Systematics of Snailfishes (Family Liparidae): Description of Two New Species from the Aleutian Islands and Investigation of the Utility of Exon Capture Techniques

Jennifer Rose Gardner

A thesis

submitted in partial fulfillment of the

requirements for the degree of

Master of Science

University of Washington

2022

Committee:

Luke Tornabene

Lorenz Hauser

James Orr

Ingrid Spies

Program Authorized to Offer Degree:

Aquatic and Fishery Sciences

University of Washington

Abstract

Systematics of Snailfishes (Family Liparidae): Description of Two New Species from the Aleutian Islands and Investigation of the Utility of Exon Capture Techniques

Jennifer Rose Gardner

Chair of the Supervisory Committee:

Luke Tornabene

School of Aquatic and Fishery Sciences

The snailfishes of the family Liparidae are a speciose and geographically widespread family of fishes residing in cold and temperate waters worldwide in both the shallow and deep ocean. Characterized by soft, scaleless, gelatinous, frequently slimy bodies, snailfishes are often difficult to work with both morphologically, as specimens are rarely able to be collected in pristine condition, as well as genetically, as there are known difficulties with extracting sufficient high quality DNA from snailfish tissues. Due to these difficulties, as well as generally conserved morphology with differences among species often based on reductive characters, phylogenetic relationships within the family are poorly understood at multiple levels. One major challenge is that there are many undescribed species, and some species complexes require taxonomic revisions. This study addresses taxonomic revision of one group of snailfishes through the

descriptions of two new snailfish species from the Aleutian Islands. Both are similar to *Careproctus candidus*, described originally from four specimens, which is redescribed here on the basis of 67 specimens in addition to the type specimens. All are allocated to the subgenus *Temnocora*. At the genus level relationships are also poorly understood, as the most recent family-wide phylogenetic analyses recover all three of the most speciose genera as paraphyletic. Most phylogenetic hypotheses for the Liparidae have been inferred based on only morphological characters, only mitochondrial gene sequences, or a combination of the two. Only one phylogeny has been inferred using genome-wide sequences. This study attempts to add to the understanding of the Liparidae phylogeny by inferring a second phylogeny based on sequences taken from across the genome using targeted exon capture. The generated exon capture dataset had widespread cross-contamination. Thus, multiple methods for detecting and attempting to remove that contamination were explored. Methods included comparing topology of well supported published trees to trees inferred based on this dataset, comparing *p*-distances based on *a priori* hypotheses coupled with iterative removal of the most contaminated specimens, and using locus heterozygosity to detect contamination. All methods confirmed the presence of widespread contamination in the dataset; however, it was so widespread that it could not be specifically removed. Finally, suggestions are made for how to avoid and or detect and remove contamination in future exon-capture studies.

Table of Contents

List of Figures

List of Tables

ACKNOWLEDGMENTS

Many thanks to my committee members L. Hauser, J. Orr, and I. Spies for their support, assistance, encouragement, many discussions about contaminated data, and critical reviews. Thanks to L. Tornabene for your guidance, assistance, being a great adviser, and letting me go on the adventure that is snailfish research.

We thank AFSC and the collectors of the specimens, including A. Abookire, E. Brown, D. Drumm, W. Flerx, R. Harrison, G. R. Hoff, B. Knoth, N. Laman, P. Logan, M. Martin, W. Palsson, K. P. Maslenikov, N. Roberson, P. Von Szalay (all from the National Marine Fisheries Service, Alaska Fisheries Science Center), G. C. Jensen (University of Washington), and R. N. Clark. We thank J. Huie, T. Buser, K. Kohen, M. Gerringer, and A. Summers for training and assistance with CT scans at Friday Harbor. We thank K. P. Maslenikov for collections support, discussions, and critical review of Chapter 1. Many thanks to L. Eakins for the specimen illustrations for the two new species.

Thanks to C. Atta for immeasurable assistance with benchwork and bioinformatic analysis for Chapter 2, as well as constant support and discussions. Thanks to C. Li and H. Yuan for discussions and assistance with contamination detection in the Assexon pipeline. Thanks to D. Arcila for assembly of data with FishLife pipeline and critical discussions of the nature of data contamination. Thanks to E. Petrou for scripts and immense assistance with variant calling and general heterozygosity analysis. Thanks to E. Miller for assistance and discussion with bioinformatic analysis and making tree figures. Thanks to S. Ghods for doing many of the snailfish extractions for previous work, and for support during the final quarter of this work.

None of this would have been possible without my amazing and supportive friends and family. Mom and Dad, thanks for always being supportive and encouraging and letting me

practice explaining the concepts even if you don't know what I'm talking about. Jamie, Rob, and Evie thanks for having the uncanny ability to always call about something completely random when I'm most in need of a break. S. Thurner, G. Thurner, and Y. Venkataraman (TC) thanks for always being there to answer all sorts of random questions and help me troubleshoot every aspect of the graduate school experience. Thanks to the SAFS graduate student body as whole for providing an amazing community to work in and always striving to make things better for the next cohort. Gratitude to all my friends who always reminded me to find fun outside of work, but also always helped to remind me why I love doing this research. Thanks to T. Swift, Y. Ma, and BiT for providing the soundtrack to countless hours of lab work and writing. Finally, infinite thanks to A. Skinner for always being there for me in every way possible and for letting me drive us both slightly mental while working six feet apart for two years.

This project was funded by the William W. and Dorothy T. Gilbert Ichthyology Research Fund, by the oVert (openVertebrate) Thematic Collection Network (TCN), NSF DBI-1701665, by a fellowship from the University of Washington School of Aquatic and Fishery Sciences, and by the Cooperative Institute for Climate, Ocean and Ecosystems (CICOES).

DEDICATION

This thesis is dedicated to my parents. You always taught me that anything was possible academically if I tried my best, and supported my theory that As are my best. Sláinte and Salud!

General Introduction

The family Liparidae (Scopoli, 1777), the snailfishes, is one of the most geographically widespread families of fishes, residing in cold marine waters worldwide (Chernova et al., 2004). They have the greatest depth range of any marine fish family; the deepest known living vertebrate is a snailfish filmed at 8,143 m (Gerringer et al., 2017), but other members of the family reside at intertidal depths (Chernova et al., 2004). As is expected from such wide-ranging geography and bathymetry, this family is extremely diverse with more than 400 recognized species in 32 genera (Chernova et al., 2004; Fricke et al., 2021). Evolutionarily, Liparidae is a relatively young (Near et al., 2013) and rapidly diversifying family (Rabosky et al., 2018) that appears to have originated in the North Pacific and then spread to cold and temperate waters worldwide (Knudsen, 2007).

Snailfish bodies are roughly tadpole shaped with larger heads that often taper to laterally compressed bodies posteriorly, with scaleless, soft skin, and often a gelatinous layer thought to assist with buoyancy (Chernova et al., 2004). The group is characterized by having pelvic fins modified to form a sucking disc, which may be lost or reduced in pelagic or deep-sea species. They possess a well-developed cephalic pore system but usually lack a lateral line canal, although many have free neuromasts either in place of the lateral line or scattered over the body. Most species do not get larger than 20–30 cm and there are many smaller species with maximum sizes of 10 cm or less.

Of the 32 genera of snailfishes, only four are represented by more than five species. *Paraliparis* is the most specious currently with 151 described species, followed by *Careproctus* with 134 species, *Liparis* with 69 species, and *Psednos* with 35 species (Fricke et al., 2021). Identifying species to the four most speciose genera is based on the number of nostrils per side,

presence and size of pelvic disc, and presence or absence of pseudobranchs. Species of *Liparis* (Scopoli, 1777) have two nostrils per side, a large pelvic disc, and pseudobranchs present. Species of *Careproctus* (Kroyer, 1862) have one nostril per side, pelvic disc present, ranging from minute to large, and pseudobranchs absent. *Paraliparis* (Collett, 1879) is diagnosed by having one nostril per side, pelvic disc absent, and pseudobranchs absent. *Psednos* (Barnard, 1927) appears similar to *Paraliparis* with absence of disc and single nostril per side, but has a unique S-curved vertebral column, and unique cephalic pore configuration (Chernova, 2001). Despite being one of the more speciose genera, *Psednos* species are extremely rare in museum collections, with 19 of the 35 species known only from type specimens (Fricke et al., 2021). The remaining 28 genera of Liparidae are each represented by five or fewer species each. The smaller genera tend to have very distinct diagnostic characters separating them from other genera such as unique jaw morphology, unique hard plates on the skull, and papillae on the cephalic pores, among others.

Among the 32 genera and over 400 currently described species, the family Liparidae itself presents a potential unique study system for looking at many different evolutionary and ecological questions. This includes investigating questions about deep sea invasion and adaptation (Gerringer, 2019), rapid diversification in cold and temperate waters (Rabosky et al., 2018), and unique reproduction strategies including laying their eggs in or on other organisms, both sessile organisms like the octocorals (Busby et al., 2006), sponges (Chernova, 2014) and xenophyophores (Levin and Rouse, 2020), and highly mobile organisms such as king crabs (Gardner et al., 2016). Preliminary work has also revealed some fascinating and potentially unique features that could have important ecological and evolutionary significance. The family may be unique among teleosts in having many species with a contractile iris (L. Britt, pers.

comm.). Some shallow species of *Liparis* also fluoresce under UV light and may possess species-specific fluorescence patterns (Gruber et al., 2021). But the complicating issues with current Liparidae taxonomy and phylogeny make studying the evolutionary trends of these traits more difficult.

Intrafamilial relationships of genera have been poorly resolved based on morphology alone. Snailfish morphology is notoriously difficult to study because they are characterized by gelatinous, scaleless bodies that are easily damaged (Burke, 1930; Chernova et al., 2004). Even pristine specimens can be difficult to work with as their skin becomes loose and their bodies very soft when preserved. This poses challenges in identification of individual specimens to species level, and thus many published species records (museum specimens and DNA sequences) have been incorrectly identified (Orr et al., 2019). These misidentifications make using resources such as sequences from repositories like GenBank and BOLD difficult if there is not a way to verify the identification of the voucher specimen. Additionally, morphology-based classifications are hampered by the fact that much of the classification within the family is based on reductive characters. While character absences can be useful in identification of individual species, they are problematic phylogenetically as characters may have been independently lost at different times throughout the evolutionary history of many species but the loss appears the same morphologically.

While genetic data are useful in uncovering independent losses of characters and recovering phylogenies, there are difficulties with genetic studies of the family as well. Obtaining high quality genetic material from snailfishes is more difficult than one would expect based on similar groups (Gardner et al., 2016). Even relatively fresh tissues often result in relatively short fragments and low quantity DNA extractions. Some of this difficulty in

extracting tissue has been hypothesized to be due to mucopolysaccharides present in snailfish slime (Orr et al., 2019), but it has also been posited that current collection methods often result in degradation of their relatively soft tissue after capture but before fixation (K. Maslenikov, Pers. Comm). Another hindrance to studies trying to resolve the relationships of the entire family is the very large geographic range and number of species paired with the rarity of certain species within the family making it logistically difficult if not impossible for reviews to encompass all, or even most, major species and genera. For this reason, most studies are focused on species from a particular region (Orr et al., 2019).

Despite the difficulties in working with snailfish, they are an evolutionarily fascinating family that deserves further study. Orr et al. (2019) inferred a phylogeny based on mitochondrial (COI) and nuclear markers from restriction enzyme associated DNA sequences (RADseq). The most comprehensive genetic study to date, this was the first time nuclear markers have been used for phylogenetic reconstruction in Liparidae (Appendix 1). The COI tree was based on 268 individuals from at least 122 species in 18 genera largely from the eastern North Pacific while the RADseq tree was based on 40 specimens from 28 species in 10 genera, entirely from the eastern North Pacific. Despite differences in sampled taxa, both tree topologies largely agreed. Based on these trees there are some major issues in the current systematics of the Liparidae family that prevent the taxonomic classification from reflecting the evolutionary history of the family, as all three of the most specious genera (*Liparis*, *Careproctus*, *Paraliparis*) were recovered as paraphyletic. *Careproctus* is especially problematic, as it contains multiple different clades that may warrant separation into multiple genera; however, morphological characters to diagnose those separate genera have not yet been identified.

In addition to taxonomy not matching the evolutionary history, the relationships of Liparidae species are further muddied by the fact that there is much undescribed diversity within the family, with many new species being described each year. Within the last decade, nearly 80 new Liparidae species have been described (e.g., Chernova, 2014; Stein, 2016; Gerringer et al., 2017; Kai et al., 2018; Orr et al., 2020; Orr, 2021). Among predominantly marine fish families, only the Gobiidae has had more species described in the last decade (Fricke et al., 2021). Due to the difficulties in collecting quality specimens in conjunction with difficulties in morphological description, there are many species still awaiting description (Gerringer, 2019; J. Orr, Pers. Comm.). It has been shown that genetic markers, including COI, can prove helpful in identifying new species (Orr et al., 2020) but suitable genetic material is not always collected with specimens. Routine collection of genetic material during all specimen collections, paired with deposition in museum collections, would greatly enhance the ability to study groups of organisms that aren't easily collected.

This work attempts to address some of the difficulties of snailfish taxonomy and systematics at two different scales. The first chapter addresses one of the many species-level areas of confusion by performing a taxonomic review of one group of *Careproctus*, specifically, the subgenus *Temonocora* as defined by Orr et al. (2019). The first species to be described from this group was *Careproctus candidus* (Gilbert and Burke, 1912), which was subsequently reallocated to the monotypic genus *Temnocora* (Burke, 1930). Both of these descriptions were based on the four type specimens. Here, *C. candidus* was redescribed on the basis of an additional 67 specimens. Additionally, two new species which appear morphologically most similar to other members of the subgenus *Temnocora* are described (Orr et al., 2019).

Secondly, the second chapter attempts to provide new phylogenetic data for the family as a whole through the use of targeted exon capture. Targeted exon capture is one of a variety of methods that aim at sequencing independent loci across the genome. Other methods include sequencing of ultra-conserved elements (UCEs; Lemmon et al., 2012) or sequencing other highly conserved loci referred to as "anchor regions" in a method called "anchored hybrid enrichment" (AHE; McCormack et al., 2013). Sequencing of single-copy exons is especially useful for phylogenetic analysis in groups without published genomes as conserved exon markers have been developed that will work across the diversity of the Actinopterygii (Jiang et al., 2019; Hughes et al., 2021). In this study we attempted to infer a phylogeny of Liparidae from exon-capture data to provide a second phylogeny of Liparidae based on genome-wide sequencing in an attempt to better understand intrafamilial relationships (Appendix 1).

Preliminary phylogenetic analysis of the exon-capture dataset generated in this study indicated that the dataset may have widespread cross-contamination. This was not unexpected as all of the above target capture methods have a drawback of requiring benchwork that increases the likelihood of cross-contamination when compared to traditional PCR, and which then may be more difficult to identify and remove from a next generation sequence (NGS) dataset when compared to Sanger sequencing. While contamination in an NGS dataset may be more difficult to identify, it can be explicitly looked for in a few different ways, including comparing tree topologies with published phylogenies, *p*-distance comparisons of related versus less related individuals within the dataset, and determining heterozygous sites across the dataset to investigate suspicious allele frequencies that would indicate the presence of multiple individuals in a single putative sample. When applied to our dataset all methods confirmed the presence of widespread contamination; however, it was so widespread that it could not be specifically

removed. We provide recommendations for addressing issues that may be encountered when attempting next generation sequencing in Liparidae and considerations for future sequencing work in Liparidae.

Chapter 1: Two New Species of Snailfishes (Cottiformes: Liparidae) from the Aleutian Islands, Alaska, and a Redescription of the Closely Related *Careproctus candidus*

Jennifer R. Gardner[1], James W. Orr[1], and Luke Tornabene[1]

[1] School of Aquatic and Fishery Sciences and Burke Museum of Natural History and Culture, University of Washington, Seattle, USA

ABSTRACT

Two new species of snailfishes are described from the Aleutian Islands. Both are similar to *Careproctus candidus*, described originally from four specimens, here redescribed on the basis of 67 specimens in addition to the types. All are allocated to the subgenus *Temnocora*. *Careproctus candidus* is distinguished from all other species of *Careproctus* by the combination of a slit pupil, an anterior dorsal-fin lobe formed by a distinct notch, and variegated reddish coloration. It is redescribed to note the presence of prickles in some specimens, a variable pectoral girdle morphology, and variable exsertions of the fin rays of the dorsal-fin lobe. *Careproctus* sp. K, new species, is distinguished by having a slit pupil, anterior dorsal fin uniquely of five to seven short anterior dorsal-fin rays nearly equal in length preceding distinctly longer succeeding rays, variegated reddish coloration, and a relatively deep body. *Careproctus* sp. M, new species, is distinguished by having a slit pupil, low vertebrae and median fin ray counts, variegated light brown and pale coloration, and the absence of an anterior dorsal-fin lobe. *Careproctus candidus* ranges across the Aleutian Islands from Attu Island to the northern Gulf of Alaska at depths from 88 to 432 m. *Careproctus* sp. K is found in the central and western Aleutian Islands from off Buldir Island in the west to Amlia Island in the east at depths from 134 to 180 m. *Careproctus*

sp. M is found in the central and eastern Aleutian Islands from Amchitka Pass to off Akutan Island at depths from 80 to 232 m.

INTRODUCTION

The snailfishes, family Liparidae, are a wide-spread and speciose family of fishes comprising 32 genera and over 440 species (Fricke et al. 2021) that reside in temperate and cold waters worldwide. The most species-rich genus is *Careproctus*, which has about 140 species worldwide (Orr et al., 2019), about 50 of which are found in the North Pacific. *Careproctus* is characterized by a distinct pelvic disk, single nostril, pseudobranchs absent, pectoral-fin rays typically fewer than anal-fin rays, and body color that is not variegated except in a few species (Orr and Maslenikov, 2007).

Surveys conducted by the U.S. National Marine Fisheries Service, Alaska Fisheries Science Center (AFSC), have provided materials leading to descriptions of many new liparid species from Alaska (Orr, 2004, 2012, 2016, 2021; Orr and Busby, 2001, 2006; Orr and Maslenikov, 2007), as well as material for genetic analysis (Orr et al., 2019). The two new species described here, as well as all specimens other than extant types for the species redescribed in this paper, were collected during these surveys from 2000 to 2018 in the Aleutian Islands along the length of the entire chain as well as in the northern Gulf of Alaska. The two new species are relatively small (<110 mm SL) and easily missed in the large benthic otter trawls used in these surveys. Both species appear similar to *Careproctus candidus* and other members of the subgenus *Temnocora* of Orr et al. (2019), which included *C. candidus, C. comus, C. faunus, C. iacchus,* and *C. staufferi*. While this subgenus was based solely on phylogenetic relatedness as inferred from mitochondrial COI and whole genome RADseq, the members share the unique trait of variegated live coloration that is not seen in any other members of the genus

Careproctus. Additionally, all except *C. comus* and *C. faunus* have pupils reduced to a horizontal ellipse or horizontal slit. In this paper, we provide diagnoses, descriptions, and distributions of two new species, a redescription of *C. candidus* based on new material, and comparative remarks on other species of the *Temnocora* clade.

MATERIALS AND METHODS

All material examined was obtained during bottom trawl operations conducted by the AFSC. The ‘‘benthic bag,’’ noted in Material Examined, is a small net attached to the footrope of the main net, designed to collect small specimens often missed in the larger main net (Orr, 2004). Counts, measurements, and descriptive terminology follow Orr and Maslenikov (2007) as modified by Orr (2016). Body depth at midpoint of anal fin was taken at half the distance from the anal-fin origin to the end of the hypural plate. For the type specimens of *Careproctus candidus* only, measurements of body depth at dorsal-fin origin and body depth at midpoint of anal fin were taken from photographs. Counts of median-fin rays and vertebrae were taken from x-rays for all specimens. We took micro computed-tomography (uCT) scans from 69 specimens at the Karel F. Liem Bioimaging Facility at the University of Washington Friday Harbor Laboratories, on a Bruker Skyscan 1173. Each fish was scanned using a 1 mm aluminum filter, at 65 kV, 123 μA, 1150 or 1170 ms exposure time, and with a voxel size between 13.4 and 34.1 μm. Voxel sizes were chosen to maximize tradeoffs between scan time and optimal resolution. All scans were uploaded online and made publicly accessible on Morphosource.org (media #s). The right gill membrane and abdomen were cut to examine the branchial and visceral cavities, except for the holotypes; right pectoral girdles were dissected and cleared and counterstained following Potthoff (1984) with a modified concentration (0.1%) of potassium hydroxide to prevent damage

to fragile girdles. The dorsoanterior hook is an extension of the basal cartilage connected to the cleithrum. Counts of gill rakers were taken from the first gill arch on the right side. The holotype of *C. candidus* was not dissected. Lengths are presented as standard length (SL), unless otherwise indicated as total length (TL), and proportions as percent SL, unless otherwise indicated as percent head length (HL), orbit length (OL), or caudal length (CL). Bony interorbital width is the narrowest bony width; greatest interorbital width is taken at the same point but including tissue extending to the orbital rim. Measurements and counts are presented in species accounts as the range for all material examined followed by the value for the holotype in parentheses when intraspecific variation is indicated and when data for the holotype was available. Sex and maturity of specimens was defined as following: ripe females contained large yolked eggs, mature females contained developed opaque eggs; immature females contained only small undeveloped white eggs less than 1 mm in diameter. Right ovaries were dissected from ripe females to count and measure eggs. Not all ripe females had right ovaries dissected. Ripe males had large swollen white testes; mature males had discernible but not enlarged testes; immature males had discernible small testes. Institutional abbreviations follow Sabaj (2020).

***Careproctus candidus* Gilbert and Burke 1912**

Bigeye Snailfish

Figures 1, 4, 7A, 7B, 8A, 8B, 9; Table 1, 2

Careproctus candidus Gilbert and Burke 1912b:77, fig. 22 (original description). Type locality: Aleutian Islands, off Attu Island, 135 fms depth.—Jordan et al., 1930:403 (checklist, Aleutian Islands).—Soldatov and Lindberg, 1930:378 (description, in key, possible in Russian waters).—Böhlke, 1953:135 (type catalog).—Kido, 1988:128 (reallocated from *Temnocora*).—Robins et

al., 1991:44 (common name).—Mecklenburg et al. 2002:605 (Alaska, illustration, description, key).—Nelson et al., 2004:124 (common names).—Love et al., 2005:101 (checklist).—Page et al., 2013:127 (common names).—Orr et al., 2019:13 (molecular phylogenetics).—Orr 2021:465 (compared with *C. maslenikovae* and *C. spiraki*).

Temnocora candida: Burke 1930:146 (new combination, description, key).—Taranetz, 1937:137 (checklist, Bering Sea).—Andriashev, 1939a:47 (not seen, cited in Parin et al. 2002, 2014).—Wilimovsky, 1954:287 (checklist, Bering Sea).—Wilimovsky, 1958:79 (Alaska, Bering Sea, in key).—Quast and Hall, 1972:32 (checklist, Bering Sea and Cook Inlet).—Fedorov, 1973a:21 (not seen, cited in Parin et al. 2002, 2014).—Fedorov, 1973b:67 (Bering Sea).—Matarese et al. 1989:475, table 1 (meristics).—Borets, 2000:98 (not seen, cited in Parin et al. 2002, 2014).—Sheiko and Fedorov 2000:33 (Russia).—Chernova et al., 2004:51 (checklist).—Orr and Maslenikov, 2007:708, figs. 1C, 2C (compared with *C. comus* and *C. faunus*).—Datsky, 2015:808 (western Bering Sea).—Orr 2021 (compared with *C. maslenikovae* and *C. spiraki*).—Fricke et al., 2021: online (valid).

Temnocara candida: Eschmeyer et al., 1998:2147, 2475 (*lapsus calami*).—Parin et al., 2002:S128 (northern Kuril Islands, southeastern Kamchatka, possibly off Commander Islands).

Holotype.—USNM 74384, 60.7 mm, Aleutian Islands, off Attu Island, 52.9278°N, 173.43°E, 247 m depth, *Albatross* Station 4784, 11 June 1906.

Paratypes.—2 specimens, 56.3–63.6 mm. USNM 74527, collected with holotype.

Non-Type Material.—67 specimens, 27–94.7 mm. Aleutian Islands: UW 48241, 2, 60.5–70.5 mm, 52.3975°N, 171.831°W, 271 m depth, F/V *Dominator*, cruise 2000-01, haul 53, benthic bag, 31 May 2000, J. W. Orr; UW 111862, 66 mm, 52.5987°N, 169.3605°W, 231 m depth, F/V *Morning Star*, cruise 2001-01, haul 1, benthic bag, 20 May 2001, J. W. Orr; UW 111864, 73.9 mm, 52.0722°N, 175.274°W, 88 m depth, F/V *Vesteraalen*, cruise 2000-01, haul 104, benthic bag, 15 June 2000, W. C. Flerx; UW 111870, 77.9 mm, 52.3430°N, 179.8847°W, 159 m depth, F/V *Dominator*, cruise 1997-01, haul 164, 21 July 1997; UW 113665, 4, 55.5–74.7 mm, 52.1634°N, 179.4879°E, 242 m depth, F/V *Vesteraalen*, cruise 2002-01, haul 176, benthic bag, 13 July 2002, R. N. Clark; UW 113666, 69.7 mm, 52.2748°N, 172.9359°W, 291 m depth, F/V *Vesteraalen*, cruise 2000-01, haul 65, benthic bag, 3 June 2000; UW 113698, 73 mm, 52.2754°N, 170.5991°W, 235 m depth, F/V *Sea Storm*, cruise 2004-01, haul 34, 14 June 2004, J. W. Orr; UW 113701, 67.7 mm, 51.9059°N, 178.2489°E, 249 m depth, F/V *Vesteraalen*, cruise 2002-01, haul 152, benthic bag, 6 July 2002, R. N. Clark; UW 113703, 70 mm, 53.7028°N, 167.3344°W, 88 m depth, F/V *Morning Star*, cruise 2002-01, haul 5, 27 May 2002, J. W. Orr; UW 113705, 2, 58.7–62.0 mm, 51.4255°N, 178.8431°E, 197 m depth, F/V *Gladiator*, cruise 2004-01, haul 160, benthic bag, 15 July 2004, G. C. Jensen; UW 113711, 66.8 mm, 51.6034°N, 176.4098°W, 237 m depth, F/V *Sea Storm*, cruise 2002-01, haul 87, 3 July 2002, J. W. Orr; UW 113712, 67.6 mm, 51.6151°N, 178.445°W, 149 m depth, F/V *Vesteraalen*, cruise 2002-01, haul 114, benthic bag, 24 June 2002, R. N. Clark; UW 117177, 90.5 mm, 52.3675°N, 171.2463°W, 320 m depth, F/V *Gladiator*, cruise 2006-01, haul 39, benthic bag, 16 June 2006, J. W. Orr; UW 117243, 3, 53–80 mm, 52.6466°N, 170.2027°W, 229 m depth, F/V *Gladiator*, cruise 2006-01, haul 35, benthic bag, 15 June 2006, J. W. Orr; UW 117558, 66 mm, 52.9288°N, 170.4111°W,

221 m depth, F/V *Sea Storm*, cruise 2006-01, haul 29, 14 June 2006, B. Knoth; UW 119399, 65.4 mm, 52.1428°N, 175.2807°W, 208 m depth, F/V *Gladiator*, cruise 2006-01, haul 66, benthic bag, 24 June 2006, J. W. Orr; UW 154479, 54.4 mm, 54.2021°N, 166.1854°W, 164 m depth, F/V *Sea Storm*, cruise 2010-01, haul 8, 11 June 2010, J. W. Orr; UW 154876, 4, 32-78 mm, 52.0271°N, 177.7521°W, 192 m depth, F/V *Sea Storm*, 2014-01, 153, 17 July 2014; UW 154881, 72.5 mm, 52.036°N, 179.3053°E, 277 m depth, F/V *Sea Storm*, cruise 2012-01, haul 137, 12 July 2012, W. C. Flerx; UW 155474, 94.7 mm, 52.3431°N, 170.7108°W, 223 m depth, F/V *Sea Storm*, cruise 2002-01, haul 199, 3 August 2002, R. C. Harrison; UW 155477, 2, 81.3–85.1 mm, 52.4024°N, 171.8333°W, 270 m depth, F/V *Sea Storm*, cruise 2002-01, haul 213, 8 August 2002, R. C. Harrison; UW 155512, 62.5 mm, 52.6987°N, 172.711°E, 169 m depth, F/V *Vesteraalen*, cruise 2000-01, haul 209, 14 July 2000, E. S. Brown; UW 155513, 72.1 mm, 52.6072°N, 172.8648°E, 206 m depth, F/V *Vesteraalen*, cruise 2000-01, haul 207, 14 July 2000, E. S. Brown; UW 155514, 70.5 mm, 52.3519°N, 174.5369°W, 109 m depth, F/V *Sea Storm*, cruise 2002-01, haul 4, benthic bag, 13 June 2002, K. E. Pearson; UW 155515, 57.2 mm, 52.4212°N, 170.2785°W, 211 m depth, F/V *Dominator*, cruise 2000-01, haul 43, benthic bag, 29 May 2000, J. W. Orr; UW 155517, 61.6 mm, 52.6475°N, 170.2191°W, 226 m depth, F/V *Dominator*, cruise 2000-01, haul 40, benthic bag, 27 May 2000; UW 156091, 2, 62–70.5 mm, 52.4766°N, 170.1449°W, 183 m depth, F/V *Sea Storm*, cruise 2012-01, haul 44, 18 June 2012, D. Drumm; UW 156094, 59 mm, 54.5625°N, 159.0424°W, 133 m depth, F/V *Alaska Provider*, cruise 2013-01, haul 72, 15 June 2013, P. Von Szalay; UW 156095, 71 mm, 53.0181°N, 170.3998°W, 297 m depth, F/V *Sea Storm*, cruise 2012-01, haul 31, 13 June 2012, D. Drumm; UW 157674, 71.2 mm, 51.3651°N, 178.9066°E, 165 m depth, F/V *Alaska Provider*, cruise 2016-01, haul 135, benthic bag, 8 July 2016; UW 157675, 67.2 mm, 51.4295°N, 178.8275°E, 214 m

depth, F/V *Alaska Provider*, cruise 2016-01, haul 138, benthic bag, 9 July 2016; UW 158239, 80 mm, 52.9358°N, 170.4509°W, 250 m depth, F/V *Ocean Explorer*, cruise 2012-01, haul 37, 17 June 2012; UW 158243, 58.7 mm, 52.2657°N, 170.8021°W, 226 m depth, F/V *Ocean Explorer*, cruise 2012-01, haul 45, 18 June 2012; UW 158244, 63.5 mm, 52.9825°N, 170.3857°W, 218 m depth, F/V *Ocean Explorer*, cruise 2012-01, haul 36, 17 June 2012; UW 158245, 59.1 mm, 52.2549°N, 173.156°W, 138 m depth, F/V *Ocean Explorer*, cruise 2012-01, haul 78, 26 June 2012; UW 158246, 65.5 mm, 58.7219°N, 151.9983°W, 146 m depth, F/V *Morning Star*, cruise 2001-01, haul 264, 22 July 2001; UW 158247, 79.1 mm, 52.288°N, 170.7933°W, 219 m depth, F/V *Ocean Explorer*, cruise 2012-01, haul 44, 18 June 2012; UW 158253, 51.7 mm, 51.8588°N, 178.4628°E, 218 m depth, F/V *Sea Storm*, cruise 2012-01, haul 147, 15 July 2012; UW 158291, 2, 73.8–78 mm, 52.1823°N, 172.4394°W, 190 m depth, F/V *Sea Storm*, cruise 2012-01, haul 75, 25 June 2012, D. Drumm; UW 158398, 64 mm, 52.2118°N, 174.9435°W, 207 m depth, F/V *Sea Storm*, cruise 2006-01, haul 243, 5 August 2006; UW 158411, 59.9 mm, 52.4759°N, 170.7595°W, 258 m depth, F/V *Ocean Explorer*, cruise 2012-01, haul 46, 19 June 2012; UW 158412, 2, 64.0–81.2 mm, 51.6261°N, 178.439°W, 178 m depth, F/V *Ocean Explorer*, cruise 2018-01, haul 107, 5 July 2018, N. E. Roberson; UW 159755, 53.3 mm, 52.9582°N, 170.4005°W, 220 m depth, F/V *Morning Star*, cruise 2002-01, haul 51, benthic bag, 4 June 2002; UW 159757, 27 mm, 51.6022°N, 177.5741°W, 226 m depth, F/V *Dominator*, cruise 2000-01, haul 116, benthic bag, 19 June 2000, K. E. Pearson; UW 159759, 67 mm, 53.1303°N, 169.9617°W, 433 m depth, F/V *Vesteraalen*, cruise 2000-01, haul 35, 26 May 2000; UW 159761, 2, 85–89 mm, 51.8454°N, 173.8995°W, 246 m depth, F/V *Sea Storm*, cruise 2004-01, haul 78, 28 June 2004, R. N. Clark; UW 159774, 76.4 mm, 51.8653°N, 178.3737°E, 222 m depth, F/V *Sea Storm*, cruise 2014-01, haul 147, 16 July 2014, G. R. Hoff; UW 159775, 76.8

mm, 51.6249°N, 178.4132°E, 154 m depth, F/V *Sea Storm*, cruise 2014-01, haul 133, 12 July 2014, G. R. Hoff; UW 159776, 66.1 mm, 52.0362°N, 178.0337°E, 198 m depth, F/V *Sea Storm*, cruise 2004-01, haul 140, 10 July 2004, R. N. Clark; UW 159777, 77.6 mm, 52.2307°N, 175.242°E, 146 m depth, F/V *Sea Storm*, cruise 2002-01, haul 138, 16 July 2002, J. W. Orr; UW 159778, 75 mm, 51.6843°N, 175.5556°E, 180 m depth, F/V *Vesteraalen*, cruise 2002-01, haul 158, benthic bag, 7 July 2002, R. N. Clark; UW 159779, 76.5 mm, 52.3032°N, 175.815°E, 243 m depth, F/V *Vesteraalen*, cruise 2002-01, haul 169, benthic bag, 10 July 2002, R. N. Clark.

Diagnosis.—*Careproctus candidus* is distinguished from all described species of *Careproctus* by having the following combination of characters: dorsal fin with anterior lobe of one to three longer rays followed by one to three shortened rays forming a distinct notch, an elliptical or slitted pupil, variegated coloration of reddish pink and white, and prickles often present. It is similar to other species in the subgenus *Temnocora* (Orr et al. 2019) but is further distinguished from those species in the presence of a dorsal-fin lobe formed by a distinct notch (lacking in all other members of the subgenus), as well as in having a lower total vertebral count than *C. comus* and *C. faunus* (46–52 vs. 56–61 and 52–56 for *C. comus* and *C. faunus* respectively). It is most similar to *Careproctus* sp. K, new species, but differs in having a narrower body at anal-fin origin 57.4–89.8 (mean 69.8) % HL (vs. 75.6–100.0 [mean 87.3] % HL in *Careproctus* sp. K) and narrower body at midpoint of anal fin 20.9–42.9 (mean 30.1) % HL (vs. 36.1–53.1 [mean 42.2] % HL in *Careproctus* sp. K), as well as the presence of a distinct notch in the dorsal fin (vs. notch lacking in *Careproctus* sp. K).

Description.—Body robust, deepest at dorsal-fin origin, tapering posteriorly, rounded in cross section anteriorly, compressed posteriorly; depth at dorsal-fin origin 76.5–110.1 (89.4) % HL; depth at anal-fin origin 57.4–89.8 (75.9) % HL; depth at midpoint of anal fin 20.9–42.9 (27.7) % HL. Head large and robust, its length 25.7–31.0 (28.0) % SL, dorsal profile rounded from nape to snout. Snout blunt, slightly projecting anterior to lower jaw. Mouth subterminal, maxilla 34.0–51.7 (40.6) % HL, extending to anterior edge of pupil or to mid-orbit, oral cleft extending to anterior rim of orbit or to a point slightly anterior to orbit. Lower jaw slightly inferior, premaxillary tooth plates matching mandibular tooth plates. Both premaxillary and mandibular teeth trilobed in 8–12 oblique rows of 4–11 teeth. Diastemata absent at symphysis of upper and lower jaws. Orbit large, 31.2–46.6 (31.2) % HL, dorsal margin below dorsal contour of head, suborbital depth to maxilla 15.3–60.4 (60.4) % OL; pupil elliptical to slitted. Interorbital space moderate, fleshy distance 14.7–38.7 (26.5) % HL, bony distance 5.5–16.7 % HL. Snout ranging from shorter than to about equal to orbit length, 56.8–109.4 (109.4) % OL, 24.4–38.3 (34.1) % HL. Nostril single, with well-developed tube at level with middle of orbit; nostril opening flush with snout or in a tube reaching to 7.3 (0) % HL.

Pores of cephalic lateralis of moderate size: nasal pores two, maxillary pores six, preoperculomandibular pores seven, suprabranchial pores two (pore pattern 2-6-7-2); chin pores either single, two in a single pit, or two in two separate pits. Interorbital pore absent. Free neuromasts (Andriashev and Stein, 1998) about 4–5, small and difficult to discern, originating above gill slit and extending to a level at mid-body to a distance about level with anus.

Gill opening small, 14.0–25.3 (19.4) % HL, upper margin level with upper rim of orbit or midorbit, extending ventrally to just above pectoral fin or to pectoral-fin ray one, two, or three. Opercular flap angular. Gill rakers 6–10, short, blunt.

Dorsal-fin rays 37–49 (48; Table 1), all rays simple. Anterior dorsal-fin lobe present, composed of about seven rays: rays one through three or four being longer, rays four or five through seven shortened to form a distinct notch, rays eight and beyond lengthening (Fig. 1.8A). Rays of anterior dorsal-fin lobe variably exserted, ranging from not exserted to exserted for up to half their length (Fig. 1.8A, B). Anteriormost dorsal-fin pterygiophore inserted between neural spines two and three, bearing a single small ray. Anterior rayless pterygiophore absent. Predorsal length 22.9–32.5 (28.8) % SL.

Anal-fin rays 35–41 (40; Table 1), anal-fin pterygiophores anterior to first haemal spine one to three, each bearing a single ray, or absent (one). All rays simple. Anal-fin origin below vertebrae 10-12 (below caudal vertebrae 1–3; below caudal vertebra 2 in holotype), preanal length 34.7–46.9 (37.9) % SL.

Pectoral fin shallowly to moderately notched, with 32–40 rays (35; Table 1). Upper lobe of 23–32 (23) rays extending to anal-fin ray one to seven, dorsalmost rays lengthening to rays 6–8, more ventral rays gradually shortening to shortest ray of notch. Dorsal margin crenulate, fin margins more emarginate ventrally. Lower lobe with 7–12 (12) rays, extending to just past anus or beyond to just before anal–fin origin; dorsal rays of lower lobe gradually lengthening to thicker and fleshy rays 2–3, ventral rays gradually shortening to ventralmost ray near pectoral symphysis. Tips of rays 10–70 % free of membrane, rays of lower lobe more strongly exserted. Rays in notch slightly more widely spaced than rays of lobes. Uppermost pectoral-fin ray level with region between midorbit and ventral rim of orbit.

Proximal pectoral radials four (3+1), robust; proximal radial one notched, radial two notched or rounded, radial three rounded, radial four a rounded square (Fig. 1.7A, B). Interradial fenestrae one or two, always one ovular, extending between scapula and proximal radial one,

when two fenestrae present, second is small, slender, extending between proximal radial two and three (Fig. 1.7B). Scapula broadly Y-shaped with robust helve; coracoid with triangular head and long helve. Dorsoanterior hook present or absent. Distal radials present at base of rays two to 29–34, ventral 6–9 rays articulating directly with fibrocartilage pad.

Pelvic disk moderately large, length 26.4–38.4 (30.0) % HL, round, slightly longer than wide, width 23.2–38.8 (29.4) % HL, anterior lobe moderately developed, flat with margins often slightly upturned, distance from tip of upper jaw to pelvic disc 11.8–20.2 (13.5) % SL. Anus closer to pelvic disk than to anal-fin origin, 4.2–20.2 (11.8) % HL from pelvic disk, distance from tip of upper jaw to anus 24.2–33.2 (25.6) % SL.

Principal caudal-fin rays 10–13 (10; Table 1), dorsal procurrent rays 1 to 3 (2), ventral procurrent rays 0 to 3 (1). Membrane of posterior dorsal-fin rays attached for shorter distance to caudal fin than anal-fin rays: dorsal-fin rays attached to caudal fin 12.9–37.7 (28.6) % CL; anal-fin rays, 19.8–48.2 (19.8) % CL. Depth at base of caudal fin 9.1-19.8 % CL.

Skin thin and somewhat loose in preservation, prickles absent or present. If present prickles often densest in patches on head, on fin membrane of anterior dorsal-fin lobe, along base of median fins, and on caudal peduncle, otherwise widely scattered. Pyloric caeca 9–14, length about 25% HL.

Vertebrae 46–52 (51), precaudal 9–11 (10), caudal 36–42 (41; Table 1). Pleural ribs 1–3 (2), present on vertebrae 8–10, those on vertebrae nine long and slender, others small when present. Hypural plate composed of dorsal and ventral plates divided by slight notch in plate. Single epural present.

Coloration.—Body variegated pink and white in life, head with white blotch at mandibular articulation, white blotches often scattered over body (Fig. 1.1). Median fins with darker reddish-orange to brownish coloring especially posteriorly. Eye greenish gold in life with pupil an elliptical slit. Eye black in preservation. Body pale in preservation, sometimes with scattered melanophores along posterior portions of median fins (Fig. 1.4A). Peritoneum, orobranchial cavity, stomach, intestines, pyloric caeca, and urogenital papilla pale.

Life history.—The largest specimen examined was a ripe female 94.7 mm (UW 155474), yolked egg diameter 2.7 mm. The smallest female with yolked eggs was 62.5 mm (UW 155512). Two smaller, spent females 54.4–62.0 mm were also found. Right ovaries were dissected from two females with yolked eggs; five to eight yolked eggs (1.9–3.7 mm diameter) were present with a roughly equal amount of developed but unyolked eggs (0.6 mm diameter) and with many very small immature eggs present (0.2–0.3 mm diameter). One ripe male was found 73.8 mm (UW 158291). A smaller, spent 54.4 mm male (UW 154479) was also found.

Distribution.—Based on our material, *Careproctus candidus* ranges from Attu Island in the Aleutian Islands to the northern Gulf of Alaska, north of Kodiak Island (Fig. 1.9). Collection depths range from 88 to 433 m. The species is also reported from the western Pacific region off southeastern Kamchatka and the northern Kuril Islands (Sheiko and Fedorov, 2000; Mecklenburg et al., 2002) and to depths of 64 and 518 m (Mecklenburg et al., 2002). Although these records are likely of members of *Temnocora*, we have not confirmed the identification of the specimens as *C. candidus*.

Etymology.—Derived from the Latin *candidus* literally meaning "shining white" or loosely as "transcendently beautiful," likely in reference to the white blotching and variegated reddish coloring of the species.

Remarks.—Three specimens (UW 155512, UW 155513, UW 158398) may represent an undescribed species. All three are ripe or spent females and the two ripe females are the smallest ripe females (62.5–64.0 mm SL) of *C. candidus* examined. They differ from *C. candidus* posteriorly in having a more rapidly tapering profile with a narrower body depth at the midpoint of the anal fin 20.9–23.0 % HL (vs. 23.6–42.9 % HL in *C. candidus senu stricto*) and a shorter predorsal length 77.0–90.2 (vs. 84.5–116.5 % HL). We postpone a description of this species until more material becomes available.

Pectoral girdle morphology varied among specimens. An interradial fenestra between proximal radials two and three may be present or absent (Fig. 1.7A, B). Slight differences in proximal radial shape, extent of distal radials, and presence and size of a hook on the dorsoanterior edge of the girdle were also seen. The girdle identified as *C. candidus* published by Orr and Maslenikov (2007; Fig. 2C, UW 111870) has a fenestra between proximal radials two and three but entirely lacks an anterior hook. However, the identification of this specimen is questionable as it lacks a clear notch in the dorsal fin. Another cleared and stained girdle from a positively identified *C. candidus* (UW 154876 1 of 4) had both the interradial fenestra and a dorsoanterior hook.

In addition to the three specimens above, *C. candidus* exhibited diversity in other characters. The rays comprising the dorsal-fin lobe, anterior to the dorsal notch, were variably exserted, ranging from no exsertion to being exserted for up to half the ray length. The anterior

mandibular chin pores were observed to be either a single pore in a single pit, two pores in the same pit, or two pores in separate pits. Prickles were variably present, with some specimens having prickles widely scattered over the body, some having prickles isolated in distinct, dense patches, some having only a few prickles in one of the patch locations, and some appearing to lack prickles entirely. Previously, prickles were not known to occur in *C. candidus*, however prickles are prone to fall off during capture, preservation, and handling. These ranges of variation in character states are not seen in other members of the subgenus *Temnocora*. Additionally, Orr et al. (2019) reported molecular diversity in cytochrome *c* oxidase subunit I (COI) sequence data among specimens identified as *C. candidus* (fig. 10, appendix table 1), although we found no morphological differences between these specimens. This morphological and genetic diversity suggests unrecognized species diversity may be present within *C. candidus.*

***Careproctus* sp. K, new species**

urn:lsid:zoobank.org:

Ramped Snailfish

Figures 2, 5, 7C, 7D, 8C, 9; Table 1, 2

***Holotype*.—**UW 200421 (out of UW 154881), 105.3 mm, male, Aleutian Islands, Petrel Bank, northeast of Semisopochnoi Island, 52.0360°N, 179.3053°E, 277 m depth, F/V *Sea Storm*, cruise 2012-01, haul 137, 12 July 2012, W. C. Flerx.

***Paratypes.*—**12 specimens, 72.4–106.0 mm: UW 117553, 106.0 mm, 52.0340°N, 176.4665°E, 152 m depth, F/V *Sea Storm*, cruise 2006-01, haul 191, 24 July 2006; UW 117556, 103.5 mm,

52.5192°N, 179.5092°W, 135 m depth, F/V *Sea Storm*, cruise 2006-01, haul 86, 29 June 2006, B. Knoth; UW 117559, 89.7 mm, 52.0537°N, 176.4118°E, 142 m depth, F/V *Sea Storm*, cruise 2006-01, haul 195, 25 July 2006; UW 155354, 86.5 mm, 51.9269°N, 176.6285°E, 149 m depth, F/V *Ocean Explorer*, cruise 2010-01, haul 124, 15 July 2010; UW 155411, 95.5 mm, 52.2281°N, 175.2933°E, 173 m depth, F/V *Sea Storm*, cruise 2004-01, haul 197, 25 July 2004, P. Von Szalay; UW 155491, 88.7 mm, 52.5150°N, 179.5188°W, 134 m depth, F/V *Vesteraalen*, cruise 2002-01, haul 181, benthic bag, 14 July 2002; UAM 4013 (ex UW 155516), 97.7 mm, 52.0874°N, 176.3232°E, 145 m depth, F/V *Vesteraalen*, cruise 2000-01, haul 175, 5 July 2000; UW 155518, 72.4 mm, 51.8447°N, 173.9294°W, 257 m depth, F/V *Vesteraalen*, cruise 2000-01, haul 88, 14 June 2000, W. C. Flerx; SIO 22-1 (ex UW 155519), 92.3 mm, 52.0756°N, 176.3549°E, 149 m depth, F/V *Vesteraalen*, cruise 2000-01, haul 177, 6 July 2000; UW 158344, 90.3 mm, 52.0381°N, 179.3166°E, 280 m depth, F/V *Ocean Explorer*, cruise 2018-01, haul 121, 11 July 2018, N. E. Roberson; UW 159772, 103.6 mm, 52.5150°N, 179.5188°W, 134 m depth, F/V *Vesteraalen*, cruise 2002-01, haul 181, 14 July 2002, A. A. Abookire; USNM 439017 (ex UW 159773), 80.0 mm, 52.3688°N, 179.9117°W, 182 m depth, F/V *Sea Storm*, cruise 2014-01, haul 118, 8 July 2014, G. R. Hoff.

Diagnosis.—*Careproctus* sp. K is a species of *Careproctus* distinguished from described species of *Careproctus* by having the following combination of characters: coloration variegated; anterior dorsal-fin lobe followed by a distinct notch absent; pupil elliptical to slitted; prickles absent; body deep at dorsal-fin origin, anal-fin origin, and the midway point of anal fin; and anus close to pelvic disk. It is most similar to *C. candidus*, with which it shares variegated coloring, an elliptical to slitted pupil, and meristic counts. It can be distinguished from *C. candidus* in having

a unique anterior dorsal-fin configuration in which the dorsal fin is composed of two contiguous sections: the anterior section of six to seven short rays of approximately equal lengths, followed by a long posterior section of rays that are longer than the anterior section and of approximately equal lengths. It also has a deeper body than *C. candidus*: depth at anal-fin origin 75.6–100.0 (mean 87.3) % HL (v 57.4–89.8 [mean 69.8] % HL in *C. candidus*), depth at midpoint of anal fin 36.1–53.1 (mean 42.2) % HL (v 20.9–42.9 [mean 30.1] % HL in *C. candidus*).

Description.—Body robust, deepest at dorsal-fin origin, tapering gradually posteriorly, rounded in cross section anteriorly, compressed posteriorly; depth at dorsal-fin origin 91.8–113.4 (102.0) % HL; depth at anal-fin origin 75.6–100.0 (100.0) % HL; depth at midpoint of anal fin 36.1–53.1 (53.1) % HL. Head large and robust, its length 27.5–29.4 (29.0) % SL, dorsal profile rounded from nape to snout. Snout blunt, slightly projecting anterior to lower jaw. Mouth subterminal, maxilla 33.5–39.8 (36.7) % HL, extending to anterior edge of pupil or to mid-orbit, oral cleft extending to anterior rim of orbit. Lower jaw slightly inferior, premaxillary tooth plates matching mandibular tooth plates. Both premaxillary and mandibular teeth trilobed in 8–13 oblique rows of 5–10 teeth. Diastemata absent at symphysis of upper and lower jaws. Orbit large, 30.4–40.5 (31.5) % HL, dorsal margin below dorsal contour of head, suborbital depth to maxilla 23.5–47.1 (43.8) % OL; pupil elliptical to slitted. Interorbital space moderate, fleshy distance 24.1–32.2 (28.2) % HL, bony distance 10.4–14.4 (10.8) % HL. Snout about equal to orbit length, 74.3–112.8 (112.5) % OL, 29.0–36.2 (35.4) % HL. Nostril single, in well-developed tube at level with middle of orbit; nostril tube length 1.2–4.4 (3.0) % HL.

Pores of cephalic lateralis of moderate size: nasal pores two, maxillary pores six, preoperculomandibular pores seven, suprabranchial pores two (pore pattern 2-6-7-2); chin pores

two in a single pit. Interorbital pore absent. Free neuromasts about 4–5, small and difficult to discern, originating above gill slit and extending to a level at mid-body to a distance about level with anus.

Gill opening small, 14.1–25.9 (25.9) % HL, upper margin level with upper rim of orbit or midorbit, extending ventrally to just above pectoral fin or to pectoral-fin ray one or two (entirely above). Opercular flap angular. Gill rakers 7–10, short, blunt.

Dorsal-fin rays 43–48 (47; Table 1), all rays simple. Anterior dorsal-fin lobe absent, dorsal-fin composed of two contiguous sections: the anterior with ray one shorter than rays two through five to seven, which are nearly equal in length, with rays either not exserted or tips barely exserted from fin membrane; the posterior section starting at rays six to eight with rays longer than those of anterior section and approximately equal in length (Fig. 1.8C). Anterior dorsal-fin often bent in fresh specimens (Fig 2A, C) resulting in the appearance of a notch when one is not actually present, as verified by examination of the preserved specimens. Anteriormost dorsal-fin pterygiophore inserted between neural spines two and three, bearing a single small ray. Anterior rayless pterygiophore present. Predorsal length 24.8–28.5 (27.1) % SL.

Anal-fin rays 37–41 (40; Table 1), anal-fin pterygiophores anterior to first haemal spine one or two (two), each bearing a single ray. All rays simple. Anal-fin origin below vertebrae 10-11 (caudal vertebrae 1), preanal length 38.0–46.0 (38.0) % SL.

Pectoral fin moderately notched, with 36–39 rays (39; Table 1). Upper lobe of 27–30 (30) rays extending to anal-fin ray two to four, dorsalmost rays lengthening to rays 6–8, more ventral rays gradually shortening to shortest ray of notch. Dorsal margin crenulate, fin margins more emarginate ventrally. Lower lobe with 7–9 (9) rays, extending to just past anus or beyond to halfway between anus and anal–fin origin; dorsal rays of lower lobe gradually lengthening to

thicker and fleshy rays 2–3, ventral rays gradually shortening to ventralmost ray near pectoral symphysis. Tips of rays 10–70 % free of membrane, rays of lower lobe more strongly exserted. Rays in notch slightly more widely spaced than rays of lobes. Uppermost pectoral-fin ray level with region between midorbit and ventral rim of orbit.

Proximal pectoral radials four (3+1), robust; proximal radials one through three round, radial four a rounded square (Fig. 1.7C, D). Interradial fenestrae one or two, one always ovular, extending between scapula and proximal radial one; when two fenestrae present, second is very small, slender, extending between proximal radials two and three but not reaching either. Scapula broadly Y-shaped with robust helve; coracoid with triangular head and long helve. Distal radials present at base of all but the first pectoral-fin ray and ventral 7–8 rays, which articulate directly with fibrocartilage pad.

Pelvic disk moderately large, length 29.5–37.1 (31.2) % HL, round, slightly longer than wide, width 26.2–35.9 (33.4) % HL, anterior lobe moderately developed, flat with margins often slightly upturned, distance from tip of upper jaw to pelvic disc 12.6–19.9 (17.2) % SL. Anus closer to pelvic disk than to anal-fin origin, 4.7–14.8 (14.8) % HL from pelvic disk, distance from tip of upper jaw to anus 22.9–28.6 (22.9) % SL.

Principal caudal-fin rays 10–11 (11; Table 1), dorsal procurrent rays 1 to 3 (2), ventral procurrent rays 0 to 2 (1). Membrane of posterior dorsal-fin rays attached for shorter distance to caudal fin than anal-fin rays: dorsal-fin rays attached to caudal fin 19.8–36.4 (28.4) % CL; anal-fin rays, 35.7–54.8 (45.2) % CL. Depth at base of caudal fin 11.9-17.8 (15.5) % CL.

Skin thin and somewhat loose in preservation, prickles absent. Pyloric caeca 17–21, length about 25% HL.

Vertebrae 47–51 (49), precaudal 9–10 (10), caudal 37–42 (39; Table 1). Pleural ribs 1–2 (1), present on vertebrae eight to nine, anteriormost small when two present, others long and slender. Hypural plate composed of dorsal and ventral plates divided by slight notch in plate. Single epural present.

Coloration.—Body variegated pinkish and pale in life, head with white blotch at mandibular articulation, white spots often scattered over body and with darker reddish to brownish along distal margins of median fins posteriorly (Fig. 1.2). Eye greenish dorsally to goldish ventrally in life with an elliptical to slitted pupil. Body pale and vaguely mottled in preservation, eye black (Fig. 1.5A). Peritoneum, orobranchial cavity, stomach, intestines, pyloric caeca, and urogenital papilla pale.

Life history.—Largest specimen examined was an immature female 106.0 mm (UW 117553), with immature eggs 0.4 mm in diameter. All females were maturing or immature, none were ripe with yolked eggs. In maturing females, only opaque, developed but unyolked eggs, and translucent, immature eggs less than 1 mm were present. In immature females, only very small, translucent immature eggs were present. Ten females were maturing or immature, the smallest being 72.4 mm (UW 155518). One male 103.5 mm (UW 117556) was ripe and two males, 92.3–105 mm, were mature.

Distribution.—*Careproctus* sp. K has been collected in the western and central Aleutian Islands from west of Buldir Island and on Petrel Bank to south of Amlia Island in the east (Fig. 1.9). Collection depths ranged from 134 to 280 m.

***Careproctus* sp. M, new species**

urn:lsid:zoobank.org:

Boisterous Snailfish

Figures 3, 6, 7E, 7F, 8D, 9; Table 1, 2

***Holotype.*—**UW 155719, 43.5 mm, ripe female, Aleutian Islands, Seguam Pass, 52.2629°N, 172.8827°W, 158 m depth, F/V *Gladiator*, cruise 2006-01, haul 52, 21 June 2006, J. W. Orr.

***Paratypes.*—**11 specimens, 27–46.9 mm: Aleutian Islands: UW 113670, 27 mm, 51.9376°N, 176.8389°W, 87 m depth, F/V *Dominator*, cruise 2000-01, haul 112, benthic bag, 17 June 2000, K. Pearson; UW 155709, 2, 40.5–42.0 mm, 51.6163°N, 178.1877°W, 126 m depth, F/V *Ocean Explorer*, cruise 2012-01, haul 117, 6 July 2012, J. W. Orr; UAM 4012 (ex UW 158399), 38.1 mm, 51.5981°N, 179.1703°W, 224 m depth, F/V *Dominator*, cruise 2000-01, haul 138, benthic bag, 24 June 2000, K. Pearson; UW 158400, 44.2 mm, 51.5001°N, 178.6552°W, 232 m depth, F/V *Vesteraalen*, cruise 2000-01, haul 123, benthic bag, 21 June 2000, W. C. Flerx; UW 158401, 38.5 mm, 51.5078°N, 178.5773°W, 168 m depth, F/V *Sea Storm*, cruise 2014-01, haul 106, 5 July 2014, G. R. Hoff; UW 158413, 46.9 mm, 52.0997°N, 175.2914°W, 105 m depth, F/V *Alaska Provider*, cruise 2014-01, haul 98, 4 July 2014, P. Logan; UW 158415, 39.0 mm, 51.6083°N, 177.6106°W, 125 m depth, F/V *Ocean Explorer*, cruise 2018-01, haul 97, 3 July 2018, N. E. Roberson; USNM 439016 (ex UW 158416), 38.6 mm, 51.9106°N, 173.7578°W, 119 m depth, F/V *Ocean Explorer*, cruise 2018-01, haul 82, 30 June 2018, N. E. Roberson; UW 158417, 35.1 mm, 51.6177°N, 178.1883°W, 122 m depth, F/V *Ocean Explorer*, cruise 2018-01,

haul 110, 6 July 2018, N. E. Roberson; SIO 22-2 (ex UW 158418), 35.3 mm, 54.0916°N, 166.2622°W, 80 m depth, F/V *Vesteraalen*, cruise 2002-01, haul 26, 23 May 2002.

Diagnosis.—*Careproctus* sp. M is distinguished from described species of *Careproctus* by having the following combination of characters: low counts of median-fin rays (dorsal 38–42, anal 32–37) and vertebrae (44–47), body coloration of variegated light brown and pale with white cheek blotch, anterior dorsal-fin lobe absent, anteriormost dorsal-fin rays very short or embedded in skin, and prickles absent. It is most similar to *C. iacchus* in overall size, variegated body coloring, and in having a white cheek blotch (Kai et al., 2018). It can be distinguished from *C. iacchus* in having 32–37 anal-fin rays (vs. 39–40 in *C. iacchus*), 44–47 vertebrae (vs. 49–52), and a pectoral fin with a shallow notch (vs. no notch). It is similar to *Careproctus staufferi* in meristic counts but can be distinguished based on coloration (variegated brown vs. brick red with golden slash at midbody). It is also similar to *Careproctus candidus*, from which it can be distinguished by lacking a dorsal-fin lobe (present in *C. candidus*), having lower counts of vertebrae (44–47 vs. 46–52), a smaller size at maturity (smallest ripe female 35.3 mm vs. 62.5 mm), and distinct body coloration (variegated brown vs. mottled pink and white).

Description.—Body robust, deepest at nape, tapering posteriorly, rounded in cross section anteriorly, compressed posteriorly; depth at dorsal-fin origin 78.9–107.1 (78.9) % HL; depth at anal-fin origin 46.5–75.4 (62.6) % HL; depth at midpoint of anal fin 25.5–30.7 (27.6) % HL. Head large and robust, its length 25.4–28.3 (28.3) % SL, dorsal profile rounded from nape to snout. Snout blunt, slightly projecting anterior to lower jaw. Mouth subterminal, maxilla 44.6–66.4 (56.1) % HL, extending to anterior edge of pupil or to mid-orbit, oral cleft extending to

anterior rim of orbit. Lower jaw slightly inferior, premaxillary tooth plates matching mandibular tooth plates. Both premaxillary and mandibular teeth trilobed in 8–10 oblique rows of 4–12 teeth. Diastemata absent at symphysis of upper and lower jaws. Orbit large, 31.1–46.5 (33.3) % HL, dorsal margin below dorsal contour of head, suborbital depth to maxilla 16.1–51.5 (24.4) % OL; pupil elliptical to slitted. Interorbital space flat, broad, fleshy distance 23.3–46.5 (41.5) % HL, bony distance 12.3–21.4 (17.9) % HL. Snout about equal to orbit length, 61.7–93.9 (85.4) % OL, 28.0–34.7 (28.5) % HL. Nostril single, in well-developed tube at level with middle of orbit; nostril tube length 1.6–4.5 (4.1) % HL.

Pores of cephalic lateralis of moderate size: nasal pores two, maxillary pores six, preoperculomandibular pores seven, suprabranchial pores two (pore pattern 2-6-7-2); chin pores two in a common wide oval depression, interspace equaling about one pore diameter. Interorbital pore absent. Free neuromasts about 4–5, small and difficult to discern, originating above gill slit and extending to a level at mid-body to a distance about level with anus.

Gill opening small, 14.4–22.8 (21.1) % HL, upper margin level with upper rim of orbit, extending ventrally to just above pectoral fin or to pectoral-fin ray one or two (entirely above pectoral fin). Opercular flap angular. Gill rakers 6–8, small, blunt.

Dorsal-fin rays 38–42 (41; Table 1), all rays simple. Anterior dorsal-fin lobe absent, first few rays very short or embedded in skin, rays longer posteriorly (Fig. 1.8D). Anteriormost dorsal-fin pterygiophore inserted between neural spines three and four, bearing a single small ray. Anterior rayless pterygiophore present, inserted between neural spines two and three. Predorsal length 26.9–31.5 (30.8) % SL.

Anal-fin rays 32–37 (34; Table 1), anal-fin pterygiophore anterior to first haemal spine one, bearing a single ray or absent (one). All rays simple. Anal-fin origin below vertebrae 10–11 (caudal vertebrae 1), preanal length 44.1–49.9 (48.7) % SL.

Pectoral fin shallow to moderately notched, with 30–36 rays (33; Table 1). Upper lobe of 23–29 (26) rays extending beyond anus to anal-fin origin, dorsalmost rays lengthening to rays 6–8, more ventral rays gradually shortening to shortest ray of notch. Dorsal margin crenulate, fin margins more emarginate ventrally. Lower lobe with 7–8 rays, extending to anus or just past; dorsal rays gradually lengthening to thicker and fleshy rays 2–3, ventral rays gradually shortening to ventralmost ray near pectoral symphysis. Tips of rays 10–50 % free of membrane, rays of lower lobe more strongly exserted. Rays in notch slightly more widely spaced than rays of lobes. Uppermost pectoral-fin ray level with region between midorbit and ventral rim of orbit.

Proximal pectoral radials four (3+1), robust; proximal radials one and two either notched and hourglass-shaped or crescent-shaped, radial three crescent-shaped, radial four rounded square (Fig. 1.7E, F). Interradial fenestrae two or three. Scapular fenestra elliptical, fenestra between proximal radials one and two elliptical or absent, fenestra extending between proximal radials two and three elongate ellipse. Scapula broadly Y-shaped with robust helve; coracoid with triangular head and long, thin helve. Dorsoanterior hook of the basal cartilage present. Distal radials present at base of rays 2 to 27–29, ventral 4–6 rays articulating directly with fibrocartilage pad.

Pelvic disk moderately large, length 32.1–40.2 (37.4) % HL, round, slightly longer than wide, width 30.2–38.4 (34.1) % HL, anterior lobe moderately developed, flat with margins often slightly upturned, distance from tip of upper jaw to pelvic disc 10.4–17.7 (13.1) % SL. Anus

closer to pelvic disk than to anal-fin origin, 5.3–19.8 (18.7) % HL from pelvic disk, distance from tip of upper jaw to anus 24.2–32.5 (28.3) % SL.

Principal caudal-fin rays 11 (11; Table 1), dorsal procurrent rays 2 (2), ventral procurrent rays 1–2 (1). Membrane of posterior dorsal-fin rays attached for shorter distance to caudal fin than anal-fin rays: dorsal-fin rays attached to caudal fin 15.1–27.9 (27.9) % CL; anal-fin rays, 25.0–40.6 (34.4) % CL. Depth at base of caudal fin 11.5–22.0 (11.5) % CL.

Skin relatively thick, prickles absent. Pyloric caeca 11–13, length about 33% HL.

Vertebrae 44–47 (46), precaudal 9–10 (10), caudal 34–37 (36; Table 1). Pleural ribs 1–3 (2), present on vertebrae 8–10, anteriormost small when three present, others long and slender. Hypural plate composed of dorsal and ventral plates divided by slight notch in plate. Single epural present.

Coloration.—Body variegated light brown and pale in life, head with white blotch at mandibular articulation, white spots often scattered over body, concentrated dark band of pigment at base of caudal fin with small white band immediately following (Fig. 1.3). Eye greenish to goldish in life with pupil a horizontal ellipse or slitted. Body with scattered melanophores in preservation. Eye black in preservation (Fig. 1.6A). Peritoneum, orobranchial cavity, stomach, intestines, pyloric caeca, and urogenital papilla pale.

Life history.—The largest specimen examined was a ripe female 46.9 mm (UW 158413), with yolked egg diameter of 2.1 mm. The smallest female with yolked eggs was 35.3 mm (UW 158418). Ovaries were dissected from two females with yolked eggs. Females had five to nine yolked eggs (2.2–3.9 mm diameter), a roughly equal number of opaque, developed but unyolked

eggs (0.7–0.9 mm diameter), and many small, translucent immature eggs (0.2–0.3 mm diameter). All males were mature or immature, none were ripe. The one mature male was 35.1 mm (UW 158417).

Distribution.—Careproctus sp. M has been collected throughout the Aleutian Islands from north of Amatignak Island in the west to west of Akutan Island in the east (Fig. 1.9). Collection depths range from 80 to 232 m.

DISCUSSION

Temnocora (Burke, 1930) was originally described as a monotypic genus containing *C. candidus* and diagnosed by its notched dorsal fin and slit-like pupil, among other characters. It was synonymized with *Careproctus* based on a morphological phylogeny (Kido, 1988), which was further supported by the molecular phylogeny by Orr et al. (2019). Orr et al. (2019) recovered *C. candidus* nested in a larger clade with other *Careproctus* species and recognized *Temnocora* as a subgenus that included *C. candidus*, *C. comus, C. faunus, C. iacchus,* and *C. staufferi*. All of these species, as well as the species we describe here, share general color characters and some species share morphological characters that support their inclusion in the *Temnocora* subgenus, including variegated body coloration (unique among other *Careproctus* species), a distinct white blotch on the cheek, elliptical to slitted pupils (with the exception of *C. comus* and *C. faunus*), and a relatively large eye (with the exception of *C. iacchus* and *C. staufferi*). The two new species and *Careproctus candidus* are most similar to each other and the other members of the subgenus *Temnocora* of Orr et al. (2019). Based on color pattern and

morphological similarity we assign both newly described species *Careproctus.* sp. K and *Careproctus* sp. M to the subgenus *Temnocora.*

Within the subgenus *Temnocora*, the two new species and *C. candidus* are readily distinguished from *C. comus* and *C. faunus* based on lower numbers of total vertebrae (46–52, 47–51, 44–47, and in *C. candidus*, *Careproctus.* sp. K, and *Careproctus* sp. M respectively vs. 56–61 and 52–56 in *C. comus* and *C. faunus* respectively; Table 2). Collectively, *C. candidus* and the two species described here have a larger orbit length than *C. iacchus* (31.2–46.6 % HL, 30.4–40.5 % HL, and 31.1–46.5 % HL in *C. candidus*, *Careproctus.* sp. K, and *Careproctus* sp. M respectively vs. 24.3–26.6 % HL in *C. iacchus*; Table 2) and have a notched pectoral fin (pectoral-fin notch absent in *C. iacchus*; Table 2). *Careproctus candidus* and *Careproctus.* sp. K can be differentiated from *C. staufferi* based on total vertebrae (46–52 and 47–51 in *C. candidus* and *Careproctus* sp. K respectively vs. 44–46 in *C. staufferi*; Table 2). *Careproctus* sp. M can be differentiated from *C. staufferi* based on lower count of pectoral-fin rays (30–36 in *Careproctus* sp. M vs. 36–44 in *C. staufferi*), narrower caudal-fin base depth (1.6–2.9 % SL in *Careproctus* sp. M vs. 2.7–4.1 % SL in *C. staufferi*), variegated brown and white live body coloration (vs. mottled white and red with a golden slash for *C. staufferi*; Table 2), body with scattered melanocytes in preservation (vs. uniformly pale in preservation in *C. staufferi*), and body size (largest *Careproctus* sp. M examined 46.9 mm SL vs. 89.6 mm SL in *C. staufferi*; Table 2). Members of the subgenus *Temnocora,* including both new species and *C. candidus*, are predominately found in the Aleutian Islands, with isolated reports of *C. candidus* from the western Pacific (Sheiko and Fedorov, 2000; Mecklenburg et al., 2002), except for *C. iacchus* which is known only from the Sea of Japan (Kai et al., 2018).

Outside of the subgenus, *Careproctus* sp. M is similar to *C. lacrima,* which is also small (43.5–60.0 mm SL) and lacks an anterior dorsal-fin lobe, but can be distinguished in having a variegated brown body coloring (vs. uniformly pale and translucent in *C. lacrima*), slender depressed body (vs. robust and compressed) and an elliptical to slitted pupil (vs. pupil reduced to pinpoint). Within the subgenus *Temnocora,* an anterior dorsal-fin notch is present only in *C. candidus*. Outside of *Temnocora* and within the clade Aenigmoliparia of Orr et al. (2019), an anterior dorsal-fin notch is also present in *Lopholiparis flerxi*, *C. maslenikovae*, and *C. spiraki* (Orr, 2004, 2021). *Careproctus candidus* is easily distinguished from *C. maslenikovae* and *C. spiraki* based on its variegated body color and slitted pupil, compared to the splotchy red body color and round pupils in *C. maslenikovae* and *C. spiraki* (Orr, 2021). It is easily distinguished from *L. flerxi* based on the lack of heavily ossified superficial head bones and higher counts of total vertebrae (46–52 vs. 40) (Orr, 2004).

We examined eight cleared-and-stained pectoral girdles of *Careproctus candidus* as well as two of *Careproctus* sp. K and three of *Careproctus* sp. M. In all three species, pectoral girdle morphology varied. In both *C. candidus* and *Careproctus* sp. K, each species had the same two morphotypes: one with an interradial fenestra between proximal radials two and three and the other without that fenestra (Fig. 1.7A–D). Other slight differences in proximal radial shape, extent of distal radials, and size and shape of the anterodorsal hook on the basal cartilage were also seen. In *C. candidus*, the anterodorsal hook on basal cartilage was present in both morphotypes with the exception of one specimen (UW 111870) which also lacked a clear notch in the dorsal fin, making its identification as *C. candidus* questionable. *Careproctus* sp. M also had two morphotypes: one with three large fenestrae between the scapula and proximal radial one, proximal radial one and two, and proximal radial two and three, with proximal radials one

and two notched and hourglass shaped (Fig. 1.7E). The second morphotype was lacking the large fenestra between proximal radials one and two and the radials were crescent shaped rather than hourglass shaped (Fig. 1.7F). Other slight differences in extent of distal radials were also seen. Current practice in snailfish studies is to examine and report on a single girdle, based on the assumption that pectoral girdle morphology is generally fixed within a species. In the case where studies have examined pectoral girdles in multiple specimens, some found no variation, supporting the assumption of fixed morphology (e.g., Orr et al. 2020); while others are similar to this study and found significant differences in numbers of proximal radials and fenestrae between individuals within a species (Chernova, 2006; Knudsen and Møller, 2008), supporting the suggestion that pectoral girdle morphology is more plastic than previously assumed.

Life history characteristics differed among the species described here. *Careproctus* sp. M has the smallest size at maturity (the smallest female with yolked eggs was 35.3 mm). The smallest female *C. candidus* with yolked eggs was 62.5 mm. Of the 13 specimens of *Careproctus* sp. K examined, 10 were female (72.4–106.0 mm SL), and none were found with yolked eggs. For the 59 specimens of *C. candidus* examined of known sex, 44 were female (27.0–94.7 mm SL) and of those, 10 (22.7%; 62.5–94.7mm SL) had yolked eggs. For *Careproctus* sp. M, nine females (35.3–46.9 mm SL) were found among the 12 specimens examined, and eight of those females (89%; 35.3–46.9 mm SL) had yolked eggs. While the proportion of females found was about 75% in each of the three species, the proportion of ripe females was substantially different (0 %, 22.7 %, 89%). Snailfish can be periodic or continuous spawners, with periodic spawners having ripe eggs of roughly the same size and only unyolked eggs otherwise present, while continuous spawners are characterized by having yolked eggs of many different sizes present at once (Stein, 1980). All *Careproctus* sp. M and *C. candidus* with yolked eggs had two other sizes

of eggs present: opaque, maturing but unyolked eggs around 1mm in diameter, and translucent, immature eggs less than 1 mm in diameter. The presence of relatively few egg class sizes suggests these species may be periodic spawners, but given all individuals were collected in only five months of the year it is difficult to be certain. No females of *Careproctus* sp. K had yolked eggs, instead having maturing eggs as well as some very small, immature eggs present, making it difficult to discern their spawning strategy. Additionally, while not definitive proof of larger size at maturity, the lack of ripe females in *Careproctus* sp. K indicates a potential spawn timing difference from *C. candidus* and *Careproctus* sp. M, as the material examined here was collected from the same region during the same period (May to August) across multiple years. This could also indicate a habitat difference among the species if, for example, ripe females of *Careproctus* sp. K are inhabiting rocky reefs where they are not able to be collected by bottom trawl. Males constituted about 25% of the specimens for each of the species, with there being 15 males out of 59 *C. candidus* specimens of known sex, three males out of 13 *Careproctus* sp. K specimens, and three males out of 12 *Careproctus* sp. M specimens. For *C. candidus* the only ripe male was larger than the smallest ripe female (73.8 mm SL) and one smaller male was spent (54.4 mm SL). For *Careproctus* sp. K, the only ripe male was among the largest specimens at 103.5 mm SL. No ripe males were found in *Careproctus* sp. M. The difference in proportions of ripe and spent males (13.3 % for *C. candidus*, 33.3 % for *Careproctus* sp. K, and 0 % for *Careproctus* sp. M) may also be indicative of a difference in spawn timing or habitat preference among the species.

The geographic ranges for the two new species and *C. candidus* overlap along the Aleutian Island chain. Based on our material, the range for *C. candidus* extends from the western end of the Aleutian Island chain to the northern Gulf of Alaska (Fig. 1.9). The ranges for

Careproctus sp. K and *Careproctus* sp. M both fully overlap with *C. candidus* but only minimally with each other. *Careproctus* sp. K is found in the western and central Aleutians as well as Petrel Bank, mostly between the Near Islands and Amchitka pass, and *Careproctus* sp. M is mostly found in the central and eastern Aleutians between Amchitka and Seguam passes. *Careproctus candidus* was collected across the widest geographic range, as well as the widest depth range from 88 m to the deepest collection depth of all three species at 433 m, with most specimens collected between 200 and 250 m. The depth of 433 m is the maximum depth record for our material but does not necessarily represent the maximum depth for *C. candidus*, as there may be regions the species inhabits that are deeper but have not been or perhaps cannot be sampled with bottom trawl gear. Nearly all specimens were obtained from trawling during biennial AFSC Aleutian Island surveys, which since 1994 have been routinely conducted at depths shallower than 500 m, with only 68 of 4600 tows conducted being greater than 450 m depth. *Careproctus* sp. M was collected at the shallowest depths (between 80 and 232 m) and *Careproctus* sp. K was collected at depths between 134 and 280 m, which was within the depth range of *C. candidus* and overlapping with but deeper on average than *Careproctus* sp. M. Mean depth was 214 m, 178 m, and 139 m for *C. candidus*, *Careproctus* sp. K, and *Careproctus* sp. M respectively. In one instance, *C. candidus* and *Careproctus* sp. K were collected in the same haul, but *Careproctus* sp. M was never collected in the same haul with either of the other species.

As noted in the description, at least three specimens identified in this study may represent a new species. In addition to these three, the morphological diversity among specimens we identified as *C. candidus* suggests there may be more diversity in this group than is currently recognized. At present, no genetic material is available for *Careproctus* sp. K or *Careproctus* sp. M, and very limited material available for *C. candidus*. Genetics, especially barcoding with COI

sequence data, may be the key to recognizing the patterns of diversity in this group, as it was with *C. ambustus* (Orr et al., 2020). Two COI sequences published in Orr et al. (2019) differ by 1.0% over a 492 bp length, showing there is likely genetic differentiation within the group. Thus, further collection of material including genetic material will continue to be crucial in working out the diversity and relationships within the subgenus *Temnocora*.

COMPARATIVE MATERIAL EXAMINED

Careproctus comus: UW 111841, 78.9 mm, holotype, 53.6741°N, 167.5344°W, 303 m depth, F/V *Dominator*, cruise 2000-01, haul 28, benthic bag, J. W. Orr, 24 May 2000.

Careproctus faunus: UW 111867, 101.9 mm, holotype, 52.2099°N, 172.2057°W, 348 m depth, F/V *Sea Storm*, cruise 2002-01, haul 157, J. W. Orr, 23 July 2002.

Careproctus iacchus: FAKU 201379, 39.2 mm, paratype, 44.3633°N, 143.9233°E, 163 m depth, R/V *Kaiyo-maru* No. 5, A. Tohkairin, 19 April 2014.

Careproctus lacrima: UW 200024, 50.2 mm, holotype, 52.0038°N, 177.8278°W, 111 m depth, F/V *Sea Storm*, cruise 2002-01, haul 118, J. W. Orr, 11 July 2002.

Careproctus maslenikovae: UW 155708, 36.9 mm, holotype, 52.6466°N, 170.2027°W, 234 m depth, F/V *Gladiator*, cruise 2006-01, haul 35, benthic bag, J. W. Orr, 15 June 2006.

Careproctus spiraki: UW 159753, 52.5 mm, holotype, 52.3255°N, 172.7466°W, 457 m depth, F/V *Sea Storm*, cruise 2004-01, haul 60, J. W. Orr, 19 June 2004.

Lopholiparis flerxi: UW 47868, 32.5 mm, holotype, 51.4593°N, 178.4612°W, 278 m depth, F/V *Vesteraalen*, cruise 2000-01, haul 120, W. C. Flerx, 20 June 2000; UW 159752, 40.5 mm, 51.8446°N, 173.9241°W, 252 m depth, F/V *Sea Storm*, cruise 2012-01, haul 81, D. Drumm, 26 June 2012.

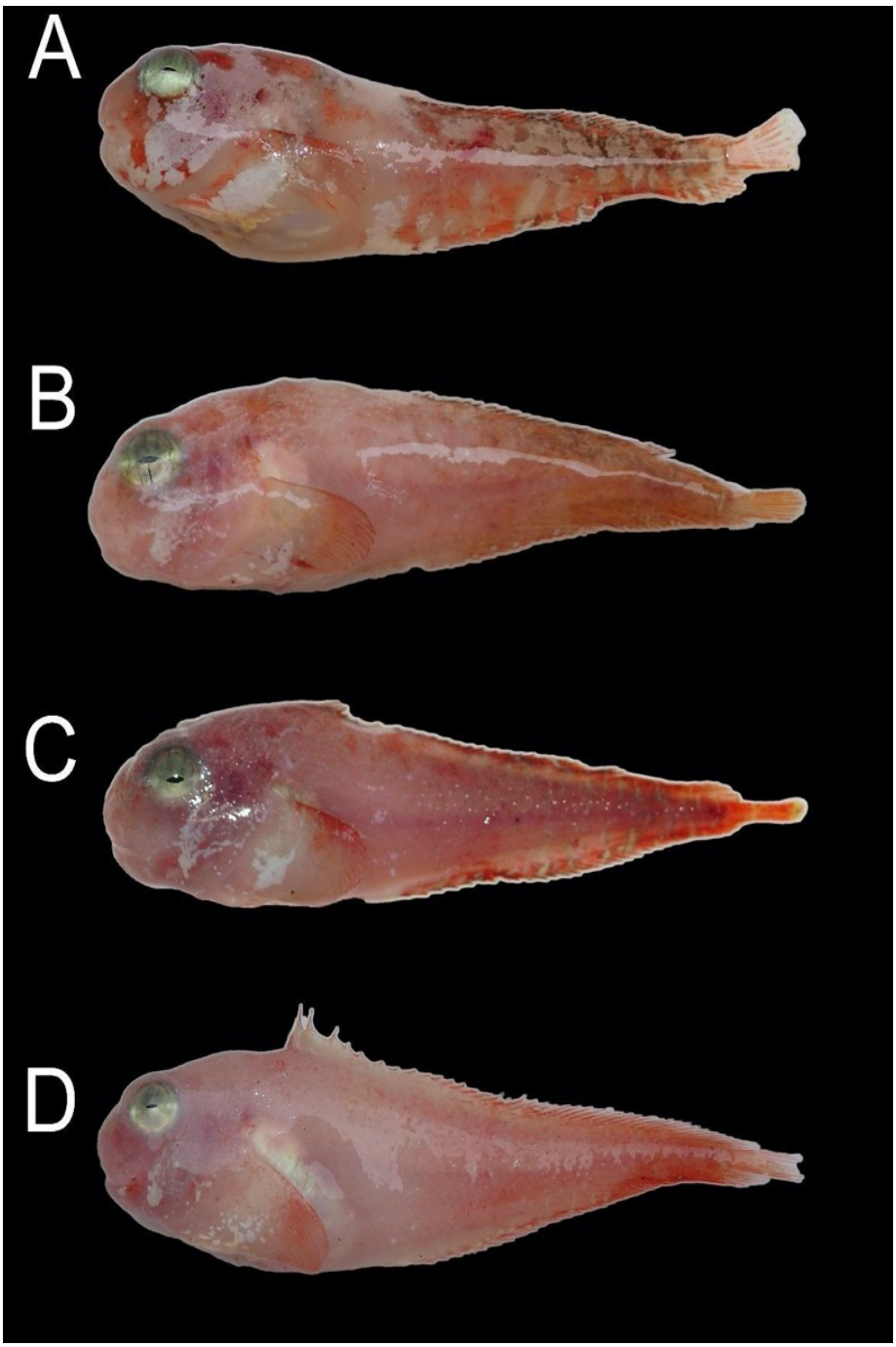

Fig. 1.1. *Careproctus candidus* (A) UW 158412, 73.4 mm, female, Aleutian Islands, Tanaga Pass, 51.6261°N, 178.439°W, 178 m depth; (B) UW 158291, 75mm, male, Aleutian Islands, south of Seguam Island, 52.1823°N, 172.4394°W, 190 m depth; (C) UW 117558, 66 mm, male, Aleutian Islands, northwest of Herbert Island, 52.9288°N, 170.4111°W, 221 m depth; (D) UW 117177, 90.5 mm, female, Aleutian Islands, south of Amukta Island, 52.3675°N, 171.2463°W, 320 m depth. All photographed before preservation.

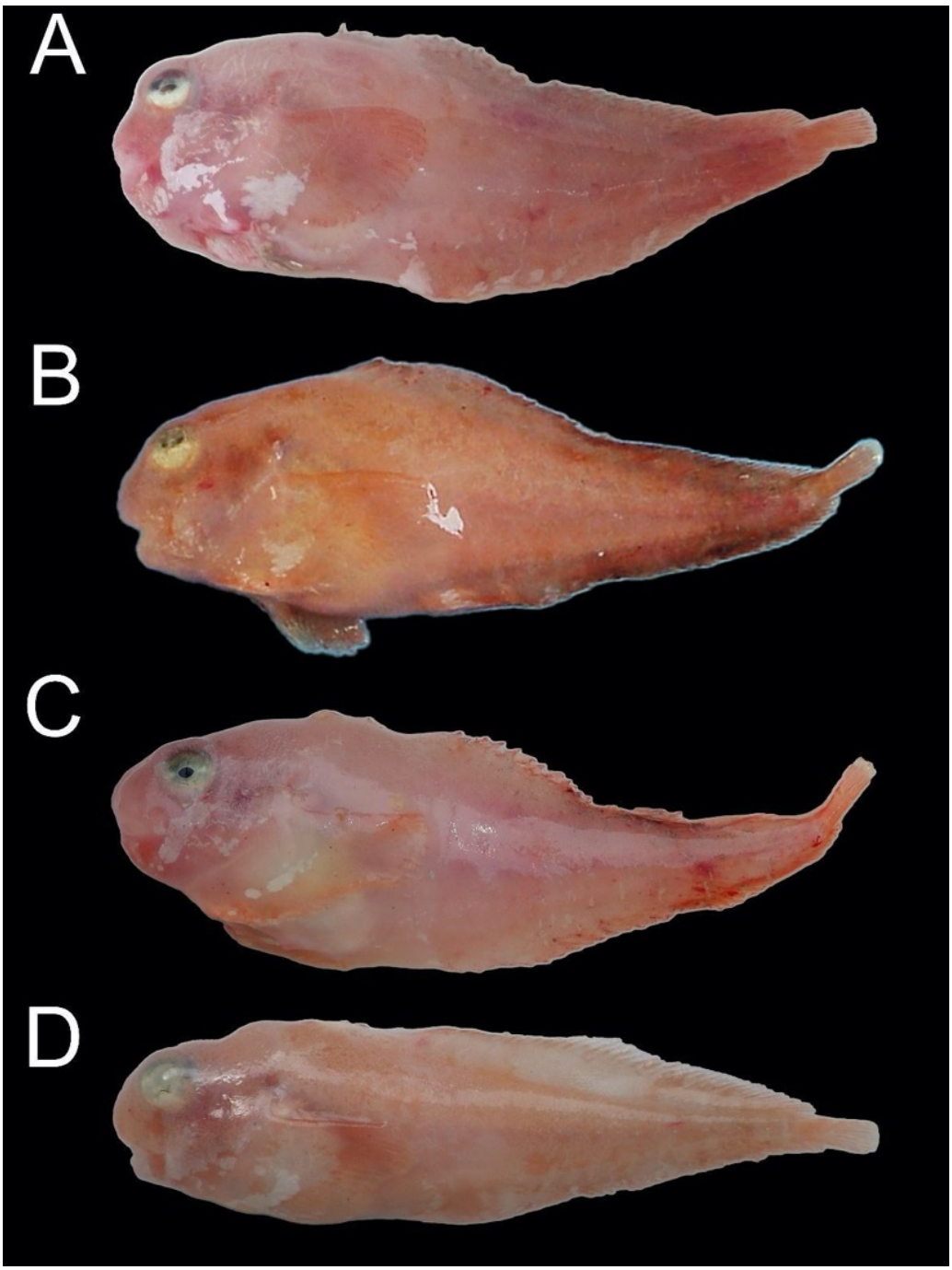

Fig. 1.2. *Careproctus* sp. K, new species: (A) UW 200421, 105.3 mm, holotype, male, Aleutian Islands, Petrel Bank, 52.0360°N, 179.3053°E, 277 m depth; (B) UW 159772, 103.6 mm, female, Aleutian Islands, Petrel Bank, 52.5150°N, 179.5188°W, 134 m depth; (C) UW 158344, 90.3 mm, female, Aleutian Islands, northwest of Semisopochnoi Island, 52.0381°N, 179.3166°E, 280 m depth; (D) UW 159773, 80.0 mm, female, Aleutian Islands, Petrel Bank, 52.3688°N, 179.9117°W, 182 m depth. All photographed before preservation.

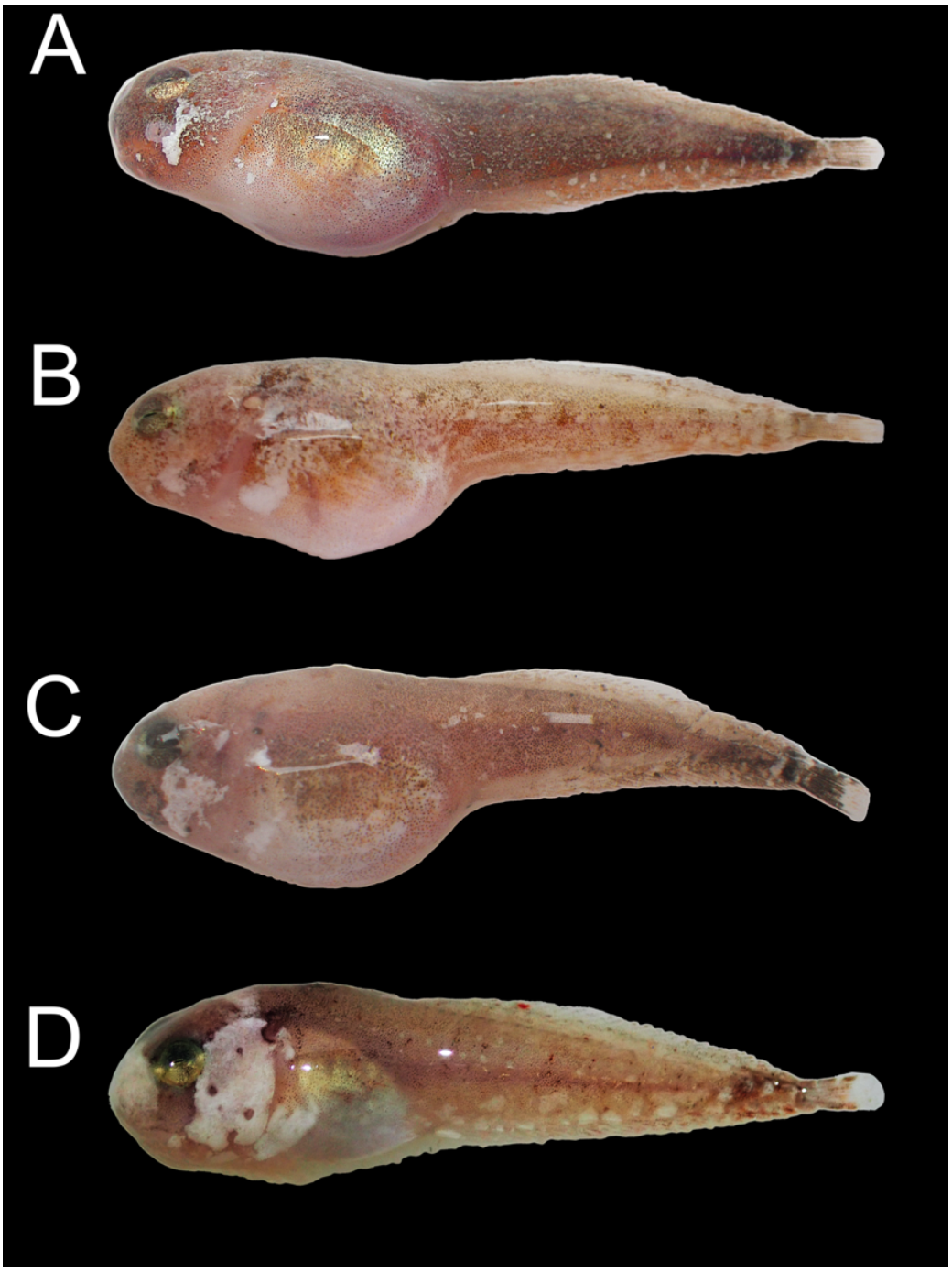

Fig. 1.3. *Careproctus* sp. M, new species: (A) UW 155719, 43.5 mm, holotype, ripe female, Aleutian Islands, Seguam Pass, 52.2629°N, 172.8827°W, 158 m depth; (B) UW 158401, 38.5 mm, ripe female, Aleutian Islands, Tanaga Pass, 51.5078°N, 178.5773°W, 168 m depth; (C) UW 155709, 2 of 2, 42.0 mm, ripe female, Aleutian Islands, Tanaga Pass, 51.6163°N, 178.1877°W, 126 m depth; (D) UW 158417, 35.1 mm, male, Aleutian Islands, Tanaga Pass, 51.6177°N, 178.1883°W, 122 m depth. All photographed before preservation.

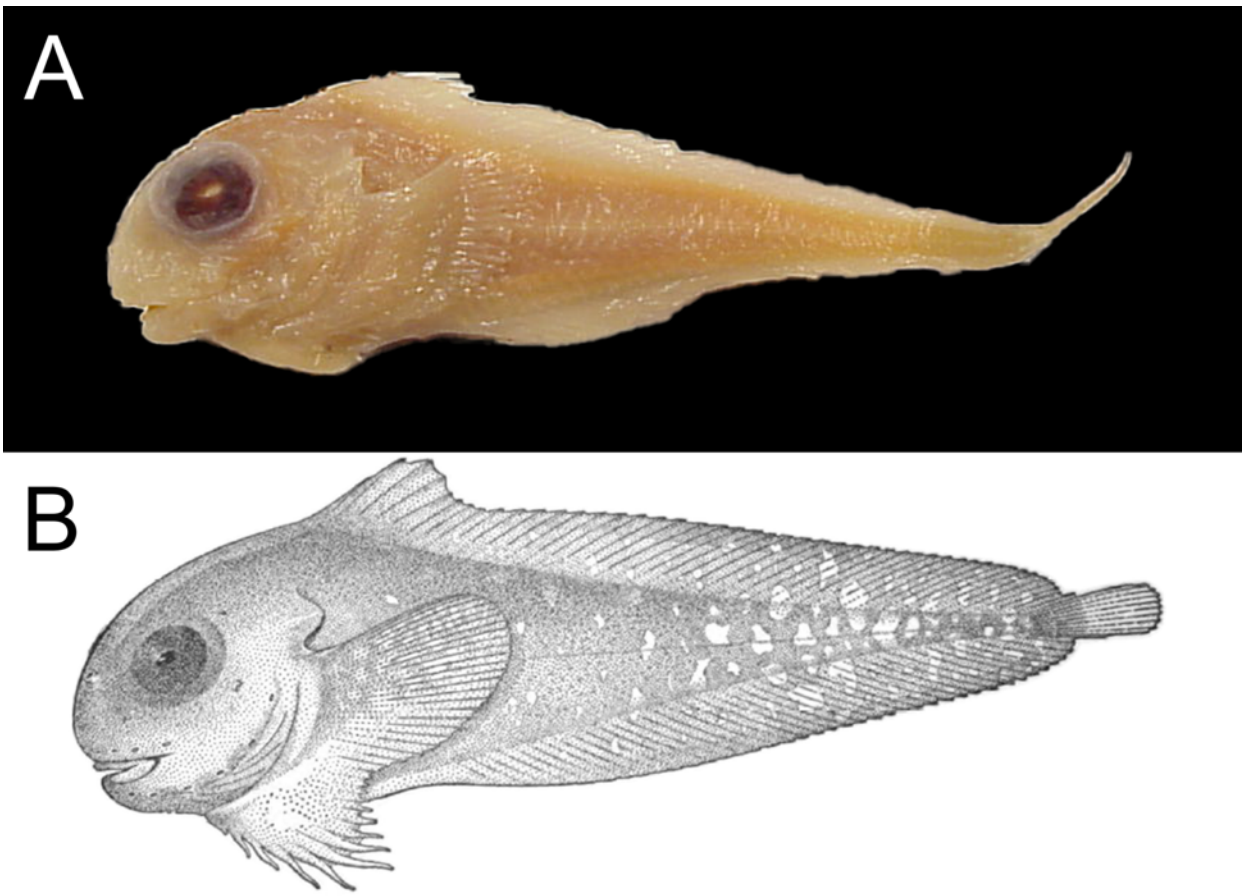

Fig. 1.4. *Careproctus candidus* (A) UW 158412 1 of 2, 73.4 mm, female, Aleutian Islands, Tanaga Pass, 51.6261°N, 178.439°W, 178 m depth, photographed after preservation; (B) USNM 74384, 60.7 mm, illustration previously published in Gilbert and Burke (1912; Fig. 22).

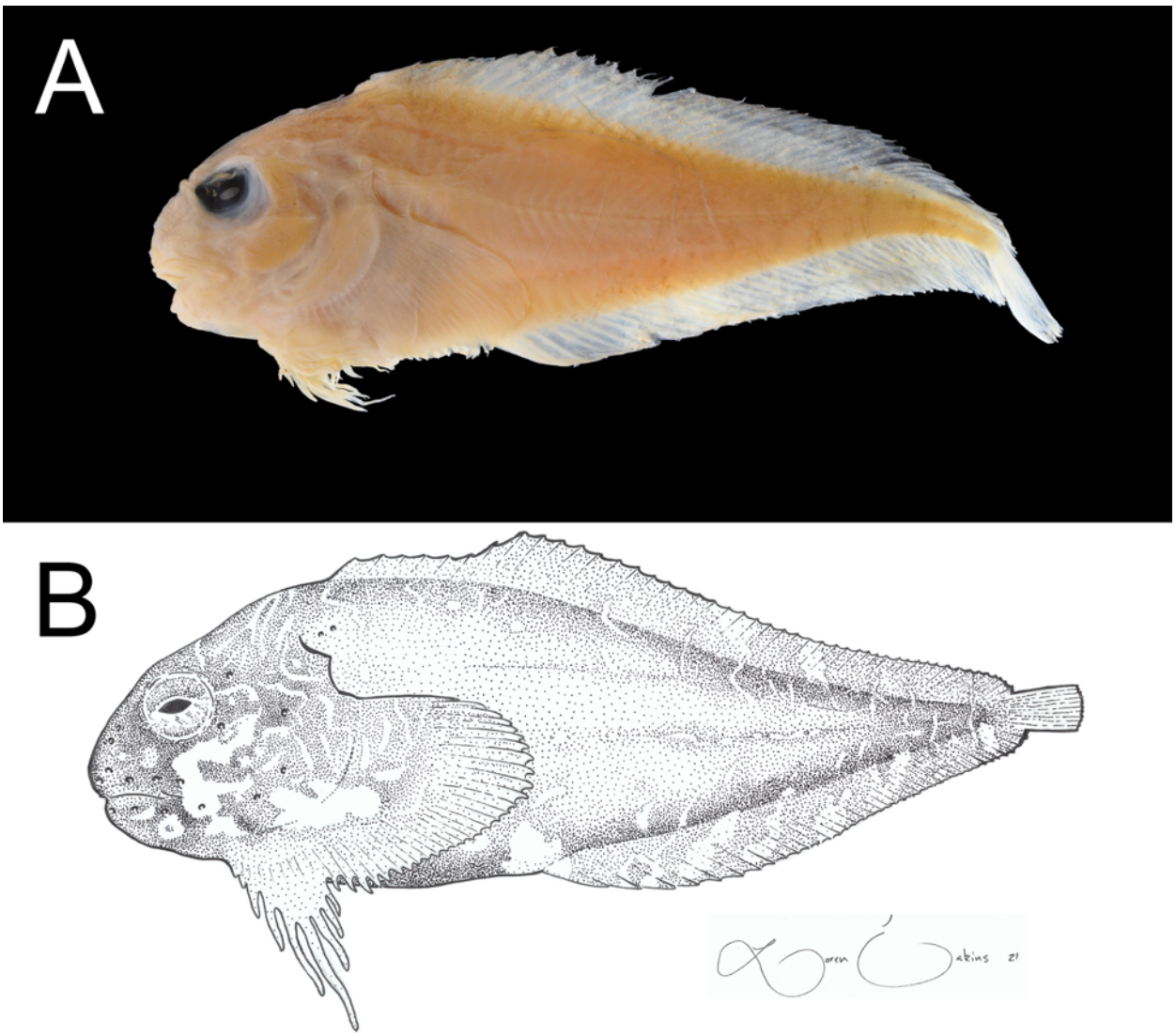

Fig. 1.5. *Careproctus* sp. K new species, UW 200421, 105.3 mm, holotype, male, Aleutian Islands, Petrel Bank, 52.0360°N, 179.3053°E, 277 m depth; (A) photographed after preservation; (B) illustration by Loren Eakins.

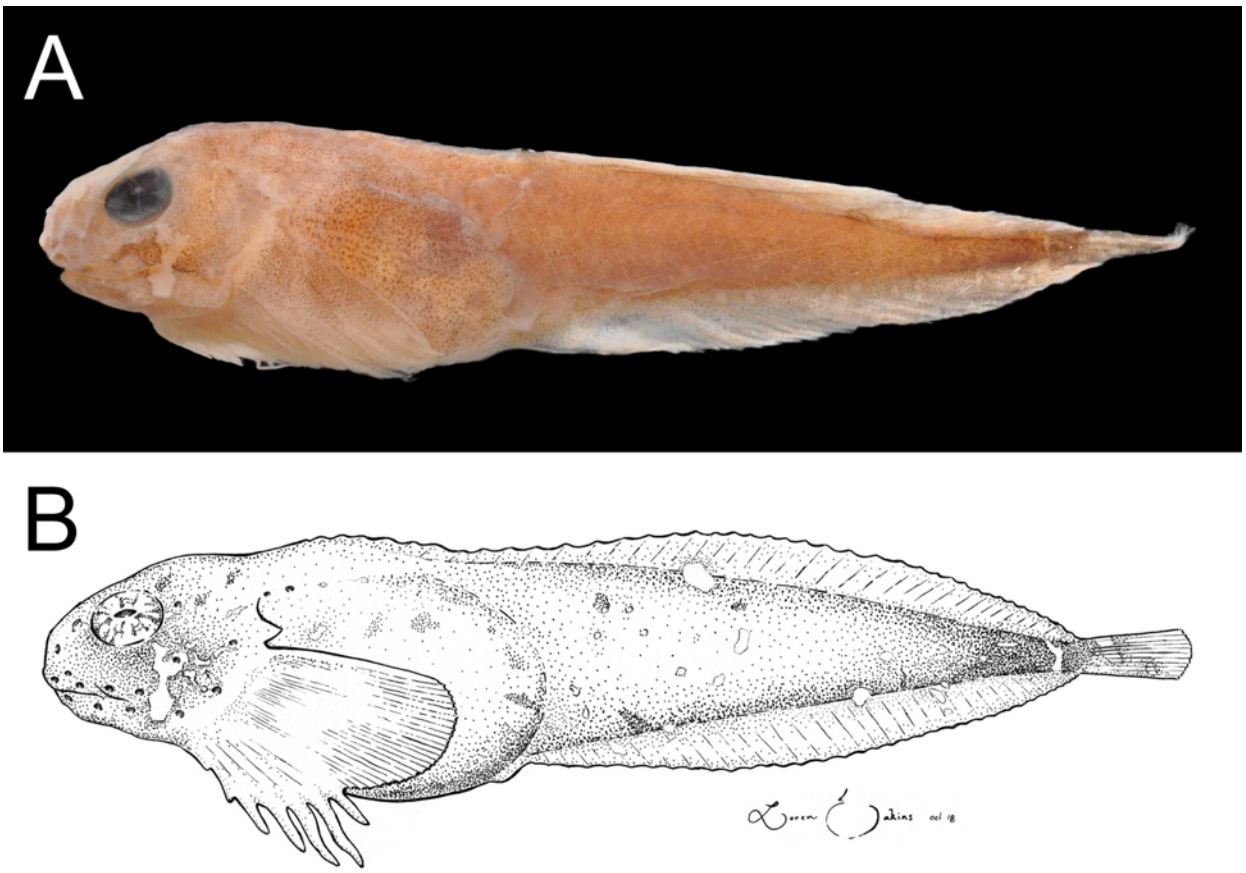

Fig. 1.6. *Careproctus* sp. M, new species, UW 155719, 43.5 mm, holotype, ripe female, Aleutian Islands, Seguam Pass, 52.2629°N, 172.8827°W, 158 m depth; (A) photographed after preservation; (B) illustration by Loren Eakins.

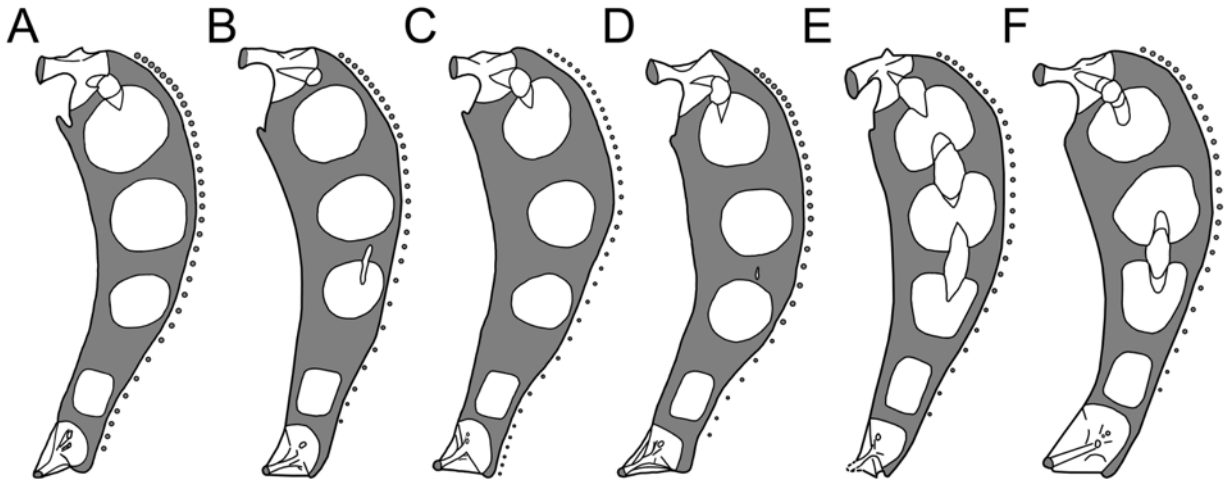

Fig. 1.7. Medial views of right pectoral girdles. (A) *Careproctus candidus,* UW 117243, 80 mm; (B) *Careproctus candidus,* UW 154876 1 of 4, 78 mm; (C) *Careproctus* sp. K, new species, paratype, UW 117553, 106.0 mm; (D) *Careproctus* sp. K, new species, paratype, UW 155491, 88.7 mm; (E) *Careproctus* sp. M, new species, paratype, UW 155709 2 of 2, 42.0 mm, dashed lines indicate where coracoid was cut during dissection; (F) *Careproctus* sp. M, new species, paratype, UW 158400, 44.2 mm. Shaded areas represent cartilage.

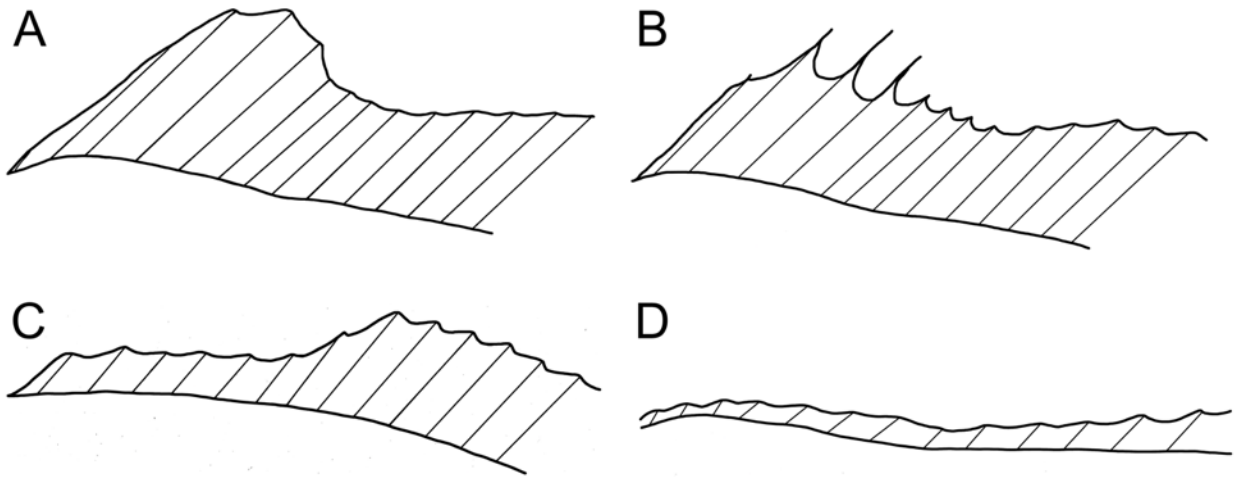

Fig. 1.8. Anterior dorsal-fin morphology for *Careproctus candidus* and two new species from the Aleutian Islands, showing (A) anterior dorsal-fin lobe with notch rays not exserted (based on *Careproctus candidus*, USNM 74384, 60.7 mm, holotype); (B) anterior dorsal-fin lobe with notch rays partially exserted (based on *Careproctus candidus,* UW 117177, 90.5 mm); (C) dorsal-fin lobe and notch absent instead fin with anterior section of rays one to seven roughly equal in length (based on *Careproctus* sp. K, new species, UW 200421, 105.3 mm, holotype); and (D) anterior dorsal-fin lobe and notch absent, rays short and embedded in skin (based on *Careproctus* sp. M, new species, UW 155719, 43.5 mm, holotype).

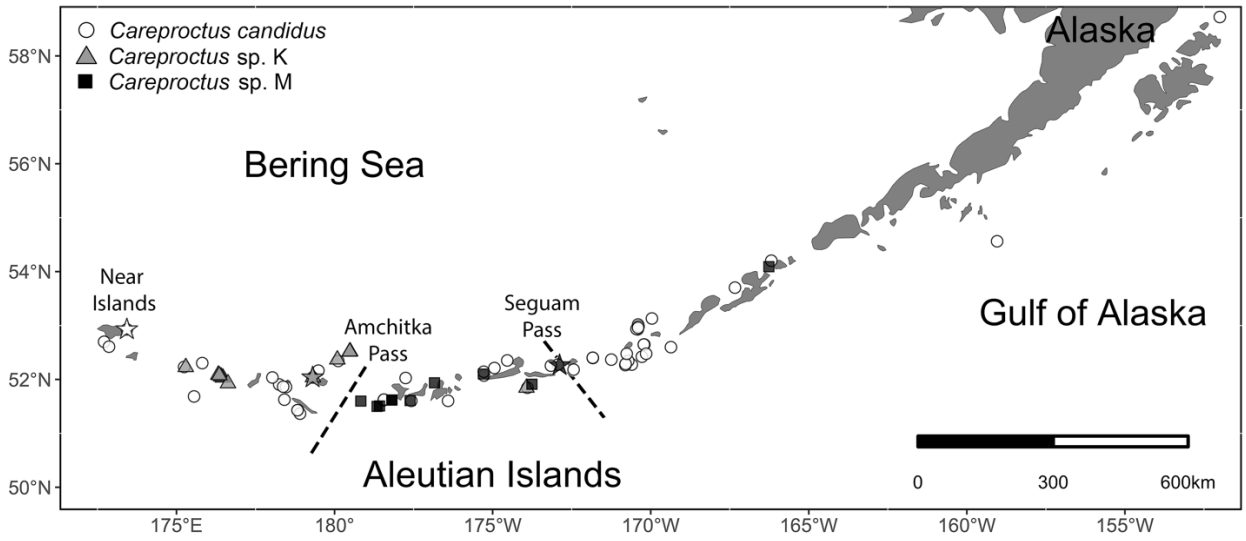

Fig. 1.9. Distribution of *Careproctus candidus* (white)*, C.* sp. K, new species (gray), and *C.* sp. M, new species (black), from the Aleutian Islands. *Careproctus candidus* distribution is based on material examined; the two new species, on all known material. Each symbol may represent more than one capture. Holotype localities represented by stars.

Table 1.1. Morphometric and meristic data for *Careproctus candidus* and two new liparid species: *Careproctus* sp. K and *Careproctus* sp. M. All species from the Aleutian Islands. Proportions in % standard length.

	Careproctus candidus			*Careproctus* sp. K			*Careproctus* sp. M		
	n	Range	Mean± SD	*n*	Range	Mean± SD	*n*	Range	Mean± SD
Standard Length	70	27.0–94.7		13	72.4–106.0		12	27.0–46.9	
Head Length	55	25.74–31.02	29.0±1.2	13	27.51–29.44	28.7±0.5	9	25.39–28.28	27.0±0.8
Head Width	55	9.67–17.66	13.5±2.1	13	11.31–15.39	13.1±1.2	9	14.29–22.76	18.2±3.0
Body depth at anal-fin origin	55	16.88–25.87	20.2±2.0	13	21.41–28.96	25.0±2.2	9	12.79–20.25	16.9±2.4
Dorsal-fin origin to anal-fin origin length	54	23.68–31.36	27.4±1.9	13	26.49–33.71	30.2±2.0	9	22.47–29.64	25.6±2.8
Body depth at pelvic disc	55	20.16–28.36	24.6±2.0	13	23.11–28.87	26.5±1.8	9	14.93–26.17	21.9±3.8
Body depth at dorsal-fin origin	55	21.24–31.63	26.8±2.4	13	25.93–32.75	29.3±2.1	8	21.37–28.31	24.8±2.4
Body depth at anal-fin midpoint	55	6.25–12.30	8.7±1.4	13	10.22–15.38	12.1±1.5	9	6.90–8.14	7.4±0.4
Snout length	55	7.13–11.35	9.1±1.1	13	8.36–10.51	9.3±0.7	9	7.47–9.38	8.2±0.6
Orbit length	55	8.73–12.97	11.1±0.8	13	8.90–11.46	9.9±0.9	9	8.57–12.34	10.1±1.3
Bony interorbital width	51	1.68–5.12	3.3±0.8	12	3.01–4.23	3.4±0.3	9	3.38–5.93	4.7±0.9
Greatest interorbital width	53	4.19–11.36	8.0±1.3	12	6.88–9.12	7.9±0.8	9	6.40–12.34	10.0±2.0
Suborbital to maxilla length	55	1.89–5.27	3.0±0.8	13	2.21–4.63	3.3±0.7	9	1.42–5.18	3.2±1.3
Suborbital to mandible length	54	3.78–8.33	5.8±0.9	13	4.75–7.17	5.6±0.7	9	4.26–6.75	5.7±0.7
Mouth width	55	7.74–15.41	10.9±1.7	13	8.80–11.51	10.2±0.7	9	9.87–18.85	15.2±3.3
Maxilla length	55	9.53–15.36	11.7±1.3	13	9.48–11.48	10.7±0.5	9	11.81–17.86	14.7±1.8
Mandible length	55	7.80–13.50	10.8±1.5	13	9.91–12.74	11.4±0.9	9	12.07–19.31	14.8±1.9
Gill slit length	55	3.93–7.52	5.6±0.8	13	4.15–7.50	5.5±1.0	9	3.85–6.17	5.4±0.8
Pectoral-fin length, upper lobe	55	17.45–24.07	20.7±1.5	13	17.23–22.54	20.3±1.7	9	14.96–21.38	18.4±2.3
Pectoral-fin length, lower lobe	55	11.22–22.21	17.6±2.5	13	14.11–23.08	17.7±2.7	9	10.62–19.51	15.8±3.0
Pectoral-fin length, notch ray length	54	4.32–9.91	7.6±1.4	13	6.13–10.37	8.4±1.3	9	6.74–9.87	8.3±1.2
Pre-dorsal-fin length	55	22.88–32.46	27.4±2.2	13	24.81–28.47	26.8±0.9	9	26.94–31.45	29.2±1.6
Pre-anal-fin length	55	34.69–46.90	40.6±3.3	13	37.99–45.99	40.6±2.3	9	44.09–49.87	46.7±2.1
Snout to pelvic disc length	55	11.78–20.16	15.6±1.8	13	12.60–19.89	15.9±2.1	9	10.37–17.66	14.7±2.3
Snout to anus length	55	24.23–33.21	27.8±2.4	13	22.89–28.59	26.2±1.6	9	24.20–32.48	28.7±2.9
Pelvic disc length	55	7.58–11.34	9.4±1.0	13	8.12–10.52	9.5±0.7	9	8.83–11.11	10.0±0.7
Pelvic disc width	55	6.61–11.25	8.5±1.0	13	7.22–10.51	8.9±0.9	9	8.31–10.62	9.1±0.7
Pelvic disc to anus length	55	1.23–6.00	3.0±1.2	13	1.33–4.27	2.5±1.0	9	1.43–5.45	3.7±1.6
Anus to anal-fin length	55	8.09–21.07	13.2±2.5	13	11.83–16.62	14.3±1.2	9	15.10–23.12	19.2±2.8
Caudal fin length	49	10.76–16.88	14.3±1.1	13	13.49–14.88	14.1±0.5	9	11.94–15.24	13.5±1.1
Dorsal-fin attachment to caudal fin length	52	1.95–5.62	3.5±0.7	12	2.89–4.91	4.2±0.6	9	1.98–3.91	2.7±0.6
Anal-fin attachment to caudal fin length	52	2.95–7.19	5.1±0.9	13	5.12–7.73	6.1±0.7	9	3.13–6.19	4.4±1.1
Caudal-fin base depth	54	1.39–3.02	1.9±0.4	13	1.68–2.51	2.1±0.2	9	1.58–2.86	2.2±0.5
Nostril tube length	55	0.00–2.17	0.9±0.4	13	0.34–1.25	0.9±0.2	9	0.43–1.23	0.9±0.3
Urogenital papilla length	53	0.00–3.17	0.7±0.6	13	0.00–2.90	0.7±0.8	9	0.00–0.57	0.2±0.2
Pyloric caeca length	7	5.52–10.62	7.8±1.9	5	5.53–8.98	7.6±1.3	3	5.18–11.40	8.9±3.3

Table 1.1 (Continued)

	Careproctus candidus			*Careproctus* sp. K			*Careproctus* sp. M		
	n	Range	Mean± SD	*n*	Range	Mean± SD	*n*	Range	Mean± SD
Pyloric caeca	9	9–14	12.0±1.3	4	17–21	19.2±2.1	3	11–13	12.0±1.0
Dorsal-fin rays	53	37–49	44.9±2.0	13	43–48	46.2±1.3	10	38–42	40.7±1.3
Anal-fin rays	52	35–41	38.3±1.5	13	37–41	39.1±1.3	12	32–37	34.2±1.4
Pectoral-fin rays	55	32–40	36.4±1.9	13	36–39	37.5±1.1	9	30–36	34.0±1.7
Rays of lower lobe of pectoral fin	60	7–12	8.1±0.8	13	7–9	8.4±0.7	9	7–8	7.1±0.3
Caudal-fin rays	40	10–13	10.9±0.6	10	10–11	10.5±0.5	10	11	11.0±0.0
Precaudal vertebrae	59	9–11	9.8±0.5	13	9–10	9.2±0.4	12	9–10	9.9±0.3
Caudal vertebrae	58	36–42	39.1±1.5	13	37–42	40.2±1.3	12	34–37	35.8±0.9
Total vertebrae	58	46–52	48.8±1.5	13	47–51	49.5±1.1	12	44–47	45.7±1.0
Tooth counts per row	32	4–11	6.9±0.9	13	5–10	7.0±1.2	8	4–12	7.5±1.5
Tooth rows	32	8–12	9.5±1.0	13	8–13	9.6±1.3	8	8–10	9.1±0.6
Gill rakers	32	6–10	8.1±0.9	6	7–10	8.3±1.5	8	6–8	6.6±0.7

Table 1.2. Diagnostic characters for species of *Careproctus* of subgenus *Temnocora* (Orr et al., 2019).

	C. candidus	***C.* sp. K**	***C.* sp. M**	***C. comus***	***C. faunus***	***C. iacchus***	***C. staufferi***
Total Vertebrae	46–52	47–51	44–47	56–61	52–56	49–52	44–46
Pectoral-fin ray count	32–40	36–39	30–36	33–39	32–38	36	36–44
Orbit Length	31.2–46.6 % HL	30.4–40.5 % HL	31.1–46.5 % HL	33.7–44.2 % HL	31.7–46.0 % HL	24.3–26.6 % HL	24.7–34.5 % HL
Smallest Ripe Female (SL)	62.5 mm	None found	35.3 mm	72.8 mm	91.3 mm	None found	71 mm
Largest specimen examined (SL)	94.7 mm	106.0 mm	46.9 mm	98.6 mm	101.9 mm	42.6 mm	89.6 mm
Anterior dorsal-fin lobe formed by distinct notch	present	absent	absent	absent	absent	absent	absent
Prickles	present or absent	absent	absent	absent	absent	absent	absent
Pupil shape	horizontal ellipse to slitted	horizontal ellipse to slitted	horizontal ellipse to slitted	elliptical to round	elliptical to round	horizontal ellipse to slitted	horizontal ellipse to slitted
Pectoral-fin notched	shallowly to moderately	moderately	shallowly to moderately	shallowly	shallowly	Not notched	shallowly
Coloration	variegated white and pink	variegated white and pink	variegated brown and white	variegated red, white, and grey	variegated red, white, and grey	variegated pinkish/ yellowish and white	mottled red and pale with golden slash along body

Chapter 2: Identifying Contamination in Phylogenomic Data from Exons for the Family Liparidae

Jennifer R. Gardner[1] and Luke Tornabene[1]

[1] School of Aquatic and Fishery Sciences and Burke Museum of Natural History and Culture, University of Washington, Seattle, USA

ABSTRACT

The development of next-generation sequencing (NGS) coupled with advances in methods to reliably target homologous loci across genomes for non-model organisms has made it possible to infer phylogenies based on genome-wide sequence data for species without published genomes. A more recently identified concern with NGS data is the difficulty in uncovering cross-contamination of samples, which is exacerbated by benchwork protocols that have greater potential for cross-contamination when compared to traditional Sanger sequencing. This can lead to publication of erroneous phylogenies and contaminated datasets . This study compiled a phylogenomic dataset based on exon capture for 4,284 markers in 97 individuals representing 64 species of snailfishes (family Liparidae). Initial data analysis resulted in a phylogeny with low support for internal branches, the failure to recover monophyletic species, and dramatically different placements of most taxa when compared to published phylogenies, which suggested the presence of widespread cross-contamination. Detection of cross-contamination was examined using a number of methods including comparing the topologies of well-supported published trees to trees inferred based on this dataset, comparing p-distances based on *a priori* hypotheses coupled with iterative removal of the most contaminated specimens, and using locus heterozygosity to detect contamination. All methods indicated the presence of widespread

contamination in the dataset; however, it was so widespread that it could not be specifically removed. The methods applied here are useful for detecting contamination in other datasets generated from NGS and recommendations are made for how to better avoid cross-contamination.

INTRODUCTION

Recently, next-generation sequencing (NGS) has enabled rapid sequencing of massive amounts of DNA and increased the feasibility of sequencing thousands of markers from across the genomes of hundreds of taxa, including non-model organisms, with less effort than what would be needed for comparably sized studies using Sanger sequencing. There are many different NGS methods but for all methods, samples are labeled with an index sequenced, pooled together, and sequenced in parallel, allowing for sequencing millions of fragments and hundreds of millions of base pairs in a single run (Glenn, 2011). The most commercially successful sequencing platform, Illumina (San Diego, USA), relies on amplification of short read fragments (up to 150 bp or 300 bp depending on sequencing specifications used) that are then sequenced based on fluorescence of reversibly attached dNTPs, deoxynucleoside triphosphates. Short read fragments are usually then assembled into longer contigs often with the use of a reference genome, or, when combined with longer raw reads from different NGS platforms, they may be assembled into genome scaffolds *de novo*.

NGS has been used with a variety of methods that aim to sequence across entire genomes to obtain data for phylogenetic analysis. Restriction Site associated DNA Sequencing (RADSeq, Baird et al., 2008) uses restriction enzymes to make many cuts throughout the genome, to which sequencing adapters can then be attached and size-selected DNA can be sequenced to identify single nucleotide polymorphisms (SNPs). Target capture is another class of methods in which

mRNA probes are designed for specific regions of the genome that are selectively enriched prior to sequencing. For target capture, the goal is to find regions of the genome that are conserved enough that probes can be designed and used across multiple species or families, but also regions that contain phylogenetic signal for the relevant taxa. Some of the regions that have been identified as useful target capture are ultra-conserved elements (UCEs; Faircloth et al., 2012), single-copy exons, and other highly conserved loci referred to as "anchor regions" of the genome (referred to as anchored hybrid enrichment, AHE; McCormack et al., 2013).

Each of these methods has different utility from a phylogenetic perspective. RADSeq targets single nucleotide polymorphisms across the genome that may be found in both coding and non-coding regions, and thus can be very useful for population genetic studies and species delimitation, as well as phylogenetic reconstructions (Baird et al., 2008). However, this method has drawbacks in distantly related taxa as mutations in the enzyme cut sites can result in missing data as well as difficulties in identifying potential paralogous (Razkin et al., 2016). As the name suggests, UCEs target ultra-conserved regions of the genome and thus can be very useful in phylogenetic reconstructions of very distantly related organisms because UCEs can be easily targeted from divergent taxa. Sequence variation is higher flanking regions of UCEs, increasing as the sequences move farther from the UCEs, which makes them useful for resolving phylogenies across many different time scales (Gilbert et al., 2015). However, UCEs can become more difficult to align as flanking regions away from the ultra-conserved core become more and more dissimilar. AHE targets anchor areas of the genome that are slightly more divergent across taxa than UCEs and may contain some coding regions but are not specifically targeting coding regions (Lemmon et al., 2012). Exon capture typically targets single-copy exons across the genome, which are generally conserved across closely related organisms and thus can be easy to

target and align, making them useful in phylogenetic reconstructions at many different evolutionary time scales (Hughes et al., 2021).

All these methods have some similarities in sample preparation and using NGS that make them all susceptible to issues with detecting and removing sample contamination. The benchwork involved in all these methods can involve multiple wash steps to remove non-target DNA, including washing using magnetic beads (Baird et al., 2008; Li et al., 2013). Wash steps especially can increase potential for cross-contamination, as each step increases the number of times samples must be pipetted. For example, the original target capture protocol from Li et al. (2013) requires at least four wash steps, each step requiring liquid to be added and removed two or three times from tubes while target DNA is bound to the wall of the tube on magnetic beads drawn to the magnetic tube rack. Each repeated opening and closing of tubes and addition or removal of liquid creates an opportunity for cross-contamination by transfer of DNA in solution through many different routes including airborne DNA droplets and failure to change contaminated pipette tips. These wash steps also require moving tubes to and from the magnetic tray, which creates an opportunity for sample mislabeling through misordering of the tubes. Overall, the benchwork of target capture methods usually results in many more pipetting steps than that of traditional Sanger sequencing. In addition, there are several PCR steps involved with most target NGS library preparation protocols, increasing the likelihood of amplifying even small amounts of cross-contaminated DNA. When PCR samples are sequenced using Sanger sequencing, samples that have been cross-contaminated can be detected prior to sequencing (e.g., the presence of multiple bands on an agarose gel) or after sequencing via detection of multiple peaks for many base pairs in sequence chromatograms. Moreover, in Sanger sequencing a single tube and PCR reaction is used for each locus per sample, and therefore widespread

contamination where hundreds or thousands of loci are contaminated from other samples would be extremely unlikely.

With target capture methods, the increased potential for mislabeling of samples and cross-contamination in benchwork, difficulty in detecting contamination prior to sequencing, and the infeasibility to examine every raw sequence chromatogram produced from an NGS run, leads to the dangerous potential for undetected cross-contamination in many studies. Because NGS data sets are very large, the main check of sequence quality comes from per base quality scores. While these scores are useful in identifying bases that may not have high sequence quality, they are not able to identify contamination in the same way that Sanger sequencing is via the presence of multiple peaks on sequence chromatograms. Thus, in NGS datasets, detecting sites where contaminated DNA might be present becomes very difficult without stringent and specific checks for contamination (Longo et al., 2011). It has been shown that contamination may be an issue for at least 10% of genetic databases deposited at NCBI (Longo et al., 2011). Because of the difficulty of detecting contamination in NGS datasets, if it is not explicitly and stringently tested for, analyses in which the investigators believe a data set to be "good" can go forward, despite the contaminated data set, leading to published erroneous phylogenies (Simion et al., 2018). Not only is this a problem when contaminated data are used to publish phylogenies, but with public data sharing, it is possible that contaminated data sets get used in larger studies by other researchers who may be even less aware of potential contamination.

Detection of aberrant samples, either from cross-contamination or simple mislabeling, typically relies on some *a priori* hypotheses of relationships. For example, comparison of phylogenetic tree topologies between a previous phylogeny for the group and a phylogeny inferred using sequence data from a new study could reveal contamination when certain samples

resolve in different places on the two trees. However, topology-based methods cannot always distinguish between different phylogenetic information being and contamination. Even in uncontaminated datasets, there can be discordance between gene trees and species trees simply because each gene has a unique evolutionary history. Incomplete lineage sorting, introgression, and other phenomena can all lead to discordance between gene trees and species trees (Felsenstein, 2004). Modern methods for inferring species trees should be robust to gene tree discordance, but it can still be difficult to tell if topology differences are due to gene tree discordance or contamination of the dataset (Arcila et al., 2021).

While topology-based methods may have difficulties identifying contamination, there can be clues pointing to potential contamination if there are dramatically different placements of taxa that do not have controversial placements in published trees. Topology-based methods are also useful for identifying sample mislabeling when two samples appear in reciprocally wrong places on a tree. Cross-contaminated samples meanwhile may have a different signal. Contaminated samples in a target enrichment study may contain loci that are entirely original source DNA, or entirely contaminating source DNA, or assembled from a mixture of reads from both original and contaminating sources. In these cases, a species tree inferred from multiple loci would result in sequence data that would be a mix of the correct sample and the contaminated sample, which could result in an unclear tree topology where samples may fall out in unexpected but not obviously wrong places on the tree.

Alternatively, contaminated individuals may behave in a tree in a way similar to ‘rogue’ taxa, where when included, will consistently lower resolution across the tree simply due to their problematic placement. Historically, when ‘rogue’ taxa have been identified in a study, it has been understood that they are behaving strangely due to a unique evolutionary history when

compared to the rest of the tree, such as vastly different mutation rates, extensive incomplete lineage sorting or introgression, or extremely long branches without close relatives (i.e., 'living fossils'), and other phenomena. Researchers often opt for identifying 'rogue' taxa and removing them iteratively in order to have a well resolved tree with the exclusion of only one taxon instead a poorly supported tree with all taxa included (e.g., Aberer et al., 2013; Betancur-R et al., 2013; Rabosky et al, 2018). It can be expected that contaminated individuals may behave in a similar way as they may represent sequences from a mix of specimens, representing an incorrect evolutionary history, causing placement on the tree to be very uncertain, lowering support over the whole tree. Therefore, iterative identification and removal of potentially contaminated taxa may be expected to increase support and resolution across a tree.

In addition to tree-based approaches, another method that uses *a priori* hypotheses of relationships to identify cross-contamination performs an analysis of pairwise genetic distances (*p*-distance) of loci between pairs of taxa in a tree, whereby sample pairs that are hypothesized to be closely related are expected to have smaller *p*-distances and samples that are not expected to be closely related would have larger *p*-distances. A contaminated sample could be found using this method if a large proportion of loci were similar to another sample that is hypothesized to be very distantly related, and conversely, we might see that many loci were very different from those of closely related species. This method relies on not only having accurate *a priori* hypotheses of relationships, but also on having an arbitrary cut-off value specifying what level of genetic similarity would be indicative of contamination. This approach also requires that the loci being tested will have adequate variation to have informative *p*-distances. For highly conserved loci, especially among closely related species, it becomes difficult to determine if a sequence has been contaminated by a different specimen or if sequences are simply identical between

specimens due actual shared evolutionary history (Simeon et al., 2020). Additionally, hypervariable loci could fail using this method as they may be so variable that closely related species pairs and distantly related species pairs have similar *p*-distances due to substitution saturation.

An approach for detecting contamination that does not require *a priori* hypotheses of relationships is to compare sequences from a target capture dataset to sequences available in public repositories (Arcila et al., 2021). There are sets of genes, both nuclear and mitochondrial that have been widely used in phylogenetic studies and as such have known sequences published in public repositories such as GenBank and BOLD (Betancur-R et al., 2013; Li et al., 2007). These genes, which in recent years have been referred to as legacy markers, are frequently included in the suite of loci targeted by NGS studies (Hughes et al., 2021; Jiang et al., 2019; Li et al., 2013). In cases of suspected contamination, legacy markers can be extracted from NGS datasets and their sequences can be compared to reliably identified sequences from public repositories to verify the identity of the exon-capture sequence or determine if the sample may be contaminated. This approach relies on not only including the legacy markers in the probe set for the study, but also that the species in question are studied well enough to have sequences of legacy markers published from specimens that are correctly identified.

Finally, an approach that is less widely used in phylogenetic datasets derived from target capture methods is checking for anomalous allele frequencies at heterozygous sites for each individual taxa in the dataset. For diploid individuals, allele frequencies should either be 0:1 or 50:50 at a single heterozygous site. A contamination signal here could appear in a few different ways; a potentially heterozygous site having three or more separate alleles for a diploid individual, or allele frequencies differing significantly from the expected 0:1 or 50:50 because

alleles from multiple individuals are mixed together. However, this method can only signal contamination when sequences from both the correct and the contaminating samples are present at that particular heterozygous gene/site. It would not be able to detect an entire locus represented by a single contaminating individual, as the allele frequencies should behave as expected in such a case.

In order to accurately estimate allele frequencies at a heterozygous site, the locus must be sequenced with sufficient read depth in order to be sure that deviations from expected allele frequencies are not due to chance or sequencing error (Nielsen et al., 2011). Reliable genotyping of SNPs from NGS data has known difficulties; many algorithms and programs have been developed to allow for statistical quantification of certainty in genotype calls (Nielsen et al., 2011, Sukumar et al., 2021). But at the simplest level, allele frequencies can be estimated by simply counting numbers of raw reads representing each variant sequence; this method especially relies on sufficient read depth at a site to have any confidence in allele frequency estimates. When using NGS, there is a tradeoff between how many loci and samples that can fit in a single sequencing reaction versus the read depth per locus/sample combination. Current priorities in phylogenetics have led to maximizing the number of loci and samples per lane and getting by with the minimum read depth needed for researchers to feel confident in base calls, given the rates of sequencing error of the NGS platform. However, it is possible that current practices do not result in enough read depth to confidently estimate allele frequencies at heterozygous sites. Moreover, most assembly pipelines for systematic work do not consider heterozygotes at all and instead pick a single allele to represent each locus (Arcila et al., 2021, Yuan et al., 2019), so achieving sufficient read depth for genotyping has not previously been a priority.

In addition to benchtop tradeoffs of how many samples to pool per lane, bioinformatic assembly processes need to also come into consideration. Different bioinformatic pipelines take different approaches to assembling raw data, including different *de-novo* assemblers (Fig. 2.1; e.g., Trinity vs. SGA vs. other *de novo* or reference-based assemblers). Pipelines differ dramatically not only in the number of loci that are able to be assembled from raw data, but also the read depth per locus (Yuan et al., 2019), which in turn can subsequently inhibit the ability to detect contamination in the form of aberrant allele frequencies at heterozygous sites.

In this study we explore the potential for detecting contamination in a NGS dataset of exon capture sequences from species from a closely related bony fish family. The species come from the family Liparidae, the snailfishes, which are relatively young evolutionarily, with a hypothesized origin of the family less than 25 mya (Near et al., 2013 Fig. S2) and may have low standing genetic variation due to their recent divergence. The study design followed the protocols of Li et al. (2013) where the potential for benchtop cross-contamination was very high. It appears that there is widespread contamination throughout the dataset based on certain attributes of the phylogenies inferred from the data, including having very low support values for internal branches, failing to recover members of the same species as sister, and topologies that are very different from published phylogenies (Fig. 2.2–2.6; Orr et al., 2019). Here we explore different methods for detecting and removing the contamination in this dataset, assess the utility of each method, and make suggestions for future work in groups that may possess similar difficulties.

MATERIALS AND METHODS

Sample collection. Specimens were collected primarily during survey operations conducted in the eastern North Pacific by the Resource Assessment and Conservation Engineering (RACE)

Division of the U.S. National Marine Fisheries Service, Alaska Fisheries Science Center (AFSC). Tissue samples were taken from fresh whole specimens at sea or from frozen specimens in the laboratory and preserved in 95% ethanol for long term storage at -80°C. Most whole specimens were fixed in 10% formalin and stored in 70% ethanol; some specimens were fixed and stored in 95% ethanol. Most specimens and tissues were archived at the Burke Museum of Natural History and Culture, University of Washington Fish Collection (UWFC). One specimen of *Careproctus iacchus* came as a loan from the Kyoto University Museum (Table 2.1).

Exon capture. Exon-sequence data were obtained for 99 specimens, representing 64 liparid species in 12 genera, and two specimens representing one outgroup species, the cyclopterid *Eumicrotremus orbis*. DNA was extracted from tissue samples using Qiagen DNEasy blood and tissue kits (Valencia, CA, USA) using individual spin columns. Extractions were performed following manufacturer directions, except samples were digested for at least eight hours at room temperature or at 56°C and eluted to a final volume of 150 µl in two or three washes. DNA extractions were performed at the Alaska Fisheries Science Center (AFSC) and at the University of Washington Molecular Ecology Research Laboratory (MERLAB). Extractions performed at AFSC were stored at -20°C for up to four years.

Exon sequences were generated following the gene-capture protocol of Li et al. (2013), targeting 4,434 putatively single copy nuclear exons identified by Jiang et al. (2019). At the time this work was done, there were no whole genomes of Liparidae to design probes from, but the double hybridization method of gene-capture has shown to be effective at recovering large quantities of genomic data from non-model organisms using probes designed from model organisms of the same class (Li et al., 2013; Kuang et al., 2018). Thus, the baits used in this study were designed from *Oreochromis niloticus* (Cichlidae), the most closely related species to

Liparidae of the eight species with well-annotated genomes used to design the bait sets for the 4,434 markers (Jiang et al., 2019). The size range of these markers was 102 to 5,803 bp with a mean size of 261 bp and a total concatenated alignment size of 1,157,304 bp.

All target-capture bench work was conducted at the MERLAB using baits and gene-capture reagents from the Arbor Biosciences myBaits Hybridization Capture Kit. DNA extractions were quantified and visualized on a 1% agarose gel to check size distribution, and then sonicated using a Diagenode Bioruptor Sonicator (Diagenode USA, Denville, NJ, USA) to ~200-500 bp prior to library preparation. Library preparation, followed by double capture hybridization was performed following Li et al. (2013). This method involves using overlapping baits 90 bp in length to fully cover the size of each of the 4,434 target markers. Broadly the steps of the protocol were size selection, library preparation, hybridization PCR to anneal baits, performed twice, and a final index PCR to anneal indexes before pooling. Each of the steps listed involved at least one wash using magnetic beads per step, with library preparation and hybridization PCR each using two wash steps, resulting in at least seven wash steps throughout the protocol. Prepared samples were grouped by similarity in final DNA concentration and pooled into two pools of 52 and 46 samples each. Sequencing was performed by the University of Oregon Center for Quantitative Life Science (Corvallis, Or; https://cqls.oregonstate.edu/core/sequencing) on two lanes of an Illumina HiSeq 3000 using paired-end 150 bp reads.

Raw reads were assembled into loci using the Assexon bioinformatics pipeline from Yuan et al. (2019) (Fig. 2.1), where first adapters and low-quality reads were trimmed using Trim Galore (https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/), duplicate sequences were removed, and then remaining reads were mapped to each locus by iteratively

blasting against the sequences of reference markers. The sorted reads were then assembled into contigs iteratively for each gene and sample using String Graph Assembler (SGA) (Simpson and Durbin, 2012), and further assembled by the Assexon perl script merge.pl. Potential paralogues were identified using the perl script reblast.pl. Two samples failed to assemble and were removed from the dataset. After removing those samples, 4,284 of the 4,343 targeted markers were captured by at least one of the 97 samples. Two samples contained <1,000 genes but were retained in the dataset for initial analyses.

To minimize the effect of missing data in tree inference, the dataset was filtered to remove genes present in too few samples. In order to evaluate the effects of missing data, two separate thresholds were used, generating two different data sets. First, genes present in at least 50% of the samples were retained resulting in a larger data set with 2,917 genes kept. A second, more stringent, threshold of retaining genes present in at least 70% of the samples resulted in 2,282 genes kept. Loci from each taxon were then aligned on amino acids using MAFFT (Katoh et al., 2002) and translated back to codon-based alignment using the perl script mafft_aln.pl. A random subset of ~250 alignments were manually inspected using Geneious prime to look for obviously misaligned sequences. For the larger dataset, sequences were similar across taxa, with a mean *p*-distance of 0.03 and standard deviation of 0.02 (Fig. 2.7).

To evaluate whether the problematic topologies and suspected contamination were an artifact of poor assembly performance, data were also assembled following Arcila et al. (2021) (Fig. 2.1) and assembling using the FishLife exon markers (Hughes et al., 2021), which are a subset of the markers in the original Jiang et al. (2019) probe set. Low-quality bases and adapters were removed using Trimmomatic v0.36 and trimmed reads were mapped against 1105 FishLife exon markers using BWA-MEM (Li and Durbin 2009). PCR duplicates were removed using

SAMtools v1.9 (Li et al., 2009). Mapped reads were assembled for each exon using Velvet v1.2.10. The longest contig produced was using as a reference sequence for aTRAM v2.0 (Allen et al., 2018) to conduct *de novo* assembly using Trinity v2.8.5. Redundant contigs were removed with CD-HIT-EST with a 99% similarity threshold. Identification of open reading frames was conducted with Exonerate using reference sequence alignments and visual inspection in Geneious prime. Contigs without open reading frames were filtered and not included in downstream analysis resulting in 956 of 1,105 FishLife exon markers kept.

Tree inference and comparison. Phylogenies were built using the two Assexon assembled filtered data sets with different numbers of retained genes as well as the FishLife assembled filtered data set. For each dataset, alignments for each gene were concatenated and used to infer maximum likelihood (ML) trees using RAxML (Stamatakis, 2014) with 100 bootstrap (BS) iterations under the GTRCAT model with the *-k* marker to record branch lengths for each iteration. All 100 BS iterations were then used to create a majority consensus tree with branch lengths using the 'consensus.edges' function from the phytools package (Revell, 2012) in R version 4.1.1(R Core Team, 2021). Phylogenies were visualized using FigTree v 1.4.4.

Constructed phylogenies were visually compared to two phylogenies of Orr et al. (2019). Both phylogenies from that study were inferred under maximum likelihood and Bayesian inference. One tree was based on 270 sequences, each 490 bp in length, of cytochrome oxidase I (COI), representing at least 122 liparid species and 10 individuals representing putative undescribed species (Fig. 2.2A). The second phylogeny was inferred based on 355 RADseq loci (33,370 bp of concatenated data) sequenced from 40 samples (25 ingroup species and one outgroup (Fig. 2.2B). Three main clades shared across these phylogenies were used for broad comparisons in this study: the *Liparis* clade, the Paraliparia clade, and the Aenigmoliparia clade.

These clades were identified on all trees visually using color codes of green, orange, and blue respectively (Fig. 2.2). While three additional larger resolution clades were recovered by Orr et al. (2019), only one species from any of these clades, *Crystallichthys cyclospilus*, was included in this analysis so only the placement of that species was considered informative for comparison (color code: pink). Two additional individual species placements were considered as informative for comparison due to their placement as being sister to most of the rest of the tree (i.e., near the 'base' of the tree); *Liparis fucensis* (color code: dark green) and *Nectoliparis pelagicus* (color code: red) (Fig. 2.2). Finer resolution clades were used to identify species as close relatives in subsequent analyses (Fig. 2.2 collapsed clades).

All three trees indicated some evidence of contamination based on low support for internal branches, failure to recover conspecifics as sister taxa throughout the tree, and failure to recover clade groupings found in the trees of Orr et al. (2019) (Fig. 2.2). The trees inferred based on the Assexon assembled data and 2,917 genes recovers all liparids as a monophyletic group but intrafamily relationships are poorly supported and do not match expected topology (Fig. 2.3). Two species, *N. pelagicus* and *L. fucensis*, which were sister to all other liparids in the Orr et al. (2019) phylogenies, are recovered deeply nested within clades of other species with relatively strong support for this placement. *Crystallichthys cyclospilus* and one individual of *Paraliparis mento* are instead sister to the rest of the liparids, with the two individuals of *C. cyclospilus* not being sister to each other. The expected larger clades of Aenigmoliparia, Paraliparia, and *Liparis* were not well recovered, with members of all clades mixed together in large polytomies in the majority rule (50%) consensus tree. The smaller clades nested within the larger clades (Fig. 2.2, collapsed clades), which were recovered with strong support in previous phylogenies, were not recovered in any of our analyses with the exception of the Elassodisca clade (Fig. 2.2, 2.3).

These patterns of low resolution and apparent widespread cross-contamination also appeared in the tree inferred from the FishLife assembled data with 956 genes used (Fig. 2.4), although topology between this tree and the tree from the Assexon assembled data (Fig. 2.3), did not match completely.

Identify contaminated individuals and loci. Based on low branch support, recovery of conspecifics not as sister taxa, and failure to recover expected clades throughout the resulting trees, we believed there to be cross-contamination or sample mislabeling present in the dataset. These possibilities were investigated in multiple different ways.

Gene Trees–To further reduce any potential the effect of missing data on tree topology, a tree was inferred based on concatenation of only the 251 genes found in all 97 samples from the data assembled using the Assexon pipeline (Fig. 2.5). Additionally, individual gene trees were constructed for each using the same RAxML methods. Gene tree topologies were then visually compared with the expected topology (Fig. 2.2) and the topology of the tree inferred from the concatenation of 2,917 genes (Fig. 2.3) to attempt to identify genes with particularly aberrant topologies or genes that came closer to matching the expected topology.

Comparison to published sequences–For Liparidae, cytochrome oxidase I (COI) is the most commonly published marker in public repositories (Appendix 1; Orr et al., 2019). COI was not targeted by the bait set used in this study, but due to the prevalence of mitochondrial DNA, sequences are often present as 'bycatch' in datasets. COI sequences were mined from the raw data using BWA v0.7.17 (Li and Durbin, 2009) and SAMtools (Li et al., 2009). Mined sequences

were compared to COI sequences from Orr et al. (2019) using BLAST+ v2.12.0 (Camacho et al., 2009) to verify sequence identity.

Detecting contamination using a priori *hypotheses of relatedness*–Potentially contaminated individuals were identified by providing *a priori* relationships and calculating relative distance using the detect_contamination.pl script from the Assexon pipeline. The script requires as input a user defined table of relationships between samples (i.e., "known" phylogenetic groupings; e.g. genera, previously well-supported clades of closely related species, known species complexes, multiple taxa of the same species, etc.), and then calculates *p*-distance between all pairs of samples not defined as closely related for each gene alignment. A sample was considered potentially contaminated when the *p*-distance between the sample and an unrelated sample is less than 0.002 for at least 80% coverage between the two sequences for a locus that has informative variation present (*p*-distances for all pairs at the locus > 0.002). The output of the analysis was a table listing species pairs and the number of genes identified as contaminated between that pair, standardized by the number of times that pair has sequences for the same gene.

The Assexon detect_contamination.pl script was run first only identifying conspecifics as close relatives, and then by using the smallest resolution clades defined by Orr et al. (2019) (Fig. 2.2, collapsed clades). The output of the clade-based detection identified 4,458 species pairs with potential contamination. The five specimens with the highest level of contamination from the clade-based detection were then removed and the filtering, alignment, and tree building steps were then redone with that data set, maintaining a filtering threshold of genes present in at least 49 specimens. The tree produced was then examined for unexpected relationships and samples were chosen for removal based on placement in the tree compared to the phylogenies of Orr et al.

(2019) (Fig. 2.2), with the hope that removing 'rogue' contaminated taxa would increase resolution of the tree. This was done twice iteratively, which resulted in the removal of 13 samples from the final dataset.

Identical Sequences–If samples were contaminated in such a way that DNA from the contaminating sample was sequenced in the place of the intended sample, we would expect the sequence data of the intended sample to be identical to the contaminating source, although the direction of contamination may not be apparent. Collectively, this would appear as two unrelated samples having more identical sequences than two closely related samples. To investigate this, we calculated the percentage of identical sequences found in each pairwise comparison of samples by calculating the number of genes in which sequences for a pair were identical (*p*-distance = 0) standardized by the number of genes sequenced in both samples of the pair. This approach is similar to the detect_contamination.pl script but specifically is looking for completely identical sequences (*p*-distance = 0), not just very similar sequences (*p*-distance < 0.002). Percentage of identical sequences was compared between conspecifics, members of the same Orr et al. (2019) small clade (Fig. 2.2, collapsed clades), and pairs of species not from the same clade to identify less-related samples with high proportions of identical sequences. We would expect that the proportion of identical sequences between any pair of taxa would be highest among conspecifics, followed by species in the same clade from Orr et al. (2019), followed by species from different clades of Orr et al. (2019). This was tested using three paired *t*-tests comparing each pair of groups (same species vs. same clade, same species vs. not same clade, same clade vs. not same clade).

Heterozygosity–We inferred allele frequencies for each putatively heterozygous site in the assembled contigs of each individual sample at each gene locus. An individual sample represents DNA from a single individual snailfish, but because there are multiple raw reads mapped to each individual locus, our null assumption was that any site with more than one possible nucleotide mapped to that position was either a heterozygous site (two alleles present in a 50:50 frequency) or possibly due to sequencing error. In the case of sequencing error, one 'allele' should be very rare (<1% of reads, based on sequencing error rate of the Illumina platform) (Stoler and Nekrutenko, 2021). Variable sites with more than two alleles or allele frequencies that differed from 50:50 may be evidence of cross-contamination.

To calculate allele frequencies from assembled reads, we used a "pseudo-reference genome" approach. A concatenated file of all loci assembled from the Assexon pipeline was constructed for each individual and used as a pseudo-reference genome. Raw reads from each individual were then aligned to their respective pseudo-reference genome using Bowtie2 v.2.4.4 (Langmead et al., 2019; Langmead and Salzberg, 2012). Resulting binary alignment and map (BAM) files were converted to sequence alignment and map (SAM) files, short reads, and low-quality reads were removed, and files were sorted and indexed using SAMtools v1.13 (Li et al., 2009). Variant sites were identified using mpileup and SNP variant calling was done using call, both from bcftools version 1.13 (Danecek et al., 2021). Resulting vcf files were analyzed using the vcfR packaged (Knaus and Grünwald, 2017) in R (R Core Team, 2021). Observed minor and third allele frequencies were calculated based on counts of high-quality reads for alleles one and two compared to total read depth of high-quality reads for each polymorphic site. Sites with less than 20 high quality reads and with minor allele frequency less than 0.05 were removed. To estimate overall allele frequencies for each gene, the SNP loci remaining after filtering were

combined based on gene name and mean minor-allele frequency was calculated for all SNPs combined for that gene. Sites where allele frequencies strongly differed from 50:50 were considered potentially contaminated loci. This was done for all 97 individuals successfully assembled with Assexon as well as three individuals from an uncontaminated exon-capture dataset from Atta et al. (2022) that used the same markers and was also assembled with Assexon.

RESULTS

All trees produced by the analysis of the two datasets assembled using Assexon, either using 2,917 (Fig. 2.3) or 2,282 retained genes, and the dataset assembled using FishLife exons with 956 retained genes (Fig. 2.4), had topologies that were very poorly resolved and were strikingly different from each other and from the phylogenies from Orr et al. (2019) (Fig. 2.2). The only clade from Orr et al. (2019) that was recovered in these trees is the Elassodisca clade. The samples from this analysis that were expected to, and did, fall within this clade are two specimens of *Careproctus bowersianus*, two specimens of *Elassodiscus* sp. cf. *caudatus*, and two specimens of *Elassodiscus tremebundus* (Fig. 2.3, 2.4). Furthermore, other aspects of the tree may indicate of wide-scale cross-contamination. Most terminal tips on the trees had very long branch lengths, even between the same species. Trees overall had low bootstrap support, especially for larger clades, and clear issues with resolving relationships within those larger clades (Fig. 2.3, 2.4). In most cases, members of the same species were not recovered as sister to each other (Fig. 2.3, 2.4).

Identify contaminated individuals and loci. The different methods attempted identified the presence of widespread contamination throughout the dataset but were unsuccessful in removing contamination in a meaningful way.

Comparison to published sequences–COI sequence data was mined from raw sequences for 30 specimens (Table 2.2). Of those, five were not useable and had no BLAST hits due to short length of sequences recovered (longest fragment was <100 bp). Of the remaining 25 sequences, the top hit was the expected species for 16 of the specimens. A further four specimens did not have sequences for their species present in the dataset of COI sequences (Orr et al. 2019) and therefore did not have high matching scores as was expected. Finally, five specimens had the top blast hit be something other than the expected sequence. *Careproctus faunus* (UW 156084) matched most closely with its sister species *C. comus* and then *C. faunus* as the second hit. This was not flagged as contamination due to the sister relationship of the species and similarity in the sequences. Two specimens of *Liparis gibbus* both matched most closely with *L. fabricii* and may represent misidentified specimens but do not appear contaminated based on downstream analyses. One specimen of *L. ochotensis* (UW 49438) matched most closely with *L. bristolensis*. This mismatch was attributed to the fact that the sequence mined from *L. ochotensis* only aligned to a short region of the reference sequences from Orr et al. (2019), and thus, *L. ochotensis* was not considered contaminated. Finally, *Rhinoliparis barbulifer* (UW 157184) had a top hit matching *C. comus* and was considered contaminated and removed from downstream analyses (Table 2.1).

Gene Trees–Visual examination of individual gene trees for the 251 genes present in all 97 specimens did not reveal any gene trees whose topology came closer to the expected topology (Fig. 2.2; Orr et al., 2019). The gene trees displayed similar problems to the trees generated using 2,917 Assexon genes or 956 FishLife markers (Fig. 2.3, 2.4) with low support for internal

branches and failure to recover monophyletic species or larger clades. Gene tree topologies differed from the expected topology in many different ways and no individual tree appeared drastically more different than any other. The species tree based on the concatenation of the 251 fully sampled genes also suffered from the same problems and was even worse at resolving conspecific relationships than the trees based on the broader data set (Fig. 2.5). In this tree, even fewer pairs of conspecifics were recovered as sister and Elassodisca, the only clade of Orr et al. (2019) recovered by the larger datasets, was not recovered. Rather, one member of that clade, *Elassodiscus* sp. cf. *caudatus*, was recovered as sister to all other liparids, while its conspecific was recovered nested within the largely unresolved sister clade (Fig. 2.5).

Detecting contamination using a priori hypotheses of relatedness–Running the detect_contamination.pl script with only conspecifics defined as close relatives resulted in a remarkable 4614 species pairs identified as contaminated, with an average of 11.2% of loci potentially contaminated between pairs. Running the script with close relatives defined by the Orr et al. (2019) clades (Fig. 2.2, collapsed clades) resulted in 4,458 species pairs identified as contaminated with a mean contamination of 10.9% of loci being potentially contaminated between pairs.

The top 5% of contaminated samples, five samples with > 16.5% contamination, were removed from downstream analysis. Visual inspection of the tree inferred from the resulting data set based on 92 specimens and 2,917 genes led to identification of three more ‘rogue’ specimens that were likely contaminated based on their highly supported placement in very unexpected spots on the tree. These specimens were removed, and the tree produced based on the resulting 89 specimens and 2,917 genes led to the identification of a further five ‘rogue’ specimens that

might be contaminated based on the same criteria used in the previous tree. A final tree was produced based on 84 specimens (N=13 specimens total removed throughout this process), and 2,917 genes (Fig. 2.6). This resulting tree topology differed from the tree produced based on all 97 specimens and 2,917 genes (Fig. 2.3) in a few key ways. Of the 21 species pairs that failed to resolve as sisters in the tree based on all 97 specimens, only 14 failed to resolve as sisters in the tree based on 84 specimens (Fig. 2.6). This is due to two species pairs (*Paraliparis ulochir* and *Careproctus ectenes*) resolving as sister and the remaining five species pairs having either one or both members removed. Similar to the tree based on 97 specimens, the tree based on 84 specimens recovers the Elassodisca clade (Orr et al., 2019) with 100% bootstrap support (Fig. 2.6). But the larger clades of Orr et al. (2019) are still not resolved and internal branch support is still quite low resulting in large polytomies (Fig. 2.6). The iterative process of 'rogue' specimen removal was stopped after 13 specimens when it became apparent the tree topology was not improving dramatically and removal of further specimens would likely not change this.

Identical Sequences–The percentage of identical sequences between samples was compared between pairs of samples: 1) within the same species, 2) within the same Orr et al. (2019) clade, and 3) not within the same Orr et al. (2019) clade (Fig. 2.8). There was no significant difference in the proportion of identical sequences between members of the same species (mean 25.2%) and members of the same clade (mean 23.6%) (paired *t*-test; $p > 0.05$) (Fig. 2.8). Species not in the same clades of Orr et al. (2019) had a significantly lower proportion identical sequences (mean 14.4%) when compared to members of the same species (paired *t*-test; $p < 0.05$) and members of the same clade (paired *t*-test; $p < 0.05$) (Fig. 2.8). Despite the significant difference in proportion of identical sequences in comparisons of related versus less related species pairs, this method

was not able to identify any specific species pairs as highly contaminated, as the ranges of proportion of identical sequences between related and unrelated groups overlapped so greatly (Fig. 2.8).

Heterozygosity–Average read depth over all 105,884 putative heterozygous sites across all taxa and all genes was 71, with 45% of sites having fewer than 20 reads (Fig. 2.9). Only 11 putatively heterozygous sites had read depth greater than 300, and only 16% of putatively heterozygous sites had read depths greater than 200. When sites with fewer than 20 reads were removed, 61,204 putatively heterozygous sites remained, with an average read depth of 117 and a standard deviation of 88. Average number of heterozygous SNPs per individual was 637.5, representing an average of 72 genes per individual. Observed rates of occurrence of minor alleles were expected to have peaks near 0.5, as any homozygotes with occasional erroneous base calls would be filtered out by removing allele frequencies less than 0.05. When sites with fewer than 20 reads and minor allele frequency less than 0.05 were removed, 28,804 putatively heterozygous sites remained. Unfiltered minor allele frequencies were low, with values greater than 0.1 nearly uniformly distributed between 0.1 and 0.5 (Fig. 2.10 B). In contrast, the uncontaminated flatfish data from Atta et al. (2022) showed the expected distribution with the peak in allele frequencies being at 0.5 with some variation around this, likely due to differences in read depths (Fig 2.10 A). For the contaminated snailfish data, the mean minor allele frequency after filtering was 0.24 with a standard deviation of 0.13. This pattern was attributed to widespread contamination, resulting in a complete lack of allele frequencies converging around 0.5, which is extremely unlikely to be due to biological phenomena or sequencing error. The widespread nature of the contamination made it impossible to determine specific sites with contamination. If low read

depth was explaining our inability to see allele frequencies close to 0.5, then we might expect that loci with greater read depth would be more likely to converge on normal allele frequencies. However, when read depth was plotted against allele frequency, the expected trend of loci with higher read depth pooling more strongly near rates of 0 or 0.5 was not found (Fig. 2.11). The quadratic line of best fit is nearly flat with an adjusted $r^2 = 0.03$, reinforcing the conclusion that the abnormal frequencies at our heterozygous sites is likely due to contamination and not lack of read depth, as sites with high read depth still have allele frequencies that do not center near zero or 0.5 (Fig. 2.11).

DISCUSSION

Next generation sequencing (NGS) allows for massive parallel sequencing of DNA and can result in very large datasets that are informative for phylogenetic analysis. Target capture methods are currently being widely applied to infer phylogenies across different groups and at different scales (e.g., Arcila et al., 2021; Atta et al., 2022; Betancur-R et al., 2017; Hughes et al., 2018). Within the study of ray-finned fishes (Actinopterygii), exon capture is an especially attractive method as a large number of orthologous loci that can be targeted in distantly related species have been identified (Hughes et al., 2021; Jiang et al., 2019). Benchtop library preparation methods as well as bioinformatic assembly methods are also improving in ways that streamline data generation, lessen the likelihood of cross-contamination, and develop new ways of detecting cross-contamination in sequenced data (Arcila et al., 2021; Wang et al., 2022). However, the current practices for detecting cross-contamination in exon-capture datasets may not be sufficient for all datasets depending on the species or groups being sequenced. The majority of contamination detection methods require *a priori* hypotheses of sample relationships, that can be used in a number of ways including checking *p*-distances between less related

samples (Yuan et al., 2019), comparing tree topologies (Arcila et al., 2021), and verifying sample identification via comparison of markers in the dataset with existing published sequences (Arcila et al., 2021). These methods may be insufficient in groups where relationships are less well known and published sequences are less reliable.

For Liparidae, there remains a number of taxa and finer resolution groups for which phylogenetic placement is uncertain but overall, there are larger clades that are relatively strongly supported that can be very useful for comparison of tree topologies (Fig 2). Orr et al. (2019) recovered six clades that are well supported with three, Aenigmoliparia, Paraliparia, and the *Liparis* clade, containing most of the taxa. Failure to recover these clades in any iteration of our data, including both datasets assembled with different methods (Assexon vs. FishLife), was evidence pointing towards a widespread level of contamination causing dramatically different placements of taxa (Fig 4,5). However, the tree topology alone could not identify if there were specific individuals that were the contaminated and behaving like 'rogue' taxa in the tree due to the dramatically different placements.

Given the widespread nature of the contamination it was not possible to identify which individuals were the most contaminated and the source of their contamination based on topology comparison alone. A more direct, albeit haphazard, approach to identifying contamination is to compare sequences of any 'legacy markers' in our dataset to published sequences with confirmed IDs on GenBank. The marker with the most published sequences for snailfishes is cytochrome oxidase I (COI) (Appendix 1; Orr et al., 2019), but as COI was not specifically targeted by the bait set used in this study it instead had to be mined from raw data with the hope that it was sequenced incidentally. This mining resulted in useable COI sequences for only 25 out of 97 taxa. Additionally, as snailfish are notoriously difficult to identify accurately to species,

published sequence data can be unreliably identified to species, and a number of sequences are indeed published with the wrong species identification as noted by Orr et al. (2019), making identification verification using sequences from public databases unreliable. This was circumvented by comparing mined COI sequences only to sequences from Orr et al. (2019) where identification was verified based on voucher specimens for every sequence published. However, this leads to another issue because these reliable sequences are relatively short (492 bp). This meant that COI sequences recovered from exon capture were not fully 1:1 comparable to already published sequences as they did not overlap completely when aligned. Even with these difficulties, we were able to verify that one specimen, *Rhinoliparis barbulifer*, was clearly contaminated, as the closest blast hit of the COI sequence was *Careproctus comus*, a very different species. In order to make this method more useful in the future, for Liparidae specifically, it would be useful to specifically target the 492 bp region of COI for which reliably identified sequences are published (Orr et al., 2019). A final caveat with the method of using COI to verify species, is that contamination may result in a mixture of genes being amplified from the target species as well as a contaminating species, which means COI could be amplified from the target species but the sample may still be have contaminated sequences present for other genes.

Because of our limited ability to rely on sequences in public databases to identify contamination, we instead attempted to quantify relative levels of contamination by comparing *p*-distances between closely related and less related species. The output of the detect_contamination.pl script returns the percentage of gene pairs that may be contaminated by comparing the number of pairs flagged as contaminated for falling below the *p*-distance threshold to the total number of gene pairs for the two taxa in question. This standardization of

contamination allowed for ranking the relative levels of contamination in order to determine the most contaminated taxa. Comparison of sequence similarity can still be a problem for this group because the family has recently diverged (< 25 mya) and is rapidly speciating (Near et al., 2013; Rabosky et al., 2018), meaning true genetic distance between distantly related species may still be low, which makes it difficult to determine if sequences in our data are very similar due to shared evolutionary history or due to contamination. This was the case here, as the majority of *p*-distances between all taxa were below 0.05 (Fig. 2.7) and there were nearly the same proportion of identical sequences between species defined as closely related as there were between less closely related species (Fig. 2.8). However, ranking contamination based on relative *p*-distance measures was the most useful way to begin identifying taxa for iterative removal from the tree. After the first removal of highly ranked individuals, tree topology changed enough that additional sets of taxa that were likely contaminated could be removed. Iterative removal resulted in a slightly more resolved tree (Fig 6), where some taxa from the same species resolved as sisters for the first time, but given widespread contamination the iterative removal process would have likely continued until the vast majority of taxa were removed before the tree was well resolved.

Identifying contamination via anomalous individual heterozygosity at each locus has the potential to circumvent problems with the above methods, at least in diploid organisms, because *a priori* hypotheses of relationships and positively identified published sequences are not needed. Instead, all that is needed is to accurately calculate allele frequencies at heterozygous sites for each individual at each locus and determine which loci for which individuals significantly deviate from expected allele frequencies of 50:50. While reliably genotyping SNPs from NGS data has known difficulties, there are many methods beyond estimating ratios from strict counts

that allow for statistical quantification of certainty in genotype calls (Nielsen et al., 2011) and would allow for recognition of potentially anomalous heterozygosity calls. Even simply estimating ratios from strict counts of allele occurrence in our raw reads, we were able to detect widespread contamination in our data set, especially when compared to an uncontaminated dataset (Fig. 2.10). In the uncontaminated data from Atta et al. (2022) the allele frequencies are distributed around 0.5, whereas in our contaminated data allele frequencies pool below 0.1 with a long tail to 0.5, indicating widespread, low levels of contamination.

Current assembly methods do not retain information on heterozygosity. For heterozygous loci these methods instead randomly select one of the alleles to represent that locus (Arcila et al., 2021, Yuan et al., 2019). To fully utilize the heterozygosity method for detecting contamination it would be best to incorporate it into the assembly step in such a way that when reads are mapped to target exon sequences at the start of the process, information would be retained about heterozygous loci within each gene before choosing a random allele to use as the reference contig (Fig 1). Retaining heterozygosity information at the assembly step would also have applications beyond detecting contamination, making exon capture even more useful for studies beyond broadscale phylogenetics such as population genetics and species delimitation.

Overall, we found that widespread contamination can be detected by a number of different methods, but removing contamination is more problematic when it is rampant throughout the dataset. Conversely, it follows that less widespread contamination may be easier to remove (i.e., by removing individual contaminated taxa). In the case of individual contaminated taxa, it could be expected that using various detection methods would result in the taxa being identified as contaminated by each method, improving confidence in the detection of contamination. But,

depending on the nature of contamination, it may be more difficult to detect and thus should be explicitly tested for in all datasets even if they appear clean.

If this study were to be redone in order to successfully infer a phylogeny for Liparidae based on uncontaminated exon capture data, slightly different methods could be used to minimize likelihood of contamination and maximize likelihood of successful sequencing. After DNA was extracted, it would be quantified and visualized and then run through the Li et al. (2013) double hybrid capture protocol with slightly modified steps. First, it would be good to use inline indices (Wang et al., 2022) to perform double indexing of samples, where one index is ligated before library prep and then a second index ligated before pooling, in order to clearly identify benchtop contamination in sequenced data via index mismatch. Additionally, it would be good to follow newer protocols in development that lessen the number of wash steps in order to limit the number of times samples are pipetted and may become contaminated (C. Li, Pers. Comm.). Once data were sequenced, the assembly pipeline used to begin data analysis would ideally also retain heterozygosity data during assembly in order to have another method of identifying potentially contaminated loci as well as retain heterozygosity data for analysis.

While exon capture is a powerful tool for systematic analysis at different taxonomic scales, this study reveals it can also be hindered by cross-contamination that becomes extremely difficult to identify and remove. Much work is being done in order to make the likelihood of cross-contamination much lower during the benchwork steps, as well as to improve methods for identifying contamination (Arcila et al., 2021, Wang et al., 2022). Here we explored the utility of calculating heterozygosity for each individual at each locus in order to detect contamination via aberrant allele frequencies, which revealed widespread low levels of cross-contamination. We also suggest that assembly methods used to analyze the data be updated to retain information

about heterozygosity in order to allow for a method of detecting contamination that does not require any *a priori* hypothesis of relationships. Taking these steps could drastically improve the utility of exon capture for systematic inference in groups known to be difficult to work with for various reasons.

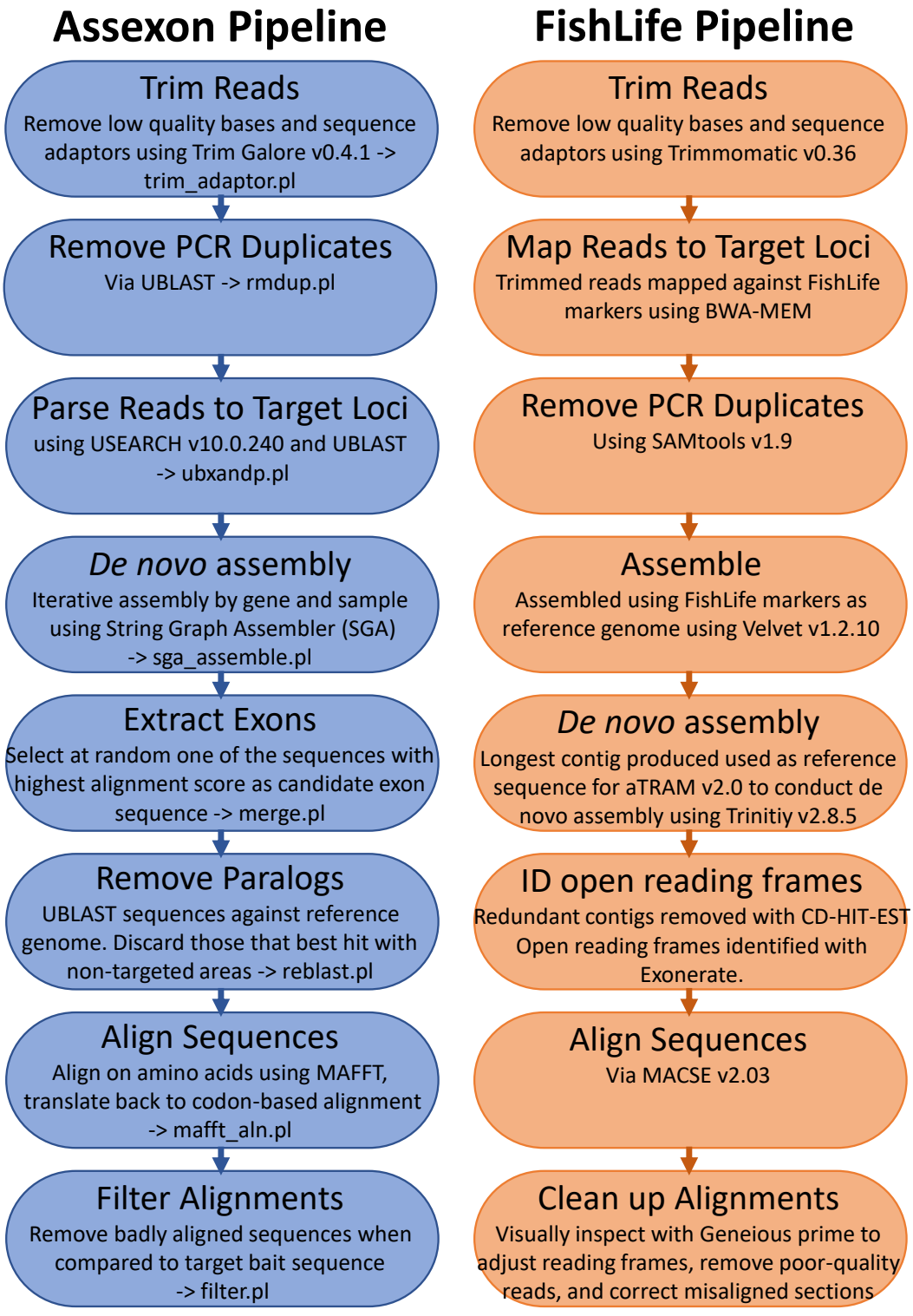

Fig 2.1. Flowchart depicting the differences in steps and programs used to assemble exon capture data using the Assexon (Yuan et al., 2019) pipeline vs. the FishLife pipeline (Arcila et al., 2021)

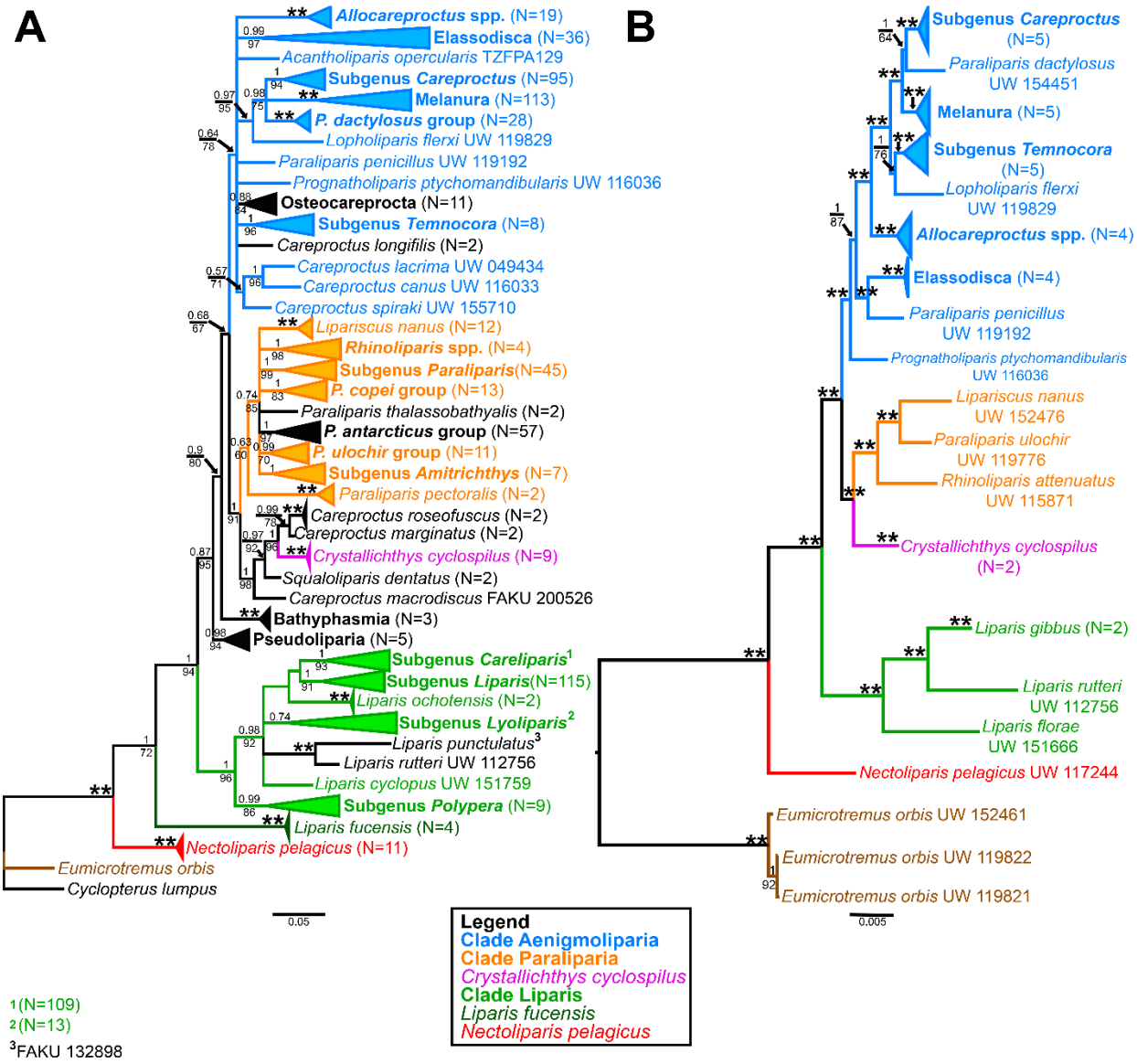

Fig 2.2. Majority rule consensus trees of Liparidae from Orr et al. (2019). A) Tree based on cytochrome *c* oxidase subunit one gene (COI) sequences. B) Tree based on RADseq sequences. Node values represent Bayesian posterior probabilities and bootstrap values from maximum likelihood analysis above and below the branches, respectively. Double asterisks indicate Bayesian posterior probabilities of 1 and maximum bootstrap values of 100.Triangles represent collapsed clades used for defining closely related groups in this study, with clade names following Orr (et al. 2019). Some well-resolved clades and individual species with important placements are highlighted using color to facilitate comparison to trees from this study. Clades and branches in black were not sampled in this analysis.

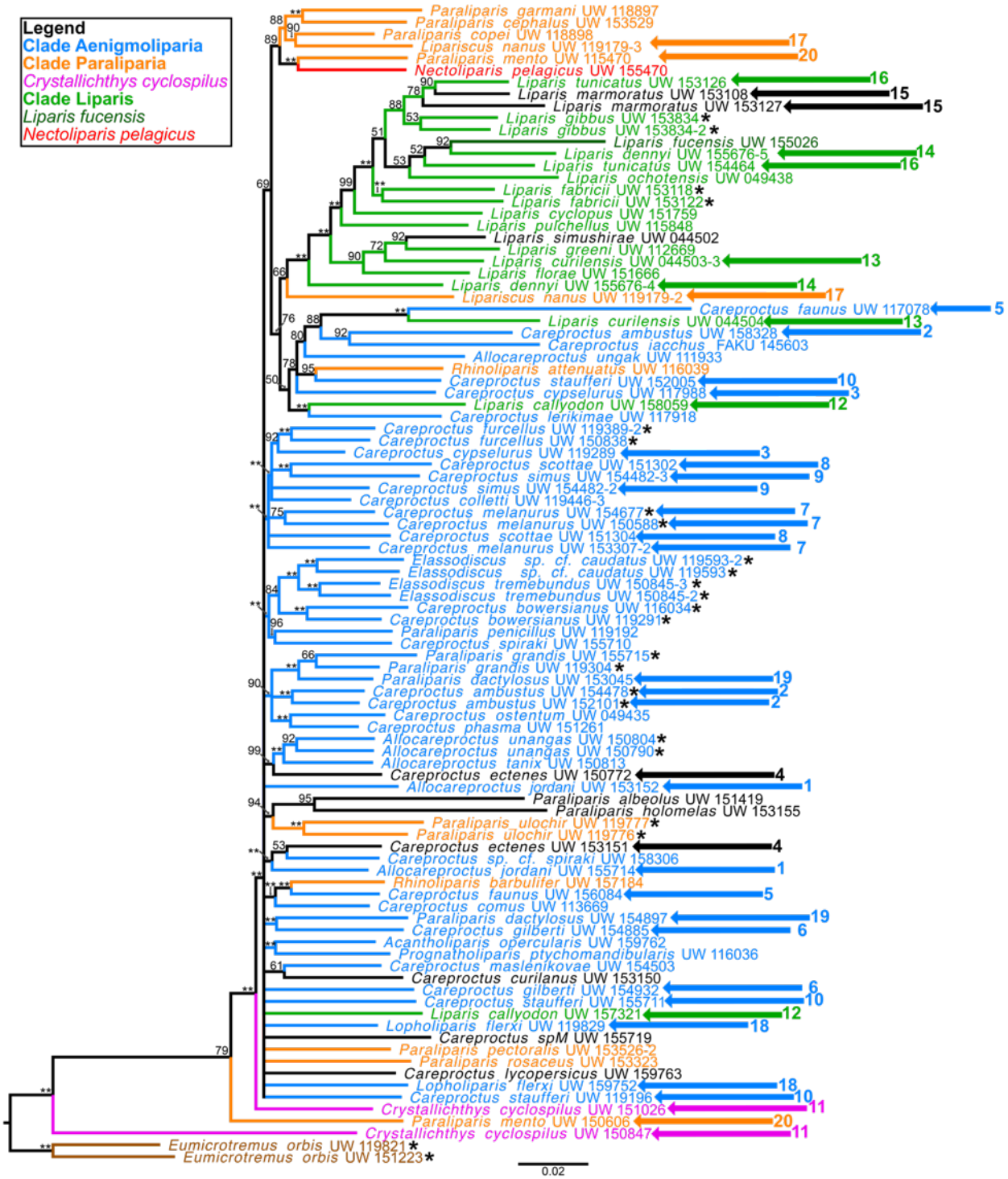

Fig 2.3. Majority rule consensus tree of Liparidae based on sequence data from the 2917 exons found in at least 50% of taxa, assembled using Assexon pipeline (Yuan et al., 2019). Node values are bootstrap values; double asterisks indicate bootstrap value of 100%. Colors follow major clades defined by Orr et al. (2019, fig 2). Species in black were not addressed by Orr et al. (2019). Large asterisks at tips indicate members of the same species recovered together. Numbered arrows indicate placement of members of the same species that were not recovered together.

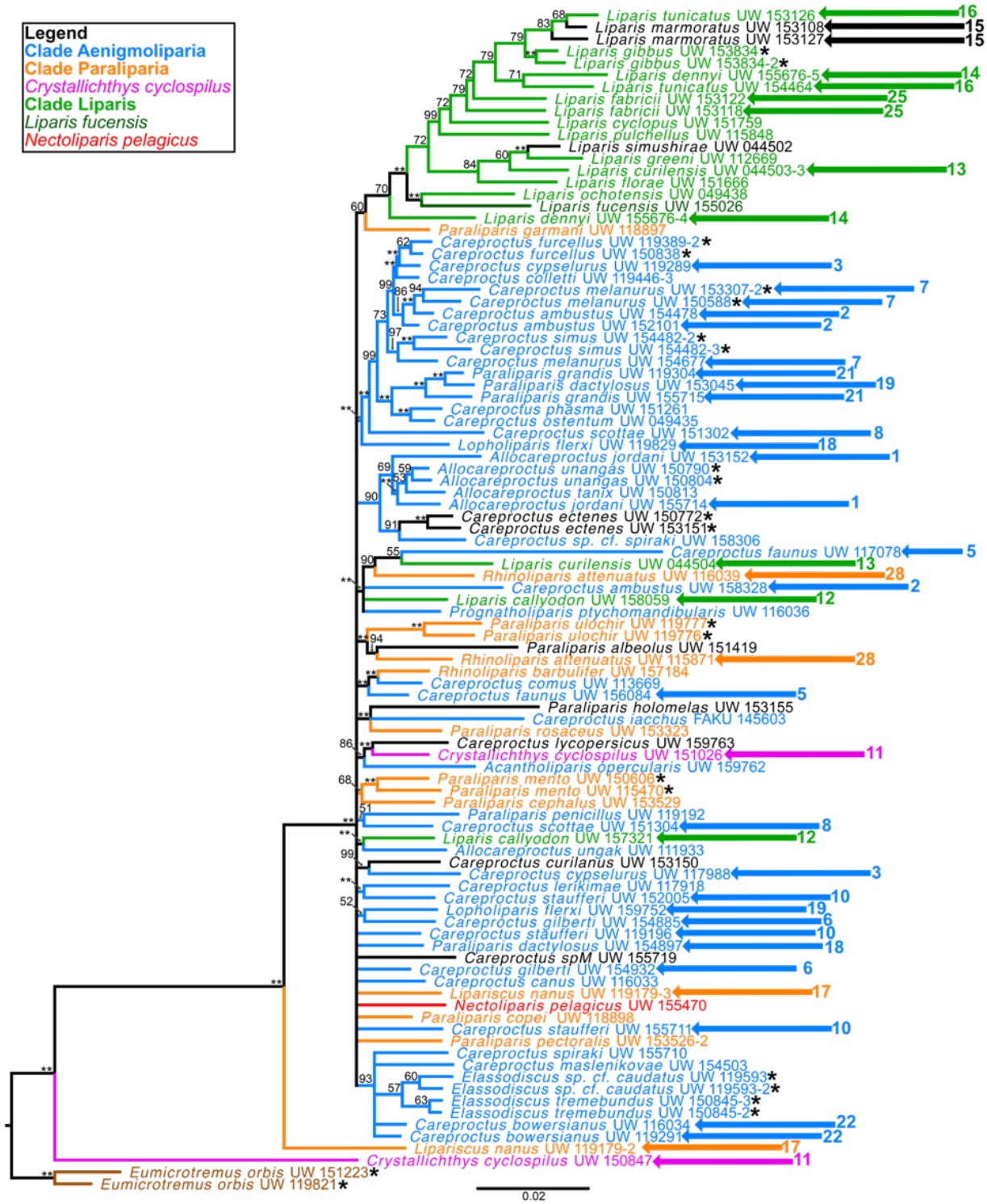

Fig 2.4. Majority rule consensus tree of Liparidae based on sequence data from the 956 exons assembled using FishLife exon pipeline (Arcila et al., 2021). Node values are bootstrap values; double asterisks indicate bootstrap value of 100%. Colors follow major clades defined by Orr et al. 2019 (Fig 2). Species in black were not addressed by Orr et al. (2019). Large asterisks at tips indicate members of the same species recovered together. Numbered arrows indicate placement of members of the same species that were not recovered together.

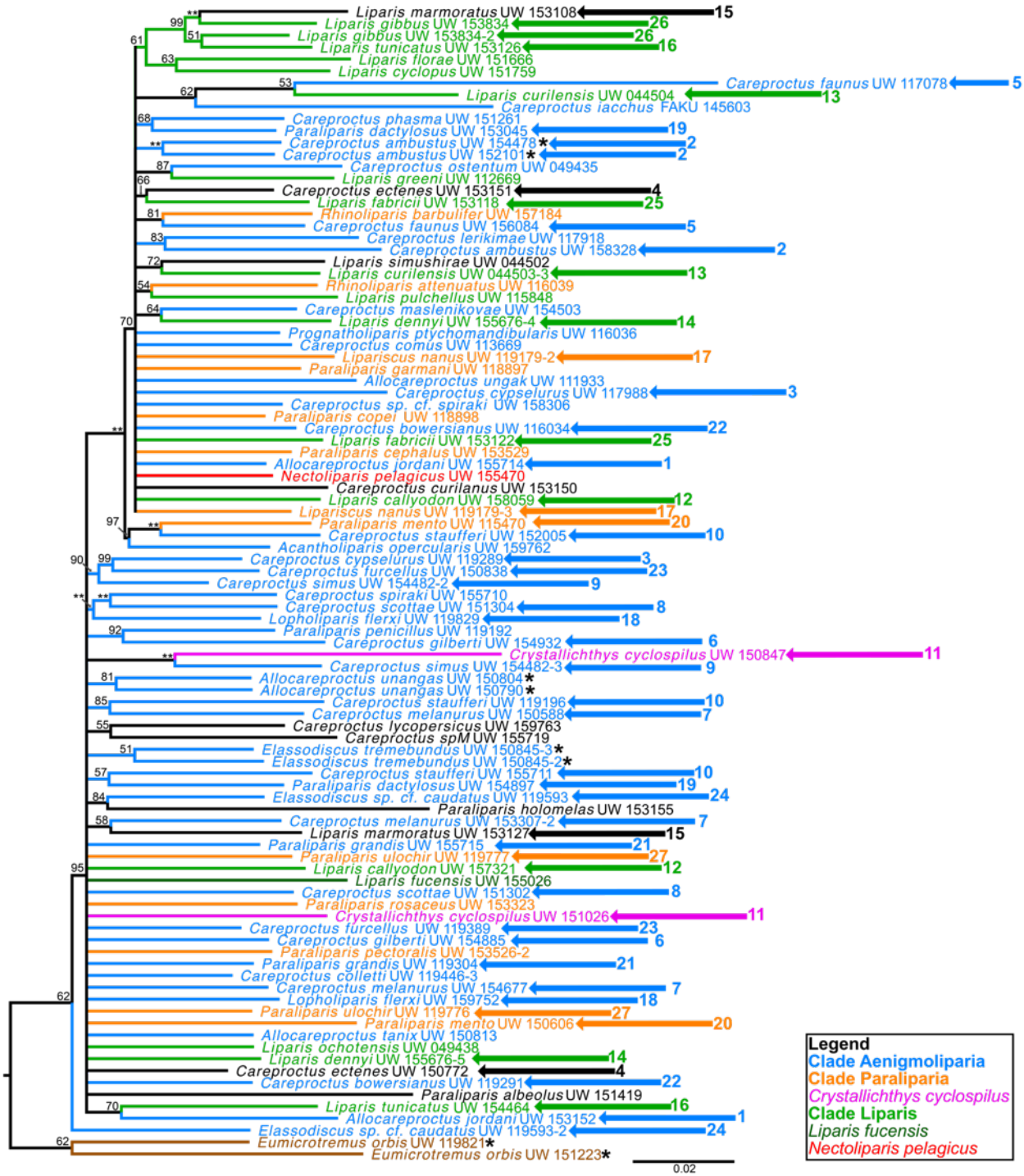

Fig 2.5. Majority rule consensus tree of Liparidae based on sequence data from the 251 exons found in all 97 individuals, assembled using Assexon pipeline (Yuan et al., 2019). Node values are bootstrap values; double asterisks indicate bootstrap value of 100%. Colors follow major clades defined by Orr et al. 2019 (Fig 2). Species in black were not addressed by Orr et al. (2019). Large asterisks at tips indicate members of the same species recovered together. Numbered arrows indicate placement of members of the same species that were not recovered together.

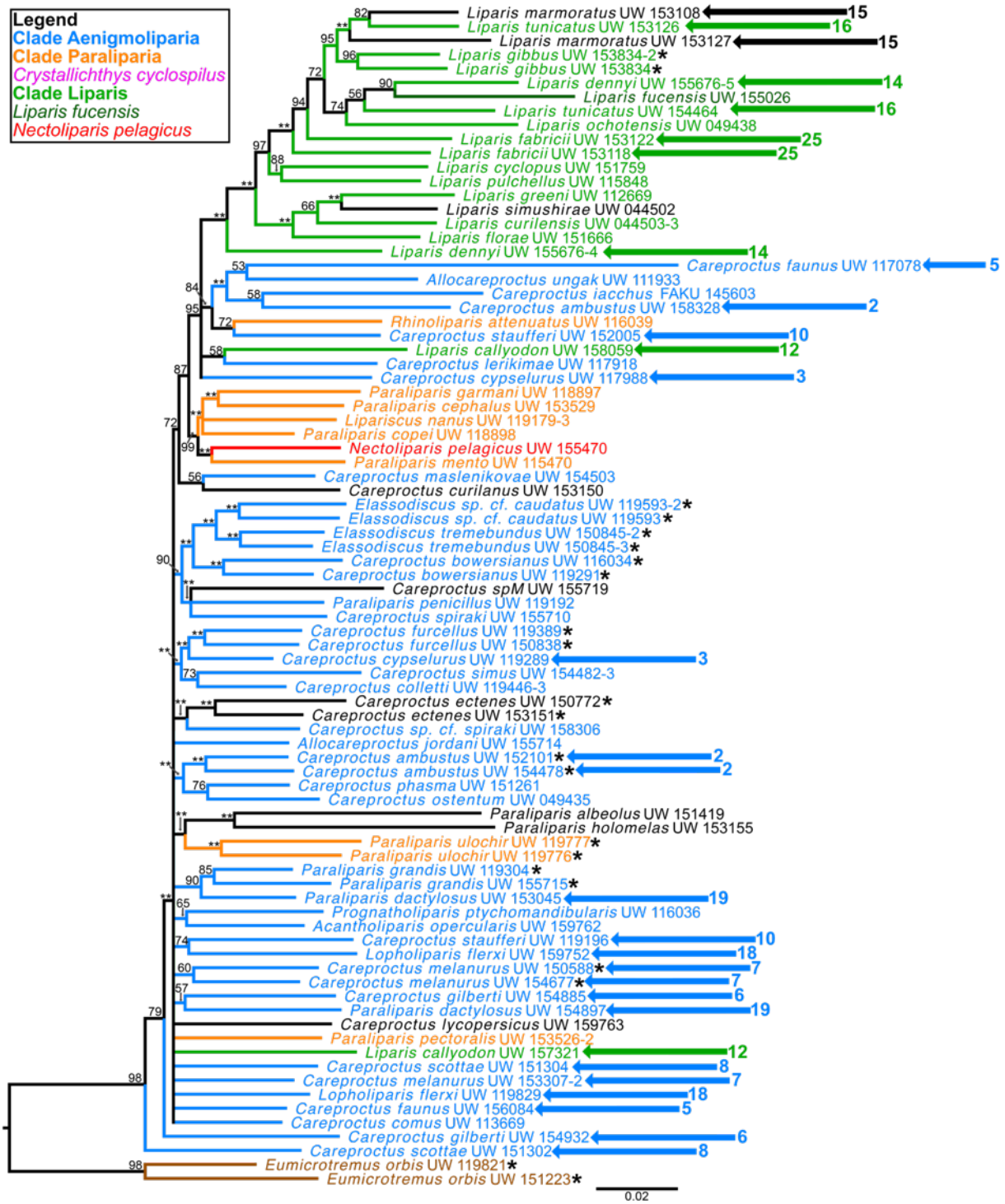

Fig 2.6. Majority rule consensus tree of Liparidae based on sequence data from the 2917 exons found in at least 50% of individuals, assembled using Assexon pipeline (Yuan et al., 2019). Tree inferred for 84 individuals with 13 removed based on suspected contamination. Node values are bootstrap values; double asterisks indicate bootstrap value of 100%. Colors follow major clades defined by Orr et al. 2019 (Fig 2). Species in black were not addressed by Orr et al. (2019). Large asterisks at tips indicate members of the same species recovered together. Numbered arrows indicate placement of members of the same species that were not recovered together.

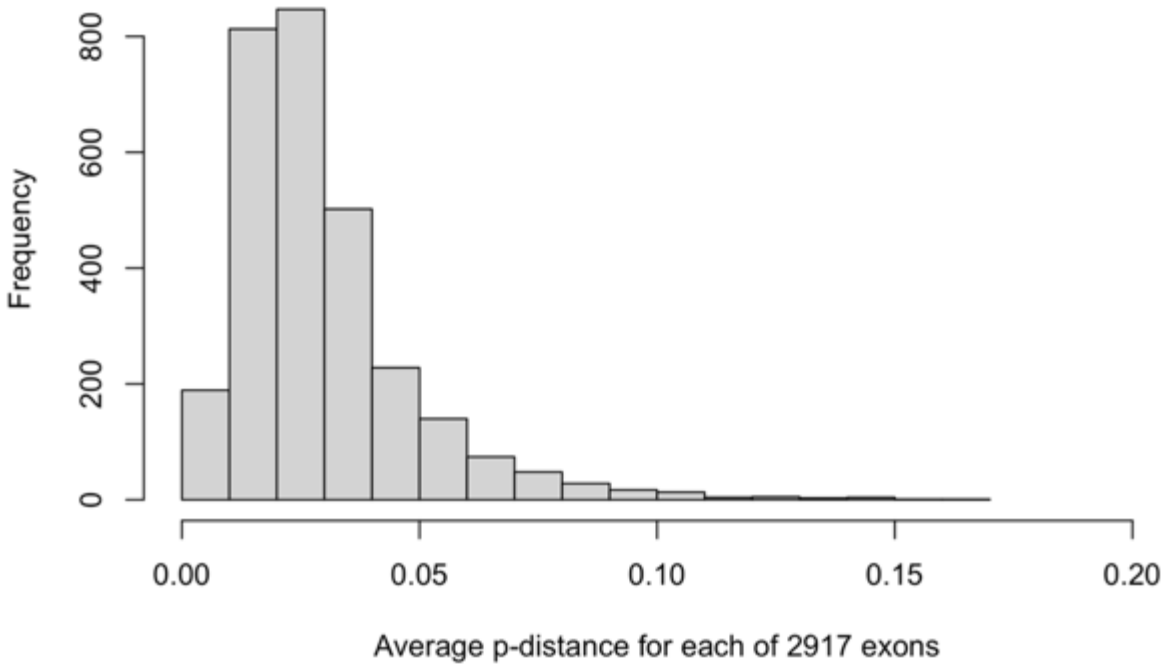

Fig 2.7. Average *p*-distance for each of 2917 exons used in phylogenetic analysis.

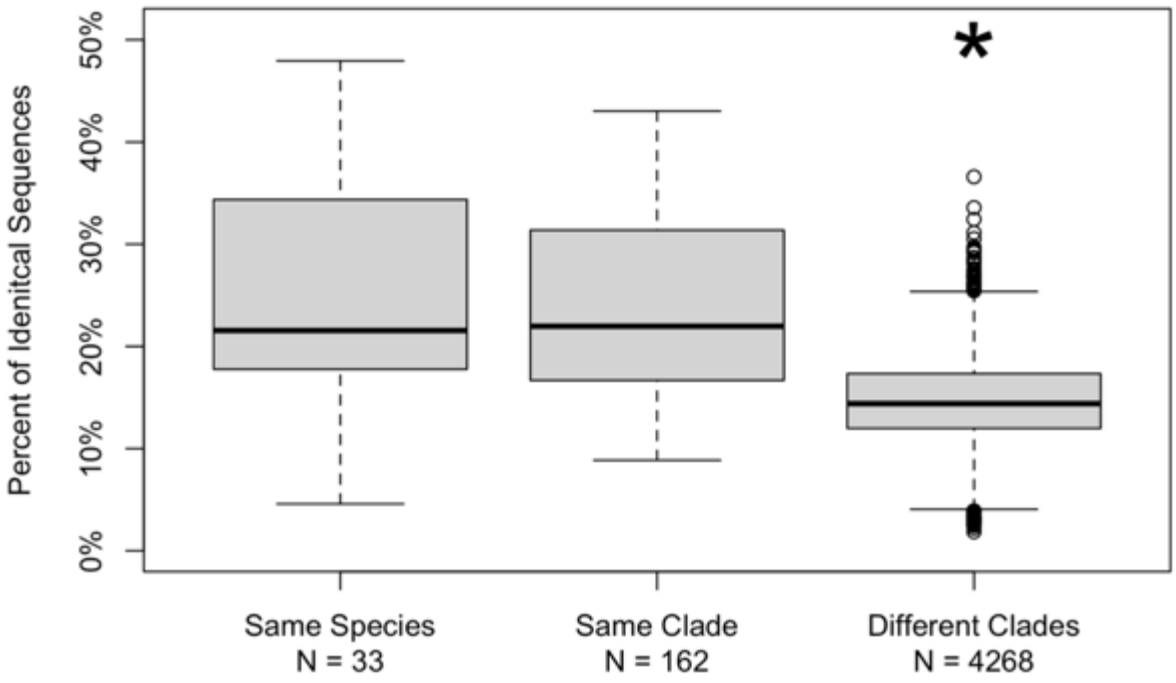

Fig 2.8. Percentage of identical sequences found between same species pairs (N = 33 pairwise comparisons), pairs of the same close relationship clade based on Orr et al. (2019) (N = 162 pairwise comparisons), and pairs not in the same close relationship clade (N = 4268 pairwise comparisons). Asterisk indicates significantly difference from other groups ($p < 0.05$) based on paired t-tests.

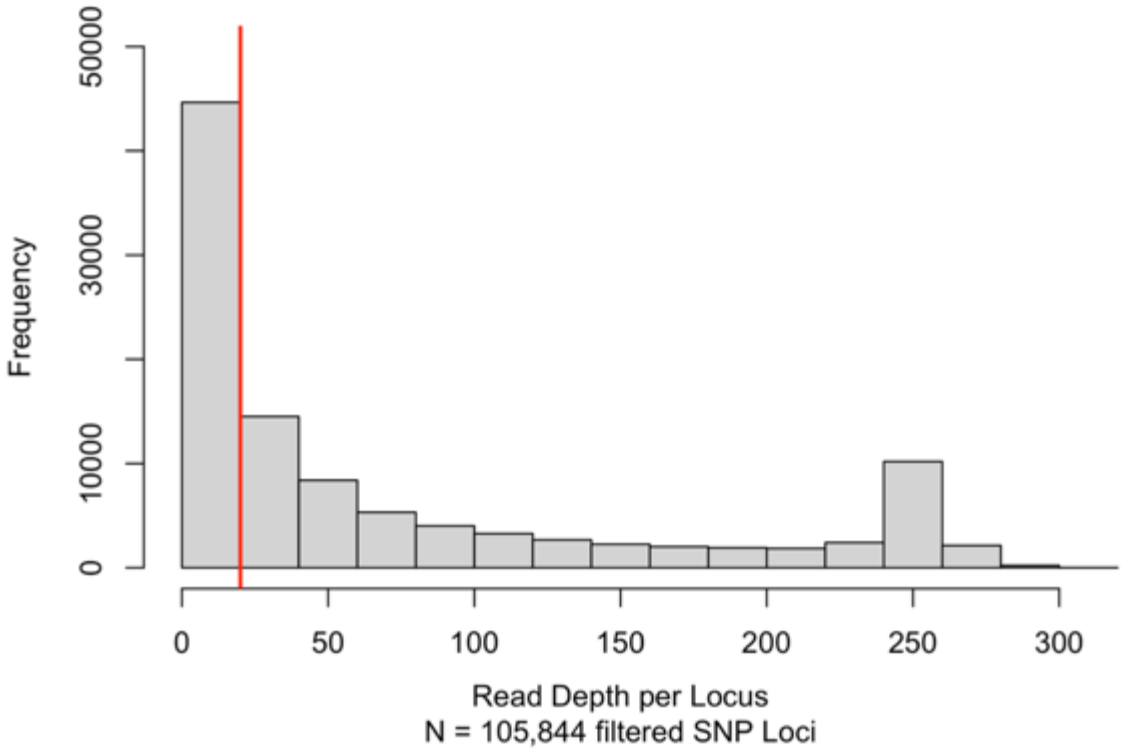

Fig 2.9. Total read depth for all 105,844 polymorphic loci found across 96 individuals. Red line indicates read depth of 20; all loci above this line were retained in further analyses.

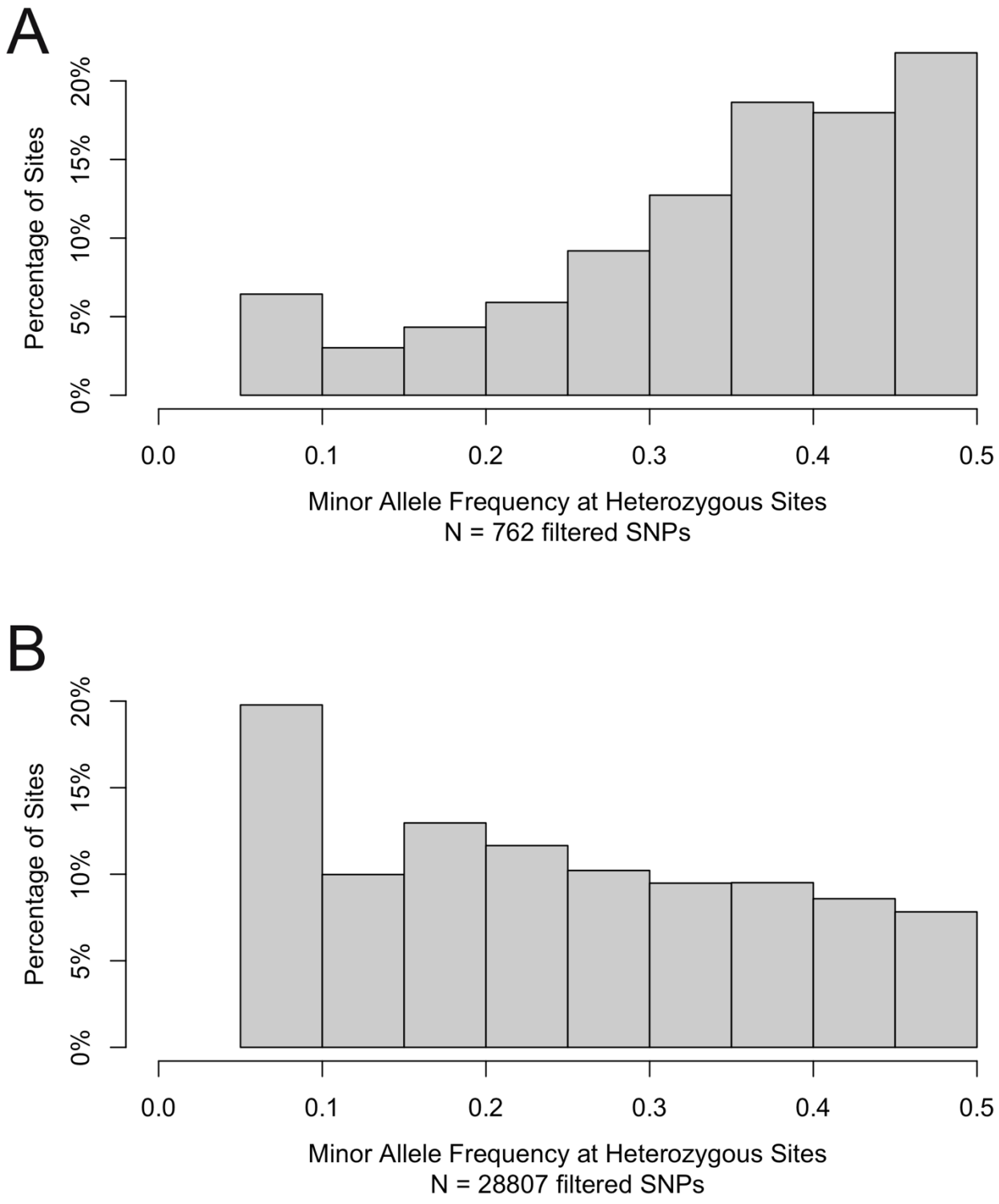

Fig 2.10. Minor allele frequency at heterozygous sites for sites with read depth greater than 20 and rate of occurrence of minor allele greater than 0.05 to remove sequencing error for A) 762 polymorphic loci in *Pleuronectes quadrituberculatus* (UW 151438) with uncontaminated exon capture data (Atta et al., 2022) and B) 28,807 polymorphic loci for 97 individuals in this study.

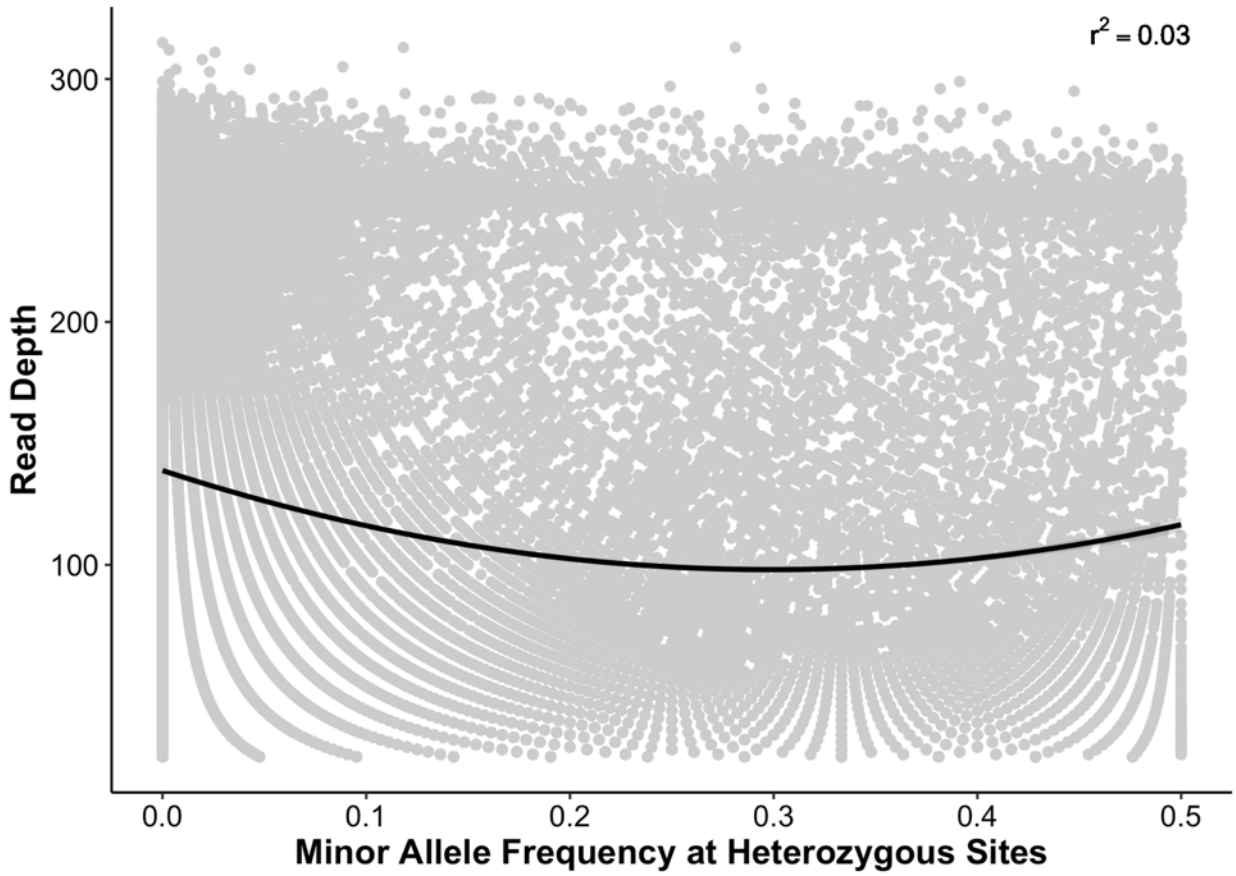

Fig 2.11. Minor allele frequency at heterozygous sites compared to read depth for 61,204 polymorphic loci with read depth greater than 20 for 96 individuals. Black line is quadratic line of best fit with an adjusted $r^2 = 0.03$.

Table 2.1. Species, catalog numbers, large clade placement (Aenigmoliparia, *Liparis*, Paraliparia) for comparisons with Orr et al. (2019), and if it was removed from downstream analysis after identifying contamination for all samples used in this study

Species	Catalog Number	Placement in Orr et al., 2019	Removed for contamination
Acantholiparis opercularis	UW 159762	Aenigmoliparia	
Allocareproctus jordani	UW 153152	Aenigmoliparia	Removed
Allocareproctus jordani	UW 155714	Aenigmoliparia	
Allocareproctus tanix	UW 150813	Aenigmoliparia	Removed
Allocareproctus unangas	UW 150790	Aenigmoliparia	Removed
Allocareproctus unangas	UW 150804	Aenigmoliparia	Removed
Allocareproctus ungak	UW 111933	Aenigmoliparia	
Careproctus ambustus	UW 152101	Aenigmoliparia	
Careproctus ambustus	UW 154478	Aenigmoliparia	
Careproctus ambustus	UW 158238	Aenigmoliparia	
Careproctus bowersianus	UW 116034	Aenigmoliparia	
Careproctus bowersianus	UW 119291	Aenigmoliparia	
*Careproctus canus**	UW 116033	Aenigmoliparia	N/A
Careproctus colletti	UW 119446 3 of 6	Aenigmoliparia	
Careproctus comus	UW 113669	Aenigmoliparia	
Careproctus curilanus	UW 153150	Absent	
Careproctus cypselurus	UW 117988	Aenigmoliparia	
Careproctus cypselurus	UW 119289	Aenigmoliparia	
Careproctus ectenes	UW 150772	Absent	
Careproctus ectenes	UW 153151	Absent	
Careproctus faunus	UW 117078	Aenigmoliparia	
Careproctus faunus	UW 156084 1 of 2	Aenigmoliparia	
Careproctus furcellus	UW 119389	Aenigmoliparia	
Careproctus furcellus	UW 150838	Aenigmoliparia	
Careproctus gilberti	UW 154885	Aenigmoliparia	
Careproctus gilberti	UW 154932	Aenigmoliparia	
Careproctus iacchus	FAKU 145603	Aenigmoliparia	
Careproctus lerikimae	UW 117918	Aenigmoliparia	
Careproctus lycoperiscus	UW 159763	Absent	
Careproctus maslenikovae	UW 154503	Aenigmoliparia	
Careproctus melanurus	UW 150588	Aenigmoliparia	
Careproctus melanurus	UW 153307 2 of 2	Aenigmoliparia	
Careproctus melanurus	UW 154677	Aenigmoliparia	
Careproctus ostentum	UW 49435	Aenigmoliparia	
Careproctus phasma	UW 151261	Aenigmoliparia	
Careproctus scottae	UW 151302	Aenigmoliparia	

Table 2.1 (Continued)

Species	Catalog Number	Placement in Orr et al., 2019	Removed for contamination
Careproctus scottae	UW 151304	Aenigmoliparia	
Careproctus simus	UW 154482 2 of 3	Aenigmoliparia	Removed
Careproctus simus	UW 154482 3 of 3	Aenigmoliparia	
Careproctus sp. cf. *spiraki*	UW 158306	Aenigmoliparia	
Careproctus sp. M	UW 155719	Absent	
Careproctus spiraki	UW 155710	Aenigmoliparia	
Careproctus staufferi	UW 119196	Aenigmoliparia	
Careproctus staufferi	UW 152005	Aenigmoliparia	
Careproctus staufferi	UW 155711	Aenigmoliparia	Removed
Crystallichthys cyclospilus	UW 150847	Crystallichthyia	Removed
Crystallichthys cyclospilus	UW 151026	Crystallichthyia	Removed
Elassodiscus sp. cf. *caudatus*	UW 119593 1 of 2	Aenigmoliparia	
Elassodiscus sp. cf. *caudatus*	UW 119593 2 of 2	Aenigmoliparia	
Elassodiscus tremebundus	UW 150845 2 of 3	Aenigmoliparia	
Elassodiscus tremebundus	UW 150845 3 of 3	Aenigmoliparia	
Eumicrotremus orbis	UW 119821 1 of 2	Outgroup	
Eumicrotremus orbis	UW 151223	Outgroup	
Liparis callyodon	UW 157321	*Liparis* clade	
Liparis callyodon	UW 158059	*Liparis* clade	
Liparis curilensis	UW 44503 3 of 5	*Liparis* clade	
Liparis curilensis	UW 44504	*Liparis* clade	Removed
Liparis cyclopus	UW 151759	*Liparis* clade	
Liparis dennyi	UW 155676 4 of 6	*Liparis* clade	
Liparis dennyi	UW 155676 5 of 6	*Liparis* clade	
Liparis fabricii	UW 153118	*Liparis* clade	
Liparis fabricii	UW 153122	*Liparis* clade	
Liparis florae	UW 151666	*Liparis* clade	
Liparis fucensis	UW 155026	Base of Tree	
Liparis gibbus	UW 153834 1 of 11	*Liparis* clade	
Liparis gibbus	UW 153834 2 of 11	*Liparis* clade	
Liparis greeni	UW 112669	*Liparis* clade	
Liparis marmoratus	UW 153108	Absent	
Liparis marmoratus	UW 153127	Absent	
Liparis ochotensis	UW 49438	*Liparis* clade	
Liparis pulchellus	UW 115848	*Liparis* clade	
Liparis simushirae	UW 44502	Absent	
Liparis tunicatus	UW 153126	*Liparis* clade	

Table 2.1 (Continued)

Species	Catalog Number	Placement in Orr et al., 2019	Removed for contamination
Liparis tunicatus	UW 154464 1 of 2	*Liparis* clade	
Lipariscus nanus	UW 119179 2 of 3	Paraliparia	Removed
Lipariscus nanus	UW 119179 3 of 3	Paraliparia	
Lopholiparis flerxi	UW 119829	Aenigmoliparia	
Lopholiparis flerxi	UW 159752	Aenigmoliparia	
Nectoliparis pelagicus	UW 155470	Base of Tree	
Paraliparis albeolus	UW 151419	Absent	
Paraliparis cephalus	UW 153529	Paraliparia	
Paraliparis copei	UW 118898	Paraliparia	
Paraliparis dactylosus	UW 153045	Aenigmoliparia	
Paraliparis dactylosus	UW 154897	Aenigmoliparia	
Paraliparis garmani	UW 118897	Paraliparia	
Paraliparis grandis	UW 119304	Aenigmoliparia	
Paraliparis grandis	UW 155715	Aenigmoliparia	
Paraliparis holomelas	UW 153155	Absent	
Paraliparis mento	UW 115470	Paraliparia	
Paraliparis mento	UW 150606	Paraliparia	Removed
Paraliparis pectoralis	UW 153526 2 of 2	Paraliparia	
Paraliparis penicillus	UW 119192	Aenigmoliparia	
Paraliparis rosaceus	UW 153323	Paraliparia	Removed
Paraliparis ulochir	UW 119776	Paraliparia	
Paraliparis ulochir	UW 119777	Paraliparia	
Prognatholiparis ptychomandibularis	UW 116036	Aenigmoliparia	
Rhinoliparis attenuatus	UW 116039	Paraliparia	
*Rhinoliparis attenuatus**	UW 115871	Paraliparia	N/A
Rhinoliparis barbulifer	UW 157184	Paraliparia	Removed

*Only present in dataset assembled using FishLife pipeline; failed to assemble with Assexon

Table 2.2. Results of BLAST+ comparison of COI sequences mined from 30 specimens from the exon-capture data with reference sequences of a 492 bp length COI with all sequences correctly identified to species by Orr et al. (2019).

Catalog Number	Specimen	Longest Sequence	Top BLAST Hit	Contaminated?
UW 150790	*Allocareproctus unangas*	66 bp	No Hits	NA – Too short to align
UW 111933	*Allocareproctus ungak*	492 bp	*Allocareproctus ungak*	No
UW 152101	*Careproctus ambustus*	573 bp	*Careproctus ambustus*	No
UW 154478	*Careproctus ambustus*	279 bp	*Careproctus ambustus*	No
UW 153150	*Careproctus curilanus*	528 bp	NA	No match in reference sequences
UW 150772	*Careproctus ectenes*	276 bp	NA	No match in reference sequences
UW 153151	*Careproctus ectenes*	708 bp	NA	No match in reference sequences
UW 117078	*Careproctus faunus*	66 bp	No Hits	NA – Too short to align
UW 156084	*Careproctus faunus*	465 bp	*Careproctus comus*	No– match is very similar sister species
UW 117918	*Careproctus lerikimae*	351 bp	*Careproctus lerikimae*	No
UW 154503	*Careproctus maslenikovae*	66 bp	No Hits	NA – Too short to align
UW 150588	*Careproctus melanurus*	348 bp	*Careproctus melanurus*	No
UW 154482	*Careproctus simus*	234 bp	*Careproctus simus*	No
UW 154482	*Careproctus simus*	243 bp	*Careproctus simus*	No
UW 151759	*Liparis cyclopus*	675 bp	*Liparis cyclopus*	No
UW 153118	*Liparis fabricii*	639 bp	*Liparis fabricii*	No
UW 151666	*Liparis florae*	480 bp	*Liparis florae*	No
UW 153834	*Liparis gibbus*	678 bp	*Liparis fabricii*	No – downstream analyses do not indicate contamination
UW 153834-2	*Liparis gibbus*	705 bp	*Liparis fabricii*	No – downstream analyses do not indicate contamination
UW 49438	*Liparis ochotensis*	375 bp	*Liparis bristolensis*	No – sequences only aligned for short fragment of 492 id reference length
UW 115848	*Liparis pulchellus*	504 bp	*Liparis pulchellus*	No
UW 153126	*Liparis tunicatus*	420 bp	*Liparis tunicatus*	No

Table 2.2 (Continued)

Catalog Number	Specimen	Longest Sequence	Top BLAST Hit	Contaminated?
UW 153155	*Paraliapris holomelas*	504 bp	NA	No match in reference sequences
UW 153529	*Paraliparis cephalus*	648 bp	*Paraliparis cephalus*	No
UW 153045	*Paraliparis dactylosus*	333 bp	*Paraliparis dactylosus*	No
UW 118897	*Paraliparis garmani*	153 bp	*Paraliparis garmani*	No
UW 150606	*Paraliparis mento*	66 bp	No Hits	NA – Too short to align
UW 116036	*Prognatholiparis ptychomandibularis*	735 bp	*Prognatholiparis ptychomandibularis*	No
UW 115871	*Rhinoliparis attenuatus*	39 bp	No Hits	NA – Too short to align
UW 157184	*Rhinoliparis barbulifer*	549 bp	*Careproctus comus*	**Yes**

REFERENCES

Aberer, A. J., D. Krompass, and A. Stamatakis. 2013. Pruning rogue taxa improves phylogenetic accuracy: an efficient algorithm and webservice. Systematic Biology 62:162–166.

Allen, J. M., R. LaFrance, R. A. Folk, K. P. Johnson, and R. P. Guralnick. 2018. aTRAM 2.0: An Improved, Flexible Locus Assembler for NGS Data. Evolutionary Bioinformatics 14:1–4.

Andriashev, A. P. 1939. Essay on the zoogeography and origin of the fish fauna of the Bering Sea and adjacent waters. Izdanie Leningradskogo Gosudarstvenogo Universiteta, Leningrad. 1–187. [In Russian, English summary [pp. 181–185]. English translation by A. Merrivale [i-lv + 1–284].]

Andriashev, A. P., and D. L. Stein. 1998. Review of the snailfish genus *Careproctus* (Liparidae, Scorpaeniformes) in Antarctic and adjacent waters. Contributions in Science (Los Angeles) 470:1–63.

Arcila, D., L. C. Hughes, B. Meléndez-Vazquez, C. C. Baldwin, W. T. White, K. E. Carpenter, J. T. Williams, M. D. Santos, J. J. Pogonoski, M. Miya, G. Ortí, and R. Betancur-R. 2021. Testing the Utility of Alternative Metrics of Branch Support to Address the Ancient Evolutionary Radiation of Tunas, Stromateoids, and Allies (Teleostei: Pelagiaria). Systematic Biology 70:1123–1144.

Atta, C. J., H. Yuan, C. Li, D. Arcila, R. Betancur-R, L. C. Hughes, G. Ortí, and L. Tornabene. 2022. Exon-capture data and locus screening provide new insights into the phylogeny of flatfishes (Pleuronectoidei). Molecular Phylogenetics and Evolution 166:107315.

Baird, N. A., P. D. Etter, T. S. Atwood, M. C. Currey, A. L. Shiver, Z. A. Lewis, E. U. Selker, W. A. Cresko, and E. A. Johnson. 2008. Rapid SNP discovery and genetic mapping using sequenced RAD markers. PloS ONE 3:e3376.

Balushkin, A. V. 1996. New genus and species of snailfish Palmoliparis beckeri (Scorpaeniformes, Liparidae) of the northern Kurile Islands with remarks on phylogeny of family. Voprosy Ikhtiologii 36:293–299. [in Russian, English translation in Journal of Ichthyology 36:281–287]

Barnard, K. H. 1927. Diagnoses of new genera and species of South African marine fishes. Annals and Magazine of Natural History 20:66–79.

Betancur-R, R., R. E. Broughton, E. O. Wiley, K. Carpenter, J. A. López, C. Li, N. I. Holcroft, D. Arcila, M. Sanciangco, J. C. Cureton Ii, F. Zhang, et al. 2013. The tree of life and a new classification of bony fishes. PLoS Currents 5. doi: 10.1371/currents.tol.53ba26640df0ccaee75bb165c8c26288

Betancur-R, R., E. O. Wiley, G. Arratia, A. Acero, N. Bailly, M. Miya, G. Lecointre, and G. Ortí. 2017. Phylogenetic classification of bony fishes. BMC Evolutionary Biology 17:162.

Böhlke, J. E. 1953. A catalogue of the type specimens of Recent fishes in the Natural History Museum of Stanford University. Stanford Ichthyological Bulletin 5:1-168.

Borets, L. A. 2000. An annotated list of fishes of the Far Eastern seas. TINRO-Center. Vladivostok. 1-192. [In Russian.]

Burke, C. V. 1930. Revision of the fishes of the family Liparidae. Bulletin of the U.S. National Museum 150:1–204.

Busby, M. S., J. W. Orr, and D. M. Blood. 2006. Eggs and late-stage embryos of *Allocareproctus unangas* (family Liparidae) from the Aleutian Islands. Ichthyological Research 53:423–426.

Camacho, C., G. Coulouris, V. Avagyan, N. Ma, J. Papadopoulos, K. Bealer, and T. L. Madden. 2009. BLAST+: architecture and applications. BMC Bioinformatics 10:421.

Chernova, N. 2001. A review of the genus *Psednos* (Pisces, Liparidae) with description of ten new species from the North Atlantic and Southwestern Indian Ocean. Bulletin of the Museum of Comparative Zoology at Harvard College 155:477–507.

Chernova, N. V. 2006. New and rare snailfishes (Liparidae, Scorpaeniformes) with the description of four new species from the Southern Hemisphere and tropical east Pacific. Journal of Ichthyology 46(supplement 1):S1–S14.

Chernova, N. V. 2014. New species of the genus *Careproctus* (Liparidae) from the Kara Sea with notes on spongiophilia, reproductive commensalism between fishes and sponges (Rossellidae). Voprosy Ikhtiologii, 54:508–519. [English translation in Journal of Ichthyology 54:501–512]

Chernova, N. V., D. L. Stein, and A. P. Andriashev. 2004. Family Liparidae Scopoli 1777 -- snailfishes. California Academy of Sciences Annotated Checklists of Fishes No. 31: 1–72.

Collett, R. 1879. Fiske fra Nordhavs-Expeditionens sidste Togt, Sommeren 1878. Forhandlinger i Videnskabs-selskabet i Christiania (for 1878) 14:1–106.

Danecek, P., J. K. Bonfield, J. Liddle, J. Marshall, V. Ohan, M. O. Pollard, A. Whitwham, T. Keane, S. A. McCarthy, R. M. Davies, and H. Li. 2021. Twelve years of SAMtools and BCFtools. GigaScience 10. doi: 10.1093/gigascience/giab008

Datsky, A. V. 2015. Ichthyofauna of the Russian exclusive economic zone of the Bering Sea: 1. Taxonomic diversity. Journal of Ichthyology 55(6):792-826.

Duhamel, G., M. Hautecoeur, A. Dettai, R. Causse, and P. Pruvost. 2010. Liparids from the eastern sector of Southern Ocean and first information from molecular studies. Cybium 34:319–343.

Eschmeyer, W. N. 1998. Catalog of fishes. California Academy of Sciences, San Francisco.

Faircloth, B. C., J. E. McCormack, N. G. Crawford, M. G. Harvey, R. T. Brumfield, and T. C. Glenn. 2012. Ultraconserved elements anchor thousands of genetic markers spanning multiple evolutionary timescales. Systematic Biology 61:717–726.

Fedorov, V. V. 1973a. Ichthyofauna of the continental slope of the Bering Sea and some aspects of its origin and formation. Izvestiia Tikhookeanskogo nauchnogo instituta rybnogo khoziaistva [Bulletins of the Pacific Science Institute] v. 87: 3-41. [In Russian. English translation by U.S. National Marine Fisheries Service.]

Fedorov, V. V. 1973b. A list of Bering Sea fishes. Izvestiia Tikhookeanskogo nauchnogo instituta rybnogo khoziaistva [Bulletins of the Pacific Science Institute] 87:42-71. [In Russian. English translation by U.S. National Marine Fisheries Service.]

Felsenstein, J. 2004. Inferring phylogenies. Sinauer associates Sunderland, MA.

Fricke, R., Eschmeyer, W. N., and R. van der Laan (Eds.). 2021. Eschmeyer's Catalog of Fishes: Genera, Species, References. (http://researcharchive.calacademy.org/research/ichthyology/catalog/fishcatmain.asp). Electronic version accessed 15 January 2022

Gardner, J. R., J. W. Orr, D. E. Stevenson, I. Spies, and D. A. Somerton. 2016. Reproductive Parasitism between Distant Phyla: Molecular identification of snailfish (Liparidae) egg masses in the gill cavities of king crabs (Lithodidae). Copeia 104:645–657.

Gerringer, M. E. 2019. On the Success of the Hadal Snailfishes. Integrative Organismal Biology:1–18.

Gerringer, M. E., T. D. Linley, A. J. Jamieson, E. Goetze, and J. C. Drazen. 2017. *Pseudoliparis swirei* sp. nov.: A newly-discovered hadal snailfish (Scorpaeniformes: Liparidae) from the Mariana Trench. Zootaxa 4358:161–177.

Gilbert, C. H., and C. V. Burke. 1912. Fishes from Bering Sea and Kamchatka. Bulletin of the Bureau of Fisheries 30:31–96.

Gilbert, P. S., J. Chang, C. Pan, E. M. Sobel, J. S. Sinsheimer, B. C. Faircloth, and M. E. Alfaro. 2015. Genome-wide ultraconserved elements exhibit higher phylogenetic informativeness than traditional gene markers in percomorph fishes. Molecular Phylogenetics and Evolution 92:140–146.

Glenn, T. C. 2011. Field guide to next-generation DNA sequencers. Molecular Ecology Resources 11:759–769.

Gruber, D. F., and J. S. Sparks. 2021. First Report of Biofluorescence in Arctic Snailfishes and Rare Occurrence of Multiple Fluorescent Colors in a Single Species. American Museum Novitates 3967:1–12.

Hughes, L. C., G. Ortí, Y. Huang, Y. Sun, C. C. Baldwin, A. W. Thompson, D. Arcila, R. Betancur-R, C. Li, L. Becker, N. Bellora, X. Zhao, X. Li…Q. Shi. 2018. Comprehensive phylogeny of ray-finned fishes (Actinopterygii) based on transcriptomic and genomic data. Proceedings of the National Academy of Sciences 115:6249–6254.

Hughes, L. C., G. Ortí, H. Saad, C. Li, W. T. White, C. C. Baldwin, K. A. Crandall, D. Arcila, and R. Betancur-R. 2021. Exon probe sets and bioinformatics pipelines for all levels of fish phylogenomics. Molecular Ecology Resources 21:816–833.

Jiang, J., H. Yuan, X. Zheng, Q. Wang, T. Kuang, J. Li, J. Liu, S. Song, W. Wang, F. Cheng, H. Li, et al. 2019. Gene markers for exon capture and phylogenomics in ray-finned fishes. Ecology and Evolution 9:3973–3983.

Jordan, D. S., B. W. Evermann, and H. W. Clark. 1930. Check list of the fishes and fishlike vertebrates of North and Middle America north of the northern boundary of Venezuela and Colombia. Report of the United States Commissioner of Fisheries (for 1928) (pt 2): 1–670.

Kai, Y., A. Tohkairin, K. Fujiwara, and T. Hamatsu. 2018. *Careproctus iacchus,* a new variegated snailfish (Liparidae) from the Seas of Japan and Okhotsk. Ichthyological Research 65:417–422.

Katoh, K., K. Misawa, K.-I. Kuma, and T. Miyata. 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Research 30:3059–3066.

Kido, K. 1988. Phylogeny of the family Liparididae, with the taxonomy of the species found around Japan. Memoirs of the Faculty of Fisheries Hokkaido University 35(2):125–256.

Knaus, B. J., and N. J. Grünwald. 2017. VCFR: a package to manipulate and visualize variant call format data in R. Molecular Ecology Resources 17:44–53.

Knudsen, S. W., and P. R. Møller. 2008. *Careproctus kidoi*, a new Arctic species of snailfish (Teleostei: Liparidae) from Baffin Bay. Ichthyology Research 55:175–182.

Knudsen, S. W., P. R. Møller, and P. Gravlund. 2007. Phylogeny of the snailfishes (Teleostei: Liparidae) based on molecular and morphological data. Molecular Phylogenetics and Evolution 44:649–666.

Krøyer, H. N. 1862. Nogle Bidrag til Nordisk ichthyologi [with subsections under separate titles]. Naturhistorisk Tidsskrift (Kjøbenhavn) (Ser. 3) 1:233–310.

Kuang, T., L. Tornabene, J. Li, J. Jiang, P. Chakrabarty, J. S. Sparks, G. J. P. Naylor, and C. Li. 2018. Phylogenomic analysis on the exceptionally diverse fish clade Gobioidei (Actinopterygii: Gobiiformes) and data-filtering based on molecular clocklikeness. Molecular Phylogenetics and Evolution 128:192–202.

Langmead, B., and S. L. Salzberg. 2012. Fast gapped-read alignment with Bowtie 2. Nature Methods 9:357–359.

Langmead, B., C. Wilks, V. Antonescu, and R. Charles. 2019. Scaling read aligners to hundreds of threads on general-purpose processors. Bioinformatics 35:421–432.

Lemmon, A. R., S. A. Emme, and E. M. Lemmon. 2012. Anchored hybrid enrichment for massively high-throughput phylogenomics. Systematic Biology 61:727–744.

Levin, L. A., and G. W. Rouse. 2020. Giant protists (xenophyophores) function as fish nurseries. Ecology 101:e02933.

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

Li, C., G. Ortí, G. Zhang, and G. Lu. 2007. A practical approach to phylogenomics: the phylogeny of ray-finned fish (Actinopterygii) as a case study. BMC Evolutionary Biology 7:44.

Li, H., and R. Durbin. 2009. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics 25:1754–1760.

Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, and 1000 Genome Project Data Processing Subgroup. 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25:2078–2079.

Longo, M. S., M. J. O'Neill, and R. J. O'Neill. 2011. Abundant human DNA contamination identified in non-primate genome databases. PloS ONE 6:e16410.

Love, M. S., C. W. Mecklenburg, T. A. Mecklenburg, and L. K. Thorsteinson. 2005. Resource Inventory of Marine and Estuarine Fishes of the West Coast and Alaska: A Checklist of North Pacific and Arctic Ocean Species from Baja California to the Alaska-Yukon Border. U. S. Department of the Interior, U. S. Geological Survey, Biological Resources Division, Seattle: i-ix + 1–276.

Matarese, A. C., A. W. Kendall, Jr., D. M. Blood, and B. M. Vinter. 1989. Laboratory guide to early life history stages of northeast Pacific fishes. NOAA Technical Report NMFS 80:1–652.

McCormack, J. E., S. M. Hird, A. J. Zellmer, B. C. Carstens, and R. T. Brumfield. 2013. Applications of next-generation sequencing to phylogeography and phylogenetics. Molecular Phylogenetics and Evolution 66:526–538.

Mecklenburg, C. W., T. A. Mecklenburg and L. K. Thorsteinson. 2002. Fishes of Alaska. American Fisheries Society, Bethesda, Maryland.

Near, T. J., A. Dornburg, R. I. Eytan, B. P. Keck, W. L. Smith, K. L. Kuhn, J. A. Moore, S. A. Price, F. T. Burbrink, M. Friedman, and P. C. Wainwright. 2013. Phylogeny and tempo of diversification in the superradiation of spiny-rayed fishes. Proceedings of the National Academy of Sciences 110:12738–12743.

Nelson, J. S., E. J. Crossman, H. Espinosa Pérez, L. T. Findley, C. R. Gilbert, R. N. Lea, and J. D. Williams. 2004. Common and Scientific Names of Fishes from the United States, Canada, and Mexico. Sixth Edition. American Fisheries Society, Special Publication 29. Bethesda, Maryland.

Nielsen, R., J. S. Paul, A. Albrechtsen, and Y. S. Song. 2011. Genotype and SNP calling from next-generation sequencing data. Nature Reviews Genetics 12:443–451.

Orr, J.W. 2004. *Lopholiparis flerxi,* a new genus and species of snailfish (Scorpaeniformes: Liparidae) from the Aleutian Islands, Alaska. Copeia 2004:551–555.

Orr, J. W. 2012. Two new species of *Careproctus* (Scorpaeniformes: Liparidae) from the eastern North Pacific. Copeia 2012:257–265.

Orr, J. W. 2016. Two new species of *Careproctus* (Liparidae) from the Aleutian Islands. Copeia 104:890–896.

Orr, J. W. 2021. Three new small snailfishes of the genus *Careproctus* (Teleostei: Cottiformes: Liparidae) from the Aleutian Islands, Alaska. Ichthyology & Herpetology 109:456–466.

Orr, J. W., and M. S. Busby. 2001. *Prognatholiparis ptychomandibularis*, a new genus and species of the fish family Liparidae (Teleostei: Scorpaeniformes) from the Aleutian Islands, Alaska. Proceedings of the Biological Society of Washington 114:51–57.

Orr, J. W., and M. S. Busby. 2006. Revision of the snailfish genus *Allocareproctus* Pitruk and Fedorov (Teleostei: Liparidae), with descriptions of four new species from the Aleutian Islands. Zootaxa 1173:1–37.

Orr, J. W., Y. Kai, and T. Nakabo. 2015. Snailfishes of the *Careproctus rastrinus* complex (Liparidae): Redescriptions of seven species in the North Pacific Ocean region, with the description of a new species from the Beaufort Sea. Zootaxa 4018:301–348.

Orr, J. W., and K. P. Maslenikov. 2007. Two new variegated snailfishes of the genus *Careproctus* (Teleostei: Scorpaeniformes: Liparidae) from the Aleutian Islands, Alaska. Copeia 2007:699–710.

Orr, J. W., D. L. Pitruk, R. Manning, D. E. Stevenson, J. R. Gardner, and I. Spies. 2020. A new species of snailfish (Cottiformes: Liparidae) closely related to *Careproctus melanurus* of the eastern North Pacific. Copeia 108:711–726.

Orr, J. W., I. B. Spies, D. E. Stevenson, G. C. Longo, Y. Kai, S. Ghods, and M. Hollowed. 2019. Molecular phylogenetics of snailfishes (Cottiformes: Liparidae) based on MtDNA and RADseq genomic analyses, with comments on selected morphological characters. Zootaxa 4642:1–79.

Page, L. M., H. Espinosa-Pérez, L. D. Findley, C. R. Gilbert, R. N. Lea, N. E. Mandrak, R. L. Mayden, and J. S. Nelson. 2013. Common and scientific Names of Fishes from the United States, Canada, and Mexico. Seventh Edition. American Fisheries Society, Special Publication 34. Bethesda, Maryland.

Parin, N. V., V. V. Fedorov, and B. A. Sheiko. 2002. An annotated catalogue of fish-like vertebrates and fishes of the seas of Russia and adjacent countries. Part 2. Order Scorpaeniformes. Journal of Ichthyology 42:S60–S135.

Parin, N. V., S. A. Evseenko, and E. D. Vasil'eva. 2014. Fishes of Russian Seas: Annotated Catalogue. KMK Scientific Press, Moscow 53:733 pp.

Potthoff, T. 1984. Clearing and staining techniques, p. 35–37. In: Ontogeny and Systematics of Fishes. H. G. Moser, W. J. Richards, D. M. Cohen, M. P. Fahay, A.W. Kendall, Jr., and S. L. Richardson (eds.). The American Society of Ichthyologists and Herpetologists, Special Publication No. 1, Lawrence, Kansas.

Quast, J. C., and E. L. Hall. 1972. List of Fishes of Alaska and Adjacent Waters with a Guide to Some of Their Literature. NOAA Technical Report NMFS SSRF-658.

R Core Team. 2021. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Rabosky, D. L., J. Chang, P. O. Title, P. F. Cowman, L. Sallan, M. Friedman, K. Kaschner, C. Garilao, T. J. Near, M. Coll, and M. E. Alfaro. 2018. An inverse latitudinal gradient in speciation rate for marine fishes. Nature 559:392–395.

Razkin, O., G. Sonet, K. Breugelmans, M. J. Madeira, B. J. Gómez-Moliner, and T. Backeljau. 2016. Species limits, interspecific hybridization and phylogeny in the cryptic land snail complex *Pyramidula*: The power of RADseq data. Molecular Phylogenetics and Evolution 101:267–278.

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

Robins, C. R., R. M. Bailey, C. E. Bond, J. R. Brooker, E. A. Lachner, R. N. Lea and W. B. Scott. 1991. Common and Scientific Names of Fishes from the United States and Canada. Fifth edition. American Fisheries Society, Special Publication 20:1–183.

Sabaj, M. H. 2020. Codes for natural history collections in ichthyology and herpetology. Copeia 108:593–669.

Scopoli, J. A. 1777. Introductio ad historiam naturalem, sistens genera lapidum, plantarum et animalium hactenus detecta, caracteribus essentialibus donata, in tribus divisa, subinde ad leges naturae. Prague. i–x + 1–506.

Sheiko, B. A., and V. V. Fedorov. 2000. Part 1. Class Cephalaspidomorphi, Class Chondrichthyes, Class Holocephali, Class Osteichthyes. Pp. 7-69. In: Moissev, R. S. (ed.) Catalog of the Vertebrates of Kamchatka and Adjacent Waters. Petropavlovsk-Kamchatsky, Kamchatskiy Petchatniy Dvor. [In Russian.]

Shen, Y., W. Dai, Z. Gao, G. Yan, X. Gan, and S. He. 2017. Molecular phylogeny and divergence time estimates using the mitochondrial genome for the hadal snailfish from the Mariana trench. Science Bulletin 62:1106–1108. supplement https://doi.org/10.1016/j.scib.2017.07.010

Simion, P., K. Belkhir, C. François, J. Veyssier, J. C. Rink, M. Manuel, H. Philippe, and M. J. Telford. 2018. A software tool "CroCo" detects pervasive cross-species contamination in next generation sequencing data. BMC Biology 16:28.

Simion, P., F. Delsuc, and H. Philippe. 2020. To what extent current limits of phylogenomics can be overcome? p. 2.1:1–2.1:34. *In*: Phylogenetics in the Genomic Era. C. Scornavacca, F. Delsuc, and N. Galtier (eds.). No commercial publisher.

Simpson, J. T., and R. Durbin. 2012. Efficient de novo assembly of large genomes using compressed data structures. Genome Research 22:549–556.

Soldatov, V. K., and G. U. Lindberg. 1930. A review of the fishes of the seas of the Far East. Izvestiia Tikhookeanskogo nauchnogo instituta rybnogo khoziaistva [Bulletins of the Pacific Science Institute] v. 5: i-xlvii+1-576, Pls. 1-15. [In Russian, English summary, and new taxa also in English.]

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

Stein, D. L. 1980. Aspects of reproduction of liparid fishes from the continental slope and abyssal plain off Oregon, with notes on growth. Copeia 1980:687–699.

Stein, D. L. 2012. A review of the snailfishes (Liparidae, Scorpaeniformes) of New Zealand, including descriptions of a new genus and sixteen new species. Zootaxa 3588:1–54.

Stein, D. L. 2016. Description of a new hadal *Notoliparis* from the Kermadec Trench, New Zealand, and Redescription of *Notoliparis kermadecensis* (Nielsen) (Liparidae, Scorpaeniformes). Copeia 104:907–920.

Steinke, D., T. S. Zemlak, J. A. Boutillier, and P. D. N. Hebert. 2009. DNA barcoding of Pacific Canada's fishes. Marine Biology 156:2641–2647.

Stoler, N., and A. Nekrutenko. 2021. Sequencing error profiles of Illumina sequencing instruments. NAR Genomics and Bioinformatics 3:lqab019.

Sukumar, S., A. Krishnan, and S. Banerjee. 2021. An Overview of Bioinformatics Resources for SNP Analysis. p. 113–135. *In*: Advances in Bioinformatics. V. Singh and A. Kumar (eds.). Springer Singapore, Singapore.

Taranetz, A. Ya. 1937. Handbook for identification of fishes of Soviet Far East and adjacent waters. Izvestiia Tikhookeanskogo nauchnogo instituta rybnogo khoziaistva [Bulletins of the Pacific Science Institute] 11:1-200 + map. [In Russian.]

Wang, Y., H. Yuan, J. Huang, and C. Li. 2022. Inline index helped in cleaning up data contamination generated during library preparation and the subsequent steps. Molecular Biology Reports 49:385–392.

Wilimovsky, N. J. 1954. List of the fishes of Alaska. Stanford Ichthyological Bulletin. 4:279–294.

Wilimovsky, N. J. 1958. Provisional Keys to the Fishes of Alaska. US Fish & Wildlife Service, Fisheries Research Laboratory, Juneau, Alaska.

Yuan, H., C. Atta, L. Tornabene, and C. Li. 2019. Assexon: Assembling Exon Using Gene Capture Data. Evolutionary Bioinformatics 15:1–13.

APPENDIX 1

Appendix 1. Summary of modern phylogenetic analyses of the family Liparidae (based on Orr et al., 2019)

Phylogenetic Analyses	Dataset based on	Specimen Description	Major Genera Monophyly
Kido, 1988	Osteology and external morphology characters	60 species in about 9 genera; focused on Japanese waters with select North Atlantic species added	*Liparis*–monophyletic *Careproctus*–paraphyletic *Paraliparis*–paraphyletic
Balushkin, 1996	7 osteological and external characters	All 27 known genera at the time	*Liparis*–monophyletic *Careproctus*–paraphyletic *Paraliparis*–not examined
Knudsen et al., 2007	Over 1000 bp of 16S and cytb combined with morphological data	24 species in 9 genera from the North Atlantic and North Pacific	*Liparis*–monophyletic *Careproctus*–monophyletic *Paraliparis*–paraphyletic
Steinke et al., 2009	Study focuses on species identification: sequences of COI 650 bp in length	78 tissues from 19 species in 8 genera from the Canadian Pacific	*Liparis*–monophyletic *Careproctus*–paraphyletic *Paraliparis*–paraphyletic
Duhamel et al., 2010	103 unique sequences of COI 668 bp in length	157 tissues from 46 species in 9 genera; Antarctic focus but other regions included from published sequences	*Liparis*–paraphyletic *Careproctus*–paraphyletic *Paraliparis*–paraphyletic
Gardner et al., 2016	Study focuses on egg identification: sequences of COI 429 bp in length	53 adult tissues from 23 species in 9 genera from the Eastern North Pacific	Large polytomy, no conclusion about monophyly or lack thereof
Shen et al., 2017 Tree 1	Sequences of 16S 550 bp in length, and cytb 400 bp in length (primarily mined from Knudsen et al., 2007)	25 species in 9 genera from the North Atlantic, Arctic, North Pacific, and Mariana Trench	*Liparis*–monophyletic *Careproctus*–monophyletic *Paraliparis*–paraphyletic
Shen et al., 2017 Tree 2	Sequences of COI 440 bp in length	88 species in 9 genera from the North Atlantic and Pacific, Arctic, Antarctic, and Indian Ocean	*Liparis*–paraphyletic *Careproctus*–paraphyletic *Paraliparis*–paraphyletic
Orr et al., 2019 Tree 1	270 unique sequences of COI 492 bp in length	122 species plus 10 individuals identified to genus only plus two outgroup species from wide geographic range	*Liparis*–paraphyletic *Careproctus*–paraphyletic *Paraliparis*–paraphyletic
Orr et al., 2019 Tree 2	40 sequences, 33,370 bp in length from 355 RADseq markers (each 94 bp after trimming and concatenation)	25 liparid species plus two outgroup specimens all from the Eastern North Pacific	*Liparis*–monophyletic *Careproctus*–paraphyletic *Paraliparis*–paraphyletic