nielsen et al., 2018; Alberdi et al., 2019). Early bat diet studies using a molecular approach described greater breadth and specificity of prey items consumed compared to traditional microscopy (Clare et al., 2009; Zeale et al., 2011). While both *in silico* (Clarke et al., 2014) and empirical (Hope et al., 2014) studies have identified potential taxa that may be missed due to PCR biases, recent modifications of primer sequences have resolved many of the amplification issues for certain taxa (Jusino et al., 2019). Subsequent applications using this molecular method have revealed key features of bat foraging in several *Myotis* species that can be used to optimize management decisions regarding habitat preservation. For instance, the genus or species-level taxonomic resolution using these molecular methods indicates prey specificity for *Myotis septentrionalis* (Dodd et al., 2012) and *M. daubentonii* (Vesterinen et al., 2016); protections for the habitats that sustain these prey items would ensure these bats have available food resources.

Metabarcoding has improved both the specificity of bat diet contents as well as potential spatial and temporal changes

in foraging patterns. For example, studies of *M. lucifugus* indicate that core dietary components can vary both by location (Clare et al., 2011) and season (Clare et al., 2014), suggesting that incorporating diet information into conservation efforts may require factoring in regional and temporal variation into management considerations. However, metabarcoding diet interpretations are complicated by whether or not a researcher chooses to link the sequence data (i.e., counts of amplicons) to species abundances (Alberdi et al., 2019; Deagle et al., 2019). We conducted our diversity analyses using both abundance-unweighted and weighted means to provide an example of how the inclusion or exclusion of sequence count information can potentially alter the subsequent inferences made from the data.

In addition, management policy would benefit by moving beyond simple lists of prey items detected in batches of guano, and evaluate if specific diet components are important to particular classes of metadata. We applied a Random Forest classifier—a supervised learning tool (a type of machine learning)—to determine what bat diet components were most important in predicting the location or site a sample was collected. These data can assist in identifying whether the same foraging areas are needed to be protected at all points of the year, and whether or not particular locations are more important for conservation with respect to Indiana bat foraging. This form of supervised learning has been applied to a range of 16S rRNA and ITS amplicon studies including identifying origins of ballast water (Gerhard and Gunsch, 2019), predicting taxonomic signatures of host fecal microbiomes (Roguet et al., 2018), understanding maternal microbiome patterns associated with preterm delivery (Dahl et al., 2017), and predicting wine metabolite profiles (Bokulich et al., 2016). Rather than summarizing the unique sequence variants of the data directly (e.g., through ordination), important sequences are identified in Random Forest classifiers by quantifying their relative contribution to the predictive accuracy of a model (Breiman, 2001; Bokulich et al., 2018b).

Guano collected as part of this study afforded an opportunity to provide the first molecular analysis of Indiana bat diets. Indiana bats are one of several threatened or endangered species in need of significant protections, and identifying trends in foraging habits serve to complement ongoing efforts to identify relevant habitat to preserve. The methods described herein offer one such means to attain improved species protections based on a detailed understanding of diet and foraging.

MATERIALS AND METHODS

Data Availability

Data, figures, and scripts applied are Available online at the GitHub repository for this project: https://github.com/devonorourke/mysosoup. **Supplementary Tables 1–3** referred to herein are available online at this repository in the "Supplementary Material" directory. We provide additional documentation for sequencing processing, database curation, classification, and diversity estimates in a "docs" folder within that GitHub repository—see the bioinformatics sections below for links to each of these documents. Raw sequences for this project are Available online at BioProject PRJNA548356. Database files are stored in the Open Source Frameworks repo of this project: https://osf.io/qju3w/. A Zenodo archive of this repository is available for download here: https://zenodo.org/badge/latestdoi/176534517.

Site Selection and Guano Collection

The Cache River Watershed comprises thousands of acres of riparian wetland forests essential to Indiana Bat foraging and roosting, and is contained within the current ∼17,000 acre Cypress Creek National Wildlife Refuge. The sampling sites were on two tracts of land approximately three miles apart: Hickory Bottoms and Egner (**Figure 1**). Each tract contained four artificial Brandenbark™ roosting habitats (Adams et al., 2015); installation of the structures was completed in 2014. These tracts consist of agricultural land mixed with mature bottomland forests containing live and standing dead trees or snags with exfoliated bark or crevices suitable for Indiana bat roosts. Both locations have adjacent riparian habitat, with Egner roosts abutting the Cache River, and Hickory Bottoms abutting Cypress Creek. Use of these structures by Indiana bats was determined through fieldwork conducted in July and August 2016 at the refuge using mist-netting, radio-telemetry, and acoustic surveys (Mangan and Mangan, 2019a).

Guano was collected at each of the eight roosts June 21, July 27, and September 15, 2017. These dates correspond to the periods prior to or during parturition and weaning in June and July respectively, and in September during expected fall migration. Plastic sheets were placed at the base of each roost the night prior to collection and replaced with new sheets before the next collection date. Up to ten guano pellets were obtained at each roost at each date using sterile forceps and were stored individually in microcentrifuge tubes. All guano was sent to the University of New Hampshire and stored at −80°C until DNA extraction. We limited our analyses to single-pellet guano samples, although bulk samples of guano containing many pellets were also collected.

DNA Extraction

Guano pellets were extracted using the Qiagen DNeasy PowerSoil kits (Qiagen, Hilden, Germany) following manufacturer guidelines. Two 96-well plates were used to process 175 individual pellets and included either 5 or 9 negative control wells. The remaining 41 individual pellets were processed with single tube extractions using the same kit chemistry. All samples were eluted with 100 μL of elution buffer.

Metabarcoding

Concentrations of guano extract DNA were estimated with a Nanodrop spectrophotometer (Thermo Fisher, Waltham, MA, USA) to guide the appropriate volumes of sample to add for subsequent normalization with SequalPrep plates following manufacturer guidelines (Applied Biosystems, Foster City, CA, USA). Highly concentrated samples were diluted so that samples were standardized to ∼2 ng/μL prior to normalization. Normalized DNA was used as input for our overlap extension PCR method that targets arthropod COI fragments. Arthropod

FIGURE 1 | Collection sites within Cypress Creek National Wildlife Refuge (CCNWR), which is situated in southern Illinois and consists of ~17,000 acres of wetland, forested, and grassland habitats. The Cache River Unit comprises just 1,136 acres of riparian wetland forests essential to Indiana Bat foraging and roosting. Guano was collected from two locations ~3 miles apart: Hickory Bottoms (Hickory) and Egner. Each collection area contained 4 artificial roosting habitats, depicted as points on each inset map. Sources: Google, ESRI, DigitalGlobe, GeoEye, Earthstar Geographics, CNES/Airbus DS, USDA, USGS, AeroGRID, IGN, and the GIS User Community.

COI gene fragments are targeted for amplification using primers detailed in Cable et al. (2020). We modified the original primer sequences to preserve the COI-specific regions, but added 5′ extensions of 17 and 19 bp, respectively. The constructs below illustrate these additional tails (bold underlined bases) as part of the modified oligos using the original Jusino sequences (not underlined):

UT-ANML-LCO1490: 5′-ACCCAACTGAATGGAGC **GGTCAACAAATCATAAAGATATTGG**-3′

UT-ANML-CO1-CFMRa: 5′-ACGCACTTGACTTGTCTTC **GGWACTAATCAATTTCCAAATCC**-3′.

Samples were amplified in 15 μL reactions, with 3 μL of normalized guano DNA extract added to 12 μL of solution containing 0.2 μM of the forward and reverse primers, 0.16 μg/μL BSA, 0.03 U/μL Platinum Taq, 0.2 mM dNTPs, 1.5 mM MgCl2, and 1.5 μL of 10X buffer (Invitrogen, Carlsbad, CA, USA). Thermal cycler settings for the reaction consisted of an initial 5 min denaturation at 94°C, followed by 5 cycles of 60 s at 94°C, 90 s at 45°C, and 90 s at 72°C; an additional 35 cycles of 60 s at 94°C, 90 s at 50°C, and 60 s at 72°C; and finally a 10 min extension at 72°C.

PCR reactions were subjected to a 1X AMPure XP bead cleanup (Agilent Technologies, Santa Clara, CA, USA) and 10 μL of the concentrated solution was normalized in SequalPrep plates (Applied Biosystems, Foster City, CA, USA). These normalized PCR products were then subject to a second amplification using custom oligos that contained the requisite Illumina adapters, a distinct 8mer barcode, and the complementary sequence to overlap with the 5′ terminus of the amplicon. The example below illustrates an example of these constructs, where the underlined portion represents an 8mer barcode, with the Illumina adapters upstream of the barcode, and the complementary overlap downstream from the barcode (in bold) to facilitate polymerase extension of the original PCR product:

Indexed-UT1-example_pair1a:
5′-AATGATACGGCGACCACCGAGATCTACACCACACAAA **GCTGGTCATCGTACCCAACTGAATGGAGC**-3′

Indexed-UT1-example_pair1b:
5′-CAAGCAGAAGACGGCATACGAGATTTTGTGTG **AGTCAGTCAGCCACGCACTTGACTTGTCTTC**-3′.

We added 2 μL of normalized PCR products (from the initial amplification) with 0.4 μM of each index primer in 25 μL reaction volumes using KAPA HiFi HotStart ReadyMix (KAPA Biosystems, Wilmington, MA, USA). Reaction conditions consisted of a 2 min denaturation at 98°C, followed by 10 cycles of 30 s at 98°C, 20 s at 60°C, and 30 s at 72°C, and a final extension for 5 min at 72°C. These final PCR products were subject to another 1X bead cleanup and normalization following

the same methods described above. We created the final library by pooling 10 μL of normalized PCR products into a single tube and concentrated to 40 μL with a 1X bead cleanup.

Library concentration was quantified by qPCR using the KAPA ROX Low Complete Kit (KAPA Biosystems, Wilmington, MA, USA). An Illumina MiSeq sequencer (Illumina, San Diego, CA, USA) with v3 chemistry generated 600 cycles of 2 × 300 bp paired-end reads.

Bioinformatics

Sequence Denoising and Filtering

Demultiplexed sequences were trimmed using Cutadapt v.1.18 (Martin, 2011) using "-m 100 –trimmed-only" parameters to retain only sequences at least 100 base pairs in length and with a detectable primer sequence. Trimmed reads were imported into a QIIME 2 v2019.10 environment (Bolyen et al., 2019) and representative sequences were identified using DADA2 v1.6.0 (Callahan et al., 2016) via the q2-dada2 QIIME 2 plugin function "qiime dada2 denoise-paired" that included retaining only the first 175 bases of the forward and reverse sequences via the "–p-trunc-len" parameter. Full details regarding sequence processing commands are described here: https://github.com/devonorourke/mysosoup/blob/master/docs/sequence_processing.md.

Because 7 of the 15 control samples (from 96-well plate DNA extractions) retained denoised sequences, we investigated whether the sequence variants present in control samples were due to contamination either through DNA extraction or PCR amplification. We found no evidence of systemic contamination, and removed the negative control samples from subsequent analysis. We justify this decision using the strategies discussed here: https://github.com/devonorourke/mysosoup/blob/master/docs/contamination_investigations.md.

Construction of Databases for Taxonomic Classification

The primers used in this study were shown previously to amplify bat COI (Jusino et al., 2019). To identify which bats contributed the guano collected in the experiment, we created a host database consisting of sequences derived from all known bat species in the region. In addition, we included all other known host reference sequences from other guano-related projects in our lab as a precaution for potential cross contamination (ultimately no unexpected host sequences were detected). Full details regarding host database design are documented here: https://github.com/devonorourke/mysosoup/blob/master/docs/host_database.md.

A second (larger) database was constructed as an additional method to identify any bat DNA missing from our smaller custom database, as well as to classify all other sequence features present in our dataset. We collected reference sequences and associated taxonomy information from two resources: BOLD (Ratnasingham and Hebert, 2007) and a GenBank-derived dataset curated by Terri Porter (Porter and Hajibabaei, 2018). Reference sequences included COI records from arthropod, chordate, and other animal taxa, as well as fungal, protist, and other microeukaryote COI records. We dereplicated the initial collection of sequences, then applied a Least Common Ancestor (LCA) process using a consensus approach to classify records that shared identical sequence information but differed with respect to taxonomic information. Additional filters included discarding references with non-standard IUPAC DNA characters, removing sequences <100 bp, and retaining only references that contained at least family-level names. The final dataset included 2,181,331 distinct sequences. The construction of this database is described here: https://github.com/devonorourke/mysosoup/blob/master/docs/database_construction.md.

Taxonomic Classification

We identified host sequences using a combination of alignment and machine learning approaches to independently confirm what bat species contributed to the guano in this experiment. The denoised representative sequences were initially aligned to our custom host database of bat sequences using VSEARCH (Rognes et al., 2016) to identify and separate host ASVs from non-host ASVs. Candidate matches were subsequently queried with NCBI BLAST (Camacho et al., 2009) to confirm host identities. We then used our larger COI database as a third means with which to discriminate among host and non-host sequences. Sequence features were classified using two methods available through the QIIME 2 plugin q2-feature-classifier (Bokulich et al., 2018a,b): first, a VSEARCH global alignment approach followed by least common ancestry taxonomy assignment with "qiime feature-classifier classify-consensus-vsearch"; and second, a supervised learning naive Bayes classifier with "qiime feature-classifier classify-sklearn." All methods identified a common set of bat-associated ASVs from the original dataset, and were used to determine the proportion of the various bat species detected in the guano. Importantly, we found that nearly all sequence data classified as host DNA belonged to *M. sodalis*, the species we expected from previous (Mangan and Mangan, 2019a) and subsequent (Mangan and Mangan, 2019b) surveillance work that concluded that the Indiana bat was the primary occupant of the artificial roosts where guano was collected. We discarded samples from our analyses for instances in which a bat host other than *M. sodalis* was assigned specifically to that sample.

For our diet analyses, representative sequences were further clustered with "qiime vsearch cluster-features-*de-novo*" using a 98.5% identity. Clustered sequence variants were classified using a hybrid approach that involved assigning taxonomic names using both naive Bayes and VSEARCH+LCA classifier methods in q2-feature-classifier. This approach prioritizes those records with exact alignments first using VSEARCH (those taxa with 100% identity and at least 94% coverage), and any clustered sequence variants that remained unclassified following this initial alignment are then classified using the naive Bayes method approach. Only those clustered sequences assigned to the Arthropoda phylum, with at least family-level taxonomic names, were retained for diversity estimates and supervised learning analyses.

Full details describing the host identification methods are described here: https://github.com/devonorourke/mysosoup/blob/master/docs/classify_sequences.md.

Diversity Estimates

We used several different approaches to generate diversity estimates, with careful attention to the suitability of the estimator in relation to the data type, making sure comparisons controlled for factors such as sequencing depth, and correcting for multiple comparisons. The dietary components identified as representative sequence clusters were rarefied to 10,000 sequences per sample for diversity estimates. Observed richness and Shannon's entropy values were calculated for these representative sequence clusters. Because the subsequent values did not follow a normal distribution (p-values <0.01 using Shapiro-Wilks test), we applied a Kruskal-Wallis non-parametric test to compare whether there were differences between groups collected in particular Site + Month combinations (e.g., Egner in June, Egner in July, Hickory in September, etc.). Significance values of pairwise differences were calculated using a Wilcoxon rank sum test, using a Benjamini-Hochberg correction for multiple testing.

Community composition among diet components was assessed using two different approaches: one using a presence-absence analysis of the sequence variants detected in each sample, and one incorporating the abundance information associated with the counts of sequence variants. Specifically, dissimilarities in composition of representative sequence variants were evaluated with non-phylogenetic binary (Dice-Sorensen) and abundance-weighted (Bray-Curtis) distances, as well as phylogeny-weighted binary and abundance-weighted distances. We explored these dissimilarities using Principal Coordinates Analysis, visualizing the first two principal components for each distance metric. Main effects of site and month on community composition were tested using the Vegan "Adonis" function; we also performed an analysis of multivariate homogeneity of group dispersions with the Vegan "betadisper" function.

Full details describing associated sequence processing, and associated R scripts used in generating the figures and data tables presented herein are described here:

https://github.com/devonorourke/mysosoup/blob/master/docs/diversity_workflow.md.

Core Features and Supervised Learning

Non-rarefied clustered sequence data was filtered to identify those variants present in at least 10% (20 or more) of guano samples using the QIIME2 "feature-table core-features" function. These core sequence variants were used in a custom R script to generate the summary figure and tables comparing the frequency of occurrence and sequence abundances for each OTU among samples.

These "core" sequence features were used in the subsequent supervised learning approach via the QIIME 2 "classify-samples-ncv" pipeline (part of the q2-sample-classifier (Pedregosa et al., 2011; Bokulich et al., 2018b) plugin) to train Random Forest classifiers. This nested cross-validation approach works in a similar fashion to standard splitting of data into testing and training subsets, but repeats the testing/training process k-times. In reshuffling the data we ensure that all sequence features are tested for relative importance to a model. Three classifiers were built and tested: a model for site, a model for month, and a model for site + month metadata classes. We increased the number of decision trees available to the model from the default (100) to 1,000 estimators, with the intention of improving the predictive accuracy. In addition, we selected an option to identify optimal feature selection (–p-parameter-tuning) which automatically selects the number of features considered during node splits on a given decision tree. Complete details for QIIME functions and associated R scripts visualizing the output are documented here:

https://github.com/devonorourke/mysosoup/blob/master/docs/diversity_workflow.md.

Additional software

Figures and statistical analyses were performed in R (R Core Team, 2018) using multiple libraries (Paradis et al., 2004; Wickham, 2007, 2017, 2018, 2019; Chamberlain and Szöcs, 2013; Kahle and Wickham, 2013; Chamberlain et al., 2014; Lumley, 2016; Ren and Russel, 2016; Wilke, 2017; Bates and Maechler, 2018; Bisanz, 2018; Garnier, 2018; Kassambara, 2018; Ogle et al., 2018; Pedersen and Crameri, 2018; Slowikowski, 2018; Graves et al., 2019; Oswaldo, 2019; Pedersen and Robinson, 2019; Wickham et al., 2020). QIIME 2 plugins for data processing and diversity analyses were also utilized (McKinney, 2010; Price et al., 2010; McDonald et al., 2012; Weiss et al., 2017; Robeson et al., 2020).

RESULTS

We applied a metabarcoding technique to amplify arthropod COI gene fragments and generated sequence data from hundreds of bat guano samples collected at artificial roosts erected at two locations in the Cypress Creek National Wildlife Refuge during the summer of 2017 (**Figure 1**). Although the primers used to amplify COI fragments were designed for arthropod sequences, other COI sequences such as host DNA often amplify as well. Thus, we first identified and separated host from non-host sequence variants. In 144 of our 196 single-pellet samples sequence variants classified exclusively to one of three bat species: Indiana bat (*M. sodalis*), little brown bat (*M. lucifugus*), and evening bat (*Nycticeius humeralis*) (**Supplementary Table 1**). The vast majority of these were classified as Indiana bat (137 samples), with rare detections of little brown bat (5 samples) and evening bat (2 samples). Those seven samples classified uniquely to little brown and evening bats were disarded from our diet analyses. In addition, 11 samples contained sequence variants from two or more species, all of which included the Indiana bat; these were included in the diet analysis. We included guano samples that lacked host classification, as many samples did not generate any host sequences. These findings corroborate previous field observations (Mangan and Mangan, 2017, 2019a) that while other species transiently occupy similar roosts, the Indiana bat is the primary occupant of the colonies where guano was collected. We acknowledge that a minor fraction of arthropod data may have come from the diet of a bat species other than Indiana bat.

The breadth of arthropod taxa detected across all samples was substantial, with 1,070 unique sequence clusters classified to 19 arthropod orders among the 189 guano samples. However,

TABLE 1 | Eight arthropod orders detected in at least 10% of samples.

Arthropod order	Fraction of samples with order detected	Fraction of OTUs in dataset
Diptera	98.4	37.1
Lepidoptera	94.2	23.2
Araneae	92.6	8.3
Hemiptera	66.7	8.4
Coleoptera	51.9	8.1
Psocodea	37.0	1.9
Trichoptera	22.2	0.7
Ephemeroptera	14.8	0.4

Fraction of samples with order detected required at least one OTU classified to that arthropod order to be present in a sample (but multiple OTUs of the same order may be present). The fraction of OTUs for each arthropod order are relative to the entire 1,070 sequence clusters classified to all arthropods in the dataset.

a particular subset of arthropods was much more likely to be observed than others. Eight orders of arthropods were identified in more than 10% of samples: Araneae, Coleoptera, Diptera, Ephemeroptera, Hemiptera, Lepidoptera, Psocodea, and Trichoptera (**Table 1**). Among these taxa, just two arthropod orders represented more than half of all sequence clusters. OTUs are defined as the most abundant exact sequence variant observed in our data amongst all exact sequences within a 98.5% identity threshold. Diptera (397 OTUs representing over 37% of all classified taxa) and Lepidoptera (248 OTUs, 23% of taxa). Interestingly, the number of distinct sequence clusters classified to a particular order did not necessarily correlate with frequency of detection. Thus, while flies and moths were detected in the most samples and contained the greatest number of unique sequence clusters, nearly as many samples contained other detectable orders, but those particular orders contained far fewer distinct sequence clusters within that particular group. Spiders, for example, were detected in 175 samples (92%) despite representing only ∼8% of all arthropod sequence clusters.

Despite generating a taxonomically broad collection of arthropod amplicons, only a small fraction of these were routinely identified. Just 56 of the 1,070 arthropod sequence clusters were identified in at least 10% of our samples, with several OTUs containing common taxonomic labels (**Table 2**). Among these "core" sequence clusters, two-thirds were classified as dipteran (37 OTUs). These dipteran OTUs are dominated by taxa known to inhabit the native riparian habitat. For example, we detected limoniid craneflies such as *Epiphragma solatrix* (112 samples) and *Erioptera caliptera* (101 samples), and tipulids such as *Nephrotoma ferruginea* (76 samples). Mosquitoes such as *Culex erraticus* (121 samples) and *Coquillettidia perturbans* (58 samples) were also frequently detected. While the majority of the core sequence clusters were classified as flies, an orb-weaving spider classified to the genus *Eustala* was the most frequently detected sequence cluster in the entire dataset (146 samples). Non-dipteran core OTUs were distributed among seven arthropod orders with just three orders containing more than one representative sequence cluster: Araneae (7 OTUs), Lepidoptera (5 OTUs) and Psocodea (2 OTUs). These molecular-level data suggest Indiana bats in the Cypress Creek National Wildlife Refuge routinely eat a diverse assortment of flies, along with a particular few representative species of other arthropods, and especially orb-weaving spiders in the genus *Eustala*.

We calculated the observed richness and Shannon's entropy of samples to investigate whether diet components were associated with the site and date a sample was collected (**Figure 2**). We applied a Kruskal-Wallis test to determine if the mean rank sums of diversity estimates of each site-date group varied, and found a significant difference for observed richness [$H(5) = 25.389$, $p < 0.001$], but not for Shannon's entropy [$H(5) = 2.174$, $p = 0.825$]. A Wilcoxon signed-rank test was applied to determine pairwise differences of the site-date group diversity estimate. Observed richness was higher among samples collected at Egner in June than either site in September, however, no differences in Shannon's diversity were detected among any site-date pair (**Figure 2**).

We next explored how community composition varied among site+date groups, and evaluated the effect of using abundance and/or phylogenetic-weighted metrics. Using a multifactorial PERMANOVA (Adonis) to test for group differences in spatial median, we found significant effects ($p < 0.01$) for both site and date using every distance metric evaluated: Dice-Sorensen (non-abundance, non-phylogenetic), Bray-Curtis (abundance-weighted, non-phylogenetic), unweighted UniFrac (non-abundance, phylogenetic-weighted), and weighted UniFrac (abundance-weighted, phylogenetic-weighted) (**Supplementary Table 2**). We also tested for dispersion differences for each group using a univariate ANOVA, PERMDISP (betadisper), and found that the effect of site ($p = 0.462$) but not date ($p < 0.001$) were non-significant for weighted UniFrac. Group dispersions for all other metrics were significant at a threshold of $p < 0.01$, while the effect of date on dispersions of Bray-Curtis distances was marginally higher at $p = 0.048$ (**Supplementary Table 3**). Because we used a balanced design, these results suggest that month and site variability in community composition occur both because of spatial group median and dispersion differences for unweighted-abundance metrics. However, the non-significant dispersion result for the Weighted UniFrac group dispersion for the effect of site suggest that there are true compositional differences between collection sites. A Principal Coordinates Analysis of these distance measures indicate that these abundance-weighted metrics provided the greatest proportion of variance in the first two principal component axes (**Figure 3**), with samples associating more by site than by date. Nevertheless, the relatively small proportion of variation shown in these ordinations also support the notion that many of the prey items that bats consume are present throughout the entire sampling period of the study, thus the overall impact of month or site differences appear minor.

A supervised learning regime was applied to the core sequence clusters by training Random Forest classifiers to each group (site, date, or site-date). For each group, we determined the accuracy of the model (that is, how often did a sample get assigned to its expected group), as well as calculated the relative importance of each OTU in building the model (OTUs with the

TABLE 2 | Taxonomic information assigned to prevalent sequence clusters (OTUs) detected in Indiana bat guano.

Order	Family	Genus	Species	Samples detected
Araneae	Araneidae	*Eustala*	sp.	142
Araneae	Tetragnathidae	*Tetragnatha*	*elongata*	34
Araneae	Araneidae	*Eustala*	*cepina*	33
Araneae	Anyphaenidae	*Anyphaena*	*pectorosa*	27
Araneae	Theridiidae	*Theridion*	*albidum*	24
Araneae	Tetragnathidae	*Leucauge*	*venusta*	22
Araneae	Araneidae	*Neoscona*	Neoscona sp. 1GAB	21
Coleoptera	Ptilodactylidae	*Ptilodactyla*	sp.	35
Diptera	Culicidae	Undetermined	sp.	133
Diptera	Limoniidae	Undetermined	sp.	130
Diptera	Limoniidae	*Rhipidia*	sp.	120
Diptera	Culicidae	*Culex*	*erraticus*	118
Diptera	Limoniidae	*Epiphragma*	*solatrix*	110
Diptera	Limoniidae	*Erioptera*	*caliptera*	101
Diptera	Chironomidae	*Chironomus*	sp.	91
Diptera	Limoniidae	*Erioptera*	parva	88
Diptera	Chironomidae	Undetermined	sp.	77
Diptera	Chironomidae	*Glyptotendipes*	sp.	68
Diptera	Tipulidae	*Nephrotoma*	*ferruginea*	67
Diptera	Limoniidae	*Helius*	*flavipes*	63
Diptera	Culicidae	*Coquillettidia*	*perturbans*	49
Diptera	Culicidae	*Uranotaenia*	*sapphirina*	49
Diptera	Chironomidae	*Glyptotendipes*	*meridionalis*	48
Diptera	Limoniidae	*Metalimnobia*	*triocellata*	48
Diptera	Tabanidae	*Tabanus*	*similis*	36
Diptera	Chaoboridae	*Chaoborus*	*punctipennis*	35
Diptera	Culicidae	*Culex*	*territans*	34
Diptera	Tabanidae	Undetermined	sp.	34
Diptera	Chironomidae	*Axarus*	*festivus*	31
Diptera	Dolichopodidae	Undetermined	sp.	30
Diptera	Psychodidae	Undetermined	sp.	28
Diptera	Tipulidae	*Nephrotoma*	*okefenoke*	26
Diptera	Tipulidae	*Tipula*	sp.	22
Diptera	Limoniidae	*Gonomyia*	sp.	20
Diptera	Limoniidae	*Pseudolimnophila*	*luteipennis*	20
Ephemeroptera	Heptageniidae	*Stenacron*	*interpunctatum*	27
Hemiptera	Flatidae	Undetermined	sp.	24
Lepidoptera	Tineidae	*Acrolophus*	*mortipennella*	49
Lepidoptera	Tortricidae	*Clepsis*	*peritana*	28
Lepidoptera	Tortricidae	*Choristoneura*	sp.	27
Lepidoptera	Gelechiidae	*Coleotechnites*	*florae*	20
Lepidoptera	Oecophoridae	*Inga*	*sparsiciliella*	20
Psocodea	Psocidae	*Metylophorus*	*novaescotiae*	51
Psocodea	Psocidae	*Blaste*	Blaste sp. 2KJEM	20
Trichoptera	Hydropsychidae	*Potamyia*	*flava*	24

Each line represents uniquely classified taxa among the 56 distinct OTUs detected in at least 10% of samples. OTUs with redundant taxonomic assignment were grouped together, and sorted by taxonomic order with the most frequently detected taxa shown first.

greatest importance are those that best discriminate samples for a grouping variable). All three classifier models were successful in predicting a sample's grouping variable from the 56 core OTUs. The model correctly predicted a sample's collection month more than 85% of the time (**Figure 4A**), as well as the site for 75% of samples (**Figure 4B**), and the specific site+date for

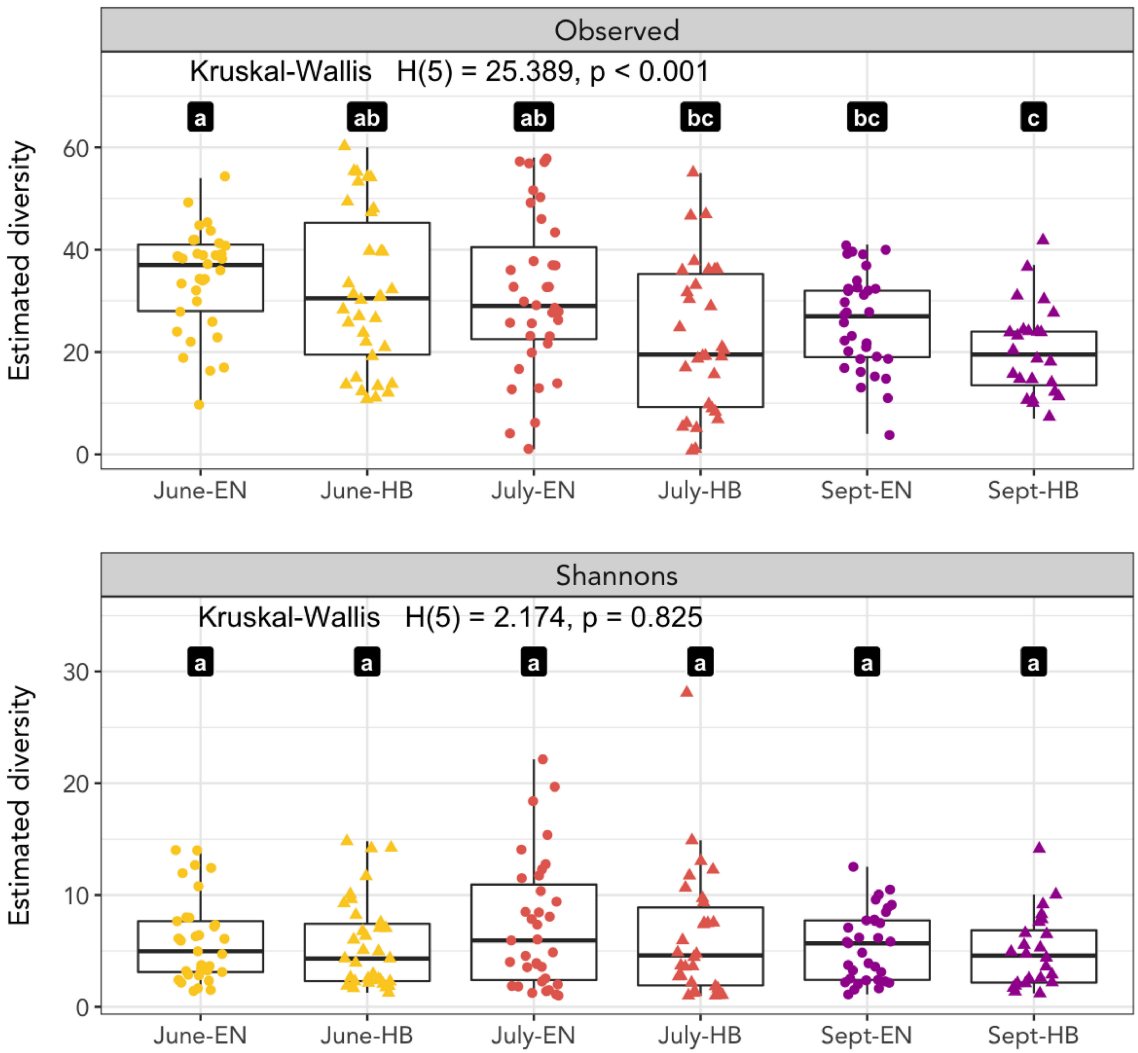

FIGURE 2 | Species richness as measured by observed richness and Shannon's entropy. Significant differences between groups of samples collected at each site-date (Egner, "EN"; Hickory Bottoms, "HB") represented by distinct lettered values.

77% of samples (**Figure 4C**). Most of these core OTUs do not play a significant role in discriminating samples between the site and date groups, as represented by their low relative importance to each model (**Figure 4D**). More than 50% of the overall importance to each model was accounted for by a few sequence clusters: 11 OTUs for site-date, 10 OTUs for site, and just 8 OTUs for date. These highly discriminant OTUs spanned a broad range of taxa, despite dipteran sequences dominating the overall dataset with respect to detections per sample and sequence cluster richness. For example, a barklice species, *Metylophorus novaescotiae* (OTU-1 in **Figure 4D**), was the most important sequence cluster for September samples at both sites (in fact, it had the highest individual importance score of any OTU for any model). A moth, *Acrolophus mortipennella* (OTU-2), was indicative of samples collected at both sites in June. A net spinning caddisfly, *Potamyia flava* (OTU-43), was the strongest indicator of a sample originating from the Hickory site. Dipteran sequence clusters were also relevant at discriminating between sampling date or site. For example, sequence clusters classified to *Glyptotendipes* (OTUs 19 and 33) predicted the sampling site, while a pair of mosquitoes, *Uranotaenia sapphirina* (OTU-4) and *Culex erraticus* (OTU-7) were discriminant for sampling date. Notably, the remaining core OTUs that failed to discriminate site or sampling month does not suggest their lack of importance to the Indiana bat diet—they simply share similar detection frequencies and sequence abundances frequencies, and therefore do not help the model differentiate a grouping variable.

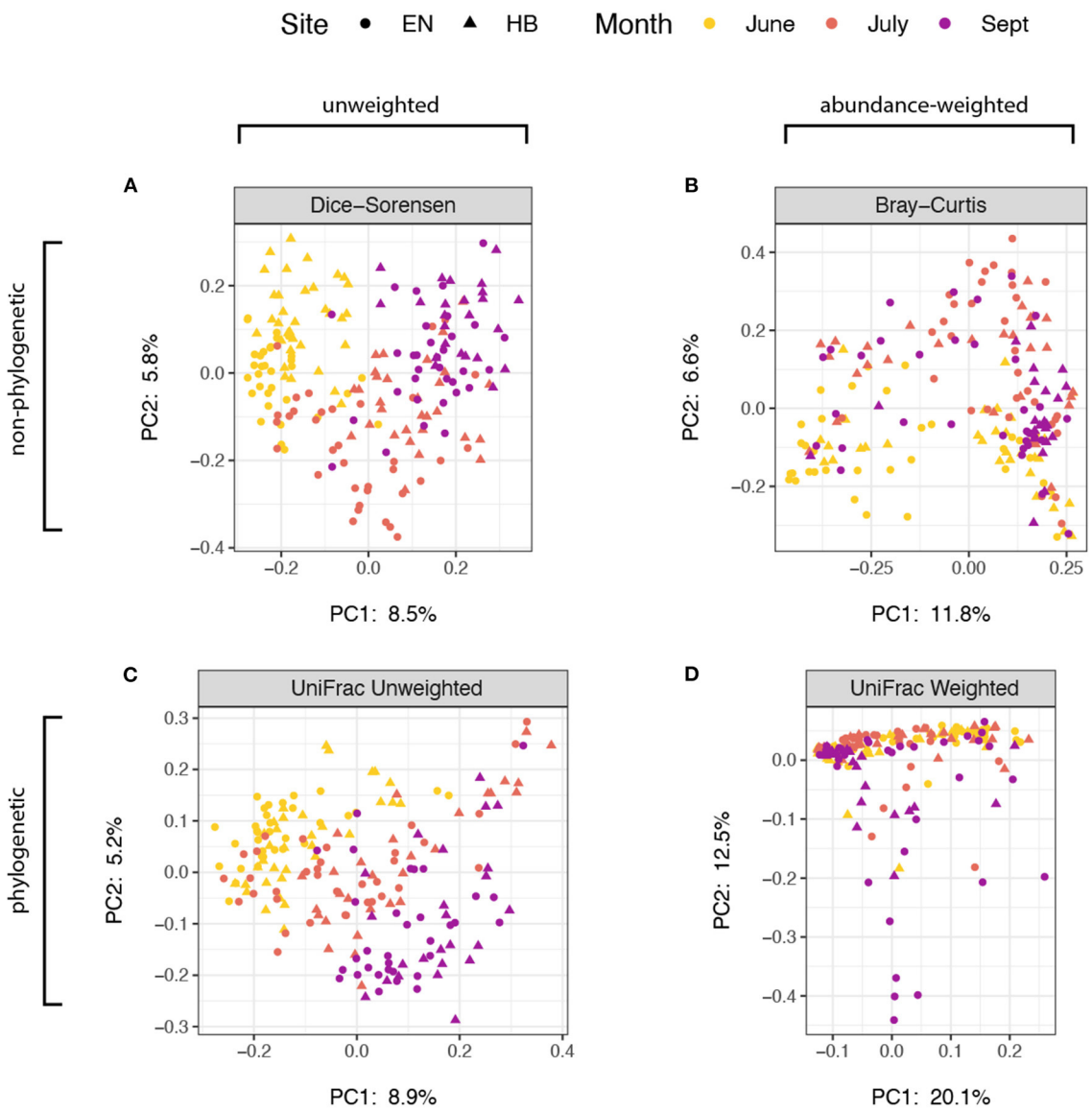

FIGURE 3 | Principal coordinates analysis of distance estimate ordinated with samples distinguished by sampling site as points (Egner, "EN"; Hickory Bottoms, "HB") and sampling month as colors. The four distance metrics varied with respect to sequence abundance and phylogenetic weights: **(A)** Dice-Sorensen, unweighted abundance and unweighted phylogenetic; **(B)** Bray-Curtis, weighted-abundance and unweighted phylogenetic; **(C)** UniFrac Unweighted, unweighted abundance, weighted phylogenetic; **(D)** UniFrac Weighted, weighted abundance and weighted phylogenetic. The proportion of variance captured by each of the first two principal component axes are shown.

DISCUSSION

Much of existing bat conservation policy in North America focuses on identifying and conserving winter hibernacula and summer maternity roosts. With the decline of insects globally, and the direct impact on aerial insectivores such as bats, the need to connect diet and foraging to habitat needs is clear. For the Indiana bat specifically, a framework to understand the particular resources essential for foraging habitats is still being developed. We found that the molecular techniques applied herein offer a rapid and cost-effective solution that is capable of achieving a greater taxonomic resolution of bat diets than previous morphological estimates. Collectively, these molecular data indicate Indiana bats are generalist predators, confirming earlier morphological analyses of guano contents that this bat species' diet consists of Coleoptera, Diptera, and Lepidoptera. However, we observed dipteran taxa as the largest proportion of fecal content using molecular methods, while most of the morphological analyses suggest Indiana bat guano consists of coleopteran and lepidopteran taxa [see Figure 1 in Sparks et al. (2005) for a review]. This disparity was also depicted in a survey conducted in Shawnee National Forest—just 20 miles east of our

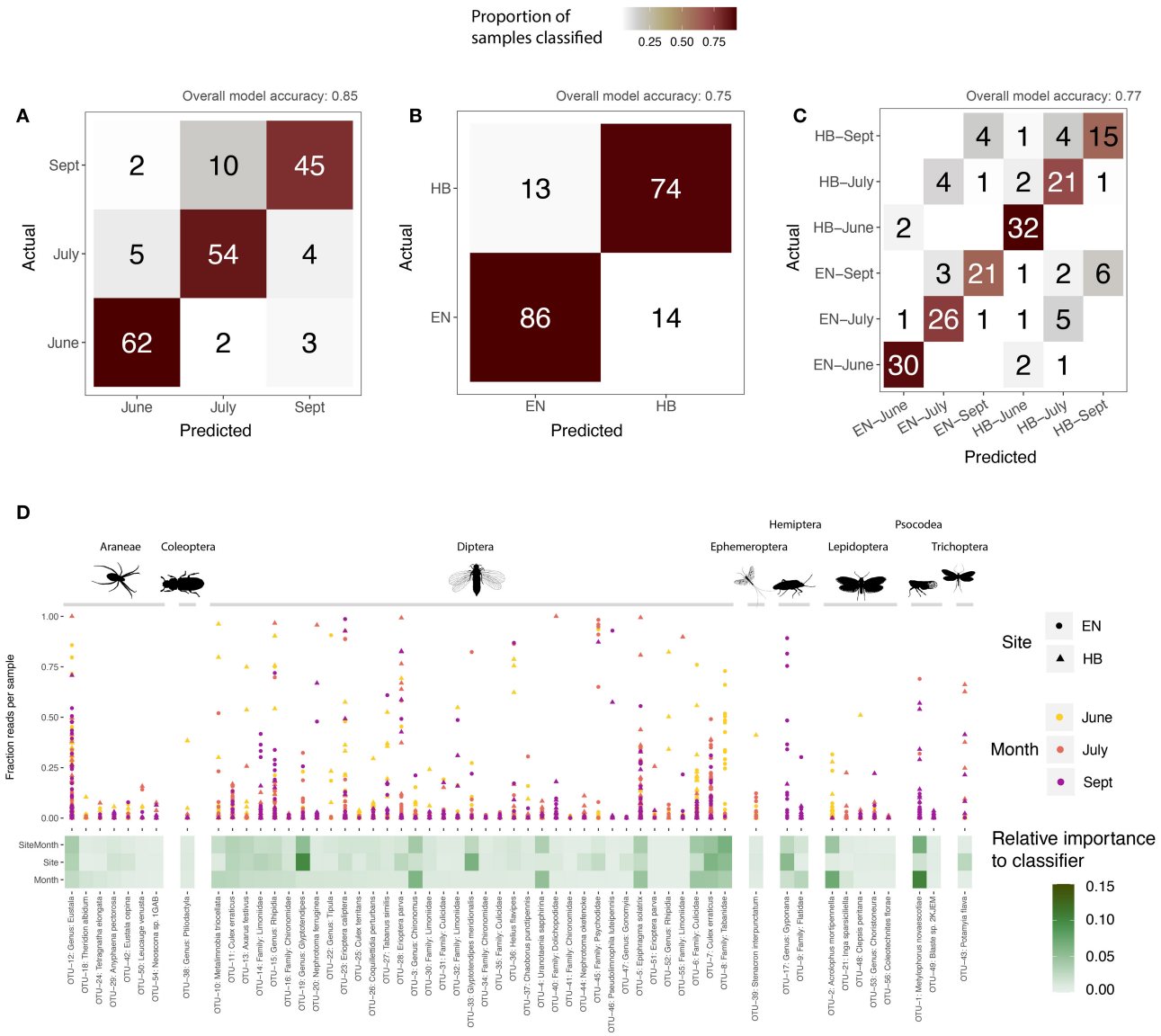

FIGURE 4 | Supervised learning of core sequence clusters (OTUs) using Random Forest classifiers. Confusion matrices illustrate model accuracy, showing the number of samples predicted for each group: **(A)** sampling month; **(B)** sampling site (Egner, "EN"; Hickory Bottoms, "HB"); **(C)** sampling site and month. Overall accuracy for each model shown at top of confusion matrix. The relative importance of each OTU, for each model, is shown in **(D)**, along with the proportion of sequences comprised of that OTU for each sample, defined by its sampling site and month.

location—suggesting that Indiana bats consume largely moths and beetles (Feldhamer et al., 2009). While it is probable that these differences are partly due to prey availability in the different sites, it is also likely that the interpretation of Indiana bat diet is influenced by the analytical tools applied.

Guano samples were collected in June, July, and September—periods aligning with the timing of parturition, weaning, and fall migration, respectively (Humphrey et al., 1977). Previous visual identification of Indiana bat guano contents from maternity colonies in Indiana demonstrated temporal shifts in diet, with increasing lepidopterans and decreasing trichopterans from June through August (Brack, 1983). Likewise, molecular diet analyses of Little brown bat maternity colonies demonstrated seasonal changes in diet (Clare et al., 2011), progressing from dipteran to lepidopteran taxa from May through September. However, we found little evidence of substantial change in diet composition across the foraging season. Instead, particular taxa were detected throughout the entire foraging season: Culicidae, Limoniidae, and Chironomidae families in the dipteran order, as well as an orb-weaving spider in the genus *Eustala*. The lack of seasonal turnover in the most frequently detected prey is likely a consequence of the proximity of the roost sites to the Cache River, and a reflection of the robust aquatic dipteran taxa available throughout the foraging season. It appears that positioning these artificial roosts within a riparian habitat—a preferred landscape for Indiana bat maternity colonies (Humphrey et al., 1977;

Garner and Gardner, 1992) but not necessarily male Indiana bats (LaVal et al., 1977)—is both sufficient for recruiting Indiana bats as well as promoting local foraging.

It is unclear whether the relatively higher proportion of dipteran and aranean-classified sequence counts are a reflection of foraging preference (i.e., biomass of prey) or an artifact of experimental design. Incorporating abundance information into fecal analyses is challenging for several reasons, including different digestion rates of arthropod prey or DNA extraction biases (Deagle et al., 2019). Observed differences in sequencing depths can also be impacted by the particular molecular tools applied. For example, *in silico* analyses (Clarke et al., 2014) and empirical tests (Braukmann et al., 2019; Jusino et al., 2019) suggest that primer choice can influence observed taxonomic diversity, as can the various choices of sequencing platform and depth of coverage (Braukmann et al., 2019), or sequence processing software (O'Rourke et al., 2020). Nevertheless, the primers we employed in this experiment were previously tested using biological mock communities and indicated only minor bias among particular arthropod orders [see Figure 1 in Jusino et al. (2019)]. Interestingly, these previously reported biases lead to marginally greater identification of coleopteran and lepidopteran sequences rather than dipteran, making it unlikely that our frequently detected spider and fly sequences are a result of preferential template binding. Thus, it does not appear that the relatively large fraction of fly and spider taxa we detected is due to a particular molecular bias.

Furthermore, we observed high proportions of sequences for individual samples among multiple core arthropod orders including OTUs classified to Araneae, Ephemeroptera, Lepidoptera, Psocodea, and Trichoptera (**Figure 4D**). Therefore, both in terms of sequence abundances and in terms of frequency of detection, the core prey items identified in these Indiana bats are congruent. Additionally, while our study may differ in prevalence of the most frequently detected arthropod orders, our work concurs with previous diet studies (for a helpful summary, see Lewis, 2007) describing Indiana bats as engaging in aerial foraging activity. Interestingly, this likely applies even to the prevalent spider detected in our study, which was classified to the genus, *Eustala*, and is known for ballooning behavior (Bell et al., 2005). Perhaps, as has been previously suggested in other bat species (Segura-Trujillo et al., 2016; Wray et al., 2020), these Indiana bats are more aptly characterized as arthropodivores.

Despite these molecular tools confirming and expanding the historical understanding of Indiana bat diets, using these data to inform actionable management practices requires further consideration regarding whether or not the relative abundances of sequences are applied in the analysis. In a presence-absence context, we find significant differences with respect to observed richness between sampling sites and dates, whereas an abundance-based measure of diversity via Shannon's entropy suggested no such difference (**Figure 2**). If the management goal was to identify priority conservation sites to optimize foraging success, and we considered optimal locations in areas where a more diverse set of taxa are available, the two frameworks may lead to alternative actions. A presence-absence context would suggest placing greater priority on sites in the Egner tract over the Hickory Bottoms tract (i.e., Egner had higher overall observed richness for each sampling month). However, a relative abundance context indicates that all sites and locations are equally useful, and no additional prioritization would be necessary. Incorporating abundance information was also a relevant factor when interpreting whether sampling site or date affects community composition. A greater proportion of variance was captured in the first two principal component axes when abundance information was applied (**Figure 3**). Analyzing these data in a presence-absence context would again imply significant site and date differences, whereas abundance-based measures point to far greater overlap in spatial and temporal dimensions.

However, these data are interpreted, our molecular diet analysis concurs with earlier work advocating for the protection of the wetland and riparian habitat of the Cypress Creek National Wildlife Refuge because of its critical role in supporting Indiana bat foraging [in particular, see Chapter 4 of Brown and Melius (2014)]. The artificial roosts used in this study were positioned between aquatic and agricultural environments (**Figure 1**), thus it was possible that a variety of taxa found in both landscapes might be routinely detected in our data. Instead, the majority of the Indiana bat samples contained dipterans like craneflies, mosquitoes, and non-biting midges, as well as caddisflies, mayflies, and other aquatic invertebrates known to inhabit the Cache River area. Furthermore, these core taxa—sequence clusters identified in at least 10% of samples—are dominated by aquatic insects (**Figure 4D**). Few of these core diet components were important to the supervised learning models built to classify samples to a particular site or date, indicating that there is an extensive dietary overlap in both season and location among these regularly consumed taxa (**Figure 4D**). Notably, the sequence clusters important to a given model often fit an expected life history for the organism. For example, populations of barklice *M. novaescotiae* are known to build throughout the season and emerge as adults on the wing in large cohorts in late summer (M. Jeffords, Personal communication).

Because our study did not conduct insect trapping at the time of guano collection it is unclear to what extent differences in spatial or temporal variability are due to selective foraging or prey availability. Clarifying such distinctions can further inform management criteria. For example, if these Indiana bats are largely selective toward particular aquatic taxa, those aquatic habitats are likely of conservation interest. Yet thoroughly sampling the available prey—particularly for a mobile and generalist consumer like the Indiana bat—is an intensive task that was beyond the scope of our study. Nevertheless, our molecular methods have identified a broad range of taxa that can assist future studies when determining what trap types are necessary to accurately capture the true extent of abundance and distribution of available prey. Indeed, a recent molecular diet study of the Little brown bat, *M. lucifugus*, found prey abundance was generally unrelated to prey consumption (Wray et al., 2020), however the authors note that their black-light trapping method likely was unable to attract certain taxa. Given the propensity for these Indiana bats to consume orb-weaving spiders, as well as some dipteran (e.g., Limoniidae) and ephemeropteran species, a combination of trap types are likely necessary to properly survey the prey items available to Indiana bats in Cypress Creek.

As with many wildlife conservation challenges, the best plans will have strong partnerships with a variety of stakeholders (Mosher et al., 2020). Molecular methods are a valuable addition to understanding the foraging requirements of the Indiana bat, but are most valuable when contextualized with contributions from land managers, field ecologists, and wildlife experts. We hope our wet bench and bioinformatic methods offer a template to bring the molecular tools into the discussion of future conservation management plans.

DATA AVAILABILITY STATEMENT

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

AUTHOR CONTRIBUTIONS

Conceptualization performed by MTM, KM, and JF. Guano collection by KM and MTM. Data curation, Bioinformatics, Formal analysis, Visualization, Writing—original draft, was performed by DO'R. Writing—review & editing performed by NB, JF, MTM, KM, and MDM. All authors contributed to the article and approved the submitted version.

ACKNOWLEDGMENTS

We thank Ryelan McDonough for performing DNA extraction and preparing libraries for sequencing, and Katy Parise for multiple levels of data management and troubleshooting. We also thank Mike Jeffords, Illinois Natural History Survey, for his perspectives describing particular arthropod life histories in the Cache River area.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fevo.2021.623655/full#supplementary-material

REFERENCES

Adams, J., Roby, P., Schwierjohann, J., Gumbert, M., and Brandenburg, M. (2015). Success of Brandenbark^{TM}, an artificial roost structure designed for use by Indiana Bats (*Myotis sodalis*). *J. Am. Soc. Min. Reclam.* 4, 1–15. doi: 10.21000/JASMR15010001

Alberdi, A., Aizpurua, O., Bohmann, K., Gopalakrishnan, S., Lynggaard, C., Nielsen, M., et al. (2019). Promises and pitfalls of using high-throughput sequencing for diet analysis. *Mol. Ecol. Resour.* 19, 327–348. doi: 10.1111/1755-0998.12960

Alberdi, A., Aizpurua, O., Gilbert, M. T. P., and Bohmann, K. (2018). Scrutinizing key steps for reliable metabarcoding of environmental samples. *Methods Ecol. Evol.* 9, 134–147. doi: 10.1111/2041-210X.12849

Bates, D., and Maechler, M. (2018). *Matrix: Sparse and Dense Matrix Classes and Methods*.

Bell, J. R., Bohan, D. A., Shaw, E. M., and Weyman, G. S. (2005). Ballooning dispersal using silk: world fauna, phylogenies, genetics and models. *Bull. Entomol. Res.* 95, 69–114. doi: 10.1079/BER2004350

Bisanz, J. E. (2018). *qiime2R: Importing QIIME2 Artifacts and Associated Data into R Sessions*.

Bokulich, N., Dillon, M., Bolyen, E., Kaehler, B., Huttley, G., and Caporaso, J. (2018b). q2-sample-classifier: machine-learning tools for microbiome classification and regression. *J. Open Source Softw.* 3:934. doi: 10.21105/joss.00934

Bokulich, N. A., Collins, T. S., Masarweh, C., Allen, G., Heymann, H., Ebeler, S. E., et al. (2016). Associations among wine grape microbiome, metabolome, and fermentation behavior suggest microbial contribution to regional wine characteristics. *mBio* 7:e00631-16. doi: 10.1128/mBio.00631-16

Bokulich, N. A., Kaehler, B. D., Rideout, J. R., Dillon, M., Bolyen, E., Knight, R., et al. (2018a). Optimizing taxonomic classification of marker-gene amplicon sequences with QIIME 2's q2-feature-classifier plugin. *Microbiome* 6:90. doi: 10.1186/s40168-018-0470-z

Bolyen, E., Rideout, J. R., Dillon, M. R., Bokulich, N. A., Abnet, C. C., Al-Ghalith, G. A., et al. (2019). Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nat. Biotechnol.* 37, 852–857. doi: 10.1038/s41587-019-0209-9

Brack, V. W. (1983). *The Nonhibernating Ecology of Bats in Indiana With Emphasis on the endangered Indiana Bat, Myotis Sodalis* (Dissertation). Purdue University, West Lafayette, IN, USA.

Brady, J. T., LaVal, R. K., Kunz, T. H., Tuttle, M. D., Wilson, D. E., and Clawson, R. L. (1983). *Recovery plan for the Indiana Bat. United States Fish and Wildlife Service*. Available online at: https://www.fws.gov/midwest/endangered/mammals/inba/pdf/inba_recplan.pdf (accessed December 31, 2020).

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

Breiman, L. (2001). Random forests. *Mach. Learn.* 45, 5–32. doi: 10.1023/A:1010933404324

Brown, M., and Melius, T. O. (2014). *Cypress Creek National Wildlife Refuge Habitat Management Plan*. Cypress Creek National Wildlife Refuge Ullin, IL: United States Fish and Wildlife Service. Available online at: https://www.fws.gov/WorkArea/DownloadAsset.aspx?id=2147557022 (accessed December 31, 2020).

Cable, A. B., O'Keefe, J. M., Deppe, J. L., Hohoff, T. C., Taylor, S. J., and Davis, M. A. (2020). Habitat suitability and connectivity modeling reveal priority areas for Indiana bat (*Myotis sodalis*) conservation in a complex habitat mosaic. *Landsc. Ecol.* doi: 10.1007/s10980-020-01125-2

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

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., et al. (2009). BLAST+: architecture and applications. *BMC Bioinformatics* 10:421. doi: 10.1186/1471-2105-10-421

Chamberlain, S., Szocs, E., Boettiger, C., Ram, K., Bartomeus, I., Baumgartner, J., et al. (2014). *Taxize: Taxonomic Information from Around the Web*. Available online at: https://cran.r-project.org/web/packages/taxize/index.html (accessed September 17, 2020).

Chamberlain, S. A., and Szöcs, E. (2013). Taxize: taxonomic search and retrieval in R. *F1000Research* 2:191. doi: 10.12688/f1000research.2-191.v1

Clare, E. L., Barber, B. R., Sweeney, B. W., Hebert, P. D. N., and Fenton, M. B. (2011). Eating local: influences of habitat on the diet of little brown bats (*Myotis lucifugus*): molecular detection of variation in diet. *Mol. Ecol.* 20, 1772–1780. doi: 10.1111/j.1365-294X.2011.05040.x

Clare, E. L., Fraser, E. E., Braid, H. E., Fenton, M. B., and Hebert, P. D. N. (2009). Species on the menu of a generalist predator, the eastern red bat (*Lasiurus borealis*): using a molecular approach to detect arthropod prey. *Mol. Ecol.* 18, 2532–2542. doi: 10.1111/j.1365-294X.2009.04184.x

Clare, E. L., Symondson, W. O. C., Broders, H., Fabianek, F., Fraser, E. E., MacKenzie, A., et al. (2014). The diet of *Myotis lucifugus* across Canada: assessing foraging quality and diet variability. *Mol. Ecol.* 23, 3618–3632. doi: 10.1111/mec.12542

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

Dahl, C., Stanislawski, M., Iszatt, N., Mandal, S., Lozupone, C., Clemente, J. C., et al. (2017). Gut microbiome of mothers delivering prematurely shows reduced diversity and lower relative abundance of Bifidobacterium and Streptococcus. *PLoS ONE* 12:e0184336. doi: 10.1371/journal.pone.0184336

Deagle, B. E., Thomas, A. C., McInnes, J. C., Clarke, L. J., Vesterinen, E. J., Clare, E. L., et al. (2019). Counting with DNA in metabarcoding studies: how should we convert sequence reads to dietary data? *Mol. Ecol.* 28, 391–406. doi: 10.1111/mec.14734

Dodd, L. E., Chapman, E. G., Harwood, J. D., Lacki, M. J., Rieske, L. K., and Stevens, R. D. (2012). Identification of prey of *Myotis septentrionalis* using DNA-based techniques. *J. Mammal.* 93, 1119–1128. doi: 10.1644/11-MAMM-A-218.1

Feldhamer, G. A., Carter, T. C., and Whitaker, J. O. (2009). Prey consumed by eight species of insectivorous bats from Southern Illinois. *Am. Midl. Nat.* 162, 43–51. doi: 10.1674/0003-0031-162.1.43

Frick, W. F., Pollock, J. F., Hicks, A. C., Langwig, K. E., Reynolds, D. S., Turner, G. G., et al. (2010). An emerging disease causes regional population collapse of a common North American bat species. *Science* 329, 679–682. doi: 10.1126/science.1188594

Garner, J. D., and Gardner, J. E. (1992). Determination of Summer *Distribution and Habitat Utilization of the Indiana Bat (Myotis sodalis) in Illinois. Division of Natural Heritage Illinois Department of Conservation; Center for Biogeographic Information (Illinois Natural History Survey)*. Available online at: https://www.ideals.illinois.edu/bitstream/handle/2142/10287/inhscbiv01992i00002_opt.pdf (accessed December 31, 2020).

Garnier, S. (2018). *viridis: Default Color Maps from "matplotlib."*

Gerhard, W. A., and Gunsch, C. K. (2019). Metabarcoding and machine learning analysis of environmental DNA in ballast water arriving to hub ports. *Environ. Int.* 124, 312–319. doi: 10.1016/j.envint.2018.12.038

Graves, S., Piepho, H.-P., and Selzer, L. (2019). *multcompView: Visualizations of Paired Comparisons*. Available online at: https://CRAN.R-project.org/package=multcompView (accessed September 17, 2020).

Hope, P. R., Bohmann, K., Gilbert, M. T. P., Zepeda-Mendoza, M., Razgour, O., and Jones, G. (2014). Second generation sequencing and morphological faecal analysis reveal unexpected foraging behaviour by *Myotis nattereri* (Chiroptera, Vespertilionidae) in winter. *Front. Zool.* 11:39. doi: 10.1186/1742-9994-11-39

Humphrey, S. R., Richter, A. R., and Cope, J. B. (1977). Summer habitat and ecology of the endangered Indiana bat, *Myotis sodalis*. *J. Mammal.* 58, 334–346. doi: 10.2307/1379332

Jachowski, D. S., Dobony, C. A., Coleman, L. S., Ford, W. M., Britzke, E. R., and Rodrigue, J. L. (2014). Disease and community structure: white-nose syndrome alters spatial and temporal niche partitioning in sympatric bat species. *Divers. Distrib.* 20, 1002–1015. doi: 10.1111/ddi.12192

Jachowski, D. S., Rota, C. T., Dobony, C. A., Ford, W. M., and Edwards, J. W. (2016). Seeing the forest through the trees: considering roost-site selection at multiple spatial scales. *PLoS ONE* 11:e0150011. doi: 10.1371/journal.pone.0150011

Johnson, C. M., and King, R. A. (2018). *Beneficial Forest Management Practices for WNS-affected Bats: Voluntary Guidance for Land Managers and Woodland Owners in the Eastern United States. White-nose Syndrome Conservation and Recovery Working Group*. Available online at: https://s3.us-west-2.amazonaws.com/prod-is-cms-assets/wns/prod/393b6360-f27c-11e8-87d8-09a30749711d-Final_Forestry_BFMPs_May31-2018.docx (accessed December 31, 2020).

Jusino, M. A., Banik, M. T., Palmer, J. M., Wray, A. K., Xiao, L., Pelton, E., et al. (2019). An improved method for utilizing high-throughput amplicon sequencing to determine the diets of insectivorous animals. *Mol. Ecol. Resour.* 19, 176–190. doi: 10.1111/1755-0998.12951

Kahle, D., and Wickham, H. (2013). ggmap: spatial visualization with ggplot2. *R J.* 5, 144–161. doi: 10.32614/RJ-2013-014

Kassambara, A. (2018). *ggpubr: "ggplot2" Based Publication Ready Plots*.

King, A. (2019). *2019 Indiana Bat (Myotis sodalis) Population Status Update*. Indiana Ecological Services Field Office: United States Fish and Wildlife Service. Available online at: https://www.fws.gov/midwest/Endangered/mammals/inba/pdf/2019_IBat_Pop_Estimate_6_27_2019a.pdf (accessed December 31, 2020).

Kunz, T. H., and Whitaker, J. O. (1983). An evaluation of fecal analysis for determining food habits of insectivorous bats. *Can. J. Zool.* 61, 1317–1321. doi: 10.1139/z83-177

LaVal, R. K., Clawson, R. L., LaVal, M. L., and Caire, W. (1977). Foraging behavior and nocturnal activity patterns of Missouri bats, with emphasis on the endangered species *Myotis grisescens* and *Myotis sodalis*. *J. Mammal.* 58, 592–599. doi: 10.2307/1380007

Lewis, L. (2007). *Indiana Bat (Myotis sodalis) Draft Recovery Plan: First Revision*. Fort Snelling: United States Fish and Wildlife Service. Available online at: https://www.govinfo.gov/content/pkg/FR-2007-04-16/pdf/07-1866.pdf (accessed December 31, 2020).

Lorch, J. M., Meteyer, C. U., Behr, M. J., Boyles, J. G., Cryan, P. M., Hicks, A. C., et al. (2011). Experimental infection of bats with *Geomyces destructans* causes white-nose syndrome. *Nature* 480, 376–378. doi: 10.1038/nature10590

Lumley, T. (2016). *xkcdcolors: Color Names from the XKCD Color Survey*. Available online at: https://CRAN.R-project.org/package=xkcdcolors (accessed September 17, 2020).

Mangan, K., and Mangan, M. T. (2017). *Bat Use of Artificial Roosting Structures (3rd Annual Report)*. Southern Illinois Sub-Office, Marion, IL: United States Fish and Wildlife Service.

Mangan, K., and Mangan, M. T. (2019a). *2016 Indiana Bat Survey Report*. Southern Illinois Ecological Services Sub-Office, Marion, IL: United States Fish and Wildlife Service. Available online at: https://www.fws.gov/midwest/endangered/mammals/inba/surveys/pdf/2016IndianaBatSummerSurveyGuidelines11April2016.pdf (accessed December 31, 2020).

Mangan, K., and Mangan, M. T. (2019b). *2018 Indiana Bat Survey Report*. Southern Illinois Ecological Services Sub-Office, Marion, IL: United States Fish and Wildlife Service.

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. J.* 17, 10–12. doi: 10.14806/ej.17.1.200

McDonald, D., Clemente, J. C., Kuczynski, J., Rideout, J. R., Stombaugh, J., Wendel, D., et al. (2012). The biological observation matrix (BIOM) format or: how I learned to stop worrying and love the ome-ome. *GigaScience* 1:7. doi: 10.1186/2047-217X-1-7

McKinney, W. (2010). "Data structures for statistical computing in python," in *Proceedings of the 9th Python in Science Conference* (Austin, TX), 56–61. doi: 10.25080/Majora-92bf1922-00a

Menzel, J. M., Ford, W. M., Menzel, M. A., Gardner, J. E., Garner, J. D., and Hofmann, J. E. (2005). Summer habitat use and home-range analysis of the endangered Indiana bat. *J. Wildl. Manag.* 69, 430–436. doi: 10.2193/0022-541X(2005)069<0430:SHUAHA>2.0.CO;2

Mosher, B. A., Bernard, R. F., Lorch, J. M., Miller, D. A., Richgels, K. L., White, C. L., et al. (2020). Successful molecular detection studies require clear communication among diverse research partners. *Front. Ecol. Environ.* 18:2141. doi: 10.1002/fee.2141

Murray, S. W., and Kurta, A. (2004). Nocturnal activity of the endangered Indiana bat (*Myotis sodalis*). *J. Zool.* 262, 197–206. doi: 10.1017/S0952836903004503

Nielsen, J. M., Clare, E. L., Hayden, B., Brett, M. T., and Kratina, P. (2018). Diet tracing in ecology: method comparison and selection. *Methods Ecol. Evol.* 9, 278–291. doi: 10.1111/2041-210X.12869

Ogle, D. H., Wheeler, P., and Dinno, A. (2018). *FSA: Fisheries Stock Analysis*.

O'Rourke, D. R., Bokulich, N. A., Jusino, M. A., MacManes, M. D., and Foster, J. T. (2020). A total crapshoot? Evaluating bioinformatic decisions in animal diet metabarcoding analyses. *Ecol. Evol.* 10, 9721–9739. doi: 10.1002/ece3.6594

O'Shea, T. J., and Bogan, M. A. (2003). *Monitoring Trends in Bat Populations of the United States and Territories: Problems and Prospects*. Fort Collins Science Center: U.S. Geological Survey. Available online at: https://pubs.er.usgs.gov/publication/itr030003 (accessed December 31, 2020).

Oswaldo, S. B. (2019). *ggsn: North Symbols and Scale Bars for Maps Created with "ggplot2" or "ggmap."* Available online at: https://CRAN.R-project.org/package=ggsn (accessed September 17, 2020).

Paradis, E., Claude, J., and Strimmer, K. (2004). APE: analyses of phylogenetics and evolution in R language. *Bioinformatics* 20, 289–290. doi: 10.1093/bioinformatics/btg412

Pedersen, T. L., and Crameri, F. (2018). *scico: Colour Palettes Based on the Scientific Colour-Maps*. Available online at: https://CRAN.R-project.org/package=scico (accessed September 17, 2020).

Pedersen, T. L., and Robinson, D. (2019). *gganimate: A Grammar of Animated Graphics*. Available online at: https://CRAN.R-project.org/package=gganimate (accessed September 17, 2020).

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: machine learning in python. *J. Mach. Learn.* 12, 2825–2830. doi: 10.5555/1953048.2078195

Pompanon, F., Deagle, B. E., Symondson, W. O. C., Brown, D. S., Jarman, S. N., and Taberlet, P. (2012). Who is eating what: diet assessment using next generation sequencing. *Mol. Ecol.* 21, 1931–1950. doi: 10.1111/j.1365-294X.2011.05403.x

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

Price, M. N., Dehal, P. S., and Arkin, A. P. (2010). FastTree 2 – Approximately maximum-likelihood trees for large alignments. *PLoS ONE* 5:e9490. doi: 10.1371/journal.pone.0009490

R Core Team (2018). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing. Available online at: https://www.R-project.org/ (accessed December 31, 2020).

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

Ren, K., and Russel, K. (2016). *formattable: Create "Formattable" Data Structures*. Available online at: https://CRAN.R-project.org/package=formattable (accessed September 17, 2020).

Robeson, M., O'Rourke, D. R., Kaehler, B., Ziemski, M., Dillon, M. R., Foster, J. T., et al. (2020). *RESCRIPt: Reference Sequence Annotation and Curation Pipeline*. doi: 10.5281/zenodo.3891931

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

Roguet, A., Eren, A. M., Newton, R. J., and McLellan, S. L. (2018). Fecal source identification using random forest. *Microbiome* 6:185. doi: 10.1186/s40168-018-0568-3

Segura-Trujillo, C. A., Lidicker, W. Z. Jr., and Álvarez-Castañeda, S. T. (2016). New perspectives on trophic guilds of arthropodivorous bats in North and Central America. *J. Mammal.* 97, 644–654. doi: 10.1093/jmammal/gyv212

Slowikowski, K. (2018). *ggrepel: Automatically Position Non-Overlapping Text Labels with "ggplot2."*

Sparks, D. W., Ritzi, C. M., Duchamp, J. E., and Whitaker, J. O. (2005). Foraging habitat of the Indiana bat (*Myotis sodalis*) at an urban-rural Interface. *J. Mammal.* 86, 713–718. doi: 10.1644/1545-1542(2005)0860713:FHOTIB2.0.CO;2

Thogmartin, W. E., King, R. A., McKann, P. C., Szymanski, J. A., and Pruitt, L. (2012). Population-level impact of white-nose syndrome on the endangered Indiana bat. *J. Mammal.* 93, 1086–1098. doi: 10.1644/11-MAMM-A-355.1

Thomson, C. E. (1982). Myotis sodalis. *Mamm. Species* 163, 1–5. doi: 10.2307/3504013

Turner, G. G., Reeder, D. M., and Coleman, J. T. H. (2011). A five-year assessment of mortality and geographic spread of white-nose syndrome in North American bats and a look to the future. *Bat Res. News* 52, 13–27.

Tuttle, N. M., Benson, D. P., and Sparks, D. W. (2006). Diet of the *Myotis sodalis* (Indiana Bat) at an urban/rural interface. *Northeast. Nat.* 13, 435–442. doi: 10.1656/1092-6194(2006)13435:DOTMSI2.0.CO;2

Udall, S. L. (1967). Endangered Species Act. *Fed. Regist.* 32. Available online at: https://www.fws.gov/endangered/species/pdf/32FedReg_pg%204001.pdf (accessed December 31, 2020).

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

Vesterinen, E. J., Ruokolainen, L., Wahlberg, N., Peña, C., Roslin, T., Laine, V. N., et al. (2016). What you need is what you eat? Prey selection by the bat *Myotis daubentonii*. *Mol. Ecol.* 25, 1581–1594. doi: 10.1111/mec.13564

Warnecke, L., Turner, J. M., Bollinger, T. K., Lorch, J. M., Misra, V., Cryan, P. M., et al. (2012). Inoculation of bats with European *Geomyces destructans* supports the novel pathogen hypothesis for the origin of white-nose syndrome. *Proc. Natl. Acad. Sci. U.S.A.* 109, 6999–7003. doi: 10.1073/pnas.1200374109

Weiss, S., Xu, Z. Z., Peddada, S., Amir, A., Bittinger, K., Gonzalez, A., et al. (2017). Normalization and microbial differential abundance strategies depend upon data characteristics. *Microbiome* 5:27. doi: 10.1186/s40168-017-0237-y

Wickham, H. (2007). Reshaping data with the reshape package. *J. Stat. Softw.* 21, 1–20. doi: 10.18637/jss.v021.i12

Wickham, H. (2017). *tidyverse: Easily Install and Load the "Tidyverse."*

Wickham, H. (2018). *scales: Scale Functions for Visualization*.

Wickham, H. (2019). *rvest: Easily Harvest (Scrape) Web Pages*. Available online at: https://CRAN.R-project.org/package=rvest (accessed September 17, 2020).

Wickham, H., Henry, L., Pedersen, T. L., Luciani, T. J., Decorde, M., and Lise, V. (2020). *svglite: An "SVG" Graphics Device*. Available online at: https://CRAN.R-project.org/package=svglite (accessed September 17, 2020).

Wilke, C. O. (2017). *cowplot: Streamlined Plot Theme and Plot Annotations for "ggplot2."*

Womack, K. M., Amelon, S. K., and Thompson, F. R. (2013). Resource selection by Indiana bats during the maternity season. *J. Wildl. Manag.* 77, 707–715. doi: 10.1002/jwmg.498

Wray, A. K., Peery, M. Z., Jusino, M., Kochanski, J. M., Banik, M. T., Palmer, J. M., et al. (2020). Predator preferences shape the diets of arthropodivorous bats more than quantitative local prey abundance. *Mol. Ecol.* doi: 10.1111/mec.15769

Zeale, M. R. K., Butlin, R. K., Barker, G. L. A., Lees, D. C., and Jones, G. (2011). Taxon-specific PCR for DNA barcoding arthropod prey in bat faeces: DNA barcoding. *Mol. Ecol. Resour.* 11, 236–244. doi: 10.1111/j.1755-0998.2010.02920.x

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: 25 May 2021
doi: 10.3389/fevo.2021.624109

Four Species Linked by Three Hybrid Zones: Two Instances of Repeated Hybridization in One Species Group (Genus *Liolaemus*)

Jared A. Grummer[1*†], *Luciano J. Avila*[2], *Mariana M. Morando*[2] *and Adam D. Leaché*[1]

[1] *Department of Biology, Burke Museum of Natural History and Culture, University of Washington, Seattle, WA, United States,*
[2] *Centro Nacional Patagónico–Consejo Nacional de Investigaciones Científicas y Técnicas, Puerto Madryn, Argentina*

OPEN ACCESS

Edited by:
Michael G. Campana, Smithsonian Conservation Biology Institute (SI), United States

Reviewed by:
Leo Joseph, CSIRO, Australia
Tyler Chafin, University of Arkansas, United States

***Correspondence:**
Jared A. Grummer
grummer@zoology.ubc.ca

†Present address:
Jared A. Grummer, Department of Zoology, Biodiversity Research Centre, University of British Columbia, Vancouver, BC, Canada

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

Received: *30 October 2020*
Accepted: *28 April 2021*
Published: *25 May 2021*

Citation:
Grummer JA, Avila LJ, Morando MM and Leaché AD (2021) Four Species Linked by Three Hybrid Zones: Two Instances of Repeated Hybridization in One Species Group (Genus Liolaemus). Front. Ecol. Evol. 9:624109. doi: 10.3389/fevo.2021.624109

Hybridization is an evolutionary process that can generate diverse outcomes, such as reinforcing species boundaries, generating new species, or facilitating the introgression of locally-adapted alleles into new genomic backgrounds. *Liolaemus* is a highly diverse clade of South American lizards with ∼260 species and as many as ten new species are described each year. Previous *Liolaemus* studies have detected gene flow and introgression among species using phylogenetic network methods and/or through comparisons of nuclear and mitochondrial DNA patterns, yet no study has systematically studied hybrid zones between *Liolaemus* species. Here, we compared three hybrid zones between four species in the *Liolaemus fitzingerii* group of lizards in Central Argentina where two species, *L. melanops* and *L. xanthoviridis*, each hybridize with two other species (*L. shehuen* and *L. fitzingerii*). We sampled three transects that were each ∼120 km in length and sequenced both mitochondrial and genome-wide SNP data for 267 individuals. In our analyses of nuclear DNA, we also compared bi-allelic SNPs to phased alleles (50 bp RAD loci). Population structure analyses confirmed that boundaries separating species are sharp, and all clines are <65 km wide. Cline center estimates were consistent between SNPs and phased alleles, but cline width estimates were significantly different with the SNPs producing wider estimates. The mitochondrial clines are narrower and shifted 4–20 km southward in comparison to the nuclear clines in all three hybrid zones, indicating that either each of the species has sex-biased dispersal (males northward or females southward), the population densities are unequal, or that the hybrid zones are moving north over time. These comparisons indicate that some patterns of hybridization are similar across hybrid zones (mtDNA clines all narrower and shifted to the south), whereas cline width is variable. Hybridization in the *L. fitzingerii* group is common and geographically localized; further studies are needed to investigate whether hybrid zones act as hard species boundaries or promoters of speciation through processes such as reinforcement. Nonetheless, this study provides insights into both biotic and abiotic mechanisms helping to maintain species boundaries within the speciose *Liolaemus* system.

Keywords: SNP, admixture, lizard, population, cline, Patagonia (Argentina), ddRAD

1. INTRODUCTION

Hybridization, or interbreeding between distinct populations, has captivated evolutionary biologists for nearly two centuries (Darwin, 1862; Harrison, 1993). It can be a means of transferring adaptive genetic diversity between lineages (Chhatre et al., 2018; Hanemaaijer et al., 2018), forming hybrid swarms and potentially collapsing lineages (Pritchard and Edmands, 2013), or conversely, creating new species through hybrid speciation (e.g., Rieseberg et al., 1995). Hybridization is indeed common across the tree of life, with documentation in 10% of animal species and 25% of plant species (Mallet, 2005). Within Tetrapods, hybridization is particularly common in squamate reptiles—the lizards and snakes (Jančúchová-Lásková et al., 2015). Given the diverse roles that hybridization can play in shaping patterns of diversity, it is important to deepen our understanding of this process in natural systems.

Hybrid zones can provide detailed information about the evolutionary and ecological interactions between species, and replicated hybrid zones offer the additional advantage of investigating the repeatability of evolutionary processes (McKinnon and Rundle, 2002). Replicate transects across a single hybrid zone can offer insights into the extrinsic and intrinsic factors that govern the dynamics of a hybrid zone (Brelsford and Irwin, 2009; Zieliński et al., 2019; Westram et al., 2021, e.g.,). Replicate hybrid zones have mainly been studied in fish, which typically show a high amount of variability of introgression rates and genomic divergence between different hybrid zones. In *Xiphophorus* swordtail fish, Culumber et al. (2011) found that linkage disequilibrium and Hardy-Weinberg equilibrium estimates varied substantially across seven transects. Nolte et al. (2009) also found little correlation in genomic differentiation between two hybrid zones of sculpin fish (*Cottus*). And similarly, hybridization rates were found to vary considerably across ten topminnow (*Fundulus*) replicate hybrid zones (Duvernell and Schaefer, 2014). These differences identified across replicate hybrid zones are typically ascribed to distinct environments that characterize each hybrid zone (Aboim et al., 2010). In this study, we investigate replicate hybrid zones in a species group of liolaemid lizards. Here, we use the term "replicate" not in the statistical sense, but to indicate that one species hybridizes with more than one other distinct species and thus represents "evolutionary replicates" given that the process of hybridization has occurred multiple times in distinct geographic areas.

The genus *Liolaemus* (family Liolaemidae) is a particularly diverse clade with ∼260 species and 5–10 new species described each year (http://www.reptile-database.org/). Some authors have recently posited that hybridization may be one of the factors responsible for generating the exceptional diversity within this clade, particularly when compared to its sister clade *Phymaturus* that only has 48 species (Olave et al., 2018, 2020; Morando et al., 2020). Indeed, several studies have detected or suggested hybridization in disparate *Liolaemus* groups including the *lineomaculatus* series (Breitman et al., 2011) and *leopardinus* clade (Esquerré et al., 2019), the *darwinii*, *kriegi*, and *petrophilus* complexes (Morando et al., 2004; Feltrin, 2013; Medina et al., 2014), and the *chiliensis* and *fitzingerii* groups (Avila et al., 2006; Grummer et al., 2018; Araya-Donoso et al., 2019).

In most cases, hybridization is inferred through incongruence of mitochondrial and nuclear phylogenies and/or morphological species designations, given the contrasting inheritance modes of the two genomes (Ballard and Whitlock, 2004). Furthermore, instances of hybridization are typically localized to areas where two distinct populations or species come into contact, "hybrid zones." It has been suggested that hybridization can play two important roles within *Liolaemus*: (1) increasing genetic and adaptive diversity following population bottlenecks, and (2) limiting specialization to maintain a generalist phenotype that is better suited to heterogeneous and unstable habitats, such as those in southern South America (Morando et al., 2020; Olave et al., 2020). Although hybridization is suspected to be relatively common in *Liolaemus*, detailed examinations of hybrid zones using thorough transect sampling and genomic data analyses are lacking.

Hybrid zones form at the interface between two distinct populations and in some cases are best described as "clines," which represent transitions in observed character states between distinct populations (Barton and Hewitt, 1985). Clines inferred from different characters that share the same center are said to be coincident, and those that share the same shape/width are said to be concordant. Clines and contact zones are often formed in ecotones where two distinct habitats fuse (Leaché and Cole, 2007). These contact zones typically occur in one area between species and therefore offer a single perspective into the evolutionary process. However, some species complexes have established themselves into loosely formed "rings" (or perhaps more aptly, horseshoes) around unsuitable habitat, where species grade into each other at contact zones, but the forms are reproductively isolated where the "ring" closes (e.g., *Ensatina* salamanders, Moritz et al., 1992; *Phylloscopus* warblers, Irwin et al., 2001). In other conceptually related instances, "mosaic" hybrid zones can be formed when individuals from distinct species repeatedly come into contact with each other across the landscape (e.g., *Helianthus* sunflowers; Rieseberg et al., 1999). In all of these cases, replicate hybrid zones are formed where one species participates in hybridization in >1 geographic area.

In hybrid zones, neutral and selected markers will respond to hybridization in distinct manners. For instance, because nDNA is biparentally inherited and mtDNA maternally inherited in vertebrates (Ballard and Whitlock, 2004), sex-biased dispersal can be seen by comparing nuclear and mitochondrial patterns in hybrid zones (but see Bonnet et al., 2017 for alternative explanations). Furthermore, many mitochondrial genes code for proteins involved in the electron transport chain and ATP production, making the whole mitochondrial genome subject to selection via linkage. Thus, a beneficial mitochondrial haplotype may sweep to fixation in both populations via selection and gene flow in the hybrid zone. However, some authors have argued for the neutral evolution of the mitochondrial genome with respect to phenotype in some systems (e.g., Rohwer et al., 2001). Assuming that mtDNA is neutrally evolving allows for the estimation of hybrid zone movement, because neutral markers geographically lag behind non-neutral markers (McGuire et al.,

2007). When a hybrid zone moves due to an invading population, the neutral mitochondrial haplotypes will be left in the wake of the invading species (Rohwer et al., 2001). Differing selection pressures and inheritance patterns of nuclear and mitochondrial genomes mean that cline shape and geographic center may in fact be distinct from one another in a given hybrid zone (e.g., Leaché et al., 2017). Depending on the concordance or discordance between nuclear and mitochondrial clines, an inference can be made about hybrid zone movement, the dispersal behavior of the two sexes, or differential selection between nuclear and mitochondrial genomes in the hybrid zone.

Here, we investigated hybrid zones in the *Liolaemus fitzingerii* species group through genome-wide single nucleotide polymorphism (SNP) data. The twelve species belonging to this group are distributed throughout the Patagonian shrub-steppe of central Argentina (Avila et al., 2006, 2008, 2010). However, a phylogenetic analysis using genome-wide SNP data and dense geographic sampling of individuals only found support for six distinct genetic groups, suggesting that species diversity in the group could be overestimated (Grummer, 2017). Nonetheless, the four species studied here are supported by both morphological and SNP data. Two putative contact zones were previously discovered through genomic analyses: one between *L. melanops* and *L. shehuen*, and a second between *L. xanthoviridis* and *L. fitzingerii* (**Figure 1**; Grummer, 2017). These contact zones are further supported by color polymorphism data and the co-occurrence of mtDNA sequences (*cytochrome B* from different species in single populations (Morando and Avila, personal communication). Although multiple lines of evidence support the presence of these hybrid zones, nothing is known regarding their geographic arrangements and limits and therefore the biotic and abiotic processes maintaining them. We studied hybrid zones in the *L. fitzingerii* species group using transect sampling to contrast patterns in both nuclear and mitochondrial genomes through population structure estimation, phylogenetic inference, cline analysis, and network analyses. Our aim is to provide an understanding of evolutionary processes at a fine-scale where the ranges of species come into contact, providing insights into the nature of speciation in a system where species boundaries are porous and blurry.

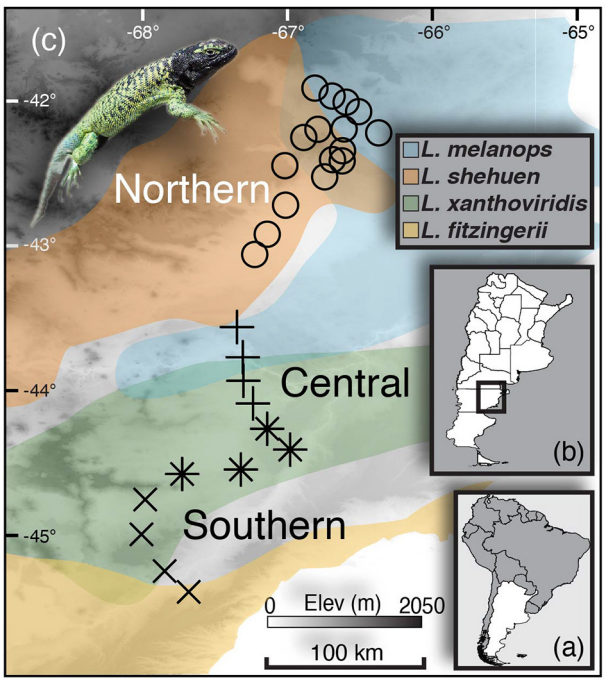

FIGURE 1 | Sampling map with insets showing **(A)** Argentina and **(B)** the provincial borders within Argentina and where the study took place **(C)**. Transect sampling localities for the Northern hybrid zone are shown as circles, whereas the "Central" and "Southern" transects are shown as +s and Xs, respectively (the localities marked with asterisks/stars were analysed in both Central and Southern transects). Colors on the map reflect population boundaries as determined by genome-wide SNP data in Grummer (2017) that largely correspond to the species *Liolaemus melanops* (blue), *L. shehuen* (orange), *L. xanthoviridis* (green), and *L. fitzingerii* (yellow).

2. MATERIALS AND METHODS

2.1. Bioethics

All research specimens were collected by hand using methods approved by the University of Washington Office of Animal Welfare (IACUC protocol number 4249-01) and in accordance with provincial permits from the Argentinean Dirección de Fauna y Flora Silvestre.

2.2. Sampling and DNA Extraction

All voucher specimens and tissues were deposited into the LJAMM-CNP herpetology collection in the Centro Patagónico Nacional (IPEEC-CONICET), Puerto Madryn, Chubut, Argentina. DNA was extracted from tissue (tails tips and liver) through a salt (NaCl) extraction method (MacManes, 2013). Prior to library sequencing preparation, we discarded samples that had low DNA concentration or had degraded genomic DNA that lacked high molecular weight DNA.

2.2.1. Northern Hybrid Zone

During January and December of 2015, we collected 169 individuals from 17 distinct localities in Rio Negro and Chubut provinces (**Figure 1**; **Supplementary Table 1**). Sampling was performed every ∼15–20 km along the transect.

2.2.2. Central and Southern Hybrid Zones

In December 2015, we collected 120 individuals from 13 distinct localities in Chubut province (**Figure 1**; **Supplementary Table 2**). Analyses revealed that what was assumed to be a single hybrid zone in the southern transect in fact represented two hybrid zones (see Results), so we therefore broke up this single transect into a northern ("Central") and southern ("Southern") transect (further detail below).

2.3. DNA Sequence Preparation

2.3.1. Nuclear DNA

We generated a nuclear dataset with the double digestion restriction site-associated DNA sequencing (ddRADseq) approach (Peterson et al., 2012). Genomic DNA was digested

with two enzymes, SbfI (8 bp recognition sequence [5′ CCTGCAGG 3′], "rare cutter"; New England Biolabs, Ipswich, MA) and MspI (4 bp recognition sequence [5′ CCGG 3′], "common cutter"; New England Biolabs, Ipswich, MA). Unique barcoded primers were ligated to all genomic DNA fragments to enable multiplex sequencing. Genomic DNA fragments between ∼365 and 465 bp (415–515 bp after ligating barcoded oligonucleotides) were size-selected with a Blue Pippin DNA fragment size selector (Sage Science, MA, USA). Samples with distinct barcodes were pooled in multiples of eight and unique indexes were applied to each pool using PCR with NEB Phusion Taq polymerase (New England Biolabs Inc., MA, USA) and the following thermocycler conditions: 98° for 0:30, (98° for 0:10, 58° for 0:30, 72° for 0:30) × 12 cycles, and a final 10:00 extension at 72°C. The amplified pools were multiplexed (up to 160 individuals per sequencing lane, some runs with individuals from other studies) and sequenced on Illumina HiSeq 2500 and 4000 machines (Illumina Inc., CA, USA) with 50 bp single-end reads at the University of California Berkeley's QB3 Vincent J. Coates sequencing facility. After de-multiplexing, each read contained 39 bp of sequenced genomic DNA.

2.3.2. Mitochondrial DNA

We targeted a fragment of the *cytochrome B* (*cytB*) gene to sequence for all individuals and contrast with patterns observed from the nuclear genome. Two sets of primers were used, an "external" pair that amplified an ∼800 bp fragment, and an "internal" pair that amplified a ∼360 bp fragment; primer sequences are given in Morando et al. (2003). Twenty-three μL of Tankara EmeraldAmp GT PCR Master Mix (Takara Bio USA, Inc.; Mountain View, CA, USA) were mixed with 2 μL genomic DNA, and amplified with the following thermocycler conditions: 95°C for 5:00, (95° for 0:45, 55° for 0:30, 72° for 1:00) × 35 cycles, with a final 10:00 extension at 72°C. If individuals did not amplify for the larger fragment, we attempted to amplify the smaller fragment with the internal primers. PCR products were sent to Genewiz (South Plainfield, NJ, USA) where they were purified and sequenced in both forward and reverse directions.

2.4. DNA Dataset Assemblies

2.4.1. ddRAD Bioinformatics and Dataset Assembly

Raw sequence reads were processed to generate "clusters" (e.g., loci) and identify SNPs with the program pyRAD v3.0.66 (Eaton, 2014). After demultiplexing individuals using their unique adapter and barcode sequences, pyRAD uses VSEARCH (Rognes et al., 2016) and MUSCLE (Edgar, 2004) to cluster and align reads into loci. Raw sequence reads were discarded if they had ≥4 bp with a Phred quality score <20. Reads were clustered within individuals and then across individuals with clustering thresholds of 90, 92, and 95%, and we ultimately chose 92% to minimize the number of paralogs while not over-splitting homologous loci given the sequence divergence across populations (Ilut et al., 2014; de Oca et al., 2017). We used a minimum depth of coverage of 10 for all loci. We set the paralog filter in pyRAD to 90%, meaning that up to 90% of individuals at a site can be heterozygous (e.g., be represented by two alleles with an IUPAC ambiguity code), as we expect many heterozygous positions to be due to shared ancestry (e.g., homology) and not due to fixed paralogs differences. We set the missing data threshold at 25%, meaning that ≥75% of individuals had data at each locus. All other parameters in pyRAD were left at their default settings.

2.4.2. Unlinked SNPs vs. Sequence Data

Unlinked SNPs can generate a maximum of four alleles per locus, but are more commonly bi-allelic with only two alleles. However, considering all variant and invariant sites together can greatly increase the number of distinct alleles at a locus. This richer allelic information might offer higher precision in delimiting population boundaries and/or inferring admixture proportions vs. SNPs, so we analyzed both datasets in parallel for comparison. PyRAD generates a ".alleles" file that contains allelic sequence data (e.g., two alleles per individual) for all loci that met all quality and assembly parameters; sequences need not be 39 bp, as indels can cause loci to be >39 bp. It is from these loci that biallelic SNPs are extracted. These ≥39 bp RAD loci were then coded as alleles (e.g., "microhaplotypes"), two per individual. We generated a custom Python script to parse the ".alleles" file into a file formatted for the program Structure (Pritchard et al., 2000), where any non-N difference at a site between alleles constituted a unique and new allele. This dataset (herein termed "alleles") was then analyzed in parallel to the unlinked SNPs dataset to compare the power of each to identify population boundaries, admixture proportions, and clines.

2.4.3. mtDNA Dataset Assembly

Raw sequence data (".ab1" chromatograms) from both sequencing directions were made into contigs and hand-edited in Geneious v10 (Biomatters; Auckland, New Zealand). Consensus sequences were exported as .fasta sequences and aligned with Clustal2 (Larkin et al., 2007) in Mesquite v3.2 (Maddison and Maddison, 2017). *Liolaemus cuyanus* was included as an outgroup to root phylogenetic trees used in cline analyses (see below).

2.5. Geographic Cline Analyses

We estimated clines for both nuclear and mitochondrial datasets to identify the geographic interface between populations, and to contrast cline patterns between markers with different inheritance patterns. To generate transect distances along a single axis between sampling localities of each hybrid zone, we first calculated pairwise distances between every sampling locality as the great circle distance with latitude-longitude coordinates in the R package *Fossil* (Vavrek and Vavrek, 2012) with the function "earth.dist." We note that this method does not consider topography when calculating distances. We then used classical multidimensional scaling through principal coordinates analysis to reduce the pairwise matrix of distances between each locality into a single distance value for each locality that retained the original overall pairwise distance structure (as in Gompert et al., 2010). This ordination represents sampling locations along a single axis where kilometer (km) 0 was converted to represent the northern-most sampling site of each transect.

2.5.1. nDNA Clines

We used Structure v2.3 (Pritchard et al., 2000; Falush et al., 2007) and the Evanno method (Evanno et al., 2005) to identify the number of populations (k) in each hybrid zone. Analyses on the Northern hybrid zone dataset were run for 250,000 generations following a 75,000 generation burn-in period with five replicates of each k value of 1–5. Because of a higher number of loci (see Results below), the Central + Southern hybrid zones dataset was run for 300,000 generations with 100,000 burn-in generations, also with five replicate runs of k 1–5.

After identifying the optimal k value, we used Structure to determine the admixture proportions (Q) of all individuals and therefore of each sampling locality. Q estimates from five replicate Structure runs were combined through CLUMPP (Jakobsson and Rosenberg, 2007) and were then used to generate geographic clines. For the combined Central + Southern hybrid zones, the Evanno et al. (2005) method selected $k = 3$ (**Supplementary Figure 1A**), where a central population intergrades with two distinct populations, one to the north and another to the south. This larger Central + Southern hybrid zone was therefore split into two separate hybrid zones, where the northern hybrid zone was designated as localities A–I and the southern hybrid zone as localities F-M (**Figure 1**). The northern half of our bigger southern transect will be referred to as the "Central" hybrid zone, and the southern portion of the southern transect will be referred to as the "Southern" hybrid zone.

With the use of the Q proportions and geographic locations as described above, we estimated geographic maximum likelihood clines, including cline centers and cline widths, in the R package *Hzar* (Derryberry et al., 2014). Cline models were tested with minimum and maximum values fixed to the observed data, without allowing exponential tails on both sides of the cline. The cline fit analysis was run for 200,000 generations and a burn-in of 40,000 generations, from which the maximum likelihood parameter estimates of the cline were generated. The best-fit cline model (along with 95% confidence interval) was then plotted as a function of geographic distance along the transect.

2.5.2. mtDNA Clines

Because mtDNA is haploid and non-recombining, haplotype frequencies were calculated in terms of the relative proportions of the distinct parental haplotypes found at each sample location. We used both tree-based and network-based approaches to determine haplotype frequencies at each sampling locality. For the tree-based approach, we used jModelTest v2.1.7 (Guindon and Gascuel, 2003; Darriba et al., 2012) to determine the optimal DNA substitution model (HKY + Γ for all datasets), which was then used to estimate maximum likelihood trees in RAxML v8.2 (Stamatakis, 2014) with 100 bootstrap iterations. For each hybrid zone, we calculated haplotype frequencies as the proportion of individuals in each locality that belonged to the "northern" clade, resulting in haplotype frequencies ranging from 0 to 1.

Our second approach was analogous to the tree-based approach, but instead was network-based. We inferred minimum-spanning networks (Bandelt et al., 1999) using the program PopART (http://popart.otago.ac.nz), and divided the network into two groups on the edge (branch) with the highest number of sequence substitutions. As in the tree-based approach, we determined haplotype frequencies by calculating the proportion of individuals from each locality that were in each of the two major groups. Cline analysis was performed with these frequencies using the same methodology as in the nDNA cline estimates.

Because we were interested in contrasting evolutionary patterns in the mitochondrial vs. nuclear genomes, we quantitatively tested how different the cline estimates were for these two datasets. To do so, we constrained the cline estimate of the nuclear data to have either the cline center or cline width that was inferred from the mtDNA, and then estimated the log likelihood of the constrained clines (for both alleles and unlinked SNPs datasets). With the maximum likelihood estimate and number of free parameters in the model, we were able to estimate AIC scores for each cline (with the "hzar.AIC.hzar.cline" function). A difference in AIC score >2 between unconstrained and constrained cline estimates indicated a significant difference between the two genomes.

3. RESULTS

After we removed individuals with poor genomic DNA or sequence data quality and filtered loci based on the parameters above, the nuclear datasets consisted of 151 individuals (2,814–15,963 loci) in the Northern hybrid zone, 73 individuals in the Central and 61 in the Southern hybrid zones (586–13,835 loci). After combining across individuals, the datasets consisted of 1,295 and 2,436 loci in the Northern and Central + Southern hybrid zones, respectively. We removed individuals from a single locality in the Northern hybrid zone because our analyses showed it to be geographically outside (to the east) of the hybrid zone. We also removed a single locality from analysis from the Central hybrid zone because this locality was represented by a single individual. Samples per locality ranged from 3 to 13 in the Northern hybrid zone with an average of 7.8, a range of 3–15 with an average of 9.1 in the Central hybrid zone, and a range of 3–11 with an average of 7.6 individuals in the Southern transect localities (**Supplementary Tables 1**, **2**). In the mitochondrial dataset (832 base pairs), the Northern transect was represented by 146 individuals, whereas the Central and Southern transects had 75 and 59 individuals, respectively (**Supplementary Tables 3**, **4**). Individuals in both transects displayed a high level of morphological variation across both age and localities (**Figure 2**). The 16 localities in the Northern transect had an average sample size of 8.75 and ranged from 2 to 13 individuals; the average number of mitochondrial samples per locality in the Central and Southern transects ranged from 2 to 15 ($\bar{x} = 9.38$) and 2 to 13 ($\bar{x} = 8.00$), respectively.

3.1. Population Identification

The numbers of unlinked biallelic SNPs used in the Northern and Central + Southern transects were 1,295 (mean number of loci per individual = 1,140) and 2,436 (mean number of loci per individual = 2,173), respectively. Coding the nuclear loci into alleles, which retains all of the SNP variation at each locus, resulted in an average of 6.6, 4.7, and 4.5 alleles per locus, with

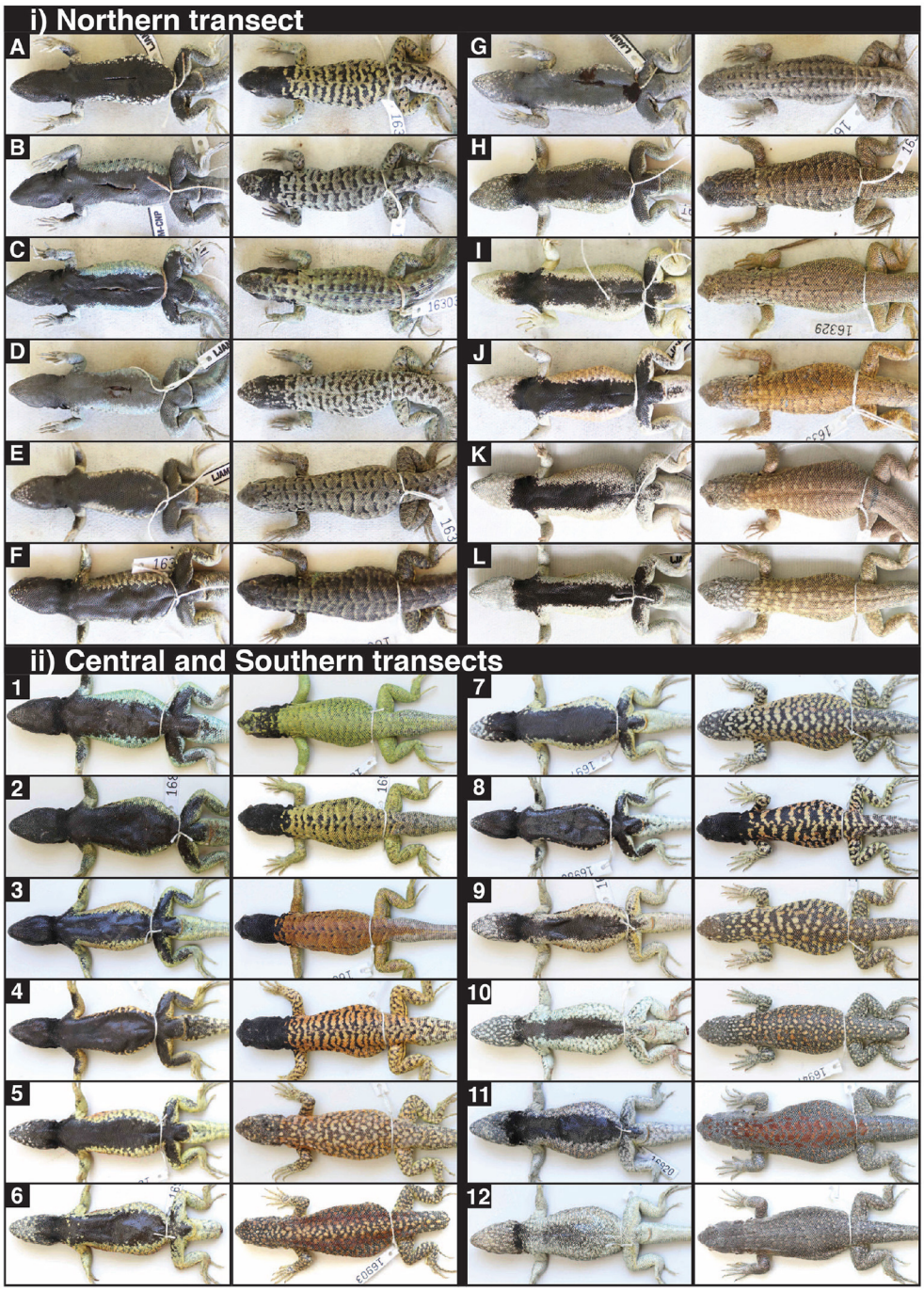

FIGURE 2 | Ventral and dorsal photos of males from localities sampled in the Northern (i) and Central + Southern (ii) hybrid zones. Letters and numbers indicate transect sampling points in **Figure 4**. Photos were not available for individuals from localities M–P in the northern transect.

maximum number of alleles of 32, 28, and 28 for Northern, Central, and Southern transects, respectively (**Figure 3**).

3.1.1. Northern Hybrid Zone

The Evanno et al. (2005) method favored two populations ($k = 2$) with the unlinked SNPs and alleles datasets alike (**Supplementary Figure 2**). The interface between the two populations is sharp and occurs on the eastern edge of the Somuncura Plateau (**Figure 4A**).

3.1.2. Central and Southern Hybrid Zones

Estimates of the optimal k value via the Evanno et al. (2005) method were in conflict: the unlinked SNPs dataset favored four populations, whereas the alleles dataset supported three

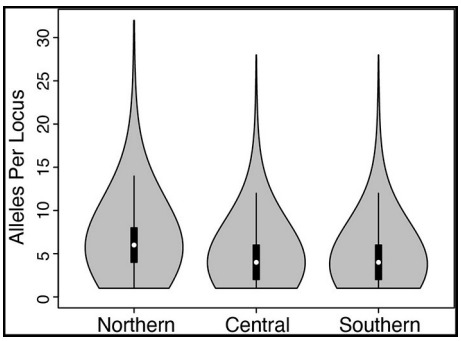

FIGURE 3 | Violin plots showing the number of alleles per locus when the sequence data were coded into alleles (vs. unlinked SNPs) for Northern, Central, and Southern hybrid zones. Density is shown by width of the "violin," whereas box plots inside depict the mean (white dot), first and third quartiles (black boxes), and 1.5× the inter-quartile range (vertical lines).

(**Supplementary Figure 1**). Visualizing the results of $k = 4$ revealed that the fourth inferred population is almost completely confined to individuals in the northern-most sampling locality ("1"; **Figure 4**; **Supplementary Figure 3**). The $k = 4$ result doesn't make biological sense and is in conflict with the results from the alleles dataset, so we therefore focused on the $k = 3$ results for the larger Central + Southern transect. Visualizing the $k = 3$ result revealed a "sandwich" hybrid zone in which individuals from the center of the transect (roughly equivalent to the described species *Liolaemus xanthoviridis*) hybridize with two distinct populations—one to the north (*L. melanops*) and one to the south (*L. fitzingerii*; **Figures 4B,C**). Furthermore, the northern population in the Central hybrid zone is the same "species," *L. melanops*, that constitutes the northern populations of the Northern transect (**Figure 1**; **Supplementary Figure 4**).

3.2. Clines

3.2.1. Northern Hybrid Zone

Cline width estimates were 30.13 and 35.27 km for the alleles and unlinked SNPs datasets, respectively (**Table 1**). Estimates of cline centers from the two nuclear datasets were ∼0.5 km different from one another in the northern hybrid zone (**Table 1**). The inferred admixture (Q) proportions were more extreme for the alleles dataset, providing admixture estimates closer to 1 or 0 at the opposite ends of the transect (**Figure 5A**). In terms of calculating haplotype frequencies from the mitochondrial data, the phylogeny, and network were in 100% agreement (**Supplementary Figure 5**). When mitochondrial and nuclear clines are compared, the mitochondrial cline is shifted ∼7 km to the south of the nDNA clines and is ∼13 km narrower at 20.64 km (**Table 1**; **Figure 5**). When the nuclear data were inferred under the constraint of the mitochondrial cline center or width estimates, the position of the center, but not the width, was inferred to be significantly different (**Table 2**).

3.2.2. Central and Southern Hybrid Zones

Central. As in the Northern hybrid zone, admixture proportions inferred with the alleles dataset were more extreme than the unlinked SNPs dataset (**Figure 5B**). The cline center inferred from the nDNA is ∼40 km to the south of the northern-most sampling locality, and ∼45 km wide (**Table 1**). As in the Northern hybrid zone, the haplotype frequencies calculated from

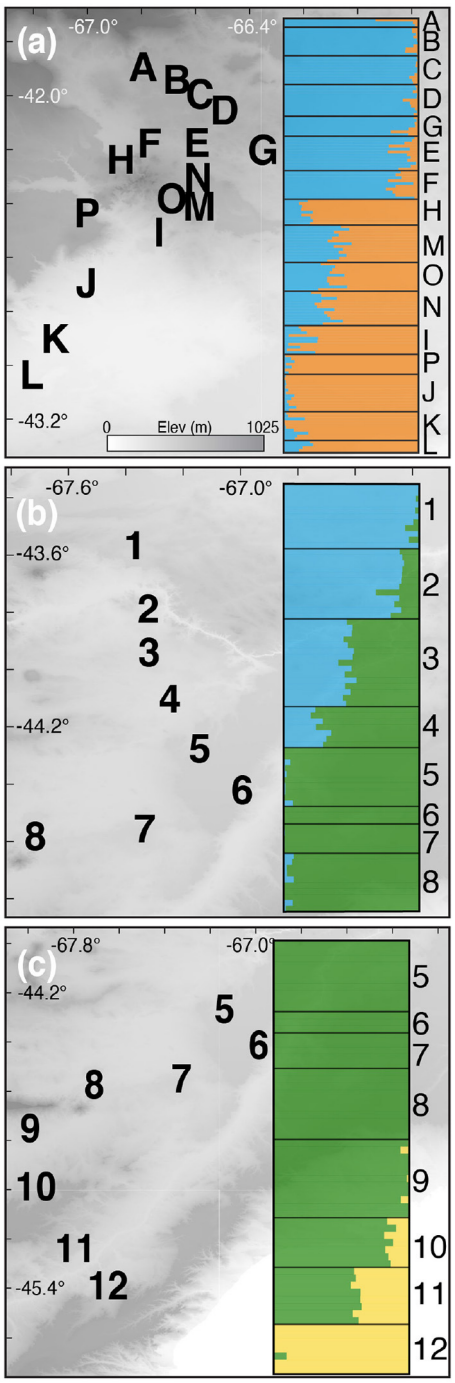

FIGURE 4 | STRUCTURE results and hybrid zone locations for the Northern **(A)**, Central **(B)**, and Southern **(C)** hybrid zones. Latitude and longitude are shown, and the elevational gradient is the same in all maps. Note that populations 5–8 are used in both Central **(B)** and Southern **(C)** transects.

TABLE 1 | Cline analysis results from Hzar for Northern, Central, and Southern hybrid zones.

Dataset	Northern		Central		Southern	
	Center	Width	Center	Width	Center	Width
Alleles	36.87 (32.14–41.62)	33.75 (20.03–53.03)	42.86 (32.90–53.16)	53.44 (27.45–93.60)	120.84 (109.28–135.01)	45.99 (19.58–101.68)
Unlinked SNPs	37.59 (31.18–44.35)	42.84 (21.77–77.17)	41.34 (27.46–55.09)	63.43 (27.44–124.20)	122.10 (108.45–145.54)	58.68 (20.21–143.68)
mtDNA	43.96 (40.59–48.21)	20.64 (14.54–35.20)	59.57 (52.84–67.74)	21.37 (16.35–54.25)	123.08 (116.50–146.81)	1.19 (5.26–86.58)

Mean estimates are shown along with the 95% credible intervals in parentheses. The cline center results represent distance from the northern-most sampling locality, and all numbers represent kilometers.

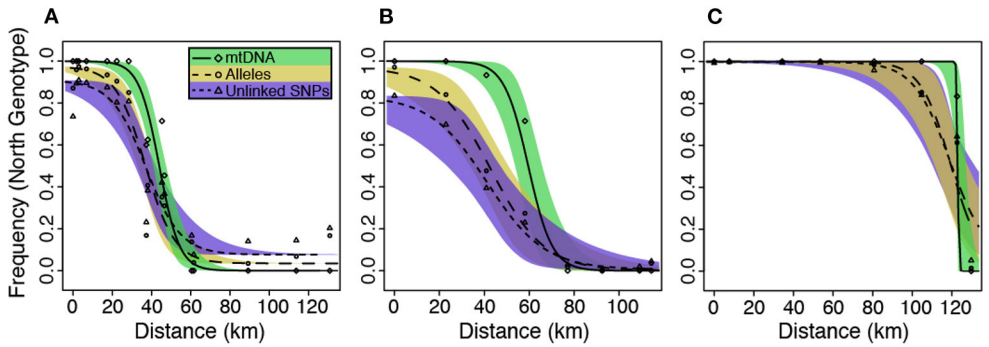

FIGURE 5 | Hzar maximum likelihood cline estimates and 95% credible intervals estimated from both nDNA and mtDNA datasets for the **(A)** Northern, **(B)** Central, and **(C)** Southern hybrid zones. The x-axis represents distances from the northern-most sampling locality.

TABLE 2 | AIC scores for genetic clines from Hzar analyses.

	Northern			Central			Southern		
	nDNA	mtDNA Center	mtDNA Width	nDNA	mtDNA Center	mtDNA Width	nDNA	mtDNA Center	mtDNA Width
Alleles	12.480	**18.932**	12.784	5.943	**13.663**	**10.257**	8.372	7.236	**296.191**
Unlinked SNPs	11.438	**14.003**	12.495	5.210	**10.696**	**7.525**	8.222	6.939	**11.890**

The nDNA value represents the AIC score when estimating the maximum likelihood (ML) cline for the nDNA, whereas the "mtDNA Center" and "mtDNA Width" columns represent AIC scores when forcing the ML estimate of the mtDNA center or width on the nDNA ML cline estimates, respectively. Bold values indicate AIC scores >2 points different in comparison to the freely estimated nDNA clines.

the mtDNA data were identical between phylogeny and network approaches (**Supplementary Figure 6**). The nDNA clines are in stark contrast to the mtDNA cline, whose center is ∼20 km to the south and less than half as wide as the nDNA clines (21.37 km). When the nDNA was constrained to fit the mtDNA cline center and width, the clines estimated from both data types were significantly different from each other in both of these characteristics (**Table 2**).

Southern. Cline center estimates were only 0.43 km different between alleles and unlinked SNP datasets. However, the alleles cline width estimate was ∼10 km narrower (27.42 vs. 37.14 km; **Table 1**). In comparison to the mitochondrial genealogies inferred for the other two hybrid zones, the phylogeny of the Southern transect individuals did not contain two strongly supported clades (**Supplementary Figure 7**). However, two distinct groups were inferred in the network that corresponded to a division created by the longest branch in the phylogeny. In contrast to the other two transects, the clines estimated in the Southern transect were in the very southern portion of the transect (**Figure 5C**). The mtDNA cline center was ∼4 km to the south and much narrower (1.19 km) when compared to the nDNA clines (**Table 1**; **Figure 5C**). Constraining the nDNA cline center to the mtDNA estimate strangely led to an improvement in model score, however, the nDNA cline widths were significantly wider than the mtDNA cline width (**Table 2**).

4. DISCUSSION

Our study marks the first in-depth study of hybrid zone dynamics within *Liolaemus*, a clade where hybridization is widespread and potentially fundamental to its evolutionary history. The arrangement of three geographically sequential hybrid zones in the *L. fitzingerii* species group, a group in which hybridization appears to be common, is unusual and provides a valuable system for analyzing hybridization in a replicated fashion. In the north,

L. melanops hybridizes with *L. shehuen* and *L. xanthoviridis*, and in the south, *L. xanthoviridis* hybridizes with both *L. melanops* and *L. fitzingerii*. Analyses revealed similarities shared across all three hybrid zones: mitochondrial clines are (1) steeper compared to nuclear clines, (2) displaced to the south of the nuclear clines, and (3) significantly different from nuclear clines in terms of cline center and/or width. Our results indicate that hybridization is common in the *L. fitzingerii* species group and the hybrid zones are well-defined. Although hybridization is common and is a potential mechanism for generating extensive phenotypic variation in the *L. fitzingerii* species group (**Figure 2**), we did not test whether hybridization enhances speciation (through a mechanism, such as reinforcement) as some authors have hypothesized because it is outside the scope of this paper (e.g., Olave et al., 2018; Morando et al., 2020).

4.1. Hybridization and Species Boundaries in *Liolaemus*

In spite of considerable progress over the past few decades, much remains to be understood about phylogeography and systematics of southern hemisphere taxa (Beheregaray, 2008). Knowledge on the taxonomic and phylogenetic diversity of Patagonian lizards specifically is incomplete, leaving room for many future studies (Brito, 2010; Diniz-Filho et al., 2013). These uncertainties manifest taxonomically and result in many species "groups" and "complexes" whose geographic distributions, and species limits, are not clearly defined. The results here indicate that hybrid zones are clearly defined in the *L. fitzingerii* group, and that in spite of many instances of interspecific hybridization, species are clear entities outside of contact zones.

Character clines in hybrid zones can vary substantially in shape—broad vs. narrow—and different shapes can provide insights into the evolutionary processes maintaining hybrid zones. A recent meta-analysis of animal hybrid zones (McEntee et al., 2020) provides some context for interpreting the mtDNA and nuclear cline widths estimated in the *L. fitzingerii* group. Across a variety of taxa, hybrid zone cline widths range from 10 m to >3,000 km (McEntee et al., 2020). In lizards (n = 95 cline estimates in McEntee et al., 2020), the reported range is ∼30–190 km with a left-skewed distribution—20% of the values are <1 km and 90% are <60 km. The hybrid zones in the *L. fitzingerii* group were estimated to be ∼35–65 km wide with nuclear data, or ∼1–20 km wide with mitochondrial data (**Figure 5**). Accordingly, the cline widths in the *L. fitzingerii* group appear to be "typical" in relation to other lizard species. We would expect much more variance in cline estimates across hybrid zones if a cline was maintained solely by selection, as opposed to a balance between dispersal and selection (Barton and Hewitt, 1985). The observation that both cline width and shape do not vary substantially between hybrid zones indicates that dispersal of parental genotypes into the contact zone is offset by selection against heterozygotes. In the *L. fitzingerii* species group, the strengths of selection and gene flow seem to be within the same order of magnitude, and similar to those seen in other squamate species (Mallet et al., 1990; McEntee et al., 2020).

4.2. Nuclear vs. Mitochondrial Clines

Geographic cline analyses revealed that the mitochondrial cline center is displaced to the south of the nuclear cline in all three hybrid zones. Furthermore, nuclear and mitochondrial clines were significantly different from each other in cline center and/or width in all three hybrid zones. Observing significantly different clines between nuclear and mtDNA is not necessarily unexpected, given that a variety of biotic and evolutionary processes can generate discordance between nuclear and mitochondrial DNA (Bonnet et al., 2017). These two genomes have different modes of inheritance (unipartental vs. biparental), recombination (mtDNA lacks recombination), and are subject to different selection pressures (Ballard and Whitlock, 2004). Additionally, the amount of gene flow between populations within a species and demographic factors affecting levels of allele "surfing" can mitigate introgression at contact zones and further complicate characterizations of hybrid zone dynamics (Petit and Excoffier, 2009).

Discordance between nuclear and mitochondrial genomes and their estimated clines can be generated by two classes of processes affecting the mitochondrial genome: selective (e.g., positive selection for the introgressing mitochondrial genome or negative pleiotropic selection on many nuclear loci) and neutral processes involving sex-related asymmetries, such as interspecific mate preference (females of taxon a preferring males of taxon b while no such preference occurs in females of taxon b), sex-biased dispersal, or differences in hybrid survival by sex (Funk and Omland, 2003; Bonnet et al., 2017). In *Liolaemus* lizards, males leave their natal ground to establish home ranges, whereas females disperse shorter distances (Kacoliris et al., 2009), arguing that sex-biased dispersal could result in a southerly shifted mtDNA via northward migration of juvenile males from the southern population into the northern population. Additionally, a southerly shifted mtDNA cline could also result from a southward migration of females from the northern population into the southern population; these two hypotheses are not mutually exclusive. Thus, although sex-biased dispersal, asymmetric mating preferences, or differential survival rates of hybrid offspring can lead to mito-nuclear discordance (Bonnet et al., 2017), we are unable to determine the relative strengths of these processes here.

A second reason for the discordant mt- and nDNA clines is that these hybrid zones could be moving. Many empirical studies have documented moving hybrid zones over time (reviewed in Buggs, 2007). Hybrid zones can move when selection against hybrids is genetically countered by dispersal of parental forms into the hybrid zone (tension zone model; Barton and Hewitt, 1985), or when a change in the external environment causes selection along a gradient to generate fitness differences (May et al., 1975). When an environmental gradient moves (e.g., as the result of a change in climate), geographic ranges and hybrid zones can shift with it (e.g., Leaché et al., 2017). As geographic ranges shift, asymmetric introgression from the expanding species into the stationary one will cause neutral markers to geographically trail behind non-neutral markers (McGuire et al., 2007). In particular, asymmetric introgression

of the mitochondrial genome and its discordance with nuclear markers has been used to deduce a moving hybrid zone (Rohwer et al., 2001). Although hybrid zone movement over time can be inferred from discordant mt- and nDNA clines sampled from a single time-point, the most convincing cases of hybrid zone movement come from studies with replicated sampling efforts over time (e.g., Carling and Zuckerberg, 2011; Taylor et al., 2014; Leaché et al., 2017). A lagging cline inferred from the putatively neutral mtDNA that is following the leading edge of an expanding population further suggests northward range expansion of *L. shehuen*, *L. xanthoviridis*, and *L. fitzingerii*.

Concluding that a hybrid zone moves because a trailing mtDNA cline has been observed assumes that the mitochondrial gene(s) under study is/are neutrally evolving in these species, which might not be true. Indirect selection on mtDNA through differential selection of the heterogametic sex (e.g., Haldane's Rule) or direct selection via cyto-nuclear incompatibilities would also impair the effective movement of mtDNA across the hybrid zone (Dasmahapatra et al., 2002). This leads to a third reason for mt-nDNA cline discordance, which is differential selection on the two genomes (Bonnet et al., 2017). If strong enough positive selection was working on any site in the mitochondrial genome, that mitochondrial haplotype could sweep through the population (due to linkage) and the cline would not lag behind as expected for a neutral marker. Although we did not explicitly test for selection, it is unlikely to affect our results because loci under selection would likely be in the minority of our dataset. Nonetheless, we agree with Dasmahapatra et al. (2002) in that "asymmetry of introgression, or lack of introgression of molecular markers, is relatively unconvincing evidence either for or against hybrid zone movement."

We performed our population structure and cline analyses on two nuclear datasets, one where a single SNP was randomly selected from each RAD locus, and another that used all invariant and variant sites present at each locus recoded into alleles ("alleles"). The alleles dataset provided many more alleles per locus than the unlinked SNPs dataset, with 6.6, 4.7, and 4.5 alleles per locus for the alleles dataset in the Northern, Central, and Southern transects, respectively, whereas the unlinked SNPs datasets contained only bi-allelic loci. Although Structure plots between the two datasets were qualitatively similar (results not shown), admixture proportions (*Q*) were more "intermediate" for the unlinked SNPs dataset, meaning that the *Q* values weren't as extreme as in the alleles dataset. This can be seen in the cline estimates (**Figure 5**), where the frequency of the northern genotype for the alleles dataset reached closer to 0.0 and 1.0. A similar pattern is seen in the cline width estimates (**Table 1**), where the widths estimated for two of the three transects from the alleles dataset were narrower by ∼5–10 km. These narrower cline estimates, and more extreme *Q* estimates, are almost certainly due to the increased information content associated with higher allelic richness in the alleles dataset. It is not possible to determine which dataset produced more accurate cline parameters without conducting a thorough simulation study where the true cline parameters are known. However, we suspect that the "alleles" data has the advantage over the bi-allelic SNP analysis because it uses all of the variation present in the data.

4.3. Replicated Hybrid Zones

In this study, two species—*L. melanops* and *L. xanthoviridis*—are each involved in two hybrid zones. First, *L. melanops* hybridizes with *L. shehuen* in the Northern hybrid zone as well as with *L. xanthoviridis* in the Central hybrid zone (**Figure 1**). In the Northern hybrid zone, the interface of *L. melanops* and *L. shehuen* occurs on the eastern edge of the Somuncura Plateau, a geological feature that is ∼25 million years old (Kay et al., 2007) and reaches an elevation of ∼1,600 m. That this geologic feature is at the interface of two populations is perhaps not surprising, however, *L. shehuen* individuals are found both below (to the east) and on top of this plateau. The elevation imposed by this plateau does seem to form a western barrier for *L. melanops*, which is found in lower elevation Patagonian shrub-steppe habitats to the east and south of the plateau. In fact, elevation explains 32% of the variance in admixture proportions (*Q*) between these two species (**Supplementary Figure 8**). Assuming equal dispersal capabilities of *L. melanops* individuals throughout the range of this species, the narrower cline width in the north (∼32 vs. 45 km) qualitatively implies stronger selection in the Northern hybrid zone. This evidence implies that exogenous selection (e.g., environmental differences) is a potential mechanism maintaining *L. melanops* and *L. shehuen* as distinct species. The boundary between *L. melanops* and *L. xanthoviridis* corresponds with the Chubut River, which is a large river and seasonally >100 m wide in this area. Although the divergence between these two species appears to be allopatric, our genetic data show that the Chubut River is in fact a porous boundary.

Second, in a similarly replicated fashion, *Liolaemus xanthoviridis* hybridizes in two separate areas: to the north with *L. melanops* and to the south with *L. fitzingerii*. The nDNA cline width in the north with *L. melanops* is ∼45 km, whereas it is ∼32 km wide in the hybrid zone with *L. fitzingerii*. Assuming these hybrid zones are best modeled as tension zones that are a balance of dispersal and selection, narrower clines could be the result of two non-mutually exclusive causes: reduced dispersal abilities, or stronger selection. In *Liolaemus* generally, we do not have good estimates of dispersal (but see Frutos and Belver, 2007 and Camargo et al., 2013 for some estimates), especially when trying to compare differing dispersal abilities between species in the *L. fitzingerii group*. In terms of selection, the narrower cline seen in the Southern hybrid zone does not seem to be the result of sexual selection via interspecific mating and a higher disparity in body sizes because both taxa are large-bodied (male max SVL = 94 vs. 102 mm for *L. xanthoviridis* and *L. fitzingerii*; Etheridge, 2000). The narrower cline in this hybrid zone, however, might be due to exogenous (environmental) causes. *Liolaemus fitzingerii* is found in loosely formed sand dunes dominated by the mesquite bush *Prosopis denudans*, whereas *L. xanthoviridis* occurs in the hardpan Patagonian shrub-steppe habitat.

5. CONCLUSIONS

In this study, we were able to compare multiple hybrid zones across *Liolaemus* lizards in central Argentina. Hybridization

appears to be common in the *L. fitzingerii* group, and the hybrid zones are narrow and geographically localized. *Liolaemus melanops* hybridizes with two species, and the hybrid zone in the north (with *L. shehuen*) is significantly narrower than in the south (with *L. xanthoviridis*), likely due to the environmental gradient (i.e., change in elevation and vegetation) posed by the Somuncura Plateau. Nonetheless, other hybrid zones in this group have formed in the absence of any obvious physical barriers, suggesting that other ecological or intrinsic factors may be playing a role in maintaining species as distinct entities. The discordance between mitochondrial and nuclear cline estimates suggests sex-biased dispersal, divergent selection across genomes, or movement of these hybrid zones over time. Re-sampling these hybrid zones in the future may help tease apart these alternative hypotheses. Lastly, although hybridization has generated novel genotypes and morphological variation in hybrid zones, it is unclear whether hybridization has reinforced species boundaries or promoted speciation within the *L. fitzingerii* group. This research has provided a fine-scale understanding of hybrid zone dynamics within the *Liolaemus fitzingerii* group, with implications not only for other *Liolaemus* species and Patagonian taxa more broadly, but for hybrid zone systems globally.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found at: https://www.ncbi.nlm.nih.gov/, PRJNA670250.

ETHICS STATEMENT

All research methods were approved by the University of Washington Office of Animal Welfare (IACUC protocol number 4249-01) and in accordance with provincial permits from the Argentinean Dirección de Fauna y Flora Silvestre.

AUTHOR CONTRIBUTIONS

JG performed the field sampling, DNA data collection, analyses, and wrote the paper. LA performed the field sampling and provided the financial support. MM and AL edited the manuscript and provided the financial support. All authors designed the research.

FUNDING

This research was funded in part by a National Science Foundation Doctoral Dissertation Improvement Grant (DEB-1500933) and UW Study Abroad fellowship to JG, ANPCyT-FONCYT PICT 2015-1252 (LA), and CONICET-PIP 2012-336 (MM).

ACKNOWLEDGMENTS

This work used the Vincent J. Coates Genomics Sequencing Laboratory at UC Berkeley, supported by NIH S10 OD018174 Instrumentation Grant. Analyses were facilitated though the use of advanced computational, storage, and networking infrastructure provided by the Hyak supercomputer system at the University of Washington. Thanks to C. Pérez and T. Avila for assistance in sample collection during the December 2015 trip.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fevo.2021.624109/full#supplementary-material

REFERENCES

Aboim, M., Mavárez, J., Bernatchez, L., and Coelho, M. (2010). Introgressive hybridization between two Iberian endemic cyprinid fish: a comparison between two independent hybrid zones. *J. Evol. Biol.* 23, 817–828. doi: 10.1111/j.1420-9101.2010.01953.x

Araya-Donoso, R., Torres-Pérez, F., Véliz, D., and Lamborot, M. (2019). Hybridization and polyploidy in the weeping lizard *Liolaemus chiliensis* (Squamata: Liolaemidae). *Biol. J. Linn. Soc.* 128, 963–974. doi: 10.1093/biolinnean/blz145

Avila, L., Morando, M., and Sites, J. (2006). Congeneric phylogeography: hypothesizing species limits and evolutionary processes in patagonian lizards of the *Liolaemus boulengeri* group (Squamata: Liolaemini). *Biol. J. Linn. Soc.* 89, 241–275. doi: 10.1111/j.1095-8312.2006.00666.x

Avila, L. J., Morando, M., and Sites J. W. Jr. (2008). New species of the Iguanian lizard genus *Liolaemus* (Squamata, Iguania, Liolaemini) from Central Patagonia, Argentina. *J. Herpetol.* 42, 186–196. doi: 10.1670/06-244R2.1

Avila, L. J., Pérez, C. H. F., Morando, M., and Sites J. W. Jr., (2010). A new species of *Liolaemus* (Reptilia: Squamata) from southwestern Rio Negro province, Northern Patagonia, Argentina. *Zootaxa* 2434, 47–59. doi: 10.11646/zootaxa.2434.1.4

Ballard, J. W. O., and Whitlock, M. C. (2004). The incomplete natural history of mitochondria. *Mol. Ecol.* 13, 729–744. doi: 10.1046/j.1365-294X.2003.02063.x

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

Barton, N. H., and Hewitt, G. M. (1985). Analysis of hybrid zones. *Annu. Rev. Ecol. Syst.* 16, 113–148. doi: 10.1146/annurev.es.16.110185.000553

Beheregaray, L. B. (2008). Twenty years of phylogeography: the state of the field and the challenges for the southern hemisphere. *Mol. Ecol.* 17, 3754–3774. doi: 10.1111/j.1365-294X.2008.03857.x

Bonnet, T., Leblois, R., Rousset, F., and Crochet, P. A. (2017). A reassessment of explanations for discordant introgressions of mitochondrial and nuclear genomes. *Evolution* 71, 2140–2158. doi: 10.1111/evo.13296

Breitman, M. F., Avila, L. J., Sites, J. W., and Morando, M. (2011). Lizards from the end of the world: phylogenetic relationships of the *Liolaemus lineomaculatus* section (Squamata: Iguania: Liolaemini). *Mol. Phylogenet. Evol.* 59, 364–376. doi: 10.1016/j.ympev.2011.02.008

Brelsford, A., and Irwin, D. E. (2009). Incipient speciation despite little assortative mating: the yellow-rumped Warbler hybrid zone. *Evolution* 63, 3050–3060. doi: 10.1111/j.1558-5646.2009.00777.x

Brito, D. (2010). Overcoming the linnean shortfall: data deficiency and biological survey priorities. *Basic Appl. Ecol.* 11, 709–713. doi: 10.1016/j.baae.2010.09.007

Buggs, R. (2007). Empirical study of hybrid zone movement. *Heredity* 99, 301–312. doi: 10.1038/sj.hdy.6800997

Camargo, A., Werneck, F. P., Morando, M., Sites, J. W., and Avila, L. J. (2013). Quaternary range and demographic expansion of *Liolaemus darwinii*

(Squamata: Liolaemidae) in the monte desert of Central Argentina using bayesian phylogeography and ecological niche modelling. *Mol. Ecol.* 22, 4038–4054. doi: 10.1111/mec.12369

Carling, M. D., and Zuckerberg, B. (2011). Spatio-temporal changes in the genetic structure of the Passerina bunting hybrid zone. *Mol. Ecol.* 20, 1166–1175. doi: 10.1111/j.1365-294X.2010.04987.x

Chhatre, V. E., Evans, L. M., DiFazio, S. P., and Keller, S. R. (2018). Adaptive introgression and maintenance of a trispecies hybrid complex in range-edge populations of populus. *Mol. Ecol.* 27, 4820–4838. doi: 10.1111/mec.14820

Culumber, Z., Fisher, H., Tobler, M., Mateos, M., Barber, P., Sorenson, M., et al. (2011). Replicated hybrid zones of xiphophorus swordtails along an elevational gradient. *Mol. Ecol.* 20, 342–356. doi: 10.1111/j.1365-294X.2010.04949.x

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

Darwin, C. (1862). Notes on the causes of cross and hybrid sterility. *Correspond. Charles Darwin* 10, 700–711.

Dasmahapatra, K. K., Blum, M. J., Aiello, A., Hackwell, S., Davies, N., Bermingham, E. P., et al. (2002). Inferences from a rapidly moving hybrid zone. *Evolution* 56, 741–753. doi: 10.1111/j.0014-3820.2002.tb01385.x

de Oca, A. N. M., Barley, A. J., Meza-Lázaro, R. N., García-Vázquez, U. O., Zamora-Abrego, J. G., Thomson, R. C., et al. (2017). Phylogenomics and species delimitation in the knob-scaled lizards of the genus *Xenosaurus* (Squamata: Xenosauridae) using ddradseq data reveal a substantial underestimation of diversity. *Mol. Phylogenet. Evol.* 106, 241–253. doi: 10.1016/j.ympev.2016.09.001

Derryberry, E. P., Derryberry, G. E., Maley, J. M., and Brumfield, R. T. (2014). HZAR: hybrid zone analysis using an r software package. *Mol. Ecol. Resour.* 14, 652–663. doi: 10.1111/1755-0998.12209

Diniz-Filho, J. A. F., Loyola, R. D., Raia, P., Mooers, A. O., and Bini, L. M. (2013). Darwinian shortfalls in biodiversity conservation. *Trends Ecol. Evol.* 28, 689–695. doi: 10.1016/j.tree.2013.09.003

Duvernell, D. D., and Schaefer, J. F. (2014). Variation in contact zone dynamics between two species of *Topminnows, Fundulus notatus* and *F. olivaceus*, across isolated drainage systems. *Evol. Ecol.* 28, 37–53. doi: 10.1007/s10682-013-9653-z

Eaton, D. A. (2014). PyRAD: assembly of *de novo* RADseq loci for phylogenetic analyses. *Bioinformatics* 30, 1844–1849. doi: 10.1093/bioinformatics/btu121

Edgar, R. C. (2004). Muscle: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* 32, 1792–1797. doi: 10.1093/nar/gkh340

Esquerré, D., Ramírez-Álvarez, D., Pavón-Vázquez, C. J., Troncoso-Palacios, J., Garín, C. F., Keogh, J. S., et al. (2019). Speciation across mountains: phylogenomics, species delimitation and taxonomy of the *Liolaemus leopardinus* clade (Squamata, Liolaemidae). *Mol. Phylogenet. Evol.* 139:106524. doi: 10.1016/j.ympev.2019.106524

Etheridge, R. (2000). A review of lizards of the *Liolaemus wiegmannii* group (Squamata, Iguania, Tropiduridae), and a history of morphological change in the sand-dwelling species. *Herpetol. Monogr.* 14, 293–352. doi: 10.2307/1467049

Evanno, G., Regnaut, S., and Goudet, J. (2005). Detecting the number of clusters of individuals using the software structure: a simulation study. *Mol. Ecol.* 14, 2611–2620. doi: 10.1111/j.1365-294X.2005.02553.x

Falush, D., Stephens, M., and Pritchard, J. K. (2007). Inference of population structure using multilocus genotype data: dominant markers and null alleles. *Mol. Ecol. Notes* 7, 574–578. doi: 10.1111/j.1471-8286.2007.01758.x

Feltrin, N. (2013). *Conservadurismo o divergencia de nicho filogentico: especies patagonicas del grupo petrophilus (Squamata: Liolaemus) como caso de estudio* (Ph.D. thesis), Universidad Nacional de Córdoba, Córdoba, Argentina.

Frutos, N., and Belver, L. C. (2007). Dominio vital de liolaemus koslowskyi etheridge 1993 (Iguania: Liolaemini) en el noroeste de la provincia de La Rioja, Argentina. *Cuad. Herpetol.* 21, 83–92. Available online at: https://core.ac.uk/download/pdf/301027745.pdf

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

Gompert, Z., Lucas, L. K., Fordyce, J. A., Forister, M. L., and Nice, C. C. (2010). Secondary contact between *Lycaeides idas* and *L. melissa* in the rocky mountains: extensive admixture and a patchy hybrid zone. *Mol. Ecol.* 19, 3171–3192. doi: 10.1111/j.1365-294X.2010.04727.x

Grummer, J. A. (2017). *Evolutionary history of the Patagonian Liolaemus fitzingerii species group of lizards* (Ph.D. thesis), University of Washington, Seattle, WA, United States.

Grummer, J. A., Morando, M. M., Avila, L. J., Sites, J. W. Jr., and Leaché, A. D. (2018). Phylogenomic evidence for a recent and rapid radiation of lizards in the Patagonian *Liolaemus fitzingerii* species group. *Mol. Phylogenet. Evol.* 125, 243–254. doi: 10.1016/j.ympev.2018.03.023

Guindon, S., and Gascuel, O. (2003). A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst. Biol.* 52, 696–704. doi: 10.1080/10635150390235520

Hanemaaijer, M. J., Collier, T. C., Chang, A., Shott, C. C., Houston, P. D., Schmidt, H., et al. (2018). The fate of genes that cross species boundaries after a major hybridization event in a natural mosquito population. *Mol. Ecol.* 27, 4978–4990. doi: 10.1111/mec.14947

Harrison, R. G. (1993). *Hybrid Zones and the Evolutionary Process*. Oxford: Oxford University Press.

Ilut, D. C., Nydam, M. L., and Hare, M. P. (2014). Defining loci in restriction-based reduced representation genomic data from nonmodel species: sources of bias and diagnostics for optimal clustering. *Biomed. Res. Int.* 2014:675158. doi: 10.1155/2014/675158

Irwin, D. E., Bensch, S., and Price, T. D. (2001). Speciation in a ring. *Nature* 409, 333–337. doi: 10.1038/35053059

Jakobsson, M., and Rosenberg, N. A. (2007). Clumpp: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. *Bioinformatics* 23, 1801–1806. doi: 10.1093/bioinformatics/btm233

Jančúchová-Lásková, J., Landová, E., and Frynta, D. (2015). Are genetically distinct lizard species able to hybridize? A review. *Curr. Zool.* 61, 155–180. doi: 10.1093/czoolo/61.1.155

Kacoliris, F. P., Williams, J. D., De Arcaute, C. R., and Cassino, C. (2009). Home range size and overlap in *Liolaemus multimaculatus* (Squamata: Liolamidae) in Pampean coastal dunes of Argentina. *South Am. J. Herpetol.* 4, 229–234. doi: 10.2994/057.004.0305

Kay, S. M., Ardolino, A., Gorring, M., and Ramos, V. (2007). The somuncura large igneous province in Patagonia: interaction of a transient mantle thermal anomaly with a subducting slab. *J. Petrol.* 48, 43–77. doi: 10.1093/petrology/egl053

Larkin, M. A., Blackshields, G., Brown, N., Chenna, R., McGettigan, P. A., McWilliam, H., et al. (2007). Clustal w and clustal x version 2.0. *Bioinformatics* 23, 2947–2948. doi: 10.1093/bioinformatics/btm404

Leaché, A. D., and Cole, C. J. (2007). Hybridization between multiple fence lizard lineages in an ecotone: locally discordant variation in mitochondrial DNA, chromosomes, and morphology. *Mol. Ecol.* 16, 1035–1054. doi: 10.1111/j.1365-294X.2006.03194.x

Leaché, A. D., Grummer, J. A., Harris, R. B., and Breckheimer, I. K. (2017). Evidence for concerted movement of nuclear and mitochondrial clines in a lizard hybrid zone. *Mol. Ecol.* 26, 2306–2316. doi: 10.1111/mec.14033

MacManes, M. (2013). Available online at: http://dx.doi.org/10.6084/m9.figshare.658946. Figshare 5.

Maddison, W. P., and Maddison, D. (2017). *Mesquite: A Modular System for Evolutionary Analysis*. version 3.2. Available online at: http://mesquiteproject.org

Mallet, J. (2005). Hybridization as an invasion of the genome. *Trends Ecol. Evol.* 20, 229–237. doi: 10.1016/j.tree.2005.02.010

Mallet, J., Barton, N., Lamas, G., Santisteban, J., Muedas, M., and Eeley, H. (1990). Estimates of selection and gene flow from measures of cline width and linkage disequilibrium in heliconius hybrid zones. *Genetics* 124, 921–936. doi: 10.1093/genetics/124.4.921

May, R. M., Endler, J. A., and McMurtrie, R. E. (1975). Gene frequency clines in the presence of selection opposed by gene flow. *Am. Nat.* 109, 659–676. doi: 10.1086/283036

McEntee, J. P., Burleigh, J. G., and Singhal, S. (2020). Dispersal predicts hybrid zone widths across animal diversity: implications for species borders under incomplete reproductive isolation. *Am. Nat.* 196, 000–000. doi: 10.1086/709109

McGuire, J. A., Linkem, C. W., Koo, M. S., Hutchison, D. W., Lappin, A. K., Orange, D. I., et al. (2007). Mitochondrial introgression and incomplete lineage

sorting through space and time: phylogenetics of crotaphytid lizards. *Evolution* 61, 2879–2897. doi: 10.1111/j.1558-5646.2007.00239.x

McKinnon, J. S., and Rundle, H. D. (2002). Speciation in nature: the threespine stickleback model systems. *Trends Ecol. Evol.* 17, 480–488. doi: 10.1016/S0169-5347(02)02579-X

Medina, C. D., Avila, L. J., Sites J. W. Jr., and Morando, M. (2014). Multilocus phylogeography of the patagonian lizard complex *Liolaemus kriegi* (Iguania: Liolaemini). *Biol. J. Linn. Soc.* 113, 256–269. doi: 10.1111/bij.12285

Morando, M., Avila, L. J., Baker, J., Sites J. W. Jr., and Ashley, M. (2004). Phylogeny and phylogeography of the *Liolaemus darwinii* complex (Squamata: Liolaemidae): evidence for introgression and incomplete lineage sorting. *Evolution* 58, 842–861. doi: 10.1111/j.0014-3820.2004.tb00416.x

Morando, M., Avila, L. J., and Sites, J. W. (2003). Sampling strategies for delimiting species: genes, individuals, and populations in the *Liolaemus elongatus*-kriegi complex (Squamata: Liolaemidae) in Andean-Patagonian South America. *Syst. Biol.* 52, 159–185. doi: 10.1080/10635150390192717

Morando, M., Medina, C. D., Minoli, I., Pérez, C. H. F., Sites, J. W., and Avila, L. J. (2020). "Diversification and evolutionary histories of Patagonian Steppe lizards," in *Lizards of Patagonia*, eds M. Morando, and L. J. Avila (Cham: Springer), 217–254. doi: 10.1007/978-3-030-42752-8_9

Moritz, C., Schneider, C. J., and Wake, D. B. (1992). Evolutionary relationships within the *Ensatina eschscholtzii* complex confirm the ring species interpretation. *Syst. Biol.* 41, 273–291. doi: 10.1093/sysbio/41.3.273

Nolte, A., Gompert, Z., and Buerkle, C. (2009). Variable patterns of introgression in two sculpin hybrid zones suggest that genomic isolation differs among populations. *Mol. Ecol.* 18, 2615–2627. doi: 10.1111/j.1365-294X.2009.04208.x

Olave, M., Avila, L. J., Sites J. W. Jr., and Morando, M. (2018). Hybridization could be a common phenomenon within the highly diverse lizard genus *Liolaemus*. *J. Evol. Biol.* 31, 893–903. doi: 10.1111/jeb.13273

Olave, M., Marín, A. G., Avila, L. J., Sites, J. W., and Morando, M. (2020). "Disparate patterns of diversification within Liolaemini lizards," in *Neotropical Diversification: Patterns and Processes*, V. Rull and A. Carnaval (Cham: Springer), 765–790. doi: 10.1007/978-3-030-31167-4_28

Peterson, B. K., Weber, J. N., Kay, E. H., Fisher, H. S., and Hoekstra, H. E. (2012). Double digest RADseq: an inexpensive method for de novo snp discovery and genotyping in model and non-model species. *PLoS ONE* 7:e37135. doi: 10.1371/journal.pone.0037135

Petit, R. J., and Excoffier, L. (2009). Gene flow and species delimitation. *Trends Ecol. Evol.* 24, 386–393. doi: 10.1016/j.tree.2009.02.011

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

Pritchard, V. L., and Edmands, S. (2013). The genomic trajectory of hybrid swarms: outcomes of repeated crosses between populations of *Tigriopus californicus*. *Evolution* 67, 774–791. doi: 10.1111/j.1558-5646.2012.01814.x

Rieseberg, L. H., Van Fossen, C., and Desrochers, A. M. (1995). Hybrid speciation accompanied by genomic reorganization in wild sunflowers. *Nature* 375, 313–316. doi: 10.1038/375313a0

Rieseberg, L. H., Whitton, J., and Gardner, K. (1999). Hybrid zones and the genetic architecture of a barrier to gene flow between two sunflower species. *Genetics* 152, 713–727. doi: 10.1093/genetics/152.2.713

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

Rohwer, S., Bermingham, E., and Wood, C. (2001). Plumage and mitochondrial DNA haplotype variation across a moving hybrid zone. *Evolution* 55, 405–422. doi: 10.1111/j.0014-3820.2001.tb01303.x

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

Taylor, S. A., White, T. A., Hochachka, W. M., Ferretti, V., Curry, R. L., and Lovette, I. (2014). Climate-mediated movement of an avian hybrid zone. *Curr. Biol.* 24, 671–676. doi: 10.1016/j.cub.2014.01.069

Vavrek, M. J., and Vavrek, M. M. J. (2012). *Package 'Fossil'*.

Westram, A. M., Faria, R., Johannesson, K., and Butlin, R. (2021). Using replicate hybrid zones to understand the genomic basis of adaptive divergence. *Mol. Ecol.* doi: 10.1111/mec.15861. [Epub ahead of print].

Zieliński, P., Dudek, K., Arntzen, J. W., Palomar, G., Niedzicka, M., Fijarczyk, A., et al. (2019). Differential introgression across newt hybrid zones: evidence from replicated transects. *Mol. Ecol.* 28, 4811–4824. doi: 10.1111/mec.15251

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

METHODS
published: 31 May 2021
doi: 10.3389/fevo.2021.629252

Cross-Species Application of Illumina iScan Microarrays for Cost-Effective, High-Throughput SNP Discovery

Emily D. Fountain[1†], Li-Chen Zhou[2†], Alyssa Karklus[3], Qun-Xiu Liu[2], James Meyers[4], Ian K. C. Fontanilla[5], Emmanuel Francisco Rafael[6], Jian-Yi Yu[2], Qiong Zhang[2], Xiang-Lei Zhu[2], En-Le Pei[2], Yao-Hua Yuan[2] and Graham L. Banes[1,7]**

[1] Wisconsin National Primate Research Center, University of Wisconsin – Madison, Madison, WI, United States, [2] Shanghai Zoological Park, Shanghai, China, [3] School of Veterinary Medicine, University of Wisconsin – Madison, Madison, WI, United States, [4] Independent Researcher, Madison, WI, United States, [5] Institute of Biology, College of Science, University of the Philippines Diliman, Quezon City, Philippines, [6] Avilon Zoo, Rodriguez, Rizal, Philippines, [7] Chinese Academy of Sciences Key Laboratory of Computational Biology, Shanghai Institute of Nutrition and Health, Chinese Academy of Sciences, Shanghai, China

OPEN ACCESS

Edited by:
Michael G. Campana, Smithsonian Conservation Biology Institute (SI), United States

Reviewed by:
Alexandra K. Fraik, Washington State University, United States
Mario Barbato, Catholic University of the Sacred Heart, Piacenza, Italy

***Correspondence:**
Graham L. Banes
banes@wisc.edu
Yao-Hua Yuan
yhyuansh@163.com

† These authors have contributed equally to this work

Specialty section:
This article was submitted to Evolutionary and Population Genetics, a section of the journal Frontiers in Ecology and Evolution

Received: *13 November 2020*
Accepted: *26 April 2021*
Published: *31 May 2021*

Citation:
Fountain ED, Zhou L-C, Karklus A, Liu Q-X, Meyers J, Fontanilla IKC, Rafael EF, Yu J-Y, Zhang Q, Zhu X-L, Pei E-L, Yuan Y-H and Banes GL (2021) Cross-Species Application of Illumina iScan Microarrays for Cost-Effective, High-Throughput SNP Discovery. Front. Ecol. Evol. 9:629252. doi: 10.3389/fevo.2021.629252

Microarrays can be a cost-effective alternative to high-throughput sequencing for discovering novel single-nucleotide polymorphisms (SNPs). Illumina's iScan platform dominates the market, but their commercial microarray products are designed for model organisms. Further, the platform outputs data in a proprietary format. This cannot be easily converted to human-readable genotypes or be merged with pre-existing data. To address this, we present and validate a novel pipeline to facilitate data analysis from cross-species application of Illumina microarrays. This facilitates the generation of a compatible VCF from iScan data and the merging of this with a second VCF comprising genotypes derived from other samples and sources. Our pipeline includes a custom script, iScanVCFMerge (presented as a Python package), which we validate using iScan data from three great ape genera. We conclude that cross-species application of microarrays can be a rapid, cost-effective approach for SNP discovery in non-model organisms. Our pipeline surmounts the common challenges of integrating iScan genotypes with pre-existing data.

Keywords: Infinium, BeadArray, BeadChip, bead chip, SNP discovery, genotyping, great apes

INTRODUCTION

Single-nucleotide polymorphisms (SNPs) are a powerful tool for population genetic studies. In contrast with mainstay mitochondrial and microsatellite markers, SNPs can be generated at higher quality and with broader genome coverage and provide equivalent or greater statistical power for downstream studies (Morin et al., 2004). High-density SNP arrays are especially simple and cost-effective for the study of model organisms. In contrast with sequencing approaches, SNP arrays have built-in SNP redundancy (Oliphant et al., 2002) and call genotypes by averaging over multiple calls to increase accuracy. Moreover, they uniformly genotype all individuals at the exact same loci. Commercial arrays are widely available, particularly for association studies in humans (Ha et al., 2014), to develop breeding programs for livestock (Goddard and Hayes, 2009), and to facilitate

crop improvement (Gupta et al., 2008). These arrays can be purchased for independent use or for application at service laboratories.

For non-model organisms, however, discovering a panel of informative SNPs can be expensive, time-consuming, and methodologically complex. Non-targeted reduced-representation sequencing approaches, such as RADSeq (Baird et al., 2008), ddRAD (Peterson et al., 2012), and genotyping-by-sequencing (GBS) (Elshire et al., 2011), can allow for finding species-specific markers on a large scale, but can suffer higher error rates than microarrays. Increasing the number of samples in a single next-generation-sequencing run also comes at the expense of decreased coverage per locus. Lower coverage can result in error rates > 2%, yielding SNPs not useful for kinship and GWAS studies (Fountain et al., 2016). Even if SNPs are successfully discovered, genotyping these on a larger scale is likely to be prohibitive: both PCR- and sequencing-based methods are either expensive (e.g., next-generation sequencing panels, or dual-probe TaqMan assays) or impractical for large sample sizes (e.g., Melt Analysis of Mismatch Amplification Mutation Assay, or Sanger sequencing). Designing and manufacturing a custom SNP chip is also unlikely to be practical, given the minimum number of chips that must be ordered. For example, Illumina's custom Infinium iSelect chips require a commitment of at least 1,152 samples, with chips manufactured in 24-sample format and comprising 3,072–700,000 markers—this will not be cost-effective for all but the largest of studies. The required buy-in can become even more inhibitive if the chosen SNPs do not amplify consistently or provide data of insufficient quality; this issue is especially problematic when genotyping degraded samples (von Thaden et al., 2020), or when the SNP markers were chosen from a small population subset.

Cross-species application of commercial SNP arrays might therefore be considered as a means to rapidly genotype SNPs at low cost and with limited equipment and skills (Miller et al., 2012). This approach to SNP discovery has been previously used in reindeer (*Rangifer tarandus*) with the BovineSNP50 and OvineSNP50 chips, respectively, intended for cattle and sheep (Kharzinova et al., 2015); Antarctic fur seals (*Arctocephalus gazella*), with the CanineHD BeadChip intended for dogs (Hoffman et al., 2013); bighorn (*Ovis canadensis*) and thinhorn (*O. dalli*) sheep, with the OvineSNP50 chip (Miller et al., 2010); and in Arabian (*Oryx leucoryx*) and scimitar-horned oryx (*O. dammah*) with the BovineSNP50 array (Ogden et al., 2011). Their success comes in varying degrees, as the number of polymorphic SNPs obtained can be expected to decline proportionately with phylogenetic distance (Miller et al., 2010). Furthermore, SNP discovery with a small sample size often results in ascertainment bias, skewing the discovery of accurate F_{ST} values to obtain population informative SNPs (Trask et al., 2011; Quinto-Cortés et al., 2018). However, this limitation has not diminished the utility of cross-species SNP-chip application. Notably, the Bovine50K SNP chip was successfully used for SNP discovery in deer (*Odocoileus* spp.), despite the >25 million-year divergence between their taxonomic families (Haynes and Latch, 2012).

A key barrier to broader adoption of the cross-species approach is that most commercial arrays produce data in proprietary formats. In particular, Illumina's Infinium assays must be processed on their iScan System platform, producing IDAT-format files that record the scanner intensities for each probe on the array. These files are intended to be opened in Illumina's proprietary GenomeStudio software, to cluster and filter human-readable genotypes—though open-source IDAT parsing tools have since been written to produce the same outcome (e.g., Smith et al., 2013). Yet most cross-species studies will require their data in VCF format, to merge with data from other populations (e.g., from published studies). GenomeStudio can export variants as a GenomeStudio text file in four strand orientations—Illumina's top-bottom, plus-minus, forward-reverse, or probe-target. Illumina's top-bottom system was designed to allow for integration even if the reference allele changes in dbSNP or the human reference, but it is often difficult to understand (Guo et al., 2014). GenomeStudio also allows for data to be exported as a PLINK report (comprising .ped and .map files) following the top-bottom format (Purcell et al., 2007), or as an Affymetrix GeneSpring text file following the dbSNP forward strand format, but even using the dbSNP format means that not all SNPs are on the plus strand. There is no way to export a VCF that maintains the standard format and guarantees correct reference alleles for the target species. It is perhaps not coincidental, therefore, that none of the previously cited studies that used microarrays merged their genotypes with pre-existing data derived from non-microarray-based methods for comparative studies. On the contrary, each study analyzed the microarray data as a "closed" population, greatly limiting the utility of these genotypes.

Here, we provide guidance for selecting the most appropriate BeadChip for cross-species use, and for pre-processing the resulting IDAT files in GenomeStudio and PLINK. We then present a custom, cross-platform Python 3 script—iScanVCFMerge.py—that can be used to merge iScan microarray data with a pre-existing VCF comprising genotypes from other sources or samples. To demonstrate the efficacy of our script, we merged iScan data derived from 58 chimpanzees (*Pan troglodytes*), eight gorillas (*Gorilla* spp.), and 82 orang-utans (*Pongo* spp.) generated in this study with publicly available VCFs derived from whole-genome sequencing endeavors (Prado-Martinez et al., 2013). We show that microarrays for non-target species are an ideal tool for rapid and inexpensive SNP discovery.

MATERIALS AND EQUIPMENT

Use of our pipeline requires Illumina microarray data in IDAT format; the accompanying software program, Illumina GenomeStudio (RRID:SCR_010973); and our custom script, iScanVCFMerge.py (RRID:SCR_021193), which was tested with Python 3.9 (RRID:SCR_008394). The script is available both on GitHub[1] and for installation as a Python package (i.e., pip install iScanVCFMerge). Though we describe methods for generating

[1]https://www.github.com/baneslab/

IDAT data (e.g., from great ape blood and tissue samples), this protocol is applicable to IDAT data generated from any cross-species application of Illumina bead chips.

METHODS

DNA Extraction, Quantification, and Bead Chip Selection

We collected whole blood (N = 81) or tissue (N = 4) samples from 85 orang-utans (*Pongo* spp.) in zoos in the United States (N = 65), China (N = 18), and the Philippines (N = 2); whole blood from 58 chimpanzees (*Pan troglodytes*) in Chinese zoos; and whole blood from eight Western lowland gorillas (*Gorilla gorilla*) in United States zoos, from 2013 to 2018. Blood was drawn into EDTA Vacutainers during routine veterinary examinations or through voluntary blood-draw training. Tissue was collected during necropsy and stored in tubes or Whirl-paks (Nasco). All samples were stored at −20°C following collection. Genomic DNA was extracted from whole blood samples using the Promega ReliaPrep^{TM} Blood gDNA Miniprep System or using the Promega Maxwell RSC Blood DNA Kit; tissue samples were extracted using the Promega Maxwell RSC Tissue DNA Kit. Extractions utilizing Maxwell RSC kits were automated on the eponymous instrument. We followed the manufacturer's standard protocols for all extractions, with one modification for tissue samples: we performed an initial overnight digestion in Tail Lysis Buffer (Promega).

We quantified DNA *via* qPCR on a Roche LightCycler 480 instrument, with SYBRGreen qPCR Master Mix [*sensu* (Fünfstück et al., 2014)] and primers targeting an 81-bp portion of the c-*myc* proto-oncogene (Morin et al., 2001). Conditions comprised an initial denaturation of 10 min at 95°C; followed by 40 cycles of 10 s at 95°C, 10 s at 60°C, and 10 s at 72°C; concluding with one cycle of 10 s at 95°C, 60 s at 65°C, and 15 s at 95°C. We derived standard curves from serially diluted human genomic DNA (Promega). Extracts were then processed on the Illumina iScan platform, following the manufacturer's standard protocols. To select the best chip for use in each species, the probe sequences were obtained from the .bed files provided by Illumina, which we mapped to the human hg18 genome. We then used BLAST to compare the probe sequences from five of Illumina's commercial Infinium human microarrays (Core 24, Omni 2.5, Omni 5, OmniExpress, and Multi-Ethnic Global) to each species' reference genome. We chose the chip with the highest proportion of total probes with the single best hit, proportional to the total size of the manifest. Subsequently, we hybridized orang-utan DNA to the Illumina Infinium Multi-Ethnic Global Bead Chip (61.27% single best hit) and chimpanzee DNA to the Illumina Infinium Omni 2.5 Bead Chip (83.21% single best hit). As Illumina probe sequences are designed from the human transcriptome, we considered these values best estimates of on-target probes.

iScan Data Analysis

We analyzed the resulting IDAT files separately for each species in GenomeStudio 2.0. A detailed description of all abbreviations for iScan quality filters is presented in **Supplementary Table 1**. We first visualized sample performance by plotting the call rate against the P10 GC value; selected any samples that fell outside the majority cluster of samples; and excluded these poorly performing samples (i.e., a call rate below 0.98). After updating SNP statistics, we then filtered out SNPs based on low call quality: those that did not clearly cluster into heterozygotes and homozygotes (based on a Cluster Sep score < 0.3) and those for which more than 25% lacked calls across samples. We again updated SNP statistics, re-clustered all remaining SNPs, and exported the resulting new cluster positions as a custom cluster file for downstream analyses.

Using the custom cluster, we reanalyzed the IDAT files by first visualizing sample performance as above. After updating SNP statistics, we then filtered out SNPs based on low call quality: Cluster Sep score < 0.3 and those for which more than 10% lacked calls across samples. As this study only utilized autosomal SNPs, we filtered out all those on the X, Y, and mitochondrial chromosomes. Next, we filtered those with an AB R Mean < 0.12 (mean of the normalized intensity—R—values for the AB genotypes) and an AB T Mean < 0.15 or >0.85 (mean of the normalized theta values of the heterozygous cluster); i.e., clustered too closely to the homozygous clusters. As the majority of our SNPs were homozygous across all individuals, we filtered SNPs with a Minor Allele Frequency (MAF) > 0.01 and <0.8. Finally, we updated SNP statistics and exported the resulting data in three formats: GenomeStudio Final Report (tab-delimited.txt) using top-bottom strand, PLINK (.ped; Purcell et al., 2007), and GeneSpring (.txt; Agilent Technologies).

Data exported in GenomeStudio and PLINK formats report the reference alleles using top–bottom strand reference. To convert the SNPs to positive strand format, we used the custom script by Robertson (2012) and the Strand and Position Files for each chip as presented by Rayner and McCarthy (2011). After converting the SNPs to the same strand, we then exported the SNPs from PLINK in VCF format for downstream analysis.

Merging iScan Calls With Reference VCF Files

We used the VCF files published by Prado-Martinez et al. (2013), who re-sequenced whole genomes from animals sourced across the natural range of the genera and mapped these to the human hg18 (NCBI Build 36.1, GCF_00000145.12) reference genome. As our iScan chips were in hg19 (GRCh37.p13, GCF_000001405.25) format, we used Picard[2] to lift-over the VCFs from hg18 to hg19. For orang-utans, we merged the separate species-specific VCFs into a single VCF using bcftools (Li, 2011).

Our script, iScanVCFMerge.py, is designed to merge two VCF files of any format into a single VCF based on matches of chromosome, position, and certain conditions of major and minor alleles. Matched rows in the two VCFs are concatenated into a single row in the output files. The concatenated row comprises data for all individuals in both VCFs. This process allows the individuals from multiple populations to be analyzed in the same dataset.

[2]http://broadinstitute.github.io/picard

Usage: iScanVCFMerge -R reference_file.vcf -I iScan_file.vcf -O output_directory.

The first VCF file (-R, –reference_vcf) should comprise the pre-existing genotypes and will be used as the source of reference for REF and ALT alleles. This step is necessary because GenomeStudio assigns the REF and ALT based on the minor allele frequencies of the population genotyped and not based on a reference genome (i.e., that of the species genotyped). Inevitably, these REF and ALT alleles will not always match; particularly when only small subsections or subpopulations of a species are typed. This VCF file must include a header. The second VCF file (-I, –iScan_vcf) should comprise the novel iScan genotypes, in which the REF and ALT alleles will be updated. A header is not required and would in any case be removed by the program: contig values exported by GenomeStudio and/or PLINK are computed from the BeadChip and will not match the true species' reference genome. Input VCF files can be in either uncompressed (.vcf) or gzipped (.gzip) format; no index or dictionary files are needed. The script will run substantially faster if the input files are sorted; however, lexicographical sorting of both VCFs is performed irrespective.

At the script's execution, both VCFs are read into data frames, and only those positions shared between each file are retained for further processing. Because GenomeStudio and PLINK list chromosomes numerically, in contrast with newer reference genomes, the script first checks for a "chr" prefix in the iScan VCF and adds this where missing. Duplicate positions in that VCF are then dropped: this step is essential, as Illumina iScan microarrays often include duplicate or multiple probes for the same position in their design. All INDELs in the iScan VCF are then dropped, as—unlike single nucleotide variants—these may require further *in vitro* validation cross-species. The iScan VCF is then checked for other GenomeStudio or PLINK anomalies that might occur during pre-processing, e.g., CHROM or POS positions with values of zero. Additional FORMAT and INFO tags are dropped, as they become inapplicable following the merge, though the ID field is retained—if present—from the iScan VCF. Thereon, each position is evaluated for the following cases, prior to one of the four subsequent actions:

Case 1: The positions are biallelic and the alleles in both VCFs match exactly.

That is, the REF and ALT in both the reference and the iScan VCF files are exactly the same. The individuals are all merged into a single row with the major and minor alleles unchanged.

Case 2: The positions are biallelic and the alleles in both VCFs match exactly when reversed.

The reference file's alleles are used as a reference and samples from this VCF are unchanged. Genotypes in the second file are re-coded to conform to the mirrored state of the REF and ALT alleles inferred by GenomeStudio. For example, where the reference VCF states REF = A and ALT = T, the iScan VCF would state REF = T and ALT = A; thus, the genotypes in that file would be flipped.

Case 3: The positions are multi-allelic; the major (REF) alleles match exactly, but the ALT allele of the iScan VCF matches an alternate allele of the reference VCF.

The reference file's alleles are used as a reference and samples from this VCF are unchanged. Genotypes from the iScan VCF are re-coded to refer to the necessary ALT allele of the reference VCF. For example, where the reference VCF file states REF = G and ALT = T,A,C, and the iScan VCF states REF = G and ALT = C, an iScan genotype of 1/1 would be re-coded to 1/3.

Case 4: The positions are multi-allelic; the ALT allele of the iScan VCF exactly matches the REF allele of the reference VCF, but the REF allele of the iScan VCF matches either the tri- or quad-ALT allele of the refeence VCF.

The reference file's alleles are used as a reference and samples from this VCF are unchanged. Genotypes from the iScan VCF are first flipped, and then re-coded to refer to the appropriate REF and ALT alleles of the reference VCF. For example, where the reference VCF file states REF = G and ALT = T,A,C and the iScan VCF states REF = C and ALT = G, an iScan genotype of 0/1 would be re-coded to 1/3.

At completion, the script will output four files containing the passing variants, plus a fifth in which all are merged for downstream analysis (merged.vcf): exact_matches_biallelic.vcf and exact_matches_multiallelic.vcf, containing either bi- or multi-allelic genotypes that matched the reference REF and ALT (or one of the ALTS) exactly; and exact_matches_rev_biallelic.vcf and exact_matches_rev_multiallelic.vcf, comprising those where the iScan REF and one ALT allele matched those of the reference once reversed. A sixth file, rejected.vcf, contains all positions that did not match, and was therefore dropped. The script reports progress and outputs summary statistics of all loci processed.

RESULTS

Following re-clustering in GenomeStudio, we recorded on-target genotyping rates of 95% for chimpanzees and 70% for gorillas and orang-utans. In total, we genotyped 2,382,209 SNPs in chimpanzees and 1,748,250 SNPs in gorillas and orang-utans (**Table 1**). Of these, the majority were homozygous, as expected, with some SNPs in which all samples were heterozygous for the same alleles: 94% for chimpanzees, 96% for gorillas, and 95% for orang-utans.

We retained all chimpanzee and gorilla samples for analyses but removed three orang-utan samples that could not cluster

TABLE 1 | On-target genotyping rates and SNP statistics for each species, including the number of reported SNPs (i.e., those previously reported in other studies based on whole-genome sequencing in the target species) and unreported SNPs (i.e., newly discovered SNPs detected in this study, using microarrays) observed in each of the retained polymorphic SNP datasets.

Species	On-target genotyping rate	Total SNPs obtained	Total number polymorphic SNPs	After merging: number of reported SNPs	After merging: number of unreported SNPs
Chimpanzee	95%	2,382,209	48,831	24,255	24,576
Orang-utan	70%	1,748,250	47,536	20,362	27,174
Gorilla	70%	1,748,250	44,389	17,305	27,084

correctly. After removing homozygous and purely heterozygous SNPs and filtering for MAF, we were left with 48,831 polymorphic SNPs for chimpanzees, 47,536 polymorphic SNPs for gorillas, and 44,389 polymorphic SNPs for orang-utans (**Table 1**).

After merging with iScanVCFMerge, our final chimpanzee VCF matched 49.6% of the published SNPs (24,255); thus, 50.4% of our SNPs were previously unreported. Our final gorilla VCF matched 36.4% of the published SNPs (17,305); thus 63.6% were newly discovered. Our final orang-utan VCF matched 45.9% of the published SNPs (20,362); thus, 54.1% of our SNPs were novel (**Table 1**). The majority of the remaining SNPs were lost during merging due to chromosome and position mismatches, i.e., SNPs were not genotyped at the same location in both the public and the iScan data. Two SNPs were rejected for chimpanzees due to REF and ALT mismatches at a chromosome and position, 28 SNPs were rejected for gorillas, and 53 SNPs were rejected for orang-utans.

TABLE 2 | Comparative costs of SNP discovery approaches in gorillas, considering either sequencing or BeadChip scanning costs, for either microarrays (this study), whole-genome sequencing (X), and reduced-representation sequencing (X).

Method	Input DNA	No. animals sequenced	No. chips or lanes	Number of variable SNPs	Average cost (USD)
Illumina iScan	200 ng (50 ng/μL)	8	1	47,536	$256
Whole genome (Prado-Martinez et al., 2013)	2 μg (50 ng/μL)	31	125	13,731,122	$350,834
Reduced Representation (Scally et al., 2013)[a]	1 μg	14	12	3,006,670	$41,298

[a]*Statistics were determined from the 12 individuals published under NCBI BioProject PRJEB2590, for which one individual was sequenced per lane.*
The microarray approach required lower input DNA volumes and was substantially cheaper than the other approaches. Cost estimates were based on UW-Madison Biotechnology Center pricing (for iScan) or Genohub average pricing (http://www.genohub.com/; for Illumina sequencing), using the same instruments, read lengths, minimum coverage, and fragment sizes as detailed in the cited studies.

DISCUSSION

Our findings reiterate that microarrays can be applied across species, and that—when utilizing our scripted pipeline—novel SNPs can be recovered and merged for downstream analyses with pre-existing data. Our polymorphic SNP recovery rates were slightly higher than in previous studies: 6% of all loci in chimpanzees, 4% in gorillas, and 5% in orang-utans, despite the former having diverged from our common ancestor *c.* 5 mya, *c.* 10 mya (Scally et al., 2012), and *c.* 14 mya, respectively (Locke et al., 2011). In contrast, the OvineSNP50 BeadChip—designed for domestic sheep (*Ovis aries*)—yielded 570 polymorphic SNPs in bighorn sheep (1.82% of the 48,230 genotyped) and 330 SNPs in thinhorn sheep (0.69% of the 48,004 genotyped), despite their much closer evolutionary history. The effect of species divergence on loci recovery emphasizes the importance of selecting the most appropriate chip. In our case, multiple human chips were available and assessed for their single best hit against the great ape genomes. In contrast, commercial sheep microarrays are less abundant, and are designed to detect recently arisen mutations useful in discerning domestic sheep breeds (Miller et al., 2010).

The utility of cross-species microarray data will depend on the yield of polymorphic SNPs. While whole-genome sequencing (for example) may yield a greater number, the lower input DNA quantities and scanning (vs. library preparation and sequencing) costs offset the disadvantage of lower yields from microarrays. In gorillas, for example, genome-wide SNPs have been obtained from whole-genome sequencing (Prado-Martinez et al., 2013), reduced representation sequencing (Scally et al., 2013), as well as with microarrays (this study). When comparing the number of polymorphic SNPs vs. input DNA and cost of sequencing, our cross-species microarray approach was substantially cheaper (**Table 2**).

Nonetheless, with only small numbers of SNPs, it can be difficult to calculate LD and runs of homozygosity (ROH), which are needed for inferring kinship or to perform QTL and GWAS studies. With a medium-density (50K) SNP array, the number of short ROH can be overestimated even when using microarrays in the species they were designed for (Ferenčaković et al., 2013; Szmatoła et al., 2020). Though it may be tempting to include all SNPs, rather than filter out monomorphic SNPs, this will falsely raise homozygosity estimates and can lead to assumptions of inbreeding—as was observed when using the Bovine50 chip to study LD in reindeer (Shafer et al., 2016). Further, large gaps in SNP coverage can lead to the detection of false ROH islands, most likely caused by ROH detection algorithms not detecting short gaps in the flanking regions of the ROH (Nandolo et al., 2018). Nonetheless, in most non-model studies, microarray data are analyzed as private populations—meaning polymorphisms when compared to other populations cannot be detected. Our pipeline might therefore serve to increase the utility of microarray data from prior studies, if used to merge their outputs with pre-existing genotypes. The present version of iScanVCFMerge does not address the creation of tri-allelic data (i.e., creating a tri-allele when the iScan population has an allele present that is not present in the publicly accessed data). In a future iteration, however, this capability could facilitate discovery of rare alleles and kinship-informative alleles only found in the study population.

CONCLUSION

Cross-species application of microarrays is a rapid, cost-effective approach for SNP discovery in non-model organisms. The use of Illumina microarrays has to date been hampered by an inability to export genotypes into VCF and combine these with a pre-existing VCF comprising additional data. Our pipeline, utilizing our custom script—iScanVCFMerge—facilitates the simple and rapid merging of such files, enabling the detection of novel SNP loci and increasing the likelihood of observing polymorphic sites.

DATA AVAILABILITY STATEMENT

The great-ape microarray data featured in this study are used as demo data with which to demonstrate the efficacy of our pipeline and script. Because the data were derived from zoo-housed animals, restrictions apply to their availability, as their source biomaterials were used under license for the current study. Data may be available from the corresponding author upon reasonable request and with the permission of each licensor. The pipeline can otherwise be independently verified using any iScan-derived dataset.

ETHICS STATEMENT

The animal study was reviewed and approved by the Chinese Academy of Sciences.

AUTHOR CONTRIBUTIONS

EF, L-CZ, Q-XL, E-LP, Y-HY, and GB designed the study. GB, L-CZ, IF, EFR, E-LP, and Y-HY collected and processed the biological samples. EF, GB, L-CZ, AK, J-YY, QZ, and X-LZ performed the laboratory work. EF, JM, and GB performed the computational analyses and wrote the script. EF and GB wrote the manuscript. All authors read and approved the final manuscript and agreed to be accountable for the content of the work.

FUNDING

This research was funded by the Shanghai Municipal Forestry Department (to Y-HY and E-LP), the Arcus Foundation, the Ronna Noel Charitable Trust, The Eppley Foundation for Research, Inc., The Orang-utan Conservation Genetics Trust (now The Orang-utan Conservation Genetics Project, Inc.) (all to GB), and the Avilon Wildlife Conservation Foundation (to EFR). AK was supported by the Morris Animal Foundation. This study capitalized on the computing resources and assistance of the UW-Madison Center for High Throughput Computing (CHTC) in the Department of Computer Sciences, which is supported by UW-Madison, the Advanced Computing Initiative, the Wisconsin Alumni Research Foundation, the Wisconsin Institute for Discovery, and the National Science Foundation. In doing so, the study utilized the Open Science Grid, which is supported by the National Science Foundation and the U.S. Department of Energy's Office of Science. Research reported in this publication was also supported in part by the Office of the Director, National Institutes of Health, under Award Number P51OD011106 to the Wisconsin National Primate Research Center, University of Wisconsin–Madison. In turn, this was conducted in part at a facility constructed with support from Research Facilities Improvement Program grant numbers RR15459-01 and RR020141-01. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. Funding bodies had no role in the design of the study; the collection, analysis, and interpretation of data; or in writing the manuscript.

ACKNOWLEDGMENTS

We would like to thank the University of Wisconsin Biotechnology Center – Gene Expression Center, Madison, WI, United States; the Carver Biotechnology Center, University of Illinois at Urbana-Champaign, Champaign, IL, United States; Genergy Bio, Shanghai, China; and the University of the Philippines – Diliman, Quezon City, Philippines, for facilitating the molecular work. Samples for this project were provided by the following zoos: ABQ BioPark, Zoo Atlanta, Audubon Zoo, Birmingham Zoo, Cameron Park Zoo, Cheyenne Mountain Zoo, Chicago Zoological Society – Brookfield Zoo, Cleveland Metroparks Zoo, Columbus Zoo, St Paul's Como Park Zoo and Conservatory, Fresno Chaffee Zoo, Fort Wayne Children's Zoo, Gladys Porter Zoo, Greenville Zoo, Utah's Hogle Zoo, Indianapolis Zoo, Little Rock Zoo, Zoo Miami, Milwaukee County Zoo, Smithsonian's National Zoo, Oklahoma City Zoo, Oregon Zoo, Philadelphia Zoo, Phoenix Zoo, Rolling Hills Zoo, Sacramento Zoo, Sedgwick County Zoo, and Seneca Park Zoo (United States); Anji Zhongnan Baicao Yuan, Chongqing Zoo, Hangzhou Wild Animal Park, Shanghai Zoo (China); and AVILON Zoo (Philippines). We thank the Chinese Association of Zoological Gardens (CAZG) and the Philippine Zoos and Aquariums Association (PHILZOOS) for facilitating sample collection in their countries, and the Orangutan, Gorilla and Chimpanzee Species Survival Plan (SSP) Steering Committees for their approval by recommendation for sample collection in US zoos.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fevo.2021.629252/full#supplementary-material

REFERENCES

Baird, N. A., Etter, P. D., Atwood, T. S., Currey, M. C., Shiver, A. L., Lewis, Z. A., et al. (2008). Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS One* 3:e3376. doi: 10.1371/journal.pone.0003376

Elshire, R. J., Glaubitz, J. C., Sun, Q., Poland, J. A., Kawamoto, K., Buckler, E. S., et al. (2011). A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS One* 6:e19379. doi: 10.1371/journal.pone.0019379

Ferenčaković, M., Sölkner, J., and Curik, I. (2013). Estimating autozygosity from high-throughput information: effects of SNP density and genotyping errors. *Genet. Sel. Evol.* 45:42. doi: 10.1186/1297-9686-45-42

Fountain, E., Pauli, J., Reid, B., Palsbøll, P., and Peery, M. (2016). Finding the right coverage: the impact of coverage and sequence quality on single nucleotide polymorphism genotyping error rates. *Mol. Ecol. Resour.* 16, 966–978. doi: 10.1111/1755-0998.12519

Fünfstück, T., Arandjelovic, M., Morgan, D. B., Sanz, C., Breuer, T., Stokes, E. J., et al. (2014). The genetic population structure of wild western lowland gorillas (*Gorilla gorilla* gorilla) living in continuous rain forest. *Am. J. Primatol.* 76, 868–878. doi: 10.1002/ajp.22274

Goddard, M. E., and Hayes, B. J. (2009). Mapping genes for complex traits in domestic animals and their use in breeding programmes. *Nat. Rev. Genet.* 10, 381–391. doi: 10.1038/nrg2575

Guo, Y., He, J., Zhao, S., Wu, H., Zhong, X., Sheng, Q., et al. (2014). Illumina human exome genotyping array clustering and quality control. *Nat. Protoc.* 9, 2643–2662. doi: 10.1038/nprot.2014.174

Gupta, P. K., Rustgi, S., and Mir, R. R. (2008). Array-based high-throughput DNA markers for crop improvement. *Heredity* 101, 5–18. doi: 10.1038/hdy.2008.35

Ha, N.-T., Freytag, S., and Bickeboeller, H. (2014). Coverage and efficiency in current SNP chips. *Eur. J. Hum. Genet.* 22, 1124–1130. doi: 10.1038/ejhg.2013.304

Haynes, G. D., and Latch, E. K. (2012). Identification of novel single nucleotide polymorphisms (SNPs) in deer (*Odocoileus* spp.) using the BovineSNP50 BeadChip. *PLoS One* 7:e36536. doi: 10.1371/journal.pone.0036536

Hoffman, J. I., Thorne, M. A. S., McEwing, R., Forcada, J., and Ogden, R. (2013). Cross-amplification and validation of SNPs conserved over 44 million years between seals and dogs. *PLoS One* 8:e68365. doi: 10.1371/journal.pone.0068365

Kharzinova, V. R., Sermyagin, A. A., Gladyr, E. A., Okhlopkov, I. M., Brem, G., and Zinovieva, N. A. A. (2015). Study of applicability of SNP chips developed for bovine and ovine species to whole-genome analysis of reindeer *Rangifer tarandus*. *J. Hered.* 106, 758–761.

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

Locke, D. P., Hillier, L. W., Warren, W. C., Worley, K. C., Nazareth, L. V., Muzny, D. M., et al. (2011). Comparative and demographic analysis of orang-utan genomes. *Nature* 469, 529–533.

Miller, J. M., Kijas, J. W., Heaton, M. P., McEwan, J. C., and Coltman, D. W. (2012). Consistent divergence times and allele sharing measured from cross-species application of SNP chips developed for three domestic species. *Mol. Ecol. Resour.* 12, 1145–1150. doi: 10.1111/1755-0998.12017

Miller, J. M., Poissant, J., Kijas, J. W., Coltman, D. W., and ISG Consortium (2010). A genome-wide set of SNPs detects population substructure and long range linkage disequilibrium in wild sheep. *Mol. Ecol. Resour.* 11, 314–322. doi: 10.1111/j.1755-0998.2010.02918.x

Morin, P. A., Chambers, K., Boesch, C., and Vigilant, L. (2001). Quantitative polymerase chain reaction analysis of DNA from noninvasive samples for accurate microsatellite genotyping of wild chimpanzees (*Pan troglodytes verus*). *Mol. Ecol.* 10, 1835–1844. doi: 10.1046/j.0962-1083.2001.01308.x

Morin, P. A., Luikart, G., Wayne, R. K., and The SNP Workshop Group (2004). SNPs in ecology, evolution and conservation. *Trends Ecol. Evol.* 19, 208–216. doi: 10.1016/j.tree.2004.01.009

Nandolo, W., Utsunomiya, Y. T., Mészáros, G., Wurzinger, M., Khayadzadeh, N., Torrecilha, R. B. P., et al. (2018). Misidentification of runs of homozygosity islands in cattle caused by interference with copy number variation or large intermarker distances. *Genet. Sel. Evol.* 50:43.

Ogden, R., Baird, J., Senn, H., and McEwing, R. (2011). The use of cross-species genome-wide arrays to discover SNP markers for conservation genetics: a case study from Arabian and scimitar-horned oryx. *Conserv. Genet. Resour.* 4, 471–473. doi: 10.1007/s12686-011-9577-2

Oliphant, A., Barker, D. L., Stuelpnagel, J. R., and Chee, M. S. (2002). BeadArray TM technology: enabling an accurate, cost-effective approach to high-throughput genotyping. *Biotechniques* 32, S56–S61.

Peterson, B. K., Weber, J. N., Kay, E. H., Fisher, H. S., and Hoekstra, H. E. (2012). Double digest RADseq: an inexpensive method for de novo SNP discovery and genotyping in model and non-model species. *PLoS One* 7:e37135. doi: 10.1371/journal.pone.0037135

Prado-Martinez, J., Sudmant, P. H., Kidd, J. M., Li, H., Kelley, J. L., Lorente-Galdos, B., et al. (2013). Great ape genetic diversity and population history. *Nature* 499, 471–475.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., et al. (2007). A tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* 81, 559–575. doi: 10.1086/519795

Quinto-Cortés, C. D., Woerner, A. E., Watkins, J. C., and Hammer, M. F. (2018). Modeling SNP array ascertainment with approximate Bayesian Computation for demographic inference. *Sci. Rep.* 8:10209.

Rayner, N. W., and McCarthy, M. I. (2011). *Genotyping Chips Strand and Build Files*. Available online at: https://www.well.ox.ac.uk/~{}wrayner/strand/ (accessed June 4, 2020).

Robertson, N. (2012). *Update_Build.sh*. Available online at: https://www.well.ox.ac.uk/~{}wrayner/strand/update_build.sh (accessed June 4, 2020).

Scally, A., Dutheil, J. Y., Hillier, L. W., Jordan, G. E., Goodhead, I., Herrero, J., et al. (2012). Insights into hominid evolution from the gorilla genome sequence. *Nature* 483, 169–175.

Scally, A., Yngvadottir, B., Xue, Y., Ayub, Q., Durbin, R., and Tyler-Smith, C. (2013). A genome-wide survey of genetic variation in gorillas using reduced representation sequencing. *PLoS One* 8:e65066. doi: 10.1371/journal.pone.0065066

Shafer, A. B. A., Miller, J. M., and Kardos, M. (2016). Cross-species application of SNP chips is not suitable for identifying runs of homozygosity. *J. Hered.* 107, 193–195. doi: 10.1093/jhered/esv137

Smith, M. L., Baggerly, K. A., Bengtsson, H., Ritchie, M. E., and Hansen, K. D. (2013). illuminaio: an open source IDAT parsing tool for Illumina microarrays. *F1000Research* 2:264. doi: 10.12688/f1000research.2-264.v1

Szmatoła, T., Gurgul, A., Jasielczuk, I., Fu, W., and Ropka-Molik, K. (2020). A detailed characteristics of bias associated with long runs of homozygosity identification based on medium density SNP microarrays. *J. Genomics* 8, 43–48. doi: 10.7150/jgen.39147

Trask, J. A. S., Malhi, R. S., Kanthaswamy, S., Johnson, J., Garnica, W. T., Malladi, V. S., et al. (2011). The effect of SNP discovery method and sample size on estimation of population genetic data for Chinese and Indian rhesus macaques (*Macaca mulatta*). *Primates* 52, 129–138. doi: 10.1007/s10329-010-0232-4

von Thaden, A., Nowak, C., Tiesmeyer, A., Reiners, T. E., Alves, P. C., Lyons, L. A., et al. (2020). Applying genomic data in wildlife monitoring: development guidelines for genotyping degraded samples with reduced single nucleotide polymorphism panels. *Mol. Ecol. Resour.* 20, 662–680. doi: 10.1111/1755-0998.13136

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: 04 June 2021
doi: 10.3389/fevo.2021.671281

Historical Demographic Processes Dominate Genetic Variation in Ancient Atlantic Cod Mitogenomes

Lourdes Martínez-García[1†], Giada Ferrari[1†], Tom Oosting[2†], Rachel Ballantyne[3†], Inge van der Jagt[4], Ingrid Ystgaard[5,6], Jennifer Harland[7], Rebecca Nicholson[8], Sheila Hamilton-Dyer[9], Helle Tessand Baalsrud[1†], Marine Servane Ono Brieuc[1†], Lane M. Atmore[1†], Finlay Burns[10], Ulrich Schmölcke[11], Kjetill S. Jakobsen[1†], Sissel Jentoft[1†], David Orton[12†], Anne Karin Hufthammer[13], James H. Barrett[3,6†] and Bastiaan Star[1*†]*

[1] Department of Biosciences, Centre for Ecological and Evolutionary Synthesis (CEES), University of Oslo, Oslo, Norway, [2] School of Biological Sciences, Victoria University of Wellington, Wellington, New Zealand, [3] McDonald Institute for Archaeological Research, Department of Archaeology, University of Cambridge, Cambridge, United Kingdom, [4] Department of Archaeology, Cultural Heritage Agency of the Netherlands, Amersfoort, Netherlands, [5] Department of Historical and Classical Studies, Norwegian University of Science and Technology (NTNU), Trondheim, Norway, [6] Department of Archaeology and Cultural History, Norwegian University of Science and Technology (NTNU) University Museum, Norwegian University of Science and Technology, Trondheim, Norway, [7] Archaeology Institute, University of the Highlands and Islands, Orkney, United Kingdom, [8] Oxford Archaeology, Oxford, United Kingdom, [9] Department of Archaeology & Anthropology, Bournemouth University, Poole, United Kingdom, [10] Marine Scotland Science, Aberdeen, United Kingdom, [11] Centre for Baltic and Scandinavian Archaeology (ZBSA), Schleswig, Germany, [12] BioArCh, Department of Archaeology, University of York, York, United Kingdom, [13] Department of Natural History, The University Museum, University of Bergen, Bergen, Norway

Ancient DNA (aDNA) approaches have been successfully used to infer the long-term impacts of climate change, domestication, and human exploitation in a range of terrestrial species. Nonetheless, studies investigating such impacts using aDNA in marine species are rare. Atlantic cod (*Gadus morhua*), is an economically important species that has experienced dramatic census population declines during the last century. Here, we investigated 48 ancient mitogenomes from historical specimens obtained from a range of archeological excavations in northern Europe dated up to 6,500 BCE. We compare these mitogenomes to those of 496 modern conspecifics sampled across the North Atlantic Ocean and adjacent seas. Our results confirm earlier observations of high levels of mitogenomic variation and a lack of mutation-drift equilibrium—suggestive of population expansion. Furthermore, our temporal comparison yields no evidence of measurable mitogenomic changes through time. Instead, our results indicate that mitogenomic variation in Atlantic cod reflects past demographic processes driven by major historical events (such as oscillations in sea level) and subsequent gene flow rather than contemporary fluctuations in stock abundance. Our results indicate that historical and contemporaneous anthropogenic pressures such as commercial fisheries have had little impact on mitogenomic diversity in a wide-spread marine species with high gene flow such as Atlantic cod. These observations do not contradict evidence that overfishing has had negative consequences for the abundance of Atlantic cod and the

OPEN ACCESS

Edited by:
Melissa T. R. Hawkins, Smithsonian Institution, United States

Reviewed by:
Martin Taylor, University of East Anglia, United Kingdom
Athanasios Exadactylos, University of Thessaly, Greece

***Correspondence:**
Lourdes Martínez-García
l.m.garcia@ibv.uio.no
Bastiaan Star
bastiaan.star@ibv.uio.no

†ORCID:
Lourdes Martínez-García
orcid.org/0000-0002-1582-3611
Giada Ferrari
orcid.org/0000-0002-0850-1518
Tom Oosting
orcid.org/0000-0002-7031-0747
Rachel Ballantyne
orcid.org/0000-0002-6506-3163
Helle Tessand Baalsrud
orcid.org/0000-0002-4161-3247
Marine Servane Ono Brieuc
orcid.org/0000-0001-8601-2122
Lane M. Atmore
orcid.org/0000-0002-8903-8149
Kjetill S. Jakobsen
orcid.org/0000-0002-8861-5397
Sissel Jentoft
orcid.org/0000-0001-8707-531X
David Orton
orcid.org/0000-0003-4069-8004
James H. Barrett
orcid.org/0000-0002-6683-9891
Bastiaan Star
orcid.org/0000-0003-0235-9810

Specialty section:
This article was submitted to Evolutionary and Population Genetics, a section of the journal Frontiers in Ecology and Evolution

Received: *26 February 2021*
Accepted: *06 May 2021*
Published: *04 June 2021*

importance of genetic variation in implementing conservation strategies. Instead, these observations imply that any measures toward the demographic recovery of Atlantic cod in the eastern Atlantic, will not be constrained by recent loss of historical mitogenomic variation.

Keywords: population structure, fisheries, human exploitation, phylogenomics, population expansion, demographic history

INTRODUCTION

Continuous human activities and a changing climate have influenced terrestrial and marine ecosystems for millennia (Venter et al., 2016; Rodrigues et al., 2019; Mitchell and Rawlence, 2021), impacting the evolutionary potential and population demography of a range of species (Seersholm et al., 2018). Ancient mitochondrial DNA (mtDNA) has been widely used to understand long-term genomic consequences of such impacts (Shapiro et al., 2004; Nyström et al., 2006; Stiller et al., 2010; Paijmans et al., 2013; Fortes and Paijmans, 2015; Casas-Marce et al., 2017). Nonetheless, most ancient mtDNA studies have focused on terrestrial species, and studies that investigate the impacts of long-term human activities and/or climatic variation on fish, using whole genome sequencing approaches, are relatively rare. Long-term commercial fisheries—covering many centuries—have contributed to the decline of economically and ecologically important marine species (Exadactylos et al., 2007; Pinnegar and Engelhard, 2008; Barrett, 2019). The consequences of intensive fishing in recent times may be difficult to assess as this requires an understanding of historical population dynamics (Selim et al., 2016). The analysis of long-term biological and demographic fluctuations can therefore help to improve guidelines for sustainable fisheries management and optimal conservation measures (Barrett, 2019). In order to provide a long-term perspective on fishing exploitation impacts, the use of archeological evidence, such as fish bone remains, is essential for those periods for which little or no historical data are available. Recent developments in whole genome aDNA methods now allow the inference of demographic histories and the estimation of genetic fluctuations over time from fishbone samples (Oosting et al., 2019; Ferrari et al., 2021). Such combined molecular analyses of historical and modern samples can potentially provide an understanding of the association between human-environmental impact and population declines (Hofman et al., 2015).

Several studies have shown the utility of temporal mtDNA analyses in the marine environment. For instance, ancient mitogenomes have investigated impacts of climate and hunting on the Atlantic walrus (Star et al., 2018; Keighley et al., 2019; Barrett et al., 2020), narwhals (Louis et al., 2020), and the extinct great auk (Thomas et al., 2019). In fish, such studies remain limited to partial mitogenome data. For example, a shift in sturgeon species distributions was detected during the Holocene in the North East Atlantic based on CytB amplicon data (Nikulina and Schmölcke, 2016). Moreover, impacts of habitat destruction and human activities during the 1800s were associated with a reduction of the mtDNA diversity of Chinook salmon from the Columbia River in the 12S and control region by comparing ancient and modern samples (Johnson et al., 2018). Similarly, impacts of human exploitation and climate oscillations were associated with losses of haplotypic CytB variation in Atlantic cod during the 15th to 16th centuries in Iceland (Olafsdottir et al., 2014). In contrast, comparable levels of ancient mtDNA genetic diversity were found between ancient and modern samples of herring specimens, despite continuous human exploitation (Speller et al., 2012). Notwithstanding these examples, human-environmental impacts and population declines remain unclear for a wide range of marine species and populations.

Atlantic cod (*Gadus morhua* L. 1758) is a benthopelagic predatory fish with high reproductive rates and with a fundamental ecological role in marine ecosystems (Barth et al., 2017; Edvardsson et al., 2019). It has been one of the most exploited fish species in the North Atlantic Ocean (Carr et al., 1995; Árnason et al., 2000; Nicholls et al., 2021). The distribution of this species extends through the cold waters of North America, across the continental shelves of Greenland and Iceland, and northern Europe (Lait et al., 2018). Relatively large population sizes have been characteristic throughout its entire distribution even during the expansion of long-distance fish trading during the 12th to 13th centuries in the eastern Atlantic and at the beginning of the 16th century in the western Atlantic (Barrett et al., 2004, 2011; Orton et al., 2014; Castañeda et al., 2020). However, intensive fishing activities during the 20th century (Mieszkowska et al., 2009; Jonsson et al., 2016; Brattey et al., 2018) resulted in the severe depletion of several stocks, for instance the North Sea stock, which was decimated from annual landings of 354,000 to 50,000 tons during this period (Bannister, 2004). In addition to past human exploitation, climatic events like the Little Ice Age—a cooling period that varied regionally in timing and duration but occurred between ca. 1300–1850 CE—may have caused large declines between the sixteenth and 17th centuries (Edvardsson et al., 2019).

The genomic consequences of such population dynamics and declines in Atlantic cod remain unclear. Based on partial and whole mtDNA data, Atlantic cod populations between the western and eastern Atlantic Ocean show significant structure (Árnason, 2004; Jørgensen et al., 2018; Lait et al., 2018), whereas low to no mtDNA differentiation has been found across a wide range of eastern Atlantic locations (Carr et al., 1995; Árnason and Palsson, 1996; Árnason et al., 1998, 2000; Sigurgíslason and Árnason, 2003). Here, we compared modern and ancient Atlantic cod mitogenomes—dated up to 6500 BCE—from different fishing locations in northern Europe. We evaluated whether Atlantic cod in the eastern Atlantic have experienced any loss

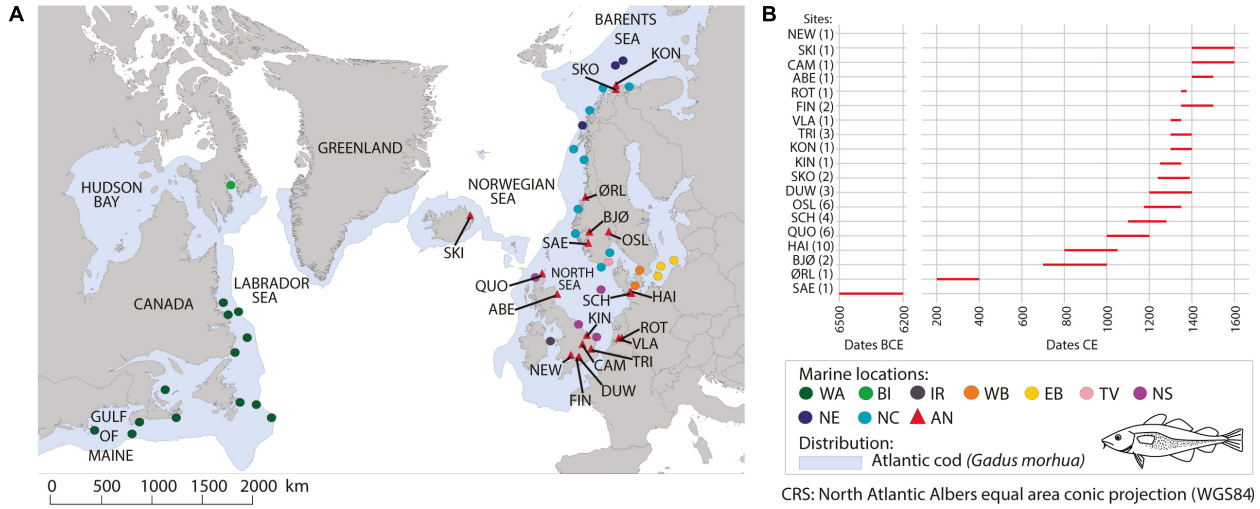

FIGURE 1 | (A) Ancient and modern Atlantic cod sampling locations. The abbreviated name of archeological sites is shown in the map (see also **Supplementary Table 1**). Modern data were obtained from Jørgensen et al. (2018); Lait et al. (2018) and Barth et al. (2019). Sample sites are grouped according to larger oceanographic location, ecotype and modern or ancient specimens: western Atlantic (WA) = dark green, Baffin Island (BI) = light green, Irish Sea (IS) = gray, North Sea (NS) = purple, western Baltic (WB) = orange, eastern Baltic (EB) = yellow, Tvedestrand fjord (TV) = pink, North East Arctic (NE) = dark blue, Norwegian coast (NC) = blue, and Ancient specimens (AN) = red (i.e., Saevarhelleren = SAE, Ørland Kampflybase = ØRL, Bjørkum = BJØ, Haithabu = HAI, Quoygrew = QUO, Schleswig Schild = SCH, Oslo Mindets Tomt = OSL, Bristol Dundas Wharf = DUW, Skonsvika = SKO, King's Lynn Raynham House = KIN, Kongshavn = KON, London Trig Lane = TRI, Vlaardingen Gat in de Markt = VLA, Bristol Finzel's Reach = FIN, Rotterdam, Hoogstraat 13–26 = ROT, Aberdeen = ABE, Cambridge Grand Arcade = CAM, Skriðuklaustur = SKI, Newport Ship = NEW). Modern distribution of Atlantic cod is shown in light blue. Fish illustration drawn by Geir Holm. **(B)** Cleveland plot showing the age range of ancient samples. Sample age is estimated based on archeological context. The number of individuals per site is represented in brackets after the site name (**Supplementary Table 1**).

of genetic variation, analyzed long term patterns of effective population size, and related any observed decline to the impact of commercial fisheries or climate change.

MATERIALS AND METHODS

Sample Collection

Ancient samples of Atlantic cod (n = 48) were obtained from 11 excavation sites (**Figure 1** and **Supplementary Table 1**) and were stored dry and unfrozen. The specimens were all supplied by the relevant archeological organizations, or sampled with permission on their premises. The shipment of Atlantic cod bones does not require CITES or other wildlife regulation permits for transport or analysis. Where practicable, only a subsample of bone was employed for the aDNA research, leaving material for other studies. Dating of the samples (**Supplementary Table 1**) was based on archeological context. Ancient samples were morphologically and genetically identified as Atlantic cod. A total of 472 available modern mitogenomes were obtained from Jørgensen et al. (2018), Lait et al. (2018), and Barth et al. (2019). Novel mtDNA sequence data from modern specimens sampled in 2016 in Orkney, United Kingdom (n = 24) were also included (**Figure 1** and **Supplementary Table 2**). The collection of the Orkney specimens complied with the Nagoya Protocol and Convention on Biological Diversity, which the United Kingdom signed up to in 2016. All specimens were deceased when the fin clip was collected.

DNA Extraction, Amplification and Sequencing

DNA extraction and library preparation from ancient samples were performed in the aDNA laboratory at the University of Oslo under rigorous measures (Cooper and Poinar, 2000; Gilbert et al., 2005). All ancient samples were processed with the same DNA extraction and library protocols according to Ferrari et al. (2021). In short, bones were UV-treated for 10 min per side and pulverized using a stainless-steel mortar (Gondek et al., 2018). Per specimen, two aliquots containing between 150 and 200 mg of bone powder were used as starting material for DNA extraction. Double-indexed blunt-end sequencing libraries were built from 15 to 16 µl of DNA extract using the Meyer-Kircher protocol (Meyer and Kircher, 2010; Kircher et al., 2012) with the modifications listed in Schroeder et al. (2015) and the single-tube (BEST) protocol (Carøe et al., 2018) with the modifications described in Mak et al. (2017). Sequencing reads were processed using PALEOMIX v1.2.13 (Schubert et al., 2014). Trimming of residual adapter contamination, filtering and collapse of reads was done using AdapterRemoval v.2.1.7 (Lindgreen, 2012). Sequencing reads shorter than 25 bp were discarded. Mapping of remaining reads was performed against the Atlantic cod GadMor3.0 nuclear genome (RefSeq assembly accession GCF_902167405.1; Star et al., 2011; Tørresen et al., 2017) and mitochondrial genome (Johansen and Bakke, 1996) using BWA v.0.7.12 (Li and Durbin, 2009) with the aln algorithm, disabled seeding and minimum quality score of 25. The resulting BAM files were indexed with *samtools* v.1.9 (Li et al., 2009) and

DNA postmortem damage assessed using MapDamage v.2.0.9 (Jónsson et al., 2013). DNA from modern Orkney samples were extracted using a DNeasy Blood & Tissue kit (Qiagen). Libraries were assembled with a TrueSeq DNA PCR-Free Preparation Kit and sequenced on an Illumina HiSeq 2,500. Modern alignment—including Orkney and Barth et al. (2019) samples, and the outgroup Alaska pollock (*Gadus chalcogrammus*; Malmstrøm et al., 2016) – was carried out using BWA v.0.7.12 with the mem algorithm, and a minimum quality score of 25.

Mitogenomic Analysis

Variant calling was performed using GATK v.4.1.4. (McKenna et al., 2010) simultaneously in all ancient, modern Orkney and Barth et al. (2019) samples, including the outgroup. gVCF files were created for each sample using GATK HaplotypeCaller (with ploidy set to 1). Individual genotypes were combined in one file using GATK CombineGVCFs and GenotypeGVCFs. Filtering was performed using *bcftools* v.1.9. (Li et al., 2009) and *vcftools* v.0.1.16. (Danecek et al., 2011) with the following thresholds: FS < 60.0, SOR < 4, MQ > 30.0, QD > 2.0, SnpGap = 10, minGQ = 15, minDP = 3, remove indels = yes, biallelic loci = yes, meanDP < 30 and read depth > 3. Consensus sequences were built using *bcftools* consensus and aligned using MAFFT v.7.429 (Katoh and Standley, 2013). Available modern mitogenomes obtained from Jørgensen et al. (2018) and Lait et al. (2018) were manually inspected using MEGA v.7 (Kumar et al., 2016) to set as missing the consistent nucleotide differences (between 50 and 100%; **Supplementary Table 3**) between their Illumina, Sanger and/or Roche 454 technologies with the Illumina sequenced mitogenomes in this study. Control region and half of the tRNA-Pro sequences from all the mitogenomes were excluded from further analyses as these two regions were not fully complete (i.e., 15,696–15,815 positions) after aligning sequences obtained from Jørgensen et al. (2018) and Lait et al. (2018) with the sequences presented in this study and Barth et al. (2019) samples. Thus, all sequences analyzed had 15,695 bp in length. Validated SNPs were annotated as transversion and/or transition using *SNP-sites* (Page et al., 2016). Checked and modified modern sequences (Jørgensen et al., 2018; Lait et al., 2018) were added and aligned to our multi-fasta alignment using MAFFT v.7.429. Unique sequences were identified with IQTREE v.1.6.12 (Nguyen et al., 2015). File formats required for different software and/or packages were obtained with seqinr and ape (i.e., nexus format; Paradis and Schliep, 2019; Charif et al., 2020), and phylotools (i.e., phylip format; Zhang et al., 2017) packages implemented in R.

Different sample combinations were used to compare the genetic diversity of the ancient samples to those of the modern conspecifics. Given the low spatial structure in the eastern Atlantic region (Árnason and Palsson, 1996; Árnason et al., 1998; Sigurgíslason and Árnason, 2003) and lack of consistent spatial structure amongst specimens (**Supplementary Figures 3, 5, 6**), all 48 ancient samples were compared as a single group to modern samples grouped into larger marine locations (according to their geographical proximity or ecotype; **Figure 1**). In addition, a comparison of subsets of multiple specimens from two archeological locations (Quoygrew and Haithabu) for which a more specific temporal pair from the same geographical region could be identified, was performed (**Supplementary Table 1**). Quoygrew specimens were locally sourced (Harland and Barrett, 2012; Star et al., 2017). Therefore, modern specimens sampled in the same area (i.e., modern Orkney) provide a logical, spatially consistent temporal comparison. However, specimens from Haithabu, were sourced from northern Norway (Star et al., 2017), and belonged to the North East Arctic ecotype. For these traded specimens, the North East Arctic ecotypes provide a spatially relevant temporal comparison, rather than North Sea or western Baltic specimens.

Haplotype (h) and nucleotide diversities (π), number of haplotypes (Nh) and number of polymorphic sites (S) were calculated using DnaSP v.6 (Rozas et al., 2017). To allow direct comparison with earlier CytB results (Árnason, 2004; Olafsdottir et al., 2014; Jørgensen et al., 2018), specific CytB haplotypes based on 250 bp gene fragment as previously reported by Árnason (2004) were identified using MEGA v.7. Demographic histories were determined by Tajima's D (TD) and Fu's F (F) neutrality in DnaSP v.6. A different number of specimens were obtained for ancient and modern locations. We corrected for such differences in sample size by randomly downsampling the modern specimens for each of the temporal comparisons (North East Arctic and Orkney) using 1,000 bootstrap replicates. A 95% confidence interval of the genetic parameters; genetic variation (π) and patterns of population demography (TD and F) was calculated from these 1,000 bootstrap replicates that were sampled using a without replacement approach with the *sample* function implemented in R (R Core Team, 2020) and the *fasta.sample* function in the FastaUtils package also in R (Salazar, 2020). For the bootstrapping test, π, TD and F from temporally spaced modern locations were re-calculated with the *pegas* (Paradis, 2010) and *PopGenome* (Pfeifer et al., 2020) packages implemented in R. Relationships among ancient and modern samples were visualized for whole mitogenome and CytB sequence data, by constructing a mitochondrial haplotype-genealogy graph using Fitchi (Matschiner, 2016) with the ML-based phylogenetic tree obtained with IQTREE v.1.6.12 as input.

Population Dynamics and Demographic Reconstruction

An analysis of molecular variance (AMOVA, 1,000,000 permutations) and population pairwise genetic distances (Φ_{ST}) were obtained in Arlequin v.3.5 (Excoffier and Lischer, 2010), to determine the distribution of variation between marine locations and temporally spaced locations. Divergence and coalescent analyses were based on unique sequences only (n = 525 sequences including the outgroup). Substitution model selection for unique sequences was performed using PHYML v.3.1 (Guindon et al., 2010) as implemented in JMODELTEST v.2.1.10 (Guindon and Gascuel, 2003; Darriba et al., 2012). Model selection was determined on the following partitions: 1st, 2nd, and 3rd codons from protein coding regions, rRNAs and tRNAs. Best-fitting models were selected according to the Aikaike Information Criterion (AIC; **Supplementary Table 4**). Based on these results, phylogenetic estimates were obtained using BEAST v.2.6.3 (Bouckaert et al., 2019).

Bayesian settings for all phylogenetic analyses included two sets of partitions: coding region and non-coding region. Three independent runs to test for chain convergence were run under the Coalescent Constant Population Tree Prior. Tip ages (ancient and modern dates) were included for each set of runs (**Supplementary Tables 1, 2**). Sample dates for ancient specimens were rounded to a midpoint date—from a given range—where necessary. To achieve high effective sample sizes (ESS = $\geq$ 200), chain lengths were run 800,000,000 under a substitution rate of 1.14×10^{-8} substitution/site/year as per Lait (2016) assuming a GTR + I (for coding regions) and TIM1 + I (for non-coding regions) models of evolution and a strict clock. Tracer v.1.71 (Rambaut et al., 2018) was used to check for convergence of MCMC and to ensure sufficient sampling. Consensus trees were obtained using TreeAnnotator v.2.6.2—implemented in BEAST v.2.6.3—after 10% burn-in. Final phylogenetic trees were viewed and edited in FigTree v.1.4.4.

Finally, a Coalescent Bayesian Skyline (CBS) analysis was completed to reconstruct the demographic history—including female effective population size (*Ne*) —of Atlantic cod through time. To assess any confounding effect of past or contemporary population structure (Heller et al., 2013), we analyzed demographic history using 6 different data sets (excluding the outgroup): (I) all 524 sequences, (II) 476 modern sequences (excluding 48 ancient samples), (III) 273 sequences (excluding clades associated with most western Atlantic and Baltic Sea samples), (IV) 368 sequences (excluding the clade associated with most Baltic Sea samples), (V) 429 sequences (excluding clades associated with most western Atlantic samples) and (VI) 48 ancient sequences (excluding all modern samples). The specific clades that were excluded in III, IV and V can be found in **Supplementary Figure 4**. We used the same MCMC sampling procedure described before with 3 independent runs reaching convergence at high effective sample sizes (ESS = $\geq$ 200). Chain lengths were run 800,000,000 for data sets I, II and V with a number of bPopSize and bGroupSize of 10; while chain length for data sets III and IV were run 500,000,000 and 50,000 for data set VI with a number of bPopSize and bGroupSize of 5.

RESULTS

Mitogenomic Variation

Sequencing reads from all ancient specimens showed the expected patterns of DNA fragmentation and deamination rates that were consistent with those of authentic aDNA (**Supplementary Figure 1**). We obtained 48 mitogenomes with at least 3-fold average coverage. We also obtained mitogenomes for 24 modern Orkney specimens (**Supplementary Table 2**). A total of 2135 SNPs (∼13% of mitogenome positions) were identified among all 545 samples – including the outgroup species Alaska pollock –: 1219 SNPs corresponded to informative sites and 916 SNPs were singletons (**Supplementary Table 5**).

Nucleotide diversity (π) between modern locations ranged between 0.002 and 0.003 (**Table 1**) and π of ancient samples did not vary from the values obtained in modern locations. The temporal comparison of specific sites (Quoygrew-Orkney and Haithabu-North East Arctic), showed limited significant differences between genetic statistics of temporally spaced ancient and modern locations (**Supplementary Table 6** and **Supplementary Figure 2**), where Haithabu has significantly lower π and higher F compared to the North East Arctic (**Supplementary Figure 2**).

Neutrality tests showed significant negative values for all Tajima's D (*TD*) and *F* statistics in most locations, except for the western location Baffin Island, and the eastern locations Tvedestrand fjord and western Baltic (**Table 1**). Overall, there were 486 haplotypes —- including the outgroup—across all 545 samples, of which only 26 were shared between individuals (**Figure 2** and **Supplementary Table 7**). Ancient CytB variation consisted of 7 different haplotypes, including four main haplotypes (A, C, D, and E) previously identified in modern mtDNA studies (Árnason, 2004; Jørgensen et al., 2018). Two novel variations of existing CytB haplotypes were found in western Baltic (haplotype ED) and North Sea (haplotype LI), while another 2 novel variations of existing CytB haplotypes were found among ancient samples (haplotypes LJ and TI). The most prevalent ancient haplotypes were A and E (∼ 40 and 38%, respectively, **Supplementary Tables 1, 7**), which were also commonly found in modern samples (**Supplementary Table 8**). The haplotype genealogy for whole mitogenome and CytB sequence data showed an extensive distribution of ancient samples across marine locations (**Figure 2** and **Supplementary Figure 4**). Limited geographic mitogenome structure was observed, except for elevated divergence between western Atlantic and eastern Atlantic locations, and between locations in the western and eastern Baltic Sea and other eastern Atlantic locations (**Figures 2B,C** and **Supplementary Figures 4B,C**). A star-like topology is observed for the whole mitogenome and CytB genealogies (**Figure 2** and **Supplementary Figure 4**).

Demographic Patterns and Population Structure

The AMOVA assigned 7.58% of the variation between marine locations (including ancient samples as a single group) while 91.47% of the variation was represented between individuals (ΦCT = 0.076, $p \leq 0.001$; ΦST = 0.085, $p \leq 0.000$). Pairwise Φ_{ST} values (**Figure 3** and **Supplementary Tables 9, 10**) showed significant differentiation levels between all ancient samples and western Atlantic, Baffin Island, western Baltic, eastern Baltic and Tvedestrand fjord. Ancient samples showed higher differentiation when compared to western Atlantic (Φ_{ST} = 0.117), and Baffin Island (Φ_{ST} = 0.192) in comparison to other eastern Atlantic locations. Among modern samples, western Atlantic, Baffin Island, western Baltic and eastern Baltic showed significant Φ_{ST} values when compared to all other locations (**Supplementary Tables 9, 10**). Φ_{ST} values were not significant between North Sea, North East Arctic, Norwegian coast and Ancient samples. Pairwise Φ_{ST} values between temporal spaced locations also showed no significant differentiation (Quoygrew and modern Orkney: Φ_{ST} = 0.000; p = 0.807; and Haithabu and North East Arctic: Φ_{ST} = 0.000; p = 0.456).

TABLE 1 | Estimates of genetic diversity statistics for Atlantic cod at whole mitogenomes from different marine locations or ecotypes in the North Atlantic (see **Supplementary Tables 1, 2, 5**).

	Location	Code	*N*	*h*	*Nh*	*S*	π	*TD*	*F*
Modern	western Atlantic	WA	124	1.000	124	759	0.002	–2.652*	–5.668*
	Baffin Island	BI	18	0.791	7	68	0.002	1.570	1.688
	western Baltic	WB	43	0.996	40	221	0.002	–1.719	–2.992*
	eastern Baltic	EB	36	1.000	36	249	0.002	–1.888*	–3.305*
	Tvedestrand (fjord)	TV	37	0.982	31	256	0.002	–1.570	–2.757*
	North Sea	NS	99	0.999	96	678	0.003	–2.350*	–4.421*
	North East Arctic	NE	97	0.999	92	716	0.003	–2.408*	–4.695*
	Norwegian coast	NC	41	1.000	41	377	0.002	–2.187*	–3.694*
Ancient	Ancient	AN	48	0.998	46	364	0.002	–2.283*	–3.906*

*Significant values are indicated with * ($p \leq 0.01$). N, sample size; h, haplotype diversity; Nh, number of haplotypes; S, number of polymorphic sites; π, nucleotide diversity; TD, Tajima's D; F, Fu's F. Locations with 1 sample are excluded for genetic analysis (i.e., Irish Sea = IS).*

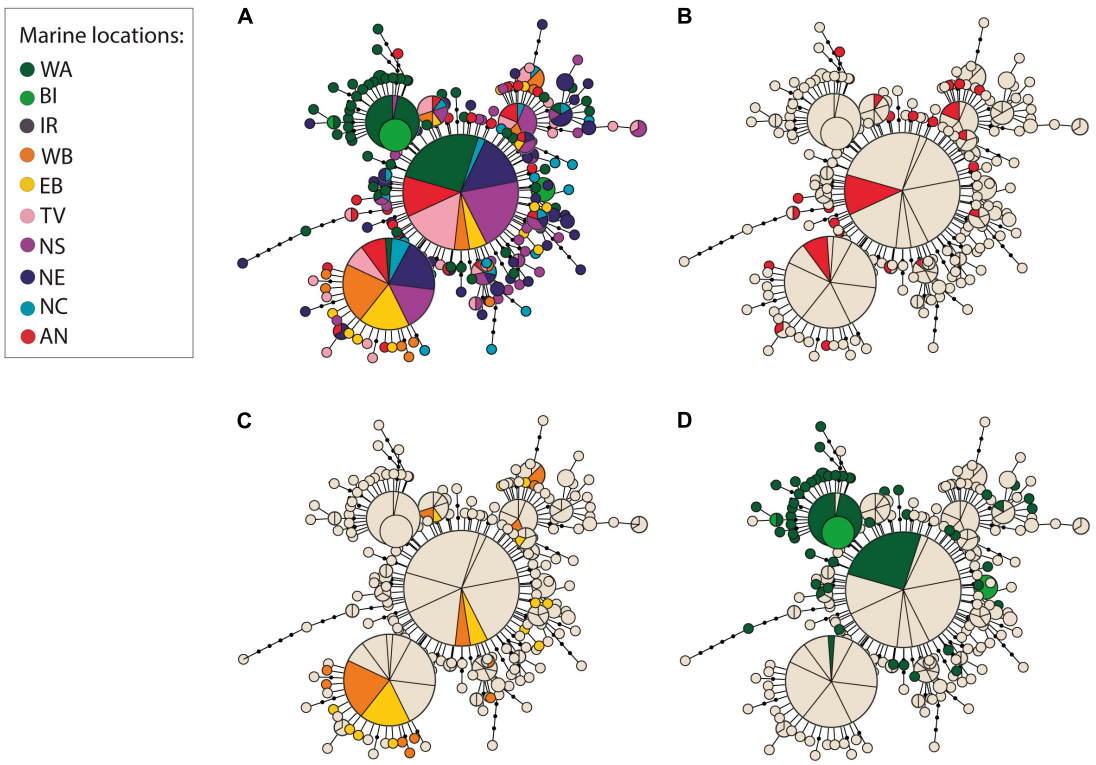

FIGURE 2 | Star-like haplotype genealogy based on complete mitogenomes of 545 samples of Atlantic cod. Only transversions (336) are depicted in this genealogy. **(A)** Haplotype genealogy colored by location. **(B)** Haplotype genealogy highlighting ancient specimens as a single group (in red; see **Figure 1** and **Supplementary Table 1**; for specific distribution of sampling sites of ancient samples see **Supplementary Figure 3**). **(C)** Haplotype genealogy highlighting samples from the western (orange) and eastern (yellow) Baltic. **(D)** Haplotype genealogy highlighting western Atlantic (green) and Baffin Island (light green) specimens. Circle size is proportional to haplotype frequency. Black dots indicate intermediate haplotypes not found among the specimens and black lines connecting black circles correspond to mutation steps between haplotypes.

The time-calibrated Bayesian phylogeny for ancient and modern Atlantic cod samples resulted in 2 main clades with an estimated divergence from the most recent common ancestor at 220 kya (95% highest posterior density (HPD) = 194,780—249,980 kya; **Figure 4**). The first clade, which is not further divided, includes mitogenomes from 6 different widely scattered localities. The second clade was composed by 16 subclades with posterior probability > 0.8, with divergence times of ca. 100 kya. Clades and subclades in the phylogeny were not geographically structured, with the exception of most samples from western Atlantic, and most samples from western and eastern Baltic, which clustered together (**Figures 2**, **4**).

The Bayesian skyline analysis using different subsets of the data revealed a consistent pattern of step-wise population expansions followed by periods of constant population size. Expansions around 150, 50, and 10 kya are present in most

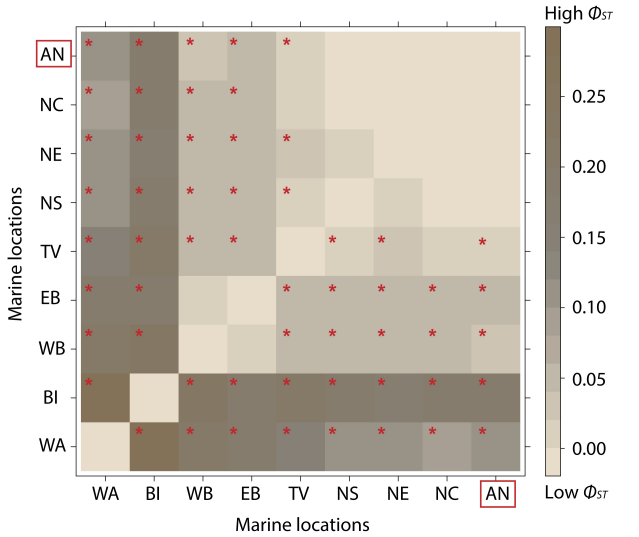

FIGURE 3 | Heatmap of differentiation pairwise (Φ_{ST}) between 10 marine locations of Atlantic cod: western Atlantic (WA), Baffin Island (BI), western Baltic (WB), eastern Baltic (EB), Tvedestrand fjord (TV), North Sea (NS), North East Arctic (NE), Norwegian coast (NC), and Ancient specimens (AN). Dark brown colors indicate higher levels of differentiation (Φ_{ST} = 0.15–0.30). Light brown corresponds to lower levels of differentiation (Φ_{ST} = ≥ 0.10). Significant *p*-values ($p \leq 0.05$) are indicated with a red *p-values are listed in **Supplementary Tables 9, 10**.

subsets (**Figure 5**). A population expansion of Atlantic cod was identified ca. 50 kya in all subsets. The most recent expansion identified (around 10 kya), is only present in data sets that include clades with most Baltic Sea specimens (**Figures 5A,B,E**). Despite such differences, all analyses agree with a high and increasing female effective population size (*Ne*) of Atlantic cod (*Ne* = ca. 1,000,000—10,000,000) during the last ca. 100 kya, with highest estimates of *Ne* during the last few millennia (**Figure 5**).

DISCUSSION

Here, we compared modern and ancient mtDNA diversity in Atlantic cod to investigate whether observed historical and contemporaneous census population declines (Hutchinson et al., 2003; Hylen et al., 2008; Limburg et al., 2008; Bartolino et al., 2012; Jonsson et al., 2016; Brattey et al., 2018) have had mitogenomic consequences. The temporal comparison of 48 ancient specimens to 496 modern conspecifics did not reveal consistent significant mitogenomic changes or measurable effective genetic population declines through time. Below, we discuss reasons why such genomic impacts may not be observed.

First, mitogenomic variation is high in modern Atlantic cod and is characterized by limited genetic differentiation between populations and incomplete lineage sorting over large spatial scales across its range in the North Atlantic (Jørgensen et al., 2018; Lait et al., 2018). Low observed genetic differentiation (Φ_{ST}) between Tvedestrand fjord and other Norwegian coastal locations, as well as between the North Sea, the North East Arctic and the Norwegian coast confirm this lack of geographic structuring over large parts of the eastern Atlantic (**Figure 3**). Indeed, the non-significant differentiation of all ancient samples with modern North Sea, North East Arctic and Norwegian coast is fully consistent with their presumed geographical origin and highlights the long-term lack of mtDNA structure in this region. Non-significant Φ*ST* values between the Norwegian coastal locations and Tvedestrand fjord indicate possible recent migration of fish between such coastal communities and more restricted fjord populations (Knutsen et al., 2011). Compared to many terrestrial ecosystems, where populations can often be isolated by physical barriers—which restrain interbreeding and dispersal—(Hauser and Carvalho, 2008; Exadactylos et al., 2019), in marine ecosystems the absence of physical barriers promotes larger panmictic populations and Atlantic cod is no exception (Berg et al., 2016, 2017; Sodeland et al., 2016; Barth et al., 2017). Thus, a combination of low spatial resolution of mtDNA data as a result of continuous gene flow and connectivity may mask any local temporal erosion of mitogenomic diversity (Welch et al., 2012) in Atlantic cod.

Second, we determined high long-term estimates of effective population size (*Ne* = ca. 1,000,000–10,000,000; **Figure 5**), which is in agreement with earlier observations in Atlantic cod (Hardie et al., 2006; Therkildsen et al., 2010; Pinsky et al., 2021). Estimates of *Ne* can remain high in economically important fish species, even if their populations have experienced a large biomass decline (Hauser and Carvalho, 2008) since it takes hundreds of generations (i.e., depending on the generation time of the species; Amos and Balmford, 2001; Frankham et al., 2002) for the actual population numbers and breeding populations to be reflected in *Ne* (Hauser and Carvalho, 2008). In fact, simulations have shown that a population with theoretical *Ne* of 100 (which is several orders of magnitude lower than observed in Atlantic cod) would retain 75% of heterozygosity after 57 generations (Frankham et al., 2002; Welch et al., 2012). Given that such population declines take a very long time to lead to measurable genomic consequences, mtDNA—as a single locus—will have limited power to record such changes in populations of high *Ne* (Allentoft et al., 2014; Johnson et al., 2018; Thomas et al., 2019; Spencer, 2020). The absence of significant genetic changes in this study is consistent with the absence of such changes in genome-wide data using historical samples of Atlantic cod from the western and eastern Atlantic (Pinsky et al., 2021) and with the absence of such changes in mitogenomic data from other taxa that have similarly high estimates of *Ne* as Atlantic cod, such as the Pacific herring (Speller et al., 2012; Moss et al., 2016), the Hawaiian petrel (Welch et al., 2012) and even extinct species such as the New Zealand moa (Allentoft et al., 2014), the passenger pigeon (Murray et al., 2017) and the great auk (Thomas et al., 2019).

In contrast, temporal losses of mitogenomic diversity and/or declines in *Ne* have been reported in species that have suffered population fragmentation (e.g., resulting in small effective population sizes) or that have experienced limited connectivity, such as the steppe bison (Shapiro et al., 2004), the Scandinavian arctic fox (Nyström et al., 2006), cave bears (Stiller et al., 2010), the Iberian lynx (Casas-Marce et al., 2017), the Iberian salmon

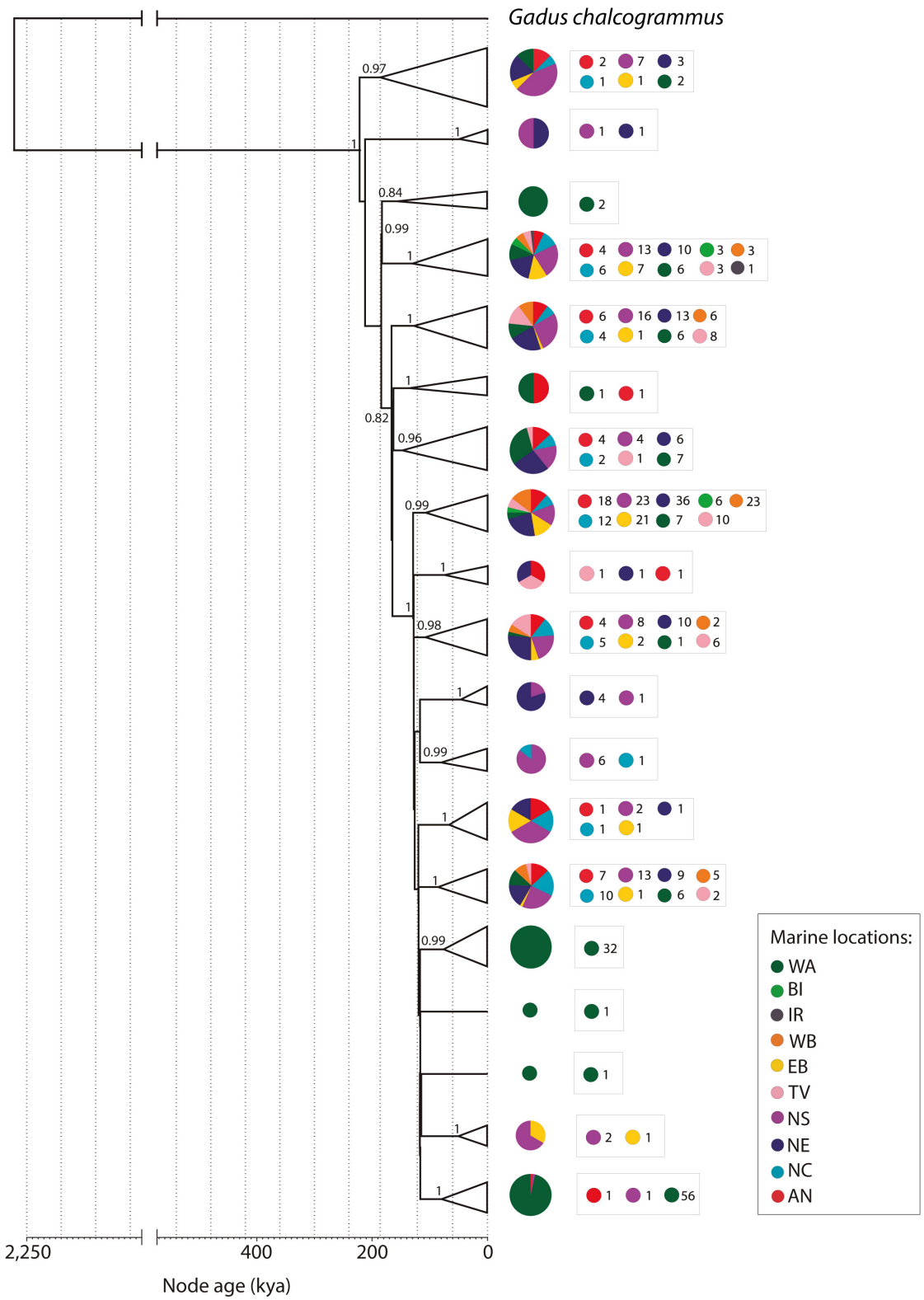

FIGURE 4 | Time calibrated collapsed Bayesian phylogeny of full mitogenomes from 525 Atlantic cod specimens using Alaska pollock (*Gadus chalcogrammus*) as an outgroup. Pie charts represent the marine locations distributed in each clade. Numbers beside pie charts indicate the number of individuals from each marine location distributed in each clade. Only branches with posterior probability > 0.8 are indicated next to the corresponding clade/subclade. For specific distribution of sampling sites of ancient samples see **Supplementary Figure 6**.

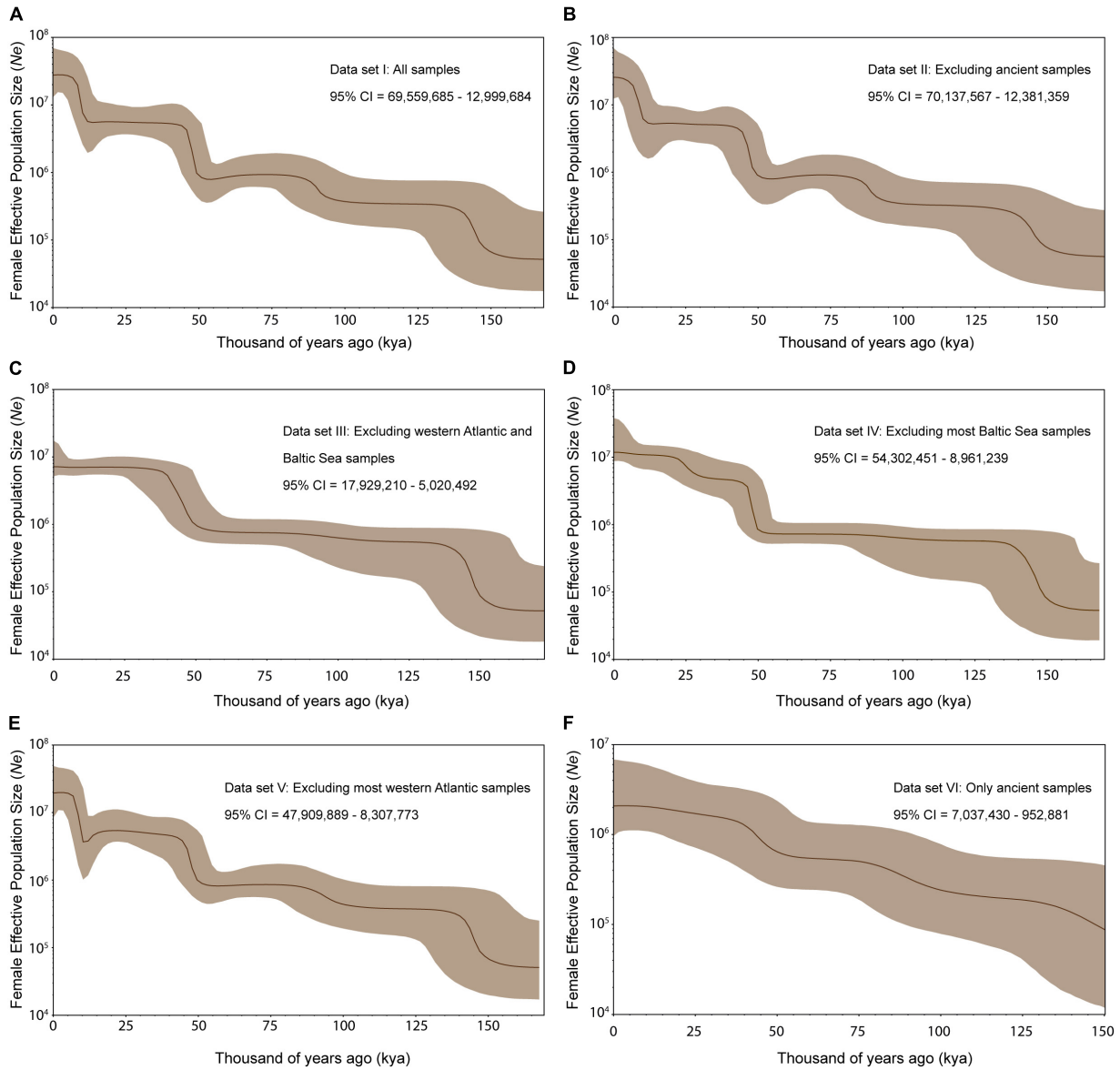

FIGURE 5 | Atlantic cod demographic history representing the fluctuations of effective female population size (*Ne;* in dark brown line) based on **(A)** data set I: all 524 sequences, **(B)** data set II: 476 modern sequences (excluding 48 ancient samples), **(C)** data set III: 273 sequences (excluding clades associated with most western Atlantic and Baltic Sea samples), **(D)** data set IV: 368 sequences (excluding the clade associated with most Baltic Sea samples), **(E)** data set V: 429 sequences (excluding clades associated with most western Atlantic samples) and **(F)** data set VI: 48 ancient sequences (excluding all modern samples). The 95% CI corresponds to the light brown area. The most recent 95% CI at time 0 is shown in each respective panel. Description of the clades excluded in data sets III, IV, and V can be found in **Supplementary Figure 7**.

(Consuegra et al., 2002) and the common bream (Ciesielski and Makowiecki, 2005). Interestingly, a loss of haplotypic variation has been identified—using CytB sequence data—for a single period (i.e., 15th to 16th centuries, out of 6 temporal periods investigated) in an Icelandic population of Atlantic cod (Olafsdottir et al., 2014). There are two potential explanations for this discrepancy. First, nearly all substitutions that comprise the CytB haplotypes can be affected by post-mortem deamination (i.e., they consist of C > T and G > A substitutions). Most of the ancient sequences (90%) investigated in Olafsdottir et al. (2014) were obtained in a single round of PCR without evaluation of such post-mortem deamination. Therefore, such bias due to post-mortem damage cannot be excluded. Second, our sampling does not include many specimens from Iceland (**Figure 1** and **Supplementary Table 1**), and it remains possible that—with 156 samples—a local effect has been observed in Olafsdottir et al. (2014), which we do not detect in our data.

Third, we do not observe major novel mtDNA lineages in the ancient data, nor observe a significant loss of such lineages over time. Instead, the majority of Atlantic cod mtDNA lineages

observed in ancient and modern samples today have originated ca. 100–150 kya (**Figure 4**), during a period of population expansion (**Figure 5**). Therefore, the gain of such lineages—and associated population expansions—in Atlantic cod is more likely caused by changes in abundance driven by major historical climatic events such as eustatic oscillations in sea level, and the interglacial and warming periods experienced during the last glacial maximum ca. 23,000 kya (Bigg et al., 2008) and the Wisconsinan (ca. 110–120 kya) and Illinoian (ca. 200–130 kya) glaciations (Gibbard and Van Kolfschoten, 2005) as described by Lait et al. (2018). For instance, we only observe the most recent population expansion ca. 10 kya (**Figures 5A,B,E**) when including those mtDNA clades which are strongly associated with the Baltic Sea. The timing of this expansion is in agreement with the development of the Baltic Sea (ca. 7,000–8,000 years; Ojaveer et al., 2010; Wenne et al., 2020) which has led to genetically distinct Atlantic cod populations that have adapted to local environmental conditions (i.e., salinity and temperature; Johannesson and Andre, 2006; Berg et al., 2015; Wenne et al., 2020). Therefore, the observed changes in *Ne* reflect past population demography rather than recent and contemporary demographic changes (Lombal et al., 2020).

It is clear from zooarcheological evidence that Atlantic cod has periodically experienced intense exploitation in the distant past, particularly around the North Sea and the Baltic Sea (Barrett et al., 1999; Enghoff, 1999; Olson and Walther, 2007; Orton et al., 2011). This fishing pressure became even greater in the 19th and 20th centuries (e.g., Thurstan et al., 2010). Landings of Atlantic cod exceeded 4,000,000 tons during 1960–1990s in the North Atlantic Ocean (Shelton and Morgan, 2014). In particular, landings surpassed 600,000 tons in Iceland by ca. 1930s (Drinkwater, 2006), 354,000 tons in the North Sea during ca. 1970s (Bannister, 2004), 200–400,000 tons in the eastern Baltic during 1960–1990s (MacKenzie et al., 2002), 650,000 tons in North East Arctic between 1937 and 1938 up to 800–1,200,000 tons in ca. 1950s (Sætersdal and Hylen, 1964; Hylen, 2002). Such high levels of exploitation led to major reductions in present abundances of most Atlantic cod populations (i.e., Food and Agriculture Organization [Fao], 2020-2021a,b). Nonetheless, for the reasons discussed above, our results indicate that such population declines of Atlantic cod did not lead to a detectable impact on the mtDNA genome on the time scale we investigated here.

Taken together, our results highlight that historical and contemporaneous anthropogenic pressures such as commercial fisheries have had little impact on the ancient mitogenomic diversity of a wide-spread marine species with high gene flow such as Atlantic cod. Future ancient DNA studies should consider the inclusion of nuclear genomic data and extensive sampling on a local scale—considering a temporal comparison of specimens from the same geographical region—to assess the effects of climate and human exploitation with greater statistical power. Finally, our observations do not contradict evidence that overfishing has had negative consequences for the abundance of Atlantic cod and they do not oppose information about the important implications of genetic variation in evolutionary biology, ecology and conservation biology. Instead, our observations suggest that conservation management measures aimed toward the demographic recovery of Atlantic cod in the eastern Atlantic, if achievable by conservation management measures, will not be constrained by recent loss of historical mitogenomic variation.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: https://www.ebi.ac.uk/ena, PRJEB42959.

AUTHOR CONTRIBUTIONS

BS and JB: conceptualization, project design, and supervision. GF and LM-G: laboratory work and data curation. LM-G with input from GF, TO, LA, BS, and JB: formal analysis. HB and MS: analytical advice. JB, IY, IJ, JH, RN, DO, SH-D, US, AH, and RB: ancient Atlantic cod specimens. FB: modern Orkney specimens. JB, IY, IJ, JH, RN, DO, SH-D, US, AH, and RB: archeological context information. LM-G and BS: data visualization. JB, BS, KJ, SJ, and DO: funding acquisition. LM-G and BS with input from JB: writing—original draft. All authors writing—review and editing.

FUNDING

This work was supported by Research Council of Norway projects "Catching the Past" (262777) and "The Aqua Genome Project" (221734), Leverhulme Trust Project MRF-2013-065 and the European Union's Horizon 2020 Research and Innovation Programme under the Marie Skłodowska-Curie grant agreement No. 813383.

ACKNOWLEDGMENTS

We thank Agata T. Gondek-Wyrozemska for help with processing the ancient specimens and Francis Neat for providing modern Orkney specimens. We also thank Oliver Kersten and Michael Matschiner for advice during analyses. Finally, we thank M. Skage, S. Kollias and A. Tooming-Klunderud at the Norwegian Sequencing Centre for sequencing and processing of samples. Analyses were performed on the SAGA Cluster using the resources and assistance from the SIGMA2 Metacenter, the Norwegian National Infrastructure for High Performance Computing and Data Storage.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fevo.2021.671281/full#supplementary-material

REFERENCES

Allentoft, M. E., Heller, R., Oskam, C. L., Lorenzen, E. D., Hale, M. L., Gilbert, M. T. P., et al. (2014). Extinct New Zealand megafauna were not in decline before human colonization. *Proc. Natl. Acad. Sci.U. S. A.* 111, 4922–4927. doi: 10.1073/pnas.1314972111

Amos, W., and Balmford, A. (2001). When does conservation genetics matter? *Heredity* 87, 257–265. doi: 10.1046/j.1365-2540.2001.00940.x

Árnason, E. (2004). Mitochondrial cytochrome b DNA variation in the high-fecundity Atlantic cod: trans-Atlantic clines and shallow gene genealogy. *Genetics* 166, 1871–1885. doi: 10.1093/genetics/166.4.1871

Árnason, E., and Palsson, S. (1996). Mitochondrial cytochrome b DNA sequence variation of Atlantic cod Gadus morhua, from Norway. *Mol. Ecol.* 5, 715–724. doi: 10.1111/j.1365-294x.1996.tb00368.x

Árnason, E., Petersen, P. H., Kristinsson, K., Sigurgíslason, H., and Pálsson, S. (2000). Mitochondrial cytochrome b DNA sequence variation of Atlantic cod from Iceland and Greenland. *J. Fish Biol.* 56, 409–430. doi: 10.1006/jfbi.1999.1167

Árnason, E., Petersen, P. H., and Pálsson, S. (1998). Mitochondrial cytochrome b DNA sequence variation of Atlantic cod, Gadus morhua, from the Baltic and the White Seas. *Hereditas* 129, 37–43. doi: 10.1111/j.1601-5223.1998.00037.x

Bannister, R. C. A. (2004). "The rise and fall of cod (*Gadus morhua, L.*) in the North Sea," in *Management of Shared Fish Stocks*, eds A. I. L. Payne, C. M. O'Brien, and S. I. Rogers (Oxford: Blackwell Publishing), 316–338. doi: 10.1002/9780470999936.ch19

Barrett, J. H. (2019). An environmental (pre) history of European fishing: past and future archaeological contributions to sustainable fisheries. *J. Fish Biol.* 94, 1033–1044. doi: 10.1111/jfb.13929

Barrett, J. H., Boessenkool, S., Kneale, C. J., O'Connell, T. C., and Star, B. (2020). Ecological globalisation, serial depletion and the medieval trade of walrus rostra. *Quat. Sci. Rev.* 229:106122. doi: 10.1016/j.quascirev.2019.106122

Barrett, J. H., Locker, A. M., and Roberts, C. M. (2004). 'Dark Age Economics' revisited: the English fish bone evidence AD 600-1600. *Antiquity* 78, 618–636. doi: 10.1017/s0003598x00113262

Barrett, J. H., Nicholson, R. A., and Cerón-Carrasco, R. (1999). Archaeo-ichthyological Evidence for Long-term Socioeconomic Trends in Northern Scotland: 3500 BC to AD 1500. *J. Archaeol. Sci.* 26, 353–388. doi: 10.1006/jasc.1998.0336

Barrett, J. H., Orton, D., Johnstone, C., Harland, J., Van Neer, W., Ervynck, A., et al. (2011). Interpreting the expansion of sea fishing in medieval Europe using stable isotope analysis of archaeological cod bones. *J. Archaeol. Sci.* 38, 1516–1524. doi: 10.1016/j.jas.2011.02.017

Barth, J. M., Berg, P. R., Jonsson, P. R., Bonanomi, S., Corell, H., Hemmer−Hansen, J., et al. (2017). Genome architecture enables local adaptation of Atlantic cod despite high connectivity. *Mol. Ecol.* 26, 4452–4466. doi: 10.1111/mec.14207

Barth, J. M. I., Villegas-Ríos, D., Freitas, C., Moland, E., Star, B., André, C., et al. (2019). Disentangling structural genomic and behavioural barriers in a sea of connectivity. *Mol. Ecol.* 28, 1394–1411. doi: 10.1111/mec.15010

Bartolino, V., Cardinale, M., Svedäng, H., Linderholm, Hans, W., Casini, M., et al. (2012). Historical spatiotemporal dynamics of eastern North Sea cod. *Can. J. Fish. Aquat. Sci.* 69, 833–841. doi: 10.1139/f2012-028

Berg, P. R., Jentoft, S., Star, B., Ring, K. H., Knutsen, H., Lien, S., et al. (2015). Adaptation to Low Salinity Promotes Genomic Divergence in Atlantic Cod (Gadus morhua L.). *Genome Biol. Evol.* 7, 1644–1663. doi: 10.1093/gbe/evv093

Berg, P. R., Star, B., Pampoulie, C., Bradbury, I. R., Bentzen, P., Hutchings, J. A., et al. (2017). Trans-oceanic genomic divergence of Atlantic cod ecotypes is associated with large inversions. *Heredity* 119, 418–428. doi: 10.1038/hdy.2017.54

Berg, P. R., Star, B., Pampoulie, C., Sodeland, M., Barth, J. M. I., Knutsen, H., et al. (2016). Three chromosomal rearrangements promote genomic divergence between migratory and stationary ecotypes of Atlantic cod. *Sci. Rep.* 6:23246. doi: 10.1038/srep23246

Bigg, G. R., Cunningham, C. W., Ottersen, G., Pogson, G. H., Wadley, M. R., and Williamson, P. (2008). Ice-age survival of Atlantic cod: agreement between palaeoecology models and genetics. *Proc. R. Soc. B Biol. Sci.* 275, 163–173. doi: 10.1098/rspb.2007.1153

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

Brattey, J., Cadigan, N., Dwyer, K. S., Healey, B. P., Ings, D. W., Lee, E. M., et al. (2018). Assessment of the Northern Cod (*Gadus morhua*) stock in NAFO Divisions 2J3KL in 2016. *DFO Can. Sci. Advis. Sec. (CSAS) Res. Doc. 2018/018* 107, 20–24.

Carøe, C., Gopalakrishnan, S., Vinner, L., Mak, S. S., Sinding, M. H. S., Samaniego, J. A., et al. (2018). Single−tube library preparation for degraded DNA. *Methods Ecol. Evol.* 9, 410–419. doi: 10.1111/2041-210x.12871

Carr, S., Snellen, A., Howse, K., and Wroblewski, J. (1995). Mitochondrial DNA sequence variation and genetic stock structure of Atlantic cod (Gadus morhua) from bay and offshore locations on the Newfoundland continental shelf. *Mol. Ecol.* 4, 79–88. doi: 10.1111/j.1365-294x.1995.tb00194.x

Casas-Marce, M., Marmesat, E., Soriano, L., Martínez-Cruz, B., Lucena-Perez, M., Nocete, F., et al. (2017). Spatiotemporal dynamics of genetic variation in the Iberian lynx along its path to extinction reconstructed with ancient DNA. *Mol. Biol. Evol.* 34, 2893–2907. doi: 10.1093/molbev/msx222

Castañeda, R. A., Burliuk, C. M. M., Casselman, J. M., Cooke, S. J., Dunmall, K. M., Forbes, L. S., et al. (2020). A Brief History of Fisheries in Canada. *Fisheries* 45, 303–318. doi: 10.1002/fsh.10449

Charif, D., Lobry, J. R., Necsulea, A., Palmeira, L., Penel, S., Perriere, G., et al. (2020). *Package 'seqinr*. URL: http://seqinr.r-forge.r-project.org/

Ciesielski, S., and Makowiecki, D. (2005). Ancient and modern mitochondrial haplotypes of common bream (Abramis brama L.) in Poland. *Ecol. Freshw. Fish* 14, 278–282. doi: 10.1111/j.1600-0633.2005.00097.x

Consuegra, S., García, de Leániz, C., Serdio, A., González Morales, M., Straus, L., et al. (2002). Mitochondrial DNA variation in Pleistocene and modern Atlantic salmon from the Iberian glacial refugium. *Mol. Ecol.* 11, 2037–2048. doi: 10.1046/j.1365-294x.2002.01592.x

Cooper, A., and Poinar, H. N. (2000). Ancient DNA: do it right or not at all. *Science* 289, 1139–1139.

R Core Team. (2020). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., et al. (2011). The variant call format and VCFtools. *Bioinformatics* 27, 2156–2158. doi: 10.1093/bioinformatics/btr330

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

Drinkwater, K. F. (2006). The regime shift of the 1920s and 1930s in the North Atlantic. *Prog. Oceanogr.* 68, 134–151. doi: 10.1016/j.pocean.2006.02.011

Edvardsson, R., Patterson, W. P., Bárðarson, H., Timsic, S., and Ólafsdóttir, G. Á (2019). Change in Atlantic cod migrations and adaptability of early land-based fishers to severe climate variation in the North Atlantic. *Quat. Res.* 2019, 1–11. doi: 10.1017/qua.2018.147

Enghoff, I. B. (1999). Fishing in the Baltic Region from the 5th century BC to the 16th century AD: evidence from Fish Bones. *Archaeofauna* 8, 41–85.

Exadactylos, A., Rigby, M. J., Geffen, A. J., and Thorpe, J. P. (2007). Conservation aspects of natural populations and captive-bred stocks of turbot (Scophthalmus maximus) and Dover sole (Solea solea) using estimates of genetic diversity. *ICES J. Mar. Sci.* 64, 1173–1181. doi: 10.1093/icesjms/fsm086

Exadactylos, A., Vafidis, D., Tsigenopoulos, C. S., and Gkafas, G. A. (2019). High Connectivity of the White Seabream (Diplodus sargus, L. 1758) in the Aegean Sea, Eastern Mediterranean Basin. *Animals* 9:979. doi: 10.3390/ani9110979

Excoffier, L., and Lischer, H. E. (2010). Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Mol. Ecol. Resour.* 10, 564–567. doi: 10.1111/j.1755-0998.2010.02847.x

Ferrari, G., Cuevas, A., Gondek-Wyrozemska, A. T., Ballantyne, R., Kersten, O., Pálsdóttir, A. H., et al. (2021). The preservation of ancient DNA in archaeological fish bone. *J. Archaeol. Sci.* 126:105317. doi: 10.1016/j.jas.2020.105317

Food and Agriculture Organization [Fao]. (2020-2021a). *International Council for the Exploration of the Sea (ICES). ICES Advice 2019. Cod - Western English Channel, Bristol Channel, Celtic Sea and Southwest of Ireland. FIRMS Reports. In: Fisheries and Resources Monitoring System (FIRMS) [online]*. Rome: Food and Agriculture Organization doi: 10.1016/j.jas.2020.105317

Food and Agriculture Organization [Fao]. (2020-2021b). *International Council for the Exploration of the Sea (ICES). ICES Advice 2020. Cod - Baltic Sea (Eastern part). FIRMS Reports. In: Fisheries and Resources Monitoring System (FIRMS) [online]*. Rome: Food and Agriculture Organization.

Fortes, G. G., and Paijmans, J. L. (2015). "Analysis of whole mitogenomes from ancient samples," in *Whole genome amplification*, (Heidelberg: Springer), 179–195. doi: 10.1007/978-1-4939-2990-0_13

Frankham, R., Ballou, J. D., and Briscoe, D. A. (2002). "Chapter 10: Loss of genetic diversity in small populations," in *Introduction to conservation genetics*, ed. K. H. McInness (Cambridge, UK: Cambridge University Press), 227–253. doi: 10.1017/cbo9780511808999.013

Gibbard, P., and Van Kolfschoten, T. (2005). "The Pleistocene and Holocene Epochs," in *A Geologic Time Scale 2004*, eds A. G. Smith, F. M. Gradstein, and J. G. Ogg (Cambridge: Cambridge University Press), 441–452. doi: 10.1017/cbo9780511536045.023

Gilbert, M. T. P., Bandelt, H.-J., Hofreiter, M., and Barnes, I. (2005). Assessing ancient DNA studies. *Trends Ecol. Evol.* 20, 541–544. doi: 10.1016/j.tree.2005.07.005

Gondek, A. T., Boessenkool, S., and Star, B. (2018). A stainless-steel mortar, pestle and sleeve design for the efficient fragmentation of ancient bone. *BioTechniques* 64, 266–269. doi: 10.2144/btn-2018-0008

Guindon, S., Dufayard, J.-F., Lefort, V., Anisimova, M., Hordijk, W., and Gascuel, O. (2010). New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: assessing the Performance of PhyML 3.0. *Syst. Biol.* 59, 307–321. doi: 10.1093/sysbio/syq010

Guindon, S., and Gascuel, O. (2003). A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst. Biol.* 52, 696–704. doi: 10.1080/10635150390235520

Hardie, D. C., Gillett, R. M., and Hutchings, J. A. (2006). The effects of isolation and colonization history on the genetic structure of marine-relict populations of Atlantic cod (Gadus morhua) in the Canadian Arctic. *Can. J. Fish. Aquat. Sci.* 63, 1830–1839. doi: 10.1139/f06-085

Harland, J., and Barrett, J. H. (2012). "Chapter 7: The Maritime Economy, Fish Bone," in *Being an Islander: Production and identity at Quoygrew, Orkney, AD 900-1600*, ed. J. H. Barrett (Cambridge: McDonald Institute for Archaeological Research), 115–138.

Hauser, L., and Carvalho, G. R. (2008). Paradigm shifts in marine fisheries genetics: ugly hypotheses slain by beautiful facts. *Fish Fish.* 9, 333–362. doi: 10.1111/j.1467-2979.2008.00299.x

Heller, R., Chikhi, L., and Siegismund, H. R. (2013). The Confounding Effect of Population Structure on Bayesian Skyline Plot Inferences of Demographic History. *PLoS One* 8:e62992. doi: 10.1371/journal.pone.0062992

Hofman, C. A., Rick, T. C., Fleischer, R. C., and Maldonado, J. E. (2015). Conservation archaeogenomics: ancient DNA and biodiversity in the Anthropocene. *Trends Ecol. Evol.* 30, 540–549. doi: 10.1016/j.tree.2015.06.008

Hutchinson, W. F., Oosterhout, C. V., Rogers, S. I., and Carvalho, G. R. (2003). Temporal analysis of archived samples indicates marked genetic changes in declining North Sea cod (Gadus morhua). *Proc. R. Soc. Lon. B Biol. Sci.* 270, 2125–2132. doi: 10.1098/rspb.2003.2493

Hylen, A. (2002). Fluctuations in abundance of Northeast Arctic cod during the 20th century. *ICES Mar. Sci. Symp.* 215, 543–550.

Hylen, A., Nakken, O., and Nedreaas, K. (2008). *Northeast Arctic cod: fisheries, life history, stock fluctuations and management. Norwegian spring-spawning herring and Northeast Arctic cod*. Trondheim: Tapir Academic Press, 83–118.

Johannesson, K., and Andre, C. (2006). INVITED REVIEW: life on the margin: genetic isolation and diversity loss in a peripheral marine ecosystem, the Baltic Sea. *Mol. Ecol.* 15, 2013–2029. doi: 10.1111/j.1365-294x.2006.02919.x

Johansen, S., and Bakke, I. (1996). The complete mitochondrial DNA sequence of Atlantic cod (Gadus morhua): relevance to taxonomic studies among codfishes. *Mol. Mar. Biol. Biotechnol.* 5, 203–214.

Johnson, B. M., Kemp, B. M., and Thorgaard, G. H. (2018). Increased mitochondrial DNA diversity in ancient Columbia River basin Chinook salmon Oncorhynchus tshawytscha. *PLoS One* 13:e0190059. doi: 10.1371/journal.pone.0190059

Jónsson, H., Ginolhac, A., Schubert, M., Johnson, P. L., and Orlando, L. (2013). mapDamage2. 0: fast approximate Bayesian estimates of ancient DNA damage parameters. *Bioinformatics* 29, 1682–1684. doi: 10.1093/bioinformatics/btt193

Jonsson, P. R., Corell, H., André, C., Svedäng, H., and Moksnes, P. O. (2016). Recent decline in cod stocks in the North Sea–Skagerrak–Kattegat shifts the sources of larval supply. *Fish. Oceanogr.* 25, 210–228. doi: 10.1111/fog.12146

Jørgensen, T. E., Karlsen, B. O., Emblem, Å, Breines, R., Andreassen, M., Rounge, T. B., et al. (2018). Mitochondrial genome variation of Atlantic cod. *BMC Res. Notes* 11:397. doi: 10.1186/s13104-018-3506-3

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

Keighley, X., Pálsson, S., Einarsson, B. F., Petersen, A., Fernández-Coll, M., Jordan, P., et al. (2019). Disappearance of Icelandic walruses coincided with Norse settlement. *Mol. Biol. Evol.* 36, 2656–2667. doi: 10.1093/molbev/msz196

Kircher, M., Sawyer, S., and Meyer, M. (2012). Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform. *Nucleic Acids Res.* 40:e3. doi: 10.1093/nar/gkr771

Knutsen, H., Olsen, E. M., Jorde, P. E., Espeland, S. H., André, C., and Stenseth, N. C. (2011). Are low but statistically significant levels of genetic differentiation in marine fishes 'biologically meaningful'? A case study of coastal Atlantic cod. *Mol Ecol.* 20, 768–783. doi: 10.1111/j.1365-294X.2010.04979.x

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

Lait, L. A. (2016). *A mitogenomic study of four at-risk maring fish species: Atlantic wolffish, spotted wolffish, northern wolffish, and Atlantic cod, with special emphasis on the waters off Newfoundland and Labrador*. Ph.D. thesis, Canada: Memorial University of Newfoundland.

Lait, L. A., Marshall, H. D., and Carr, S. M. (2018). Phylogeographic mitogenomics of Atlantic cod Gadus morhua: variation in and among trans−Atlantic, trans−Laurentian, Northern cod, and landlocked fjord populations. *Ecol. Evol.* 8, 6420–6437. doi: 10.1002/ece3.3873

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

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics* 25, 2078–2079. doi: 10.1093/bioinformatics/btp352

Limburg, K. E., Walther, Y., Hong, B., Olson, C., and Storå, J. (2008). Prehistoric versus modern Baltic Sea cod fisheries: selectivity across the millennia. *Proc. R. Soc. B Biol. Sci.* 275, 2659–2665. doi: 10.1098/rspb.2008.0711

Lindgreen, S. (2012). AdapterRemoval: easy cleaning of next-generation sequencing reads. *BMC Res. Notes* 5:337. doi: 10.1186/1756-0500-5-337

Lombal, A. J., O'dwyer, J. E., Friesen, V., Woehler, E. J., and Burridge, C. P. (2020). Identifying mechanisms of genetic differentiation among populations in vagile species: historical factors dominate genetic differentiation in seabirds. *Biol. Rev.* 95, 625–651. doi: 10.1111/brv.12580

Louis, M., Skovrind, M., Samaniego Castruita, J. A., Garilao, C., Kaschner, K., Gopalakrishnan, S., et al. (2020). Influence of past climate change on phylogeography and demographic history of narwhals. Monodon monoceros. *Proc. R. Soc. B* 287:20192964. doi: 10.1098/rspb.2019.2964

MacKenzie, B. R., Alheit, J., Conley, D. J., Holm, P., and Kinze, C. C. (2002). Ecological hypotheses for a historical reconstruction of upper trophic level biomass in the Baltic Sea and Skagerrak. *Can. J. Fish. Aqua. Sci.* 59, 173–190. doi: 10.1139/f01-201

Mak, S. S. T., Gopalakrishnan, S., Carøe, C., Geng, C., Liu, S., Sinding, M. S., et al. (2017). Comparative performance of the BGISEQ-500 vs Illumina HiSeq2500 sequencing platforms for palaeogenomic sequencing. *Gigascience* 6, 1–13. doi: 10.1093/gigascience/gix049

Malmstrøm, M., Matschiner, M., Tørresen, O. K., Star, B., Snipen, L. G., Hansen, T. F., et al. (2016). Evolution of the immune system influences speciation rates in teleost fishes. *Nat. Genet.* 48, 1204–1210. doi: 10.1038/ng.3645

Matschiner, M. (2016). Fitchi: haplotype genealogy graphs based on the Fitch algorithm. *Bioinformatics* 32, 1250–1252. doi: 10.1093/bioinformatics/btv717

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., et al. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* 20, 1297–1303. doi: 10.1101/gr.107524.110

Meyer, M., and Kircher, M. (2010). Illumina sequencing library preparation for highly multiplexed target capture and sequencing. *Cold Spring Harb. Protoc.* 2010:db.rot5448.

Mieszkowska, N., Genner, M. J., Hawkins, S. J., and Sims, D. W. (2009). "Chapter 3 Effects of Climate Change and Commercial Fishing on Atlantic Cod Gadus morhua," in *Advances in Marine Biology*, ed. D. W. Sims (Cambridge: Academic Press), 213–273. doi: 10.1016/s0065-2881(09)56003-8

Mitchell, K. J., and Rawlence, N. J. (2021). Examining Natural History through the Lens of Palaeogenomics. *Trends Ecol. Evol.* 36, 258–267. doi: 10.1016/j.tree.2020.10.005

Moss, M. L., Rodrigues, A. T., Speller, C. F., and Yang, D. Y. (2016). The historical ecology of Pacific herring: tracing Alaska Native use of a forage fish. *J. Archaeol. Sci. Rep.* 8, 504–512. doi: 10.1016/j.jasrep.2015.10.005

Murray, G. G., Soares, A. E., Novak, B. J., Schaefer, N. K., Cahill, J. A., Baker, A. J., et al. (2017). Natural selection shaped the rise and fall of passenger pigeon genomic diversity. *Science* 358, 951–954. doi: 10.1126/science.aao0960

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

Nicholls, J., Allaire, B., and Holm, P. (2021). The Capacity Trend Method: a new approach for enumerating the Newfoundland cod fisheries (1675–1790). *Histori. Methods J. Quan. Interdiscip. History* 2021 1–14. doi: 10.1080/01615440.2020.1853643

Nikulina, E. A., and Schmölcke, U. (2016). Reconstruction of the historical distribution of sturgeons (Acipenseridae) in the eastern North Atlantic based on ancient DNA and bone morphology of archaeological remains: implications for conservation and restoration programmes. *Divers. Distrib.* 22, 1036–1044. doi: 10.1111/ddi.12461

Nyström, V., Angerbjörn, A., and Dalén, L. (2006). Genetic consequences of a demographic bottleneck in the Scandinavian arctic fox. *Oikos* 114, 84–94. doi: 10.1111/j.2006.0030-1299.14701.x

Ojaveer, H., Jaanus, A., MacKenzie, B. R., Martin, G., Olenin, S., Radziejewska, T., et al. (2010). Status of Biodiversity in the Baltic Sea. *PLoS One* 5:e12467. doi: 10.1371/journal.pone.0012467

Olafsdottir, G. A., Westfall, K. M., Edvardsson, R., and Palsson, S. (2014). Historical DNA reveals the demographic history of Atlantic cod (Gadus morhua) in medieval and early modern Iceland. *Proc. Biol. Sci.* 281:20132976. doi: 10.1098/rspb.2013.2976

Olson, C., and Walther, Y. (2007). Neolithic cod (*Gadus morhua*) and herring (*Clupea harengus*) fisheries in the Baltic Sea, in the light of fine-mesh sieving: a comparative study of subfossil fishbone from the late Stone Age sites at Ajvide, Gotland, Sweden and Jettböle, Åland, Finland. *Environ. Archaeol.* 12, 175–185. doi: 10.1179/174963107x226435

Oosting, T., Star, B., Barrett, J. H., Wellenreuther, M., Ritchie, P. A., and Rawlence, N. J. (2019). Unlocking the potential of ancient fish DNA in the genomic era. *Evol. Appl.* 12, 1513–1522. doi: 10.1111/eva.12811

Orton, D. C., Makowiecki, D., de Roo, T., Johnstone, C., Harland, J., Jonsson, L., et al. (2011). Stable Isotope Evidence for Late Medieval (14th–15th C) Origins of the Eastern Baltic Cod (Gadus morhua) Fishery. *PLoS One* 6:e27568. doi: 10.1371/journal.pone.0027568

Orton, D. C., Morris, J., Locker, A., and Barrett, J. H. (2014). Fish for the city: meta-analysis of archaeological cod remains and the growth of London's northern trade. *Antiquity* 88, 516–530. doi: 10.1017/S0003598X00101152

Page, A. J., Taylor, B., Delaney, A. J., Soares, J., Seemann, T., Keane, J. A., et al. (2016). SNP-sites: rapid efficient extraction of SNPs from multi-FASTA alignments. *Microb. Genom.* 2:e000056. doi: 10.1099/mgen.0.000056

Paijmans, J. L., Gilbert, M. T. P., and Hofreiter, M. (2013). Mitogenomic analyses from ancient DNA. *Mol. Phylogenet. Evol.* 69, 404–416. doi: 10.1016/j.ympev.2012.06.002

Paradis, E. (2010). pegas: an R package for population genetics with an integrated–modular approach. *Bioinformatics* 26, 419–420. doi: 10.1093/bioinformatics/btp696

Paradis, E., and Schliep, K. (2019). ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics* 35, 526–528. doi: 10.1093/bioinformatics/bty633

Pfeifer, B., Wittelsbuerger, U., Li, H., Handsaker, B., and Pfeifer, M. B. (2020). *Package 'PopGenome'*.URL: https://CRAN.R-project.org/package=PopGenome

Pinnegar, J. K., and Engelhard, G. H. (2008). The 'shifting baseline'phenomenon: a global perspective. *Rev. Fish Biol. Fish.* 18, 1–16. doi: 10.1007/s11160-007-9058-6

Pinsky, M. L., Eikeset, A. M., Helmerson, C., Bradbury, I. R., Bentzen, P., Morris, C., et al. (2021). Genomic stability through time despite decades of exploitation in cod on both sides of the Atlantic. *Proc. Natl. Acad. Sci. U. S. A.* 118:e2025453118. doi: 10.1073/pnas.2025453118

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

Rodrigues, A. S., Monsarrat, S., Charpentier, A., Brooks, T. M., Hoffmann, M., Reeves, R., et al. (2019). Unshifting the baseline: a framework for documenting historical population changes and assessing long-term anthropogenic impacts. *Philos. Trans. R. Soc. B* 374:20190220. doi: 10.1098/rstb.2019.0220

Rozas, J., Ferrer-Mata, A., Sánchez-DelBarrio, J. C., Guirao-Rico, S., Librado, P., Ramos-Onsins, S. E., et al. (2017). DnaSP 6: DNA sequence polymorphism analysis of large data sets. *Mol. Biol. Evol.* 34, 3299–3302. doi: 10.1093/molbev/msx248

Salazar, G. (2020). *FastaUtils: Utilities for DNA/RNA sequence processing*. URL: https://github.com/GuillemSalazar/FastaUtils.git

Schroeder, H., Ávila-Arcos, M. C., Malaspinas, A.-S., Poznik, G. D., Sandoval-Velasco, M., Carpenter, M. L., et al. (2015). Genome-wide ancestry of 17th-century enslaved Africans from the Caribbean. *Proc. Natl. Acad. Sci. U. S. A.* 112, 3669–3673. doi: 10.1073/pnas.1421784112

Schubert, M., Ermini, L., Der Sarkissian, C., Jónsson, H., Ginolhac, A., Schaefer, R., et al. (2014). Characterization of ancient and modern genomes by SNP detection and phylogenomic and metagenomic analysis using PALEOMIX. *Nat. Protoc.* 9:1056. doi: 10.1038/nprot.2014.063

Seersholm, F. V., Cole, T. L., Grealy, A., Rawlence, N. J., Greig, K., Knapp, M., et al. (2018). Subsistence practices, past biodiversity, and anthropogenic impacts revealed by New Zealand-wide ancient DNA survey. *Proc. Natl. Acad. Sci. U. S. A.* 115:7771. doi: 10.1073/pnas.1803573115

Selim, S. A., Blanchard, J. L., Bedford, J., and Webb, T. J. (2016). Direct and indirect effects of climate and fishing on changes in coastal ecosystem services: a historical perspective from the North Sea. *Reg. Environ. Change* 16, 341–351. doi: 10.1007/s10113-014-0635-7

Shapiro, B., Drummond, A. J., Rambaut, A., Wilson, M. C., Matheus, P. E., Sher, A. V., et al. (2004). Rise and fall of the Beringian steppe bison. *Science* 306, 1561–1565. doi: 10.1126/science.1101074

Shelton, P. A., and Morgan, M. J. (2014). Impact of maximum sustainable yield-based fisheries management frameworks on rebuilding North Atlantic cod stocks. *J. Northwest Atlant. Fish. Sci.* 46, 15–25. doi: 10.2960/j.v46.m697

Sigurgíslason, H., and Árnason, E. (2003). Extent of mitochondrial DNA sequence variation in Atlantic cod from the Faroe Islands: a resolution of gene genealogy. *Heredity* 91, 557–564. doi: 10.1038/sj.hdy.6800361

Sodeland, M., Jorde, P. E., Lien, S., Jentoft, S., Berg, P. R., Grove, H., et al. (2016). "Islands of Divergence" in the Atlantic Cod Genome Represent Polymorphic Chromosomal Rearrangements. *Genome Biol. Evol.* 8, 1012–1022. doi: 10.1093/gbe/evw057

Speller, C. F., Hauser, L., Lepofsky, D., Moore, J., Rodrigues, A. T., Moss, M. L., et al. (2012). High Potential for Using DNA from Ancient Herring Bones to Inform Modern Fisheries Management and Conservation. *PLoS One* 7:e51122. doi: 10.1371/journal.pone.0051122

Spencer, H. G. (2020). Beyond Equilibria: the Neglected Role of History in Ecology and Evolution. *Quart. Rev. Biol.* 95, 311–321. doi: 10.1086/711803

Star, B., Barrett, J. H., Gondek, A. T., and Boessenkool, S. (2018). Ancient DNA reveals the chronology of walrus ivory trade from Norse Greenland. *Proc. R. Soc. B Biol. Sci.* 285:20180978. doi: 10.1098/rspb.2018.0978

Star, B., Boessenkool, S., Gondek, A. T., Nikulina, E. A., Hufthammer, A. K., Pampoulie, C., et al. (2017). Ancient DNA reveals the Arctic origin of Viking Age cod from Haithabu. Germany. *Proc. Natl. Acad. Sci. U. S. A.* 114, 9152–9157. doi: 10.1073/pnas.1710186114

Star, B., Nederbragt, A. J., Jentoft, S., Grimholt, U., Malmstrøm, M., Gregers, T. F., et al. (2011). The genome sequence of Atlantic cod reveals a unique immune system. *Nature* 477, 207–210.

Stiller, M., Baryshnikov, G., Bocherens, H., Grandal, d'Anglade, A., Hilpert, B., et al. (2010). Withering Away—25,000 Years of Genetic Decline Preceded Cave Bear Extinction. *Mol. Biol. Evol.* 27, 975–978. doi: 10.1093/molbev/msq083

Sætersdal, G., and Hylen, A. (1964). The decline of the skrei fisheries: a review of the landing statistics 1866-1957 and an evaluation of the effects of the postwarincrease in the total exploitation of the arctic cod. *Fiskeridirektoratets havforskningsinstitutt* 13, 56–69.

Therkildsen, N. O., Nielsen, E. E., Swain, D. P., and Pedersen, J. S. (2010). Large effective population size and temporal genetic stability in Atlantic cod (Gadus morhua) in the southern Gulf of St. Lawrence. *Can. J. Fish. Aquat. Sci.* 67, 1585–1595. doi: 10.1139/f10-084

Thomas, J. E., Carvalho, G. R., Haile, J., Rawlence, N. J., Martin, M. D., Ho, S. Y. W., et al. (2019). Demographic reconstruction from ancient DNA supports rapid extinction of the great auk. *Elife* 8:e47509. doi: 10.7554/eLife.47509

Thurstan, R. H., Brockington, S., and Roberts, C. M. (2010). The effects of 118 years of industrial fishing on UK bottom trawl fisheries. *Nat. Commun.* 1:15. doi: 10.1038/ncomms1013

Tørresen, O. K., Star, B., Jentoft, S., Reinar, W. B., Grove, H., Miller, J. R., et al. (2017). An improved genome assembly uncovers prolific tandem repeats in Atlantic cod. *BMC Genomics* 18:95. doi: 10.1186/s12864-016-3448-x

Venter, O., Sanderson, E. W., Magrach, A., Allan, J. R., Beher, J., Jones, K. R., et al. (2016). Sixteen years of change in the global terrestrial human footprint and implications for biodiversity conservation. *Nat. Commun.* 7:12558. doi: 10.1038/ncomms12558

Welch, A. J., Wiley, A. E., James, H. F., Ostrom, P. H., Stafford, Jr., and Fleischer, R. C. (2012). Ancient DNA reveals genetic stability despite demographic decline: 3,000 years of population history in the endemic Hawaiian petrel. *Mol. Biol. Evol.* 29, 3729–3740. doi: 10.1093/molbev/mss185

Wenne, R., Bernaś, R., Kijewska, A., Poćwierz-Kotus, A., Strand, J., Petereit, C., et al. (2020). SNP genotyping reveals substructuring in weakly differentiated populations of Atlantic cod (Gadus morhua) from diverse environments in the Baltic Sea. *Sci. Rep.* 10, 1–15.

Zhang, J., Pei, N., Mi, X., and Zhang, M. J. (2017). *Package 'phylotools'. dimension* 12. URL: https://github.com/helixcn/phylotools

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.

Citation: Martínez-García L, Ferrari G, Oosting T, Ballantyne R, van der Jagt I, Ystgaard I, Harland J, Nicholson R, Hamilton-Dyer S, Baalsrud HT, Brieuc MSO, Atmore LM, Burns F, Schmölcke U, Jakobsen KS, Jentoft S, Orton D, Hufthammer AK, Barrett JH and Star B (2021) Historical Demographic Processes Dominate Genetic Variation in Ancient Atlantic Cod Mitogenomes. Front. Ecol. Evol. 9:671281. doi: 10.3389/fevo.2021.671281

frontiers in Ecology and Evolution

ORIGINAL RESEARCH
published: 12 July 2021
doi: 10.3389/fevo.2021.659134

Population Genomics of the Commercially Important Gulf of Mexico Pink Shrimp *Farfantepenaeus duorarum* (Burkenroad, 1939) Support Models of Juvenile Transport Around the Florida Peninsula

Laura E. Timm[1,2], Thomas L. Jackson[3], Joan A. Browder[3] and Heather D. Bracken-Grissom[1]*

[1] CRUSTOMICS Laboratory, Department of Biological Sciences, Institute of Environment, Florida International University, Miami, FL, United States, [2] Computational Biosciences Program, University of Colorado, Aurora, Aurora, CO, United States, [3] Southeast Fisheries Science Center, NOAA National Marine Fisheries Service, Miami, FL, United States

OPEN ACCESS

Edited by:
Melissa T. R. Hawkins, Smithsonian Institution, United States

Reviewed by:
Nick Wade, Commonwealth Scientific and Industrial Research Organisation (CSIRO), Australia
Gonzalo Gajardo, University of Los Lagos, Chile

***Correspondence:**
Laura E. Timm
ltimm004@fiu.edu

Specialty section:
This article was submitted to Evolutionary and Population Genetics, a section of the journal Frontiers in Ecology and Evolution

Received: *27 January 2021*
Accepted: *17 June 2021*
Published: *12 July 2021*

Citation:
Timm LE, Jackson TL, Browder JA and Bracken-Grissom HD (2021) Population Genomics of the Commercially Important Gulf of Mexico Pink Shrimp Farfantepenaeus duorarum (Burkenroad, 1939) Support Models of Juvenile Transport Around the Florida Peninsula. Front. Ecol. Evol. 9:659134. doi: 10.3389/fevo.2021.659134

The Gulf of Mexico pink shrimp, *Farfantepenaeus duorarum*, supports large fisheries in the United States and Mexico, with nearly 7,000 tons harvested from the region in 2016. Given the commercial importance of this species, management is critical: in 1997, the southern Gulf of Mexico pink shrimp fishery was declared collapsed and mitigation strategies went into effect, with recovery efforts lasting over a decade. Fisheries management can be informed and improved through a better understanding of how factors associated with early life history impact genetic diversity and population structure in the recruited population. *Farfantepenaeus duorarum* are short-lived, but highly fecund, and display high variability in recruitment patterns. To date, modeling the impacts of ecological, physical, and behavioral factors on juvenile settlement has focused on recruitment of larval individuals of *F. duorarum* to nursery grounds in Florida Bay. Here, we articulate testable hypotheses stemming from a recent model of larval transport and evaluate support for each with a population genomics approach, generating reduced representation library sequencing data for *F. duorarum* collected from seven regions around the Florida Peninsula. Our research represents the first and most molecular data-rich study of population structure in *F. duorarum* in the Gulf and reveals evidence of a differentiated population in the Dry Tortugas. Our approach largely validates a model of larval transport, allowing us to make management-informative inferences about the impacts of spawning location and recruitment patterns on intraspecific genetic diversity. Such inferences improve our understanding of the roles of non-genetic factors in generating and maintaining genetic diversity in a commercially important penaeid shrimp species.

Keywords: pink shrimp, *Penaeus duorarum*, *Farfantepenaeus duorarum*, Gulf of Mexico, ddRADSeq, population genomics, fisheries management

INTRODUCTION

The Gulf pink shrimp, *Farfantepenaeus duorarum* (Burkenroad, 1939) supports multiple, international fisheries along its described geographic range, representing millions of dollars of economic activity (Sheridan, 1996; Ramírez-Rodríguez et al., 2003; Hart et al., 2012). Over 7,000 tons of pink shrimp were harvested across fisheries in the Gulf of Mexico in 2016, the last year for which such data are available (Hart, 2017). Given the economic and social influence of the large-scale fishing effort directed at *Farfantepenaeid* species in the Gulf, proper management is critical to the sustained stability of the species and protection of economic interests in the region: all of the species within the *Farfantepenaeus* group are targeted by fisheries to some extent (see Timm et al., 2019 for more information).

Management of fished species requires understanding the biology and ecology of the organism, including assessments of intraspecific biodiversity and the evolutionary processes that drive it (Bernatchez, 1995). Management of *F. duorarum* by Mexico and the United States of America makes such insight particularly crucial: shrimp fisheries have supported regional Mexican economies for decades, and pink shrimp have contributed substantially to these fisheries, with 90% of fished shrimp in the 1990s being *F. duorarum* (Arreguín-Sánchez et al., 2008). In the late 1990s, however, the *F. duorarum* fishery in the southern Gulf of Mexico was declared collapsed (Arreguín-Sánchez et al., 1997). Investigation of possible underlying causes of the collapse found evidence for decreased stock-recruitment (Arreguín-Sánchez et al., 1997, 1999), and efforts were undertaken to promote recovery (Arreguín-Sánchez et al., 2008). Such events have occurred in United States fisheries as well, resulting in the closure of the northern brown shrimp (*F. aztecus*) fishery along the Texas coast in the 1980s (Klima et al., 1987). The co-occurrence of several, economically important species of *Farfantepenaeus* along the coasts of the Gulf of Mexico further complicate management. Specifically, juvenile individuals of *F. brasiliensis* and *F. duorarum* look very similar, and the ability to confidently identify juvenile individuals taxonomically by reproductive structure morphology (Pérez-Farfante, 1988) is nearly impossible (Ditty and Alvarado Bremer, 2011; Teodoro et al., 2016). A recent study found cryptic diversity within *F. brasiliensis,* identifying two distinct populations (one occupying United States coasts and the other present along the coasts of South America). The study called for additional efforts to better understand population structure and evolutionary history within managed species (Timm et al., 2019). A break in species composition exists between the Gulf of Mexico and the greater Atlantic; divided by prevailing environmental features (Avise, 1992; Young et al., 2002). Studies focused on genetic diversity and population connectivity in species that span this break (such as *F. duorarum*) might prove particularly informative.

Life history can be significant in determining the composition and structure of adult assemblages, especially in species with complicated development cycles. Adults of *F. duorarum* spawn year-round in aggregations offshore of the Dry Tortugas and the Marquesas on the southwest Florida shelf (Cummings, 1961; Roberts, 1986). There is a distinct spawning aggregation on the Sanibel grounds as well, and, despite geographic overlap between Sanibel and Dry Tortugas nursery grounds, a division between shrimp originating from these two spawning grounds has been noted near Indian Key (i.e., between Sanibel and Dry Tortugas; Costello and Allen, 1966; Robblee et al., 1999): shrimp emanating from Sanibel nursery grounds only rarely migrate into the Dry Tortugas trawling grounds south of Indian Key and vice versa. After hatching, larvae rapidly progress through 11 developmental stages [nauplii (5), protozoea (3), and mysis (3)] in approximately 15 days (Dobkin, 1961). During this time, larval individuals exhibit a vertical migration pattern, alternating between deeper waters and surface waters (Rothlisberg, 1982; Rothlisberg et al., 1995, 1996; Condie et al., 1999). For the first 15 days of development, vertical migration is triggered by light [diel vertical migration (DVM)]. During the subsequent 15 days, as individuals pass through several postlarval stages (3–6 stages; Ewald, 1965), vertical migration is timed to tidal movement [selective tidal-stream transport (STST)], allowing postlarvae to take advantage of tidal movement toward nursery grounds and avoid tidal movement in the opposite direction (Forward and Tankersley, 2001; Queiroga and Blanton, 2005). A modeling study by Criales et al. (2015) suggests that these two behaviors, DVM and STST, facilitate movement from spawning grounds toward primary nursery grounds in Florida Bay and mangrove estuaries along the southwest coast (Tabb et al., 1962; Costello and Allen, 1966; Browder and Robblee, 2009), where they grow through the juvenile stage, returning to spawning grounds as young adults. Environmental factors such as salinity and temperature on their nursery grounds affect their rate of growth and mortality (Browder et al., 1999, 2002; Ehrhardt and Legault, 1999), potentially influencing recruitment to the offshore fishery.

Two routes have been proposed for larval/postlarval migration (**Figure 1**): larvae may drift east and northeast along the Florida Current, to enter Florida Bay through the Florida Keys (Munro et al., 1968; Criales et al., 2003). The other route posits that larvae move more directly across the southwest Florida shelf, entering Florida Bay at its northwest side (Jones et al., 1970; Criales et al., 2006). Recently, Criales et al. (2015) found support for both suggested migration routes with a biophysical model utilizing Lagrangian modeling to display larval-to-postlarval behaviors, receiving output from a physical oceanographic model providing the drivers. The modeling system supported an investigation into the influence of spawning location, larval traits, and oceanographic features (tides, winds, and currents) on larval transport. Virtual larvae were released near the water column's surface from the Dry Tortugas and the Marquesas areas, mimicking a combination of DVM and STST behavior, and allowed to be transported for 28–30 days according to current speeds and directions and larval position in the water column (i.e., bottom vs. middle to surface). Finally, a benthic habitat module reflected larval aggregations on the offshore spawning grounds and suitable settlement habitats near the coast. The biophysical and physical oceanographic model developed by Criales et al. (2015) indicated that recruitment success was largely determined by season and spawning ground: generally, larvae

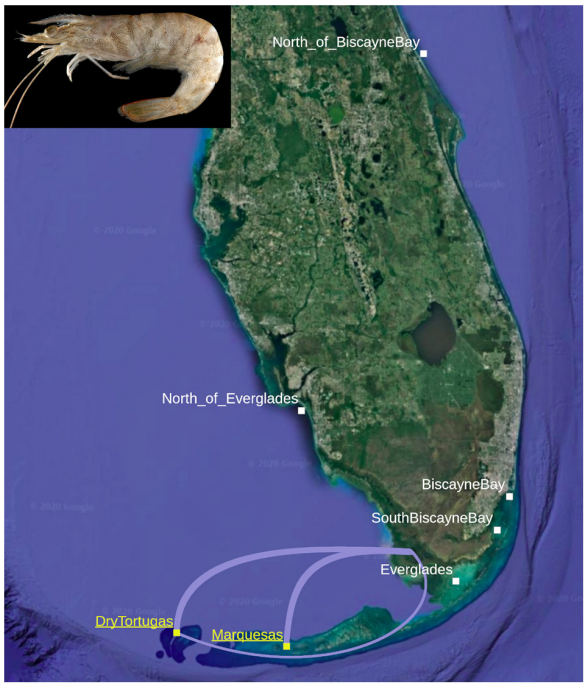

FIGURE 1 | Individuals of *Farfantepenaeus duorarum* were collected from seven regions around the Florida Peninsula and Florida Keys, including two spawning aggregations, DryTortugas and Marquesas (yellow and underlined). The major (thick lines) and minor (thin line) migratory routes described by Criales et al. (2015) are depicted between the spawning aggregations and the major nursery grounds in Florida Bay (Photograph of *F. duorarum* from wikimedia commons. Base map from Google Earth).

simulated from the Marquesas were several times more likely to reach nursery habitat than those simulated from the Dry Tortugas, and summer simulations consistently resulted in higher larval settlement compared to winter simulations. Simulated larvae were most likely to settle in nursery habitat when they were released from the Marquesas in the summer, migrating east-northeast across the southwest Florida shelf. When simulated larvae originated from the Dry Tortugas, they were likely to become entrained in the Florida Current, exiting the Gulf of Mexico entirely and entering the greater Atlantic. The few simulated larvae released from the Dry Tortugas that successfully reached Florida Bay did so through both hypothesized routes, while those simulated larvae successfully recruited to the Florida Bay recruitment area from the Marquesas never migrated through the Florida Keys. These results provide expectations of population dynamics that can be tested with molecular methods.

The model of larval transport and migration developed by Criales et al. (2015) leads to testable, if relatively qualitative, hypotheses. Under the null hypothesis, all pink shrimp around the Florida Peninsula represent a single, genetically homogeneous population, originating from spawning aggregations offshore of the Dry Tortugas and the Marquesas, traveling either migratory route (**Figure 1**), and reaching adulthood on nursery grounds around the Florida Peninsula. From a fishery management perspective, this would be the simplest conclusion: lacking differentiated intraspecific diversity, all fisheries targeting the species can be managed as one. The alternative hypothesis, however, posits that the two spawning aggregations and different migratory routes to the nursery grounds support at least two genetically differentiated populations. If the alternative hypothesis holds, the Dry Tortugas and the Marquesas represent separate spawning aggregations to some extent, maintaining at least two distinct populations (these may be characterized by spawning aggregation, i.e., Dry Tortugas vs. Marquesas, or migratory route, i.e., the more-traveled east-northeast "major" route across the southwest Florida shelf vs. the less-traveled south-southeast "minor" route through the Florida Keys), and more complex management strategies would be needed to protect both populations during the stock-recruitment phase.

A better understanding of these two routes, major and minor, is of primary concern to researchers focused on sustainable fishing of pink shrimp (Browder et al., 1999, 2002; Ehrhardt and Legault, 1999; Criales et al., 2000, 2003, 2006, 2007, 2010, 2015; Ehrhardt et al., 2001; Ogburn et al., 2013). The major route, which traverses the southwest Florida shelf, crosses through a regional fishery operating year-round near the Dry Tortugas and Key West (Klima et al., 1987; Upton et al., 1992; Hart et al., 2012), catching both fully mature and young adult shrimp (Ehrhardt and Legault, 1999; Browder et al., 2002). The co-localization of these large, highly productive pink shrimp fisheries with spawning grounds and out-migrating larvae makes an understanding of population dynamics in the region especially important to long-term species sustainability. Here, we utilize a next-generation sequencing method, double digest Restriction-site Associated sequencing (ddRADseq) to investigate the fine-scale population structure of *F. duorarum* in the eastern Gulf of Mexico. Our overall objective is to characterize diversity and connectivity in terms of the larval migration and transport within the area for the purpose of informing and improving fishery management. To accomplish this, we: (1) validate the biophysical oceanographic modeling results of Criales et al. (2015) with an independent data type (ddRADseq data); (2) investigate any evidence of population differentiation within *F. duorarum* in the region, including whether postlarvae recruited to Biscayne Bay originate from the Dry Tortugas; and (3) contextualize the population genomics results in terms of fisheries management.

MATERIALS AND METHODS

Because the migratory routes between pink shrimp spawning aggregations and nursery habitat span a relatively small geographic range, our sampling effort targeted proximal locations around the Florida Peninsula. Over 100 postlarval, juvenile, and adult specimens of *Farfantepenaeus* were collected from several sites representing seven regions around the Florida Peninsula between 2011 and 2015 (**Figure 1**): New Smyrna ("North_of_BiscayneBay" or "NBB"), Hobie Beach, Bear Cut, and South Virginia Key ("BiscayneBay" or "BB"), NOAA sampling stations 2.1–2.3 and 7.1–7.3 ("SouthBiscayneBay" or "SBB"), Bradley Key ("Everglades" or "EVG"), Pumpkin Bay, Estero Bay,

Fakahatchee Bay, and Pine Island Sound ("North_of_Everglades" or "NEVG"), Fort Jefferson to Key West, which sampled across the Marquesas spawning ground ("Marquesas" or "MQ"), and the Dry Tortugas ("DryTortugas" or "DT"). The majority of samples collected from nursery habitats around the Florida Peninsula were acquired by Jackson as part of a collaboration between the Ecosystems Investigations Unit of the Southeast Fisheries Science Center (SEFSC) in Miami. South Biscayne Bay samples were collected as part of a nearshore southwestern Biscayne Bay monitoring project. Because some of the samples, primarily those representing spawning aggregations, were obtained from shrimping vessels, exact collection coordinates were not obtained. Sampled specimens were frozen after collection and shipped to the Ecosystems Investigations Lab at SEFSC for taxonomic identification, specifically focused on the diagnostic characters associated with reproductive morphology (gonopore, thelyca, and petasmata; see Pérez-Farfante, 1969, 1970, 1988; Pérez-Farfante and Kensley, 1997). After identification to species, 105 frozen individuals identified as *F. duorarum* or likely to be *F. duorarum* (labeled *F.* sp.) were transferred to the CRUSTOMICS Lab in North Miami, Florida, where each was given a unique voucher ID in the Florida International University Crustacean Collection (FICC). The ID and all metadata associated with collection were entered into the FICC database. Samples were thawed and muscle tissue was plucked from each specimen by lifting the integument of the second abdominal segment and removing a few milligrams of tissue, using care to avoid puncturing the digestive tract. Tissue was stored at –20°C in 70% EtOH. The intact whole-specimens were preserved in 70% EtOH and stored in the FICC. All specimens included in the study presented here, including all relevant metadata, are presented in **Supplementary Table 1**.

DNA Extraction and Next-Generation Sequencing Library Preparation

Juveniles and adults were targeted for DNA extraction; postlarvae were excluded to ensure individuals collected had survived their initial migratory journey. Juveniles were expected in nursery areas and adults on spawning grounds. Only adults would be present on the spawning grounds as they return to spawn. DNA was extracted from the plucked abdominal muscle tissue with the DNeasy Blood and Tissue kit (Qiagen), following the manufacturer's instructions. To ensure a sufficient amount of DNA had been obtained from an extraction for downstream ddRADseq library prep, DNA was quantified with the Qubit dsDNA High Sensitivity Analysis kit (ThermoFisher). Gel electrophoresis was used to confirm the presence of intact, high molecular weight DNA: DNA extractions were run through a 2% agarose gel for 90 min at 100 V, visualized with GelRed (Biotium). Only samples with more than 500 ng of unfragmented DNA were considered for ddRADseq library prep.

Of the 105 *F. duorarum* specimens that underwent DNA extraction, a subset were found to meet the criteria described above. Of these, 68 were chosen for next-generation sequencing library prep (∼10 samples per sampled region). Reduced representation libraries were prepared following the double digest Restriction-site Associated DNA sequencing (ddRADseq) method published by Peterson et al. (2012). Briefly, we began with a series of enzyme trials to determine the optimal enzyme combination and size selection range to provide adequate genomic coverage at adequate sequencing depth. At least 500 ng of extracted DNA was digested with *Sph*I-HF and *Eco*RI-HF (New England Biolabs) for 3 h at 37°C. Enzymatic activity was stopped with a 30 min hold at 65°C. Custom barcode adapters (synthesized at Integrated DNA Technologies) were ligated to the double-digested fragments using T4 ligase (New England Biolabs). Following barcode adapter ligation, samples were pooled into nine samples of eight, uniquely barcoded libraries. Fragments between 270 and 330 bp, including adapter length, were size selected on a PippinPrep with a 1.5% Agarose Gel Cassette (Sage Science). To reduce the impact of PCR bias, each size-selected sample was subdivided into five parallel PCR amplification reactions and a negative control was used to ensure reagents were not contaminated. Using the Phusion Hi-Fidelity Polymerase (Thermo Scientific), the PCR reactions went for 10 cycles and incorporated i7 indices and Illumina adapters into every amplified fragment, allowing for pooling of all libraries into a single sample. This final sample was quality-checked on an Agilent BioAnalyzer 2100 (Agilent Technologies) immediately prior to sending it for sequencing with the Illumina HiSeq4000, SE150, at the University of Texas' Genomic Sequencing and Analysis Facility.

Data Assembly and Quality Filtering

Initial quality checks of the raw data were conducted with fastQC (Andrews, 2010) before data assembly began in STACKS v1.45 (Catchen et al., 2013) on Florida International University's High Performance Computing Cluster (FIU HPCC). Given the risk of data assembly decisions resulting in a biased data set, recent literature was consulted before beginning the complex task of generating datasets from ddRADseq data (Mastretta-Yanes et al., 2015; Paris et al., 2017; Rochette and Catchen, 2017; O'Leary et al., 2018). Data assembly followed the recommended core pipeline for *de novo* data: process_radtags to demultiplex the reads, ustacks to align reads within each individual, cstacks to catalog these reads, sstacks to query putative loci against this catalog, and rxstacks to utilize population data to correct individual genotype calls. As any individual dataset, assembled according to the authors' best judgment, can reflect biases stemming from assembly decisions, nine datasets were generated, differing in the maximum Hamming distance allowed between stacks (ustacks' –M), the minimum depth required to designate a stack (ustacks' –m), and the maximum Hamming distance allowed between sample loci (cstacks' –n). The data assembly parameters for each dataset are presented in **Table 1**. These datasets are referred to as "batches" and reflect the parameter settings that generated them: "batch161" is the dataset assembled with a maximum Hamming distance of 1 allowed between stacks, a minimum stack depth of 6, and a maximum Hamming distance of 1 allowed between sample loci (–M 1 –m 6 –n 1).

Quality filtering of the VCFs output from STACKS was accomplished with VCFtools (Danecek et al., 2011) on the FIU HPCC. First, the minimum read depth was set to 10×. Next, sites with ≥50% missing data were removed, followed by individuals with ≥90% missing data. The resulting VCF files

TABLE 1 | Details of data assembly in STACKS v1.45 are provided below, including flags and settings used at every step of the pipeline.

process_radtags	–renz_1 sphI								
	–renz_2 ecoRI								
	–q								
	–r								
ustacks	–M 1			–M 3			–M 5		
	–m 6			–m 4			–m 2		
cstacks	–n 1	–n 3	–n 5	–n 1	–n 3	–n 5	–n 1	–n 3	–n 5
	–report_matches								
sstacks	N/A								
rxstacks	–lnl_filter								
	–lnl_limit -15.0								
	–conf_filter								
	–prune_haplo								
populations	–write-random-snp								
	–vcf								
Dataset ID	*batch161*	*batch163*	*batch165*	*batch341*	*batch343*	*batch345*	*batch521*	*batch523*	*batch525*

Note the differences between data sets in –M (ustacks), –m (ustacks), and –n (cstacks).

were reformatted in PGDSpider v2.0.5.2 (Lischer and Excoffier, 2012) for analysis in BayeScan v2.1 (Foll and Gaggiotti, 2008; Foll et al., 2010; Fischer et al., 2011), which identifies loci which may be under natural selection, as well as GenAlEx v6.501 (Peakall and Smouse, 2006, 2012).

Population Genomics Analyses

Pairwise measures between regions, including Nei's unbiased genetic distances, which describe allelic differences assuming genetic drift and mutation are in equilibrium (Nei, 1972, 1987, and F_{ST} values, which quantifies the proportion of genetic variation explained by population structure (Wright, 1950), were calculated in the Excel data analysis suite, GenAlEx v6.501. GenAlEx was also used to identify private alleles within each region and conduct the Analyses of Molecular Variance (AMOVAs). The number of private alleles identified for every region were normalized by each region's sample size (PA_{norm}). Pairwise F_{ST} values were calculated alongside the AMOVAs utilizing GenAlEx's "AMOVA" option. Standard permutation was selected to calculate statistical significance of results over 999 permutations. Missing data were not imputed. Neighbor Joining (NJ) trees and Principal Component Analyses (PCAs) were constructed in the R package, *adegenet* (Jombart, 2008; Jombart and Ahmed, 2011). Three principal components (PCs) were calculated for each dataset, plotting the primary and secondary PCs with *ggplot2* (Wickham, 2016). Ellipses, encompassing the 0.95 confidence levels, were added for each region. Finally, using the "dapc" command in *adegenet*, Discriminant Analyses of Principal Components (DAPCs) were built from the first three PCs for each dataset.

Population structure was tested in STRUCTURE v2.3.4 (Pritchard et al., 2000) with K taking values between 2 and 7, each tested 10× under the admixture model with allele frequencies correlated among populations. Initially, each analysis ran for 100,000 generations, and the first 25% were discarded as burn-in. Review of preliminary results found high agreement between replicates, indicating that this number of generations was sufficient to achieve convergence. After STRUCTURE analyses were complete, results were collated in STRUCTURE HARVESTER v0.6.94 (Earl and VonHoldt, 2012). Within STRUCTURE HARVESTER, the optimal K value was inferred using *ad hoc* posterior probability models (Pritchard et al., 2000) and the Evanno Method (Evanno et al., 2005). STRUCTURE plots were generated within the R package *pophelper* (Francis, 2017).

Validating the Existing Biophysical Oceanographic Model

While Criales et al. (2015) presented a suite of models, for simplicity, here we focus only on the model that incorporates the larval behaviors of DVM and STST and describes the major and minor routes, as this was the only model that resulted in successful recruitment. Testing the hypotheses indicated by the modeling work of Criales et al. (2015) may be accomplished through patterns of unique haplotypes (private alleles), measures of genetic distance (Nei's unbiased distance), population differentiation (F_{ST}), and components of genetic variance (AMOVA). It is important to note that statistical tests are performed on values calculated from pseudoreplicated datasets (batches), not fully independent data.

Expectations under the null hypothesis: Most of the survivorship research on *F. duorarum* has focused on recruitment success (Browder et al., 1999, 2002; Ehrhardt and Legault, 1999; Criales et al., 2006, 2007, 2015), describing a density-dependent trend (Ehrhardt et al., 2001). By definition, the spawning aggregation represents the highest population density of sexually mature, spawning shrimp. Adults found on nearshore nursery grounds have matured on those grounds or in nearby estuaries and will soon return to spawning grounds for

their turn at spawning. Spawning aggregations hold greater genetic diversity than found on any one nursery ground when spawners come from several nursery locations. Under the null hypothesis, we expect the highest number of private alleles-per-individual (PA_{norm}) to come from sites representing a spawning aggregation. A *t*-test, assuming unequal variance, was utilized to statistically compare PA_{norm} for spawning (DT and MQ) vs. nursery (NBB, BB, SBB, EVG, and NEVG) regions. Most estuaries from which samples were collected for this study represent nursery areas, although young shrimp may move out of an estuary to avoid disruptive changes in conditions such as storms or cold snaps (e.g., see Tabb et al., 1962, pp. 26–27).

Finally, under the null, we expect little-to-no statistically significant pairwise population differentiation between regions; the vast majority of genetic variance should come from differences between individuals (F_{IT}). Pairwise F_{ST} values and AMOVA results will provide support in this regard.

Expectations under the alternative hypothesis: If the Dry Tortugas and the Marquesas support population-specific spawning aggregations, we expect statistically significant pairwise population differentiation between these sites, which was tested with an ANOVA comparing pairwise F_{ST} values by region type: spawning-spawning (DryTortugas-Marquesas), spawning-nursery (all region pairs containing DryTortugas or Marquesas), and nursery-nursery (all region pairs that do not contain DryTortugas or Marquesas). Moreover, while the majority of molecular variance may be attributable to variance among individuals (F_{IT}), F_{ST} should be greater than zero and statistically significant.

In additional to the statistical tests described, PCAs, DAPCs, and STRUCTURE results were evaluated for evidence of population structure. Any results, quantitative or qualitative, contradictory to both hypotheses will be considered as contradictions to the validity of the model presented by Criales et al. (2015), and the relative strength of such contradictions will be assessed in the context of the full study presented here.

RESULTS

The preparation of next-generation sequencing libraries occurred for 68 individuals collected from 19 sites representing seven regions. Over 117 million SR150 raw reads were returned from the Illumina HiSeq4000. Demultiplexed data were submitted to the NCBI SRA database under BioProject PRNJA554161 and are also publicly available through the Gulf of Mexico Research Initiative's Information and Data Cooperative (doi: 10.7266/n7-hhnq-kh83; Timm, 2019). Nine parameterizations of STACKS yielded nine data assemblies (batches, see **Table 1**) with 11,971–20,820 single nucleotide polymorphisms (SNPs). Additional quality filtering was executed in *vcftools*: when minimum read depth was set to 10×, 4,025–13,267 SNPs remained; applying a missing data filter (<90% individual missingness and <50% missing SNP data allowed) resulted in 740–800 high-confidence SNPs. BayeScan identified no loci under selection. See **Table 2** for a detailed report of this information.

The sample sizes across regions included in the research presented here could be considered low compared to traditional population genetics studies of microsatellites or multilocus datasets. However, reduced representation library (RRL) approaches, such as ddRADseq, generate vastly more data, sampled from across the genome of each individual, and this increase in genomic data for each individual empowers the detection of fine-scale population structure with substantially fewer samples (Willing et al., 2012; Jeffries et al., 2016; Nazareno et al., 2017).

TABLE 2 | Details of the assembled and quality-filtered ddRADseq datasets are presented.

		batch 161	*batch 163*	*batch 165*	*batch 341*	*batch 343*	*batch 345*	*batch 521*	*batch 523*	*batch 525*
Data assembly	Raw reads					117,257,163				
	Passed STACKS	16,315	16,868	17,205	20,617	20,332	20,820	12,005	11,971	11,974
	Passed minDP 10×	12,292	12,979	13,267	10,584	10,281	10,278	4,083	4,034	4,025
	Passed missing data filter	799	800	795	761	746	763	771	740	756
	Passed BayeScan	799	800	795	761	746	763	771	740	756
Sample sizes	*N*	57	56	57	57	57	56	57	57	57
	North of Biscayne Bay	9	9	9	9	9	9	9	9	9
	Biscayne Bay	9	9	9	9	9	9	9	9	9
	South Biscayne Bay	3	3	3	3	3	3	3	3	3
	Everglades	9	9	9	9	9	9	9	9	9
	North of Everglades	7	7	7	7	7	7	7	7	7
	Marquesas	10	9	10	10	10	9	10	10	10
	Dry Tortugas	10	10	10	10	10	10	10	10	10

Information about the numbers of reads and SNPs passing each step of data assembly and filtering are reported in the upper section of the table, including the number of raw reads, the number of SNPs assembled within STACKS, and the number of SNPs that passed quality filtering (including minimum read depth of 10×, site missingness of 50%, individual missingness of 90% allowed, and removal of sites under selection). The lower section of the table reports final sample sizes for each region in the datasets that were analyzed in this study.

Population Genomics Analyses

Nine datasets were analyzed to better understand the robustness of results to data assembly decisions, however, results across datasets were highly similar. While all results are reported in the **Supplementary Materials**, only results from batch161, the dataset with the highest number of samples and SNPs (N = 57, SNPs = 799) are presented in-text. Because very few samples representing the SouthBiscayneBay region were retained following quality filtering (n = 3), these samples were removed for calculation of Nei's unbiased distance calculation, estimation of pairwise F_{ST} between populations, and AMOVAs. The SouthBiscayneBay samples were included in PA_{norm}, PCAs, NJ trees, and STRUCTURE analyses.

Estimates of Nei's unbiased genetics distance between all region-pairs, excluding SouthBiscayneBay, ranged from 0.003 to 0.006, with the highest value attributable to the comparison between BiscayneBay and North_of_Everglades (**Figure 2A** and **Supplementary Figure 1**). The shortest genetic distance was calculated for multiple region-pairs: North_of_BiscayneBay compared to either spawning region (DryTortugas and Marquesas), Everglades compared to either spawning region, and North_of_BiscayneBay compared to Everglades. All other region-pair distances fell between 0.004 and 0.005 (**Supplementary Table 2**).

Pairwise comparisons between regions, excluding SouthBiscayneBay, were also examined through estimates of F_{ST} (**Figure 2B** and **Supplementary Figure 2**), which ranged from 0.000 to 0.102. Many region-pairs returned null F_{ST} values: all comparisons including BiscayneBay or North_of_Everglades and a region associated with the nursery range of *F. duorarum* (Everglades and North_of_BiscayneBay), as well as the Marquesas-DryTortugas region pair. With the exception of the North_of_BiscayneBay-Marquesas region pair, all non-zero F_{ST} were characteristic of region pairs that included a spawning region, with the highest F_{ST} values calculated between DryTortugas and Everglades (**Supplementary Table 3**), though recall that DryTortugas-Everglades had a very low genetic distance.

Analyses of Molecular Variance across all nine datasets, excluding SouthBiscayneBay, yielded an average among-population variance value of 1.69% (standard deviation 0.71%, **Table 3**). The vast majority of molecular variance was attributable to differences among individuals (88.49%, standard deviation 1.68%) and the remainder came from differences within individuals. Overall average F_{ST} (the proportion of total genetic variance found within a population), F_{IS} (the proportion of genetic variance in a population which is found within an individual from that population), and F_{IT} (the proportion of

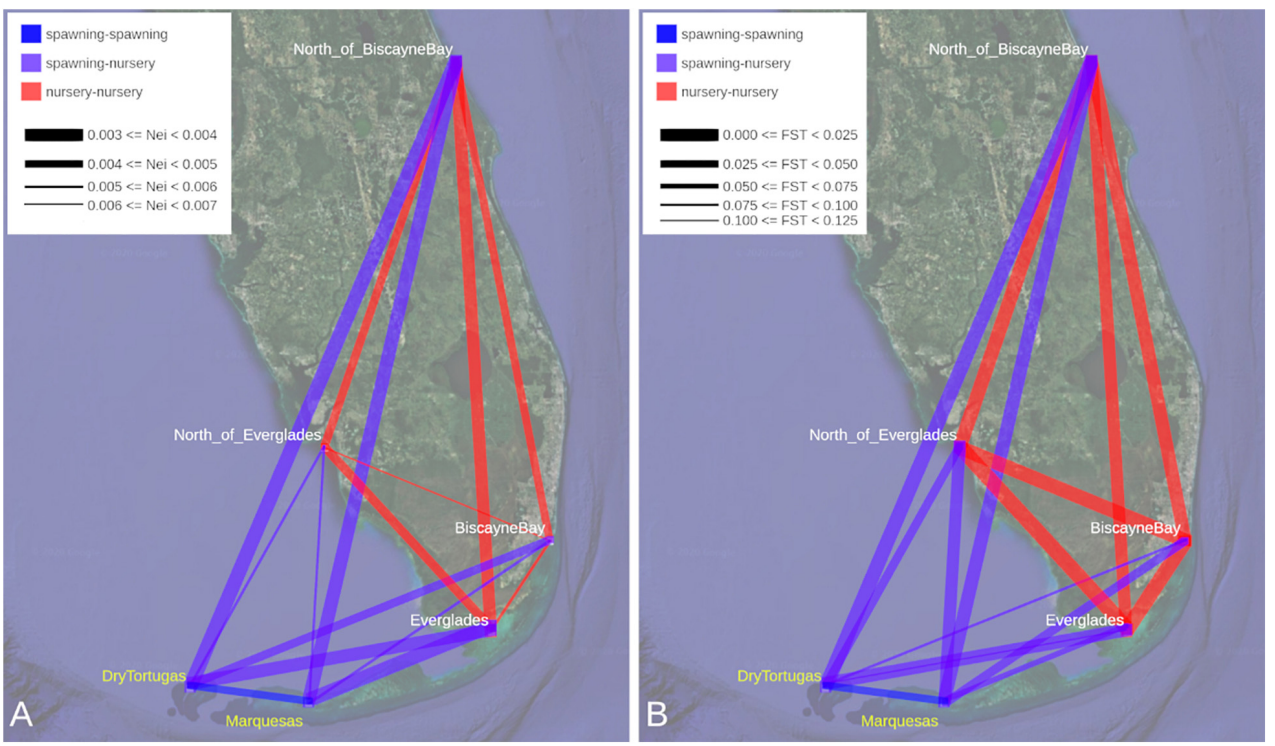

FIGURE 2 | (A) Nei's unbiased distance values and **(B)** pairwise F_{ST} values are depicted between all regions. Due to the small sample size of the SouthBiscayneBay region (n = 3), samples from this region have been excluded from these analyses. Detailed results for each dataset are presented in **Supplementary Tables 2, 3** (Nei's and F_{ST}, respectively) and distributions are plotted in **Supplementary Figures 1, 2** (Nei's and F_{ST}, respectively). Line color indicates the type of region pair: nursery region to nursery region (red), nursery region to spawning region (purple), or spawning region to spawning region (blue). Line width indicates genetic distance with greater genetic distance or higher F_{ST} illustrated with a narrower connecting line. Please note that these lines solely represent pairwise values, not movement of individual shrimp between regions. Spawning regions are labeled in yellow (Base map from Google Earth).

TABLE 3 | Analyses of molecular variance (AMOVA) results and *F* statistics.

	batch 161	*batch 163*	*batch 165*	*batch 341*	*batch 343*	*batch 345*	*batch 521*	*batch 523*	*batch 525*	**AVG**	**SD**
Among populations	0.99%	2.41%	1.28%	1.42%	1.47%	1.30%	3.37%	1.64%	1.30%	1.69%	0.70%
Among individuals	91.21%	89.08%	90.61%	89.23%	87.87%	88.75%	85.75%	87.14%	86.79%	88.49%	1.68%
Within individuals	7.79%	8.51%	8.11%	9.34%	10.66%	9.95%	10.88%	11.23%	11.91%	9.82%	1.38%
FST	0.010	**0.024***	0.013	0.014	0.015	0.013	**0.034***	0.016	0.013	0.017	0.007
FIS	**0.921***	**0.913***	**0.918***	**0.905***	**0.892***	**0.899***	**0.887***	**0.886***	**0.879***	0.900	0.014
FIT	**0.922***	**0.915***	**0.919***	**0.907***	**0.893***	**0.900***	**0.891***	**0.888***	**0.881***	0.902	0.014

**Indicates statistical significance.*

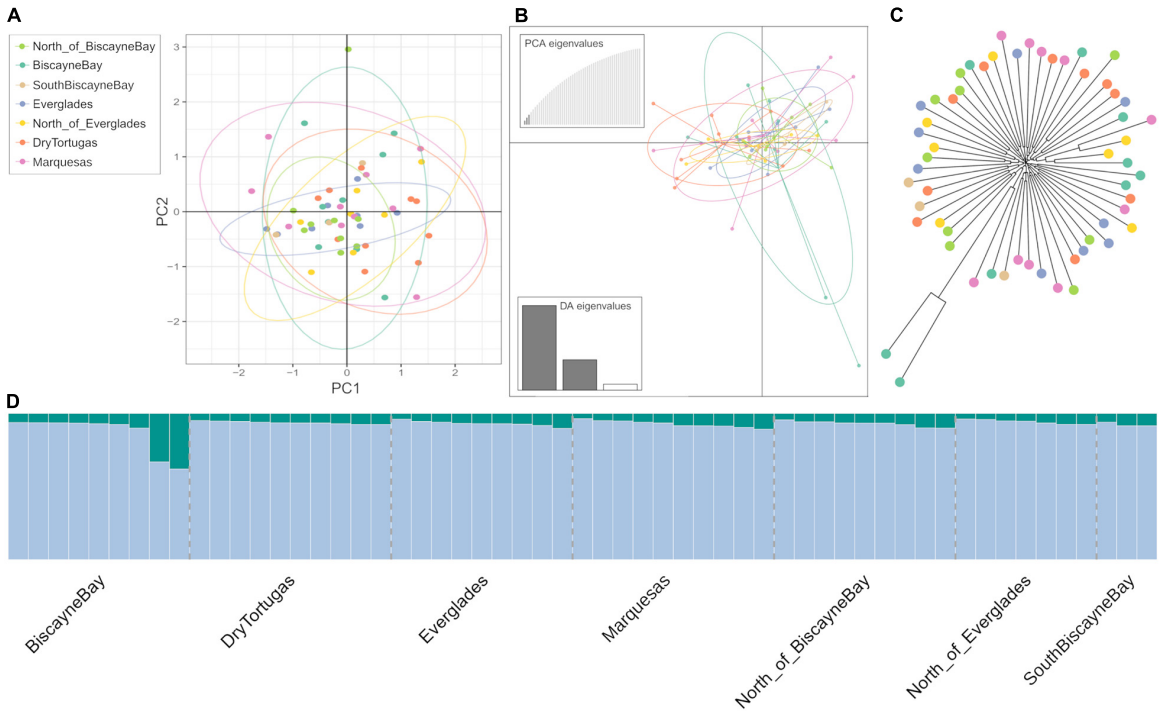

FIGURE 3 | (A) Principal component analysis (PCA; PC1 and PC2 explain 4.1 and 4.0%, respectively). **(B)** Discriminant Analysis of Principal Components (DAPC), and **(C)** Neighbor Joining Tree (NJ) of the batch161 dataset. All seven regions are represented and points are color-coded by collection site. **(D)** STRUCTURE results are reported for all seven regions (reported beneath the plot) for batch161. *K* = 2 was deemed optimal using the Evanno method. Results across all nine datasets are presented in **Supplementary Figures 3–6** (PCAs, DAPCs, Njs, and STRUCTURE plots, respectively).

total genetic variance found within an individual) reflected these values as well (0.017 ± 0.007, 0.900 ± 0.014, and 0.902 ± 0.014, respectively). Across the nine AMOVAs, F_{ST} was statistically significant in two cases, while F_{IS} and F_{IT} were statistically significant in every case.

Results from Principal Component Analysis (batch161 presented in **Figure 3A**; all batches presented in **Supplementary Figure 3**) and DAPCs (batch161 presented in **Figure 3B**, all batches presented in **Supplementary Figure 4**) include all samples, revealing a large, central cluster. However, samples from the DryTortugas are slightly shifted from the center. The NJ results (batch161 presented in **Figure 3C**; all batches presented in **Supplementary Figure 5**), which included samples from SouthBiscayneBay, show little structure. Across NJ trees, only two BiscayneBay samples are differentiated from the otherwise unstructured tree, but the other BiscayneBay samples do not reflect a larger separation of the region from the rest of the samples. To ensure these individuals did not represent contamination, we confirmed the taxonomic identification of these two samples as *F. duorarum*.

Further examining relationships between samples with the *K*-means clustering method STRUCTURE, the Evanno method was applied to identify the optimal *K* in each analysis. Across datasets, the Evanno method identified *K* = 2 as the optimal number of clusters within the data (**Supplementary Figure 6**). The two BiscayneBay samples differentiated in the NJ trees are clearly seen in the STRUCTURE plots as representing higher proportions of the minor cluster, otherwise all individuals appear highly similar, regardless of collection region (**Figure 3D**).

Normalized counts of private alleles within each region (PA_{norm}), including SouthBiscayneBay, ranged from 7.3 (standard deviation 0.7, Everglades) to 10.4 (standard deviation

1.1, BiscayneBay) private alleles per sampled individuals (**Figure 4** and **Supplementary Table 4**). With the exception of BiscayneBay, spawning regions had higher normalized private allele counts (Marquesas = 10.2 ± 1.2, DryTortugas = 10.1 ± 0.6) than regions from the nursery range.

Validating the Existing Biophysical Oceanographic Model

Expectations under the existing model were evaluated through several tests of significance (**Table 4**): to begin, we evaluated whether PA_{norm} differed significantly between spawning regions and nursery regions. A one-tailed, two-sample *t*-test assuming unequal variances between PA_{norm} of nursery regions (North_of_BiscayneBay, BiscayneBay, SouthBiscayneBay, Everglades, and North_of_Everglades) and spawning regions (DryTortugas and Marquesas) indicated significantly higher PA_{norm} in spawning regions ($t_{stat} = -4.46$; $p = 2.23 \times 10^{-5}$). Next, we performed two single-factor ANOVAs to test whether Nei's unbiased genetic distances or pairwise F_{ST} values differed significantly between types of region-pairs: spawning-spawning, spawning-nursery, and nursery-nursery (**Table 4**). The ANOVA analyzing Nei's unbiased distances between region-pairs did not detect a statistically significant difference between region-pair types ($F_{stat} = 1.95$; $p = 0.15$). The ANOVA analyzing pairwise F_{ST} values, however, yielded a statistically significant result ($F_{stat} = 42.83$; $p = 4.63 \times 10^{-15}$). We followed the ANOVAs with three two-tailed, paired *t*-tests comparing PA_{norm}, Nei's unbiased genetic distances, and pairwise F_{ST} between all nursery regions and DryTortugas to all nursery regions and Marquesas. The normalized number of private alleles and Nei's distances did not differ significantly by spawning region (PA_{norm} $t_{stat} = 2.31$, $p = 0.91$; Nei's $t_{stat} = 0.48$, $p = 0.63$), while pairwise F_{ST} values were significantly higher between region-pairs including DryTortugas compared to region-pairs including Marquesas ($t_{stat} = -8.22$; $p = 1.09 \times 10^{-9}$).

DISCUSSION

The study presented here used next-generation sequencing data to inform management strategies by characterizing the population dynamics of *F. duorarum* around the Florida Peninsula, with specific focus on the role of migration from spawning aggregations to nursery grounds. Much of this work was motivated by the biophysical oceanographic model of

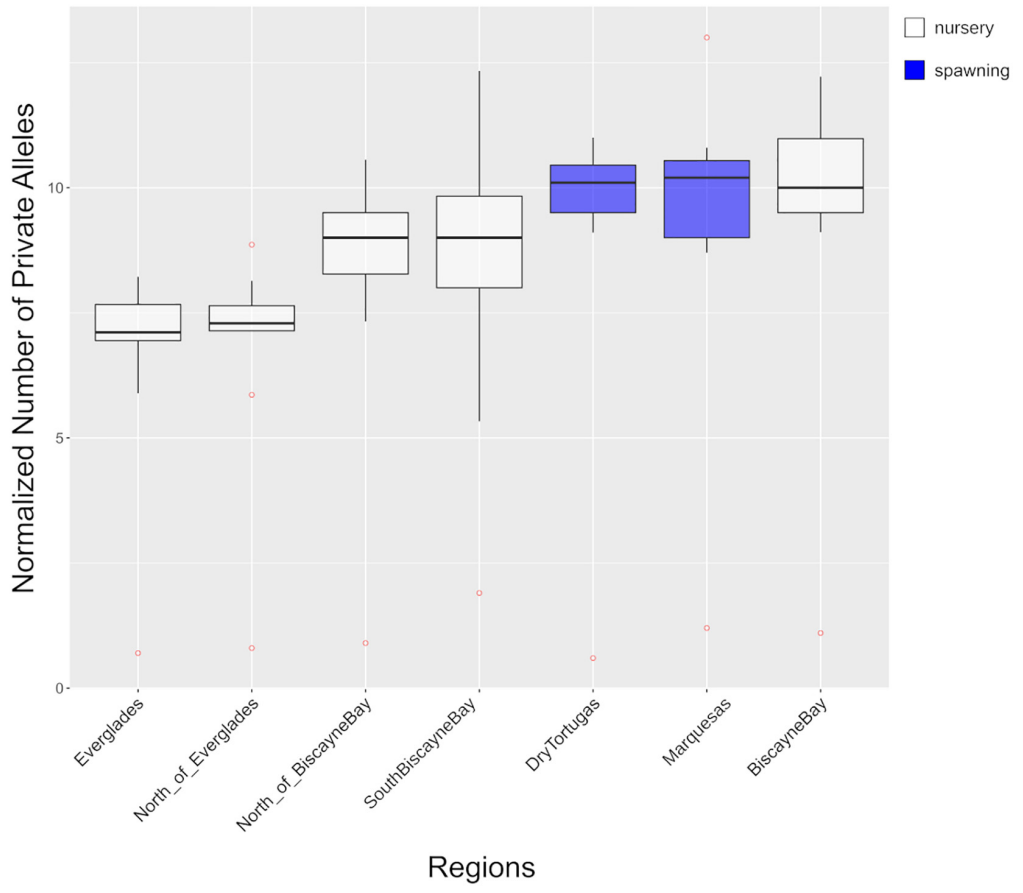

FIGURE 4 | The number of private alleles, normalized by each region's sample size (referred to as PA_{norm} in-text), are presented for each region. Detailed results for each dataset are presented in **Supplementary Table 4**. Box color indicates the type of region: nursery region (white) or spawning region (blue). Outliers are red.

TABLE 4 | Results of all significance tests comparing region types (spawning vs. nursery), region-pair types (spawning-spawning vs. spawning-nursery vs. nursery-nursery region pairs), and spawning regions (Marquesas vs. DryTortugas).

Significance test	Samples	Statistic value	*p*
One-tailed two-sample *T*-test assuming unequal variance*	**PA_{norm} spawning vs. PA_{norm} nursery**	**t_{stat} = –4.46**	**2.23E-5**
Single factor ANOVA	Spawning-spawning vs. spawning-nursery vs. nursery-nursery region pairs (Nei's unbiased genetic distance)	F_{stat} = 1.95	0.15
Single factor ANOVA*	**Spawning-spawning vs. spawning-nursery vs. nursery-nursery region pairs (pairwise F_{ST} values)**	**F_{stat} = 42.83**	**4.63E-15**
Two-tailed Paired *T*-test	PA_{norm} Marquesas vs. PA_{norm} DryTortugas	t_{stat} = 2.31	0.91
Two-tailed Paired *T*-test	Marquesas-all vs. DryTortugas-all (Nei's unbiased genetic distance)	t_{stat} = 0.48	0.63
Two-tailed Paired *T*-test*	**Marquesas-all vs. DryTortugas-all (pairwise F_{ST} values)**	**t_{stat} = -8.22**	**1.09E-9**

*In all cases, the significance test, comparison of interest (Samples), test statistic value, and p-value are presented. *Indicates a statistically significant test (where $p < \alpha$ when $\alpha = 0.05$).*

larval transport from spawning aggregations offshore of the Dry Tortugas and Marquesas to nursery grounds in Florida Bay (Criales et al., 2015). The model supported two migration routes from spawning regions to nursery grounds: the major route crosses the southwest Florida shelf in a fairly direct east-northeast path (Munro et al., 1968; Criales et al., 2003); the minor route involves downstream transport along the Florida Current, bringing larvae east-northeast with the Current and then breaking with the Florida Current to move west-northwest toward Florida Bay through the passes in the Middle and Lower Florida Keys (Jones et al., 1970; Criales et al., 2006). These two routes have the potential to sustain population differentiation within the species, representing overlooked biodiversity. Independent analysis of next-generation sequencing data revealed some population differentiation associated with the Dry Tortugas. With some caveats, the work presented here provides strong support for the model of larval migration and recruitment developed by Criales et al. (2015).

Utilizing Population Genomics Data to Validate a Biophysical Oceanographic Model

There is no paucity of potentially confounding variables when modeling current- and tide-mediated transport of dispersing larvae: the oversimplification of active swimming behaviors and the disparity between potential and realized dispersal has been described previously, including the biological importance of single individuals occasionally dispersing long distances (Shanks, 2009). However, biophysical modeling can be used in concert with genetic evidence to improve our understanding of the dynamic relationships between marine organisms and their environment (Liggins et al., 2013; Timm et al., 2020). Such an integrative approach has been utilized in studies of marine invertebrate populations (Dawson et al., 2005), including a recent study investigating the causes of population structure in an economically important decapod, the spiny lobster *Panulirus argus* (Truelove et al., 2017).

The biophysical oceanographic model developed by Criales et al. (2015) describes two migratory routes, which differ in their origin (Dry Tortugas and Marquesas vs. Dry Tortugas only), usage (many vs. few individuals, represented as particles), and recruitment success (majority of particles are successfully recruited to Florida Bay vs. few particles are successfully recruited). These differences have the potential to maintain intraspecific diversity via population differentiation. It is important to note that no model perfectly reflects reality; while the model developed by Criales et al. (2015) accounts for direction and velocity across water depth, this information is not discussed in the work. However, the model provides three questions that can be addressed with population genomics: Is there independent support for the model? Do the modeled spawning aggregations sufficiently explain the genomic results? Do we see evidence that the minor route sustains a differentiated population?

Next-generation sequencing data provided strong support for the existing model of larval transport: across analyses, samples collected from the Marquesas and Dry Tortugas were clearly part of a larger population present across the Florida Peninsula (see **Figure 3**). The presence of significantly more private alleles in the spawning regions compared to the nursery sites (**Table 4**) further supports the model of larvae originating from the Dry Tortugas and the Marquesas. It is worth explicitly addressing the two Biscayne Bay outliers identified throughout clustering analyses in **Figure 3**, which suggest recruitment to Biscayne Bay from a spawning aggregation that was not sampled in this study. In this regard, the existing model, which simulates spawning aggregations in the Marquesas and the Dry Tortugas, may not be complete and an additional spawning site contributes recruits to the region.

Evidence of Population Structure in the Study Region

Under the null hypothesis, we expect one homogeneous population present throughout the study region. While cluster analyses (PCA, DAPC, and STRUCTURE) do not clearly delineate populations, we see some shifting of samples from the Dry Tortugas (**Figure 3**), and statistical tests of population differentiation (global and pairwise) indicate low levels of structure throughout (though these values are only rarely statistically significant). This structure provides evidence for the alternative hypothesis: the separate spawning aggregations and

migratory routes (major and minor) support genetic structure in the pink shrimp population around the Florida Peninsula.

With few exceptions, significant pairwise population differentiation was highest and statistically significant when regions from the nursery range were compared to spawning regions (**Figure 2**, **Supplementary Figure 2**, and **Supplementary Table 3**). Examining pairwise population differentiation by region pair type (spawning-spawning vs. spawning-nursery vs. nursery-nursery) revealed significant differences (**Table 4**), with differentiation between the spawning-spawning pair < nursery-nursery pair < spawning-nursery pair. To a large extent, the Dry Tortugas seems to be driving this trend: analyses of population differentiation indicate the Marquesas region is better genetically connected to the nursery regions than the Dry Tortugas region is (**Figure 2B** and **Supplementary Table 3**). It should be noted that the highest significant pairwise population differentiation calculated in this study was relatively low, but low, statistically significant F_{ST} estimates are fairly common in the marine realm (Waples, 1998; Waples and Gaggiotti, 2006; Hauser and Carvalho, 2008; Therkildsen et al., 2013; Timm et al., 2020). This would also explain the lack of clear structuring in clustering analyses.

The presence of a differentiated population in the Dry Tortugas (hereafter referred to as the "Dry Tortugas subpopulation") is unexpected. Recall that larvae are spawned offshore of the Dry Tortugas and the Marquesas. Larvae pass through a series of developmental stages as they migrate, taking the major or minor route (**Figure 1**), to estuarine nursery grounds around the Florida Peninsula where they complete their maturation into adults. Year-round, these adults migrate back to the spawning aggregations to reproduce, which should, theoretically, lead to sufficient mixing to result in a single, genetically homogeneous population. We suspect the maintenance of a Dry Tortugas subpopulation may be the result of geographic or temporal separation of spawning populations. By the geographic mechanism, the Dry Tortugas subpopulation spawns exclusively in the Dry Tortugas and solely utilizes the minor migratory route, while the larger population spawns in the Dry Tortugas and the Marquesas and utilizes the major migratory route. However, the lack of clearly defined genetic structure separating the Dry Tortugas subpopulation from the larger population suggests this geographic mechanism is not sufficient to explain the results presented here.

The population structure we identify may also be the result of a temporal mechanism: since the 1980s, Key West shrimpers have reported anecdotal evidence of two spawning surges annually for the past several decades (pers. comm.) and unpublished data of Robblee suggest two peaks in population abundance of juvenile pink shrimp in western Florida Bay (pers. comm.). Costello and Allen (1966) also remark on the seasonal nature of juvenile pink shrimp in the region. Without additional data, it is difficult to characterize this mechanism further; however, if adults of the Dry Tortugas subpopulation arrive at the Dry Tortugas spawning ground before or after the larger aggregation, they will only be able to reproduce with each other. Moreover, depending on the seasonal timing of this second spawning surge, larvae originating from the Dry Tortugas subpopulation may utilize the minor migratory route to Florida Bay, leading to higher mortality and lower recruitment success. Given the lack of a clearly distinguishable Dry Tortugas subpopulation in the clustering analyses, it may be that such a mechanism results in the differentiation of the Dry Tortugas subpopulation, with occasional gene flow between it and the larger population preventing strong genetic structuring.

Either mechanism, geographic or temporal, might be facilitated by local recruitment of the Dry Tortugas population to the Dry Tortugas or a region not represented by samples collected for this study. In line with the alternative hypothesis, we find evidence of an unsampled spawning aggregation contributing individuals to Biscayne Bay and the Everglades: both regions show low-but-significant differentiation from the Marquesas and the Dry Tortugas, but no differentiation between themselves. The Loop Current's episodic influence may bring migrants into nearshore currents, bringing recruits to Biscayne Bay from the Caribbean (Saloman et al., 1968). Alternatively, migrants may be contributed from the Sanibel spawning aggregation. A previous mark-recapture study found that, while geographic ranges of stocks from the Dry Tortugas and Sanibel overlap in nursery grounds, there is only evidence of weak, one-way migration of Sanibel stocks to the Dry Tortugas (Costello and Allen, 1966). Such separation between spawning grounds could provide a basis for population differentiation. Interestingly, this mark-recapture study did not find any evidence of shrimp migration between Biscayne Bay and the Sanibel grounds, nor between Biscayne Bay and the Dry Tortugas; indeed, all individuals marked and released within Biscayne Bay were only ever recovered from Biscayne Bay. The results presented here contradict this study, finding gene flow between the Dry Tortugas and Biscayne Bay (though Biscayne Bay may also receive recruits from a spawning aggregation that was not sampled in the current study).

Spawning-recruitment relationships of pink shrimp in south Florida appear to be more complex than previously believed and additional research is needed to investigate the mechanisms we hypothesize here. Representative sampling of *F. duorarum* from Sanibel, Cuba, and the Bahamas would be needed to further investigate the relative support for these potential sources of postlarval migrants. Anecdotal evidence of spawning surges, and the role this may play in the population structure of pink shrimp around the Florida Peninsula, would require a longitudinal study to better understand this mechanism.

Relevance to Fisheries Management

The fisheries supported by *F. duorarum* contribute to economies internationally (Sheridan, 1996; Ramírez-Rodríguez et al., 2003; Hart et al., 2012), and the continued exploitation of this natural resource is critically dependent on the stability and sustainability of the species in the Gulf of Mexico and around the Florida Peninsula. One crucial factor contributing to species stability is successful larval recruitment: the movement of larval and postlarval individuals from spawning aggregations to nursery grounds.

Our results support the biophysical oceanographic model developed by Criales et al. (2015), which indicates a major route, traversed by larvae from the Dry Tortugas and the Marquesas,

and a minor route, which only resulted in successful recruitment when larvae originated from the Dry Tortugas. Moreover, we find evidence that samples from the Dry Tortugas represent a differentiated population. Co-located with this region is a pink shrimp fishery (Klima et al., 1987; Hart et al., 2012), which harvests mature and young adult shrimp year-round on the lower southwest Florida shelf (Ehrhardt and Legault, 1999; Browder et al., 2002), perhaps with important implications for intraspecific genetic diversity: individuals harvested near the Dry Tortugas may represent the subpopulation indicated by our analyses. The removal of these individuals could undermine the subpopulation's stability by reducing the density of juveniles and subsequently decreasing recruitment success (Ehrhardt et al., 2001).

Additional work is needed to further characterize the role of these two spawning grounds and migration routes, particularly by including individuals collected from the Sanibel grounds and the Caribbean. Such research will assist in determining whether the species should be managed as a single stock or if more complex management is required. Enhancing our understanding of larval recruitment success in *F. duorarum* will ultimately improve the long-term sustainability of these fisheries while protecting diversity within the species.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: https://www.ncbi.nlm.nih.gov/, PRJNA554161; https://data.gulfresearchinitiative.org, doi: 10.7266/n7-hhnq-kh83.

AUTHOR CONTRIBUTIONS

TJ collected samples for inclusion in this study. JB and HB-G provided financial support for the research. LT performed the lab work and data analysis, generating an early version of this manuscript, including all tables, figures, and **Supplementary Materials**. All authors contributed to the study design and final manuscript.

FUNDING

This research was made possible by a grant from the Gulf of Mexico Research Initiative (GOMRI) and NOAA RESTORE project (to Bracken-Grissom), as well as the Colorado Biomedical Informatics Training Program (to Timm, NIH T15 LM009451).

ACKNOWLEDGMENTS

Logistic support for this study was provided by the Protected Resources and Biodiversity Division of the Southeast Fisheries Science Center, National Marine Fisheries Service, Miami, Florida. The authors also extend their gratitude to the shrimpers who donated samples for inclusion in this study, as well as to the port agents (Thomas Herbert, Ft. Myers Beach, and Edwin Pulido, Key West) who collected the samples from the shrimpers and shipped them to Miami. The authors further thank Ms. Michelle Harangody for her assistance in taxonomically identifying putative specimens of *Farfantepenaeus duorarum*, Ms. Shaina Simon for her early work on this project, as well as Emily Warschefsky for her support in the library preparation stage of the study. Special gratitude is due to Ed Little for providing the initial anecdotal evidence of temporally staggered spawning events in the region in 1995. Finally, the authors thank Maria Criales and the reviewers for their feedback on earlier versions of this manuscript. This is contribution #285 from the Coastlines and Oceans Division of the Institute of Environment at Florida International University. The scientific results and conclusions, as well as any views or opinions expressed herein, are those of the authors and do not necessarily reflect those of NOAA or the Department of Commerce.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fevo.2021.659134/full#supplementary-material

REFERENCES

Andrews, S. (2010). *FastQC: A Quality Control Tool for High Throughput Sequence Data*. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc (accessed April 2020).

Arreguín-Sánchez, F., Sánchez, J. A., Flores-Hernández, D., Ramos-Miranda, J., Sánchez-Gil, P., and Yáñez-Arancibia, A. (1999). "Stock–recruitment relationships (SRRs): a scientific challenge to support fisheries management in the Campeche Bank, Mexico," in *The Gulf of Mexico Large Marine Ecosystem*, eds H. Kumpf and K. Sherman (Malden, MA: Blackwell Science), 225–235.

Arreguín-Sánchez, F., Schultz-Ruiz, L. E., Sanchez, J. A., Gracia, A., and Alarcon, T. (1997). "Estado actual y perspectvas del recurso camaron en prensa," in *Analisis y Diagnostico de los Recursos Pesqueros Criticos del Gulfo de Mexico*, 7th Edn, eds D. Flores-Hernandez, P. Sanchez-Gil, J. C. Seija, and F. Arreguin-Sanchez (Campeche: EPOMEX Serie Cientifica.), 185–203.

Arreguín-Sánchez, F., Zetina-Rejón, M., and Ramírez-Rodríguez, M. (2008). Exploring ecosystem-based harvesting strategies to recover the collapsed pink shrimp (*Farfantepenaeus duorarum*) fishery in the southern Gulf of Mexico. *Ecol. Model.* 214, 83–94. doi: 10.1016/j.ecolmodel.2007.11.021

Avise, J. C. (1992). Molecular population structure and the biogeographic history of a regional fauna: a case history with lessons for conservation biology. *Oikos* 63, 62–76. doi: 10.2307/3545516

Bernatchez, L. (1995). "A role for molecular systematics in defining evolutionarily significant units in fishes," in *Evolution and the Aquatic Ecosystem: Defining Unique Units in Population Conservation*, ed. J. L. Nielsen (Bethesda, MA: American Fisheries Society), 114–132.

Browder, J. A., Restrepo, V. R., Rice, J. K., Robblee, M. B., and Zein-Eldin, Z. (1999). Environmental influences on potential recruitment of pink shrimp, *Farfantepenaeus duorarum*, from Florida Bay Nursery Grounds. *Coast. Estuar. Res. Federat.* 22, 484–499. doi: 10.2307/1353213

Browder, J. A., and Robblee, M. B. (2009). Pink shrimp as an indicator for restoration of everglades ecosystems. *Ecol. Ind.* 9(Suppl.), 17–28. doi: 10.1016/j.ecolind.2008.10.007

Browder, J. A., Zein-Eldin, Z., Criales, M. M., Robblee, M. B., Wong, S., Jackson, T. L., et al. (2002). Dynamics of pink shrimp (*Farfantepenaeus duorarum*) recruitment potential in relation to salinity and temperature in Florida Bay. *Coast. Estuar. Res. Federat.* 25, 1355–1371. doi: 10.1007/bf02692230

Burkenroad, M. D. (1939). Further observations on the *Penaeidae* of the northern Gulf of Mexico. *Bull. Bingham Oceanogr. Collect.* 6, 1–62.

Catchen, J., Hohenlohe, P. A., Bassham, S., Amores, A., and Cresko, W. A. (2013). Stacks: an analysis tool set for population genomics. *Mol. Ecol.* 22, 3124–3140. doi: 10.1111/mec.12354

Condie, S. A., Loneragan, N. R., and Die, D. J. (1999). Modelling the recruitment of tiger prawns *Penaeus esculentus* and *P. semisulcatus* to nursery grounds in the Gulf of Carpentaria, northern Australia: implications for assessing stock-recruitment relationships. *Mar. Ecol. Prog. Ser.* 178, 55–68. doi: 10.3354/meps178055

Costello, T. J., and Allen, D. M. (1966). Migrations and geographic distribution of pink shrimp, *Penaeus duorarum*, of the Tortugas and Sanibel grounds, Florida. *Fish. Bull.* 65, 449–459.

Criales, M. M., Bello, M. J., and Yeung, C. (2000). Diversity and recruitment of penaeoid shrimps (*Crustacea*: *Decapoda*) at Bear Cut, Biscayne Bay, Florida, USA. *Bull. Mar. Sci.* 67, 773–788.

Criales, M. M., Browder, J. A., Mooers, C. N. K., Robblee, M. B., Cardenas, H., and Jackson, T. L. (2007). Cross-shelf transport of pink shrimp larvae: interactions of tidal currents, larval vertical migrations and internal tides. *Mar. Ecol. Prog. Ser.* 345, 167–184. doi: 10.3354/meps06916

Criales, M. M., Cherubin, L. M., and Browder, J. A. (2015). Modeling larval transport and settlement of pink shrimp in South Florida: dynamics of behavior and tides. *Mar. Coast. Fish.* 7, 148–176. doi: 10.1080/19425120.2014.1001541

Criales, M. M., Robblee, M. B., Browder, J. A., Cárdenas, H., and Jackson, T. L. (2010). Nearshore concentration of pink shrimp (*Farfantepenaeus duorarum*) postlarvae in northern Florida bay in relation to nocturnal flood tide. *Bull. Mar. Sci.* 86, 53–74.

Criales, M. M., Wang, J. D., Browder, J. A., Robblee, M. B., Jackson, T. L., and Hittle, C. (2006). Variability in supply and cross-shelf transport of pink shrimp (*Farfantepenaeus duorarum*) postlarvae into western Florida Bay. *Fish. Bull.* 104, 60–74.

Criales, M. M., Yeung, C., Jones, D. L., Jackson, T. L., and Richards, W. J. (2003). Variation of oceanographic processes affecting the size of pink shrimp (*Farfantepenaeus duorarum*) postlarvae and their supply to Florida Bay. *Estuar. Coast. Shelf Sci.* 57, 457–468. doi: 10.1016/S0272-7714(02)00374-8

Cummings, W. C. (1961). Maturation and spawning of the pink shrimp, *Penaeus duorarum* Burkenroad. *Trans. Am. FIsh. Soc.* 90, 462–468. doi: 10.1577/1548-8659(1961)90[462:masotp]2.0.co;2

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., et al. (2011). The variant call format and VCFtools. *Bioinformatics* 27, 2156–2158. doi: 10.1093/bioinformatics/btr330

Dawson, M. N., Gupta, A. S., and England, M. H. (2005). Coupled biophysical global ocean model and molecular genetic analyses identify multiple introductions of cryptogenic species. *Proc. Natl. Acad. Sci. U.S.A.* 102, 11968–11973. doi: 10.1073/pnas.0503811102

Ditty, J. G., and Alvarado Bremer, J. R. (2011). Species discrimination of postlarvae and early juvenile brown shrimp (*Farfantepenaeus aztecus*) and pink shrimp (*F. duorarum*) (*Decapoda*: *Penaeidae*): coupling molecular genetics and comparative morphology to identify early life stages. *J. Crust. Biol.* 31, 126–137. doi: 10.1651/10-3304.1

Dobkin, S. (1961). Early developmental stages of the pink shrimp *Penaeus duorarum* from Florida waters. *Fisheries* 1, 321–348.

Earl, D. A., and VonHoldt, B. M. (2012). STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. *Conserv. Genet. Resour.* 4, 359–361. doi: 10.1007/s12686-011-9548-7

Ehrhardt, N. M., and Legault, C. M. (1999). Pink shrimp, *Farfantepenaeus duorarum*, recruitment variability as an indicator of Florida Bay dynamics. *Estuaries* 22, 471–483. doi: 10.2307/1353212

Ehrhardt, N. M., Legault, C. M., and Restrepo, V. R. (2001). Density-dependent linkage between juveniles and recruitment for pink shrimp (*Farfantepenaeus duorarum*) in southern Florida. *ICES J. Mar. Sci.* 58, 1100–1105. doi: 10.1006/jmsc.2001.1101

Evanno, G., Regnaut, S., and Goudet, J. (2005). Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Mol. Ecol.* 14, 2611–2620. doi: 10.1111/j.1365-294X.2005.02553.x

Ewald, J. J. (1965). The laboratory rearing of pink shrimp. *Penaeus duorarum* Burkenroad. *Bull. Mar. Sci.* 15, 436–449.

Fischer, M. C., Foll, M., Excoffier, L., and Heckel, G. (2011). Enhanced AFLP genome scans detect local adaptation in high-altitude populations of a small rodent (*Microtus arvalis*). *Mol. Ecol.* 20, 1450–1462. doi: 10.1111/j.1365-294X.2011.05015.x

Foll, M., Fischer, M. C., Heckel, G., and Excoffier, L. (2010). Estimating population structure from AFLP amplification intensity. *Mol. Ecol.* 19, 4638–4647. doi: 10.1111/j.1365-294X.2010.04820.x

Foll, M., and Gaggiotti, O. (2008). A genome-scan method to identify selected loci appropriate for both dominant and codominant markers: a Bayesian perspective. *Genetics* 180, 977–993. doi: 10.1534/genetics.108.092221

Forward, R. B. J., and Tankersley, R. A. (2001). "Selective tidal-stream transport of marine animals," in *Oceanography and Marine Biology: An Annual Review*, eds R. N. Gibson, M. Barnes, and R. J. A. Atkinson (Milton Park: Taylor & Francis Inc), 305–353.

Francis, R. M. (2017). pophelper: an R package and web app to analyse and visualize population structure. *Mol. Ecol. Resour.* 17, 27–32. doi: 10.1111/1755-0998.12509

Hart, R. A. (2017). *Stock Assessment Update for Pink Shrimp (Farfantepenaeus duorarum) in the U.S. Gulf of Mexico for the 2016 Fishing Year (Issue December)*. Avaliable online at: http://www.galvestonlab.sefsc.noaa.gov/publications/pdf/936.pdf (accessed September 2020).

Hart, R. A., Nance, J. J. M., and Primrose, J. A. (2012). The US Gulf of Mexico Pink Shrimp, *Farfantepenaeus duorarum*, fishery: 50 years of commercial catch statistics. *Mar. Fish. Rev.* 74, 1–6.

Hauser, L., and Carvalho, G. R. (2008). Paradigm shifts in marine fisheries genetics: ugly hypotheses slain by beautiful facts. *Fish Fish.* 9, 333–362. doi: 10.1111/j.1467-2979.2008.00299.x

Jeffries, D. L., Copp, G. H., Handley, L. L., Håkan Olsén, K., Sayer, C. D., and Hänfling, B. (2016). Comparing RADseq and microsatellites to infer complex phylogeographic patterns, an empirical perspective in the Crucian carp, *Carassius carassius*, L. *Mol. Ecol.* 25, 2997–3018. doi: 10.1111/mec.13613

Jombart, T. (2008). adegenet: a R package for the multivariate analysis of genetic markers. *Bioinformatics* 24, 1403–1405. doi: 10.1093/bioinformatics/btn129

Jombart, T., and Ahmed, I. (2011). adegenet 1.3-1: new tools for the analysis of genome-wide SNP data. *Bioinformatics* 27, 3070–3071. doi: 10.1093/bioinformatics/btr521

Jones, A. C., Dimitriou, D. E., Ewald, J. J., and Tweedy, J. H. (1970). Distribution of early developmental stages of pink shrimp. *Penaeus duorarum*, in Florida waters. *Bull. Mar. Sci.* 20, 634–661.

Klima, E. F., Nance, J. M., Sheridan, P. F., Baxter, K. N., Patella, F. J., and Koi, D. B. (1987). *Review of the 1986 Texas Closure for the Shrimp Fishery off Texas and Louisiana*. Galveston, TX: NOAA Technical Memorandum. NMFS-SEFC-197.

Liggins, L., Treml, E. A., and Riginos, C. (2013). Taking the plunge: an introduction to undertaking seascape genetic studies and using biophysical models. *Geogr. Compass* 7, 173–196. doi: 10.1111/gec3.12031

Lischer, H. E. L., and Excoffier, L. (2012). PGDSpider: an automated data conversion tool for connecting population genetics and genomics programs. *Bioinformatics* 28, 298–299. doi: 10.1093/bioinformatics/btr642

Mastretta-Yanes, A., Arrigo, N., Alvarez, N., Jorgensen, T. H., Piñero, D., and Emerson, B. C. (2015). Restriction site-associated DNA sequencing, genotyping error estimation and *de novo* assembly optimization for population genetic inference. *Mol. Ecol. Resour.* 15, 28–41. doi: 10.1111/1755-0998.12291

Munro, J. L., Jones, A. C., and Dimitriou, D. (1968). Abundance and distribution of the larvae of the pink shrimp (*Penaeus duorarum*) on the Tortugas Shelf of Florida, August 1962-October 1964. *Fish. Bull.* 67, 165–181.

Nazareno, A. G., Bemmels, J. B., Dick, C. W., and Lohmann, L. G. (2017). Minimum sample sizes for population genomics: an empirical study from an

Amazonian plant species. *Mol. Ecol. Resour.* 17, 1136–1147. doi: 10.1111/1755-0998.12654

Nei, M. (1972). Genetic distance between populations. *Am. Nat.* 106, 283–292.

Nei, M. (1987). *Molecular Evolutionary Genetics.* New York, NY: Columbia University Press.

Ogburn, M. B., Criales, M. M., Thompson, R. T., and Browder, J. A. (2013). Endogenous swimming activity rhythms of postlarvae and juveniles of the penaeid shrimp *Farfantepenaeus aztecus*, *Farfantepenaeus duorarum*, and *Litopenaeus setiferus*. *J. Exp. Mar. Biol. Ecol.* 440, 149–155. doi: 10.1016/j.jembe.2012.12.007

O'Leary, S. J., Puritz, J. B., Willis, S. C., Hollenbeck, C. M., and Portnoy, D. S. (2018). These aren't the loci you're looking for: principles of effective SNP filtering for molecular ecologists. *Mol. Ecol.* 27, 3193–3206. doi: 10.1111/mec.14792

Paris, J. R., Stevens, J. R., and Catchen, J. M. (2017). Lost in parameter space: a road map for STACKS. *Methods Ecol. Evol.* 8, 1360–1373. doi: 10.1111/2041-210X.12775

Peakall, R., and Smouse, P. E. (2006). GENALEX 6: genetic analysis in Excel. Population genetic software for teaching and research. *Mol. Ecol. Notes* 6, 288–295. doi: 10.1111/j.1471-8286.2005.01155.x

Peakall, R., and Smouse, P. E. (2012). GenAlEx 6.5: genetic analysis in Excel. Population genetic software for teaching and research-an update. *Bioinformatics* 28, 2537–2539. doi: 10.1093/bioinformatics/bts460

Pérez-Farfante, I. (1969). *Western Atlantic Shrimps of the Genus Penaeus*. Washington, DC: U.S. Fish and Wildife Service Fishery Bulletin, 461–591.

Pérez-Farfante, I. (1970). *Claves Ilustradas Para la Identificación de los Camarones Comerciales de la América Latina*. No. Folleto 1605. Mexico.

Pérez-Farfante, I. (1988). Illustrated key to the penaeoid shrimps of commerce in the Americas. *NOAA Tech. Rep* 64:32.

Pérez-Farfante, I., and Kensley, B. F. (1997). Penaeoid and Sergestoid shrimps and prawns of the world: keys and diagnoses for the families and genera. *Memoires Du Museum National d'Histoire Naturelle* 175, 1–233. doi: 10.3897/zookeys.418.7629

Peterson, B. K., Weber, J. N., Kay, E. H., Fisher, H. S., and Hoekstra, H. E. (2012). Double digest RADseq: an inexpensive method for *de novo* SNP discovery and genotyping in model and non-model species. *PLoS One* 7:e37135. doi: 10.1371/journal.pone.0037135

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

Queiroga, H., and Blanton, J. (2005). Interactions between behaviour and physical forcing in the control of horizontal transport of decapod crustacean larvae. *Adv. Mar. Biol.* 47, 107–214. doi: 10.1016/s0065-2881(04)47002-3

Ramírez-Rodríguez, M., Arreguín-Sánchez, F., and Lluch-Belda, D. (2003). Recruitment patterns of the pink shrimp *Farfantepenaeus duorarum* in the southern Gulf of Mexico. *Fish. Res.* 65, 81–88. doi: 10.1016/j.fishres.2003.09.008

Robblee, M., Fry, B., Fourqurean, J. W., and Mumford, P. L. (1999). "Relationships among inshore sources of the pink shrimp, *Penaeus duorarum*, and the offshore tortugas and sanibel fisheries," in *US Geological Survey Program on the South Florida Ecosystem - Proceedings of the Technical Symposium*. Fort Lauderdale, FL, 94.

Roberts, T. W. (1986). Abundance and distribution of pink shrimp in and around the Tortugas Sanctuary, 1981-1983. *N. Am. J. Fish. Manag.* 6, 311–327. doi: 10.1577/1548-8659(1986)6<311:aadops>2.0.co;2

Rochette, N. C., and Catchen, J. M. (2017). Deriving genotypes from RAD-seq short-read data using Stacks. *Nat. Protoc.* 12, 2640–2659. doi: 10.1038/nprot.2017.123

Rothlisberg, P. C. (1982). Vertical migration and its effect on dispersal of penaeid shrimp larvae in the Gulf of Carpentaria. *Fishe. Bull.* 80, 541–554.

Rothlisberg, P. C., Church, J. A., and Fandry, C. B. (1995). A mechanism for near-shore concentration and estuarine recruitment of post-larval *Penaeus plebejus* Hess (*Decapoda. Penaeidae*). *Estuar. Coast. Shelf Sci.* 40, 115–138. doi: 10.1016/s0272-7714(05)80001-0

Rothlisberg, P. C., Craig, P. D., and Andrewartha, J. R. (1996). Modelling penaeid prawn larval advection in Albatross Bay, Australia: defining the effective spawning population. *Mar. Freshw. Res.* 47, 157–168. doi: 10.1071/mf9960157

Saloman, C. H., Allen, D. M., and Costello, T. J. (1968). Distribution of three species of shrimp (genus *Penaeus*) in waters contiguous to southern Florida. *Bull. Mar. Sci.* 18, 343–350.

Shanks, A. L. (2009). Pelagic larval duration and dispersal distance revisited. *Biol. Bull.* 216, 373–385. doi: 10.2307/25548167

Sheridan, P. (1996). Forecasting the fishery for pink shrimp, *Penaeus duorarum*, on the Tortugas grounds, Florida. *Fish. Bull.* 94, 745–755.

Tabb, D. C., Dubrow, D. L., and Jones, A. E. (1962). *Studies on the biology of the pink shrimp, Penaeus duorarum Burkenroad, in Everglades National Park, Florida*. Coral Gables, FL: University of Miami. State of Florida Board of Conservation Technical Series No. 37.

Teodoro, S. S. A., Terossi, M., Mantelatto, F. L., and Costa, R. C. (2016). Discordance in the identification of juvenile pink shrimp (*Farfantepenaeus brasiliensis* and *F. paulensis*: family *Penaeidae*): an integrative approach using morphology, morphometry and barcoding. *Fish. Res.* 183, 244–253. doi: 10.1016/j.fishres.2016.06.009

Therkildsen, N. O., Hemmer-Hansen, J., Hedeholm, R. B., Wisz, M. S., Pampoulie, C., Meldrup, D., et al. (2013). Spatiotemporal SNP analysis reveals pronounced biocomplexity at the northern range margin of Atlantic cod *Gadus morhua*. *Evol. Appl.* 6, 690–705. doi: 10.1111/eva.12055

Timm, L. E. (2019). *Raw ddRADseq Data in Fastq Format for Population Genomic Analysis of the Gulf Pink Shrimp (Farfantepenaeus duorarum) From 2007-02-17 to 2015-07-15*. Corpus Christi, TX: Texas A&M University – Corpus Christi, doi: 10.7266/n7-hhnq-kh83

Timm, L. E., Browder, J. A., Simon, S., Jackson, T. L., Zink, I. C., and Bracken-Grissom, H. D. (2019). A tree money grows on: the first inclusive molecular phylogeny of the economically important pink shrimp (*Decapoda*: *Farfantepenaeus*) reveals cryptic diversity. *Invertebr. Syst.* 33, 488–500. doi: 10.1071/IS18044

Timm, L. E., Isma, L. M., Johnston, M. W., and Bracken-Grissom, H. D. (2020). Comparative population genomics and biophysical modeling of shrimp migration in the Gulf of Mexico reveals current-mediated connectivity. *Front. Mar. Sci.* 7:19. doi: 10.3389/fmars.2020.00019

Truelove, N. K., Kough, A. S., Behringer, D. C., Paris, C. B., Box, S. J., Preziosi, R. F., et al. (2017). Biophysical connectivity explains population genetic structure in a highly dispersive marine species. *Coral Reefs* 36, 233–244. doi: 10.1007/s00338-016-1516-y

Upton, H. F., Hoar, P., and Upton, M. (1992). *The Gulf of Mexico Shrimp Fishery: Profile of a Valuable National Resource*. Washington, DC: Center for Marine Conservation.

Waples, R. S. (1998). Separating the wheat from the chaff: patterns of genetic differentiation in high gene flow species. *J. Hered.* 89, 438–450. doi: 10.1093/jhered/89.5.438

Waples, R. S., and Gaggiotti, O. (2006). What is a population? An empirical evaluation of some genetic methods for identifying the number of gene pools and their degree of connectivity. *Mol. Ecol.* 15, 1419–1439. doi: 10.1111/j.1365-294X.2006.02890.x

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis.* New York, NY: Springer-Verlag. Availble online at: https://ggplot2.tidyverse.org

Willing, E. M., Dreyer, C., and van Oosterhout, C. (2012). Estimates of genetic differentiation measured by Fst do not necessarily require large sample sizes when using many SNP markers. *PLoS One* 7:e42649. doi: 10.1371/journal.pone.0042649

Wright, S. (1950). Genetical structure of populations. *Nature* 166, 247–249. doi: 10.1038/166247a0

Young, A. M., Torres, C., Mack, J. E., and Cunningham, C. W. (2002). Morphological and genetic evidence for vicariance and refugium in Atlantic and Gulf of Mexico populations of the hermit crab *Pagurus longicarpus*. *Mar. Biol.* 140, 1059–1066. doi: 10.1007/s00227-002-0780-2

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

frontiers in Ecology and Evolution

ORIGINAL RESEARCH
published: 26 July 2021
doi: 10.3389/fevo.2021.678110

Phylogenomic Assessment of Biodiversity Using a Reference-Based Taxonomy: An Example With Horned Lizards (*Phrynosoma*)

Adam D. Leaché[1,2], Hayden R. Davis[1,2], Sonal Singhal[3], Matthew K. Fujita[4], Megan E. Lahti[5] and Kelly R. Zamudio[6]*

[1] Burke Museum of Natural History and Culture, Seattle, WA, United States, [2] Department of Biology, University of Washington, Seattle, WA, United States, [3] Department of Biology, California State University-Dominguez Hills, Carson, CA, United States, [4] Department of Biology, Amphibian and Reptile Diversity Research Center, The University of Texas at Arlington, Arlington, TX, United States, [5] Department of Biology, Truckee Meadows Community College, Reno, NV, United States, [6] Department of Ecology and Evolutionary Biology, Cornell University, Ithaca, NY, United States

OPEN ACCESS

Edited by:
Michael G. Campana, Smithsonian Conservation Biology Institute (SI), United States

Reviewed by:
Kevin De Queiroz, Smithsonian National Museum of Natural History (SI), United States
Susan Tsang, American Museum of Natural History, United States

***Correspondence:**
Adam D. Leaché
leache@uw.edu

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

Received: 08 March 2021
Accepted: 14 June 2021
Published: 26 July 2021

Citation:
Leaché AD, Davis HR, Singhal S, Fujita MK, Lahti ME and Zamudio KR (2021) Phylogenomic Assessment of Biodiversity Using a Reference-Based Taxonomy: An Example With Horned Lizards (Phrynosoma). Front. Ecol. Evol. 9:678110. doi: 10.3389/fevo.2021.678110

Phylogenomic investigations of biodiversity facilitate the detection of fine-scale population genetic structure and the demographic histories of species and populations. However, determining whether or not the genetic divergence measured among populations reflects species-level differentiation remains a central challenge in species delimitation. One potential solution is to compare genetic divergence between putative new species with other closely related species, sometimes referred to as a reference-based taxonomy. To be described as a new species, a population should be at least as divergent as other species. Here, we develop a reference-based taxonomy for Horned Lizards (*Phrynosoma*; 17 species) using phylogenomic data (ddRADseq data) to provide a framework for delimiting species in the Greater Short-horned Lizard species complex (*P. hernandesi*). Previous species delimitation studies of this species complex have produced conflicting results, with morphological data suggesting that *P. hernandesi* consists of five species, whereas mitochondrial DNA support anywhere from 1 to 10 + species. To help address this conflict, we first estimated a time-calibrated species tree for *P. hernandesi* and close relatives using SNP data. These results support the paraphyly of *P. hernandesi;* we recommend the recognition of two species to promote a taxonomy that is consistent with species monophyly. There is strong evidence for three populations within *P. hernandesi*, and demographic modeling and admixture analyses suggest that these populations are not reproductively isolated, which is consistent with previous morphological analyses that suggest hybridization could be common. Finally, we characterize the population-species boundary by quantifying levels of genetic divergence for all 18 *Phrynosoma* species. Genetic divergence measures for western and southern populations of *P. hernandesi* failed to exceed those of other *Phrynosoma*

species, but the relatively small population size estimated for the northern population causes it to appear as a relatively divergent species. These comparisons underscore the difficulties associated with putting a reference-based approach to species delimitation into practice. Nevertheless, the reference-based approach offers a promising framework for the consistent assessment of biodiversity within clades of organisms with similar life histories and ecological traits.

Keywords: multispecies coalescent, *Phrynosoma*, phylogeography, species delimitation, systematics, taxonomy comparative species delimitation

INTRODUCTION

One of the most difficult aspects of species delimitation is determining when genetic divergence is sufficient for the recognition of new species. Many methods have been developed to help determine the boundary between populations and species using genetic data (Yang and Rannala, 2010; Jones et al., 2015; Kapli et al., 2017; Smith and Carstens, 2020; Sukumaran et al., 2021), yet the question still remains whether or not the delimited units should be recognized as populations or species (Sukumaran and Knowles, 2017; Leaché et al., 2019). This is an important question, because as the ease of genomic data collection increases so does the resolution at which populations can be distinguished. This has the potential to lead to over-splitting species and artificially inflating biodiversity estimates (Carstens et al., 2013; Rannala, 2015).

One potential solution to this problem is to measure and compare the levels of genetic divergence for putative taxa to those observed among other closely related species (Sites and Marshall, 2003, 2004; Galtier, 2019). This reference-based taxonomic approach uses levels of divergence among species to define a potential shared boundary between population and species (Tobias et al., 2010). Comparing levels of genetic divergence using a reference-based taxonomy allows us to answer the question, "Are putative species more or less divergent compared to reference species?" If a clear population-species transition point is identified, then it could be used to establish a more effective and reliable "yardstick" for conducting quantitative taxonomic comparisons (Sukumaran et al., 2021). This approach requires a thorough understanding of a group's taxonomy so that existing biases are not perpetuated onto a revised taxonomy. Further, although low levels of genetic divergence may provide weak evidence in favor of the new species, other sources of data such as morphology and ecology could be integrated to strengthen the case for species identity (de Queiroz, 2007; Padial et al., 2010). In doing so, reference-based taxonomy builds on the existing data available for a species group and moves species delimitation into a comparative framework (Galtier, 2019).

Reference-based approaches are not a new idea (Mayr, 1969). Some DNA barcoding approaches routinely use heuristic cutoffs for species delimitation (i.e., thresholds of genetic divergence) based on levels of divergence among species (Hebert et al., 2004; Hebert and Gregory, 2005). However, these approaches are limited by the use of a single, idiosyncratic locus (typically mtDNA coding genes) and their requirement for reciprocal monophyly (Moritz and Cicero, 2004: Hickerson et al., 2006). A modern approach based on genome-wide data can overcome these limitations by incorporating multiple independent loci and a coalescent model to accommodate incomplete lineage sorting. Multilocus data and coalescent models provide a more thorough perspective on the genetic divergence and demographic history of populations and species (Yang and Rannala, 2017). Yet, like its predecessors, this genome-wide approach can still falter when there is introgression or hybridization (Jiao and Yang, 2021), or when different axes of divergence disagree (e.g., morphological vs. genetic).

Modernizing reference-based taxonomic approaches to leverage genomic data can provide an empirical perspective on how genetic divergence relates to the "speciation continuum" (Chan and Grismer, 2019; Poelstra et al., 2021). A reference-based taxonomy could use any number of genetic diversity measures ranging from pairwise genetic distances to more sophisticated coalescent-based metrics. An advantage of coalescent units is that they provide an expectation for the amount of genealogical discordance produced by different combinations of species tree branch lengths and population sizes (Pamilo and Nei, 1988). One such coalescent-based metric is the genealogical divergence index (*gdi;* Jackson et al., 2017). The *gdi* measures genetic divergence between two populations, reflecting the combined effects of genetic isolation and gene flow (Jackson et al., 2017). Higher *gdi* values indicate that populations are more evolutionarily independent and can be used as evidence to distinguish between populations and species.

Here, we use genomic data (ddRADseq) to estimate genetic divergence among species to develop a reference-based taxonomy for Horned Lizards (*Phrynosoma*) to conduct comparative species delimitation within the Greater Short-horned Lizard species complex (*P. hernandesi*). A previous phylogeographic study of *P. hernandesi* using mitochondrial DNA (mtDNA) identified three major clades with relatively strong geographical structure (**Figure 1**). These mtDNA clades did not correspond to existing subspecies boundaries defined by morphology, precluding their recognition as species (Zamudio et al., 1997). A systematic study of *P. hernandesi* based on morphometric analyses of morphological traits recommended the recognition of five species (two of which contained two subspecies; Montanucci, 2015; **Figure 1**). The morphological study provided indirect evidence of gene flow and identified large geographic regions of putative hybridization (Montanucci, 2015). A subsequent species delimitation analysis of mtDNA data supported anywhere from 1 to 10 species, and although the validity of the morphological species were questioned, no formal taxonomic

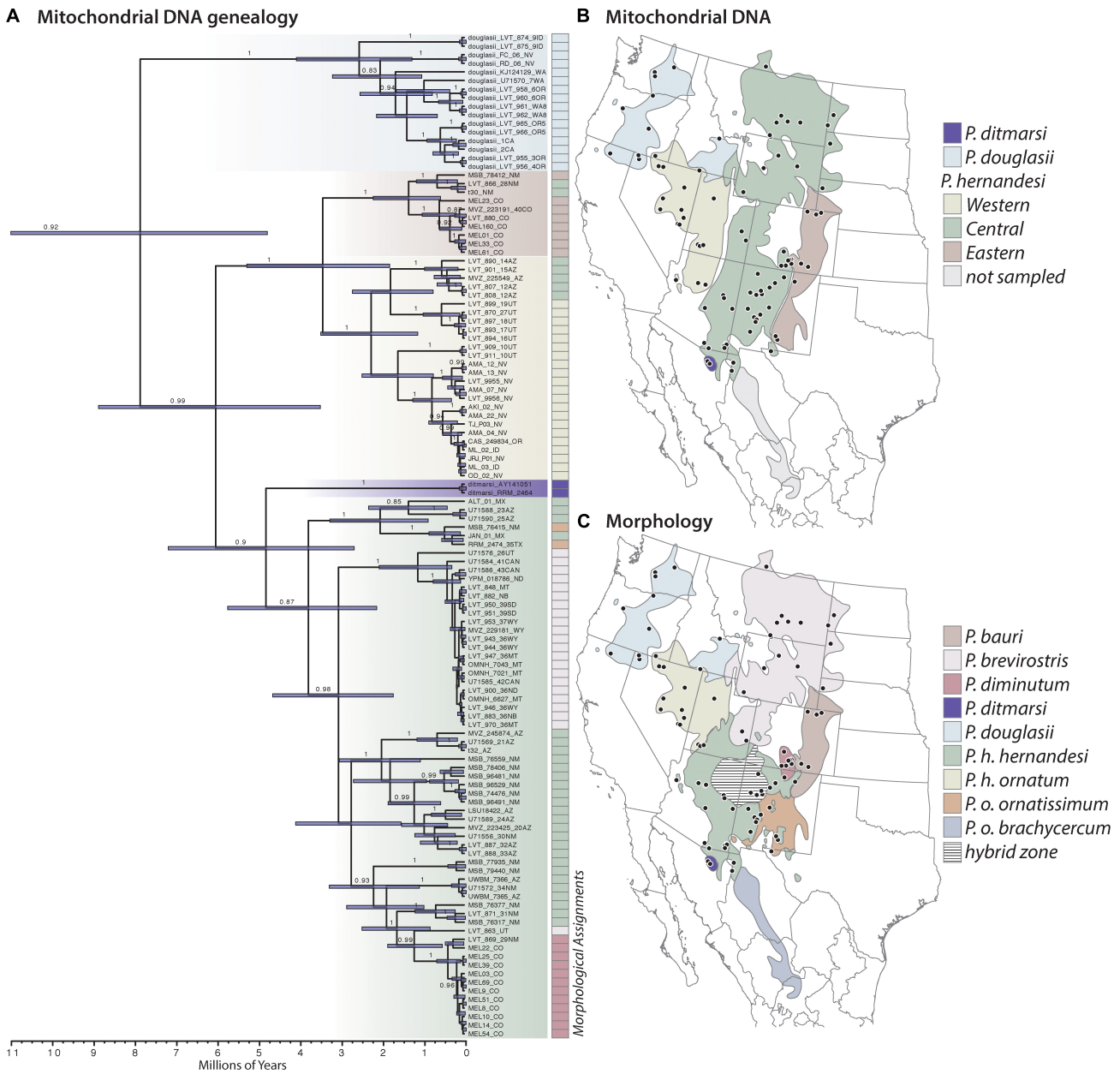

FIGURE 1 | Summary of previous systematic studies of *Phrynosoma hernandesi* and closely related species belonging to the *Tapaja* clade, which includes *P. ditmarsi, P. douglasii, P. hernandesi,* and *P. orbiculare* (not shown). **(A)** Mitochondrial DNA genealogy estimated with Bayesian inference (posterior probability values ≥ 80% are shown). The tree was calibrated using a secondary fossil calibration information on the age of *Tapaja* (12.7 Ma; 95% CI = 10.8–14.7). Node bars show divergence time estimation uncertainty. The genealogy is color-coded to illustrate the species-level taxonomy and phylogeographic groups within *P. hernandesi* supported by the mtDNA genealogy and described by Zamudio et al. (1997). **(B)** Geographic distributions of the mtDNA clades within *P. hernandesi* (modified from Zamudio et al., 1997). Unsampled area of taxonomic importance (*P. o. brachycercum*) is shown in gray. **(C)** Geographic distributions of species and subspecies based on morphological delimitations (modified from Montanucci, 2015). The hatched area in the Colorado Plateau is one of several putative hybrid zones hypothesized to occur between species. Mapping the morphological taxonomy onto the mtDNA genealogy illustrates their discordances.

recommendations were made (Blair and Bryson, 2017). Because mtDNA and morphological species assignments conflict and there is evidence of hybridization, we collected multilocus nuclear data to investigate phylogeny, phylogeography, demography, and species delimitation in the *P. hernandesi* species complex. First, we characterize population structure and phylogeny in the *P. hernandesi* species complex and three other closely related *Phrynosoma* species. We then use coalescent models to infer the demographic history of *P. hernandesi* populations. Finally, we analyze patterns of genetic divergence among all *Phrynosoma* species to develop a reference-based taxonomy and to delimit *P. hernandesi* populations.

MATERIALS AND METHODS

Ethics Statement

Tissue samples were obtained as loans from natural history museum collections. We also included mtDNA sequence data from *Phrynosoma hernandesi* that were used in a previous phylogeographic study (Zamudio et al., 1997) and an unpublished dissertation (Lahti, 2010). All animal research protocols presented in this study were approved by the University of Washington Institutional Animal Care and Use Committee (UW IACUC #4367–03).

Taxon Sampling

Analyses were conducted at three different levels: (1) phylogeny of the *P. hernandesi* species complex and other closely related species in the *Tapaja* clade, (2) phylogeographic and demographic history of *P. hernandesi* populations, and (3) genetic divergence comparisons among *Phrynosoma* to develop a reference-based taxonomy for delimiting the *P. hernandesi* populations.

Phylogeny of *Tapaja*

These analyses focused on estimating the phylogenetic relationships within and among species of *Tapaja,* which is the name referring to the crown clade originating in the last common ancestor of *P. ditmarsi, P. douglasii, P. hernandesi,* and *P. orbiculare* (Leaché and McGuire, 2006). Molecular phylogenetic studies provide strong evidence for the monophyly of *Tapaja* (Leaché and McGuire, 2006; Leaché et al., 2015; Leaché and Linkem, 2015). The species within *Tapaja* share several life history and morphological characteristics including viviparity (give birth to live young) and short to extremely reduced cranial horns. Separate phylogenetic analyses were conducted with mtDNA and nuclear data. The mtDNA dataset included 118 samples (**Supplementary Table 1**): *P. ditmarsi* ($n = 2$), *P. douglasii* ($n = 16$), *P. hernandesi* ($n = 99$), and *P. orbiculare* ($n = 1$). The ddRADseq dataset included 118 samples (**Supplementary Table 2**): *P. ditmarsi* ($n = 3$), *P. douglasii* ($n = 17$), *P. hernandesi* ($n = 94$), and *P. orbiculare* ($n = 4$).

Phylogeographic and Demographic History of *P. hernandesi*

To investigate the population structure and demography of *P. hernandesi*, we conducted focused analyses of the ddRADseq data on range-wide *P. hernandesi* samples (90 samples from 73 unique locations) from across Western and Central North America. These analyses excluded four samples belonging to an early diverging lineage containing four samples that cause *P. hernandesi* to be paraphyletic with respect to *P. douglasii*.

Reference-Based Taxonomy

The final taxon sampling set was used to establish a reference-based taxonomy for *Phrynosoma*, and included multiple samples for all 17 species in the genus (**Table 1**). A total of 83 samples were included with 24 of the samples representing the *P. hernandesi* species complex (**Supplementary Table 3**).

TABLE 1 | Species included in the reference-based taxonomic analysis of *Phrynosoma.*

Species	Samples
P. asio	4
P. blainvillii	4
P. braconnieri	4
P. cerroense	4
P. cornutum	4
P. coronatum	4
P. ditmarsi	3
P. douglasii	4
P. goodei	4
P. hernandesi	20
P. "hernandesi"	4
P. mcallii	4
P. modestum	2
P. orbiculare	4
P. platyrhinos	3
P. sherbrookei	4
P. solare	3
P. taurus	4

Voucher specimen information is provided in ***Supplementary Table 3****.*

Molecular Methods

Genomic DNA was extracted from fresh tissue samples using QIAGEN DNeasy extraction kits (QIAGEN Inc.). We collected mtDNA sequence data from the *ND4* gene to build on the existing mtDNA genealogy (**Figure 1**; Zamudio et al., 1997). We followed standard PCR amplification and sequencing protocols with primers used in a previous *Phrynosoma* study (Leaché and McGuire, 2006). To obtain multilocus nuclear data, we collected ddRADseq data following the protocol described by Peterson et al. (2012) using a slightly modified protocol with the restriction enzymes *Sbf* I and *Msp*I (Leaché et al., 2015). Short sequence reads (51 base pairs) were obtained using single-end sequencing with an Illumina HiSeq 4,000 at the QB3 facility at UC Berkeley.

Bioinformatics

For the mtDNA data, we edited and aligned the raw *ND4* sequences using Geneious (Kearse et al., 2012). The *ND4* protein-coding gene contained no indels making alignment with existing sequences trivial. For the ddRADseq data, we processed raw Illumina reads using the program iPyRAD v.0.7.30 (Eaton and Overcast, 2020). We de-multiplexed samples using their unique barcode and adapter sequences, and sites with Phred quality scores under 99.95% (Phred score = 33) were changed into "N" characters and reads with ≥10% N's were discarded. The filtered reads were clustered using a threshold of 90%. Consensus sequences that had low coverage (<6 reads), excessive undetermined or heterozygous sites (>5), or too many haplotypes (>2 for diploids) were discarded. We removed putative paralogs by filtering out loci with excessive shared heterozygosity among samples (paralog filter = 0.5). We then assembled separate datasets for each of the three taxon sampling sets to minimize the amount of missing data. For each dataset,

we controlled levels of missing data by adjusting the minimum individual (min. ind.) value, which specifies the minimum number of individuals that are required to have data present at a locus for that locus to be included in the final matrix. Details on the levels of missing data for each assembly are provided in the relevant methods sections below.

Phylogeny of *Tapaja*

The mitochondrial *ND4* data were analyzed using BEAST v2.6.4 (Bouckaert et al., 2019). We used the GTR nucleotide substitution model with gamma distributed rate variation (five categories), following previous studies of *P. hernandesi* using the same locus (Zamudio et al., 1997; Blair and Bryson, 2017). Time calibration was accomplished with a relaxed log normal clock model calibrated using a secondary fossil calibration information from a phylogenomic analysis of *Phrynosoma* that estimated the crown age of *Tapaja* at 12.7 Ma (Leaché and Linkem, 2015). We implemented a normal distribution with a mean = 12.7 Ma on the age of *Tapaja* with a 95% confidence interval of 10.8–14.7 Ma to accommodate divergence time estimation errors. We conducted two replicate analyses (10 million generations each) and assessed convergence by comparing posterior distributions of parameters and checking for high ESS values (>200). The posterior distributions were combined using LogCombiner, and a maximum clade credibility (MCC) tree was summarized using TreeAnnotator after discarding the first 20% of samples as burn-in.

The ddRADseq data were assembled with a maximum of 15% missing data at a locus (min.ind. = 100). To identify genetic structure within and among species, we used Adegenet (Jombart, 2008) to conduct a principal component analysis (PCA) using all variable sites from across all loci. The genetic clusters identified by PCA were used in the subsequent species tree analysis. PCA does not make any assumptions about the underlying population genetic model, making it a useful approach for visualizing genetic differences among populations and species.

The concatenated ddRADseq data were analyzed using ML with RAxML (Stamatakis, 2014) using the GTRGAMMA substitution model and 100 bootstrap replicates. To determine phylogenetic relationships among the genetic clusters identified in the PCA, we estimated a time-calibrated species tree from the unlinked and biallelic SNPs using the multispecies coalescent model in the program SNAPP v1.5.0 (Bryant et al., 2012) implemented in BEAST v2.5.2 (Bouckaert et al., 2019). Divergence-time estimation was accomplished with a strict clock model calibrated using secondary fossil calibration information from a phylogenomic analysis of *Phrynosoma* that estimated the crown age of *Tapaja* at 12.7 Ma (Leaché and Linkem, 2015). We implemented a normal distribution with a mean = 12.7 Ma on the divergence of *Tapaja* with a 95% confidence interval of 10.8–14.7 Ma to accommodate divergence time estimation errors. The input files were generated using methods described by Stange et al. (2018) using the snapp_prep.rb scripts available on GitHub[1]. To reduce computational time, the number of samples included for *P. douglasii* was reduced to eight (one sample from each unique sampling locality), and the number of *P. hernandesi* samples was reduced to 12 (**Supplementary Table 4**). Two independent analyses were run for 200,000 generations each, sampling every 50 generations. The posterior distributions were combined using LogCombiner, and a MCC tree was summarized using TreeAnnotator after discarding the first 20% of samples as burn-in.

Phylogeographic and Demographic History of *P. hernandesi*

Given the conflict between mitochondrial and morphological species delimitations for *P. hernandesi*, we conducted a focused exploration of the phylogeography and population demographics of this species. Our phylogenetic analysis of the *Tapaja* clade revealed that *P. hernandesi* is paraphyletic with respect to *P. douglasii,* with an early diverging lineage containing four samples from three relatively low elevation locations in the southern portion of the range in New Mexico, Texas, and Mexico (**Figure 2**). We excluded this early diverging lineage of *P.* "*hernandesi*" and *P. douglasii* from subsequent phylogeographic and demographic analyses and focused on the remaining 90 samples of *P. hernandesi.* The SNP data assembly allowed a maximum of 50% missing data at a locus (min. ind. = 45).

Population structure was estimated using the maximum likelihood method ADMIXTURE v1.3.0 (Alexander et al., 2009) to determine the optimal number of populations (K) and admixture proportions of samples. This analysis is necessary for identifying putative hybrids with mixed population ancestry; previous morphological data indicate that taxa in the *P. hernandesi* species complex hybridize (Montanucci, 2015). Samples were considered admixed with assignment probabilities $\leq$0.90. To determine the best-fit model, we compared analyses for K = 1 through K = 10 and selected the analysis that minimized group assignment error; e.g., the K with the lowest cross-validation error was considered the best-fit model. The analyses were repeated 10 times to measure uncertainty in cross-validation error estimation. After selecting the K value with the lowest cross-validation error, the 10 replicate runs were combined to summarize the admixture proportions for each sample.

Phylogeographic studies often present intraspecific genealogical relationships among samples, but in the context of nuclear loci that segregate independently the concept of a single bifurcating tree relating all samples is misleading. Network methods can depict relationships that are not necessarily bifurcating and can also help identify admixed samples (Blair and Ané, 2020). A genetic network was constructed from the concatenated SNP data (uncorrected "p" distances; all constant and variable sites were included) using the NeighborNet algorithm (Bryant and Moulton, 2004) in SplitsTree v4.6 (Huson and Bryant, 2006).

We estimated the phylogenetic relationships among populations using SNAPP using the population assignments estimated from the top-ranked ADMIXTURE model. We limited the number of samples assigned to each population to reduce computational times (**Supplementary Table 5**). An

[1]https://github.com/mmatschiner/snapp_prep

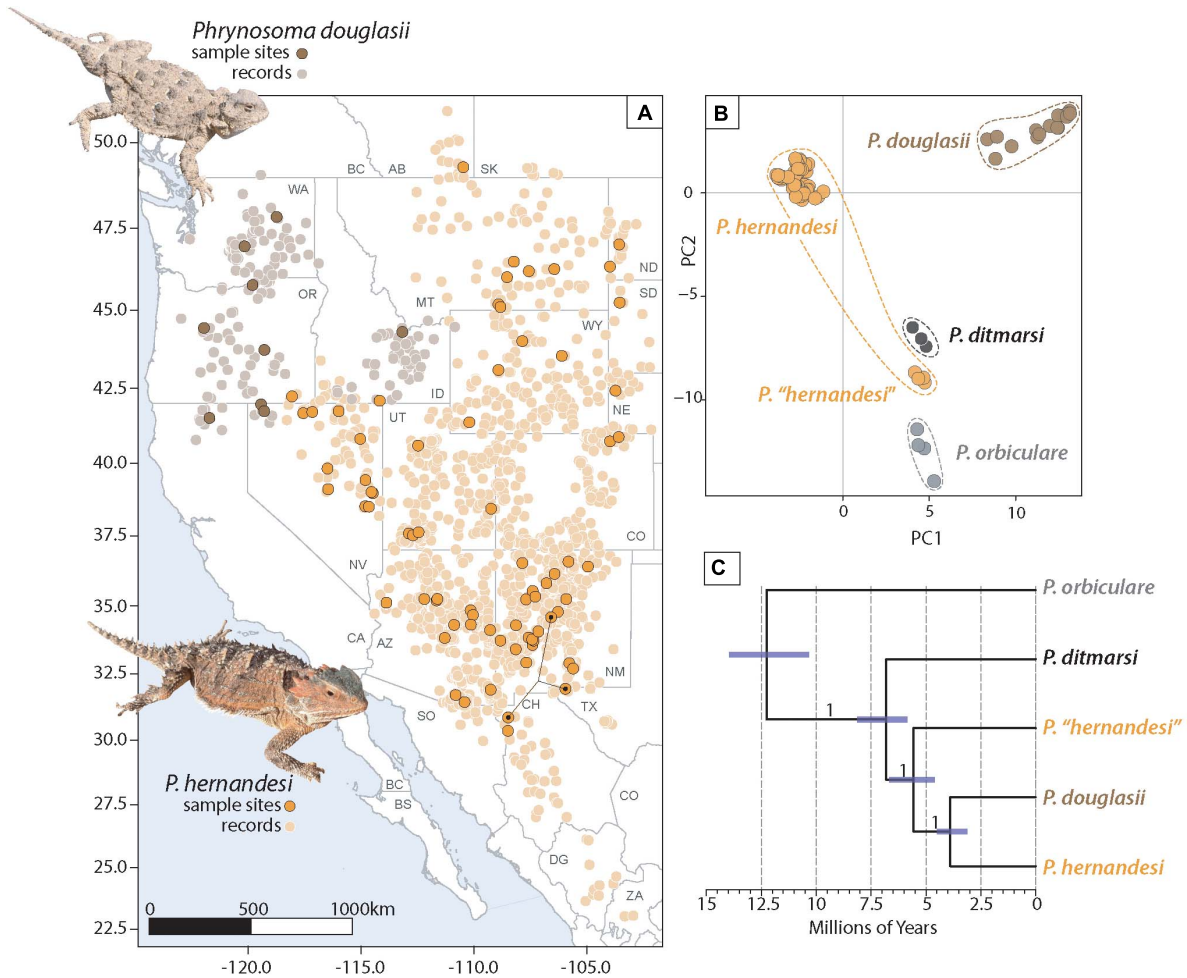

FIGURE 2 | Analyses of multilocus nuclear data focusing on *Phrynosoma douglasii* and *P. hernandesi*. **(A)** Geographic distribution of *P. douglasii* and *P. hernandesi* in Western and Central North America. Distributional data are based on published locality records and distribution maps (Montanucci, 2015) and augmented with records downloaded from VertNet.org and iNaturalist. Detailed information on sample sites included in the phylogeographic component of this study are provided in **Supplementary Table 2**. **(B)** Principal components analysis of genetic variation (5,715 biallelic SNPs) for the species belonging to the *Tapaja* clade of *Phrynosoma*. The 94 *P. hernandesi* samples form two genetic clusters; the majority of the samples belong to *P. hernandesi* (*n* = 90), while *P.* "*hernandesi*" only includes four samples from the southern portion of the range with localities in New Mexico, Texas, and Mexico (shown on map with connected black dots). **(C)** A time-calibrated species tree analysis using the biallelic SNP data supports the paraphyly of *P. hernandesi* with respect to *P. douglasii*. Numbers on nodes are posterior probability values.

estimate of the nuclear genome-wide substitution rates for lizards (7.7×10^{-10} substitutions per site per year; Perry et al., 2018) was used to convert branch length estimates to absolute time. This is a strong assumption that directly influences the divergence dates being estimated. We compared the divergence times estimated for *P. hernandesi* between this analysis, which assumes a substitution rate calibration, to the estimate obtained independently using a divergence time prior in the species tree analysis of *Tapaja*. Two independent analyses were run for 200,000 generations each, sampling every 50 generations. The posterior distributions were combined using LogCombiner, and an MCC tree was summarized using TreeAnnotator after discarding the first 20% of samples as burn-in.

We compared three demographic models to better understand the history of gene flow among populations of *P. hernandesi* (**Supplementary Figure 1**). In particular, we tested for gene flow and secondary contact during divergence and additionally estimated divergence times (τ), population sizes (θ), the amounts and directions of gene flow (scaled by population size—Nm), and timing of secondary contact. The first model was a simple isolation model with no gene flow during divergence. The second model was a standard isolation-migration model (IM) that allowed gene flow among all contemporary and ancestral populations. The final model, the secondary contact model (SC), allowed for gene flow after an initial period of divergence in isolation. We fit these models to a phylogeny for the three *P. hernandesi* populations [north, (south, west)] using fastsimcoal2 (Excoffier et al., 2013), which can model multiple populations using simulations under the joint site frequency spectra (JSFS). JSFS were made from unlinked SNPs sampled

from a VCF file using easySFS[2]. The full data were projected down to a smaller number of chromosomes per population to account for missing data and maximize the number of segregating sites in the JSFS (**Supplementary Table 6**). Parameters were converted to demographic units using the same mutation rate assumptions as the species tree analysis (mutation rate of 7.7×10^{-10} substitutions per site per generation). Models were optimized using 10 replicate searches (100,000 simulations each). The best-fit run from each of 10 replicates was ranked using the Akaike information criterion (AIC), and Akaike weights were used as a measure of statistical confidence of the top-ranked model. Finally, uncertainty in the point estimates for parameters of the best-fit model were obtained by non-parametric bootstrapping. Unlinked SNPs in the VCF file were sampled with replacement (50 replicates), and each bootstrap dataset was optimized in fastsimcoal2 with 10,000 simulations.

Reference-Based Taxonomy

To generate a reference-based taxonomy for *Phrynosoma*, we calculated levels of genetic divergence across all species in the clade. A total of 83 samples of *Phrynosoma* were included for the reference-based taxonomic analysis (**Supplementary Table 3**). Multiple samples were included for all 17 species of *Phrynosoma* (**Table 1**) with the addition of multiple samples for the *P. hernandesi* species complex. The genetic divergence values for the *P. hernandesi* species complex were compared to the values calculated for all other *Phrynosoma* species. The final SNP data assembly allowed a maximum of 40% missing data at a locus (min.ind. = 52).

All sites from the phased alleles (variable and constant sites) were used to calculate four measures of genetic diversity. For the first two measures, we calculated F_{ST} and d_{xy} among all population-pairs (Nei, 1987; Reich et al., 2009). Our third and fourth measures were *gdi* values and the population divergence times in coalescent units ($2\tau/\theta$) for population and species using the multispecies coalescent model implemented in BPP v4.3.0 (Flouri et al., 2018). The species tree topology was fixed to match a previous species tree estimated for *Phrynosoma* from phylogenomic data (Leaché and Linkem, 2015). The phylogenetic relationships within *Tapaja* were updated to reflect the results of this study at both the species-level and for populations within *P. hernandesi* (**Figures 2**, **3**). Posterior probability distributions for τ and θ were estimated with BPP using analysis A00 (Yang, 2015). Two replicate runs were conducted and compared to check for convergence, with each analysis sampling 200,000 steps (sample frequency = 2) after a burnin period of 100,000 steps. The priors were set for $\theta \sim$inversegamma (3, 0.01) and $\tau \sim$inversegamma (3, 0.04), which provide mean values of 0.005 and 0.02, respectively. We calculated population divergence times in coalescent units ($2\tau/\theta$) for each species and population using all samples from the combined posterior distributions. We calculated *gdi* for each species using equation $gdi = 1 - e^{-2\tau/\theta}$ (Leaché et al., 2019). Although the *gdi* can measure the combined effects of genetic isolation and gene flow (Jackson et al., 2017), we analyze the data under a multispecies coalescent model assuming no gene flow, which

[2]https://github.com/isaacovercast/easySFS

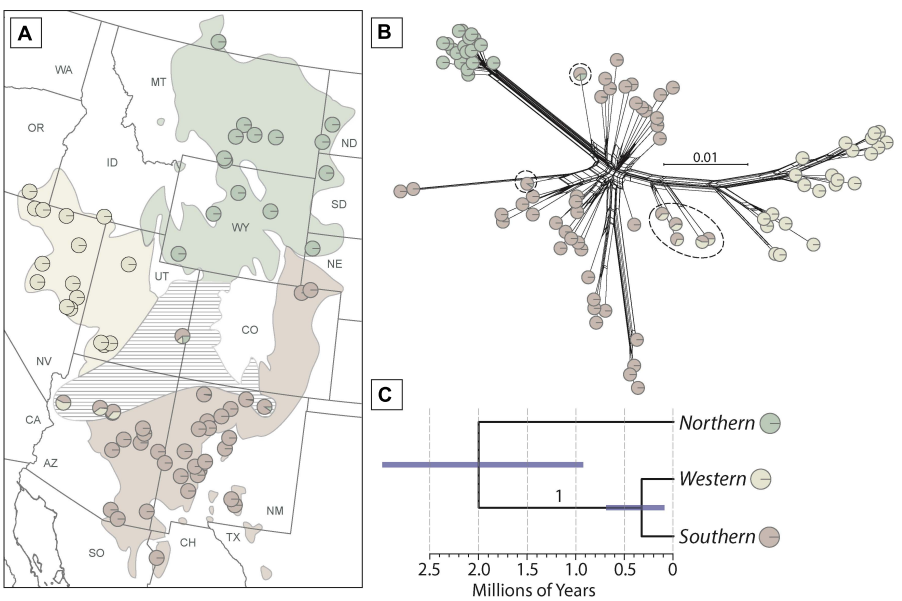

FIGURE 3 | Phylogeographic analysis supports at least three populations within *Phrynosoma hernandesi*. **(A)** Population structure analysis with the optimal $K = 3$ model estimated with ADMIXTURE. Each sample is represented by a pie-chart with colors indicating ancestry proportions. Hatching indicates an area where admixed samples are found. **(B)** Network analysis inferred from the program SplitsTree with admixture proportions assigned to each sample. Admixed samples are encircled with dashed lines. **(C)** A species tree analysis inferred from the unlinked and biallelic SNP data for the three populations. The posterior probability value supporting the west and south populations as a clade is 1.0. Branch lengths (error bars are 95% highest posterior densities) were converted to absolute time assuming a rate of 7.7×10^{-10} substitutions per site per year (see "Materials and Methods" section for details).

has been shown to provide accurate species delimitations using computer simulation (Leaché et al., 2019). The *gdi* is continuous between 0 (panmixia) and 1 (strong divergence from the sister group), and thus can indicate where a population lies on the path to speciation. Although there is no fixed "delimitation cutoff" between populations and species, Jackson et al. (2017) suggested that *gdi* < 0.2 = single species, *gdi* >0.7 = different species, and a broad range of intermediate values represent ambiguous delimitation results.

RESULTS

Phylogeny of *Tapaja*

The final alignment of the mtDNA data (*ND4*) included 118 sequences and 851 base pairs. The mtDNA gene tree estimated using Bayesian inference provides strong support for a sister relationship between *P. douglasii* (monophyletic) and *P.* "*hernandesi*" (paraphyletic with respect to *P. ditmarsi*; **Figure 1**). The phylogenetic patterns within *P. hernandesi* match those from previous studies (Zamudio et al., 1997; Blair and Bryson, 2017), most notably the support for three clades, which we refer to as the western, central, and eastern clades. The western clade includes localities in Oregon, Idaho, Nevada, Utah, and Arizona. The central clade includes localities in the Colorado Plateau, Wyoming Basins, and the Northern Great Plains. The eastern clade is primarily in the eastern piedmont (foothills) of the Rockies in New Mexico and Colorado. Mapping the morphological delimitations onto the mtDNA genealogy provides weak evidence in support of the morphological species, which are not monophyletic, and indicates that instances of conflict involve samples from the geographic boundaries between populations/species (**Figure 1**).

The PCA analysis of 5,715 biallelic SNPs (**Figure 2**) supports five clusters corresponding to (1) *P. douglasii,* (2) *P. ditmarsi,* (3) *P. orbiculare,* (4) *P. hernandesi*, and (5) *P.* "*hernandesi*" The four samples grouping in *P.* "*hernandesi*" are from locations at relatively low elevations in the Rio Grande River Valley in the southern portion of the species range (Texas, New Mexico, Chihuahua, MX). Samples from nearby locations are from relatively higher elevations and are grouped with *P. hernandesi*.

The phylogenetic analysis of the concatenated ddRADseq data (118 samples and 52,171 base pairs) supports the monophyly of *P. douglasii* and the paraphyly of *P. hernandesi,* which is divided into two separate lineages (**Supplementary Figure 2**). One lineage contains the four samples representing *P.* "*hernandesi*" and is placed sister to *P. ditmarsi* with weak bootstrap support (52%; **Supplementary Figure 2**). This clade (*P. ditmarsi* + *P.* "*hernandesi*") is sister to a clade containing *P. douglasii* and the remaining 90 samples of *P. hernandesi*.

The time-calibrated species tree estimated with 1,321 unlinked and biallelic SNPs using SNAPP is strongly supported with posterior probability values of 1.0 for each clade (**Figure 2**). The species tree is asymmetric (**Figure 2**) with a root age for *Tapaja* of 12.3 mya (95% HPD = 10.3–14.0 mya), followed by the divergence of *P. ditmarsi* at 6.8 mya (95% HPD = 5.8–8.1 mya), then the divergence of *P.* "*hernandesi*" at 5.6 mya (95% HPD = 4.6–6.7 mya), and finally the divergence between *P. hernandesi* and *P. douglasii* at 3.9 mya (95% HPD = 3.1–4.5).

Phylogeographic and Demographic History of *P. hernandesi*

Population structure analysis of *P. hernandesi* (excluding the four low elevation *P.* "*hernandesi*" samples) with ADMIXTURE using 90 samples and 5,823 unlinked SNPs (sampled from 6,531 loci) supports $K = 3$ as best-fit population model according to cross-validation scores, and this result is supported across all 10 replicate analyses (**Supplementary Figure 3**). The three phylogeographic groups are partitioned into northern, western and southern populations, and samples with mixed ancestry are located at the geographic boundaries between populations (**Figure 3**). Three different geographic regions contain admixed samples, including (1) northern Arizona between the west and south populations, (2) northern New Mexico between northern and southern populations, and (3) eastern Utah with evidence of admixture among all three populations. The samples belonging to the western population are relatively congruent with the western mtDNA clade; however, the geographic distributions of the southern and northern populations are discordant with respect to mtDNA groups (**Figures 1**, **3** and **Supplementary Figures 5,6**).

The genetic network analysis (90 samples and 261,618 base pairs) shows similar clustering into three populations (**Figure 3**). Genetic diversity (as represented by clustering of samples) is greatest in the southern population, followed by the western population, and the lowest level in the northern population. Admixed samples (as estimated by the ADMIXTURE analysis) are placed in positions intermediate to these three populations in the genetic network (**Figure 3**).

The species tree analysis using 20 samples and 4,949 unlinked and biallelic SNPs (**Figure 3**) supports a close relationship between the west and south populations (posterior probability = 1.0) with a shallow divergence time of 324 kya (95% HPD = 90–649 kya). The estimated divergence at the root of the tree between the northern population and the remaining samples is 2 mya with a broad confidence interval (95% HPD) from 900 kya to 4 mya (**Figure 3**).

Demographic modeling strongly supported the secondary contact model as the best-fit model with a weighted AIC score of 1.0, followed by the IM model (**Table 2**). Divergence time is estimated at 313 kya (63–453 kya), which is younger than the phylogenetic estimate from the SNAPP analysis (**Figure 3**). The divergence time between the west and south populations is 67 kya (32–91 kya), and the timing of secondary contact is 2,267 kya (561–10,347 kya; **Table 3**). Migration rates are highest from the south to north (1.808 migrants per generation) and south to west (1.427 migrants per generation), and also from north to the common ancestor of west + south (1.299 migrants per generation).

Reference-Based Taxonomy

The ddRADseq dataset used for the reference-based taxonomy contained 83 samples (partitioned into 17 species; **Table 1**),

TABLE 2 | Demographic model selection results for the north, south and west populations of *Phrynosoma hernandesi*.

Demographic model	LL	*K*	AIC	ΔAIC	wAIC
Secondary contact	–4015.64	16	8063.28	0.00	1.00
Isolation-migration	–4024.87	15	8079.75	16.46	0.00
Isolation	–4061.83	7	8137.65	74.37	0.00

LL, log likelihood; K, model parameters; AIC, Akaike information criterion; wAIC, AIC weights.
Visual model descriptions are provided in ***Supplementary Figure 1****. The Akaike information criterion (AIC) was used to rank the models and identify the best-fit model.*

TABLE 3 | Demographic parameter estimates for *Phrynosoma hernandesi* populations (north, west, south) under the secondary contact model.

Parameter	Point estimate (95% CI)
N_POP_{north}	25,063 (12,865–34,367)
N_POP_{west}	98,391 (48,858–157,332)
N_POP_{south}	933,165 (475,131–1,291,854)
$N\_ANC_{west+south}$	253,446 (41,522–443,199)
$N\_ANC_{north+west+south}$	23,699 (15,806–209,131)
TDIV_SC	2,267 (561–10,347)
$TDIV_{west+south}$	67,361 (32,225–91,885)
$TDIV_{north+west+south}$	313,335 (63,816–453,024)
$NM_{north \rightarrow west}$	0.009 (0.000–0.051)
$NM_{west \rightarrow north}$	0.005 (0.000–0.127)
$NM_{north \rightarrow south}$	0.059 (0.000–0.138)
$NM_{south \rightarrow north}$	1.808 (0.328–7.271)
$NM_{west \rightarrow south}$	0.968 (0.229–4.146)
$NM_{south \rightarrow west}$	1.427 (0.000–5.052)
$NM_{north \rightarrow west+south}$	1.299 (0.004–15.109)
$NM_{west+south \rightarrow north}$	0.272 (0.000–7.158)

All estimates assume diploid genomes, a 1-year generation time, and a nuclear mutation rate of 7.7×10^{-10} *(Perry et al., 2018). Point estimates are from the best-fit run of the 100 model selection replicates. The 95% confidence intervals were calculated using 50 bootstrap replicates (sampling with replacement) of the unlinked SNP variant call file. Parameter codes: N_POP (contemporary population size), N_ANC (ancestral population size), TDIV_SC (secondary contact time), TDIV (divergence time), NMij (migration estimates, the number of migrants entering population i from population j going backwards in time).*

and the concatenated ddRADseq data contained 35,677 base pairs for 909 loci. Analysis of 500 loci in BPP on the fixed species tree (**Figure 4**) provided estimates for population sizes (θ) and divergence times (τ) used to calculate genetic divergence values *gdi* and coalescent units (**Supplementary Table 7**). Values for *gdi* ranged from a low of 0.2 to nearly 1.0 for species of *Phrynosoma* (**Figure 4**). Species with the lowest values of *gdi* included *P. blainvillii, P. cerroense, P. taurus, P. goodei,* and *P. platyrhinos.* The remaining species had relatively higher *gdi* values >0.8 (**Figure 4**). In comparison, values of *gdi* for *P.* "*hernandesi*" were high (>0.9) and exceeded values for at least 10 other species (**Figure 4**). The *gdi* values for the three populations of *P. hernandesi* were mixed with low values (<0.3) for the south and west populations and high (>0.9) for the north population (**Figure 4**). Comparison of coalescent units ($2\tau/\theta$) produced similar patterns (**Figure 4**). Comparisons of the *P. hernandesi* species complex using F_{ST} and d_{xy} show lower overall levels of genetic divergence compared to nominal species-pairs (**Figure 5**). Again, *P.* "*hernandesi*" is relatively divergent compared to other species of *Phrynosoma* (**Figure 5**).

DISCUSSION

Systematics of the *Phrynosoma hernandesi* Species Complex

Using genomic data in a comparative taxonomic framework, this study resolves relationships within *Phrynosoma* and builds on previous studies using mitochondrial DNA (Zamudio et al., 1997) and morphological (Montanucci, 2015) data. The multilocus nuclear data support *P. hernandesi* being composed of at least two species. One of these species, *P. hernandesi* sensu stricto (referred to as *P. hernandesi*; **Figure 2**), has a broad distribution and contains at least three populations that diverged from *P. douglasii* approximately 3.9 mya. The other species, which up until now we have referred to as *P.* "*hernandesi*" (**Figure 2**) diverged earlier at approximately 5.6 mya and has a relatively restricted distribution in the southern portion of the range (**Figure 2**). This study supports the morphology-based taxonomy that described this divergent lineage as *P. ornatissimum* (Montanucci, 2015). Recognizing *P. hernandesi* and *P. ornatissimum* as two independent evolutionary lineages (= species) follows the general lineage concept of species (de Queiroz, 1998). Our phylogenetic analysis suggests that these lineages are distinct and divergent in relation to other species of *Phrynosoma*.

Phrynosoma ornatissimum was originally described by Girard (1858), and the current type locality is restricted to "Rio Grande Valley at Albuquerque, Bernalillo County, New Mexico" (Montanucci, 2015). This species has a unique combination of morphological characteristics, including but not limited to a relatively short tail, truncate snout, and large rounded dorsal spots with light-colored borders (Montanucci, 2015). *Phrynosoma ornatissimum* occurs at relatively low elevations (1,436 m–2,134 m) and primarily in arid short-grass plains of southern New Mexico, Texas, and northern Mexico (Montanucci, 2015). *Phrynosoma ornatissimum* is replaced by *P. hernandesi* at higher elevations (1,916–3,475 m) where juniper-pinyon woodland habitats dominate (Montanucci, 2015). The replacement of these species along elevation and habitat gradients results in a peculiar distributional pattern where montane populations of *P. hernandesi* are surrounded by *P. ornatissimum* occupying the adjacent short-grass plains (**Supplementary Figure 4**). If the isolated montane populations of *P. hernandesi* are reproductively isolated from one another, then it is possible that they could represent independent evolutionary lineages. Morphological data indicate that *P. hernandesi* and *P. ornatissimum* hybridize at habitat ecotones, but we found no evidence for admixture between these species based on the small number of *P. ornatissimum* samples included in our analyses. Interestingly, according to the mtDNA genealogy, the samples assigned to *P. ornatissimum* are nested within *P. hernandesi* (**Figure 1**), suggesting that mitochondrial introgression may have occurred at some point, or, that the genealogical discordance could be a consequence of

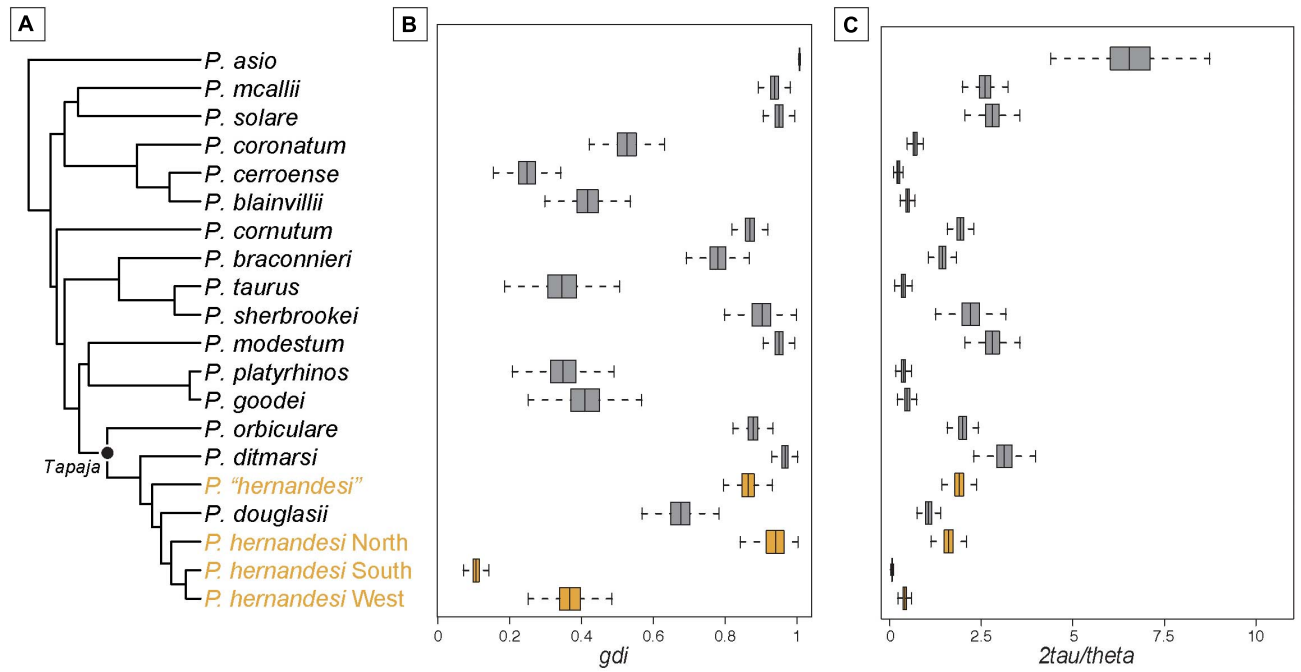

FIGURE 4 | Comparison of *Phrynosoma* populations and species using measures of divergence. **(A)** Guide-tree used for the multispecies coalescent (MSC) estimation of population sizes (θ) and divergence times (τ) in BPP. The topology was estimated using 584 nuclear loci (Leaché and Linkem, 2015). The topology for the Short-horned Lizards (*Tapaja*) has been updated to include the populations shown in **Figures 1**, **2**. Box and whisker plots for *gdi* **(B)** and coalescent units **(C)** calculated using the posterior probability distributions for θ and τ estimated with BPP.

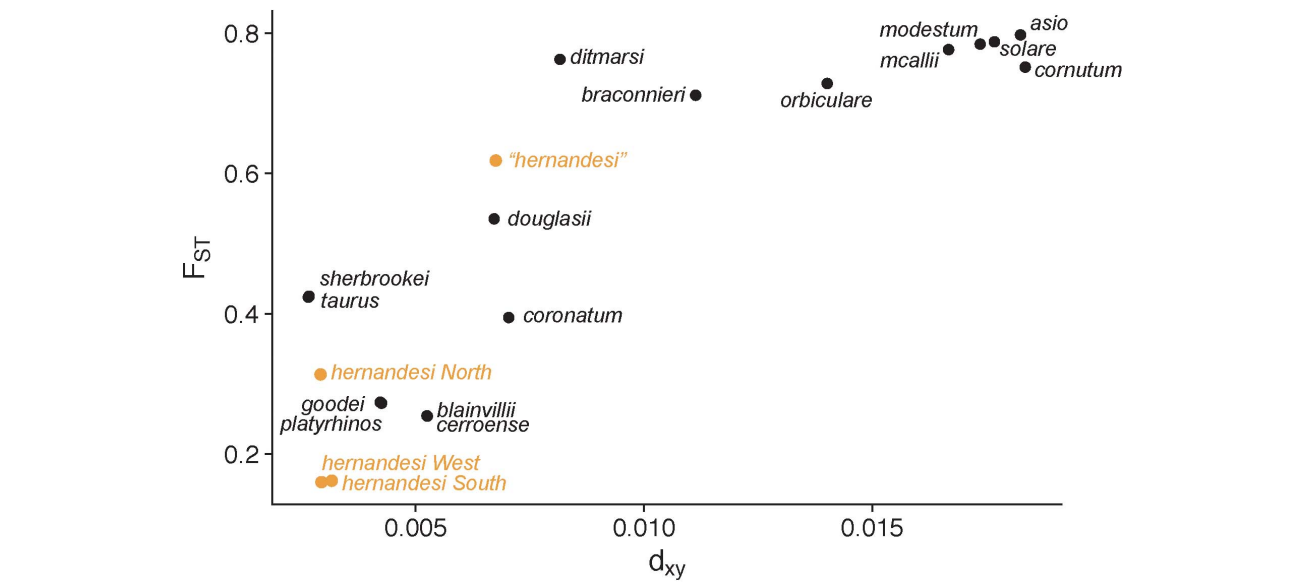

FIGURE 5 | F_{ST} and d_{xy} comparisons among *Phrynosoma* species (black dots) and four *P. hernandesi* populations (orange dots). F_{ST} and d_{xy} were estimated among all population-pairs. For each taxon, shown are the lowest F_{ST} and d_{xy} values across all its population comparisons. The three *P. hernandesi* populations (north, south, west) show markedly lower levels of genetic divergence compared to currently described species. The phylogenetically divergent *P. "hernandesi"* lineage has relatively high levels of genetic divergence in comparison to other described species.

incomplete lineage sorting. We were not able to include samples for *P. o. brachycercum* from Mexico, and additional work is needed to clarify how this taxon is related to *P. ornatissimum* and *P. hernandesi*. Based on geography and morphological similarities (Montanucci, 2015), it is likely *P. o. brachycercum* will be verified as conspecific with *P. ornatissimum*.

The results presented here call into question several of the morphology-based taxa described by Montanucci (2015), including *P. bauri, P. brevirostris, P. diminutum,* and *P. hernandesi ornatum* (**Figure 1**). We propose that these taxa should be lumped and placed within *P. hernandesi* sensu stricto (**Supplementary Figure 5**). *Phrynosoma hernandesi* contains at least three populations that are apparently connected by gene flow. Previous studies of *P. hernandesi* identified large geographic regions containing putative hybrid individuals with intermediate morphological characteristics (Montanucci, 2015). The nuclear data presented here provides additional support for hybridization. Clustering analyses revealed admixed individuals occurring in regions between populations (**Figure 3**), and demographic modeling inferred significant migration rates (>1 migration per generation). In addition, demographic modeling suggests that gene flow occurred during secondary contact following a period of divergence in isolation (**Table 2**). Theoretically, recent secondary contact can reinforce reproductive isolation as the offspring of the reconnected populations often have reduced fitness (Servedio, 2004). Alternatively, lineage fusion could be a possibility given some migration estimates exceed 1 migrant per generation (**Table 3**). Significant sampling gaps remain throughout the regions where admixed samples occur; collecting more specimens and data from these gaps will inform our understanding of the frequency of hybridization and introgression.

The genetic diversity of the three populations within *Phrynosoma hernandesi* is uneven, which has a direct influence on the coalescent-based estimates of genetic divergence. The west and south populations have relatively large population sizes (θ), and, together with their recent divergence time, this places them both at the low end of the speciation continuum in comparison to most other *Phrynosoma* (**Figure 4**). In contrast, the north population of *P. hernandesi* could be considered a separate species based on the coalescent estimates of genetic divergence in the reference-based taxonomy (**Figure 4**), but we argue that this result is driven primarily by small population size (θ). Genetic diversity is low for the north population, likely resulting from a recent bottleneck and/or recent population expansion into northern latitudes (Leung et al., 2014). However, comparisons of F_{ST} and d_{xy} values suggest that the northern population of *P. hernandesi* is at the lower end of the *Phrynosoma* speciation continuum along with the southern and western populations (**Figure 5**). This disparity among genetic divergence measures highlights the problematic nature that population size estimates can have on heuristic species delimitation. Recent simulation work has shown that population histories that include drastically different population sizes and asymmetric migration rates can create an anomaly zone with skewed gene tree probabilities that mislead species delimitation (Jiao and Yang, 2021). This situation could apply to *P. hernandesi* populations, which have drastically different population sizes and asymmetric migration rates.

The evidence presented here for admixture and gene flow among *P. hernandesi* populations suggests that these populations are incompletely separated and that they may not represent independent evolutionary lineages. Given that the nature of population admixture and hybridization can and should have an important influence on species delimitation (Burbrink and Ruane, in press), it would be premature to describe these populations as species. Simulation studies have shown that sparse sampling and isolation by distance can lead to inaccurate species delimitations (Mason et al., 2020). Further, it is too early to tell if hybridization will lead to reinforcing or fusing of population boundaries. There is an active discussion on how to treat incompletely/partially separated lineages in species delimitation. Lineages such as these have been argued to be species by some authors (Frost and Hillis, 1990), and subspecies by others (Hillis, 2020), while still others argue that they are both species and subspecies (de Queiroz, 2020). Here, we take a conservative approach; we do not recognize these populations as subspecies or species. Given the strong evidence for lack of reproductive isolation among populations, future studies of this species complex will benefit from increased sampling at population boundaries.

A morphologically distinctive population of *Phrynosoma hernandesi* occurs in the San Luis Valley in southern Colorado and northern New Mexico (Lahti, 2010). This population was described as a new species (*P. diminutum*) by Montanucci (2015). The San Luis Valley is a broad and relatively flat valley (20,700 km^2) at the headwaters of the Rio Grande River located between the Sangre de Cristo Mountain Range to the east and the San Juan Mountain Range to the west. The population of *P. hernandesi* in the San Luis Valley is morphologically distinctive with a significantly smaller body size and proportions compared to populations in surrounding areas (Hahn, 1968; Lahti, 2010; Montanucci, 2015). The mtDNA genealogy (**Figure 1**) suggests that all samples from the San Luis Valley form a recently diverged clade (≤0.5 mya) within the central clade of *P. hernandesi* that is closely related to samples from northern New Mexico. While morphologically distinct, the recent divergence of the San Luis Valley *P. hernandesi* suggests that this population is not a unique evolutionary lineage. Additional studies of the demographic and phylogenetic history of *P. hernandesi* in the San Luis Valley are needed.

Discordance between the nuclear and mtDNA data results in conflicting interpretations of *P. hernandesi* monophyly (**Supplementary Figure 6**). The nuclear phylogeny supports the monophyly of *P. hernandesi,* whereas the mtDNA genealogy supports *P. hernandesi* paraphyly with respect to *P. ditmarsi* and *P. ornatissimum* (**Supplementary Figure 6**). These results are similar to previous analyses of mtDNA data (Zamudio et al., 1997; Leung et al., 2014; Blair and Bryson, 2017) and nuclear data (Leaché and Linkem, 2015; Leaché et al., 2015), although previous genetic studies have not considered *P. ornatissimum* as a separate species. This discordance highlights an obvious problem with using a single genetic locus to delimit species: A taxonomy based on the mtDNA data will reflect the idiosyncratic history of a single genetic locus instead of the evolutionary history of the populations and species. Incomplete lineage sorting and introgression of the mtDNA genome can lead to phylogenetic discordance (Toews and Brelsford, 2012), resulting in using mtDNA for species delimitation unreliable.

By ignoring these issues and not incorporating any published nuclear data, a recent species delimitation study of *Phrynosoma* based solely on mtDNA reduced the number of species from 17 to 12 (Köhler, 2021). We recommend the use of an 18 species taxonomy for *Phrynosoma,* which considers and builds upon all available data. This taxonomy is outlined in **Table 1**, with the addition of *P. ornatissimum* in place of *P. "hernandesi."*

The Potential and Challenges of Comparative Species Delimitation

A reference-guided approach to species delimitation has two primary benefits. First, comparative species delimitation ensures that the ultimate output—a taxonomy of species—results in comparable units within the designated clade (Fujita and Leaché, 2011). Having standardized units is essential for downstream uses of a species taxonomy, including comparative analyses of diversification, biogeographic reconstruction, and trait evolution (Ruane et al., 2014) and conservation aims (Fujita et al., 2012). Second, researchers can define the appropriate phylogenetic scale for determining the threshold (Hey and Pinho, 2012; Galtier, 2019). If set at the appropriate phylogenetic scale, this threshold can reflect shared biogeographic history, which might also affect the rates at which populations transition to species (Mittelbach et al., 2007). In this study, we compared genetic diversity metrics among *Phrynosoma*; all species in the clade have similar life history characteristics (Sherbrooke, 2003) and speciated across the same general biogeographic arena (Scarpetta et al., 2020). However, if species vary in the rate at which reproductive isolation evolves (Rabosky and Matute, 2013; Campillo et al., 2020), lineages will acquire evolutionary independence at different rates, making it difficult to identify a fixed threshold. In *Phrynosoma*, the lowest *gdi* among species is ∼0.3 for *P. cerroense* (**Figure 4**). Applying this as our threshold for a population-species boundary would lead to the recognition of some *P. hernandesi* populations as species (north and west), but not the south (**Figure 4**). However, this assumes that all the species in *Phrynosoma* achieve evolutionary independence at similar rates, a yet untested assumption.

There are several potential weaknesses of comparative species delimitation as implemented here. This approach assumes that the existing taxonomy is robust. If the existing taxonomy consists of overly split or overly lumped species, a reference-guided taxonomy would perpetuate these biases into the new species delimitations. This can be further complicated if the initial taxonomy was defined along an axis not included in the current study. For example, imagine a taxonomy initially defined by differences in external morphology and color pattern, which is pertinent to the case of the *Phrynosoma*. If a subsequent reference-guided approach measured genetic divergence among species, and if external morphology and color pattern are uncorrelated with genetic divergence, then a reference-based approach would be less useful. One solution might be to be selective in which species are included in the reference taxonomy—e.g., only including species that exist in sympatry with close relatives (Tobias et al., 2010). This approach is likely to be overly conservative in estimating separately evolving lineages, given that species that occur in sympatry are often relatively far along in the process of lineage divergence. Fortunately, the existing taxonomy of species is robust in *Phrynosoma,* and many described species exhibit high levels of genetic divergence that are indicative of species-level differences (*gdi* > 0.7; **Figure 4**).

More generally, reference-guided taxonomy works best when divergence across different axes are correlated. But, empirical examples of speciation indicate that divergence can be inconsistent across axes. Most notably perhaps are cases of ecological speciation, in which species often exhibit pronounced phenotypic divergence but limited genetic divergence. How do we best reconcile conflicting signals from multiple axes, such as those can arise from conflicts between molecular and morphological data? One solution might be to apply integrative approaches to species delimitation that can accommodate different lines of evidence in a joint analysis (Solís-Lemus et al., 2015). The present study is an extreme version of this issue; here, we see inconsistencies across multiple metrics of the same axis of divergence: Genetic divergence (**Figures 4**, **5**). The northern population of *P. hernandesi* is distinct using the *gdi* metric (**Figure 4**) but not with other genetic metrics (**Figure 5**). Because we identified that metrics relying on population size can sometimes be problematic (*gdi*; **Figure 4**), we took a conservative approach and concluded for now that the northern population of *P. hernandesi* does not meet the criteria for being named a species (**Figure 5**).

Finally, comparative species delimitation does not solve some of the most persistent and thorny issues in species delimitation. Sampling gaps can create the illusion of discrete, evolutionarily independent species units (Barley et al., 2018). However, as shown in the current study, even sparse sampling throughout parapatric population borders can reveal gene flow between putative taxa, complicating our understandings of species boundaries. Introgression more generally poses a challenge for species delimitation (Burbrink et al., 2021; Jiao and Yang, 2021). An influx of genomic data has revealed that introgression is common during population divergence and between species (Edwards et al., 2016). However, determining how much introgression is too much is not clear and might depend on the underlying genomic architecture of gene flow (Harrison and Larson, 2014; Barth et al., 2020). For example, should the relatively high exchange of migrants among *P. hernandesi* populations be sufficient to preclude species status? How should our interpretation change if introgression is heterogeneous across the genome?

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: https://www.ncbi.nlm.nih.gov/genbank/, MW662366 - MW662452; https://www.ncbi.nlm.nih.gov/, BioProject PRJNA704386; https://datadryad.org/stash, doi:10.5061/dryad.gtht76hmq; https://www.ncbi.nlm.nih.gov/genbank/, MZ313846 - MZ313861.

ETHICS STATEMENT

All animal research protocols presented in this study were approved by the University of Washington Institutional Animal Care and Use Committee (UW IACUC #4367-03).

AUTHOR CONTRIBUTIONS

KZ and MEL contributed the samples. HD collected the data. AL and SS performed the analyses. All authors wrote the manuscript and designed the study.

FUNDING

This project was supported by the National Science Foundation grants to AL (NSF-SBS-2023723), SS (NSF-SBS-2023979), and MF (NSF-SBS-2024014).

ACKNOWLEDGMENTS

We thank the Arizona Game and Fish Department for permission to collect specimens for scientific research (LIC# SP568189). We thank K. Epperly for assisting in ddRADseq and mtDNA data collection. We thank the following collections for tissue loans: Louisiana State University Museum of Natural Science; Museum of Southwestern Biology, University of New Mexico; Burke Museum of Natural History and Culture, University of Washington; California Academy of Sciences; Museum of Vertebrate Zoology, University of California, Berkeley; Oklahoma Collection of Genomic Resources, Sam Noble Oklahoma Museum of Natural History, The University of Oklahoma; Dr. Rafael Alejandro Lara Reséndiz, Instituto de Biología, UNAM; Peabody Museum of Natural History, Yale University; and Dr. B. Riddle, University of Las Vegas. We thank Keaka Farleigh and Dr. Tereza Jezkova for sharing data for *Phrynosoma platyrhinos*. MEL thanks to Drs. E.D. Brodie and M. E. Pfrender for their mentorship and lab use for *Phrynosoma hernandesi* samples from the San Luis Valley. This work used the Vincent J. Coates Genomics Sequencing Laboratory at UC Berkeley, supported by NIH S10 OD018174 Instrumentation Grant.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fevo.2021.678110/full#supplementary-material

REFERENCES

Alexander, D. H., Novembre, J., and Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* 19, 1655–1664. doi: 10.1101/gr.094052.109

Barley, A. J., Brown, J. M., and Thomson, R. C. (2018). Impact of model violations on the inference of species boundaries under the multispecies coalescent. *Syst. Biol.* 67, 269–284. doi: 10.1093/sysbio/syx073

Barth, J. M., Gubili, C., Matschiner, M., Tørresen, O. K., Watanabe, S., Egger, B., et al. (2020). Stable species boundaries despite ten million years of hybridization in tropical eels. *Nat. commun.* 11, 1–13.

Blair, C., and Ané, C. (2020). Phylogenetic trees and networks can serve as powerful and complementary approaches for analysis of genomic data. *Syst. Biol.* 69, 593–601. doi: 10.1093/sysbio/syz056

Blair, C., and Bryson, R. W. (2017). Cryptic diversity and discordance in single-locus species delimitation methods within horned lizards (Phrynosomatidae: *Phrynosoma*). *Mol. Ecol. Res.* 17, 1168–1182. doi: 10.1111/1755-0998.12658

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

Bryant, D., and Moulton, V. (2004). Neighbor-net: an agglomerative method for the construction of phylogenetic networks. *Mol. Biol. Evol.* 21, 255–265. doi: 10.1093/molbev/msh018

Bryant, D., Bouckaert, R., Felsenstein, J., Rosenberg, N. A., and RoyChoudhury, A. (2012). Inferring species trees directly from biallelic genetic markers: bypassing gene trees in a full coalescent analysis. *Mol. Biol.Evol.* 29, 1917–1932. doi: 10.1093/molbev/mss086

Burbrink, F. T., Gehara, M., McKelvy, A. D., and Myers, E. A. (2021). Resolving spatial complexities of hybridization in the context of the gray zone of speciation in North American ratsnakes (*Pantherophis obsoletus* complex) *in press*. *Evolution* 75, 260–277. doi: 10.1111/evo.14141

Campillo, L. C., Barley, A. J., and Thomson, R. C. (2020). Model-based species delimitation: are coalescent species reproductively isolated? *Syst. Biol.* 69, 708–721. doi: 10.1093/sysbio/syz072

Carstens, B. C., Pelletier, T. A., Reid, N. M., and Satler, J. D. (2013). How to fail at species delimitation. *Mol. Ecol.* 22, 4369–4383. doi: 10.1111/mec.12413

Chan, K. O., and Grismer, L. L. (2019). To split or not to split? Multilocus phylogeny and molecular species delimitation of southeast Asian toads (family: Bufonidae). *BMC Evol. Biol.* 19:95. doi: 10.1186/s12862-019-1422-3

de Queiroz, K. (1998). "The general lineage concept of species, species criteria, and the process of speciation: a conceptual unification and terminological recommendations," in *Endless Forms: Species And Speciation*, eds D. J. Howard and S. H. Berlocher (Oxford: Oxford University Press), 57–75.

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

de Queiroz, K. (2020). An updated concept of subspecies resolves a dispute about the taxonomy of incompletely separated lineages. *Herpetol. Rev.* 51, 459–461.

Eaton, D. A., and Overcast, I. (2020). ipyrad: interactive assembly and analysis of RADseq datasets. *Bioinformatics* 36, 2592–2594. doi: 10.1093/bioinformatics/btz966

Edwards, S. V., Potter, S., Schmitt, C. J., Bragg, J. G., and Moritz, C. (2016). Reticulation, divergence, and the phylogeography-phylogenetics continuum. *Proc. Nat. Acad. Sci. U.S.A.* 113, 8025–8032. doi: 10.1073/pnas.1601066113

Excoffier, L., Dupanloup, I., Huerta-Sánchez, E., Sousa, V. C., and Foll, M. (2013). Robust demographic inference from genomic and SNP data. *PLoS Genetics* 9:e1003905. doi: 10.1371/journal.pgen.1003905

Flouri, T., Jiao, X., Rannala, B., and Yang, Z. (2018). Species tree inference with BPP using genomic sequences and the multispecies coalescent. *Mol. Biol. Evol.* 35, 2585–2593. doi: 10.1093/molbev/msy147

Frost, D. R., and Hillis, D. M. (1990). Species in concept and practice: herpetological applications. *Herpetologica* 46, 87–104.

Fujita, M. K., and Leaché, A. D. (2011). A coalescent perspective on delimiting and naming species: a reply to Bauer et al. *Proc. R. Soc. B: Biol. Sci.* 278, 493–495. doi: 10.1098/rspb.2010.1864

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

Galtier, N. (2019). Delineating species in the speciation continuum: a proposal. *Evol. Appl.* 12, 657–663. doi: 10.1111/eva.12748

Girard, C. (1858). Herpetology. United States Exploring Expedition. During the years 1838, 1839, 1840, 1841, 1842 under the command of Charles Wilkes, U.S.N. J. B. Lippincott and Co., Philadelphia, PA., 20, xviii + 496 pp. Philadelphia, PA: C. Sherman & Son.

Hahn, D. E. (1968). *A Biogeographic Analysis of the Herpetofauna of the San Luis Valley, Colorado. Master's Thesis*. Baton Rouge, LA Louisiana State University.

Harrison, R. G., and Larson, E. L. (2014). Hybridization, introgression, and the nature of species boundaries. *J. Hered.* 105, 795–809. doi: 10.1093/jhered/esu033

Hebert, P. D. N., Stoeckle, M. Y., Zemlak, T. S., and Francis, C. M. (2004). Identification of birds through DNA barcodes. *PLoS Biol.* 2:e312. doi: 10.1371/journal.pbio.0020312

Hebert, P. D. N., and Gregory, T. R. (2005). The promise of DNA barcoding for taxonomy. *Syst. Biol.* 54, 852–859. doi: 10.1080/10635150500354886

Hey, J., and Pinho, C. (2012). Population genetics and objectivity in species diagnosis. *Evolution* 66, 1413–1429. doi: 10.1111/j.1558-5646.2011.01542.x

Hickerson, M. J., Meyer, C. P., and Moritz, C. (2006). DNA barcoding will often fail to discover new animal species over broad parameter space. *Syst. Biol.* 55, 729–739. doi: 10.1080/10635150600969898

Hillis, D. M. (2020). The detection and naming of geographic variation within species. *Herpetol. Rev.* 51, 52–56.

Huson, D. H., and Bryant, D. (2006). Application of phylogenetic networks in evolutionary studies. *Mol. Biol. Evol.* 23, 254–267. doi: 10.1093/molbev/msj030

Jackson, N. D., Carstens, B. C., Morales, A. E., and O'Meara, B. C. (2017). Species delimitation with gene flow. *Syst. Biol.* 66, 799–812.

Jiao, X., and Yang, Z. (2021). Defining species when there is gene flow. *Syst. Biol.* 70, 108–119. doi: 10.1093/sysbio/syaa052

Jombart, T. (2008). adegenet: a R package for the multivariate analysis of genetic markers. *Bioinformatics* 24, 1403–1405. doi: 10.1093/bioinformatics/btn129

Jones, G., Aydin, Z., and Oxelman, B. (2015). DISSECT: an assignment-free Bayesian discovery method for species delimitation under the multispecies coalescent. *Bioinformatics* 31, 991–998. doi: 10.1093/bioinformatics/btu770

Kapli, P., Lutteropp, S., Zhang, J., Kobert, K., Pavlidis, P., Stamatakis, A., et al. (2017). Multi-rate Poisson tree processes for single-locus species delimitation under maximum likelihood and Markov chain Monte Carlo. *Bioinformatics* 33, 1630–1638.

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

Köhler, G. (2021). Taxonomy of horned lizards, genus *Phrynosoma* (Squamata. Phrynosomatidae). *Taxonomy* 1, 83–115. doi: 10.3390/taxonomy1020009

Lahti, M. E. (2010). *The Status Of Dwarfed Populations Of Short-Horned Lizards (Phrynosoma hernandesi)* and Great Plains Toads (*Anaxyrus cognatus*) in the san Luis Valley, Colorado. Dissertation. Logan (UT): Utah State University.

Leaché, A. D., and McGuire, J. A. (2006). Phylogenetic relationships of horned lizards (*Phrynosoma*) based on nuclear and mitochondrial data: evidence for a misleading mitochondrial gene tree. *Mol. Phylogenet. Evol.* 39, 628–644. doi: 10.1016/j.ympev.2005.12.016

Leaché, A. D., and Linkem, C. W. (2015). Phylogenomics of horned lizards (Genus: *Phrynosoma*) using targeted sequence capture data. *Copeia* 103, 586–594. doi: 10.1643/ch-15-248

Leaché, A. D., Banbury, B. L., Felsenstein, J., De Oca, A. N. M., and Stamatakis, A. (2015). Short tree, long tree, right tree, wrong tree: new acquisition bias corrections for inferring SNP phylogenies. *Syst. Biol.* 64, 1032–1047. doi: 10.1093/sysbio/syv053

Leaché, A. D., Zhu, T., Rannala, B., and Yang, Z. (2019). The spectre of too many species. *Syst. Biol.* 68, 168–181. doi: 10.1093/sysbio/syy051

Leung, M. N.-Y., Paszkowski, C. A., and Russell, A. P. (2014). Genetic structure of the endangered greater short -horned lizard (*Phrynosoma hernandesi*) in Canada: evidence from mitochondrial and nuclear genes. *Can. J. Zoo.* 92, 875–883. doi: 10.1139/cjz-2014-0079

Mason, N. A., Fletcher, N. K., Gill, B. A., Funk, W. C., and Zamudio, K. R. (2020). Coalescent-based species delimitation is sensitive to geographic sampling and isolation by distance. *Syst. Biodivers.* 18, 269–280. doi: 10.1080/14772000.2020.1730475

Mayr, E. (1969). *Principles of Systematic Zoology*. New York, NY: McGraw-Hill.

Mittelbach, G. G., Schemske, D. W., Cornell, H. V., Allen, A. P., Brown, J. M., Bush, M. B., et al. (2007). Evolution and the latitudinal diversity gradient: speciation, extinction and biogeography. *Ecol. Lett.* 10, 315–331.

Montanucci, R. R. (2015). A taxonomic revision of the *Phrynosoma douglasii* species complex (Squamata: Phrynosomatidae). *Zootaxa* 4015, 1–177. doi: 10.11646/zootaxa.4015.1.1

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

Nei, M. (1987). *Molecular Evolutionary Genetics*. New York, NY: Columbia University Press.

Padial, J. M., Miralles, A., De la Riva, I., and Vences, M. (2010). The integrative future of taxonomy. *Front. Zoo.* 7:1–14. doi: 10.1186/1742-9994-7-16

Pamilo, P., and Nei, M. (1988). Relationships between gene trees and species trees. *Mol. Biol. Evol.* 5, 568–583.

Perry, B. W., Card, D. C., McGlothlin, J. W., Pasquesi, G. I., Adams, R. H., Schield, D. R., et al. (2018). Molecular adaptations for sensing and securing prey and insight into amniote genome diversity from the garter snake genome. *Genome Biol. Evol.* 10, 2110–2129. doi: 10.1093/gbe/evy157

Peterson, B. K., Weber, J. N., Kay, E. H., Fisher, H. S., and Hoekstra, H. E. (2012). Double digest RADseq: an inexpensive method for de novo SNP discovery and genotyping in model and non-model species. *PloS One* 7:e37135. doi: 10.1371/journal.pone.0037135

Poelstra, J., Salmona, J., Tiley, G. P., Schüßler, D., Blanco, M. B., Andriambeloson, J. B., et al. (2021). Cryptic patterns of speciation in cryptic primates: microendemic mouse lemurs and the multispecies coalescent. *Syst. Biol.* 70, 203–218. doi: 10.1093/sysbio/syaa053

Rabosky, D. L., and Matute, D. R. (2013). Macroevolutionary speciation rates are decoupled from the evolution of intrinsic reproductive isolation in *Drosophila* and birds. *Proc. Nat. Acad. Sci. U.S.A.* 110, 15354–15359. doi: 10.1073/pnas.1305529110

Rannala, B. (2015). The art and science of species delimitation. *Curr. Zoo.* 61, 846–853. doi: 10.1093/czoolo/61.5.846

Reich, D., Thangaraj, K., Patterson, N., Price, A. L., and Singh, L. (2009). Reconstructing Indian population history. *Nature* 461, 489–494. doi: 10.1038/nature08365

Ruane, S., Bryson, R. W., Pyron, R. A., and Burbrink, F. T. (2014). Coalescent species delimitation in milksnakes (genus *Lampropeltis*) and impacts on phylogenetic comparative analyses. *Syst. Biol.* 63, 231–250. doi: 10.1093/sysbio/syt099

Scarpetta, S. G., Ledesma, D. T., Llauger, F. O., and White, B. A. (2020). Evolution of North American lizards. *eLS* 1, 705–717. doi: 10.1002/9780470015902.a0029078

Servedio, M. R. (2004). The what and why of research on reinforcement. *PLoS Biol.* 2:e420. doi: 10.1371/journal.pbio.0020420

Sherbrooke, W. C. (2003). *Introduction to Horned Lizards of North America (No. 64)*. Berkeley: University of California Press.

Sites, J. W., and Marshall, J. C. (2003). Delimiting species: a Renaissance issue in systematic biology. *Trends Ecol. Evol.* 18, 462–470. doi: 10.1016/s0169-5347(03)00184-8

Sites, J. W., and Marshall, J. C. (2004). Operational criteria for delimiting species. *Ann. Rev. Ecol. Evol. Syst.* 35, 199–227. doi: 10.1146/annurev.ecolsys.35.112202.130128

Smith, M. L., and Carstens, B. C. (2020). Process-based species delimitation leads to identification of more biologically relevant species. *Evolution* 74, 216–229. doi: 10.1111/evo.13878

Solís-Lemus, C., Knowles, L. L., and Ané, C. (2015). Bayesian species delimitation combining multiple genes and traits in a unified framework. *Evolution* 69, 492–507. doi: 10.1111/evo.12582

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

Stange, M., Sánchez-Villagra, M. R., Salzburger, W., and Matschiner, M. (2018). Bayesian divergence-time estimation with genome-wide single-nucleotide polymorphism data of sea catfishes (Ariidae) supports Miocene closure of the Panamanian Isthmus. *Syst. Biol.* 67, 681–699. doi: 10.1093/sysbio/syy006

Sukumaran, J., and Knowles, L. L. (2017). Multispecies coalescent delimits structure, not species. *Proc. Nat. Acad. Sci. U.S.A.* 114, 1607–1612. doi: 10.1073/pnas.1607921114

Sukumaran, J., Holder, M. T., and Knowles, L. L. (2021). Incorporating the speciation process into species delimitation. *PLoS Comput. Biol.* 17:e1008924. doi: 10.1371/journal.pcbi.1008924

Tobias, J. A., Seddon, N., Spottiswoode, C. N., Pilgrim, J. D., Fishpool, L. D., and Collar, N. J. (2010). Quantitative criteria for species delimitation. *Ibis* 152, 724–746. doi: 10.1111/j.1474-919x.2010.01051.x

Toews, D. P., and Brelsford, A. (2012). The biogeography of mitochondrial and nuclear discordance in animals. *Mol. Ecol.* 21, 3907–3930. doi: 10.1111/j.1365-294x.2012.05664.x

Yang, Z. (2015). The BPP program for species tree estimation and species delimitation. *Curr. Zoo.* 61, 854–865. doi: 10.1093/czoolo/61.5.854

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

Yang, Z., and Rannala, B. (2017). Bayesian species identification under the multispecies coalescent provides significant improvements to DNA barcoding analyses. *Mol. Ecol.* 26, 3028–3036. doi: 10.1111/mec.14093

Zamudio, K. R., Jones, K. B., and Ward, R. H. (1997). Molecular systematics of short-horned lizards: biogeography and taxonomy of a widespread species complex. *Syst. Biol.* 46, 284–305. doi: 10.1093/sysbio/46.2.284

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

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

frontiers in Ecology and Evolution

ORIGINAL RESEARCH
published: 15 October 2021
doi: 10.3389/fevo.2021.758348

Fine-Scale Spatial Structure of Soil Microbial Communities in Burrows of a Keystone Rodent Following Mass Mortality

*Chadwick Kaufmann and Loren Cassin-Sackett**†*

Department of Integrative Biology, University of South Florida, Tampa, FL, United States

Soil microbial communities both reflect and influence biotic and abiotic processes occurring at or near the soil surface. Ecosystem engineers that physically alter the soil surface, such as burrowing ground squirrels, are expected to influence the distribution of soil microbial communities. Black-tailed prairie dogs (*Cynomys ludovicianus*) construct complex burrows in which activities such as nesting, defecating, and dying are partitioned spatially into different chambers. Prairie dogs also experience large-scale die-offs due to sylvatic plague, caused by the bacterium *Yersinia pestis*, which lead to mass mortality events with potential repercussions on microbial communities. We used 16S sequencing to examine microbial communities in soil that was excavated by prairie dogs from different burrow locations, and surface soil that was used in the construction of burrow entrances, in populations that experienced plague die-offs. Following the QIIME2 pipeline, we assessed microbial diversity at several taxonomic levels among burrow regions. To do so, we computed community similarity metrics (Bray–Curtis, Jaccard, and weighted and unweighted UniFrac) among samples and community diversity indexes (Shannon and Faith phylogenetic diversity indexes) within each sample. Microbial communities differed across burrow regions, and several taxa exhibited spatial variation in relative abundance. Microbial ecological diversity (Shannon index) was highest in soil recently excavated from within burrows and soils associated with dead animals, and was lowest in soils associated with scat. Phylogenetic diversity varied only marginally within burrows, but the trends paralleled those for Shannon diversity. *Yersinia* was detected in four samples from one colony, marking the first time the genus has been sampled from soil on prairie dog colonies. The presence of *Yersinia* was a significant predictor of five bacterial families and eight microbial genera, most of which were rare taxa found in higher abundance in the presence of *Yersinia*, and one of which, *Dictyostelium*, has been proposed as an enzootic reservoir of *Y. pestis*. This study demonstrates that mammalian modifications to soil structure by physical alterations and by mass mortality can influence the distribution and diversity of microbial communities.

Keywords: environmental microbiome project, nutrient pulse, grasslands, pathogens, extirpation, spatial partitioning

OPEN ACCESS

Edited by:
Susana Caballero,
University of Los Andes, Colombia

Reviewed by:
Huan Li,
Lanzhou University, China
Ismail Kudret Saglam,
Koç University, Turkey

****Correspondence:***
Loren Cassin-Sackett
cassin.sackett@gmail.com

†Present address:
Loren Cassin-Sackett,
Department of Biology,
University of Louisiana, Lafayette, LA,
United States

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

Received: *13 August 2021*
Accepted: *24 September 2021*
Published: *15 October 2021*

Citation:
Kaufmann C and Cassin-Sackett L (2021) Fine-Scale Spatial Structure of Soil Microbial Communities in Burrows of a Keystone Rodent Following Mass Mortality. Front. Ecol. Evol. 9:758348. doi: 10.3389/fevo.2021.758348

INTRODUCTION

Microbial communities are diverse assemblages of microbiotic species that, through interactions with each other and with the physical and chemical components of their abiotic environments, have substantial impacts on global processes. Microbes play an important role in global nutrient cycling (Treseder et al., 2016; Heijboer et al., 2018) and energy flow through ecosystems (Konopka, 2009). In turn, microbial communities are structured by the physical and chemical properties (Leff et al., 2015; Garcia et al., 2020; Xia et al., 2020) of the soil substrate, including soil moisture, C:N ratio, pH, and total carbon content (Shen et al., 2013; Li et al., 2017).

In addition to their interactions with abiotic processes, soil microbiota structure biotic diversity and regulate the health of hosts that house the microbial communities (Ichinohe et al., 2011; Shen et al., 2018). Soil microbes influence plant and animal communities (Lau and Lennon, 2011; Seastedt et al., submitted[1]) through mechanisms such as increasing plant nutrient acquisition (Hestrin et al., 2019) and resistance to desiccation (Xi et al., 2018) and inhibiting or facilitating the establishment of pathogens (Perez et al., 2008; van Elsas et al., 2012). Soil microbes are in turn governed by the actions of plants (Zak et al., 2003; Prescott and Grayston, 2013; Lange et al., 2015) and animals (Kandeler et al., 1999; Cline et al., 2017; Bray et al., 2019), creating feedbacks between soil microbial and aboveground communities (Bartelt-Ryser et al., 2005).

Biotic and abiotic processes that influence soil characteristics may be predicted to govern microbial diversity. For instance, ecosystem engineers that influence sediment abiotic properties (e.g., bioturbating shrimp, Laverock et al., 2010; *Populus*, Ciadamidaro et al., 2013) or soil nutrients (e.g., prairie dogs, Anacker et al., 2021) should thus also determine the microbial communities present (Gutiérrez and Jones, 2006; Cregger et al., 2018; Zotti et al., 2020). Similarly, mass mortality events in animals supply nutrient pulses that should alter microbial communities and contribute to terrestrial nutrient cycling (Metcalf et al., 2016b). Mass mortality in ecosystem engineers or keystone species, which influence the abundance of other (typically plant and animal) taxa, could have an especially pronounced effect. Soil microbiota can regulate the microbial pathogens causing such mass mortality, for instance if soil microbial communities contain animal pathogens or reservoirs for animal pathogens (Markman et al., 2018) or, conversely, microbes that inhibit establishment of animal pathogens. Through facilitation or inhibition of pathogens (Perez et al., 2008; van Elsas et al., 2012), soil microbes thus contribute to the maintenance of biodiversity of plants and animals.

Black-tailed prairie dogs (*Cynomys ludovicianus*) are social, fossorial ground squirrels inhabiting North American grasslands that build extensive underground burrows. Burrows typically range in length from 5 to 10 m long and extend as deep as 3–4 m below ground (Wilcomb, 1954; Hoogland, 1995). Burrows maximize air and water flow through the burrow and minimize water retention within the burrow, thus creating a moist but not wet environment. Their burrows increase soil porosity (Gedeon et al., 2012, which can facilitate deeper penetration of precipitation (Munn, 1993). Prairie dogs also increase the total nitrogen content and productivity of soils inside or near their burrows, leading to higher plant growth and diversity (Whicker and Detling, 1988; Holland and Detling, 1990).

More than half of a prairie dog's life is spent within its burrow: prairie dogs use their burrows for reproducing, storing food, and escaping from both predators and the environment (Hoogland, 1995). Therefore, burrows are complex and heterogeneous in structure, and include spatially segregated chambers with various purposes, including nesting, hibernating (in species that hibernate; Cooke and Swiecki, 1992), defecating, and burying or isolating dead kin (Burns et al., 1989). Prairie dogs can die in their burrows over winter as a result of insufficient resources, and at other times of year from causes such as infectious disease. The primary disease affecting prairie dogs is sylvatic plague, caused by the Gram-negative bacterium *Yersinia pestis*. Typically transmitted by fleas, the pathogen is extremely virulent to prairie dogs, with individual colonies undergoing severe population declines ranging between 85% and complete extinction (Cully et al., 2010). These die-offs can thus result in hundreds of kilograms of carcasses appearing over the course of several weeks. In between epizootics, the plague reservoir is unknown: Some have hypothesized the pathogen persists in an alternative mammalian (Salkeld et al., 2010) host or flea vector (Webb et al., 2006) while others have posited that the reservoir is telluric (Drancourt and earlier; Eisen et al., 2008), residing in an invertebrate such as a nematode or amoeba (Markman et al., 2018).

Prairie dogs regularly clean out their burrows, leaving piles of nesting material, scat, and bones near some entrances of burrows. This excavated soil provides an opportunity to non-invasively explore the microbial composition of various locations within prairie dog burrows. We hypothesize that prairie dogs structure soil microbial communities through their functional partitioning of burrows, and that this structure may be pronounced after mass mortality caused by the pathogen *Y. pestis*. This study is the first to characterize the fine-scale spatial variation in microbial communities in the complex structure of prairie dog burrows.

MATERIALS AND METHODS

Soil Sampling and Processing

Seventy-nine soil samples were collected in 2009 from six prairie dog colonies (named 1A, 12A, 17A, 19A, 30A, and 47A after Bai et al., 2008; Sackett et al., 2013; **Supplementary Figure 1**) located in Boulder County, CO (United States). All six colonies experienced die-offs from plague in 2006, and recolonization had begun in 2007 (five colonies) or 2008 (one colony; Sackett et al., 2013). Samples were collected from several locations, targeting different regions of the inner burrow (**Figure 1**; designed after Wilcomb, 1954): (1) loose soil on or adjacent to the burrow mound that had been recently excavated from within the burrow,

[1]Seastedt, T. R., Porazinska, D. L., Gendron, E. M. S., and Schmidt, S. K. (submitted). An annual grass restructures the soil food web and alters soil carbon sequestration of a perennial grassland. *Plant Soil*

"adjacent"; (2) soil at the burrow entrance that had been excavated from within the burrow along with prairie dog bones, "bones"; (3) soil at the burrow entrance that had been excavated along with prairie dog bones and scat, "bones + scat"; (4) soil that had been excavated along with remnants of a dead prairie dog, or soil at the entrance of a burrow emitting the smell of a dead animal, "dead"; (5) soil collected from within the mouth/entrance of the burrow, "entrance"; (6) loose soil from burrows containing plague-exposed animals (Sackett et al., 2013) in previous years, "plague"; and (7) soil at the burrow entrance located next to prairie dog scat (usually scat had been excavated from within the burrow), "scat." Whenever possible (in all but five cases, where dry soils were sampled from beneath bones), we sampled soil that was still moist (indicated by visible moisture). Soils were stored frozen in 15 mL vials or plastic ziploc bags until nutrient analysis and DNA extraction.

Nutrient analysis was performed at the Institute for Arctic and Alpine Research and at the Mountain Research Station at the University of Colorado. Total carbon, total nitrogen content, and C:N ratios were assessed on a CHN analyzer (LECO Corp., St. Joseph, MI, United States) with a standard run in between every 10 samples. Soil moisture was estimated by drying ∼1–2 g soil in an oven at 105°C for 5 days, weighing the samples before and after drying, and dividing the water weight by the wet soil weight. pH was measured using a ∼1:2 ratio of soil:water.

Variation in pH, water content, total nitrogen, total carbon, and carbon:nitrogen ratio among colonies and among regions within burrows were assessed using one-way ANOVA tests computed in R (R Core Team, 2018). A Tukey *post hoc* test was subsequently conducted for factors that varied significantly. These soil properties were included as covariates in the models below.

Sequencing and Quality Control

DNA was extracted from soil samples in duplicate using a PowerSoil extraction kit (MO Bio Laboratories Inc., Carlsbad, CA, United States) following manufacturer's protocol. Sample processing, 16S sequencing, and core amplicon data analysis were performed by the Earth Microbiome Project[2] (Thompson et al., 2017), and all amplicon sequence data and metadata have been made public through the EMP data portal[3] (Qiita study 11519) and through the European Bioinformatics Institute (EBI) as project ERP106314.

The raw fastq files were compiled into a QIIME2 archive and all analyses were performed using Qiita (Gonzalez et al., 2018) and QIIME2 (RRID:SCR_021258, version 2017.8 or later). Sequences were demultiplexed using the demux plugin of QIIME2 and denoised using DADA2 (Callahan et al., 2016). The median Phred score of the sequences never dropped below 30; therefore, 3 bp were trimmed from the beginning and 5 bp from the end of each sequence to ensure all adapter sequences were removed. Both a feature table and its representative sequences were produced following denoising.

[2]www.earthmicrobiome.org

[3]qiita.microbio.me/emp

Analysis and Visualization

Taxonomic analysis of the soil samples was performed using a naive Bayesian classifier (Wang et al., 2007) trained using the Greengenes 13_8 99% OTUs (DeSantis et al., 2006; McDonald et al., 2012). This classifier was used along with the representative soil sequences in the q2-feature-classifier plugin (Bokulich et al., 2018) of QIIME2 (Bolyen et al., 2019) to assign taxonomies. Differences in the most abundant taxon in each burrow region were examined with a Chi-square test in R using different taxonomic levels.

Sequences were aligned and masked using mafft (Katoh and Standley, 2013), and an unrooted phylogenetic tree was generated using FastTree (Price et al., 2010). The tree was then rooted at its midpoint using the QIIME2's phylogeny plugin. Using the rooted midpoint tree and the core-metrics plugin of QIIME2, the previously created feature table was rarefied with a sampling depth of 22,000 using the q2-diversity plugin to assess Bray–Curtis dissimilarity and Jaccard distance estimates and conduct a weighted (Lozupone et al., 2007) and unweighted UniFrac (Lozupone and Knight, 2005) diversity principal coordinates analysis (PCoA). All PCoA results were plotted using QIIME2's Emperor plugin (Vázquez-Baeza et al., 2013) and visualized for clustering by burrow region.

We used a generalized linear modeling approach to determine the best predictors of ecological (Shannon) and phylogenetic (Faith) diversity. To do so, we modeled diversity as a function of burrow region, using pH, water content, and soil nutrients (C, N, and C:N) as covariates. We also tested models that included colony, excluded single nutrients, included relative abundance of Enterobacteriaceae, and included the presence of *Yersinia*, and we selected the best model using AIC. Next, we assessed whether Enterobacteriaceae was unique in its contribution to model fit (see section "Results") by testing separately whether the addition of each of 565 microbial families also improved model fit. We evaluated model fit by comparing AIC values, irrespective of whether there was a significant relationship between a single taxon and diversity.

To assess the effect of burrow region on relative abundance of Enterobacteriaceae, we conducted a generalized linear model that included all predictors except colony. Next, to determine whether taxa in general varied in relative abundance at small spatial scales, we evaluated each taxon separately (565 families and 990 genera) in a generalized linear model with the same structure as the Enterobacteriaceae model. The significance of effects was determined using the Benjamini–Yekutieli false discovery rate correction (Benjamini and Yekutieli, 2001) for p-values returned from the glm.

Finally, we aimed to determine whether the presence of *Yersinia* (see section "Results") was correlated with relative abundance of other taxa or the overall diversity of the sample. To do so, we first performed a non-parametric Kruskal–Wallis test on each taxon separately at the taxonomic levels of both family and genus and assessed significance using the Benjamini–Yekutieli false discovery rate correction (Benjamini and Yekutieli, 2001). Next, we assessed whether *Yersinia* presence was associated with levels of microbial diversity by performing

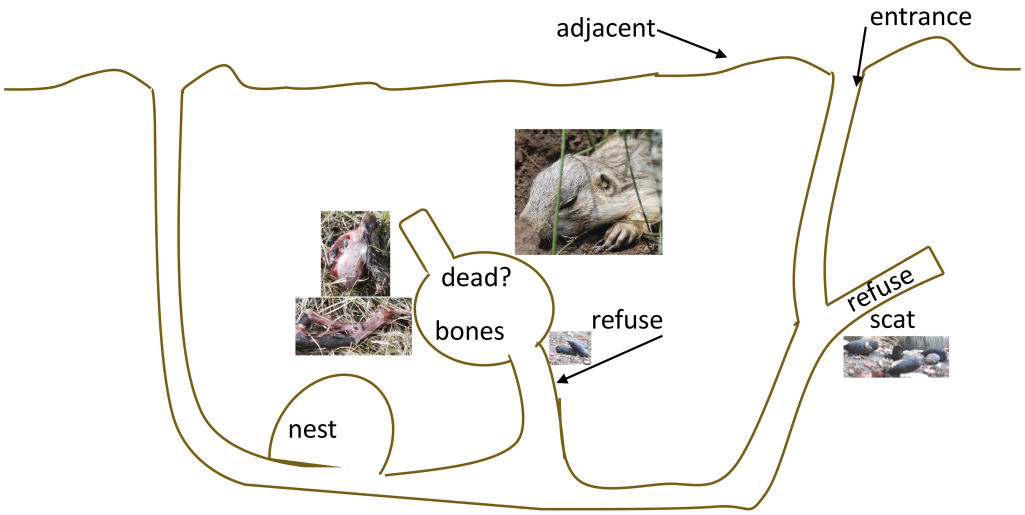

FIGURE 1 | Schematic of the regions of a prairie dog burrow (not to scale), designed after Wilcomb (1954), showing chambers used for various purposes including nesting, defecating, and isolating dead individuals.

a Kruskal–Wallis test on two measures of diversity at the genus level: the Shannon diversity index and the Faith phylogenetic diversity index (Faith, 1992). All R scripts are available on GitHub: https://github.com/CassinSackett/soilmicrobes/.

RESULTS

We obtained 79 soil samples from 64 burrows in 6 colonies. Nitrogen content averaged 0.267% (range 0.062–0.685%) and carbon content averaged 3.56% (range 0.645–7.59%); mean C:N ratio was 14.8% (range 9.18–41.5%). Mean soil water content was 0.075 g/g (range 0.004–0.22) and mean pH was 7.89 (range 6.18–9.06). There was significant variation in soil properties among colonies (**Supplementary Figure 1**) and among burrow regions (**Supplementary Figure 2**). In particular, soil moisture was significantly higher in colony 30A and lower in colony 1A than other sites, and pH was significantly lower in colony 12A than several other sites (but sample sizes in 12A and 30A were small). Soil moisture was significantly higher in soil collected from the burrow entrance than in excavated soil containing prairie dog bones. The C:N ratio was significantly higher in soil sampled from excavated soil containing bones and scat than all other regions except those with scat. Total carbon, total nitrogen, and pH did not vary across sampling regions.

All clustering methods produced highly similar results, with the UniFrac unweighted method resulting in the highest proportion of variance explained by the first three axes. Samples collected from recently excavated soil adjacent to burrows clustered slightly on Axis 1, but samples from different regions were largely overlapping (**Supplementary Figure 3**).

The best initial model (excluding single taxa) of Shannon's ecological diversity included the predictors: burrow region, pH, water content, and interactions between pH and water content and between carbon and nitrogen content (AIC 254.98). Colonies did not differ in ecological diversity, and inclusion of colony as a predictor worsened the model (AIC 263.65). Inclusion of *Yersinia* presence as a predictor worsened the model, but not significantly (AIC 256.97). All variables in the model significantly influenced diversity (**Supplementary Table 1**). Diversity was lowest in soil collected in the presence of scat, followed by soil with bones and scat, and was highest in soil recently excavated from burrows and from those with plague-positive animals (**Figure 2**). The best model of phylogenetic diversity included the same predictor variables (in this case, inclusion of colony as a predictor led to a worse, but statistically indistinguishable model: AIC with colony = 1125.9, AIC without colony = 1125.7). Inclusion of *Yersinia* presence as a predictor resulted in a statistically indistinguishable model (AIC 1125.8). All predictors significantly influenced phylogenetic diversity except for burrow location, which had a marginally significant effect (p = 0.074). Similar to the pattern observed for ecological diversity, phylogenetic diversity exhibited a trend toward lower diversity in soil collected in the presence of scat, followed by soil with bones and scat, and higher diversity in soil recently excavated from burrows, soil from burrows inferred to contain dead animals, and soil excavated from burrows with plague-positive animals (**Figure 2**).

Adding the relative abundance of Enterobacteriaceae improved the AIC of both models (Shannon diversity AIC 252.05, significant improvement; Faith diversity AIC 1124.3, marginal improvement). The relative abundance of Enterobacteriaceae had a negative effect on both ecological and phylogenetic diversity, although this effect seemed to be driven by an outlier with a relatively high proportion of Enterobacteriaceae and low diversity. Removing the outlier changed the magnitude (and significance) of the relationship, but the trend toward an inverse relationship persisted. The improvement of model fit with the inclusion of relative abundance of Enterobacteriaceae was not unique to this family; in fact, the inclusion of 157 single taxa significantly improved model fit (reducing AIC by more than

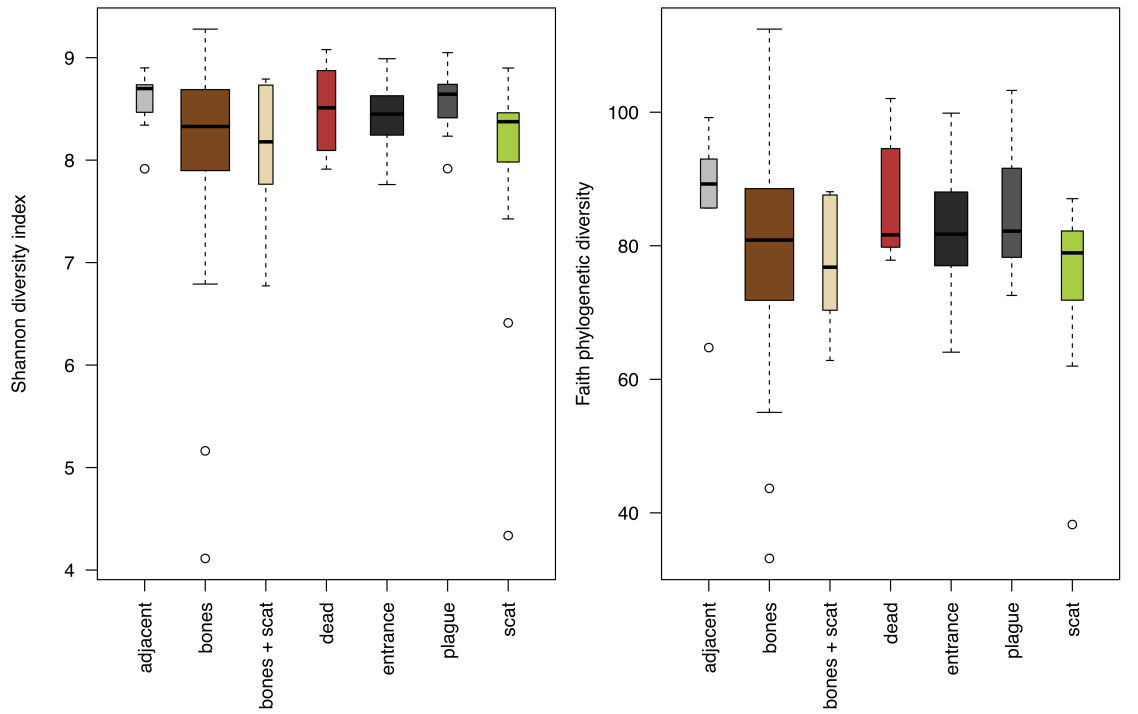

FIGURE 2 | Fine-scale variation in microbial diversity among regions of the prairie dog burrow; width of boxes represents sample size. Left: ecological diversity measured by the Shannon index. Right: phylogenetic diversity measured by the Faith phylogenetic diversity index. Width of boxes represents relative sample sizes.

2); 59 families (one at a time) reduced AIC values by >10. In particular, the best single-taxon model of Shannon's diversity included relative abundance of Planococcaceae (AIC 188.53) in addition to the previous predictors, and the only other model within 10 AIC was a model including relative abundance of Micrococcaceae (AIC 194.10). Both of these taxa exhibited a strong negative relationship with Shannon's diversity. Similarly, the inclusion of 157 single taxa significantly improved model fit (reducing AIC by more than 2) for phylogenetic diversity, and 67 families reduced AIC values by >10. The best single-taxon model of phylogenetic diversity included the relative abundance of an unknown family in order WD2101 (class Phycisphaerae, AIC 1064.83) in addition to the previous predictors, and the only other model within 10 AIC was a model including relative abundance of an unknown family in order iii1–15 (class Acidobacteria-6, AIC 1071.36). Both of these taxa exhibited a positive relationship with phylogenetic diversity.

Burrow regions differed significantly in the most abundant taxa at all taxonomic levels (phylum, class, order, family, and genus; **Supplementary Figure 4**). Across all samples, the dominant family averaged 12.5% of the total sequences per sample, and ranged from comprising 5.6–49.2% of the total sequences per sample. In soils collected from burrows inferred to currently contain dead animals, Firmicutes were more abundant than expected (Chi-square = 74.528, df = 30, p-value = 1.172e−05). Burrows with dead animals contained more Bacilli (and Bacillales) and Rubrobacteria (and Rubrobacterales) than expected, while soils containing bones were characterized by a lower abundance of Alphaproteobacteria than expected (Chi-square = 182.36, df = 72, p-value = 1.545e−11). Soils containing bones and scat possessed a lower abundance of Rhizobiales than expected (Chi-square = 234.44, df = 96, p-value = 1.382e−13).

Forty-eight bacterial families and 76 bacterial genera varied significantly in relative abundance across burrow regions (**Supplementary Tables 2–5**). Among the taxa most significantly varying across burrow regions were an unknown family and genus in the AKIW781 order of class Chloroflexi, which was an order of magnitude higher in soil with bones and scat (**Figure 3A**); *Deinococcus* (Deinococcaceae), which was an order of magnitude higher in soil with bones and scat and an order of magnitude lower in soil associated with dead animals (**Figure 3B**); an unknown genus in Planococcaceae, which was highest in soil associated with dead animals (**Figure 3C**); and *Cellulosimicrobium* (Promicromonosporaceae, Actinomycetales), which was highest in soils sampled with scat (**Figure 3D**).

Enterobacteriaceae, the family containing *Y. pestis*, was found in all soil samples, but at low proportions (never exceeding 3%). The proportion of Enterobacteriaceae sequences was significantly higher in samples with higher C:N (p = 0.0002) and in burrow regions associated with dead animals than in other regions (p = 0.013). Although we ran this model first due to our particular interest in the family, we also aimed to determine the extent to which spatial variation in abundance was characteristic shared by many microbial taxa. When we ran separate models for all

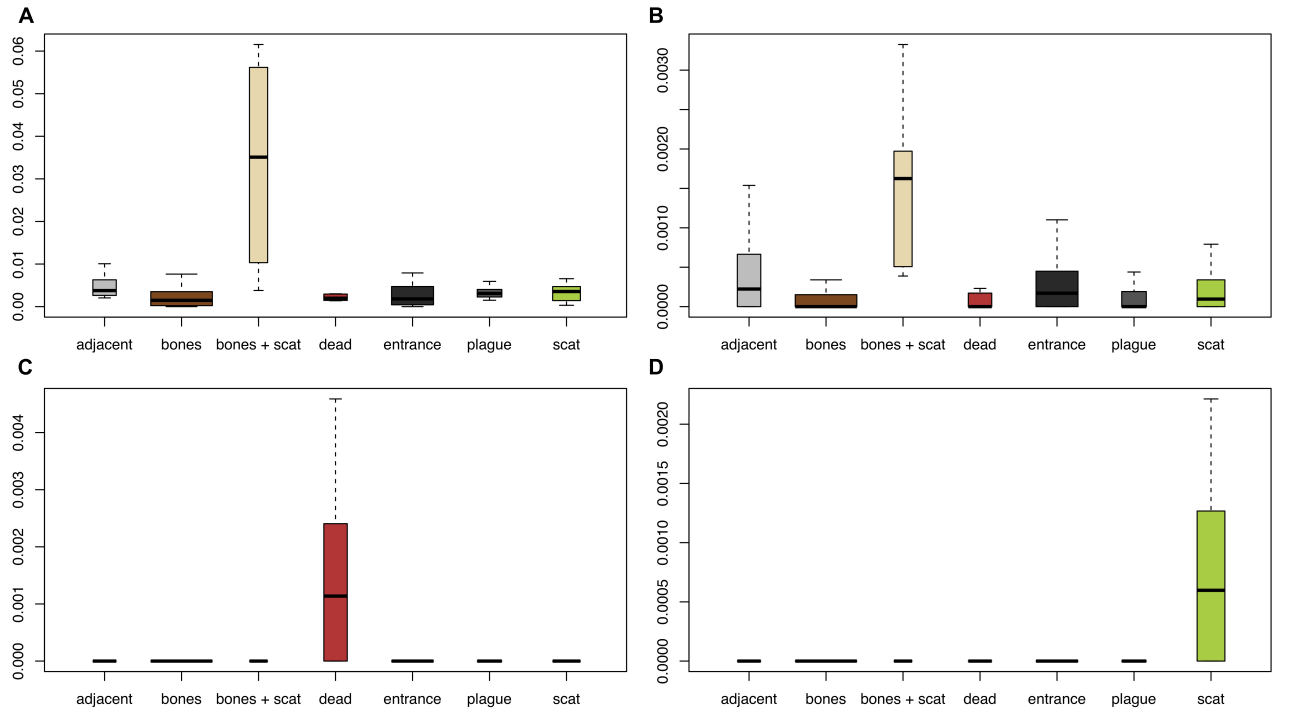

FIGURE 3 | Variation in relative abundance of four representative microbial taxa among regions of the prairie dog burrow. **(A)** An unknown family and genus in the AKIW781 order of class Chloroflexi, **(B)** *Deinococcus* (Deinococcaceae), **(C)** an unknown genus in Planococcaceae, and **(D)** *Cellulosimicrobium* (Promicromonosporaceae, Actinomycetales). *Y*-axes are the percentage of total microbial sequences that comprise represented taxa. Width of boxes represents relative sample sizes.

565 families and 990 genera, the false discovery rate correction led to a loss of statistical significance for spatial variation in Enterobacteriaceae (data not shown). *Yersinia* was identified in four samples from two burrows in one colony (19A). All *Yersinia*-containing samples were collected from waste chambers. Presence of this genus was a significant predictor of the relative abundance of five bacterial families and eight microbial genera (**Figure 4** and **Tables 1**, **2**). All of these taxa were found in significantly higher abundance in samples where *Yersinia* was present. Many of these genera (e.g., 9 out of the 10 strongest associations) were extremely rare taxa that appeared only or primarily in the samples containing *Yersinia*. The presence of *Yersinia* in a sample was associated with slightly, but not significantly, lower microbial diversity within samples (Shannon without *Yersinia* 8.303, Shannon with *Yersinia* 8.059, $p = 0.14$; Faith PD without *Yersinia* 81.807, Faith PD with *Yersinia* 75.111, $p = 0.17$; **Figure 5**).

DISCUSSION

Microbial communities as a whole varied – and many specific taxa differed in relative abundance – at small spatial scales among regions of a prairie dog burrow following a mass mortality event. Dominant taxa were consistent with predictions of microbial succession following the nutrient pulse that occurs during decomposition of mammalian corpses (Metcalf et al., 2016a,b). In addition, several taxa were significantly associated with the presence of *Yersinia* in soil samples, primarily as a result of taxa of low abundance found at higher abundance when *Yersinia* was present. Both ecological and phylogenetic diversity resulted from the combined influences of soil properties and burrow region.

Other studies have shown similar degrees of fine-scale spatial structure in microbial communities resulting from niche differentiation (Zhuang et al., 2020), particularly in microbial communities associated with plant roots (Aas et al., 2019) and other plant tissues (Cregger et al., 2018). Niche diversification may be particularly likely when niches are divergent even at small spatial scales, when specific microbes present in high abundance in certain environments exert selection on other microbial taxa (e.g., predatory microbes) or when microenvironments are less hospitable (e.g., very dry). In this system, microbial communities associated with scat may be specialized for living in the mammalian gut, metabolizing plant tissues, or both. Soil collected with bones were the driest soils we sampled, thereby potentially exerting strong selection on microbial communities in these soils.

Fine-scale spatial structure could also arise from community assembly (Nemergut et al., 2013) and succession processes such as colonization of a deceased animal from soil microbiota (Metcalf et al., 2016b), particularly if animals died in a spatially structured way or were moved to specific locations after death – scenarios that are consistent with the few existing observations of deceased

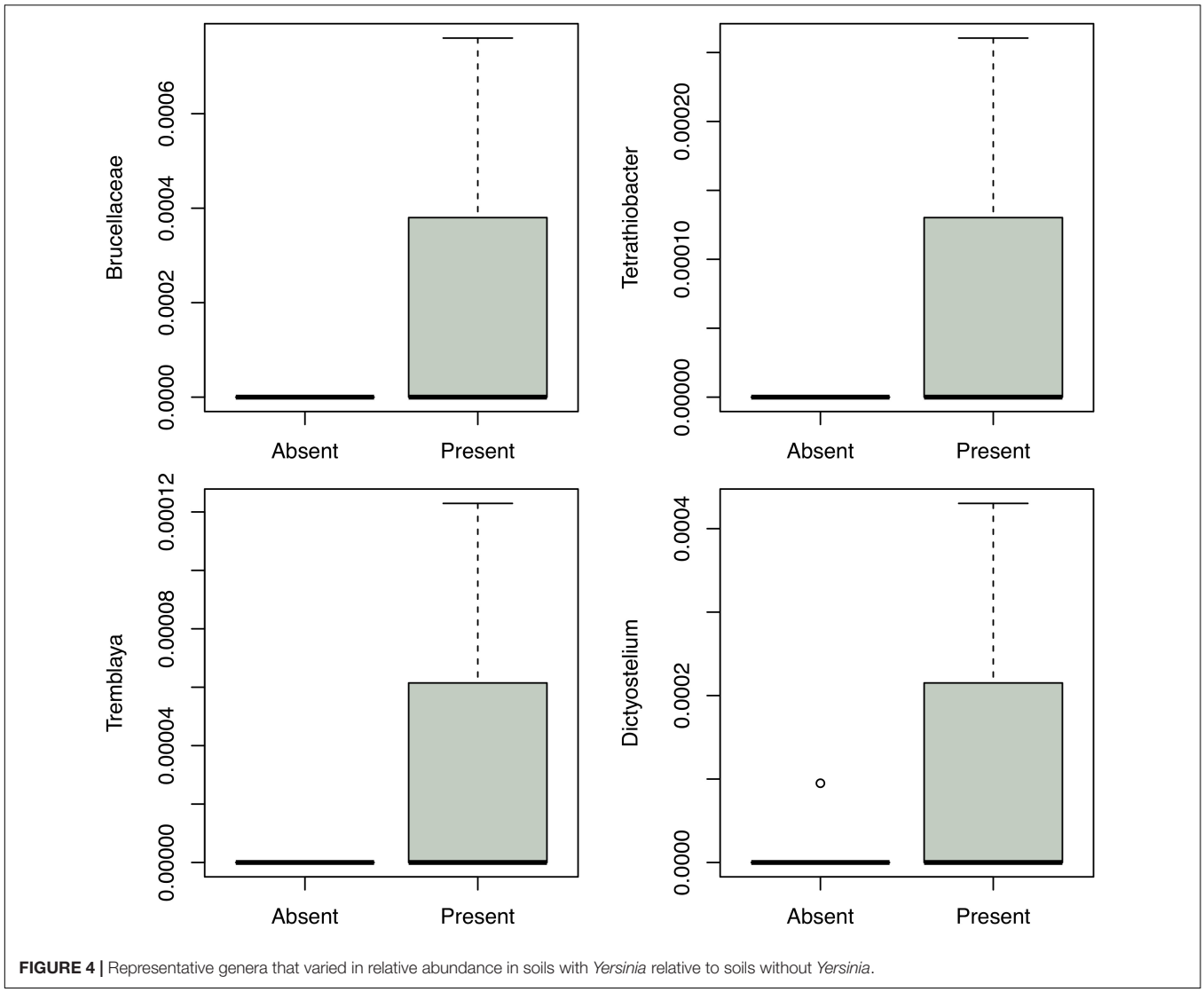

FIGURE 4 | Representative genera that varied in relative abundance in soils with *Yersinia* relative to soils without *Yersinia*.

TABLE 1 | Classification of bacterial families found at significantly higher abundance in soil samples containing *Yersinia*.

Phylum	Class	Order	Family
Proteobacteria	Betaproteobacteria	Tremblayales	Tremblayaceae
TM7 (Saccharibacteria)	TM7-3	I025	Unknown
Chloroflexi	Ktedonobacteria	Ktedonobacterales	Ktedonobacteraceae
Bacteroidetes	Bacteroidia	Bacteroidales	Rikenellaceae
Proteobacteria	Gammaproteobacteria	Legionellales	Unknown

prairie dogs within burrows (Burns et al., 1989). For instance, Enterobacteriaceae are abundant in the early stages of corpse decay, while Planococcaceae become more abundant as corpse decay progresses (Metcalf et al., 2016a). This is consistent with our observation of significantly higher abundance of both taxa in soils collected near dead animals.

Keystone microbial taxa (Banerjee et al., 2018) can influence the abundance of other community members based on ecological interactions (Herren and McMahon, 2018) including the prevention of pathogen establishment (Trivedi et al., 2017). We found >50 taxa that significantly influenced ecological or phylogenetic diversity among samples, with some having particularly strong effects. Four single taxa [Planococcaceae, Micrococcaceae, unknown family in WD2101 (Planctomycetes), and unknown family in iii1–15 (Acidobacteria)] were statistically separated as predictors of diversity (in conjunction with abiotic soil properties) from other taxa, indicating their potential role as keystone taxa. A negative relationship between Planococcaceae

TABLE 2 | Classification of microbial genera found at significantly higher abundance in soil samples containing *Yersinia*.

Phylum	Class	Order	Family	Genus
Proteobacteria	Alphaproteobacteria	Rhizobiales	Brucellaceae	Unknown
Proteobacteria	Betaproteobacteria	Burkholderiales	Alcaligenaceae	*Tetrathiobacter*
Proteobacteria	Betaproteobacteria	Tremblayales	Tremblayaceae	*Tremblaya*
Amoebozoa	Dictyostelia	Dictyosteliida	Dictyosteliidae	*Dictyostelium*
Bacteroidetes	Cytophagia	Cytophagales	Flammeovirgaceae	Unknown
TM7	TM7-3	I025	Unknown	Unknown
Chloroflexi	Ktedonobacteria	Ktedonobacterales	Ktedonobacteraceae	Unknown
Actinobacteria	Actinobacteria	Actinomycetales	Thermomonosporaceae	*Actinocorallia*
Bacteroidetes	Bacteroidia	Bacteroidales	Rikenellaceae	Unknown

In the original classification system, Dictyostelium was classified as a mitochondrially derived Rickettsiales; we have instead reported the accepted taxonomy for the genus.

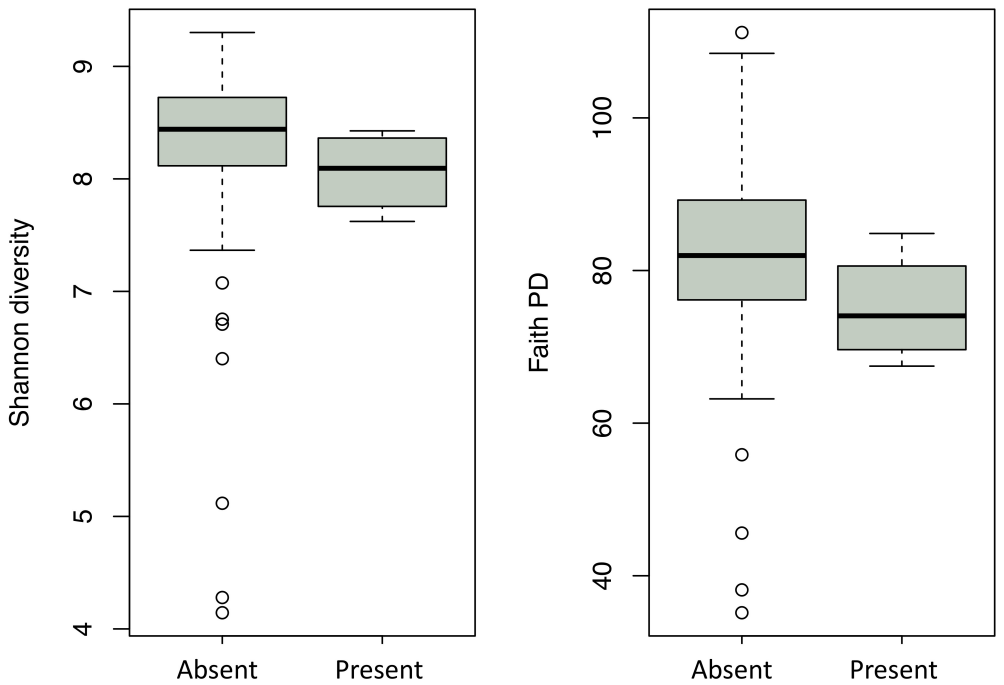

FIGURE 5 | Microbial diversity in soils with and without *Yersinia*. Left: ecological diversity measured by the Shannon index. Right: phylogenetic diversity measured by the Faith phylogenetic diversity index.

and ecological diversity supports previous findings of this family becoming more abundant after disturbance of an ecological community (Aanderud et al., 2019). Micrococcaceae have been associated with increased plant growth (Hong et al., 2016), which could cause feedbacks with microbial diversity, although the mechanism underlying this potential relationship is not clear. Both WD2101 and iii1–15 are among the most abundant soil bacteria globally (Delgado-Baquerizo et al., 2018). The lack of taxon sharing within WD2101 (<2% shared OTUs) even in similar environments (Dedysh et al., 2020) and the low degree of genomic match to characterized sequences (Delgado-Baquerizo et al., 2018) suggest a large amount of cryptic diversity in the group that could be a driving force behind the high phylogenetic diversity we found here. The abundance of iii1–15 responds to soil moisture (Barnard et al., 2013), which could provide a mechanism for its relationship with phylogenetic diversity (Brockett et al., 2012).

Among the taxa that varied spatially within prairie dog burrows was an unknown member of the AKIW781 class (order Chloroflexi), found here with bones and scat, which has previously been described in soils from deserts in North and South America (Mogul et al., 2017; Lucas et al., 2020) and is likely adapted to dry conditions. Similarly, we found Deinococcus to be higher in soils with bones and scat, which may be not only drier but more exposed to sunlight than soils excavated from other parts of the burrow. Deinococcus is resistant to solar radiation and increases in relative abundance in irradiated soils (Ogwu et al., 2019). An unknown genus in Planococcaceae was highest in soils associated with dead animals, consistent with previous description of the abundance of this family in later stages

of the decomposition process (Metcalf et al., 2016a). Finally, *Cellulosimicrobium* was found at highest relative abundance in soils containing scat, which supports the role of this genus in breaking down plant material (Bakalidou et al., 2002; Schumann and Stackebrandt, 2015).

In line with other studies of pathogens and soil microbial diversity (van Elsas et al., 2012), the presence of *Yersinia* in a sample was negatively associated with microbial diversity (although the relationship here was not significant). The most notable microbial association with *Yersinia* was with *Dictyostelium*, an amoeba that consumes bacteria. Previous experimental work has shown that *Y. pestis* can escape phagocytosis by and replicate within *D. discoideum* for at least 48 h (in comparison with control bacteria, which were consumed within 1 h; Markman et al., 2018). The prevalence of *Dictyostelium* (present in 2 of 158 samples) and another amoeba, Acanthamoeba (10 of 158 samples), in our soils was lower than that recovered in Markman et al. (2018), although the methods of recovery differed. To our knowledge, this is the first study to detect *Yersinia* in soil samples collected from prairie dog colonies. Although we were unable to classify the sequences at the species level due to read length constraints, this suggestive finding adds to the collective evidence that *Y. pestis* is present in prairie dog colonies in the absence of epizootics (3 years after the prairie-dog population die-off) and that soil amoebae may be a potential reservoir for plague in inter-epizootic intervals.

Our results show that variation in soil microbial communities occurs at fine spatial scales in relation to functional partitioning of below-ground space by a social mammalian herbivore. This fine-scale structure likely interacts with mass mortality events, for example by sudden drastic increases in input to certain physical burrow regions (e.g., chambers used for quarantining dead individuals). The existence of fine-scale spatial structure in community diversity in this and other studies suggests that estimates of beta-diversity should account for fine-scale structure in order to accurately estimate the true degree of diversity. Collectively, our results demonstrate how soil microbial communities can interact with animal pathogens (van Elsas et al., 2012; Trivedi et al., 2017) to shape above- and below-ground biodiversity in grasslands.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: https://qiita.microbio.me/emp, 11519; https://www.ebi.ac.uk/ena/browser/view/ERP106314.

AUTHOR CONTRIBUTIONS

LC-S conceived, designed, and oversaw this study. CK and LC-S analyzed the data and wrote the manuscript. Both authors contributed to the article and approved the submitted version.

FUNDING

This project was funded through a crowdfunding initiative hosted by RocketHub. Open Access charges were funded through a startup award to LC-S by the University of Louisiana.

ACKNOWLEDGMENTS

We are grateful to Se Jin Song for guidance and discussions regarding methodology, to Holly Archer for extracting DNA from the soils, and to Tim Seastedt, Stower Beals, Robin Reibold, and Shivani Ehrenfeucht for measuring soil moisture, pH, and CHN. Sample processing, 16S sequencing, and core amplicon data analysis were performed by the Earth Microbiome Project (www.earthmicrobiome.org). This project was funded through a crowdfunding initiative and we are grateful to each funder for making it possible.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fevo.2021.758348/full#supplementary-material

REFERENCES

Aanderud, Z. T., Bahr, J., Robinson, D. M., Belnap, J., Campbell, T. P., Gill, R. A., et al. (2019). The burning of biocrusts facilitates the emergence of a bare soil community of poorly-connected chemoheterotrophic bacteria with depressed ecosystem services. *Front. Ecol. Evol.* 7:467. doi: 10.3389/fevo.2019.00467

Aas, A. B., Andrew, C. J., Blaalid, R., Vik, U., Kauserud, H., and Davey, M. L. (2019). Fine-scale diversity patterns in belowground microbial communities are consistent across kingdoms. *Fems Microbiol. Ecol.* 95:fiz058. doi: 10.1093/femsec/fiz058

Anacker, B., Seastedt, T. R., Halward, T. M., and Lezberg, A. L. (2021). Soil carbon and plant richness relationships differ among grassland types, disturbance history and plant functional groups. *Oecologia* 196, 1153–1166. doi: 10.1007/s00442-021-04992-x

Bai, Y., Kosoy, M. Y., Ray, C., Brinkerhoff, R. J., and Collinge, S. K. (2008). Temporal and spatial patterns of Bartonella infection in black-tailed prairie dogs (*Cynomys ludovicianus*). *Microb. Ecol.* 56, 373–382. doi: 10.1007/s00248-007-9355-6

Bakalidou, A., Kämpfer, P., Berchtold, M., Kuhnigk, T., Wenzel, M., and König, H. (2002). Cellulosimicrobium variabile sp. nov., a cellulolytic bacterium from the hindgut of the termite Mastotermes darwiniensis. *Int. J. Syst. Evol. Microbiol.* 52, 1185–1192. doi: 10.1099/00207713-52-4-1185

Banerjee, S., Schlaeppi, K., and van der Heijden, M. G. (2018). Keystone taxa as drivers of microbiome structure and functioning. *Nat. Rev. Microbiol.* 16, 567–576. doi: 10.1038/s41579-018-0024-1

Barnard, R. L., Osborne, C. A., and Firestone, M. K. (2013). Responses of soil bacterial and fungal communities to extreme desiccation and rewetting. *ISME J.* 7, 2229–2241. doi: 10.1038/ismej.2013.104

Bartelt-Ryser, J., Joshi, J., Schmid, B., Brandl, H., and Balser, T. (2005). Soil feedbacks of plant diversity on soil microbial communities and subsequent plant growth. *Perspect. Plant Ecol. Evol. Syst.* 7, 27–49. doi: 10.1016/j.ppees.2004.11.002

Benjamini, Y., and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Ann. Stat.* 29, 1165–1188. doi: 10.1214/aos/1013699998

Bokulich, N. A., Kaehler, B. D., Rideout, J. R., Dillon, M., Bolyen, E., Knight, R., et al. (2018). Optimizing taxonomic classification of marker-gene amplicon sequences with QIIME 2's q2-feature-classifier plugin. *Microbiome* 6, 1–17. doi: 10.1186/s40168-018-0470-z

Bolyen, E., Rideout, J. R., Dillon, M. R., Bokulich, N. A., Abnet, C. C., Al-Ghalith, G. A., et al. (2019). Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nat. Biotechnol.* 37, 852–857. doi: 10.1038/s41587-019-0209-9

Bray, N., Kao-Kniffin, J., Frey, S. D., Fahey, T., and Wickings, K. (2019). Soil macroinvertebrate presence alters microbial community composition and activity in the rhizosphere. *Front. Microbiol.* 10:256. doi: 10.3389/fmicb.2019.00256

Brockett, B. F., Prescott, C. E., and Grayston, S. J. (2012). Soil moisture is the major factor influencing microbial community structure and enzyme activities across seven biogeoclimatic zones in western Canada. *Soil Biol. Biochem.* 44, 9–20. doi: 10.1016/j.soilbio.2011.09.003

Burns, J. A., Flath, D. L., and Clark, T. W. (1989). On the structure and function of white-tailed prairie dog burrows. *Great Basin Nat.* 49, 517–524.

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

Ciadamidaro, L., Madejón, E., Puschenreiter, M., and Madejón, P. (2013). Growth of Populus alba and its influence on soil trace element availability. *Sci. Tot. Envviron.* 454, 337–347. doi: 10.1016/j.scitotenv.2013.03.032

Cline, L. C., Zak, D. R., Upchurch, R. A., Freedman, Z. B., and Peschel, A. R. (2017). Soil microbial communities and elk foraging intensity: implications for soil biogeochemical cycling in the sagebrush steppe. *Ecol. Lett.* 20, 202–211. doi: 10.1111/ele.12722

Cooke, L. A., and Swiecki, S. R. (1992). Structure of a white-tailed prairie dog burrow. *Great Basin. Natural.* 52:12.

Cregger, M. A., Veach, A. M., Yang, Z. K., Crouch, M. J., Vilgalys, R., Tuskan, G. A., et al. (2018). The Populus holobiont: dissecting the effects of plant niches and genotype on the microbiome. *Microbiome* 6, 1–14. doi: 10.1186/s40168-018-0413-8

Cully, J. F. Jr., Johnson, T. L., Collinge, S. K., and Ray, C. (2010). Disease limits populations: plague and black-tailed prairie dogs. *Vector Borne Zoonotic Dis.* 10, 7–15. doi: 10.1089/vbz.2009.0045

Dedysh, S. N., Beletsky, A. V., Ivanova, A. A., Kulichevskaya, I. S., Suzina, N. E., Philippov, D. A., et al. (2020). Wide distribution of Phycisphaera-like planctomycetes from WD2101 soil group in peatlands and genome analysis of the first cultivated representative. *Environ. Microbiol.* 23, 1510–1526. doi: 10.1111/1462-2920.15360

Delgado-Baquerizo, M., Oliverio, A. M., Brewer, T. E., Benavent-González, A., Eldridge, D. J., Bardgett, R. D., et al. (2018). A global atlas of the dominant bacteria found in soil. *Science* 359, 320–325. doi: 10.1126/science.aap9516

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

Eisen, R. J., Petersen, J. M., Higgins, C. L., Wong, D., Levy, C. E., Mead, P. S., et al. (2008). Persistence of Yersinia pestis in soil under natural conditions. *Emerg. Infect. Dis.* 14, 941–943. doi: 10.3201/eid1406.080029

Faith, D. P. (1992). Conservation evaluation and phylogenetic diversity. *Biol. Conserv.* 61, 1–10. doi: 10.1016/0006-3207(92)91201-3

Garcia, M. O., Templer, P. H., Sorensen, P. O., Sanders-DeMott, R., Groffman, P. M., and Bhatnagar, J. M. (2020). Soil microbes trade-off biogeochemical cycling for stress tolerance traits in response to year-round climate change. *Front. Microbiol.* 11:616. doi: 10.3389/fmicb.2020.00616

Gedeon, C. I., Drickamer, L. C., and Sanchez-Meador, A. J. (2012). Importance of burrow-entrance mounds of Gunnison's prairie dogs (*Cynomys gunnisoni*) for vigilance and mixing of soil. *Southwest. Nat.* 57, 100–104.

Gonzalez, A., Navas-Molina, J. A., Kosciolek, T., McDonald, D., Vázquez-Baeza, Y., Ackermann, G., et al. (2018). QIITA: rapid, web-enabled microbiome meta-analysis. *Nat. Methods* 15, 796–798. doi: 10.1038/s41592-018-0141-9

Gutiérrez, J. L., and Jones, C. G. (2006). Physical ecosystem engineers as agents of biogeochemical heterogeneity. *BioScience* 56, 227–236. doi: 10.1641/0006-3568(2006)056[0227:PEEAAO]2.0.CO;2

Heijboer, A., de Ruiter, P. C., Bodelier, P. L., and Kowalchuk, G. A. (2018). Modulation of litter decomposition by the soil microbial food web under influence of land use change. *Front. Microbiol.* 9:2860. doi: 10.3389/fmicb.2018.02860

Herren, C. M., and McMahon, K. D. (2018). Keystone taxa predict compositional change in microbial communities. *Envviron. Microbiol.* 20, 2207–2217. doi: 10.1111/1462-2920.14257

Hestrin, R., Hammer, E. C., Mueller, C. W., and Lehmann, J. (2019). Synergies between mycorrhizal fungi and soil microbial communities increase plant nitrogen acquisition. *Commun. Biol.* 2, 1–9. doi: 10.1038/s42003-019-0481-8

Holland, E. A., and Detling, J. K. (1990). Plant response to herbivory and belowground nitrogen cycling. *Ecology* 71, 1040–1049. doi: 10.2307/1937372

Hong, S. H., Ham, S. Y., Kim, J. S., Kim, I. S., and Lee, E. Y. (2016). Application of sodium polyacrylate and plant growth-promoting bacterium, Micrococcaceae HW-2, on the growth of plants cultivated in the rooftop. *Int. Biodeterioration Biodegradation* 113, 297–303. doi: 10.1016/j.ibiod.2016.04.018

Hoogland, J. L. (1995). *The Black-Tailed Prairie Dog: Social Life of a Burrowing Mammal*. Chicago, IL: University of Chicago Press.

Ichinohe, T., Pang, I. K., Kumamoto, Y., Peaper, D. R., Ho, J. H., Murray, T. S., et al. (2011). Microbiota regulates immune defense against respiratory tract influenza A virus infection. *Proc. Natl. Acad. Sci. U.S.A.* 108, 5354–5359. doi: 10.1073/pnas.1019378108

Kandeler, E., Kampichler, C., Joergensen, R. G., and Mölter, K. (1999). Effects of mesofauna in a spruce forest on soil microbial communities and N cycling in field mesocosms. *Soil Biol. Biochem.* 31, 1783–1792. doi: 10.1016/S0038-0717(99)00096-6

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

Konopka, A. (2009). What is microbial community ecology? *ISME J.* 3, 1223–1230. doi: 10.1038/ismej.2009.88

Lange, M., Eisenhauer, N., Sierra, C. A., Bessler, H., Engels, C., Griffiths, R. I., et al. (2015). Plant diversity increases soil microbial activity and soil carbon storage. *Nat. Commun.* 6, 1–8. doi: 10.1038/ncomms7707

Lau, J. A., and Lennon, J. T. (2011). Evolutionary ecology of plant–microbe interactions: soil microbial structure alters selection on plant traits. *New Phytol.* 192, 215–224. doi: 10.1111/j.1469-8137.2011.03790.x

Laverock, B., Smith, C. J., Tait, K., Osborn, A. M., Widdicombe, S., and Gilbert, J. A. (2010). Bioturbating shrimp alter the structure and diversity of bacterial communities in coastal marine sediments. *ISME J.* 4, 1531–1544. doi: 10.1038/ismej.2010.86

Leff, J. W., Jones, S. E., Prober, S. M., Barberán, A., Borer, E. T., Firn, J. L., et al. (2015). Consistent responses of soil microbial communities to elevated nutrient inputs in grasslands across the globe. *Proc. Natl. Acad. Sci. U.S.A.* 112, 10967–10972. doi: 10.1073/pnas.1508382112

Li, Y., Adams, J., Shi, Y., Wang, H., He, J. S., and Chu, H. (2017). Distinct soil microbial communities in habitats of differing soil water balance on the Tibetan Plateau. *Sci. Rep.* 7:46407.

Lozupone, C., and Knight, R. (2005). UniFrac: a new phylogenetic method for comparing microbial communities. *Appl. Environ. Microbiol.* 71, 8228–8235. doi: 10.1128/AEM.71.12.8228-8235.2005

Lozupone, C. A., Hamady, M., Kelley, S. T., and Knight, R. (2007). Quantitative and qualitative β diversity measures lead to different insights into factors that structure microbial communities. *Appl. Environ. Microbiol.* 73, 1576–1585. doi: 10.1128/AEM.01996-06

Lucas, D. A., Cahill, T. M., and Marshall, P. A. (2020). A microbiome analysis of soil samples from three abandoned lead-silver mines in the arizona sonoran desert. *J. Arizona-Nevada Acad. Sci.* 49, 1–15. doi: 10.2181/036.049.0101

Markman, D. W., Antolin, M. F., Bowen, R. A., Wheat, W. H., Woods, M., Gonzalez-Juarrero, M., et al. (2018). Yersinia pestis survival and replication in potential ameba reservoir. *Emerg. Infect. Dis.* 24:294. doi: 10.3201/eid2402.171065

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

Metcalf, J. L., Xu, Z. Z., Weiss, S., Lax, S., Van Treuren, W., Hyde, E. R., et al. (2016b). Microbial community assembly and metabolic function during mammalian corpse decomposition. *Science* 351, 158–162. doi: 10.1126/science.aad2646

Metcalf, J. L., Carter, D. O., and Knight, R. (2016a). Microbiology of death. *Curr. Biol.* 26, R561–R563. doi: 10.1016/j.cub.2016.03.042

Mogul, R., Vaishampayan, P., Bashir, M., McKay, C. P., Schubert, K., Bornaccorsi, R., et al. (2017). Microbial community and biochemical dynamics of biological soil crusts across a gradient of surface coverage in the Central Mojave Desert. *Front. Microbiol.* 8:1974. doi: 10.3389/fmicb.2017.01974

Munn, L. C. (1993). "Effects of prairie dogs on physical and chemical properties of soils," in *Management of Prairie Dog Complexes for the Reintroduction of the Black-Footed Ferret*, ed. J. L. Oldemeyer (Washington, DC: US Department of the Interior), 11–17.

Nemergut, D. R., Schmidt, S. K., Fukami, T., O'Neill, S. P., Bilinski, T. M., Stanish, L. F., et al. (2013). Patterns and processes of microbial community assembly. *Microbiol. Mol. Biol. Rev.* 77, 342–356. doi: 10.1128/MMBR.00051-12

Ogwu, M. C., Srinivasan, S., Dong, K., Ramasamy, D., Waldman, B., and Adams, J. M. (2019). Community ecology of Deinococcus in irradiated soil. *Microb. Ecol.* 78, 855–872. doi: 10.1007/s00248-019-01343-5

Perez, C., Dill-Macky, R., and Kinkel, L. L. (2008). Management of soil microbial communities to enhance populations of Fusarium graminearum-antagonists in soil. *Plant Soil* 302, 53–69. doi: 10.1007/s11104-007-9455-6

Prescott, C. E., and Grayston, S. J. (2013). Tree species influence on microbial communities in litter and soil: current knowledge and research needs. *For. Ecol. Manage.* 309, 19–27. doi: 10.1016/j.foreco.2013.02.034

Price, M. N., Dehal, P. S., and Arkin, A. P. (2010). FastTree 2–approximately maximum-likelihood trees for large alignments. *PLoS One* 5:e9490. doi: 10.1371/journal.pone.0009490

R Core Team, (2018). *R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria*. Vienna: R Core Team.

Sackett, L. C., Collinge, S. K., and Martin, A. P. (2013). Do pathogens reduce genetic diversity of their hosts? Variable effects of sylvatic plague in black-tailed prairie dogs. *Mol. Ecol.* 22, 2441–2455 .

Salkeld, D. J., Salathé, M., Stapp, P., and Jones, J. H. (2010). Plague outbreaks in prairie dog populations explained by percolation thresholds of alternate host abundance. *Proc. Natl. Acad. Sci. U.S.A.* 107, 14247–14250. doi: 10.1073/pnas.1002826107

Schumann, P., and Stackebrandt, E. (2015). *Cellulosimicrobium. Bergey's Manual of Systematics of Archaea and Bacteria*. 1–9. doi: 10.1002/9781118960608.gbm00127

Shen, C., Xiong, J., Zhang, H., Feng, Y., Lin, X., Li, X., et al. (2013). Soil pH drives the spatial distribution of bacterial communities along elevation on Changbai Mountain. *Soil. Biol. Biochem.* 57, 204–211. doi: 10.1016/j.soilbio.2012.07.013

Shen, Z., Penton, C. R., Lv, N., Xue, C., Yuan, X., Ruan, Y., et al. (2018). Banana Fusarium wilt disease incidence is influenced by shifts of soil microbial communities under different monoculture spans. *Microb. Ecol.* 75, 739–750. doi: 10.1007/s00248-017-1052-5

Thompson, L. R., Sanders, J. G., McDonald, D., Amir, A., Jansson, J. K., Gilbert, J. A., et al. (2017). A communal catalogue reveals Earth's multiscale microbial diversity. *Nature* 551, 457–463. doi: 10.1038/nature24621

Treseder, K. K., Marusenko, Y., Romero-Olivares, A. L., and Maltz, M. R. (2016). Experimental warming alters potential function of the fungal community in boreal forest. *Glob. Change Biol.* 22, 3395–3404. doi: 10.1111/gcb.13238

Trivedi, P., Delgado-Baquerizo, M., Trivedi, C., Hamonts, K., Anderson, I. C., and Singh, B. K. (2017). Keystone microbial taxa regulate the invasion of a fungal pathogen in agro-ecosystems. *Soil. Biol. Biochem.* 111, 10–14. doi: 10.1016/j.soilbio.2017.03.013

van Elsas, J. D., Chiurazzi, M., Mallon, C. A., Elhottovā, D., Krištůfek, V., and Salles, J. F. (2012). Microbial diversity determines the invasion of soil by a bacterial pathogen. *Proc. Natl. Acad. Sci. U.S.A.* 109, 1159–1164. doi: 10.1073/pnas.1109326109

Vázquez-Baeza, Y., Pirrung, M., Gonzalez, A., and Knight, R. (2013). EMPeror: a tool for visualizing high-throughput microbial community data. *Gigascience* 2:16. doi: 10.1186/2047-217X-2-16

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. Microb.* 73, 5261–5267. doi: 10.1128/AEM.00062-07

Webb, C. T., Brooks, C. P., Gage, K. L., and Antolin, M. F. (2006). Classic flea-borne transmission does not drive plague epizootics in prairie dogs. *Proc. Natl. Acad. Sci. U.S.A.* 103, 6236–6241. doi: 10.1073/pnas.0510090103

Whicker, A. D., and Detling, J. K. (1988). Ecological consequences of prairie dog disturbances. *BioScience* 38, 778–785. doi: 10.2307/1310787

Wilcomb, M. J. (1954). *A Study of Prairie Dog Burrow Systems and the Ecology of Their Arthropod Inhabitants in Central Oklahoma*. Doctoral dissertation. Oklahoma, OK: University of Oklahoma.

Xi, N., Chu, C., and Bloor, J. M. (2018). Plant drought resistance is mediated by soil microbial community structure and soil-plant feedbacks in a savanna tree species. *Environ. Exp. Bot.* 155, 695–701. doi: 10.1016/j.envexpbot.2018.08.013

Xia, Z., Yang, J., Sang, C., Wang, X., Sun, L., Jiang, P., et al. (2020). Phosphorus reduces negative effects of nitrogen addition on soil microbial communities and functions. *Microorganisms* 8:1828. doi: 10.3390/microorganisms8111828

Zak, D. R., Holmes, W. E., White, D. C., Peacock, A. D., and Tilman, D. (2003). Plant diversity, soil microbial communities, and ecosystem function: are there any links? *Ecology* 84, 2042–2050. doi: 10.1890/02-0433

Zhuang, W., Yu, X., Hu, R., Luo, Z., Liu, X., Zheng, X., et al. (2020). Diversity, function and assembly of mangrove root-associated microbial communities at a continuous fine-scale. *NPJ Biofilms Microbiomes* 6, 1–10. doi: 10.1038/s41522-020-00164-6

Zotti, M., De Filippis, F., Cesarano, G., Ercolini, D., Tesei, G., Allegrezza, M., et al. (2020). One ring to rule them all: an ecosystem engineer fungus fosters plant and microbial diversity in a Mediterranean grassland. *New Phytol.* 227, 884–898. doi: 10.1111/nph.16583

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

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

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