UNIVERSITY OF CALIFORNIA
SANTA CRUZ

SYSTEMATICS AND POPULATION GENOMICS OF CLINID FISHES FROM SOUTHERN AFRICA AND NORTH AMERICA

A dissertation submitted in partial satisfaction
of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

ECOLOGY AND EVOLUTIONARY BIOLOGY

by

Daniel Wright

September 2022

The Dissertation of Daniel B. Wright
is approved:

_____
Professor Giacomo Bernardi, Chair

_____
Professor Peter Raimondi

_____
Professor Sophie von der Heyden

_____
Professor Rauri Bowie

_____
Adjunct Associate Professor Devon Pearse

_____
Peter Biehl
Vice Provost and Dean of Graduate Studies

Table of Contents

List of Tables and Figures

Abstract

Systematics and population genomics of clinid fishes from Southern Africa and North America

by

Daniel B. Wright

Biodiversity has faced a steep decline during the Anthropocene and as human influence on ecosystems through the direct impacts of development and indirect impact of climate change continue to increase the threat to global biodiversity is higher than ever. Coastal marine ecosystems are particularly susceptible to human activities due to their proximity to human settlements and ease of access. In order to protect coastal oceans, it is important to accurately catalog the biodiversity of these recreationally and commercially valuable ecosystems, but marine systems differ in the mechanisms that drive population structure and incipient speciation common to terrestrial systems. Large population size, a lack of physical barriers, and high dispersal potential make it a challenge to identify reproductively isolated populations in marine environments. Molecular techniques have provided solutions to some of these challenges as they are capable of determining where populations fall on the continuum between panmixia and speciation and allow investigators to identify previously undescribed biodiversity (interspecific and intraspecific genetic variation) in marine ecosystems. In this dissertation I use genome-wide sequencing techniques to search for patterns of cryptic speciation and explore the population dynamics of fishes from the family Clinidae in two dynamic coastal ecosystems, Southern Africa

and Western North America.

In **Chapter 1**, we investigate the systematics and population structure of fishes in the genus *Gibbonsia* along North America's west coast using restriction-site associated DNA sequencing (RADseq). We find that while there are three widely recognized species in the genus phylogenetic inference produces four genetic lineages, with G. elegans samples collected from Guadalupe Island separating into their own clade. These findings support the description of a Guadalupe Island endemic sub-species originally described as *G. e. erroli* by C.L. Clark. Among the other two species we find high levels of genetic connectivity leading to a panmictic population of *G. metzi*, but when analyzing outlier loci we see geographic partitioning for *G. montereyensis*. In **Chapter 2**, we present the assembled genome of the Southern African clinid *Clinus superciliosus* and use it to align sequence reads and search for phylogenetic patterns of cryptic speciation for *C. superciliosus* and *Muraenoclinus dorsalis*. The assembled genome was highly contiguous (no gaps) and complete with a BUSCO of 94% and represents a valuable tool for genomic investigation into the clinid fishes of South Africa. Phylogenetic analysis identified three clades among the *C. superciliosus* samples and two clades among *M. dorsalis* samples consistent with prior genetic studies. We identify samples to the species level for *C. superciliosus* using published phylogenies but were unable to do so for *M. dorsalis* thus supporting the existence of an undescribed cryptic *Muraenoclinus* species. In **Chapter 3**, we use the same RADseq dataset to investigate the population structure for three species of clinid from Southern Africa, *C. superciliosus*, *M.*

dorsalis, and *Clinus cottoides*. We find all three populations are highly structured at short distances along South Africa's Western Cape. Our results support the presence of biogeographic breaks at Cape Point and Cape Agulhas and False Bay as a secondary contact zone for populations separated during the last glacial maxima.

The results of the dissertation highlight the extent of undescribed biodiversity present in coastal marine ecosystems at the interspecific and intraspecific level and demonstrates the power of modern molecular techniques to identify cryptic species and genetically distinct populations for conservation consideration.

Acknowledgements

I would like to start by thanking my advisor Giacomo Bernardi for being such a kind and supportive presence during my Ph.D. and for turning me into an unapologetic fish guy. I will forever be grateful for his confidence in my ideas and his willingness to chase after them with me. I have enjoyed my time in Giacomo's lab more than I can express in words and I will wear the Bernardi Lab label with pride throughout my career. I would also like to thank the rest of my committee starting with Sophie von der Heyden, who has had to put up with me academically for a full decade now. This dissertation would not be possible without Sophie's knowledge and support. She is my guide when I am in South Africa and has opened her lab to me when I've needed it and even trusted me take her graduate students into the field knowing that means I must drive them around on the left side of the road. I cannot thank her enough for her support and I look forward to continuing to collaborate with her in the future. I need to thank Peter Raimondi, Rauri Bowie, and Devon Pearse for sharing their considerable experience in the fields of marine ecology and genomics. Rauri provided valuable insight into the South African marine systems and molecular techniques used in the dissertation. Devon has taught me more about the genome than anyone else and been a sounding board for my ideas for my entire Ph.D. And a special thank you to Pete, who has been on all my academic committees and has kept me grounded in what I can and cannot say with statistical certainty about the difference between populations and species.

I would like to thank EEB for their funding and support during my 5 years

here. I have always felt at home here and found the faculty approachable and supportive through global pandemics and wildfires. I am grateful to be a part of the UCSC EEB family.

I would like to thank my family for their unwavering support during my (ridiculously) long, I'll call it addiction to learning. They have always trusted me to make good decisions, even when it wasn't obvious to them that I was indeed making good decisions. I love you all.

I need to give a special thanks to my lab mates again without whom this dissertation would not be possible. To Jacob, Allyson, and Cerise, thank you for fun dives, fieldwork, inspiration during lab meetings, and for making my time here so enjoyable. A special thanks to Eric Garcia for walking me through my first library preparation. Also, special thanks to Jason Toy, Remy Gatins, and May Roberts who were all pioneering new genomic techniques for the lab and saved me incalculable amounts of time by generously sharing their pipelines and expertise with me. And while not officially member of our lab, I'd like to thank Melissa Cronin and Kat Dale whose own dissertation work offered me opportunities to advance my own understanding and skill in the lab while having fun with two amazing people.

There is a whole cast of Santa Cruz characters that I would like to thank. To the best triathlon team Santa Cruz has ever seen, Tom and Andrew who provided plenty of beer drinking, camping, and ping pong distractions during my Ph.D. To my first roomie Mari who got me settled in Santa Cruz and introduced me to most of the people I love here, including my wife. To the rest of the Spanish/MJA crew, Ale,

Mani, Mauricio, Roberto, Diana, Rafa, Maria, Diya, and Paul zie German thanks for the many Wednesday trips to the wine bar and all the wonderful dinners and straight up throw downs on Alice Street. A special thank you to Wes, Regina, and Rufus and Adam, Kathy, and little Wes who have been amazing friends in every aspect of my life here, love you guys.

Saving the best for last I would like to thank my wife Pauline and our cat Moochi. Thank you, Pauline, for taking this leap with me and for all the countless ways you have improved my life. You have supported me through the entire Ph.D. and now I can't imagine doing it without you. You are the best. And to Moochi, the silly gray fluff ball that sits in my lap while I decompress after a long day, I say psstpsstpsstpsstpsst. She knows what I mean.

Introduction

Molecular methods

Over the last two decades, the field of molecular ecology has seen something of a revolution in the scale of DNA sequencing data that are now accessible to investigators (Ellegren 2014). Molecular sequencing technologies have transitioned from first generation to second and third generations (commonly referred to as next generation sequencing, NGS) opening the full breadth of the genome to investigation. First generation sequencing used Sanger sequencing (Sanger et al. 1977) a highly accurate method of DNA sequencing, but the technology is limited in its throughput as options for parallel sequencing improvements are limited. Molecular markers using first generation sequencing target specific regions of the mitochondrial or nuclear genome using specialized primers and polymerase chain reactions (PCR) to amplify the targeted regions. This method of sequencing results in high depth of coverage (many copies of the same region) which leads to high confidence in the accuracy of base calling, but it is extremely limited in the breadth of coverage across the genome. Studies using single or few markers require prior knowledge of the marker characteristics as the mutation rate and expression status (i.e. neutral or expressed) play an important role in the outcome of analysis (Zhang and Hewitt 2003).

NGS platforms have developed methods for massively parallel sequencing to achieve high throughput, thus addressing the primary limitation of Sanger sequencing. The development of massively parallel sequencing platforms coupled

with a parallel advance in computational capacity (e.g. super computers, cloud analytics) have led to a revolution in molecular ecology (Ellegren 2014) allowing for an increase in breadth of coverage from single targeted regions to the entire genome. This increase in breadth of coverage makes it possible to identify anonymous molecular markers such as single nucleotide polymorphisms (SNPs), from neutral and expressed regions of the genome widening the potential range of analyses and testable hypotheses (Davey et al. 2011).

Because the amount of data generated in each sequence run is fixed for NGS platforms, their remains an important tradeoff between breadth of coverage, depth of coverage, and the number of specimens included in the study. High breadth of coverage (e.g. whole genome resequencing) provides the highest possible number of polymorphic loci, but requires a lot of sequence data for 1X coverage for 1 individual, especially for species with large genomes (Ekblom and Wolf 2014). This level of sequencing is appropriate for questions that require a relatively low number of sampled individuals to address (e.g., seascape ecology) but are currently too expensive to use for population level analysis (Alex Buerkle and Gompert 2013).

Systematic and population genomics studies that require larger numbers of specimens have turned to genotype by sequencing (GBS) which produces a reduced representation library and strikes a balance of genome-wide polymorphic markers with good depth of coverage (generally >10X) for many individuals (Fuentes-Pardo and Ruzzante 2017; Campbell et al. 2018). There are several GBS methods (e.g. restriction enzyme, target probe, transcriptome sequencing), but restriction-site

associated DNA sequencing (RADseq) has emerged as the most popular form (Davey et al. 2013; Narum et al. 2013) and has been applied to numerous phylogenomic and population genomic studies. RADseq targets the region of DNA adjacent to restriction enzyme cut sites producing short raw read sequence data (100-150 bp) from across the breadth of the genome. Because the region sequenced is much shorter than the distance between cut sites, most of the genome is not sequenced allowing for greater depth of coverage and more individuals included in one sequence run while still providing tens of thousands of polymorphic loci for analysis.

A challenge of using anonymous markers such as SNPs over traditional targeted loci is that sophisticated statistical tools are required to filter the data in order to identify haplotypes for comparison between individuals, and this filtering can significantly affect downstream analyses and results (Catchen et al. 2013). Ideally an assembled genome is available to aligning sequence reads to, which yields the highest number of loci for a given dataset, but *de novo* mapping for novel species is possible with RADseq data (Davey et al. 2011). The ability to produce *de novo* catalogs represents an advantage over whole genome resequencing techniques, which require an assembled genome to analyze. This combination of coverage, affordability, and the lack of need for additional genomic resources (i.e., an assembled genome) have made RADseq a powerful tool for analysis of population genomics, systematics, biogeography, adaptation, and speciation in non-model organisms.

Cryptic speciation in the sea

Speciation is at the core of evolutionary biology but has proven elusive subject to understand. The best understood scenario for speciation is allopatric speciation when a vicariant event separates a relatively small population from the larger source population (Coyne and Orr 2004a). This barrier prevents gene flow for long enough that a combination of drift and selection lead to prezygotic or postzygotic barriers to reproduction preventing introgression even if populations come back into contact (Futuyma and Kirkpatrick 2017). This model of speciation is highly applicable in terrestrial and in freshwater systems, which are divided by geological and climactic events. Marine organisms, however, tend to have large ranges, high population sizes, and few physical barriers to dispersal (Palumbi 1994). Additionally, many marine species undergo a larval phase, making it possible to reach distant localities and leading to the hypothesis that marine organisms, in general, are genetically well connected. Speciation in marine systems requires less conspicuous barriers to gene flow such as behavior, selection, isolation by distance, courtship, use of pheromones, spawning time, and geological history (Palumbi 1994). The somewhat cryptic nature of these mechanisms may, when combined with other factors, facilitate the phenomenon of cryptic speciation, which can be common in marine systems (von der Heyden 2011).

Cryptic speciation is when a population becomes reproductively isolated enough to genetically diverge from other populations but lacks a corresponding morphological change (Unmack et al. 2021). There are hypothesized mechanisms for cryptic speciation that are consistent with traits shared by many marine organisms

(Bickford et al. 2007). Sexual selection can lead to reproductive isolation and may rely on a cue that is not visible to the human eye. Marine species often use pheromones to attract a mate and are known to rely on markings that occur outside of the visible spectrum such as ultra-violet light (Macías Garcia and de Perera 2002; Wyatt 2003). Courtship behavior can also lead to reproductive isolation and would only be apparent to investigators if they were observing the target species during a courtship event. These traits alone can lead to reproductive isolation and eventually cryptic speciation, but the likelihood increases when environmental factors act to limit the range of successful morphotypes.

Marine environments can place selection pressure leading to convergent body forms and coloration in marine organisms (Bale et al. 2015). An excellent example of this is the rocky intertidal, where inhabitants are exposed to tidal forces, wave action, desiccation, and daily fluctuations in pH, salinity, and temperature. These environmental forces have led to convergent body forms and coloration in intertidal regions globally (Horn and Gibson 1988). The combination of mate selection mechanisms that require specialized equipment to observe and an environment that limits morphotypes create a scenario where reproductive isolation can lead to insipient speciation, without an apparent change in morphology (i.e., cryptic speciation). Shallow marine systems appear to be rich in examples of cryptic species, especially among invertebrates (Knowlton 1986; Williams et al. 2001; Teske et al. 2007). There are fewer examples for shallow water fishes but there is evidence for cryptic speciation in the family Clinidae (von der Heyden et al. 2011).

Fishes of the family Clinidae

The family Clinidae (clinids) encompasses 3 tribes, 20 genera, 71 species of fishes (Nelson 2006). They have a disjunct global distribution with the Clinini and Ophiclinini tribes occurring in South Africa, Australia, and Indian/Pacific Ocean islands and the Myxodini tribe occurring along the coasts of North America, South America, and in the Mediterranean Sea (Stepien 1992). Clinids are nearshore (intertidal or shallow subtidal) species that predominantly inhabit the benthos. They are highly cryptic living either under rocks and boulder or among algae that grows in tide pools or on rocky reefs in the shallow subtidal zone. They employ camouflage to avoid predation and their coloration closely matches the sand, rock, or algae they inhabit. Studies have even shown some species are able to change their coloration to match their habitat (Stepien 1988).

Southern hemisphere Clinini and Ophiclinini clinids are live bearing giving birth to post-flexion larvae with virtually no larval duration (Branch et al. 2016). By contrast, Myxodini species are brooders, attaching their eggs to nests and the offspring are thought to have a relatively long (~2 month) larval duration (Stepien 1992). This difference in reproductive strategy leads to contrasting predictions of dispersal potential and contrasting patterns of population structure and potential for instances of cryptic speciation, making clinids an interesting comparative study system for phylogenomic and population genomic investigation.

Conservation Genomics and Chapter Description

In a world increasingly impacted by climate change and human activities, that

requires intervention and active management to maintain extant biodiversity, understand the extent of that biodiversity and the population dynamics that shape its evolutionary trajectory is critical (Sintayehu 2018; Weiskopf et al. 2020). In the marine environment, conservation management often takes the form of marine protected areas (MPAs) and reserves that limit the extraction of marine resources. Marine protected areas are most effective when they connect populations through dispersal via larval recruitment or adult migration leading to an increase in the management of MPAs as networks of reserves rather than individual units (Gleason et al. 2013; Kirkman et al. 2021). Conservation management efforts have been hampered by incomplete catalogs of biodiversity and deficient understanding of population dynamics that dictate contemporary distributions. Genetic, and more recently genomic, studies have provided clarity into some of the primary inputs to prioritizing conservation efforts. Molecular techniques can identify genetic variation (a component of biodiversity) among populations that are morphologically indistinguishable at the interspecific (i.e., cryptic species) and intraspecific (i.e., population structure) levels (Hoban et al. 2013). Genetic methods provide powerful tools for identifying otherwise invisible biogeographic barriers that limit dispersal and are capable of identifying genetically diverse units that are important for conservation (Angeloni et al. 2012).

In an effort to better understand the key components of conservation management, we investigate the systematics and population dynamics of fishes in the family Clinidae in two locations, the southwest coast of South Africa and the west

coast of North America. In chapter one we focus on the genus *Gibbonsia* (*G. montereyensis, G. metzi, G. elegans*) on North America's west coast. We use RAD SNPs to produce a phylogeny to assess the number of genetic lineages present among our samples and explicitly test for the presence of cryptic species. For two of the study species (*G. metzi, G. montereyensis*), we use the same SNP dataset to compare levels of population structure between the closely related congeners and between the full dataset (all SNPs) and putatively adaptive loci (outlier loci).

In chapter two we change systems from the west coast of North America to South Africa's southwest coast. We first present the assembled genome of *Clinus superciliosus* which represents the first published genome for a clinid species. We use the genome to align a panel of SNPs for two Southern African clinids *C. superciliosus* and *M. dorsalis* including newly sampled specimens for this study. We produce phylogenies with the resulting dataset to search for the pattern of cryptic speciation previously identified using first generation genetic markers (von der Heyden et al. 2011). Finally, we use other published phylogenies to attempt to identify specimen to the species level for *C. superciliosus*.

For chapter three we build off the work from chapter 2 to assess population structure among 3 species of Southern African clinid (*C. superciliosus, M. dorsalis, C. cottoides*). We use the phylogenies produced in chapter two to ensure we are comparing between intraspecific populations and generate a panel of SNPs aligned to the *C. superciliosus* genome for all three species. We then apply multiple clustering analyses to explore population structure and the presence of biogeographic breaks

along South Africa’s Western Cape.

Chapter 1: A genomic contribution to the classification of Blennioid fishes of the family Clinidae

Abstract

There are four presently recognized species of clinid (kelpfish family: Clinidae) on the west coast of North America *Heterostichus rostratus* (Girard), *Gibbonsia elegans* (Cooper), *G. montereyensis* (Hubbs), and *G. metzi* (Hubbs). Clark L. Hubbs suggested that there were endemic subspecies in the genus *Gibbonsia* on Guadalupe Island in his 1952 publication "A contribution to the classification of the blennioid fishes of the family Clinidae, with a partial revision of the Eastern Pacific forms". Later investigation using allozymes showed high levels of gene flow between populations and treated subspecies names as synonyms, leading to the present-day status. Here, we use high throughput sequencing and a combination of nuclear and mitochondrial markers to revisit the systematics and population structure of the genus *Gibbonsia*. We find that Guadalupe Island specimens that are morphologically identified as *G. elegans* form a distinct genetic lineage for both nuclear and mitochondrial markers, supporting Hubb's hypothesis of a Guadalupe Island endemic he named *G. erroli*. Further, we find that while dispersal potential is high for all species in the genus, genome-wide markers and outlier analyses show population structure at short geographic distances suggesting fine scale environmental variation leading to local adaptation.

Introduction

Gene flow is a well-understood mechanism for maintaining genetic similarity

between intraspecific populations. When gene flow is high, populations are panmictic and cannot be genetically distinguished from one location to another (Hahn 2018). The opposite extreme is when populations become reproductively isolated and gene flow between populations ceases, which, given enough time, results in speciation (Coyne and Orr 2004a; Futuyma and Kirkpatrick 2017). In between there is a wide range of population structuring, with a decrease in gene flow sufficient to identify the sampling origin of individuals by their genetic background but maintaining a sufficient level of gene flow to prevent speciation. Varying levels of genetic connectivity, based on a species' dispersal potential, leads to a gradient of population structuring, with panmixia at one end, and incipient speciation at the other. Identifying where populations fall along this spectrum has long been the subject of molecular investigations in many taxa (Meirmans and Hedrick 2011; Hahn 2018) including marine fish (Teske et al. 2011; Bernardi 2013a; Waples et al. 2020; Caballero-Huertas et al. 2022).

Most marine fishes have a bipartite life history with a pelagic larval stage lasting from days to months depending on the species, and a sedentary adult stage (Leis 1991). This life history trait, coupled with the lack of conspicuous barriers to dispersal in marine environments led many investigators to hypothesize that populations of marine fishes were likely to experience high levels of gene flow and therefore low levels of population structure (Caley et al. 1996; Selkoe et al. 2016). Genetic techniques have allowed investigators to measure the genetic differentiation between populations and many studies have revealed high levels of

population structure in marine fishes based on habitat (Kelly and Palumbi 2010; von der Heyden et al. 2013), life history (Selkoe et al. 2014; Wright et al. 2015) and biogeography (Teske et al. 2011; Dalongeville et al. 2022). In particular, marine fishes in shallow, near-shore, marine habitats can exhibit high levels of population structure, and multiple studies have found evidence of high levels of population structure in the family Clinidae (Teske et al. 2011; von der Heyden et al. 2013; Wright et al. 2015).

Fishes from the family Clinidae have a disjunct global distribution comprised of 20 genera, 71 species, 3 tribes (Stepien 1992; Nelson 2006). Most clinid species are part of the Clinini and Ophiclinini tribes and occur in South Africa and Australia, but there are presently eight recognized species in the tribe Myxodini that occur along the coasts of the Americas and in the Mediterranean Sea (Stepien 1992). Clinids are generally nearshore (intertidal or shallow subtidal), benthic species that are highly cryptic and capable of changing their coloration to match their surroundings (Stepien et al. 1988). Their reproductive strategy can be either viviparous with low larval dispersal (e.g., genus *Clinus* from South Africa) or oviparous with high larval dispersal (e.g., *Gibbonsia* from North America). In oviparous species, territoriality and parental care have been observed in the form of males guarding an oviposition and tending the broods of eggs (Love 2011).

On the west coast of North America there are presently 4 recognized species of clinid in two genera: *Heterostichus rostratus* (Girard), *Gibbonsia elegans* (Cooper), *G. montereyensis* (Hubbs), and *G. metzi* (Hubbs). In the genus *Gibbonsia*

the three species are separated into the northern species that inhabit cooler water provinces north of Point Conception, *G. montereyensis* and *G. metzi*, and the warmer water Southern California province inhabitant *G. elegans* (including the offshore island of Guadalupe, Mexico). However, all three species have been found to co-occur in parts of southern California and in areas of upwelling in Punta Clara, Mexico (Stepien and Rosenblatt 1991).

Much of what is known about *Gibbonsia* species comes from Clark L. Hubbs (Hubbs 1952), and later work led by Carol Stepien (Stepien 1986; Stepien et al. 1988; Stepien and Rosenblatt 1991; Stepien 1992). C.L. Hubbs described the three presently recognized forms of *Gibbonsia* along with a proposed new species *G. erythra* several proposed subspecies *G. montereyensis vulgaris, G. elegans verifera*, *G. elegans rubrior*, and two Guadalupe Island endemics *G. montereyensis norea*, and *G. elegans erroli*. Systematics of the genus based on a combination of morphological character and allozyme analysis suggested *Gibbonsia* populations are well connected by gene flow with little evidence of population structure attributed to the relatively long larval duration (2 months) and corresponding high dispersal potential (Stepien and Rosenblatt 1991). The new species *G. erythra*, and subspecies *G. e. verifera* and *G. m. norea* showed lower differentiation compared *to G. montereyensis* and *G. elegans* samples and other congeneric clinids. Meristic and morphometric analyses showed that character differences likely corresponded to previously undescribed sexual dimorphism and depth segregation (Stepien and Rosenblatt 1991). Ultimately, all sub-species were synonymized with *G. montereyensis* and *G. elegans*, resulting in

the present status of 3 recognized species in the genus.

Here we employ higher resolution molecular techniques and a combination of nuclear and mitochondrial markers to revisit the systematics and population structure of species in the genus *Gibbonsia*. We use restriction associated DNA sequencing (RADseq) to produce a genome-wide set of nuclear single nucleotide polymorphisms (SNPs) to test for fine scale population structure. We use the same dataset along with the 16S rRNA mitochondrial marker to infer phylogenetic relationships among species. We test 1) if the combination of nuclear and mitochondrial markers provides molecular evidence to support the subspecies described by C.L. Hubbs (1952) and 2) if by using genome-wide SNPs we can uncover genetic structure between populations of *G. metzi* and *G. montereyensis*. 3) for the presence of adaptive variation, as indicated by outlier loci.

Materials & Methods

Sample collection and DNA extraction

We sampled fish specimens from locations in the Monterey Bay area, Catalina Island in Southern California, and Guadalupe Island, Mexico (Figure 1.1). The number of specimens captured from each sample location can be viewed in Table 1.2. Fish specimens were netted at low tide either in tide pools or in the shallow subtidal. Fin clips and muscle tissue were immediately preserved in 95% ethanol and stored at -20 C. We extracted DNA from tissue using DNeasy Blood and Tissue kits (Qiagen) or chloroform/isopropanol extraction protocol and assessed DNA concentrations using a Qubit 4.0 fluorometer.

Library preparation and sequencing

We constructed restriction site associated DNA (RAD) libraries for all samples and amplified the 16S mitochondrial region for a subset of specimens. We constructed the RAD libraries using a variation of the original protocol (Miller et al. 2007; Baird et al. 2008) using restriction enzyme SBFI described in (Miller et al. 2012) and NEBNext reagents (New England Biolabs). We started the library preparation with 100 ng of genomic DNA (gDNA). We multiplexed and sheared libraries to roughly 500 bp lengths on a Biorupter Sonicator using four cycles of 30 seconds. We used Dynabeads (Invitrogen 11206D) to remove non-tagged DNA and used SPRI beads (DeAngelis et al. 1995) for purification and size selection. We carried out the final polymerase chain reaction (PCR) amplification on 16 ul reaction volumes for 10 amplification cycles. Samples used in the study were sequenced in one of two libraries, each containing individually barcoded samples. RAD libraries were sequenced on a NovaSeq 6000 S4 at Novogene Corporation, Sacramento, CA.

To investigate phylogenetic relationships based on mitochondrial markers, we amplified and sequenced the 16S rRNA region for two specimens each of *G. montereyensis* and *G. elegans*, 3 samples from Guadalupe Island, and one specimen of *Heterostichus rostratus* as an outgroup. For one of the samples from Guadalupe Island we were able to amplify 16S, but did not produce a successful RAD library. We amplified the 16S rRNA segments using the universal primers 16SAR-16SBR (Palumbi et al. 1991) with 35 cycles at a denaturation temperature of 94 ◦C

for 30 s, an annealing temperature of 52 ◦C, and an extension of 30 s at 72 ◦C. We purified the PCR product following the manufacturer's protocol and had them Sanger sequenced in both directions with the primers used in PCR amplification at the UC Berkeley DNA Sequencing Facility.

Quality filtering and marker discovery

For all specimens included in the study we trimmed, filtered, and genotyped raw reads using the software program Stacks version 2.2 (Rochette et al. 2019). We split raw reads according to their 6 bp unique barcode, removed the barcodes and trimmed reads on the 3' end to a final sequence length of 91 bp using the program *process_radtags*. Sequences were removed if the quality scores fell below 90%, a raw phred quality score of 10. We generated a denovo catalog using the Stacks 2.0 shell programs *ustacks*, *cstacks*, and *sstacks* on a subset of the specimens including the most data rich individuals from each species (20 specimens total).

We ran the rest of the denovo_map.pl pipeline manually (*tsv2bam*, *gstacks*) and set the minimum stack depth (-m), maximum mismatches per loci for each individual (-M) to three and allowed up to seven mismatches when building catalog loci (-n). We ran the Stacks program *populations* to generate output files for downstream phylogenetic analysis with each individual being treated as a population for phylogenetic inference. We retained only a single SNP per locus (write_single_snp option in Stacks).

We aligned the 16S rRNA sequence reads in Geneious v. 2020.2.4 (https://www.geneious.com) and converted the aligned reads into Nexus format for

downstream analyses.

PCA and phylogenetic inference

We first ran a principal components analysis (PCA) using the RAD dataset for all samples in the ADEGENET package in R (Jombart 2008) with the variant call format file (VCF) output from Stacks *populations* command. To generate a phylogenetic tree, we used the phylip format output from the *populations* in Stacks, which contain nucleotides that are fixed within groups but variable between groups (in this case individuals). We treated each individual as its own population and only retained SNPs that were present in ≥80% of individuals. We converted the resulting phylip files into Nexus format and filtered the list of nucleotides to the informative loci and performed maximum-likelihood (ML) implemented in Phyml (http://www.atgc-montpellier.fr/phyml/) and maximum-parsimony (MP), and neighbor-joining (NJ) phylogenetic inference as implemented in Paup* (Wilgenbusch and Swofford 2003). We generated support values for nodes using 1000 non-parametric bootstrap replicates.

For the 16S alignment we used Neighbor-Joining and Maximum-Likelihood methods with bootstrapping support as described above. Phylogenetic trees and support values were visualized using FigTree v1.4.4. (http://tree.bio.ed.ac.uk/software/figtree/).

Population structure analysis

To explore genetic differences between populations we employed two lines of cluster analysis. For bayesian cluster analysis we ran STRUCTURE 2.3.4 (Pritchard

and Wen 2010) using the structure format output file from Stacks for each species and all loci. We used a parameter set of 10,000 iterations as the burn-in with 200,000 iterations under the admixture model. We ran 10 repetitions for K values ranging from 1 to two more than the number of sample locations. We used the Evanno method (Evanno et al. 2005) implemented in Structure Harvester (Earl and vonHoldt 2012) to identify the highest likelihood for K.

Additionally, we ran a discriminant analysis of principal components (DAPC) (Jombart et al. 2010). We performed the analysis using the ADEGENET package in R (Jombart 2008) with the variant call format file (VCF) output from Stacks *populations* command. We determined the number of principal components to retain using the cross-validation tool xvalDapc as implemented in ADEGENET.

Outlier analysis

To examine if there was evidence of local adaptation leading to genetic population structure, we identified outlier loci, which are candidate loci that are putatively under natural selection. We used the genepop format output file from the Stacks *populations* command and converted it into Bayescan format using PGDspider2 version 2.1.1.5 (http://www.cmpg.unibe.ch/software/PGDSpider/). We ran Bayescan2.1 using the *G. montereyensis* specimens captured at Ocean Cove and Pacific Grove, and *G. metzi* specimens from Half Moon Bay and Davenport. There was not a sufficient number of specimens or sample locations for an outlier analysis for *G. elegans*. We visualized the output from Bayescan2.1 in R and retained loci with a critical threshold of ≤ 0.05. We then reran the *populations* command from

Stacks using the *white list* option, using loci identified by Bayescan as outliers, to produce a structure format file to repeat the STRUCTURE analysis.

Results

Nuclear and mitochondrial sequencing

RAD library sequencing generated > 50 Mb of filtered and cleaned data for all specimens. Denovo alignment resulted in a catalog of 102,315 consensus sequence reads across all species. Average depth of coverage for all specimens was 21x (min 4x, max 63x). The populations command identified 33,578; 24,962, and 19,822 SNPs for *G. montereyensis*, *G. metzi*, and *G. elegans* respectively. The number of SNPs exactly corresponds to the number of loci as only one SNP was kept for each locus.

For the mitochondrial 16S region we generated an alignment of 557 bp for two *G. montereyensis*, two *G. metzi* and seven *G. elegans*.

Phylogenetic inference and PCA

We were able to obtain a fully resolved phylogeny using RAD markers (Figure 1.3) with specimens separating into 4 clades and each methodology produced identical tree topologies (ML, NJ, MP). Specimens collected from the Monterey Bay area were identified as *G. montereyensis* or *G. metzi*, however the island specimens previously identified as *G. elegans* based on morphological characters separated into two clades consistent with their sampling location (Catalina Island or Guadalupe Island). A separate phylogeny using the 16S marker was mostly consistent with the RAD phylogeny again separating *G. elegans* samples from Catalina Island and Guadalupe Island into distinct clades. *G. montereyensis* and *G. metzi* fell into the

same clade with one nucleotide difference between one of the *G. metzi* samples and the other three specimens (Figure 1.4).

Principal component analysis based on the RAD dataset separated *G. montereyensis* and *G. metzi* samples into clusters along PC1 and separated the *G. elegans* samples into separate clusters along PC2 for a total of 4 distinct genetic clusters (Figure 1.5A).

Population Structure analysis

STRUCTURE analysis when run using all SNPs revealed no population structure for *G. montereyensis* or *G. metzi*. The Evanno method identified K = 2 for *G. montereyensis* and K = 5 for *G. metzi* as the most likely number of genetic clusters. Individuals were indistinguishable from each other across all locations for both species (Figure 1.6).

The discriminant analysis of principal components was concordant with STRUCTURE analysis for *G. metzi* and failed to separate individuals between sample locations. However, for *G. montereyensis,* Pacific Grove samples were separated from Ocean Cove samples (Figure 1.5B, C). Cross validation analysis retained 12 principal components for *G. montereyensis* and 10 for *G. metzi*.

Outlier analysis

To further investigate the DAPC results, we performed an outlier analysis. Bayescan2.1 identified 20 candidate loci for *G. montereyensis* but returned no outliers for *G. metzi* below the critical threshold of 0.05 so *G. metzi* was removed from downstream outlier analysis. STRUCTURE analysis using only the outlier loci

for *G. montereyensis* assigned all individuals from Ocean Cove to one genetic cluster and individuals from Pacific Grove as > 50% assigned to the second genetic cluster. The Evanno method identified K = 2 as the most likely number of genetic clusters (Figure 1.6C).

Discussion

Among the three species in the genus *Gibbonsia* we found genetic evidence supporting the full range of possibilities for gene flow from panmixia to incipient speciation. *Gibbonsia metzi* appears panmictic with genome-wide SNPs failing to distinguish one population from another. For *Gibbonsia montereyensis* however, there is a signal of population structure driven by a subset of putatively adaptive loci. *Gibbonsia elegans* samples from Guadalupe Island have become reproductively isolated enough that they correspond to a unique genetic lineage using both nuclear and mitochondrial markers.

Phylogenetic inference and taxonomy

Presently, there are 3 recognized species in the genus *Gibbonsia*, but phylogenetic inference and the principal components analysis consistently produces four distinct clades. *Gibbonsia montereyensis* and *G. metzi* each produce monophyletic clades, however, *G. elegans* specimens from Guadalupe Island and Catalina Island belong to separate genetic lineages and results were concordant for analyses based on nuclear (SNPs) and mitochondrial (16S) markers (Figures 1.3 and 1.4). This finding was further supported by the presence of four distinct clusters in the PCA (Figure 1.5A). All lines of evidence indicate that the Guadalupe Island *G.*

elegans are more closely related to Catalina Island specimens than either *G. montereyensis* or *G. metzi*. Phylogenetic relationships are further supported by morphological character analysis. Guadalupe island *G. elegans* have fewer dorsal fin rays (modal 7, range 5-8) than *G. metzi* (modal 9, range 7-10) and there are scales present on the caudal fin, which are not present in *G. montereyensis*, so identifications using a dichotomous key result in identification as *G. elegans* (Love and Passarelli 2020).

The combination of these lines of evidence suggests that the Guadalupe Island specimens from this study are examples of the Guadalupe Island endemic *Gibbonsia elegans erroli*, originally described as a sub-species by C.L. Hubbs in 1952. The primary morphological character used to differentiate *G. erroli* from *G. elegans* was the number of dorsal ray spines (DRS), with *Gibbonsia erroli* having a mode of 32 DRS (range 31-33), while *G. elegans* have a mode of 34 DRS (range 31-35) (Figure 1.7; Hubbs 1952). All three of the Guadalupe Island samples included in this study have 32 DRS, which is consistent with the description of *G. erroli.*

While there is overlap in the number of DRS between the *G. erroli* and *G. elegans*, this meristic trait when taken in context with the molecular evidence supports the existence of *Gibbonsia erroli* as a separate species from *G. elegans*. This appears to be a classic example of allopatric speciation by a founding population that recruited to an oceanic island (Guadalupe Island) and became isolated from mainland populations. The most likely mechanism for isolation is the California current, which has a consistent southward flow between Guadalupe Island and the

Baja California Peninsula and may prevent larvae from being carried far enough to the west to reach Guadalupe Island (Hickey 1979; Stepien and Rosenblatt 1991; Checkley and Barth 2009) except for those rare occasions that resulted in the original colonization. Future studies are warranted to ascertain whether there are populations of *G. elegans* on Guadalupe Island, or if all specimens are the endemic *G. erroli*, and to include samples from San Benito and Cedros islands and Punta Clara on the Baja Peninsula for genetic comparison.

Population Structure

Previous studies of population structure for *G. metzi* and *G. montereyensis* have shown little to no impediment to gene flow between populations and indeed our findings largely support this finding (Stepien and Rosenblatt 1991; Stepien 1992). *Gibbonsia metzi* in particular appears to be completely panmictic among the sampling locations included in this study. Analysis using STRUCTURE and DAPC failed to distinguish sample locations suggesting they may be considered as a single genetic population (Figures 1.5, 1.6). This finding was reinforced when Bayescan outlier analysis failed to identify loci below a critical threshold of 0.05 for *G. metzi*. *Gibbonsia metzi* can tolerate the widest range of temperature of any of the *Gibbonsia* species (Davis 1977; Stepien and Rosenblatt 1991). Genetically panmictic metapopulations favor phenotypic plasticity over local adaptation and are thus likely to be less susceptible to fine scale changes to sea surface temperature caused by cold water upwelling than other species.

Gibbonsia montereyensis appears to have an intermediate level of gene flow

when compared to *G. elegans* and *G. metzi*. Similar to the results for *G. metzi*, when all SNPs are included in the analysis for *G. montereyensis*, STRUCTURE detects no difference between sample locations. DAPC, which can detect population differentiation at a finer scale than a structure analysis (Jombart et al. 2010), did however separate Ocean Cove specimens from Pacific Grove when analyzing all SNPs (Figure 1.5C). This finding led to an outlier analysis producing 20 putatively adaptive loci. When STRUCTURE analysis was rerun using just the outlier loci, it identified 2 genetic clusters, which corresponded to sampling locations (Figure 1.6C). The sample locations of Ocean Cove and Pacific Grove are separated by approximately 317 km of coastline. In between there is suitable habitat for *G. montereyensis* where samples have been collected (Figure 1.8) creating the potential for a steppingstone pattern of connectivity. This, coupled with a two-month larval duration suggests that local adaptation is likely to play an important role in maintaining genetic variation between populations. A seascape ecology approach is out of scope for this study, but our findings suggest that the environment experienced by *G. montereyensis* is different enough between Ocean Cove and Pacific Grove to produce differential fitness between genotypes.

Clarity around the findings of this study could benefit from broader sampling. *Gibbonsia metzi* and *G. montereyensis* were found to rarely co-occur, limiting the number of locations included for each species. In particular, *G. metzi* samples were found along a relatively short stretch of coastline. The sample locations where *G. metzi* were sampled experience less intense upwelling events than where *G.*

montereyensis were found. It is possible that *G. metzi'*s tolerance of warmer water (Davis 1977) could explain the observed pattern. It is also possible that the species segregate by depth or rely on the presence of certain species of algae. Increasing the scale of sampling is warranted to test if the finding of genetic panmixia holds across a broader section of the species range.

Conclusion

In this study, we have presented a combination of genetic and morphological characters that suggest there is a Guadalupe Island endemic *Gibbonsia* species, *G. erroli*, as originally described by C.L. Hubbs (1952). Using next generation sequencing technology, we found that the other *Gibbonsia* species have varying levels of population structure from panmixia (*G. metzi*) to shallow but significant population structure (*G. montereyensis*). Within one genus we find the full range of possibilities from genetically indistinguishable populations to speciation-level reproductive isolation highlighting the important role adaptation and biogeography play in structuring populations, even among closely related congeners.

Tables and Figures

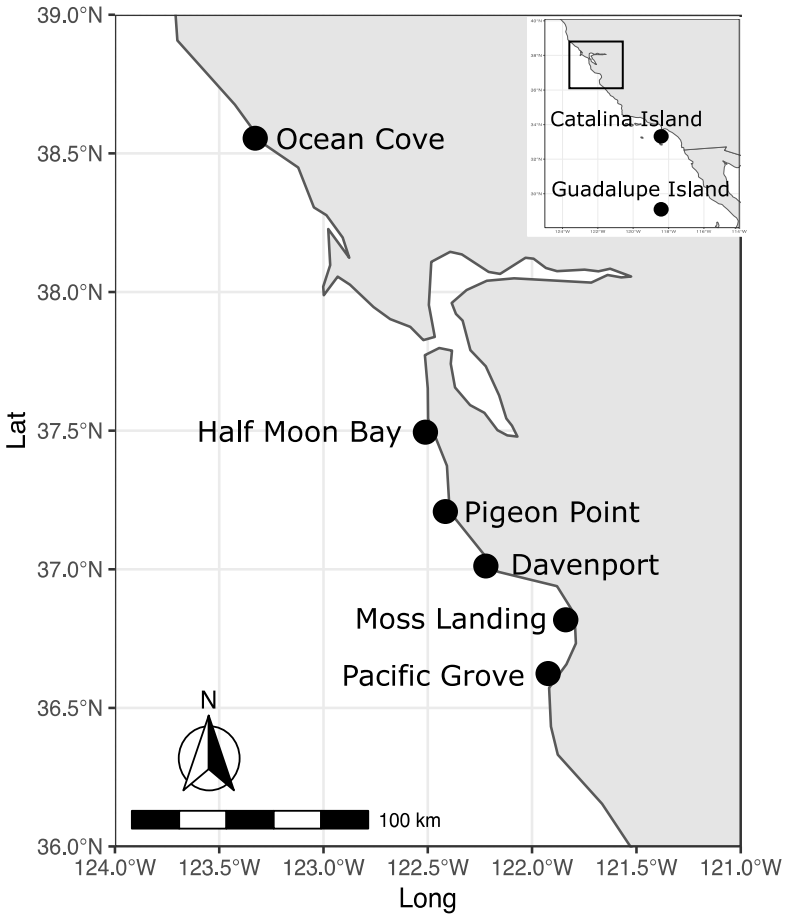


Figure 1.1. Map of sample locations. Main map shows locations in the Monterey Bay area, from north to south Ocean Cove, Half Moon Bay, Pigeon Point, Davenport, Moss Landing, Pacific Grove. Inset shows sample locations Catalina Island (northernmost point), and Guadalupe Island (southernmost point)

Table 1.2. A list of the study species the number of specimens collected from each sample location. * Guadalupe Island samples referred to as *Gibbonsia erroli* in text.

Species	Common name	Ocean Cove	Half Moon Bay	Pigeon Point	Davenport	Moss Landing	Pacific Grove	Catalina Island	Guadalupe Island	Total N
Gibbonsia montereyensis	Crevice kelpfish	14	-	4	-	-	8	-	-	26
Gibbonsia metzi	Striped kelpfish	-	6	3	9	1	-	-	-	19
Gibbonsia elegans	Spotted kelpfish	-	-	-	-	-	-	4	3*	7

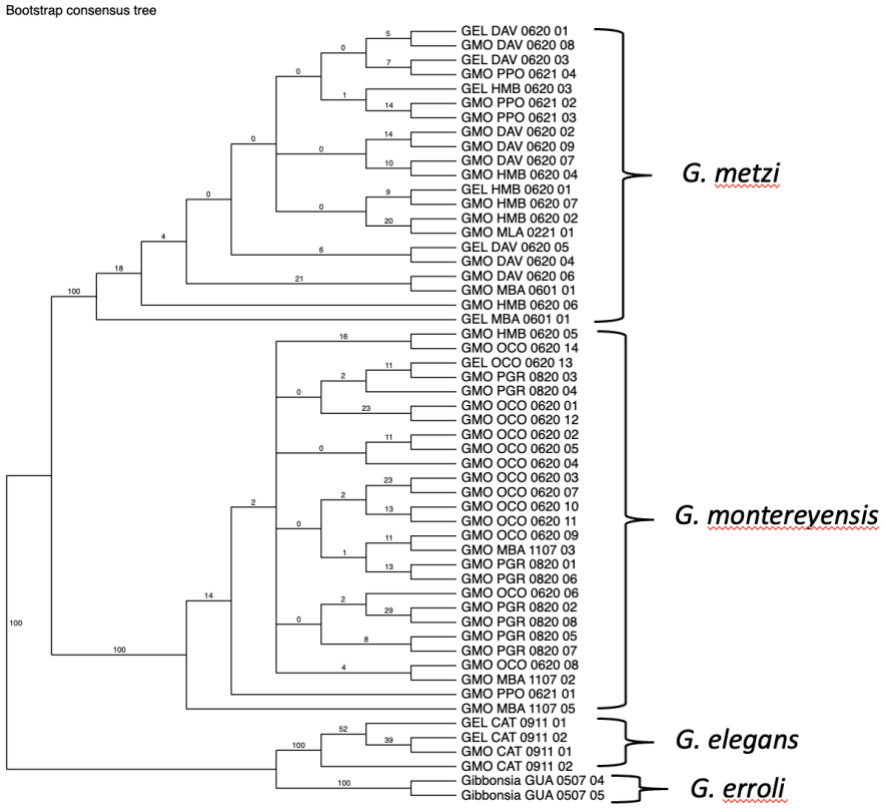


Figure 1.3. Phylogenetic tree inferred using RADseq SNPs. Guadalupe Island samples labeled as *G. erroli*.

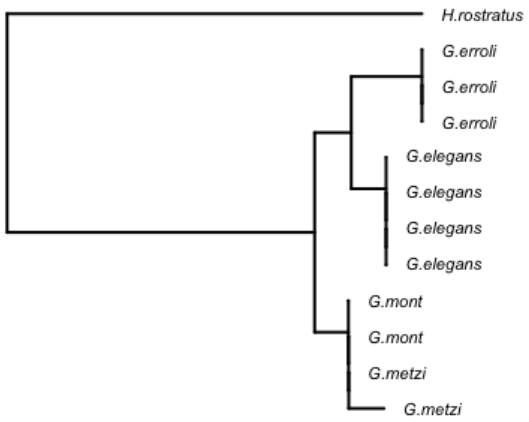


Figure 1.4. Neighbor joining phylogenetic tree using mitochondrial 16S rRNA marker with *Heterostichus rostratus* as an outgroup. Samples from Guadalupe Island, Mexico are labeled *G. erroli*.

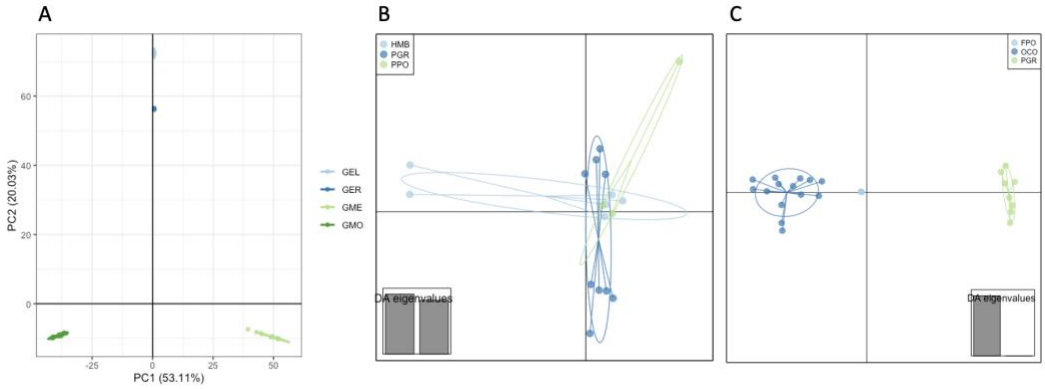


Figure 1.5. A) Principal component analysis of all species separating *G. montereyensis* (GMO) and *G. metzi* (GME) on PC1 and *G. elegans* (GEL) and *G. erroli* (GER) on PC2. B) Discriminant analysis of principal components (DAPC) for *G. metzi* and C) DAPC for *G. montereyensis*.

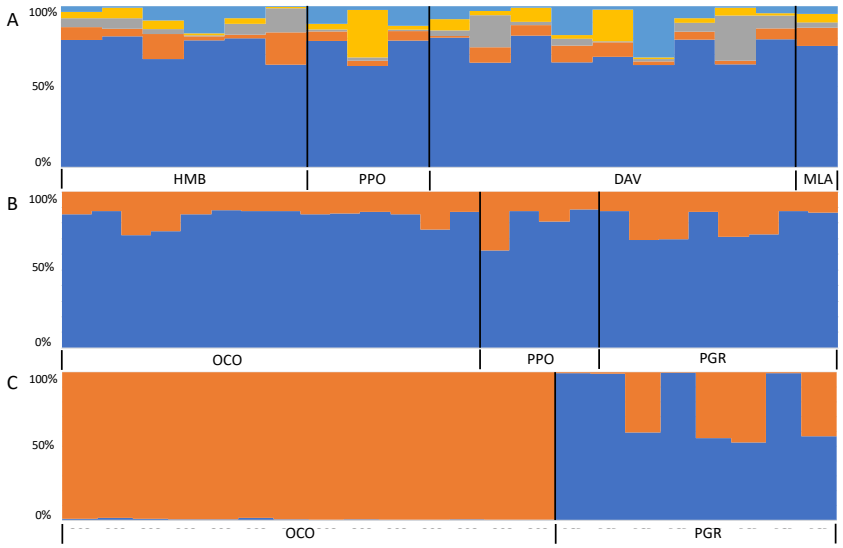
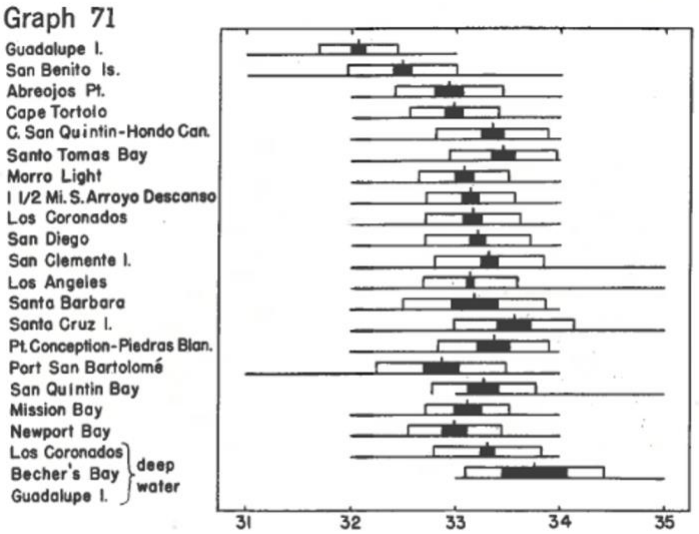


Figure 1.6. Structure plots for A) *G. metzi* using all loci B) *G. montereyensis* using all loci and C) *G. montereyensis* using outlier loci. Sample locations are separated by black lines. Abbreviations are as follows: HMB = Half Moon Bay, PPO = Pigeon Point, DAV = Davenport, MLA = Moss Landing, OCO = Ocean Cove, PGR = Pacific Grove.



GRAPHS 71-72. Subspecies of *Gibbonsia elegans*. Graph 71. Dorsal spines. Graph 72. Standard length into pectoral length, in percents.

Figure 1.7. Image of graph showing the distribution of the number of dorsal spines for sub-species of *G. elegans* from Hubbs (1952) including original figure description. First row shows data for the Guadalupe Island endemic *G. erroli*.

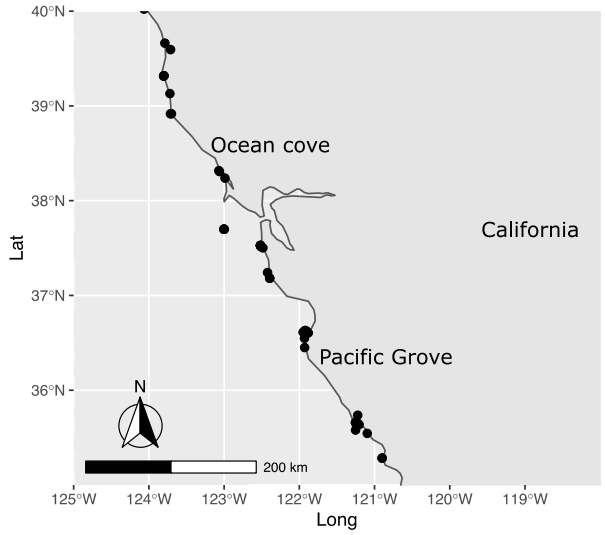


Figure 1.8. Map of central California. Black dots are collection locations for *G. montereyensis* specimens archived in Eschmeyer's Catalog of Fishes at the California Academy of Sciences. Sample locations for *G. montereyensis* included in this study are labeled.

Chapter 2: A genome and RADseq phylogeny support cryptic speciation in *Clinus superciliosus* and *Muraenoclinus dosalis*

Abstract

Speciation in marine systems present challenges to the classic model of allopatric speciation due to the large ranges, high population sizes, and lack of obvious barriers to dispersal for many marine organisms. The inconspicuous nature of marine speciation coupled with strong environmental pressures leading to convergent body forms and adaptations in marine systems may, in part, be responsible for the relatively high instance of cryptic speciation found among marine organisms. Here, we present two distinct examples of cryptic speciation from South Africa's rocky intertidal for *Clinus superciliosus* and *Muraenoclinus dorsalis* (family Clinidae). We first present a highly contiguous assembled genome for *C. superciliosus*, with a BUSCO of 94%, the first for a South African clinid, as a tool to assist in genomic investigations into population structure, biogeography, and local adaptation in this study system. We then use genome-wide single nucleotide polymorphisms (SNPs) to test for concordance in phylogenetic pattern with mitochondrial and nuclear markers from a prior study while attempting to identify specimens using published phylogenies. We find that mtDNA markers and SNPs produce concordant phylogenies for both study species. For *C. superciliosus* we were able to identify all genetic clades by comparing 16S rRNA to a published phylogeny

and highlight the difficulty of identifying samples for species with strikingly similar morphology at various life stages. For *M. dorsalis*, there is no known species with a similar morphology present in the intertidal making it a prime example of cryptic speciation in a shallow marine system.

Introduction

Speciation is central to the field of evolutionary biology and understanding the mechanisms that drive biodiversity and biogeography has been a central focus (Darwin 1872), (Dobzhansky 1937) (Mayr 1942). The classic model for speciation typically involves the emergence of a barrier to gene flow and a combination of drift and selection leading to genetic divergence between a small founder population and a larger source population (Coyne and Orr 2004b). The process of allopatric speciation is well understood in freshwater systems, which can be separated by land and natural or man-made barriers, but conditions are more complex in marine systems. Marine species have few apparent physical barriers to gene flow, tend to have large geographic ranges and population sizes, and many have a larval phase which, aided by ocean currents, allows for dispersal to distant localities (Palumbi 1994; Lester et al. 2007; Faria et al. 2021). As such, marine speciation is likely to occur through a number of mechanisms that are difficult to observe directly, such as invisible barriers to dispersal (e.g., currents and eddies), isolation by distance, selection, and behavior (e.g., courtship) (Palumbi 1994). The challenges associated with observing the mechanisms of speciation in marine systems make them cryptic in nature and their cryptic nature may in part, explain the relative prevalence of cryptic speciation in

marine systems (Bickford et al. 2007; von der Heyden 2011).

Cryptic species are populations that have no apparent difference in morphological character or whose morphological variation is within well-accepted intraspecific levels but represent genetically distinct evolutionary lineages. Cryptic species present challenges to investigators and conservation managers alike. For example, researchers may not anticipate the presence of cryptic species in their study design and are typically limited to field guides and dichotomous keys which rely on morphological characters to identify species. When not explicitly accounted for in an ecological model, this can lead to fixed interspecific variation being treated as intraspecific variation and can artificially alter the measured response to a treatment. Measures of species richness and diversity are necessarily impacted by the presence of cryptic species causing them to be undercounted. This is of particular importance to conservation management planning, which often relies on species richness metrics as a critical input for prioritizing protection (Myers et al. 2000; von der Heyden 2009). Biodiversity conservation efforts are hampered by an incomplete catalog of species and with the increased pressure human activity is placing on marine systems, knowing what needs to be conserved is critical to mitigation and conservation management (Hortal et al. 2015; Walters et al. 2021).

Intertidal fishes represent an excellent study system for cryptic speciation as the rocky intertidal environment with its variability in tidal inundation, temperature, salinity, pH, and wave action demands a common suite of adaptive traits, coloration, and body form from its inhabitants (Bernardi 2013b). Fishes of the rocky intertidal

are typically demersal, lack swim bladders, are desiccation-tolerant, and frequently employ camouflage to avoid detection (Horn and Gibson 1988). This collection of traits and closely shared habitat likely results in convergent body types and coloration for species and made intertidal and shallow subtidal fishes notoriously difficult to distinguish from one another (e.g., sculpins, gobies, kelpfish, clinids, triplefins, rockfishes), particularly in the field. Intertidal fishes may share habitat in the rocky pools for long periods of the day and while there is some degree of separation through intertidal zonation (Kelly and Palumbi 2010; von der Heyden et al. 2013), it is common to find multiple species in a single small tidepool. In addition to morphological traits that enable adaptation to intertidal environments, South African clinids share limited dispersal potential as all species are live-bearing, releasing post-flexion larvae throughout the year, but likely with some peak spawning times (Veith 1979; Moser 2007). This unique set up thus creates opportunities for limiting gene flow and, for several species, populations show varying degrees of population level genetic structuring across southern Africa (von der Heyden et al. 2011; Wright et al. 2015).

Phylogenetic techniques, based on molecular markers, are valuable tools for delineating cryptic species (Bickford et al. 2007). A variety of molecular techniques and markers have been used to identify cryptic species including mitochondrial (e.g., COI, Dloop, ND2, 16S) and nuclear markers (e.g., rhodopsin, GPI, EF-1α, microsatellites). More recently, next generation sequencing techniques have added single nucleotide polymorphism markers (SNPs) (Morin et al. 2004; Miller et al.

2007; Leaché et al. 2014) and while these studies remain comparatively rare, SNPs have proven to be a powerful marker for phylogenetic inference (Edwards et al. 2015; Andrews et al. 2016) and the detection of cryptic species and have resolved phylogenies for complex study systems (Wagner et al. 2013; Georges et al. 2018). Genomic sequencing methods allow for de novo assemblies of sequence catalogs (stacks in the case of RADseq) but are most effective when aligned to an assembled genome of the target species or a closely related congener (Catchen et al. 2013; Rochette et al. 2019). Aligning to a reference genome allows putative loci to be visualized positionally rather than being examined independently as they would be in *de novo* analyses, which cuts down on the number of false positives and increases the chance of detecting true outliers (Rochette and Catchen 2017). RADseq samples the genome more densely than first generation techniques and allows investigators to observe patterns of genetic variation resulting from the full range of evolutionary processes acting across the breadth of the genome including drift, selection, recombination, and mutation (Narum et al. 2013). In this study, we use RAD sequencing coupled with a newly assembled genome to investigate cryptic speciation in the klipfishes of South Africa's rocky intertidal.

The Super klipfish, *Clinus superciliosus* and Nosestripe klipfish, *Muraenoclinus dorsalis* were first identified as candidate cryptic species in von der Heyden et al. (2011), with Holleman et al. (2012) further delineating species within the '*Clinus superciliosus* species complex' using mtDNA markers and the nuclear marker Rhodopsin. We expand this investigation by applying genomic methods to a

combination of samples from the 2011 study and newly collected specimens. We first present an assembled genome for *Clinus superciliosus*, the first assembly for a South African marine fish, which represents a valuable resource as research into this system transitions from classic genetic methods to genomic techniques. We generate RAD sequence data for samples for *C. superciliosus* and *M. dorsalis* specimens aligning raw reads to the *C. superciliosus* genome to test if phylogenetic patterns from genome-wide SNPs are concordant with those using traditional markers and attempt to identify specimens as part of the *Clinus superciliosus* species complex described in Holleman et al. (2012).

Materials & Methods

Sample collection and DNA extraction

Samples of *Clinus superciliosus* and *Muraenoclinus dorsalis* were collected across three separate sampling efforts in 2009, 2013, and 2019. Sample locations are presented in Figure 2.1. Fin clips and muscle tissue were preserved in 95% ethanol and stored at -20 C. Samples were collected under the following permit: RES2021/68. DNA was extracted using DNeasy Blood and Tissue kits (Qiagen) or chloroform/isopropanol extraction protocol and assessed DNA concentrations using a Qubit 4.0 fluorometer.

Whole-genome library construction and sequencing

For long-read sequencing we sheared the DNA for individual CSU_WPO_1119_01 to 10 Kb using the Covaris g-TUBE following the manufacturer's protocol. We prepared two separate Oxford Nanopore Technologies

(ONT) libraries with 1.5 ug of DNA using the SQK-LSK109 library preparation protocol (Oxford Nanopore Technologies, Oxford, UK). We sequenced each library on a R9.4 flow cell using the MinION DNA sequencer.

To obtain high-accuracy short-read sequences for consensus polishing, we sent a 250 ng aliquot of DNA from the same individual to Novogene Inc. for Illumina whole-genome library preparation and sequencing. The libraries were prepared using the NEBNext Ultra II DNA Library Preparation Kit following the manufacturer's protocol. Sequencing was performed on the NovaSeq 6000 PE 150. We targeted 350 bp raw read length and 50x genome coverage.

Genome assembly

We used Guppy v5.0.15 to base call the raw data from the output of both MinION flow cells. We then concatenated long-read files into one large fastq file which we trimmed with Porechop v. 0.2.3 (https://github.com/rrwick/Porechop). We used Nanofilt v. 2.5.0 (https://github.com/wdecoster/nanofilt) to generate 2 filtered datasets to help the contiguity of the final assembly. The first filtered dataset kept the longest reads for initial assembly (Nanofilt parameter -q 5) and the second targeted shorter, higher quality reads for downstream polishing (-q 7). We used Wtdbg2 v2.5 (Ruan and Li 2020) on the first filtered dataset for the initial assembly, then performed two rounds of consensus correction using the q 7 dataset by mapping reads to the draft genome with Minimap2 v. 2.17 and polishing with Racon v. 1.4.7.

We used the shorter (150 paired end), but more accurate Illumina reads to further polish the draft assembly. We adapter-trimmed the raw sequences with

Trimmomatic v. 0.39 (Bolger et al. 2014) and quality checked before and after trimming using FastQC v 0.11.8 (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). We performed two rounds of polishing by mapping the trimmed reads to the draft assembly using BWA v 0.7.17 (Li 2013), sorted and indexed with Samtools v 1.9 (Li et al. 2009), and consensus corrected using Pilon v 1.23 (Walker et al. 2014). We checked for contamination using BlobToolKit framework and removed contaminated contigs (Challis et al. 2020). We assessed assembly completeness using Benchmarking Universal Single-Copy Orthologs (BUSCO v3.0.2) (Simão et al. 2015; Waterhouse et al. 2018).

Nuclear and mitochondrial sequencing

We constructed restriction site associated DNA (RAD) libraries for all samples and amplified the 16S mitochondrial rRNA region for a subset of specimens. We constructed the RAD libraries using a variation of the original protocol (Miller et al. 2007; Baird et al. 2008) using restriction enzyme SBFI and NEBNext reagents (New England Biolabs) (Miller et al. 2012). We started the library preparation with 100 ng of genomic DNA (gDNA). We multiplexed and sheared libraries to roughly 500 bp lengths on a Biorupter Sonicator using four cycles of 30 seconds. We used Dynabeads (Invitrogen 11206D) to remove non-tagged DNA and used SPRI beads (DeAngelis et al. 1995) for purification and size selection. We carried out the final polymerase chain reaction (PCR) amplification on 16 ul reaction volumes for 10 amplification cycles (reference). Samples used in the study were

sequenced in one of five libraries, each containing 96 individually barcoded samples. Libraries were either sequenced on an Illumina HiSeq 4000 or NovaSeq 6000 S4 at the Vincent J. Coates Genomics Sequencing Laboratory at UC Berkeley supported by NIH S10 OD018174 Instrumentation Grant, or at Novogene Corporation, Sacramento, CA.

To investigate phylogenetic relationships at the level of the mitochondria for *Clinus superciliosis*, we amplified and sequenced the 16S mitochondrial region for a subset of specimens from each of three phylogenetic clades for comparison with 16S sequences from Holleman et al. 2012. We amplified the 16S rRNA segments using the universal primers 16SAR-16SBR (Palumbi et al. 1991) with 35 cycles at a denaturation temperature of 94°C for 30 s, an annealing temperature of 52°C, and an extension of 30 s at 72°C. We purified the PCR product following the manufacturer's protocol and had them sequenced in both directions with the primers used in PCR amplification and at the UC Berkeley DNA Sequencing Facility for Sanger sequencing.

Quality filtering and marker discovery

For all specimens included in the study we trimmed, filtered, and genotyped raw reads using the software program Stacks version 2.2 (Rochette et al. 2019). We split raw reads according to their 6 bp unique barcode, removed the barcodes and trimmed reads on the 3' end to a final sequence length of 91 bp using the program *process_radtags*. Sequences were removed if the quality scores fell below 90%, a raw phred score of 10. We aligned the filtered sequences to the assembled *C.*

superciliosus genome using Bowtie 2 (Langmead and Salzberg 2012) with the preset option of --very-sensitive (-D 20 -R 3 -N 0 -L 20 -i S,1,0.50) to generate BAM files which we sorted and converted to SAM files with SAMtools (Li et al. 2009). We then used the Stacks 2.0 shell program *ref_map.pl* to call single nucleotide polymorphisms (SNPs) and generate population-level summary statistics. We ran the Stacks program *populations* to generate output files for downstream phylogenetic analysis with every individual being treated as a population for phylogenetic inference. We retained only a single SNP per locus (write-single_snp option in Stacks).

We aligned the 16S rRNA sequence reads generated for this study with Holleman et al (2012) sequence reads downloaded from Genbank in the program MAFFT implemented in Geneious v. 2020.2.4 (Kearse et al. 2012) and converted the aligned reads into Nexus format for downstream analysis.

Phylogenetic inference

For each genus we exported phylip format files containing nucleotides that are fixed within groups but variable between groups (in this case individuals) using the *populations* package in Stacks. We treated each individual as its own population and only retained SNPs that were present in ≥80% of individuals. We converted the resulting phylip files into Nexus format and included additional clinid species commonly found in the same tidepools as *C. superciliosus* and *M. dorsalis* as outgroups (*Clinus agilis, Clinus cottoides*). We filtered the list of nucleotides to the informative loci and performed maximum-likelihood (ML) implemented in Phyml

(http://www.atgc-montpellier.fr/phyml/) and maximum-parsimony (MP), and neighbor-joining (NJ) phylogenetic inference as implemented in Paup* (https://paup.phylosolutions.com/). We generated support values for nodes using 1000 non-parametric bootstrap replicates.

For the 16S alignment we used Neighbor-Joining and Maximum-Likelihood methods with bootstrapping support as described above. Phylogenetic trees and support values were visualized using FigTree v1.4.4. (https://github.com/rambaut/figtree/releases).

Results

Genome assembly

MinION sequencing generated approximately 20 Gb of long reads (approx. 33X coverage) while Illumina short-read sequencing produced 30 Gb (approx. 50X coverage) paired end reads. The final assembled genome of *Clinus superciliosus* yielded a total size of ~590 Mb gathered in 6,696 contigs with the largest contig at 982 Kb and an N50 of 175 Kb (Table 2.2). The benchmark universal single-copy orthologs (BUSCO) completeness for the final assembly was 94% in the Actinopterygii dataset (3,640) improved from an initial completeness of 87.4% with ONT and Illumina polishing (Figure 2.3).

Nuclear and mitochondrial sequencing results

Sequencing generated > 50 Mb of filtered and cleaned data for all specimens. Alignment to the *C. superciliosus* genome resulted in an average alignment of 97% (ranging from 96.1 to 98) for C. superciliosus and 90% for (ranging

from 88.2 to 91.4) *M. dorsalis*. RAD libraries for *C. superciliosus* and *M. dorsalis* resulted in mean coverage of approximately 32X and 13X respectively with a final catalog containing 152,150 loci (CSU) and 90,462 loci (MDO). In addition to the ingroups, specimen of *C. cottoides* and *C. agilis* were included as outgroups. The *populations* script identified 55,822 SNPs for *C. superciliosus* and 26,143 SNPs for *M. dorsalis*, which is also the number of loci as only one SNP was kept for each locus.

For the mitochondrial 16S region we generated an alignment of 559 bp for seven *C. superciliosus* specimens which we added to aligned sequence reads from Holleman (2012).

Phylogenies

Clinus superciliosus

We were able to obtain a fully resolved phylogeny using RAD markers (Figure 2.4). Our phylogeny split ingroup specimens into three sister clades to the outgroup with 100% support for each node regardless of methodology (ML, NJ, MP). The inferred tree was consistent with a published mitochondrial phylogeny for *C. superciliosus* from von der Heyden et al. (2011) with the majority of specimens falling into a large cluster (Clade 1), with a smaller cluster (Clade 2) and in this study, only three individuals in the third cluster (Clade 3).

The RAD phylogeny was also consistent with the phylogeny using the 16S rRNA marker (Figure 2.5). This phylogeny was constructed with samples from this study (seven individuals) and from Holleman (2012) (12 individuals). The Holleman

(2012) phylogeny included samples from two previously recognized species (*Clinus superciliosus* and *Clinus arborescens*) and two species described for the first time in their publication (*Clinus exasperatus* and *Clinus musaicus*). Specimens from this study, when added to the phylogeny, grouped with three of the four clusters from Holleman (2012). Clade one in the SNP phylogeny fell into the *Clinus superciliosus* species group, clade two (SNP phylogeny) individuals grouped with *Clinus exasperatus* and specimens from the SNP phylogeny clade three clustered with *Clinus arborescens* (Figure 2.5).

Muraenoclinus dorsalis

Similar to the *C. superciliosus* result, we were able to construct a fully resolved phylogeny for *M. dosalis* using the RAD markers (Figure 2.6). This phylogeny separated ingroup individuals into two sister clades to the outgroups with 100% support for nodes regardless of methodology. The inferred tree was consistent with a published mitochondrial tree from von der Heyden et al. (2011).

Discussion

The study presents the first genome assembly for a South African marine fish *Clinus superciliosus* and uses a SNP phylogeny to resolve instances of cryptic speciation in two clinid species. As such, our study provides novel biological information and contributes to growing the field of ecological genomics in the Southern Africa.

Genome assembly

The genome assembly of *Clinus superciliosus* is the first published genome of

a fish in the family Clinidae and represents a valuable resource in the continued investigation of this fascinating study system. While the assembly is fragmented (~6 K contigs), it has zero gaps and its completeness (94% BUSCO) is comparable with other high quality fish assemblies (Sayers et al. 2022). It has proven a valuable tool for genomic analysis for South African clinids as we were able to align sequence reads of several species (*Clinus agilis*, *Clinus cottoides*, *Muraenoclinus dorsalis*) to the assembled genome with an alignment rate > 85%. Studies have shown that alignment of RAD reads to a reference genome increases the number of recovered polymorphic loci when compared to de novo pipelines (Kunvar et al. 2021). This is particularly important when investigating closely related species and lays the groundwork for future investigation of population structure and the genomic basis for local adaptation (Nielsen et al. 2020).

Phylogenetic pattern of cryptic speciation in C. superciliosus

The phylogeny inferred using RAD SNPs for *C. superciliosus* shows concordance with the mtDNA phylogeny from von der Heyden et al. (2011) and Holleman et al. (2012). Regardless of methodology, support values unambiguously identified three clades in the *C. superciliosus* complex; clade one was the by far the largest (131 total individuals, Figure 2.7), suggesting that the species represented by clades two and three were comparatively rare across the intertidal zones sampled for this study. Concordance in phylogenetic patterns between mtDNA and nuclear markers is expected (Avise 2004), though examples of discordance are not uncommon (Toews and Brelsford 2012). Incomplete lineage sorting, sex-biased

dispersal, selection, and introgression have all been implicated in mito-nuclear discordance (Godinho et al. 2008; Rašić et al. 2015) and there are examples of discordance in marine fishes (Larmuseau et al. 2010), cryptic marine species in South Africa (Amor et al. 2019), and in fishes in the family Clinidae (kelpfishes, unpublished data). However, the concordance shown here indicates that the divergence between mtDNA clades correspond to difference in the nuclear genome resulting in three phylogenetic species with fully sorted lineages.

The 16S phylogeny inferred by aligning samples from this study to those from Holleman et al. (2012) identified each of the three clades from the SNP phylogeny to the species level. We were able to identify clade one individuals as *Clinus superciliosus*, clade two as *Clinus exasperatus* and clade three as *Clinus arborescens*. The species level identification may in part explain the relative abundance of samples collected in each clade. *C. superciliosus* is among the most common fishes found in South Africa's rocky shore and that clade encompasses the majority of samples collected in the regions of the intertidal sampled for this study (Prochazka and Griffiths 1992). *Clinus exasperatus* was named using a latinized form of exasperate because of how difficult it is to obtain specimens (see description in (Holleman et al. 2012) however the exact range and habitat preference is unknown so while it is not well represented in our tide pool samples, *C. exasperatus* may be more abundant in other locations. *Clinus arborescens* is a predominantly subtidal species that uses the lower intertidal as a nursery for juveniles and is much more commonly found in the subtidal as adults and therefore not as common as *C.*

superciliosus in tide pools (Holleman et al. 2012).

The *Clinus superciliosus* complex is an excellent example of species that are cryptic in a particular life stage, in this case, as juveniles when specimens can be just a few centimeters long (standard length, SL). *Clinus exasperatus* is a recently described species (Holleman et al. 2012) and it is currently known from two specimens, one of which is an adult. It has unique characteristics that enable identification as adults, but they are primarily based on body depth and coloration, both of which can be difficult to distinguish in the field. The ratio of body depth to SL appears to fall outside the range seen in *C. superciliosus*, but with only one specimen described the statistical range of values for body depth to SL is presently unknown. There is also currently no information on how or if the ratio of body depth to SL changes between life stages. Coloration is a particularly challenging character for identification as it can be highly variable in clinids generally (Stepien 1987; Stepien et al. 1988), and coloration is impacted by preservation (i.e. ethanol, commonly used to preserve tissue for molecular analysis), so potentially distinguishing character can be lost before a detailed morphology can be determined.

The identification of *C. arborescens* is comparatively straight forward in adults, in large part because of their life history as a predominantly subtidal species, making the location of their capture a good indicator for species identification. However, as juveniles, when they are most likely to be captured as part of a sampling effort in the intertidal, the distinguishing characteristics are more difficult to tease apart. *C. arborescens* has a mean of 6 dorsal fin rays compared to

the average of 8 for *C. superciliosus*, but the ranges for the two overlap (5-7 for *C. arborescens*, 6-9 for *C superciliosus*) (Holleman et al. 2012). An additional confounding factor is that the number of dorsal fin rays for *C. superciliosus* varies with water temperature with specimens from the warmer east coast of South Africa with a mode of seven rays (Holleman et al. 2012). Other identifying characteristics include differences in the small orbital cirrus and the relative size of the first dorsal spine to the subsequent two. Both characteristics, while potentially evident in adult specimens, are difficult to distinguish between juveniles, especially in the field, and require that whole specimens be collected and carefully preserved for examination under magnification.

The *Clinus superciliosus* species complex exemplifies a system where the environmental pressures experienced by all species may have constrained the body forms that can survive in the rocky intertidal (Buser et al. 2017; Egan et al. 2021). It stands as a warning that commonly cited characters such as coloration and meristic and morphometric measures can overlap and are at best difficult if not impossible to distinguish reliably in a field setting. Even with experienced investigators, the combination of characters required to correctly identify species at different life stages can be unreliable without assistance from molecular techniques.

Cryptic speciation in Muraenoclinus dorsalis

M. dorsalis represents a contrasting version of cryptic speciation from *C. superciliosus*. The phylogeny inferred using SNPs, as with *C. superciliosus* corresponds with mtDNA markers for a completely resolved tree with two clades as

divergent as those from *C. superciliosus*. We again see an asymmetry in the number of specimens collected, with most specimens encompassed by clade one (Figure 2.8). However, unlike *C. superciliosus*, there is no described species that is morphologically similar to *M. dorsalis*. *M. dorsalis* has a unique position as the only eel-like body form in South Africa's intertidal (Branch et al. 2016). They inhabit the highest intertidal zone and are commonly found outside of tidal pools between tides surviving under rocks on moist sand (von der Heyden et al. 2013). Both in habitat and morphological character, they are easily distinguished from all other described intertidal species.

For this reason, *M. dorsalis* appears to adhere to the most rigorous definition of a cryptic species (Unmack et al. 2021). Specimens from each genetic clade occur sympatrically in their location (Gaansbai, Moullie Point) and in the intertidal zone where they were sampled (high intertidal). There is no known morphological characteristic or life history trait that separates them yet there are fixed alleles between the populations. *M. dorsalis* is the only species in its genus, so there are no congeners to confuse them with and the two clades are consistently sister clades when outgroups from the family clinidae are included in the analysis (*Clinus cottoides, Clinus superciliosus, Clinus agilis*).

Attempts have been made to collect additional specimens from clade two including efforts for this study, but they were ultimately unsuccessful, making the lineage represented in clade two known only from tissue samples collected in 2009. The subsequent sampling efforts were extensive enough to suggest members of

clade two are relatively rare. The reasons for this are still a mystery, but it is possible that there is a temporal component to their abundance. While there is nothing about time itself that would account for this result, it may be that environmental conditions will favor one lineage over another, and those conditions may cycle through time. There have been changes in Southwest summer winds along the west coast, driving an increase in the instances of cold-water upwelling (Rouault et al. 2010) and changes in the warm water Aguhlas current along the southern coast (Rouault et al. 2009) in recent decades. However, given that samples from clade two were found on the west coast (Seapoint) and southern coast (Gansbaai), there is no known single environmental variable that explains the pattern. A longitudinal seascape ecology analysis that explicitly tests for correlation between genetic and environmental variation is out of scope for the specimens included in this study but may be possible in the future.

An alternative hypothesis would be that there is hybridization and that members of clade one and hybrid offspring have increased fitness, leading to the gradual extinction of clade two. Presently there is no evidence for hybrids, with all samples included in this study cleanly separating into distinct genetic clusters in primary component analysis and STRUCTURE analysis (data not shown here), but we cannot rule out the possibility that hybrids are also relatively rare and simply have not been sampled to date. Further sampling and analysis are required to test these hypotheses.

Cryptic speciation in the sea

The klipfishes of South Africa share traits that fit well into the paradigm of speciation in marine systems laid out by Palumbi (1994). They are viviparous giving birth to post flexion larvae limiting their potential to disperse as larvae (Moser 2007). They inhabit a dynamic coastal marine system with strong currents and biogeographic breaks (von der Heyden et al. 2013) further suppressing dispersal potential and creating "invisible" barriers to gene flow and thereby the potential for populations to become reproductively isolated. South Africa's submerged coastline may also play a role in isolating populations as models have suggested that a shallow continental shelf (Agulhas Bank) off the south coast created a vicariant barrier between east and west coast populations of rocky shore inhabitants during the last glacial maxima (Toms et al. 2014). While the exact mechanism leading to species level divergence for *C. superciliosus* and *M. dorsalis* remains unknown, the combination of life history, habitat, and geographic history are consistent with mechanisms of marine speciation.

Our study adds to a growing list of publications on cryptic speciation in marine systems (Williams et al. 2001; Lima et al. 2005; Teske et al. 2007; Warner et al. 2015; Hirase et al. 2021). Traits shared among klipfishes may explain the cryptic nature of speciation observed here. *C. superciliosus* and *M. dorsalis* live in an environment that is constantly experiencing heavy wave action from the incoming or outgoing tide. The environment of the intertidal zone places strong and consistent adaptive demands on their inhabitants (Horn and Gibson 1988). Many of their adaptations require convergent morphologies such as loss of air bladders for a benthic lifestyle, elongate bodies for hiding from predators under rocks and in small crevices,

and cryptic coloration for camouflage. The environment of the intertidal may apply selective pressure preventing conspicuous morphological changes (body shape, coloration) even as populations diverge genetically. While additional work explicitly testing this hypothesis are warranted *C. superciliosus* and *M. dorsalis* support prior work suggesting small bodied shallow marine species are prone to cryptic speciation (von der Heyden 2011) and highlights the importance of testing for cryptic speciation when working in marine systems.

Conclusion

Here we've published a novel genome assembly adding an important resource to assist in transitioning research on South Africa's clinids into the genomics era. We've added examples of cryptic species in a shallow marine environment that favors convergent body forms and colorations fitting a broader pattern of environmental conditions that may facilitate cryptic speciation. We have also demonstrated that, while genome-wide nuclear markers (SNPs) are a powerful tool for phylogenetic inference, comparatively inexpensive mtDNA markers are likely sufficient for identifying cryptic species. While molecular techniques are out of scope for many studies and biodiversity assessments, we recommend that investigators that work in potentially cryptic systems incorporate molecular techniques, when possible, to explicitly test for cryptic speciation. Genome-wide techniques are important for testing phylogenetic hypotheses across the full range of evolutionary processes impacting the genome, but our study supports the use of less resource intensive barcoding methods to test for the presence of cryptic species in a study system.

Tables and Figures

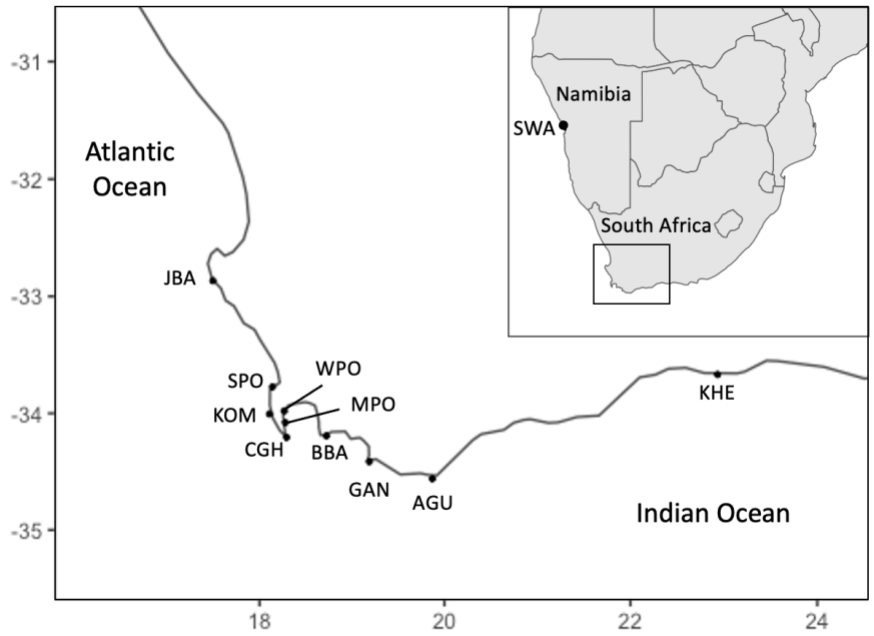


Figure 2.1. Map of sampling locations. On the main map, from west to east, abbreviations are Jacobsbaai (JBA), Sea Point (SPO), Kommetjie (KOM), Cape of Good Hope (CGH), Miller's Point (MPO), Woolley's Pool (WPO), Betty's Bay (BBA), Gansbaai (GAN), Cape Agulhas (AGU), Knysna Heads (KHE). In the inset map, the abbreviation SWA is Swakopmund where samples of *C. superciliosus* were collected.

Table 2.2. Summary statistics for the genome assembly of the Superklipfish, *Clinus superciliosus*. Top table shows statistics on the length of the genome and contigs, bottom table shows statistics on completeness of the assembly using benchmark universal single-copy orthologs (BUSCO).

	Length	N Contigs	Avg. Len.	Largest
TOTAL	590,622,898	6,703	88,113	982,327
N50	175,421	995		
N60	138,639	1,373		
N70	105,366	1,866		
N80	74,003	2,533		
N90	43,427	3,559		
N100	1,696	6,703		
Gaps	0			

	Complete	Single match	Double Match	Fragmented	Missing
BUSCO	94.00%	93.30%	0.70%	1.90%	4.10%
Complete	3,422				
Single	3,397				
Double	25				
Fragmented	70				
Missing	148				
Total	3,640				

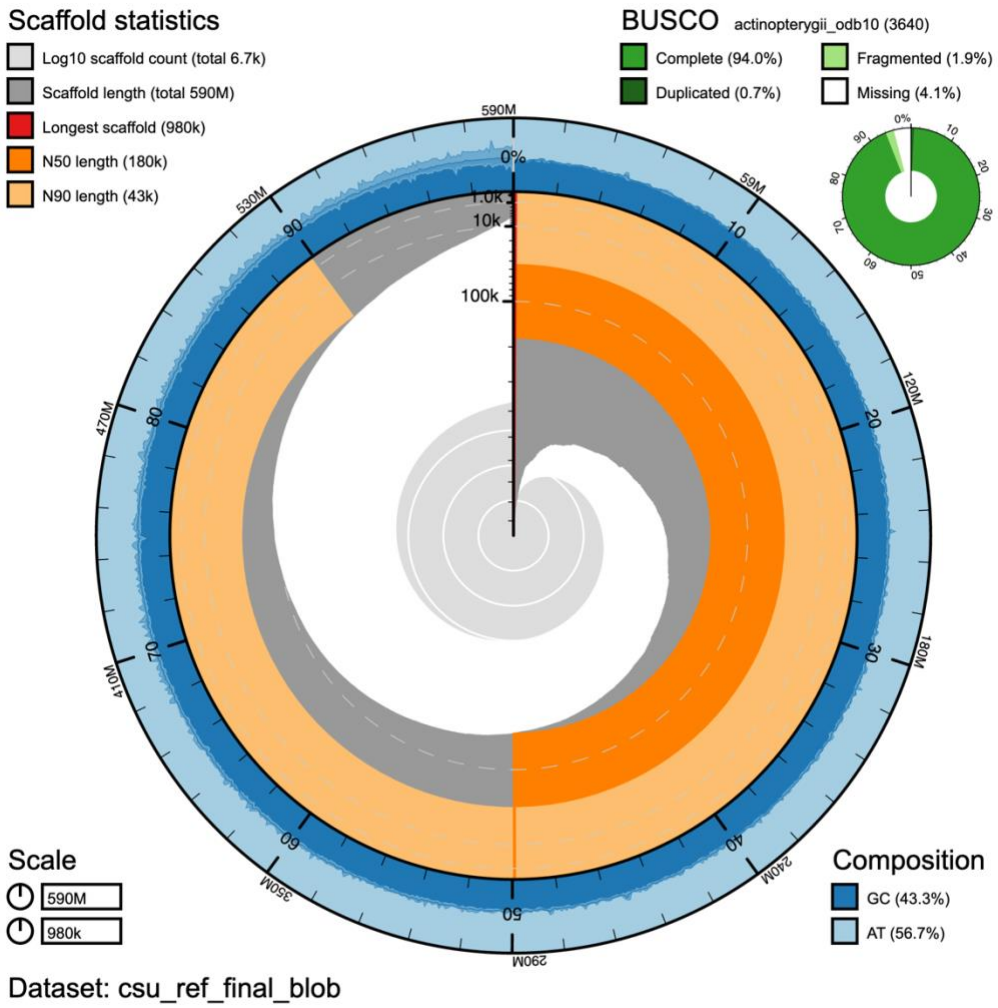


Figure 2.3. BlobToolKit Snail plot showing a graphical representation of the quality metrics presented in Table 1 for *Clinus superciliosus* genome assembly. The plot circle represents the length of the genome (~590 bp). The inside circle of the plot shows the length-related metrics with a red line showing the length of the longest contig. All other contigs (in gray) are arranged in order of size clockwise around the plot. The N50 value is represented in dark orange and the N90 in light orange. The dark vs. light blue area around the outside of the plot shows mean, maximum and minimum GC vs. AT content at 0.1% intervals. BUSCO scores from table 1 are shown in the upper right corner (Challis et al. 2020).

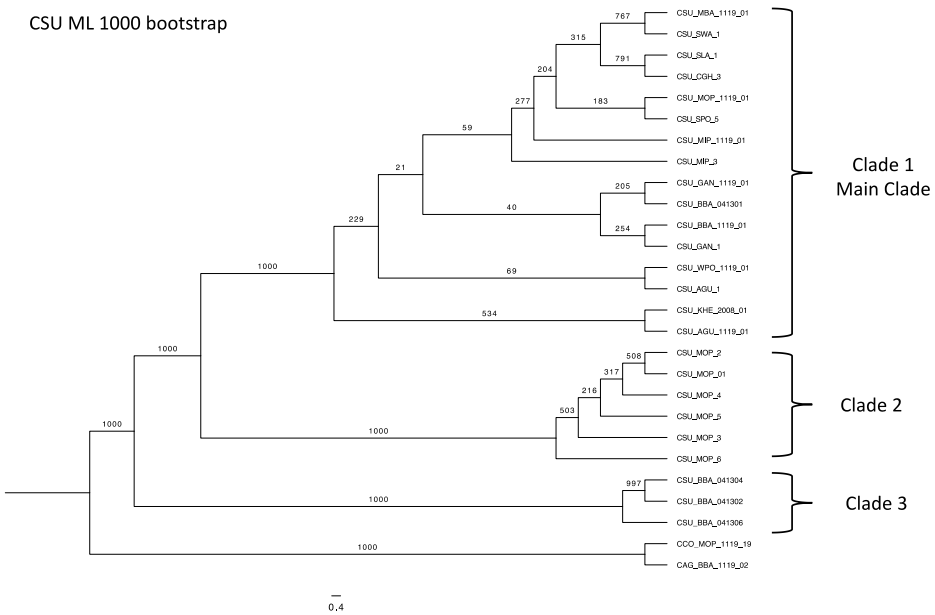


Figure 2.4. Phylogenetic tree for *C. superciliosus* complex inferred from 55,822 SNPs. Maximum likelihood support values are shown for each node. *Clinus cottoides* is included as an outgroup.

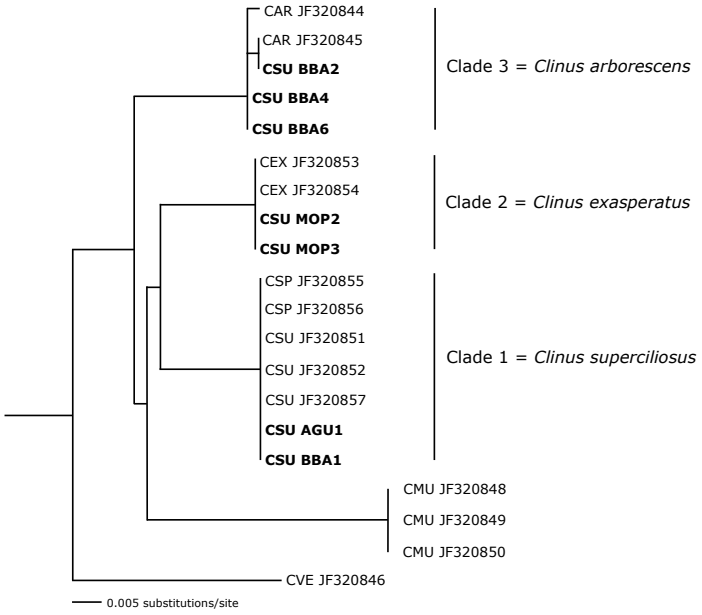


Figure 2.5. Phylogenetic tree for *C. superciliosus* complex using 16S marker. Data are combined with data from Holleman et al. 2012. Samples from this study are in bold.

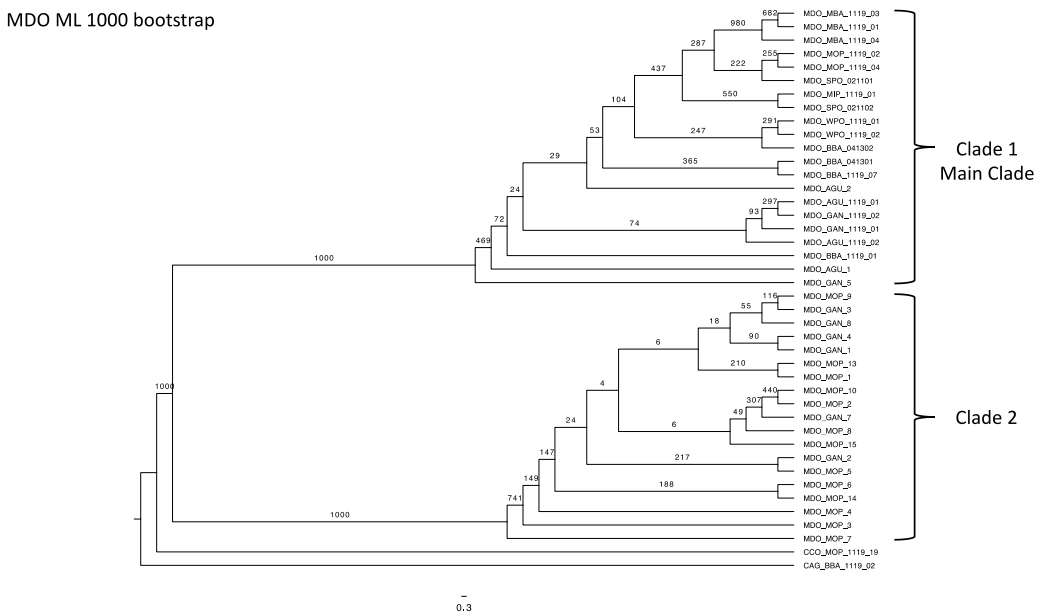


Figure 2.6. Phylogenetic tree for *M. dorsalis* complex inferred from 26,143 SNPs. Maximum likelihood support values are shown for each node. *Clinus cottoides* and *Clinus agilis* are included as outgroups.

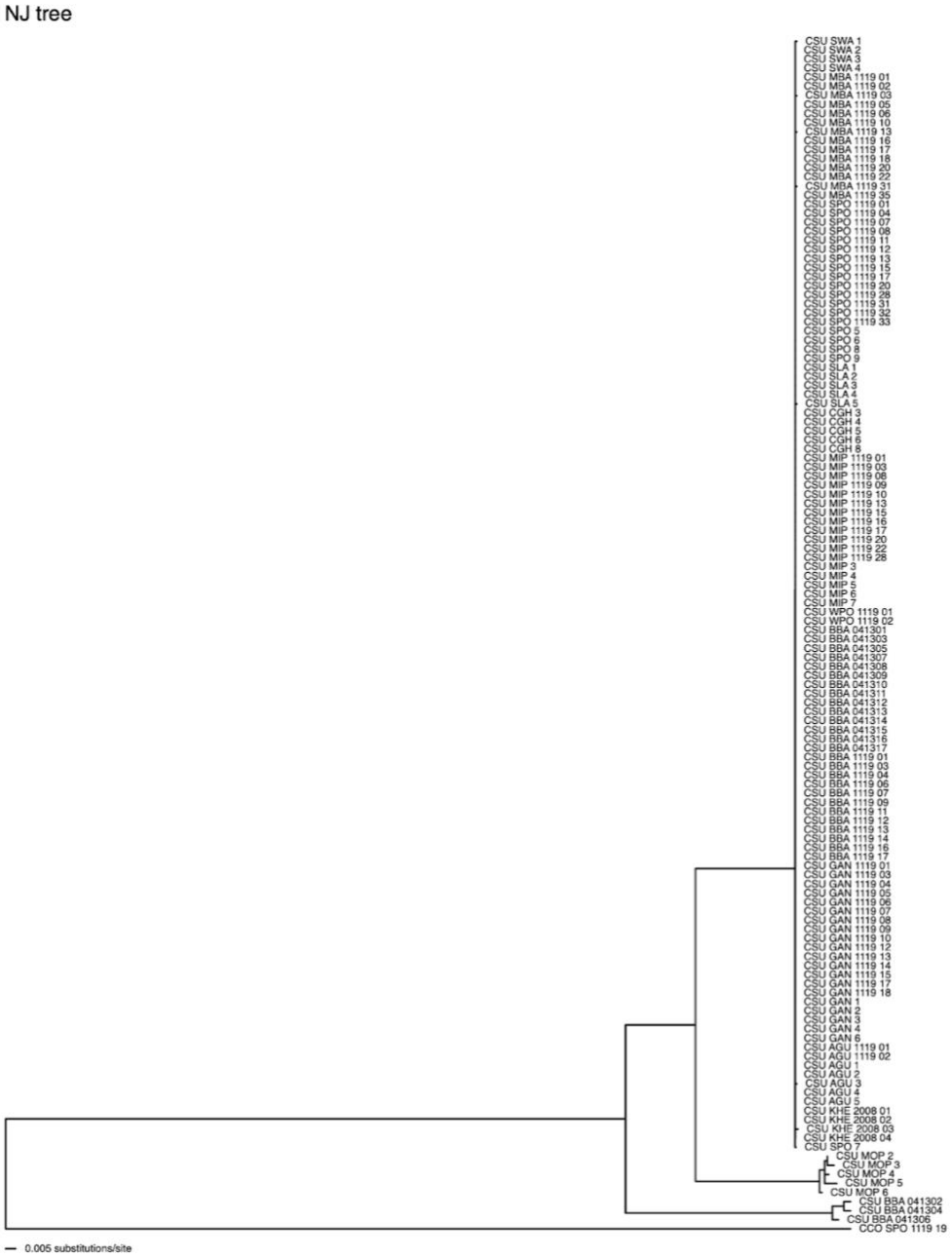


Figure 2.7. Neighbor joining tree generated with 133 specimens of *C. superciliosus* and outgroup specimen *Clinus cottoides*. Clades from top to bottom are Clade 1 (124 specimens), Clade 2 (5 specimens, and Clade 3 (4 specimens).

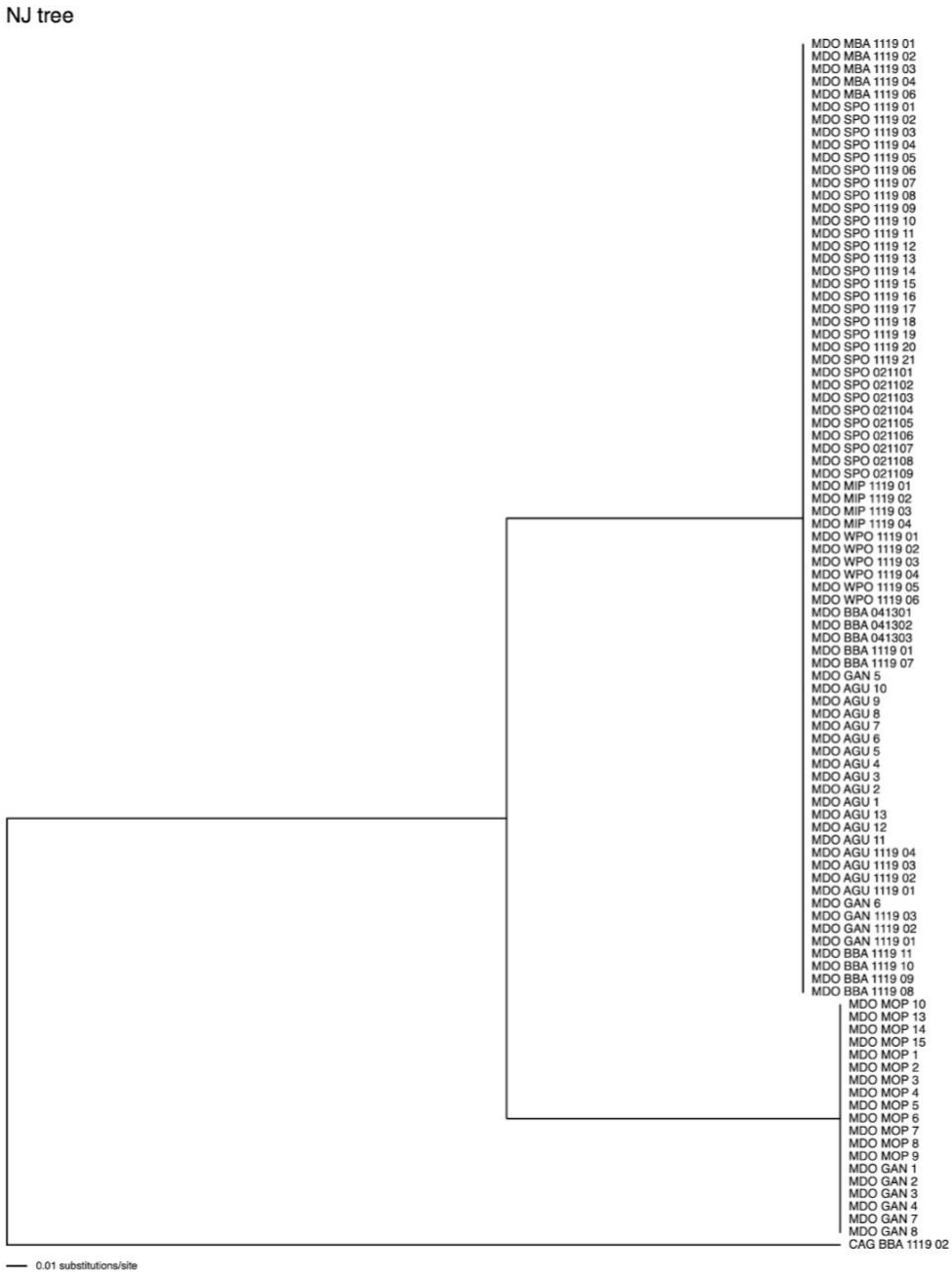


Figure 2.8. Neighbor joining tree generated with 97 specimens of *M. dorsalis* and outgroup specimen *Clinus agilis*. Clades from top to bottom, clade 1 (78 specimens), clade 2 (19 specimens).

Chapter 3: Population genomics of South African klipfishes *Clinus superciliosus, Clinus cottoides*, and *Muraenoclinus dorsalis* (Family: Clinidae)

Abstract

South Africa's southwest coast is a dynamic marine ecosystem where the Indian and Atlantic oceans meet. The interaction of contemporary environmental pressures and geological history have made this region an excellent system for investigating the population dynamics and biogeography of coastal marine species. In this study we apply next generation sequencing (NGS) techniques to explore the population structure of three species of live bearing intertidal klipfishes; *Muraenoclinus dorsalis*, *Clinus cottoides*, and *Clinus superciliosus*. Restriction site associated sequencing (RADseq) reveals a strong isolation by distance signal in all three species and supports the hypothesis of a secondary contact zone between lineages that were separated by the continental shelf along the southern coast during the last glacial maxima. Further, the high-resolution analysis of population structure supports the classification of Cape Agulhas as an important biogeographic break, which was inconsistently detected in previous studies of intertidal fishes. This study demonstrates the potential for investigating fine scale structure between coastal populations and highlights the importance of a network approach to marine protection to preserve the genetic diversity present in South Africa's coastal marine ecosystem.

Introduction

Molecular techniques have advanced at an accelerated rate over the last 15 years (Slatko et al. 2018). Since 2007 the sequencing cost per mega base has decreased at a rate faster than that predicted by Moore's law (genome.gov). Next generation sequencing (NGS), driven by massively parallel sequencing technologies and a corresponding increased availability of supercomputing clusters and cloud-based bioinformatic tools have made genome-wide sequencing affordable to many investigators (Ellegren 2014). Short-read reduced representation libraries like restriction site associated DNA (RAD) and amplified fragment length polymorphisms (AFLPs) are capable of producing tens of thousands of single nucleotide polymorphism (SNP) markers and require funding comparable to traditional single-marker studies (Narum et al. 2013). Long-read sequencing machines small enough to carry into the field (ONT Nanopore) that are sufficiently accurate for de novo genome assembly (Lu et al. 2016) are now available to an increasing number of investigators whereas de novo genome assembly was a process previously reserved for well-funded labs working on model organisms (Michael et al. 2018; Choi et al. 2020). The result is a growing number of studies that incorporate NGS methods to gain unparalleled resolution into population structure and genetic connectivity in non-model systems. While SNP markers have been used to investigate previously unexplored study systems, there is also an opportunity to revisit complex study systems such as the dynamic marine ecosystem of the Western Cape of South Africa and test if analysis with SNPs corroborates that of previous studies and if there are new insights to be gained.

South Africa's Western Cape sits at the confluence of the Atlantic and Indian oceans and the region's marine ecosystem is driven by a complex mix of oceanographic features. The width of the continental shelf varies along the coast and is narrow on the east coast, moderate along the west coast and extends nearly 200 km offshore the southern coast in a region known as the Agulhas Bank (Griffiths et al. 2010) (Figure 3.1). The shape of the continental shelf is an important driver of coastal sea temperature as it guides the region's major sea currents. The coastal region of the Western Cape is dominated by two currents, the Agulhas current which runs along the narrow continental shelf of the east coast carrying warm waters from the Indian ocean before pushing offshore along the Agulhas bank and the Benguela current which pushes cold, nutrient rich water north along South Africa's west coast (Griffiths et al. 2010). The two currents converge between Cape Point and Cape Agulhas mixing and producing an area of anticyclonic eddies known as the Agulhas rings which transports Indian Ocean water into the Benguela current (Griffiths et al. 2010). The Agulhas and Benguela currents produce a change in sea temperature along South Africa's coast splitting the coastal region into 4 biogeographic provinces (Teske et al. 2011). The tropical and subtropical provinces occur along the east coast, and the warm temperate and cool temperate provinces along the south and west coast, with a region of mixing that spatially coincides with the Agulhas Rings. In addition to the influence of the currents on sea temperature, strong seasonal winds can produce areas of cold-water upwelling, particularly along the western and southwestern coastline, and have been known to rapidly change sea surface temperatures by as

much as 6 degrees Celsius over a distance of a few kilometers (Goschen and Schumann 1995). The combination of currents, upwelling events, and oceanographic features makes the coast of the Western Cape one of the most dynamic marine ecosystems in the world. These features place a high demand on the marine flora and fauna and influence the biogeography of the region (Griffiths et al. 2010). The impact of these features on population structure and biogeography have been studied using a number of taxa (Neethling et al. 2008; Wright et al. 2015; Nielsen et al. 2018; Dalongeville et al. 2022) but are perhaps no more apparent than in the clinid fishes (klipfishes) of the rocky intertidal.

Clinids are the most abundant fishes found in tide pools around the western cape comprising, in some regions, 88% to 98% of the intertidal fish community. Among the clinids, *Muraenoclinus dorsalis* (Bleeker, 1860), *Clinus cottoides* (Valenciennes, 1836), and *Clinus superciliosus* (Linnaeus, 1758), are the most numerically abundant (Prochazka and Griffiths 1992). While they occur sympatrically and share common traits (live bearing, cryptic coloration, demersal) they are not identical in their adaptations to the intertidal. The eel-like intertidal resident *M. dorsalis* lives in the high shore (littorina and upper balanoid) intertidal zone and is capable of spending time between tides sheltering on moist sand out of tidal pools under rocks or kelp. *C. cottoides* has a cylindrical tapering body form and is most commonly found in the mid-shore zone (upper balanoid to lower balanoid) (Bennett and Griffiths 1984). *C. superciliosus*, the most abundant of the three, is bilaterally compressed and while juveniles can be found in all zones, adults

predominantly inhabit the lower intertidal and shallow subtidal zones (Prochazka and Griffiths 1992). The sympatric distribution and mix of commonalities and interspecific adaptations have made the rocky shore clinids a productive system for molecular ecology. Studies have previously addressed demography (von der Heyden et al. 2015), speciation (von der Heyden et al. 2011; Holleman et al. 2012), biogeography (von der Heyden et al. 2013), and population structure (von der Heyden et al. 2008; Wright et al. 2015). To date, all genetic studies of South African clinids have been restricted to a combination of mitochondrial DNA (mtDNA) markers (Dloop, COI, 16S, 12S) and single nuclear markers (rhodopsin, S7 intron) and were subject to the limitations of inference and resolution inherent to those markers (matrilineal inheritance (mtDNA), putatively neutral, mutation rate).

In this study, we aim to apply NGS techniques to further explore the population structure and biogeography of South African clinids. Specifically, we take advantage of a draft assembly of the genome of *C. superciliosus* to align RAD sequence reads from all three clinid species and identify tens of thousands of genome-wide SNPs. With this large set of nuclear markers, we aim to 1) test if genome-wide anonymous nDNA markers reveal similar genetic patterns based on life history (Wright et al. 2015) and biogeography (von der Heyden et al. 2013; Toms et al. 2014) previously identified by mtDNA and single nDNA markers. And 2) assess if, with added resolution, we can uncover patterns previously obscured at the level of mitochondrial or single nuclear markers.

Materials & Methods

The *M. dorsalis* and *C. superciliosus* sample used in this study are the same as those from Chapter II so the sample collection and DNA extraction, library preparation, and quality filtering and marker discovery steps are identical. The *C. cottoides* dataset was added for this study.

Sample collection and DNA extraction

We sampled fish specimens across the core of their distributional ranges from Jaccobsbaai on the west coast to Cape Agulhas on the southwest coast for all species (Figure 3.2). For *C. superciliosus* additional samples were collected from Swakopmund, Namibia and Knysna Heads, South Africa. We collected a total of 65, 26, and 121 specimens for *Muraenoclinus dorsalis*, *Clinus* cottoides, and *Clinus superciliosus* respectively between 2009 and 2019. Sample locations and N for each species can be found in Table 3.2. Fin clips and muscle tissue were immediately preserved in 95% ethanol and stored at -20 C. We extracted DNA from tissue using DNeasy Blood and Tissue kits (Qiagen) or chloroform/isopropanol extraction protocol and assessed DNA concentrations using a Qubit 4.0 fluorometer.

Library preparation and sequencing

To prepare the extracted DNA for sequencing, we constructed restriction site associated DNA (RAD) libraries for all samples and amplified the 16S mitochondrial region for a subset of specimens. We constructed the RAD libraries using a variation of the original protocol (Miller et al. 2007; Baird et al. 2008) using restriction enzyme SBFI described in (Miller et al. 2012) and NEBNext reagents (New England Biolabs). We started the library preparation with 100 ng of genomic DNA

(gDNA). We multiplexed and sheared libraries to roughly 500 bp lengths on a Biorupter Sonicator using four cycles of 30 seconds. We used Dynabeads (Invitrogen 11206D) to remove non-tagged DNA and used SPRI beads (DeAngelis et al. 1995) for purification and size selection. We carried out the final polymerase chain reaction (PCR) amplification on 16 ul reaction volumes for 10 amplification cycles. Samples used in the study were sequenced in one of five libraries, each containing 96 individually barcoded samples. Libraries were either sequenced on an Illumina HiSeq 4000 or NovaSeq 6000 S4 at the Vincent J. Coates Genomics Sequencing Laboratory at UC Berkeley supported by NIH S10 OD018174 Instrumentation Grant, or at Novogene Corporation, Sacramento, CA.

Quality filtering and marker discovery

For all specimens included in the study we trimmed, filtered, and genotyped raw reads using the software program Stacks version 2.2 (Rochette et al. 2019). We split raw reads according to their 6 bp unique barcode, removed the barcodes and trimmed reads on the 3' end to a final sequence length of 91 bp using the program *process_radtags*. Sequences were removed if the quality scores fell below 90%, a raw phred score of 10. We aligned the filtered sequences to the assembled *C. superciliosus* genome using Bowtie 2 (reference) with the preset option of --very-sensitive (-D 20 -R 3 -N 0 -L 20 -i S,1,0.50) to generate BAM files which we sorted and converted to SAM files with SAMtools (Langmead and Salzberg 2012). We then used the Stacks 2.0 shell program *ref_map.pl* to call single nucleotide polymorphisms (SNPs) and generate population-level summary statistics. We ran the Stacks program

populations to generate output files for downstream phylogenetic analysis with every individual being treated as a population for phylogenetic inference. We retained only a single SNP per locus (write-single_snp option in Stacks2).

Population structure analysis

To identify haplotype clusters we ran STRUCTURE 2.3.4 (Pritchard and Wen 2010) using the structure format output file from STACKS2 for each species and all loci. We used a parameter set of 10,000 iterations as the burn-in with 200,000 iterations under the admixture model. We ran 10 repetitions for K values ranging from 1 to the number of sample locations +2 (up to 11 in the case of CSU). We used the Evanno method (Evanno et al. 2005) implemented in Structure Harvester (Earl and vonHoldt 2012) to identify the highest likelihood for K.

To further assess genetic differences between populations we ran a discriminant analysis of principal components (DAPC) (Jombart et al. 2010). We performed the analysis using the ADEGENET package in R (Jombart 2008) with the variant call format file (VCF) output from STACKS2 *populations* command. We determined the number of principal components to retain using the cross-validation tool xvalDapc as implemented in Adegenet.

For comparison with F_{st} values from previous studies, we performed a pairwise differentiation analysis on the SNP dataset to calculate pairwise F_{st} between populations in GENODIVE v 3.03 (Meirmans 2020).

Isolation by Distance

To check for isolation by distance we performed a Mantel test to compare the

significance of the relationship between genetic distance (F_{st}) with a geographic distance for all species and sample locations. F_{st} values were produced by the pairwise differentiation analysis described above. We measured geographic distances following the coastline between sample sites using Google Maps distance calculator (www.mapdevelopers.com). To visualize the data, we plotted the relationship between F_{st} and geographic distance in R and ran a linear regression to test the goodness of fit (R^2) and to calculate p-values. To make the results more comparable in geographic scale we removed the Swakopmund population for *C. superciliosus* from the analysis.

Results

Nuclear and mitochondrial sequencing results

Sequencing generated > 50 Mb of filtered and cleaned data for all specimens. Alignment to the *C. superciliosus* genome resulted in an average alignment of 85% (range 74.2 to 87.8) for *M. dorsalis*, 87% (range 83 to 88.7) for *C. cottoides*, and 97% (range 81.3 to 97.8) for *C. superciliosus*. RAD libraries resulted in mean coverage of 13.79X, 31.81X, and 32.95X respectively with a final catalog containing 90,462 loci (*M. dorsalis*), 73,120 (*C. cottoides*), and 152,150 loci (*C. superciliosus*). The *populations* script identified 26,143 SNPs for *M. dorsalis*, and 21,214 SNPs for *C. cottoides*, and 55,822 SNPs for *C. superciliosus*, which is also the number of loci as only one SNP was kept for each locus.

For the mitochondrial 16S region we generated an alignment of 559 bp for 8 *M. dorsalis*, 6 *C. cottoides*, and 7 *C. superciliosus*.

Population structure analysis

For each of the three species, population structure was based on all orthologous loci identified by STACKS2. A likelihood assessment obtained by using the Evanno method on ten replicates for each value of K showed that maximum likelihood was reached when K = 2 for *M. dorsalis* and *C. cottoides* and when K = 3 for *C. superciliosus*. The STRUCTURE analysis showed genetic clustering partitioned by sample location (Figure 3.4). For all three species there was a nearly panmictic population (all samples 97-100% assigned to a genetic cluster) at Cape Agulhas and for points to the east with a stair step pattern of introgression for sample locations across False Bay and Cape Point. *M. dorsalis* and *C. cottoides* again showed a panmictic genetic cluster (all samples >99% assigned to a genetic cluster) along the West coast whereas *C. superciliosus* remained highly structured across the mixing region and cool temperate province to the northwestern most sample site in Swakopmund, Namibia.

Population pairwise F_{st} values were relatively high for all species ranging from 0.013 - 0.175 for *M. dorsalis*, 0.032 - 0.248 for *C. cottoides*, and 0.002 - 0.24 for *C. superciliosus*. P-values were significant for all F_{st} values at $P < 0.035$ (Table 3.5).

Isolation by distance

Regression analysis revealed a highly significant (P-value < 0.001) positive relationship between F_{st} and geographic distance for all three species. The linear model calculated an R^2 value of 0.562 for *M. dorsalis*, 0.399 for *C. cottoides*, and

0.879 for *C. superciliosus* (Figure 3.6).

Discriminant analysis of principal components

The DAPC, which can detect population differentiation at a finer scale than a structure analysis (Jombart et al. 2010) was able to cleanly separate nearly all sample locations for all species. The only sample locations that overlapped in the analysis were Gansbaai and Betty's Bay, and Kommetji and Cape of Good hope for *C. superciliosus,* and Sea Point and Miller's Point for *C. cottoides* (Figure 3.7). The analysis separated populations that appear panmictic in STRUCTURE plots revealing fine scale population genetic structure along the southwest coast. For *C. superciliosus,* samples from Agulhas were unambiguously separated from Betty's Bay and Gansbaai as were specimens from Agulhas and Knysna Heads.

Discussion

High throughput sequencing has made it possible to generate genome-wide datasets increasing the statistical power and resolution of population-level analysis, transitioning from population genetics to population genomics. We have employed NGS techniques to revisit populations of intertidal clinids and compare results to previously published outcomes while identifying novel inference made possible through advances in sequencing and analytical tools.

Population structure and biogeography

All avenues of investigation into population structure produced results that reveal high levels of genetic population structure for all study species. The results for STRUCTURE analyses and pairwise F_{st} at the level of genome-wide SNPs largely

corroborates hypotheses from previously published mtDNA or single nuclear marker studies (von der Heyden et al. 2013; Wright et al. 2015). The concordance between markers (mtDNA and nDNA) and methods (F_{st}, IBD) supports highly structured populations, which is in line with expectations based on species biology and suggests the system is well suited for clusters analysis (Gilbert 2016). Evanno analysis as implemented in Structure Harvester reached maximum likelihood at K=2 for *C. cottoides* and *M. dorsalis*, separating populations into one genetic cluster for locations in the cool temperate province (Kommetjie and Sea Point) and a second cluster for populations in the warm temperate province (Agulhas). In between, there is an area of genetically mixed populations with haplotypes belonging to both genetic clusters (Figure 3.8). Toms et al. (2014) hypothesized that sea level changes during glacial periods exposed the currently submerged Agulhas Bank which is composed of less rocky substrate than the present-day coastline. Sea level change may have created a vicariant barrier producing long stretches of coastline without rocky shore habitat ultimately leading to geographically isolated populations. This vicariant event is estimated to have lasted for ~40 k years allowing for genetic divergence in allopatry as the result of a combination of drift and adaptation to environmental conditions (e.g., sea surface temperature). As sea levels have risen to their current extent (~9 kya) the two genetic lineages have come into secondary contact, creating a hybrid zone in the region between Cape Point and Cape Agulhas. This southwest region of coastline has previously been identified as an area of relatively high genetic diversity (von der Heyden et al. 2013; Wright et al. 2015) which is consistent with the pattern

expected from admixture in a secondary contact zone. In this study, the correspondence of the sample locations with genetically mixed populations to the location of habitat loss during glacial maxima supports the hypothesis of a hybrid zone formed by contemporary habitat availability for multiple species.

Clinus superciliosus also reveals a pattern of genetic mixing along the southwestern coast but the pattern is distinct from the other two species. Firstly, while there is still a genetic break at Cape Agulhas, the nearby populations of Betty's Bay and Gansbaai are genetically more similar than seen in *M. dorsalis* and *C. superciliosus*. The more conspicuous pattern is that on the west coast, there continues to be a clear signal of population structure all the way to the northernmost population in Swakopmund, NA (Figure 3.9) and, unlike *M. dorsalis* and *C. cottoides*, *C. superciliosus* lacks a genetically panmictic cluster on the west coast. The most likely explanation for this distinction is the difference in life histories between the species. *C. superciliosus* is found in the mid and upper intertidal as juveniles, but they move to the lower intertidal and shallow subtidal zones as adults. The less specialized habitat needs may have allowed *C. superciliosus* to maintain better connectivity across the same region that produced a vicariant barrier to *M. dorsalis* and *C. cottoides*. This suggests that the more important driver of contemporary genetic structure for *C. superciliosus* may be a combination of dispersal potential and adaptation leading to a genetically mixed population that occurs across the entire cool temperate province (including the mixing region) and a distinct genetic lineage that occurs in the warm temperate province. A seascape ecology approach to explicitly

test for correlation between environmental conditions and genetic variance is outside the scope of this study but is a possible direction for future investigation.

Isolation by Distance and dispersal

In a secondary contact zone, where genetic lineages are hybridizing and dispersal potential is high, we would expect to see individuals in populations across the contact zone with genetic proportions that follow classic cross experiments based on their generation (F1, F2, etc.). Here, however, we see a genetically mixed region with sample locations that produce a "staircase" pattern where populations that are geographically closer to one of the genetic lineages are more similar to that genetic cluster. This is consistent with a pattern of isolation by distance. Because of the limited dispersal potential of clinids who are oviparous with minimal larval duration and predominantly sessile as adults, IBD has been previously hypothesized and tested as an important driver of genetic structure (Wright et al. 2015). There are published concerns around statistical bias in IBD analyses (Meirmans 2012) and we tried to address these concerns through an approach using multiple analyses (cluster analysis and F_{st}). The results of the linear model suggest high correlation of genetic differentiation (F_{st}) and geographic distance with statistical support for all species (MDO = 0.562, CCO = 0.399, CSU = 0.879 p-value < 0.001) (Figure 3.6). Additionally, discriminant analysis of principal components separated all but the most geographically proximate sample locations into distinct genetic populations (Figure 3.7). For all species the populations were organized in the plot along the x axis in order of their geographic location with the west coast populations (SWA,

KOM) consistently farthest from the eastern locations (AGU, KHE) in principal component space. Indeed, the shape of the plots closely follows the shape of South Africa's coastline from west to east. Our analysis cannot rule out the possibility that fine scale environmental variation has led to local adaptation preventing migrants from establishing in new locations. However, the convergence of evidence points towards isolation by distance driven by limited dispersal (a function of reproductive strategy and habitat specificity) as an important driver of genetic population structure in South African clinids.

Higher resolution SNP dataset

Thus far we've discussed how analysis on genome-wide SNP datasets has supported hypotheses previously tested with mtDNA markers, single nuclear markers, or a combination of the two. While there is value in validating the inferences made from prior studies it is also important to identify where the added resolution of NGS and computationally intensive analyses have contributed to questions that were previously untenable. The importance of intertidal zonation has been tested in the same three species of fishes (von der Heyden et al. 2013) using COI to reveal extremely high pairwise PHIst for the highshore *M. dorsalis*, but lower levels of population structure for the low-shore generalist *C. superciliosus* and obligate intertidal and mid-shore specialist *C. cottoides*. At the level of genome-wide nuclear markers we are able to determine that while all three species are quite highly structured across the same spatial scale, *C. cottoides* and *M. dorsalis* have similar values of pairwise F_{st} and higher levels of structure than *C. superciliosus* (Table

3.5). This pattern of pairwise population differentiation better fits the expectation that mid and high-shore species would be more genetically structured than low-shore or subtidal species. The higher resolution afforded by a genome-wide SNP dataset has provided clarity to a previously ambiguous result, adding to the evidence supporting the role of tidal elevation on genetic structure in clinid fishes.

There is also insight to be gained from comparing multiple analyses using the same genetic dataset. In plots produced by Bayesian clustering it is not possible to distinguish between sample locations along the southern shore (Betty's Bay to Knysna Heads) for *C. superciliosus*. However, the discriminant analysis of principal components, which uses a multivariate approach to explore variation between clusters while minimizing that within clusters (Jombart et al. 2010) clearly separates Knysna Heads and Agulhas from all other populations for *C. superciliosus* (Figure 3.7). The fine scale structure along the southern coast has led to debate around the classification of Cape Agulhas as a biogeographic break as it has been detected as such in some species, but not in others (Teske et al. 2011). The combination of genome-wide markers and discriminant analysis better matches the scale of variation found among sample locations and can reliably distinguish Agulhas and Kynsna Head individuals from other populations for all species, supporting the classification of Cape Agulhas as a phylogeographic break for South Africa's marine fauna and the western edge of the warm temperate biogeographic province. This fine scale variation suggests that there may be a subset of adaptive loci that are responsible for the genetic variance along the southern shore. Outlier analysis is out of the scope of this study but

identifying putatively adaptive loci to assess the importance of local adaptation is a promising direction for future study.

Conclusion

The pace of technological advance that has defined the field of molecular ecology for the last two decades shows no sign of slowing, and as NGS techniques become more affordable genomic datasets are likely to become ubiquitous in molecular studies. Here, we have applied NGS sequencing and analyses to a system, which at the level of mtDNA and single nDNA markers is well researched. Largely, we have demonstrated that anonymous genome-wide markers support inferences made from earlier studies, with analyses from a SNP dataset complimenting results obtained from classic genetic markers. While this validation is important, we have also shown that the additional statistical power paired with more sophisticated analytical tools can elucidate previously untenable or ambiguous results. We have just begun to explore the possible applications for this dataset with many promising avenues of investigation still available. As techniques in the field of molecular ecology advances, we will continue to revisit the fishes of South Africa's dynamic and fascinating rocky intertidal.

Tables and Figures

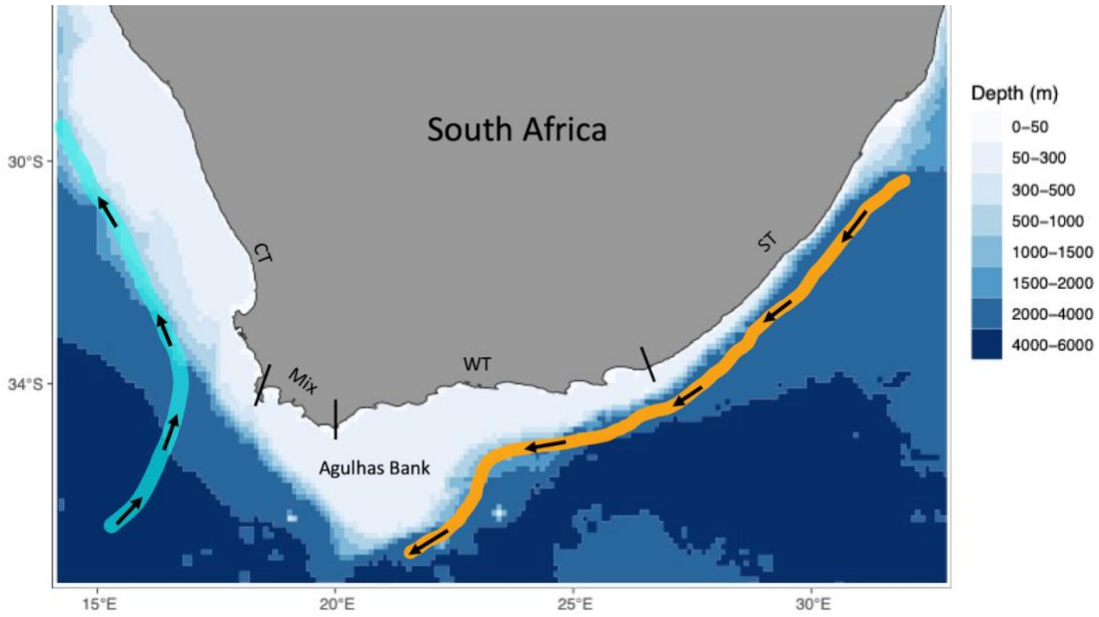


Figure 3.1. Map depicting the coast of South Africa (in gray) and the sea floor depth in meters on a gradient between white and dark blue. The continental shelf is depicted in white. The Agulhas current running from north to south (orange) and the south to north flowing Benguela current (blue). Arrows indicate the direction of flow. Solid black lines separate the coastline into biogeographic provinces. From west to east there is the cool temperate province (CT), a mixing region (Mix), the warm temperate province (WT), and the subtropical province (ST).

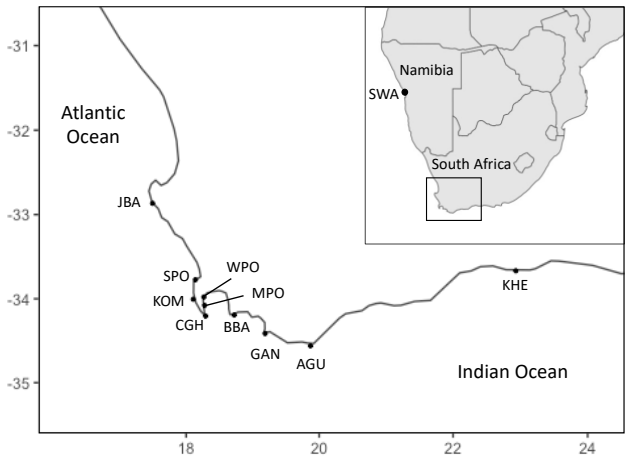


Figure 3.2. Map of sampling locations. Following the coastline from west to east, abbreviations are Jacobsbaai (JBA), Sea Point (SPO), Kommetjie (KOM), Cape of Good Hope (CGH), Miller's Point (MPO), Woolley's Pool (WPO), Betty's Bay (BBA), Gansbaai (GAN), Cape Agulhas (AGU), and Knysna Heads (KHE). Inset shows the Swakopmund (SWA) location in Namibia.

Table 3.3. Number of specimens (N) collected at each sample location. Sample locations are oriented from west to east and names are abbreviated as follows: Swakopmund (SWA), Jacobsbaai (JBA), Sea Point (SPO), Kommetjie (KOM), Cape of Good Hope (CGH), Miller's Point (MPO), Woolley's Pool (WPO), Betty's Bay (BBA), Gansbaai (GAN), Cape Agulhas (AGU), Knysna Heads (KHE).

Species	Common name	SWA	JBA	SPO	KOM	CGH	MPO	WPO	BBA	GAN	AGU	KHE	Total N
Clinus superciliosus	Super klipfish	4	14	17	5	5	18	-	26	20	7	5	121
Clinus cottoides	Bluntnose klipfish	-	5	4	-	-	7	-	-	5	5	-	26
Muraenoclinus dorsalis	Nosestripe klipfish	-	4	24	-	-	3	4	9	4	17	-	65

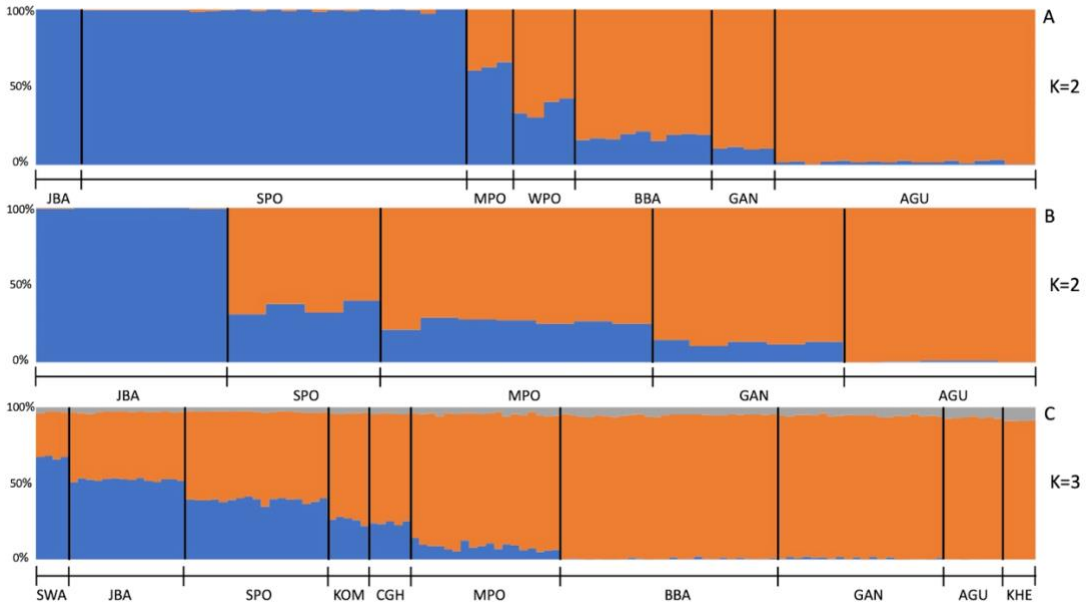


Figure 3.4. Structure plots with Bayesian assignment of A) *Muraenoclinus dorsalis*, B) *Clinus cottoides*, C) *Clinus superciliosus*. Sample locations are organized from west to east. Location abbreviations are as follows: Swakopmund (SWA), Jacobsbaai (JBA), Sea Point (SPO), Kommetjie (KOM), Cape of Good Hope (CGH), Miller's Point (MPO), Woolley's Pool (WPO), Betty's Bay (BBA), Gansbaai (GAN), Cape Agulhas (AGU), Knysna Heads (KHE).

Table 3.5. Population pairwise F_{st} values for *Clinus cottoides* (upper table, above the diagonal), *Muraenoclinus dorsalis* (upper table, below the diagonal), and *Clinus superciliosus* (lower table). All values are significant at $P < 0.03$.

Sample locality	Jacobsbaai	Sea Point	Miller's Point	Woolley's Pool	Betty's Bay	Gansbaai	Agulhas
Jacobsbaai	--	0.188	0.194	--	--	0.207	0.248
Sea Point	0.063	--	0.032	--	--	0.060	0.108
Miller's Point	0.135	0.07	--	--	--	0.053	0.103
Woolley's Pool	0.158	0.107	0.040	--	--	--	--
Betty's Bay	0.135	0.129	0.047	0.024	--	--	--
Gansbaai	0.175	0.167	0.064	0.043	0.013	--	0.045
Agulhas	0.152	0.164	0.073	0.041	0.027	0.015	--

Sample locality	Swakopmund	Jacobsbaai	Sea Point	Slankop	Cape of Good Hope	Miller's Point	Betty's Bay	Gansbaai	Agulhas	Kynsna Heads
Swakopmund	--									
Jacobsbaai	0.057	--								
Sea Point	0.085	0.030	--							
Slankop	0.101	0.040	0.015	--						
Cape of Good Hope	0.103	0.044	0.019	0.002	--					
Miller's Point	0.132	0.077	0.049	0.029	0.029	--				
Betty's Bay	0.136	0.085	0.055	0.036	0.035	0.021	--			
Gansbaai	0.136	0.085	0.055	0.035	0.033	0.023	0.004	--		
Agulhas	0.182	0.122	0.097	0.082	0.076	0.047	0.040	0.040	--	
Kynsna Heads	0.240	0.174	0.150	0.143	0.135	0.098	0.094	0.091	0.030	--

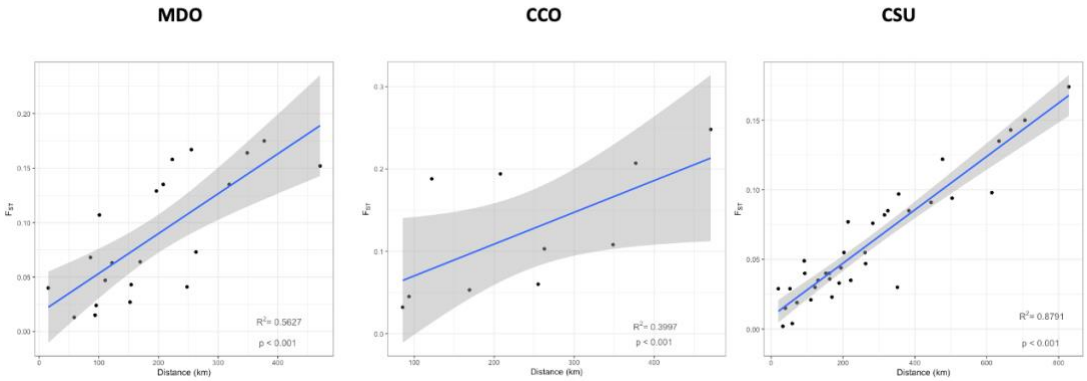


Figure 3.6. Linear regression showing the correlation between genetic distance (F_{st}) and geographic distance from left to right of M. dorsalis (MDO), C. cottoides (CCO), and C. superciliosus (CSU). R^2 and p values are shown in the bottom right corner of each plot.

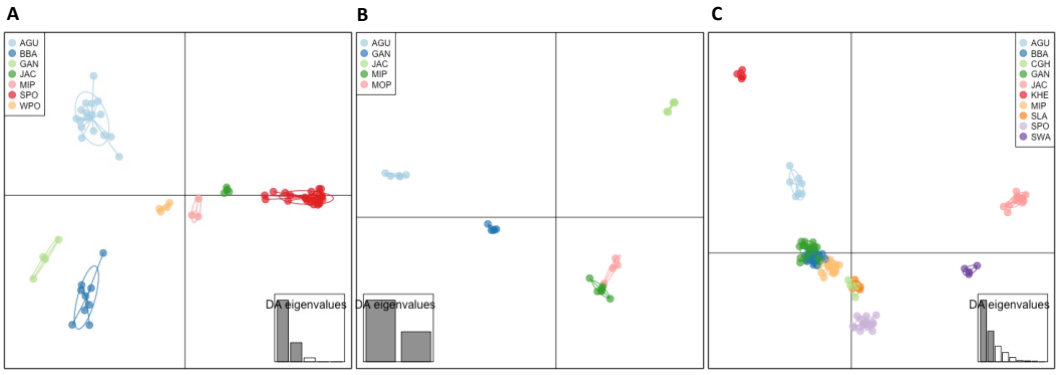


Figure 3.7. Discriminate analysis of principal components (DAPC) cluster plot of *M. dorsalis* (MDO), *C. cottoides* (CCO), and *C. superciliosus* (CSU). Plots were created retaining 5 principal components. Plots illustrate the first and second axes.

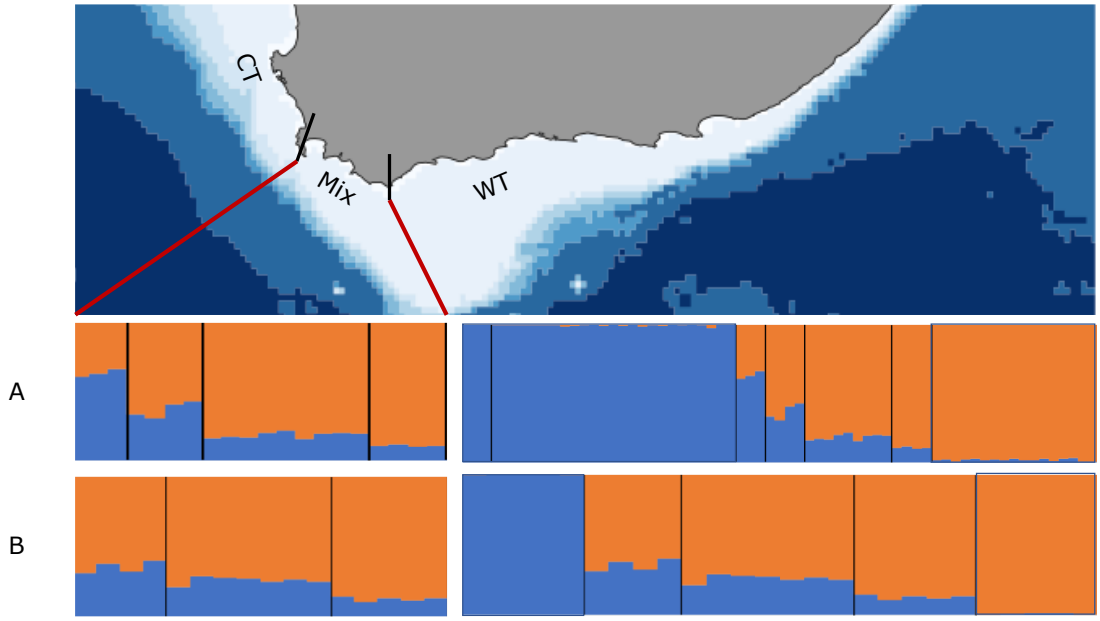


Figure 3.8. Map of South Africa with continental shelf in white. Abbreviations indicate biogeographic provinces (CT) cool temperate and (WT) warm temperate. The mixed region in between is in labeled Mix. A) Full structure plot of *M. dorsalis* on the right. On the left is the section of the plot that corresponds to the mixed province on the map (connected to map by red lines). B) Same layout as A but for *C. cottoides*.

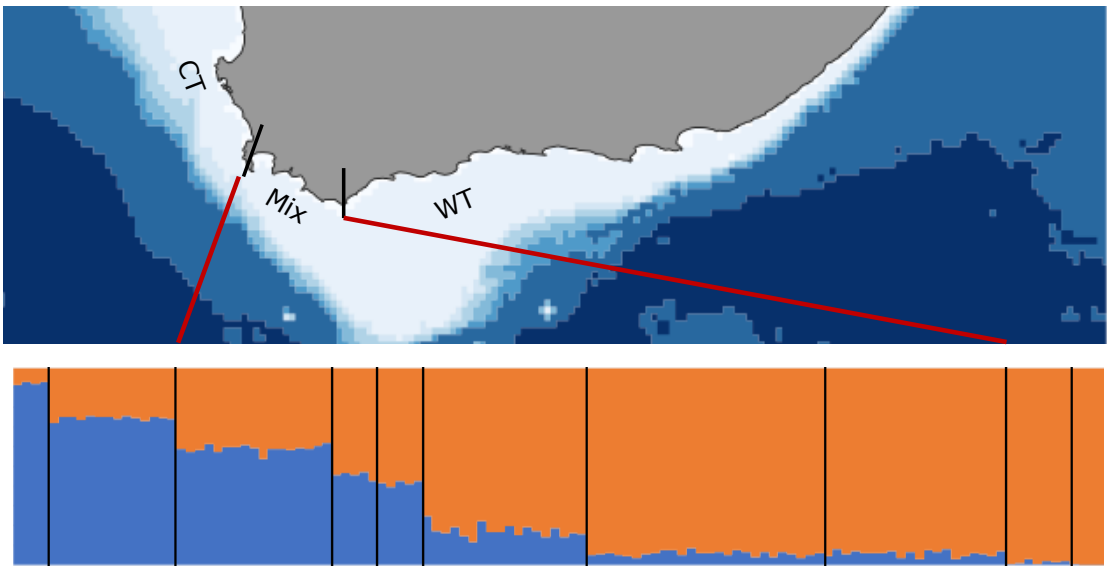


Figure 3.9. Map of South Africa with continental shelf in white. Abbreviations indicate biogeographic provinces (CT) cool temperate and (WT) warm temperate.

The mixed region in between is in labeled Mix. Bottom figure is structure plot for *C. superciliosus* with locations that fall within the mixed province indicated by red lines.

Synthesis

The evolutionary forces that shape the abundance and distribution of marine species are dynamic and difficult to directly observe (Palumbi 1994). Biogeography, isolation by distance, local adaptation, geological history, life history, and dispersal all influence connectivity between populations of marine organisms. The degree of genetic connectivity between populations can vary along a spectrum between panmictic (one genetic population across all locations) to insipient speciation (separate genetic lineages). Understanding the systematics and population structure of marine organisms is an important input for the prioritization of locations for conservation (Gleason et al. 2013; Kirkman et al. 2021). Undescribed species or cryptic species that are undiscovered pose a challenge to conservation managers who don't have a full view of the biodiversity they are trying to conserve (von der Heyden 2011). Biogeographic breaks and limits to dispersal can lead to genetically unique populations that deserve protection as their own conservation unit but lack apparent morphological differences for identification.

Molecular techniques have offered solutions to some of the challenges to observing the population dynamics of marine species. Genetic methods can identify cryptic species, estimating population demographics, and quantifying genetic connectivity between populations. As the discipline of molecular ecology transitions from traditional genetics to genomics we have generated genomic resources and applied genomic analyses to better understand the systematics and population dynamics of fishes in the family Clinidae and present examples across the spectrum of genetic connectivity.

In **Chapter 1**, we explored the systematics and population structure of clinids in the genus *Gibbonsia* from North America's west coast. We found 4 genetic lineages among our samples, which is one more than the three described species. Our genomic study supports the

description of a sub species of *Gibbonsia elegans* by C.L. Clark that is endemic to Guadalupe Island that he named *G. e. erroli*. We further investigated the population structure of *G. metzi* and *G. montereyensis* collected from the San Francisco Bay region of California. We found that *G. metzi* populations fit with the expectation of panmixia based on their long larval duration, but *G. montereyensis* showed low but significant levels of population structure over relatively short distances suggesting fine scale local adaptation. Our work better resolves the systematics of *Gibbonsia*, adding a fourth phylogenetic species and identified fine scale differentiation in a system thought to be entirely panmictic.

In **Chapter 2**, we took the first steps into transitioning the exploration the population dynamics from traditional genetics to population genomics for the klipfishes of South Africa. We sequenced and assembled the genome of the Super klipfish, *Clinus superciliosus* resulting in a highly contiguous (no gaps), and complete (94%) genome. The assembled genome represents the first in the family Clinidae and will be a valuable tool for genomic analysis of South Africa's clinids. We further produced phylogenies for *C. superciliosus* and the Nosestripe klipfish *Muraenoclinus dorsalis* and found that the phylogenies using SNPs is concordant with those using mtDNA in identifying three genetic clades for *C. superciliosus* and two for *M. dorsalis*, but we failed to collect any new specimens from the minor clades for either. We were able to identify all three genetic clades for *C. superciliosus* using a published phylogeny but maintain that tissue samples for *M. dorsalis* represent a yet undescribed cryptic species. Our study contributes valuable genomic resources to the investigation of a dynamic coastal marine ecosystem and highlights the potential challenges in identifying species using morphological characters alone in shallow marine systems.

In **Chapter 3**, we built upon the results of Chapter 2 to investigate population structure within a genetic lineage for *C. superciliosus*, *M. dorsalis*, plus a third species, the

Bluntnose klipfish *Clinus cottoides*. We found that all three showed high levels of population structure across the transition zone between cool temperate province of the west coast and the warm temperate province east of Cape Agulhas. We suggest that this region is a contemporary hybridization zone between genetic lineages that diverged during the last glacial maxima. Our work supports the presence of impactful biogeographic breaks at Cape Point and Cape Agulhas and the importance of the regions geological history in shaping contemporary population dynamics. We highlight the benefits of using genome-wide nuclear markers to capture the full range of evolutionary processes acting across the genome. Our work supports the need for a network of marine protected areas to preserve the genetic diversity present in South Africa's coastal marine ecosystem.

Going forward, we plan to perform an outlier analysis South African clinids to identify putatively adaptive loci. By mapping these to the genome of C. superciliosus we hope to identify potentially adaptive regions of the genome to target for future investigation. The addition of the assembled genome to the toolkit also opens the possibility of whole genome resequencing broadening the search for adaptive loci.

For *Gibbonsia*, we plan to return to Guadalupe Island (GI) to collect additional specimens of *G. e. erroli* and collect samples of putative *G. elegans* from other offshore islands to try and assess the species' range. We intend to collect *G. montereyensis* samples from Guadalupe Island as well as we have mitochondrial evidence that there is another GI endemic previously described by C.L. Hubbs as *G. m. norae*. On mainland North America, we plan to expand the sample locations to better cover the species range for *G. montereyensis* and *G. metzi* and further explore their patterns of population structure.

Bibliography

Alex Buerkle C, Gompert Z. 2013. Population genomics based on low coverage sequencing: how low should we go? Mol Ecol. 22(11):3028–3035. doi:10.1111/mec.12105.

Amor MD, Doyle SR, Norman MD, Roura A, Hall NE, Robinson AJ, Leite TS, Strugnell JM. 2019. Genome-wide sequencing uncovers cryptic diversity and mito-nuclear discordance in the *Octopus vulgaris* species complex. Molecular Biology. [accessed 2022 Aug 6]. http://biorxiv.org/lookup/doi/10.1101/573493.

Andrews KR, Good JM, Miller MR, Luikart G, Hohenlohe PA. 2016. Harnessing the power of RADseq for ecological and evolutionary genomics. Nature Reviews Genetics. 17(2):81–92. doi:10.1038/nrg.2015.28.

Angeloni F, Wagemaker N, Vergeer P, Ouborg J. 2012. Genomic toolboxes for conservation biologists: Genomic toolboxes for conservation biologists. Evolutionary Applications. 5(2):130–143. doi:10.1111/j.1752-4571.2011.00217.x.

Avise JC. 2004. Molecular markers, natural history, and evolution. 2nd ed. Sunderland, Mass: Sinauer Associates.

Baird NA, Etter PD, Atwood TS, Currey MC, Shiver AL, Lewis ZA, Selker EU, Cresko WA, Johnson EA. 2008. Rapid SNP Discovery and Genetic Mapping Using Sequenced RAD Markers. PLOS ONE. 3(10):e3376. doi:10.1371/journal.pone.0003376.

Bale R, Neveln ID, Bhalla APS, MacIver MA, Patankar NA. 2015. Convergent Evolution of Mechanically Optimal Locomotion in Aquatic Invertebrates and Vertebrates. Hedenström A, editor. PLoS Biol. 13(4):e1002123. doi:10.1371/journal.pbio.1002123.

Bennett BA, Griffiths CL. 1984. Factors affecting the distribution, abundance and diversity of rock-pool fishes on the Cape Peninsula, South Africa. South African Journal of Zoology. 19(2):97–104. doi:10.1080/02541858.1984.11447865.

Bernardi G. 2013a. Speciation in fishes. Molecular Ecology. 22(22):5487–5502. doi:https://doi.org/10.1111/mec.12494.

Bernardi G. 2013b. Speciation in fishes. Mol Ecol. 22(22):5487–5502. doi:10.1111/mec.12494.

Bickford D, Lohman DJ, Sodhi NS, Ng PKL, Meier R, Winker K, Ingram KK, Das I. 2007. Cryptic species as a window on diversity and conservation. Trends in Ecology & Evolution. 22(3):148–155. doi:10.1016/j.tree.2006.11.004.

Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 30(15):2114–2120.

doi:10.1093/bioinformatics/btu170.

Branch G, Griffiths C, Branch M, Beckley L. 2016. Two Oceans. Fourth. Cape Town, South Africa: Struik Nature.

Buser TJ, Burns MD, López JA. 2017. Littorally adaptive? Testing the link between habitat, morphology, and reproduction in the intertidal sculpin subfamily Oligocottinae (Pisces: Cottoidea). PeerJ. 5:e3634. doi:10.7717/peerj.3634.

Caballero-Huertas M, Frigola-Tepe X, Coll M, Muñoz M, Viñas J. 2022. The current knowledge status of the genetic population structure of the European sardine (Sardina pilchardus): uncertainties to be solved for an appropriate fishery management. Rev Fish Biol Fisheries. doi:10.1007/s11160-022-09704-z. [accessed 2022 Jun 28]. https://link.springer.com/10.1007/s11160-022-09704-z.

Caley MJ, Carr MH, Hixon MA, Hughes TP, Jones GP, Menge BA. 1996. Recruitment and the local dynamics of open marine populations. Annu Rev Ecol Syst. 27(1):477–500. doi:10.1146/annurev.ecolsys.27.1.477.

Campbell CR, Poelstra JW, Yoder AD. 2018. What is Speciation Genomics? The roles of ecology, gene flow, and genomic architecture in the formation of species. Biological Journal of the Linnean Society. 124(4):561–583. doi:10.1093/biolinnean/bly063.

Catchen J, Hohenlohe PA, Bassham S, Amores A, Cresko WA. 2013. Stacks: an analysis tool set for population genomics. Molecular Ecology. 22(11):3124–3140. doi:10.1111/mec.12354.

Challis R, Richards E, Rajan J, Cochrane G, Blaxter M. 2020. BlobToolKit – Interactive Quality Assessment of Genome Assemblies. G3 Genes|Genomes|Genetics.:14. doi:10.1534/g3.119.400908.

Checkley DM, Barth JA. 2009. Patterns and processes in the California Current System. Progress in Oceanography. 83(1):49–64. doi:10.1016/j.pocean.2009.07.028.

Choi JY, Lye ZN, Groen SC, Dai X, Rughani P, Zaaijer S, Harrington ED, Juul S, Purugganan MD. 2020. Nanopore sequencing-based genome assembly and evolutionary genomics of circum-basmati rice. Genome Biol. 21(1):21. doi:10.1186/s13059-020-1938-2.

Coyne JA, Orr HA. 2004. Speciation. Sunderland, MA: Sinauer Associates, Inc.

Dalongeville A, Nielsen ES, Teske PR, Heyden S. 2022. Comparative phylogeography in a marine biodiversity hotspot provides novel insights into evolutionary processes across the Atlantic-Indian Ocean transition. Guo B (HOTSP), editor. Diversity and Distributions.:ddi.13534. doi:10.1111/ddi.13534.

Darwin CR. 1872. On the Origin of Species by Natural Selection. 5th ed. New York: Appleton.

Davey JW, Cezard T, Fuentes-Utrilla P, Eland C, Gharbi K, Blaxter ML. 2013. Special features of RAD Sequencing data: implications for genotyping. Mol Ecol. 22(11):3151–3164. doi:10.1111/mec.12084.

Davey JW, Hohenlohe PA, Etter PD, Boone JQ, Catchen JM, Blaxter ML. 2011. Genome-wide genetic marker discovery and genotyping using next-generation sequencing. Nat Rev Genet. 12(7):499–510. doi:10.1038/nrg3012.

Davis BJ. 1977. Distribution and temperature adaptation in the teleost fish genus Gibbonsia. Mar Biol. 42(4):315–320. doi:10.1007/BF00402193.

DeAngelis MM, Wang DG, Hawkins TL. 1995. Solid-phase reversible immobilization for the isolation of PCR products. Nucl Acids Res. 23(22):4742–4743. doi:10.1093/nar/23.22.4742.

Dobzhansky T. 1937. Genetics and the Origin of Species. New York: Columbia University Press.

Earl DA, vonHoldt BM. 2012. STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. Conservation Genet Resour. 4(2):359–361. doi:10.1007/s12686-011-9548-7.

Edwards SV, Shultz AJ, Campbell-Staton SC. 2015. Next-generation sequencing and the expanding domain of phylogeography. Folia Zoologica. 64(3):187–206. doi:10.25225/fozo.v64.i3.a2.2015.

Egan JP, Buser TJ, Burns MD, Simons AM, Hundt PJ. 2021. Patterns of Body Shape Diversity and Evolution in Intertidal and Subtidal Lineages of Combtooth Blennies (Blenniidae). Integrative Organismal Biology. 3(1):obab004. doi:10.1093/iob/obab004.

Ekblom R, Wolf JBW. 2014. A field guide to whole-genome sequencing, assembly and annotation. Evol Appl. 7(9):1026–1042. doi:10.1111/eva.12178.

Ellegren H. 2014. Genome sequencing and population genomics in non-model organisms. Trends in Ecology & Evolution. 29(1):51–63. doi:10.1016/j.tree.2013.09.008.

Evanno G, Regnaut S, Goudet J. 2005. Detecting the number of clusters of individuals using the software structure: a simulation study. Mol Ecol. 14(8):2611–2620. doi:10.1111/j.1365-294X.2005.02553.x.

Faria R, Johannesson K, Stankowski S. 2021. Speciation in marine environments: Diving under the surface. J of Evolutionary Biology. 34(1):4–15. doi:10.1111/jeb.13756.

Fuentes-Pardo AP, Ruzzante DE. 2017. Whole-genome sequencing approaches for conservation biology: Advantages, limitations and practical recommendations. Molecular Ecology. 26(20):5369–5406. doi:10.1111/mec.14264.

Futuyma DJ, Kirkpatrick M. 2017. Evolution. 4th ed. Sunderland, MA: Sinauer Associates, Inc.

Georges A, Gruber B, Pauly GB, White D, Adams M, Young MJ, Kilian A, Zhang X, Shaffer HB, Unmack PJ. 2018. Genomewide SNP markers breathe new life into phylogeography and species delimitation for the problematic short-necked turtles (Chelidae: Emydura) of eastern Australia. Molecular Ecology. 27(24):5195–5213. doi:10.1111/mec.14925.

Gilbert KJ. 2016. Identifying the number of population clusters with structure: problems and solutions. Mol Ecol Resour. 16(3):601–603. doi:10.1111/1755-0998.12521.

Gleason M, Fox E, Ashcraft S, Vasques J, Whiteman E, Serpa P, Saarman E, Caldwell M, Frimodig A, Miller-Henson M, et al. 2013. Designing a network of marine protected areas in California: Achievements, costs, lessons learned, and challenges ahead. Ocean & Coastal Management. 74:90–101. doi:10.1016/j.ocecoaman.2012.08.013.

Godinho R, Crespo E g., Ferrand N. 2008. The limits of mtDNA phylogeography: complex patterns of population history in a highly structured Iberian lizard are only revealed by the use of nuclear markers. Molecular Ecology. 17(21):4670–4683. doi:10.1111/j.1365-294X.2008.03929.x.

Goschen WS, Schumann EH. 1995. Upwelling and the occurrence of cold water around Cape Recife, Algoa Bay, South Africa. South African Journal of Marine Science. 16(1):57–67. doi:10.2989/025776195784156520.

Griffiths CL, Robinson TB, Lange L, Mead A. 2010. Marine Biodiversity in South Africa: An Evaluation of Current States of Knowledge. PLOS ONE. 5(8):e12008. doi:10.1371/journal.pone.0012008.

Hahn MW. 2018. Molecular population genetics. New York : Sunderland, MA: Oxford University Press ; Sinauer Associates.

von der Heyden S. 2009. Why do we need to integrate population genetics into South African marine protected area planning? African Journal of Marine Science. 31(2):263–269. doi:10.2989/AJMS.2009.31.2.14.886.

von der Heyden S. 2011. ‘Carry on sampling!’– assessing marine fish biodiversity and discovery rates in southern Africa. Diversity and Distributions. 17(1):81–92. doi:10.1111/j.1472-4642.2010.00712.x.

von der Heyden S, Bowie RCK, Prochazka K, Bloomer P, Crane NL, Bernardi G.

2011. Phylogeographic patterns and cryptic speciation across oceanographic barriers in South African intertidal fishes: Phylogeographic patterns and cryptic speciation. Journal of Evolutionary Biology. 24(11):2505–2519. doi:10.1111/j.1420-9101.2011.02382.x.

von der Heyden S, Gildenhuys E, Bernardi G, Bowie RCK. 2013. Fine-scale biogeography: tidal elevation strongly affects population genetic structure and demographic history in intertidal fishes. [accessed 2019 Jun 7]. https://scholar.sun.ac.za:443/handle/10019.1/90896.

von der Heyden S, Prochazka K, Bowie RCK. 2008. Significant population structure and asymmetric gene flow patterns amidst expanding populations of Clinus cottoides (Perciformes, Clinidae): application of molecular data to marine conservation planning in South Africa. Molecular Ecology. 17(22):4812–4826. doi:10.1111/j.1365-294X.2008.03959.x.

von der Heyden S, Toms J, Teske P, Lamberth S, Holleman W. 2015. Contrasting signals of genetic diversity and historical demography between two recently diverged marine and estuarine fish species. Mar Ecol Prog Ser. 526:157–167. doi:10.3354/meps11191.

Hickey BM. 1979. The California current system—hypotheses and facts. Progress in Oceanography. 8(4):191–279. doi:10.1016/0079-6611(79)90002-8.

Hirase S, Yamasaki YY, Sekino M, Nishisako M, Ikeda M, Hara M, Merilä J, Kikuchi K. 2021. Genomic Evidence for Speciation with Gene Flow in Broadcast Spawning Marine Invertebrates. Crandall K, editor. Molecular Biology and Evolution. 38(11):4683–4699. doi:10.1093/molbev/msab194.

Hoban SM, Hauffe HC, Pérez-Espona S, Bertorelle G, Frith K, Galbusera P, Hoelzel AR, Nichols R, Primmer C, Siegismund HR, et al. 2013. How to bring genetic diversity to the forefront of conservation policy and management. Conservation Genet Resour. 5:593–598.

Holleman W, Von Der Heyden S, Zsilavecz G. 2012. Delineating the fishes of the Clinus superciliosus species complex in southern African waters (Blennioidei: Clinidae: Clinini), with the validation of Clinus arborescens Gilchrist & Thompson, 1908 and Clinus ornatus Gilchrist & Thompson, 1908, and with descriptions of two new species. Zool J Linn Soc. 166(4):827–853. doi:10.1111/j.1096-3642.2012.00865.x.

Horn MH, Gibson RN. 1988. Intertidal Fishes. Scientific American. 258(1):64–71.

Hortal J, de Bello F, Diniz-Filho JAF, Lewinsohn TM, Lobo JM, Ladle RJ. 2015. Seven Shortfalls that Beset Large-Scale Knowledge of Biodiversity. Annu Rev Ecol Evol Syst. 46(1):523–549. doi:10.1146/annurev-ecolsys-112414-054400.

Hubbs C. 1952. A contribution to the classification of the Blennioid fishes of the

family Clinidae.

Jombart T. 2008. adegenet: a R package for the multivariate analysis of genetic markers. Bioinformatics. 24(11):1403–1405. doi:10.1093/bioinformatics/btn129.

Jombart T, Devillard S, Balloux F. 2010. Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. BMC Genet. 11(1):94. doi:10.1186/1471-2156-11-94.

Kearse M, Moir R, Wilson A, Stones-Havas S, Cheung M, Sturrock S, Buxton S, Cooper A, Markowitz S, Duran C, et al. 2012. Geneious Basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics. 28(12):1647–1649. doi:10.1093/bioinformatics/bts199.

Kelly RP, Palumbi SR. 2010. Genetic Structure Among 50 Species of the Northeastern Pacific Rocky Intertidal Community. Sandin SA, editor. PLoS ONE. 5(1):e8594. doi:10.1371/journal.pone.0008594.

Kirkman S, Mann B, Sink K, Adams R, Livingstone T-C, Mann-Lang J, Pfaff M, Samaai T, van der Bank M, Williams L, et al. 2021. Evaluating the evidence for ecological effectiveness of South Africa's marine protected areas. African Journal of Marine Science. 43(3):389–412. doi:10.2989/1814232X.2021.1962975.

Knowlton N. 1986. Cryptic and Sibling Species Among the Decapod Crustacea. Journal of Crustacean Biology. 6(3):356–363. doi:10.1163/193724086X00208.

Kunvar S, Czarnomska S, Pertoldi C, Tokarska M. 2021. In Search of Species-Specific SNPs in a Non-Model Animal (European Bison (Bison bonasus))—Comparison of De Novo and Reference-Based Integrated Pipeline of STACKS Using Genotyping-by-Sequencing (GBS) Data. Animals. 11(8):2226. doi:10.3390/ani11082226.

Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. Nat Methods. 9(4):357–359. doi:10.1038/nmeth.1923.

Larmuseau MHD, Raeymaekers JAM, Hellemans B, Van Houdt JKJ, Volckaert FAM. 2010. Mito-nuclear discordance in the degree of population differentiation in a marine goby. Heredity. 105(6):532–542. doi:10.1038/hdy.2010.9.

Leaché AD, Fujita MK, Minin VN, Bouckaert RR. 2014. Species Delimitation using Genome-Wide SNP Data. Systematic Biology. 63(4):534–542. doi:10.1093/sysbio/syu018.

Lester SE, Ruttenberg BI, Gaines SD, Kinlan BP. 2007. The relationship between dispersal ability and geographic range size. Ecol Letters. 10(8):745–758. doi:10.1111/j.1461-0248.2007.01070.x.

Li H. 2013. Aligning sequence reads, clone sequences and assembly contigs with

BWA-MEM. [accessed 2022 Jul 8]. http://arxiv.org/abs/1303.3997.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup. 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 25(16):2078–2079. doi:10.1093/bioinformatics/btp352.

Lima D, Freitas JEP, Araujo ME, Solé-Cava AM. 2005. Genetic detection of cryptic species in the frillfin goby Bathygobius soporator. Journal of Experimental Marine Biology and Ecology. 320(2):211–223. doi:10.1016/j.jembe.2004.12.031.

Love M. 2011. Certainly more than you want to know about the fishes of the Pacific coast. Santa Barbara: Really Big Press.

Love M, Passarelli J. 2020. Miller and Lea's guide to the coastal marine fishes of California. 2nd ed. UC Agriculture and Natural Resources Publications.

Lu H, Giordano F, Ning Z. 2016. Oxford Nanopore MinION Sequencing and Genome Assembly. Genomics, Proteomics & Bioinformatics. 14(5):265–279. doi:10.1016/j.gpb.2016.05.004.

Macías Garcia C, de Perera T. 2002. Ultraviolet-based female preferences in a viviparous fish. Behavioral Ecology and Sociobiology. 52(1):1–6. doi:10.1007/s00265-002-0482-2.

Mayr E. 1942. Systematics And The Origin Of Species. New York: Columbia University Press.

Meirmans PG. 2012. The trouble with isolation by distance: NEWS AND VIEWS: OPINION. Molecular Ecology. 21(12):2839–2846. doi:10.1111/j.1365-294X.2012.05578.x.

Meirmans PG. 2020. GENODIVE version 3.0: Easy-to-use software for the analysis of genetic data of diploids and polyploids. Mol Ecol Resour. 20(4):1126–1131. doi:10.1111/1755-0998.13145.

Meirmans PG, Hedrick PW. 2011. Assessing population structure: FST and related measures. Molecular Ecology Resources. 11(1):5–18. doi:10.1111/j.1755-0998.2010.02927.x.

Michael TP, Jupe F, Bemm F, Motley ST, Sandoval JP, Lanz C, Loudet O, Weigel D, Ecker JR. 2018. High contiguity Arabidopsis thaliana genome assembly with a single nanopore flow cell. Nat Commun. 9(1):541. doi:10.1038/s41467-018-03016-2.

Miller MR, Brunelli JP, Wheeler PA, Liu S, Rexroad CE, Palti Y, Doe CQ, Thorgaard GH. 2012. A conserved haplotype controls parallel adaptation in geographically distant salmonid populations. Molecular Ecology. 21(2):237–249. doi:10.1111/j.1365-294X.2011.05305.x.

Miller MR, Dunham JP, Amores A, Cresko WA, Johnson EA. 2007. Rapid and cost-effective polymorphism identification and genotyping using restriction site associated DNA (RAD) markers. Genome Res. 17(2):240–248. doi:10.1101/gr.5681207.

Morin PA, Luikart G, Wayne RK, the SNP workshop group. 2004. SNPs in ecology, evolution and conservation. Trends in Ecology & Evolution. 19(4):208–216. doi:10.1016/j.tree.2004.01.009.

Moser HG. 2007. Reproduction in the viviparous South African clinid fish Fucomimus mus. African Journal of Marine Science. 29(3):423–436. doi:10.2989/AJMS.2007.29.3.10.340.

Myers N, Mittermeier RA, Mittermeier CG, da Fonseca GAB, Kent J. 2000. Biodiversity hotspots for conservation priorities. Nature. 403(6772):853–858. doi:10.1038/35002501.

Narum SR, Buerkle CA, Davey JW, Miller MR, Hohenlohe PA. 2013. Genotyping-by-sequencing in ecological and conservation genomics. Molecular Ecology. 22(11):2841–2847. doi:10.1111/mec.12350.

Neethling M, Matthee CA, Bowie RC, von der Heyden S. 2008. Evidence for panmixia despite barriers to gene flow in the southern African endemic, Caffrogobius caffer (Teleostei: Gobiidae). BMC Evol Biol. 8(1):325. doi:10.1186/1471-2148-8-325.

Nelson JS. 2006. Fishes of the world. 4th ed. Hoboken, N.J: John Wiley.

Nielsen ES, Henriques R, Beger M, Toonen RJ, von der Heyden S. 2020. Multi-model seascape genomics identifies distinct environmental drivers of selection among sympatric marine species. BMC Evol Biol. 20(1):121. doi:10.1186/s12862-020-01679-4.

Nielsen ES, Henriques R, Toonen RJ, Knapp ISS, Guo B, von der Heyden S. 2018. Complex signatures of genomic variation of two non-model marine species in a homogeneous environment. BMC Genomics. 19(19):347. doi:10.1186/s12864-018-4721-y.

Palumbi SR. 1994. Genetic Divergence, Reproductive Isolation, and Marine Speciation. :27.

Palumbi SR, Martin A, Romano S, McMillan WO, Stice L, Grabowski G. 1991. The Simple Fool's Guide to PCR. University of Hawaii.

Pritchard JK, Wen W. 2010. Documentation for structure software: Version 2. :30.

Prochazka K, Griffiths CL. 1992. The Intertidal fish fauna of the west coast of South Africa — species, community and biogeographic patterns. South African Journal of Zoology. 27(3):115–120. doi:10.1080/02541858.1992.11448271.

Rašić G, Schama R, Powell R, Maciel-de Freitas R, Endersby-Harshman NM,

Filipović I, Sylvestre G, Máspero RC, Hoffmann AA. 2015. Contrasting genetic structure between mitochondrial and nuclear markers in the dengue fever mosquito from Rio de Janeiro: implications for vector control. Evol Appl. 8(9):901–915. doi:10.1111/eva.12301.

Rochette NC, Catchen JM. 2017. Deriving genotypes from RAD-seq short-read data using Stacks. Nature Protocols. 12(12):2640–2659. doi:10.1038/nprot.2017.123.

Rochette NC, Rivera-Colón AG, Catchen JM. 2019. Stacks 2: Analytical methods for paired-end sequencing improve RADseq-based population genomics. Molecular Ecology. 28(21):4737–4754. doi:10.1111/mec.15253.

Rouault M, Penven P, Pohl B. 2009. Warming in the Agulhas Current system since the 1980's. Geophys Res Lett. 36(12):L12602. doi:10.1029/2009GL037987.

Rouault M, Pohl B, Penven P. 2010. Coastal oceanic climate change and variability from 1982 to 2009 around South Africa. African Journal of Marine Science. 32(2):237–246. doi:10.2989/1814232X.2010.501563.

Ruan J, Li H. 2020. Fast and accurate long-read assembly with wtdbg2. Nat Methods. 17(2):155–158. doi:10.1038/s41592-019-0669-3.

Sanger F, Air GM, Barrell BG, Brownt NL, Coulson AR, Fiddes C, Hutchison CA. 1977. Nucleotide sequence of bacteriophage. 265:9.

Sayers EW, Bolton EE, Brister JR, Canese K, Chan J, Comeau DC, Connor R, Funk K, Kelly C, Kim S, et al. 2022. Database resources of the national center for biotechnology information. Nucleic Acids Research. 50(D1):D20–D26. doi:10.1093/nar/gkab1112.

Selkoe KA, Gaggiotti OE, ToBo Laboratory, Bowen BW, Toonen RJ. 2014. Emergent patterns of population genetic structure for a coral reef community. Mol Ecol. 23(12):3064–3079. doi:10.1111/mec.12804.

Selkoe KA, Gaggiotti OE, Treml EA, Wren JLK, Donovan MK, Hawai'i Reef Connectivity Consortium, Toonen RJ. 2016. The DNA of coral reef biodiversity: predicting and protecting genetic diversity of reef assemblages. Proceedings of the Royal Society B: Biological Sciences. 283(1829):20160354. doi:10.1098/rspb.2016.0354.

Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics. 31(19):3210–3212. doi:10.1093/bioinformatics/btv351.

Sintayehu DW. 2018. Impact of climate change on biodiversity and associated key ecosystem services in Africa: a systematic review. Ecosystem Health and Sustainability. 4(9):225–239. doi:10.1080/20964129.2018.1530054.

Slatko BE, Gardner AF, Ausubel FM. 2018. Overview of Next-Generation Sequencing Technologies. Current Protocols in Molecular Biology. 122(1). doi:10.1002/cpmb.59. [accessed 2022 Aug 1]. https://onlinelibrary.wiley.com/doi/10.1002/cpmb.59.

Stepien CA. 1986. Regulation of color morphic patterns in the giant kelpfish, Heterostichus rostratus Girard: genetic versus environmental factors. Journal of Experimental Marine Biology and Ecology. 100(1):181–208. doi:10.1016/0022-0981(86)90162-0.

Stepien CA. 1987. Color Pattern and Habitat Differences Between Male, Female and Juvenile Giant Kelpfish (Blennioidei: Clinidae). [accessed 2019 Sep 16]. https://www.ingentaconnect.com/content/umrsmas/bullmar/1987/00000041/00000001/art00005.

Stepien CA. 1992. Evolution and Biogeography of the Clinidae (Teleostei: Blennioidei). Copeia. 1992(2):375. doi:10.2307/1446198.

Stepien CA, Glattke M, Fink KM. 1988. Regulation and Significance of Color Patterns of the Spotted Kelpfish, Gibbonsia elegans Cooper, 1864 (Blennioidei: Clinidae). Copeia. 1988(1):7–15. doi:10.2307/1445916.

Stepien CA, Rosenblatt RH. 1991. Patterns of Gene Flow and Genetic Divergence in the Northeastern Pacific Clinidae (Teleostei: Blennioidei), Based on Allozyme and Morphological Data. Copeia. 1991(4):873. doi:10.2307/1446084.

Teske, Papadopoulos, Griffiths, Edkins. 2007. Morphological and genetic analyses suggest that southern African crown crabs, Hymenosoma orbiculare, represent five distinct species. Crustaceana. 80(6):667–683. doi:10.1163/156854007781360694.

Teske PR, Von der Heyden S, McQuaid CD, Barker NP. 2011. A review of marine phylogeography in southern Africa. S Afr J Sci. 107(5/6):11 pages. doi:10.4102/sajs.v107i5/6.514.

Toews DPL, Brelsford A. 2012. The biogeography of mitochondrial and nuclear discordance in animals. Molecular Ecology. 21(16):3907–3930. doi:10.1111/j.1365-294X.2012.05664.x.

Toms JA, Compton JS, Smale M, von der Heyden S. 2014. Variation in palaeo-shorelines explains contemporary population genetic patterns of rocky shore species. Biol Lett. 10(6):20140330. doi:10.1098/rsbl.2014.0330.

Unmack PJ, Adams M, Hammer MP, Johnson JB, Gruber B, Gilles A, Young M, Georges A. 2021. Plotting for change: an analytical framework to aid decisions on which lineages are candidate species in phylogenomic species discovery. Biological Journal of the Linnean Society.:blab095. doi:10.1093/biolinnean/blab095.

Veith WJ. 1979. Reproduction in the live-bearing teleost *Clinus Superciliosus*. South

African Journal of Zoology. 14(4):208–211. doi:10.1080/02541858.1979.11447671.

Wagner CE, Keller I, Wittwer S, Selz OM, Mwaiko S, Greuter L, Sivasundar A, Seehausen O. 2013. Genome-wide RAD sequence data provide unprecedented resolution of species boundaries and relationships in the Lake Victoria cichlid adaptive radiation. Mol Ecol. 22(3):787–798. doi:10.1111/mec.12023.

Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, Cuomo CA, Zeng Q, Wortman J, Young SK, et al. 2014. Pilon: An Integrated Tool for Comprehensive Microbial Variant Detection and Genome Assembly Improvement. Wang J, editor. PLoS ONE. 9(11):e112963. doi:10.1371/journal.pone.0112963.

Walters AD, Cannizzaro AG, Trujillo DA, Berg DJ. 2021. Addressing the Linnean shortfall in a cryptic species complex. Zoological Journal of the Linnean Society. 192(2):277–305. doi:10.1093/zoolinnean/zlaa099.

Waples RS, Naish KA, Primmer CR. 2020. Conservation and Management of Salmon in the Age of Genomics. :29.

Warner PA, Oppen MJH van, Willis BL. 2015. Unexpected cryptic species diversity in the widespread coral Seriatopora hystrix masks spatial-genetic patterns of connectivity. Molecular Ecology. 24(12):2993–3008. doi:10.1111/mec.13225.

Waterhouse RM, Seppey M, Simão FA, Manni M, Ioannidis P, Klioutchnikov G, Kriventseva EV, Zdobnov EM. 2018. BUSCO Applications from Quality Assessments to Gene Prediction and Phylogenomics. Molecular Biology and Evolution. 35(3):543–548. doi:10.1093/molbev/msx319.

Weiskopf SR, Rubenstein MA, Crozier LG, Gaichas S, Griffis R, Halofsky JE, Hyde KJW, Morelli TL, Morisette JT, Muñoz RC, et al. 2020. Climate change effects on biodiversity, ecosystems, ecosystem services, and natural resource management in the United States. Science of The Total Environment. 733:137782. doi:10.1016/j.scitotenv.2020.137782.

Wilgenbusch JC, Swofford D. 2003. Inferring Evolutionary Trees with PAUP*. Current Protocols in Bioinformatics. 00(1). doi:10.1002/0471250953.bi0604s00. [accessed 2022 Aug 3]. https://onlinelibrary.wiley.com/doi/10.1002/0471250953.bi0604s00.

Williams ST, Knowlton N, Weigt LA, Jara JA. 2001. Evidence for Three Major Clades within the Snapping Shrimp Genus Alpheus Inferred from Nuclear and Mitochondrial Gene Sequence Data. Molecular Phylogenetics and Evolution. 20(3):375–389. doi:10.1006/mpev.2001.0976.

Wright D, Bishop JM, Matthee CA, von der Heyden S. 2015. Genetic isolation by distance reveals restricted dispersal across a range of life histories: implications for biodiversity conservation planning across highly variable marine environments. Burridge C, editor. Diversity and Distributions. 21(6):698–710.

doi:10.1111/ddi.12302.

Wyatt TD. 2003. Pheromones and animal behaviour: communication by smell and taste. Cambridge, UK ; New York: Cambridge University Press.

Zhang D-X, Hewitt GM. 2003. Nuclear DNA analyses in genetic studies of populations: practice, problems and prospects. Mol Ecol. 12(3):563–584. doi:10.1046/j.1365-294X.2003.01773.x.