Spring 2021

A Test of Rad Capture Sequencing on Ethanol-Preserved Centennial and Contemporary Specimens of Philippine Fishes

Madeleine I. Kenton
Old Dominion University, mkent001@odu.edu

A TEST OF RAD CAPTURE SEQUENCING ON ETHANOL-PRESERVED CENTENNIAL AND CONTEMPORARY SPECIMENS OF PHILIPPINE FISHES

by

Madeleine I. Kenton
B.S. May 2016, University of Tampa

A Thesis Submitted to the Faculty of
Old Dominion University in Partial Fulfillment of the
Requirements for the Degree of

MASTER OF SCIENCE

BIOLOGY

OLD DOMINION UNIVERSITY
May 2021

Approved by:

Kent E. Carpenter (Director)

Daniel Barshis (Member)

Christopher E. Bird (Member)

ABSTRACT

A TEST OF RAD CAPTURE SEQUENCING ON ETHANOL-PRESERVED CENTENNIAL AND CONTEMPORARY SPECIMENS OF PHILIPPINE FISHES

Madeleine I. Kenton
Old Dominion University, 2021
Director: Dr. Kent E. Carpenter

Understanding the relationship between ecological characteristics and genetic change in natural populations in different time scales can reveal how anthropogenic stressors affect natural populations and can improve the success of conservation strategies. The purpose of the Philippines Partnerships for International Research and Education (PIRE) project is to examine levels of genetic change between historical fish samples collected by the USS *Albatross* expedition in the early 1900s in the Philippines and contemporary populations collected at the same localities. This study tests genetic protocols to process historical and contemporary DNA for simultaneous comparison. Two DNA library preparation methods, single digest RADseq ("un-baited" sequences) and Rapture or capture probes designed from the initial RADseq tags ("baited" sequences), and two filtering pipelines, dDocentHPC and ANGSD are tested using four fishes with different life history traits. Sequencing RADseq libraries produced a range of contigs from contemporary and historic DNA across species. Sequencing baited libraries did not improve the depth of coverage for either *Albatross* or contemporary results. However, the ANGSD pipeline did improve our ability to work with and conduct analyses on the resulting low-coverage data, unlike dDocentHPC where fewer sequences passed all respective filters. This study was successful in providing the first assessment of sequencing and bioinformatics

methodologies and paves the way for developing methods to improve data that can be obtained from the historical *Albatross* specimens for future PIRE project research.

This thesis is dedicated to my parents Jeffrey and Karen Kenton, who have provided me with endless amounts of support and encouragement, no matter what goals I set out to achieve.

"Real Change, enduring change, happens one step at a time"

-Ruth Bader Ginsburg

ACKNOWLEDGMENTS

I would like to thank my advisor, Dr. Kent Carpenter for his invaluable direction and support throughout my graduate school experience. It was such a privilege to work under one of the world's leading Ichthyologists. The experience to work in his lab led me to so many amazing experiences both at ODU and in the Philippines. I have learned so many valuable lessons over the past 3 years. I would also like to thank my committee members Dr. Daniel Barshis and Dr. Christopher Bird for their support and for all of the trainings in laboratory techniques and computing.

I would also like to thank members of both the molecular systematics lab and the GMSA lab, who have provided me with amazing support and encouragement over the past 3 years. I owe a majority of my knowledge of molecular laboratory techniques to Dr. Amanda Ackiss and Ellen Biesack, who were members of the Carpenter Lab when I first started. I would especially like to thank Dr. Eric Garcia, the postdoc of the Carpenter Lab for his constant encouragement and hours spent on zoom guiding me through the entire process of this thesis.

A special thanks to all of the PIRE project principal investigators; Dr. Kent Carpenter, Dr. Beth Polidoro, Dr. Malin Pinsky, Dr. Daniel Barshis, and Dr. Chris Bird. I learned something new and incredible from each mentor and I am honored to have worked with them all. The Filipino participants on the projects supported our learning and experiences in the Philippines, Dr. Angel Alcala, Dr. Mudjekeewis Santos, and Dr. Richard Muallil, assisted in our collections and aided all of our efforts in the Philippines. I would especially like to thank Abner Bucol for teaching us so much about the culture of the Philippines and for making sure that every aspect of the project that operated in there ran without a hitch.

Dr. Jeffrey Williams provided me with incredible experiences with the Smithsonian Institution, which led me to the field of molecular systematics and this position with Dr. Carpenter. René Clark, a PhD candidate from Rutgers, provided me with much needed support and friendship, both in the Philippines and back in the United States. I would not have been able to do all of this without the never-ending support of my family and friends.

The main funding for this project came from the National Science Foundation's Partnerships in International Research and Education grant (OISE-1743711). *Albatross* specimens for the project were loaned to us by the U.S. Smithsonian Institution National Museum of Natural History Collections with the incredible help of everyone in the Division of Fishes.

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

INTRODUCTION

The National Science Foundation funded Philippines Partnerships for International Research and Education (PIRE) initiative investigates novel scientific questions about the evolutionary impacts of marine overexploitation and habitat loss. Comparing DNA from historical tissues housed in the Smithsonian Institution's National Museum of Natural History's (NMNH) collections to present-day DNA samples from corresponding populations, the project aims to reveal changes in genetic diversity of marine fishes of the Philippines that took place over the past century when substantial human impacts occurred. The current study consists of an assessment of different molecular and bioinformatics techniques to establish a successful pipeline to reach the overall objectives of this PIRE project.

The Albatross Expedition

The NMNH houses one of the greatest ichthyology collections in the world. This collection contains more than 6 million ethanol preserved specimens and a wide variety of osteological preparations and tissues preserved for genetic analyses. The largest accession ever made by the museum's fish collection includes the specimens acquired by the expedition of the U.S. Research Vessel *Albatross* (hereafter referred to as the *Albatross*). Over the course of just two years — 1907 to 1909 — the voyage of the *Albatross* resulted in the acquisition of 91,000 fish specimens (hereafter referred to as the *Albatross* specimens) contained in 28,440 cataloged single species, single locality jars or 'lots' (Smith & Williams, 1999). On this voyage the *Albatross* spent most its time exploring the natural resources of the Philippines (Smith & Williams, 1999).

The Philippines is located at the apex of the "Coral Triangle" (Allen & Werner 2002). This region is positioned along the equator, between the Indian and Pacific Oceans and includes the countries of Indonesia, Malaysia, the Philippines, Papua New Guinea, Timor-Leste, and Solomon Islands (Asaad *et al*., 2018). This area is a global hotspot of marine biodiversity and contains over 2,600 species of reef fishes (Tornabene *et al*., 2015). Out of the 6 countries that constitute the Coral Triangle, the Philippine archipelago serves as the epicenter of the world's marine biodiversity, containing more marine species per unit area than anywhere else on Earth (Carpenter & Springer, 2005).

In the Philippines, the extensive biodiversity does not only serve as a point of pride, but also substantially contributes to ecosystem services (Tamayo et al. 2018; Pinheiro et al. 2019). Many communities benefit from fisheries (both commercial and artisanal) and marine eco-tourism (White *et al*., 2000). However, Philippine marine ecosystems are also known to be some of the most impacted by anthropogenic stressors (Roberts *et al.*, 2002; Nanola et al., 2010). With the number and intensity of these stressors constantly on the rise, it is important to trace how the genetic variation of natural populations is affected as this can directly influence conservation and management efforts.

Recent advances in molecular genetic approaches allow us to closely study populations and the origins of biodiversity. Similarly, new molecular techniques have improved our ability to contrast historical DNA from museum specimens with present-day samples (Wandeler *et al*., 2007). The use of specimens from the *Albatross* collection offers the Philippines PIRE project the unique opportunity to investigate how anthropogenic impacts have affected marine species over the past century in the epicenter of marine biodiversity.

In many museums around the world, preserved specimens such as those that make up the *Albatross* collection, are often stored for long periods of time. However, if the storage conditions are not closely monitored and the preservation method is not ideal, it is not likely that they will be good candidates for molecular analysis (Chakraborty *et al*., 2006). One of the most important details concerning the *Albatross* collection is that all specimens were fixed and preserved in ethanol (Smith & Williams, 1999). This is an important distinction to make because currently the most common method of fixing fresh specimens is with formalin. Formalin is known to cause significant alterations to DNA making it challenging to obtain viable genetic material from many archival natural history collections (Chakraborty *et al*., 2006; Baloglu *et al*., 2007). However, ethanol as a method of fixation and preservation leads to significantly less DNA damage over time when compared to other common options (Chakraborty *et al*., 2006; Shiozawa *et al*., 1992). Over the past century therefore, the *Albatross* specimen's DNA molecules will have sustained less damage than those from similar collections, making them potential candidates for molecular analysis.

In this study, I explore a suite of Next Generation Sequencing (NGS) and bioinformatics techniques in order to assess strategies for successfully sequencing historic *Albatross* fish DNA with a depth of coverage that would allow us to detect the level of genetic change in response to anthropogenic stress. I first examine the performance of a common pipeline, Restriction-site Associated DNA sequencing (RADseq; Miller *et al.*, 2007; Baird *et al.*, 2008) and dDocentHPC (https://github.com/cbirdlab/dDocentHPC; a variation of dDocent, Puritz et al., 2014), on providing contemporary data that can be associated with species habitat preference. Subsequently, I compare the performance of this pipeline on both contemporary and historic data, with that of Rapture (RADseq) and Capture, (Ali *et al.*, 2016) a sequencing approach that

increases depth of coverage and ANGSD (Korneliussen *et al*., 2014), a bioinformatics software specifically designed for low coverage data.

Next Generation Sequencing and the use of RADseq

Next generation sequencing platforms perform massively parallel sequencing producing millions of fragments of DNA (Grada & Weinbrecht, 2013). These platforms generate large amounts of data but often produce high sequence error rates at the same time (Korneliussen *et al*, 2014). One method to reduce such error rates and further reduce sequencing costs is to employ genome reduction techniques (Hoffberg *et al*, 2016). Restriction-site Associated DNA sequencing (RADseq) is one such technique, which reduces the genome by sequencing thousands of DNA fragments located near specific restriction enzyme cut sites (Miller *et al*., 2007; Baird *et al*., 2008; Davey *et al*., 2011).

The RADseq methodology employed during the library preparation stage relies on a restriction enzyme digestion and a Polymerase Chain Reaction (PCR) step to provide high-resolution population genomic data at low cost (Shafer *et al*., 2017). Not only can RADseq be successful with a minimal amount of starting material but a reference genome is not required, and a wide variety of population genomic approaches such as outlier scans, linkage mapping, and demographic analyses can be conducted (Shafer *et al.*, 2017). As a result, this methodology has become a common and important component of ecological and evolutionary studies.

To explore the effectiveness of the Philippines PIRE project's proposed methods, this study began by sequencing single Sbf1-digest RADseq libraries of contemporary specimens on an Illumina platform and filtered the output using dDocentHPC. This was conducted to analyze our ability to compute and compare population genetic and neutrality test statistics in a total of

four marine fishes classified into two groups with distinct habitat preference: a demersal group including *Siganus spinus* and *Ambassis urotaenia,* and a near shore pelagic group consisting of *Spratelloides delicatulus* and *Atherinomorus endrachtensis.* These fishes are representative of the different types of species that will be used in the wider PIRE project (Table 1). *Spratelloides delicatulus* and *Atherinomorus endrachtensis* were only utilized in this first objective of the study since we did not have sequence data from their respective *Albatross* counterparts.

Study Species and Life History Characteristics

Table 1 Life history characteristics of studied species

Species (code)	Life History Trait			Reference
	Feeding Type	Depth distribution	Habitat preference	
Ambassis urotaenia (Aur)	Zooplanktivore	Benthic	Mostly river mouths/in brackish waters, Amphidromous	Need ref
Siganus spinus (Ssp)	Herbivorous, diurnal feeders	Benthic	Marine, reef-associated	Need ref
Atherinomorus endrachtensis (Aen)	Zooplanktivore	Semi-pelagic	Marine, brackish; reef associated, lagoons and inner parts of reefs	Need ref
Spratelloides delicatulus (Sde)	Planktivore	Semi-pelagic	Marine, reef-associated lagoons and along costal margins	Need ref

The Little Spinefoot, *Siganus spinus,* are demersal, marine, reef-associated fish in the family Siganidae (Laviña & Alcala, 1974). This family is distinguished by the presence of venomous spines. Siganids are widely distributed throughout the tropical, subtropical, and temperate Indo-West Pacific region and the Indian Ocean (Iwamoto *et al*., 2009). Both adults and juveniles are primarily diurnal feeders. They feed almost continuously on algae and other benthic plants during the daytime (Soliman *et al*., 2010). They are often found in small schools but may browse individually or in pairs, sometimes accompanied by other siganids, scarids, and acanthurids. This species has a planktonic larval duration (PLD) of 17 days, a restricted settlement period of 1–3 days, and spawns on or around the new moon (Harahap *et al*., 2002). *Siganus spinus* are economically important and attract attention from the aquaculture industry due to their quick growth, herbivorous lifestyle and high commercial value (Randall *et al*., 1990). Additionally, siganids constitute one of the more important food resources for local consumption in many small island nations, such as the Philippines (Laviña & Alcala, 1974). They are typically fished by spearing or throw-net with the aid of a flashlight at night.

Ambassis urotaenia is in the family Ambassidae, which are known as the "Asiatic Glassfishes" and are distinguished by their transparent bodies (Martin & Heemstra, 1988). *Ambassis* is a genus of closely related species, which inhabit the tropical and sub-tropical coastal waters and estuaries of the Indo-Pacific (Martin & Blaber, 1983). In general, *Ambassis* species are demersal zooplanktivorous occurring in schools (Martin & Blaber, 1983). They are mainly found in brackish water at the mouths of rivers, and typically amphidromous, migrating from salt water to freshwater streams (Riede, 2004).

Spratelloides delicatulus is in the Clupeidae family, which includes the herrings, shads, sardines, and menhadens (Mohan & Kunhikoya, 1985). This is a near shore pelagic marine

species that is associated with coastal reefs and lagoons, and typically occurs in small schools that feed near the surface on zooplankton (Mohan & Kunhikoya, 1985). These fish are an important part of artisanal fisheries in the Philippines, served either dried and salted or fried. They also serve as an important baitfish for the tuna fishing industry throughout the Indo-Pacific region (Jones, 1960). This species has a very short life span of around four months (Milton et al. 1991) and the occurrence of juveniles for a longer period also suggests that *S. delicatulus* may spawn more than once in a spawning season (Mohan & Kunhikoya, 1985.

Atherinomorus endrachtensis is a member of the family Atherinidae, which are known for a distinctive silver stripe that runs horizontally near their lateral line (Kimura *et al*., 2001). This species is nearshore pelagic, associated with marine and brackish waters, and inhabit lagoons and reefs but are rarely seen along the open coast (Ivantsoff and Crowley, 2000). *Atherinomorus endrachtensis* is a zooplanktivore that tends to occur in schools (Kimura *et al*., 2001). Atherinids are known to have demersal eggs (Takemura *et al*., 2004). More than 27 species of marine atherinid fishes are found in the Indo-Pacific (Ivantsoff, 1984; Ivantsoff and Crowley, 2000).

The genetic makeup of these species is compared in order to determine if genetic patterns can be associated with habitat usage and to explore the variety of population genetic signatures that are likely to be encountered in the wider PIRE project. I predict that similar patterns of heterozygosity and nucleotide diversity will be observed within the species that share habitat preferences.

Comparison of RADseq and Rapture Methodologies

Previous preliminary results from the Philippines PIRE project revealed that sequences from historical species yielded low numbers of contigs with data. In order to increase the effectiveness of our sequencing efforts, a method known as Rapture (Capture from initial RADseq libraries, Ali *et al*, 2016) was performed for both *Albatross* and Contemporary specimens of *A. urotaenia* and *S. spinus* to provide uniform, comparable sets of data.

Rapture separates RAD tag isolation and sequencing library preparation into two distinct steps and uses an in-solution capture of chosen RAD tags to target the sequencing of desired loci (Ali *et al*, 2016). This RAD methodology combines the benefits of both RAD and sequence capture into a very inexpensive and rapid library preparation that can include many individuals as well as high specificity in the number and location of genomic loci analyzed. It also tends to result in higher recovery of more unique (nonclonal) RAD fragments than other RAD protocols (Ali *et al*, 2016). The type of RAD data typically produced with Rapture was expected to provide an adequate coverage and amount of single nucleotide polymorphisms (SNPs) to detect for instance, fishing-induced declines in genetic diversity (Pinsky & Palumbi 2014).

The second aim of this study was to compare RADseq ("unbaited" sequences) and Rapture ("baited" sequences) methodologies for *Albatross* and contemporary specimens of *A. urotaenia* and *S. spinus*. Baited sequences are expected to show an increased depth of coverage with higher number of sites and contigs remaining after filtering than unbaited sequences (Peñalba *et al*, 2014).

Comparison of the Filtering Pipelines ANGSD and dDocentHPC

Low coverage data can also be optimized by choosing the appropriate data analysis pipeline. An adaptation of dDocent (Puritz *et al*., 2014) called dDocentHPC

(https://github.com/cbirdlab/dDocentHPC) was utilized for quality trimming, de novo reference assembly, mapping, and variant calling to compare contemporary populations with adequate coverage. The dDocentHPC pipeline results can be compared to the ANGSD pipeline results, which is designed to be useful for low coverage data and for non-model organisms that lack a reliable reference population (Korneliussen *et al.*, 2014). ANGSD is intended as a novel and efficient program that allows user-friendly access to methods for population genetics while working directly on *de novo*-estimated genotype likelihoods (GL). ANGSD is unique in that it allows different types of input data, however, to run all of the available analyses the input must be sequence data. It is also noteworthy because it enables users to perform a large number of common population genetic analyses (Durvasula *et al.*, 2016).

Both unbaited and baited genetic output for *S. spinus* and *A. urotaenia* were filtered using ANGSD and dDocentHPC, in order to compare their output and ability to compute analyses from sequence data. Given that ANGSD is tailored to maximize the output from low coverage data, this study hypothesizes that as opposed to dDocentHPC, results from ANGSD will generate a considerably higher number of sequences after filters and consequently, more metrics will be available to analyze focal species.

It is important to optimize methods for molecular studies as there are many variations in methodologies and the type of input data used can cause optimization strategies to vary dramatically by study (O'Leary *et al*., 2018). Methods that provide flexibility in the number of loci and individuals analyzed are necessary to facilitate effective genetic analysis (Ali *et al*, 2016). The findings of this study will further our understanding of genetic changes throughout the past century of major anthropogenic impacts. Each of the study objectives will lay the foundation for future studies on contemporary and museum specimens from several species

spanning spatial and temporal ranges as part of the larger Philippines PIRE project.

MATERIALS AND METHODS

Sampling Design

A total of four species were sampled to complete this study. From these, *A. urotaenia* and *S. spinus* were sampled and extracted from both *Albatross* and contemporary collections, while *S. delicatulus* and *A. endrachtensis* were sampled and extracted from contemporary specimens only. The sites and number sampled are displayed in Table 2 and Figure 1.

Table 2 Sampling information. Sample sites are listed by corresponding library preparation method, species, and time period of collection. The number of specimens sent for sequencing is also given. Ssp=*Siganus spinus*, Aur=*Ambassis urotaenia*, Sde=*Spratelloides delicatulus*, Aen=*Atherinomorus endrachtensis*

Library Method	Species	Time period	Collection Site	Site Code	Collection Dates	Number Sequenced
RADseq "unbaited"	Ssp	Contemporary	Albay Gulf	CGub	8-Nov-2017	52
		Albatross	Atulayan Bay	AAtu	17-Jun-1909	96
	Sde	Contemporary	Matnog Bay	CMat	8-Nov-2017	90
	Aen	Contemporary	Batangas Bay	CBat	19-Nov-2018	96
	Aur	Contemporary	Sorsogon Bay	CRag	8-Nov-2017	90
Rapture "baited"	Ssp	Contemporary	Albay Gulf	CGub	8-Nov-2017	52
		Albatross	Atulayan Bay	AAtu	17-Jun-1909	96
	Aur	Contemporary	Hamilo Cove	CHam	25-Mar-2019	96
		Albatross	Pagapas Bay	APag	20-Feb-1909	42

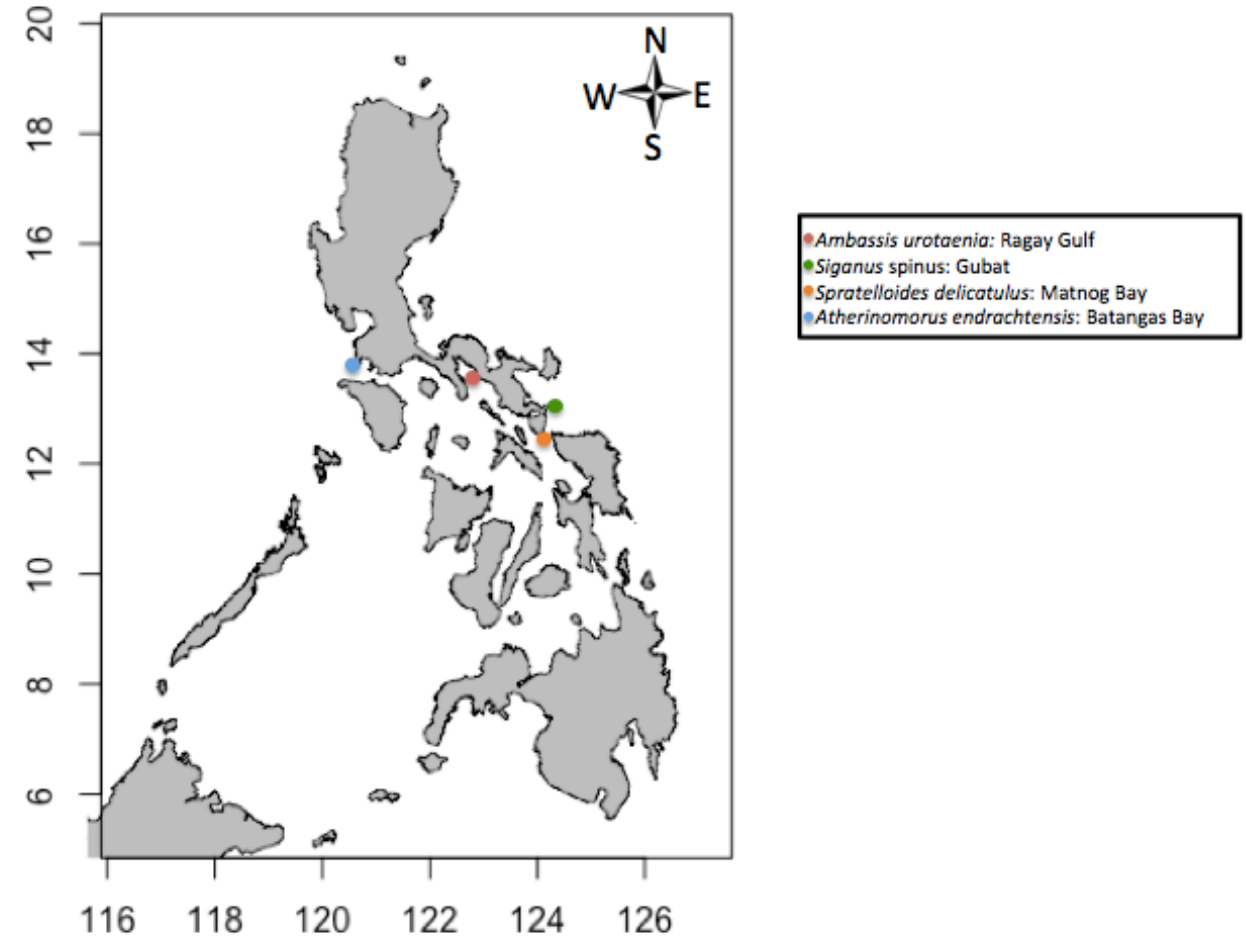

Fig. 1 Map of contemporary collection sites.

This study began by optimizing extraction and sequencing methods on contemporary specimens so that *Albatross* specimen DNA, from irreplaceable samples, would not be exhausted during testing. Contemporary collections of the four species of interest were made from sites that corresponded to existing *Albatross* collections of the same species. This methodology ensured that contemporary and *Albatross* counterparts could be compared to analyze population genetic change over the last century, which is a main focus of the Philippines PIRE Project. In order to accurately gauge the potential to reach this goal, we tested the success of proposed pipelines in

extracting and sequencing the DNA of these contemporary specimens. Each species was collected from a unique single site around the Philippines. To test our ability to successfully sequence DNA from ethanol preserved *Albatross* specimens, only single historical populations of *A. urotaenia* and *S. spinus* were processed. These two species were selected due to the success of sequencing their contemporary counterparts. The unique historical collections were borrowed from the NMNH's division of fishes collection. The *S. spinus* lot was USNM lot number 182997 and the *A. urotaenia* samples came from USNM lot number 180062.

Contemporary samples for these species were purchased from fish markets at their respective collection sites around the Philippines between 2017 and 2018. Fishes were either purchased whole from markets and landings, or fin clips were collected from vendors. Collections were made only when the original location of the harvest was verified. Specimens designated for genetic study were fixed and preserved in 95% molecular grade ethanol. Muscle tissue was subsampled using forceps, a scalpel and an alcohol lamp for sterilization.

DNA Extraction, Library Preparation and Sequencing

Muscle tissue was removed and stored in vials with 95% ethanol while whole specimens were placed into a tube with a unique identifier (so that it could be matched back to the extracted tissues) and preserved in 75% ethanol. During transport, samples were stored at room (<23°C) or refrigerated temperatures (4°C), and kept out of direct sunlight, until they could be permanently stored in a freezer (-80°C). DNA was extracted using QIAGEN DNeasy Blood and Tissue® kits with minor modifications to best accommodate both *Albatross* and Contemporary tissues. Comparison tests were run to determine optimal digestion times and amount of starting tissue. Initially results indicated that 20mg of muscle tissue, with a digestion time of 90 minutes for

Albatross specimens and 157 minutes for contemporary specimens was optimal for the highest DNA yield. This tissue amount (20mg) was extracted and utilized for all of the unbaited populations. Subsequent tests indicated that 50mg of tissue helped with low DNA yields and therefore, 50mg of tissue was extracted and utilized for all baited populations. The DNA was eluted 4 times, using 100 μL of elution buffer (buffer AE). Each DNA elution for a subset of samples were visualized via gel electrophoresis on a 1% agarose gel stained with SYBER Safe (Invitrogen, Thermo Fisher Scientific) in order to confirm high-quality extracts.

The elutions were then shipped to the Texas A&M University – Corpus Christi (TAMUCC) Genomics Core Laboratory, where RADseq libraries for Illumina sequencing were prepared. Extracted DNA was enriched for high molecular weight fragments, using Beckman-Coulter SPRI-Select paramagnetic beads. Size selection of DNA was regulated with respect to the frequency distribution of fragment lengths. The concentration of all DNA samples was quantified using a Spectramax M3 fluorescent plate reader and the Biotium AccuBlue kit.

The first libraries followed a single digest RADseq protocol using New England Biolabs SbfI-HF restriction enzyme. A biotinylated, inline barcode was ligated to digested DNA prior to sonication with a Diagenode Bioruptor in order to adjust the average DNA fragment size to ~300bp. Target biotinylated DNA was then isolated using Thermo Fisher Scientific M-280 Streptavidin Dynabeads. A second SbfI digestion was performed to remove the biotin-Dynabead complex. Illumina adapters were ligated to the samples using the KAPA Biosystems Hyper Plus DNA prep kit, as in ezRAD (Toonen *et al.*, 2013). The DNA concentration of each library was quantified using a KAPA qPCR library quantification kit on an Applied Biosystems Incorporated StepONEplus real-time thermocycler. Pooled libraries within a species were normalized and combined prior to capture.

Sequencing was completed by the Novogene facility (UC Davis, CA). The size of fragments in the final libraries were selected using a Sage Science BluePippin pulsed-field electrophoresis rig, and the DNA concentration was quantified using a KAPA qPCR library quantification kit. All libraries were sequenced using an Illumina HiSeq 4000 sequencer at a target depth of 3 million reads per individual.

Data from sequenced RADseq libraries were then bioinformatically processed (see below) to produce filtered *de novo* references for *A. urotaenia* and *S. spinus*, which were sent to Daicel Arbor Biosciences laboratories (ArborBio) where probe baits were designed for subsequent Rapture analyses. The Rapture protocol utilized custom 120bp MYcroarray MYbaits kits, where every nucleotide in each RAD locus is targeted by an average of three baits. Each kit contains custom biotinylated capture baits for one species. After the completion of the probe design, Rapture libraries were prepared by the TAMUCC Genomics Core Laboratory and sequencing was performed in the same sequencing facility as with the RADseq libraries.

Filtering and SNP Discovery

After sequencing, reads were re-associated with each sample (demultiplexing) using the process_radtags function in STACKS (Catchen *et al*., 2013). All of the following processes until the genotyping were performed within the newest version of the dDocentHPC application wrapper. Trimmomatic (Bolger *et al.*, 2014) was used to trim adapters and low-quality reads from datasets, *de novo* genome assemblies were carried out using Rainbow (Chong *et al*., 2012) for each species (since no reference genome was available for any), reads were mapped to the *de novo* reference using Burrows-Wheeler Aligner (Li & Durbin, 2009), and filtering improper pairs and PCR clones was completed using samtools (Li *et al*. 2009). Finally, the same Binary

Alignment Map (BAM) files were separately used to obtain genotype calls and likelihoods in FreeBayes (Garrison 2010) and ANGSD (Korneliussen *et al.*, 2014), respectively, for pipeline comparisons.

The dDocentHPC pipeline utilized a combination of samtools (Li *et al*. 2009), VCFtools (Danecek *et al.*, 2011) filters to parse loci and samples for minimum alternate allele depth and frequency, minimum nucleotide and mapping quality score, minimum mean read depth, missing data, and PCR clones (see Appendix A for settings). Data from individuals were then aggregated by location and time for "sample aware" filtering of loci and sampled based upon missing data, reference allele frequency in heterozygotes, strand bias, imbalanced proportions of forward and reverse reads, imbalanced mapping quality between allelic states, proper pairing, deflated locus quality scores (Li 2014), maximum mean read depth, and Hardy-Weinberg equilibrium. Haplotypes for each RAD locus were assembled using rad_haplotyper, which additionally filtered loci for paralogs, missing data, low depth of coverage, genotyping errors, and excess haplotypes. The loci filtered by rad_haplotyper were excluded from the curated Variant Call Format (VCF) files, and SNPs with more than two allelic states were also removed. From the final filtered data set, ArborBio targeted approximately 5000 loci at random from each population (after the above in-house quality control) for capture bait design. Filters and settings for each species are provided in Appendix A. In order to allow for direct comparisons, filters applied to all four contemporary species in the habitat preference analysis were optimized using the *S. delicatulus* dataset, which had the lowest number of resulting sites and contigs. In contrast, for the comparison of filtering methods as well as baited and unbaited results, filter settings were optimized for each species and time period individually (Appendices B-H). Subsequently, individuals below a threshold of contigs with data were dropped manually which generated a

secondary dataset for each comparison. During the process of making the VCF files, jobs were run with both *Albatross* and contemporary individuals when applicable and split into separate *Albatross* and contemporary runs for filtering of the VCF files.

Rapture processing followed the filtering process described above except that capture data consisted of only individuals (no pools) and the assembly of a reference genome was not necessary as baits were mapped to the original *de novo* reference created from contemporary individuals.

Genetic Diversity in relation to habitat preference

The VCF output from the dDocentHPC pipeline of all four unbaited contemporary species datasets was used to determine if genetic patterns were observed in relation to habitat preference. VCFtools (Danecek *et al.*, 2011) was run on the final VCF files produced by filtering in order to determine the mean sequencing depth and nucleotide diversity (Pi) of populations. The program PGDspider (Lischer & Excoffier, 2012) was used to convert VCF files into STRUCTURE format to calculate number of alleles (nAlleles), effective number of alleles (nEffAlleles), and heterozygosity with the program Genodive (Meirmans & Van Tienderen, 2004). PGDspider (Lischer & Excoffier, 2012) was again used to convert VCF files into FASTA format in order to calculate effective population sizes (Ne) using NeEstimator (Do *et al*., 2014).

Comparison of RADseq and Rapture Datasets

The program PGDspider (Lischer & Excoffier, 2012) was used to convert VCF files into STRUCTURE format for downstream analyses using the package “adegenet” (Jombart & Ahmed, 2011) in R (R Core Team, 2020) and Genodive (Meirmans & Van Tienderen, 2004),

and FASTA formats to provide input files for MEGA (Kumar *et al*., 2008) and NeEstimator (Do *et al*., 2014). Number of alleles (nAlleles), effective number of alleles (nEffAlleles), observed and expected heterozygosity, and inbreeding coefficient (Ho, Hs, Gis, respectively) were calculated in Genodive and the program MEGA was utilized to calculate Tajima's D. Ne estimator was utilized to calculate Ne with 95% confidence intervals. The R packages adegenet and heirfstat (Goudet, 2005) were used to compute principal component analyses (PCAs) and fixation indices (FSTs), respectively, from VCF files manually merged using tidyverse (Wickham *et al*., 2019) and custom scripts.

Comparison of dDocentHPC and ANGSD Filtering Pipelines

The pipeline ANGSD was run on the baited and unbaited BAM files generated for both *Albatross* and contemporary. The settings are listed in Appendices I and J. Filter settings for ANGSD were optimized individually for each species and library preparation method. ANGSD was used to calculate site frequency spectrum, neutrality test statistics, and FSTs between populations. Custom R scripts were utilized to calculated nucleotide diversity, Tajima's D, and principal component analyses from the ANGSD output.

RESULTS

Genetic Diversity in Relation to Habitat Preference

The *Siganus spinus* dataset produced the highest number of sites and contigs from the most individuals and had the highest number of contigs with data per individual after calculating coverage (Table 3). In contrast, *S. delicatulus* showed the least number of final sites, contigs, and final individuals (Table 3). Depth of coverage was similar in all species except for *A. endrachtensis* which had the lowest mean coverage (Table 3).

There were distinctly higher number of effective alleles and inbreeding coefficient values for the near-shore pelagic species, *A. endrachtensis* and *S. delicatulus*, than there were for the demersal species, *A. urotaenia* and *S. spinus* (Table 4). However, demersal species illustrated higher levels of observed heterozygosity even when this was expected to be lower than that of pelagic species (Table 4). There was no clear correlation between nucleotide diversity (Pi) and habitat preference. *Atherinomorus endrachtensis* had highest nucleotide diversity, while *A. urotaenia* displayed the lowest. *Spratelloides delicatulus* had the only effective population size (Ne) that was not infinite.

Table 3 Sequencing and filtering results for the four focal species**.** Final number of sites, contigs and individuals for focal species after filtering using the same set of filters optimized for *S. delicatulus*. Mean depth was calculated using VCFtools from the dDocentHPC pipeline. Ssp=*Siganus spinus*, Aur=*Ambassis urotaenia*, Sde=*Spratelloides delicatulus*, Aen=*Atherinomorus endrachtensis*.

Species	Final Sites	Final Contigs	Final Individuals	Mean Depth
Aur	9883	4723	20	47.71
Ssp	40168	11786	24	56.7
Aen	2371	1429	20	22.05
Sde	1106	584	18	49.97

Table 4 Diversity metrics for the focal 4 species. nAlleles, number of alleles; nEffAlleles effective number of alleles; Ho, observed and Hs, expected heterozygosity; Gis, inbreeding coefficient; Pi, nucleotide diversity; Ne, effective population size; and mean depth are displayed. Allele, heterozygosity, and inbreeding estimates were calculated in Genodive. Pi and Ne were calculated using VCFtools and Ne estimator, respectively. Ssp=*Siganus spinus*, Aur=*Ambassis urotaenia*, Sde=*Spratelloides delicatulus*, Aen=*Atherinomorus endrachtensis.*

Species	Habitat	nAlleles	nEffAlleles	Ho	Hs	Gis	Pi	Ne
Aur	demersal	1.981	1.322	0.221	0.217	-0.02	0.217	∞
Ssp	demersal	1.944	1.335	0.18	0.228	0.211	0.28	∞
Aen	pelagic	1.999	1.497	0.108	0.332	0.674	0.321	∞
Sde	pelagic	1.988	1.391	0.173	0.272	0.363	0.267	727

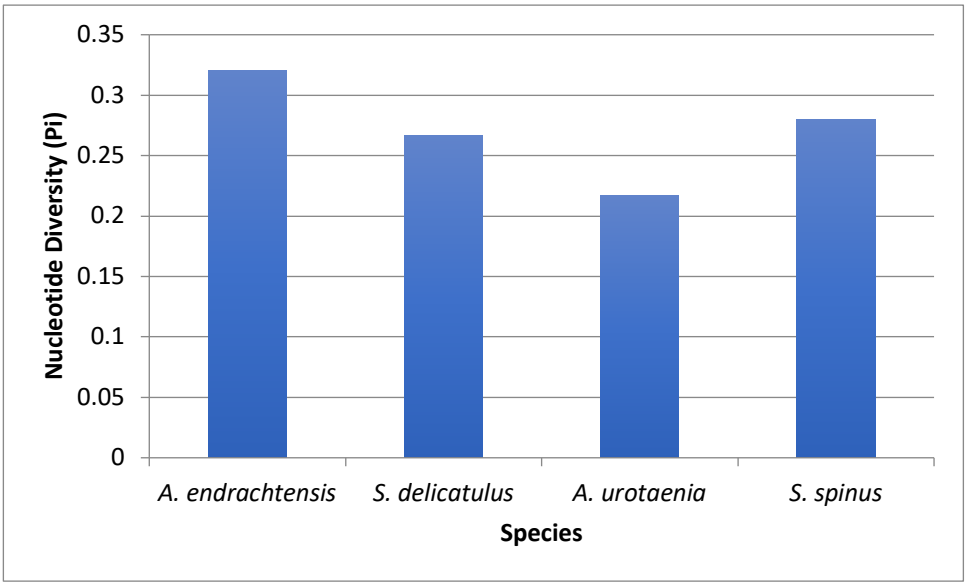

Fig. 2 Nucleotide diversity (Pi) for each of the four focal species.

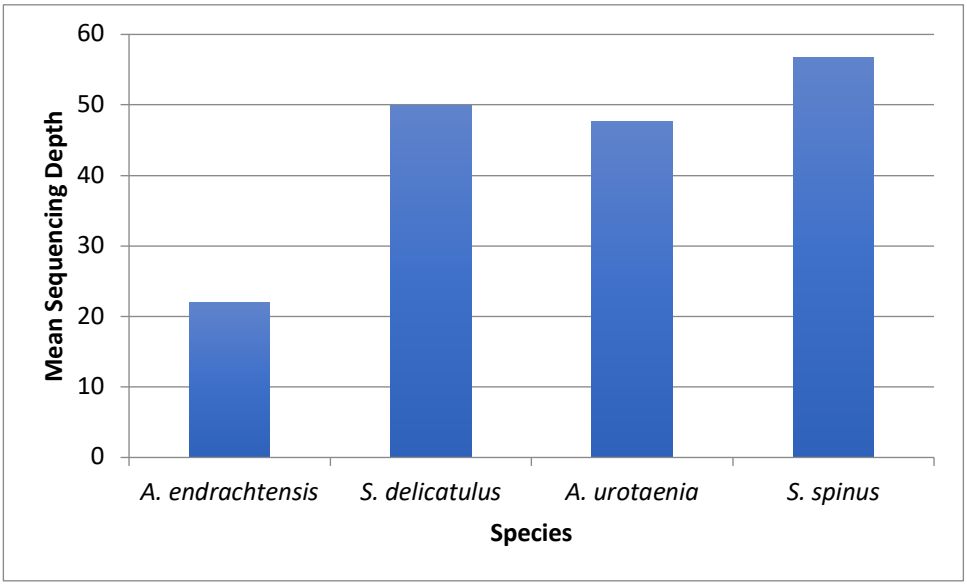

Fig. 3 Mean sequencing depth for each of the four focal species.

Comparison of RADseq and Rapture Datasets

A total of 6666 and 5047 capture baits were designed for *S. spinus* and *A. urotaenia,* respectively. However, the Rapture baited output did not perform well for any of the filter strategies used (see appendices B-H for settings). Baited *Albatross* specimens produced few or no useable sites or contigs using filter settings optimized for *S. delicatulus;* (Table 5). When filtering was optimized by species for baited *Albatross* specimens by lowering filter thresholds, additional contigs were provided (Table 6). However, filters had to be very relaxed in order to optimize *Albatross* populations and when these files were loaded into analysis programs such as Genodive, adegenet, or Ne Estimator, no useable data was present.

Principal component analyses were constructed to compare the baited and unbaited contemporary dDocentHPC output. Only the first principal component was significant for all populations (Figure 4). *Siganus spinus* and *A. urotaenia* had a very similar spread in the unbaited PCA. Similarly, the baited PCA also had a very similar spread for both species. However, there was a much wider spread in unbaited PCA than the baited ones. *Albatross* individuals could only be analyzed for PCA for unbaited *S. spinus* where they show a notable separation with their contemporary counterparts (Figure 4).

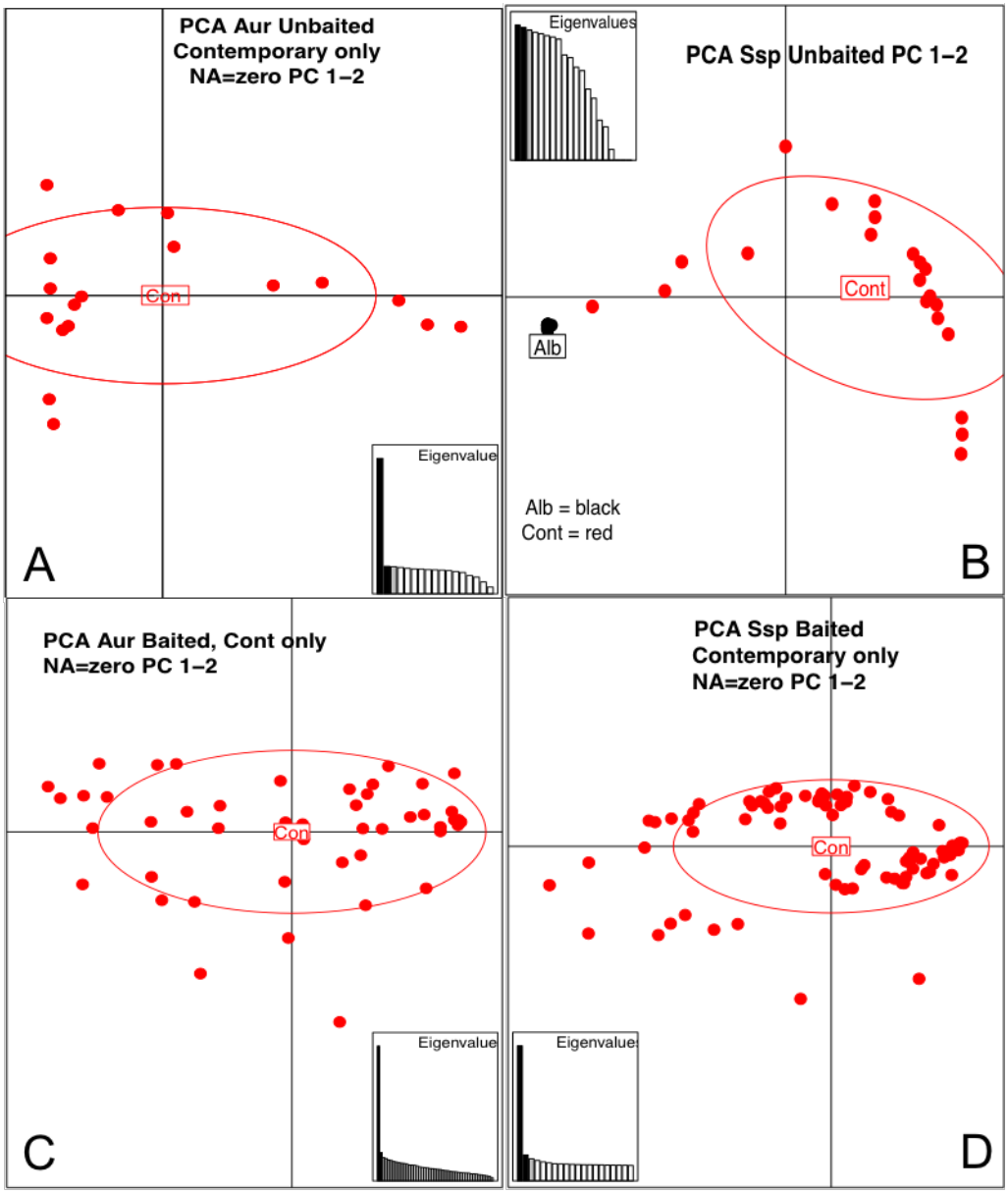

Fig. 4 Principal component analysis from the dDocentHPC output of unbaited (A, *A. urotaenia;* B, *S. spinus)* and baited datasets (C, *A. urotaenia;* D, *S. spinus).* Only principal component 1 was significant for all.

Table 5 Final sites, individuals, and contigs using filter settings optimized for *Spratelloides delicatulus* filtering for 15 individuals. (Ssp=*Siganus spinus*, Aur=*Ambassis urotaenia*, Sde=*Spratelloides delicatulus*, Aen=*Atherinomorus endrachtensis*). This includes manual individual dropping based on coverage for each population as well (see Appendix A for settings)

Library Method	Species	Time Period	Final Sites	Final Contigs	Individuals
RADseq "unbaited"	Ssp	Contemporary	40168	11786	24
		Albatross	11	9	8
	Aur	Contemporary	9883	4723	20
Rapture "baited"	Ssp	Contemporary	5808	2404	96
		Albatross	0	0	0
	Aur	Contemporary	319	126	45
		Albatross	0	0	0

Table 6 Final sites, individuals, and contigs using filter settings optimized individually for each population sequenced. (Ssp=*Siganus spinus*, Aur=*Ambassis urotaenia*, Sde=*Spratelloides delicatulus*, Aen=*Atherinomorus endrachtensis*). This includes manual individual dropping based on coverage for each population. (see Appendices B-H for settings)

Library Method	Species	Time Period	Final Sites	Final Contigs	Individuals
RADseq "Unbaited"	Ssp	Contemporary	211754	35298	21
		Albatross	2220	556	8
	Aur	Contemporary	6212	3200	21
Rapture "Baited"	Ssp	Contemporary	19621	4708	81
		Albatross	56654	12026	7
	Aur	Contemporary	621	194	46
		Albatross	2727	777	5

Comparison of dDocentHPC and ANGSD Filtering Pipelines

All *Albatross* populations had at least a few ending sites and contigs when ANGSD was utilized (Table 7). *Siganus spinus* consistently ended analysis with the highest number of

individuals, sites, and final contigs compared to *A. urotaenia* (Table 7). Filtering with dDocentHPC was highly successful for all of the unbaited data sets (Figures 5 and 6). and was more successful at producing reads for the baited *Albatross* populations than ANGSD (Tables 6 and 7). However, ANGSD was more successful in producing useable data for all analyses (Table 8). In addition, optimizing the dDocentHPC filtering required required highly relaxed settings to salvage as many contigs as possible. Overall, when compared to the dDocentHPC results (Table 6), ANGSD (Table 7) was more successful in producing analyzable reads, especially for the low coverage data provided by the populations produced with Rapture libraries.

Nucleotide diversity (Pi) and Tajima's D were higher across all dDocentHPC results when compared to ANGSD results (Table 8). Ne Estimator produced infinite Ne values for most dDocentHPC results, while the values produced by ANGSD were much smaller with bounded 95% confidence intervals. Values for FSTs were higher when produced by ANGSD and many of the values were not produced by dDocentHPC due to a lack of usable baited *Albatross* data.

The Ne calculated from ANGSD output provides the estimated population size for the temporal midpoint between the *Albatross* and contemporary populations. Therefore, they are not equal to the dDocentHPC Ne calculations and only roughly comparable. I could not directly calculate heterozygosity using the data produced by ANGSD (Table 8 list as NA).

Only the first principal component for all PCAs produced with ANGSD, was significant and there was no separation between Albatross and contemporary populations (Figures 7 to 10). Unbaited *A. urotaenia* did not have a corresponding *Albatross* population and was graphed independently. However, there was a much wider spread in baited PCA of the contemporary populations than the *Albatross* populations (Figures 9 and 10) because of the small size of the *Albatross* specimen data.

Table 7 Filtering results for the ANGSD pipeline. Final Sites, contigs, and minimum represented individuals after filtering with ANGSD. Ssp=*Siganus spinus*, Aur=*Ambassis urotaenia*, Sde=*Spratelloides delicatulus* Aen=*Atherinomorus endrachtensis*

Library Method	Species	Time Period	Final Sites	Final Contigs	Individuals
Unbaited	Ssp	Contemporary	193644	30266	20
		Albatross	876	276	6
	Aur	Contemporary	42311	7283	20
Baited	Ssp	Contemporary	93314	8041	30
		Albatross	82	20	2
	Aur	Contemporary	5091	596	30
		Albatross	18	9	2

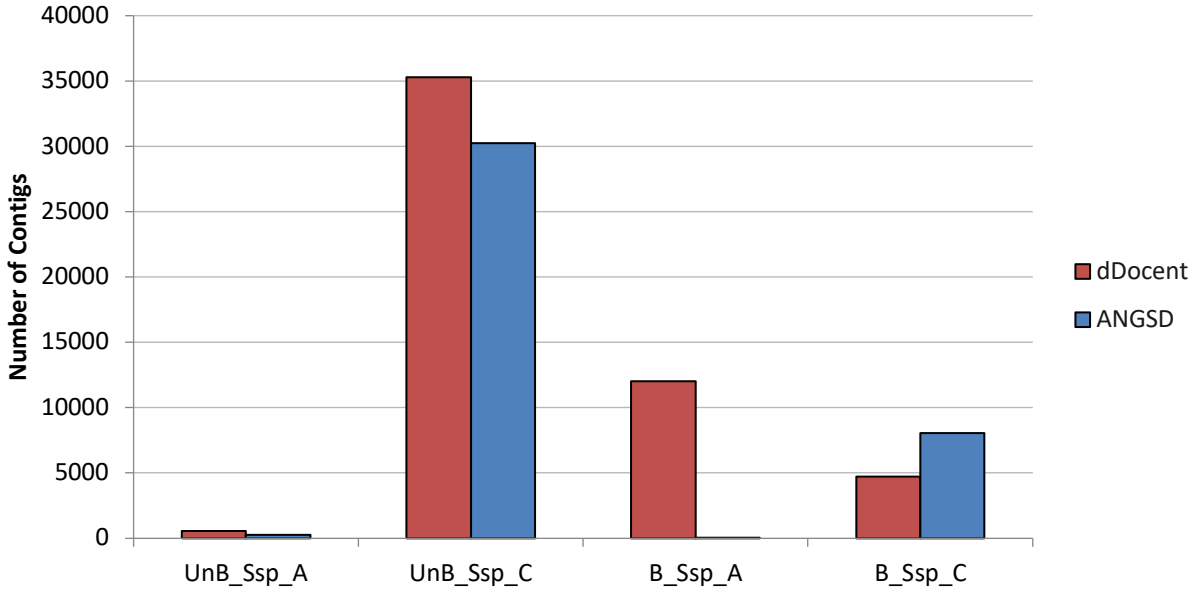

Fig. 5 Final *Siganus spinus* Contigs after Filtering using Two Different Pipelines (UnB_Ssp_A= Unbaited *Siganus spinus Albatross*, UnB_Ssp_C= Unbaited *Siganus spinus* Contemporary, B_Ssp_A= Baited *Siganus spinus Albatross*, B_Ssp_C=Unbaited *Siganus spinus* Contemporary).

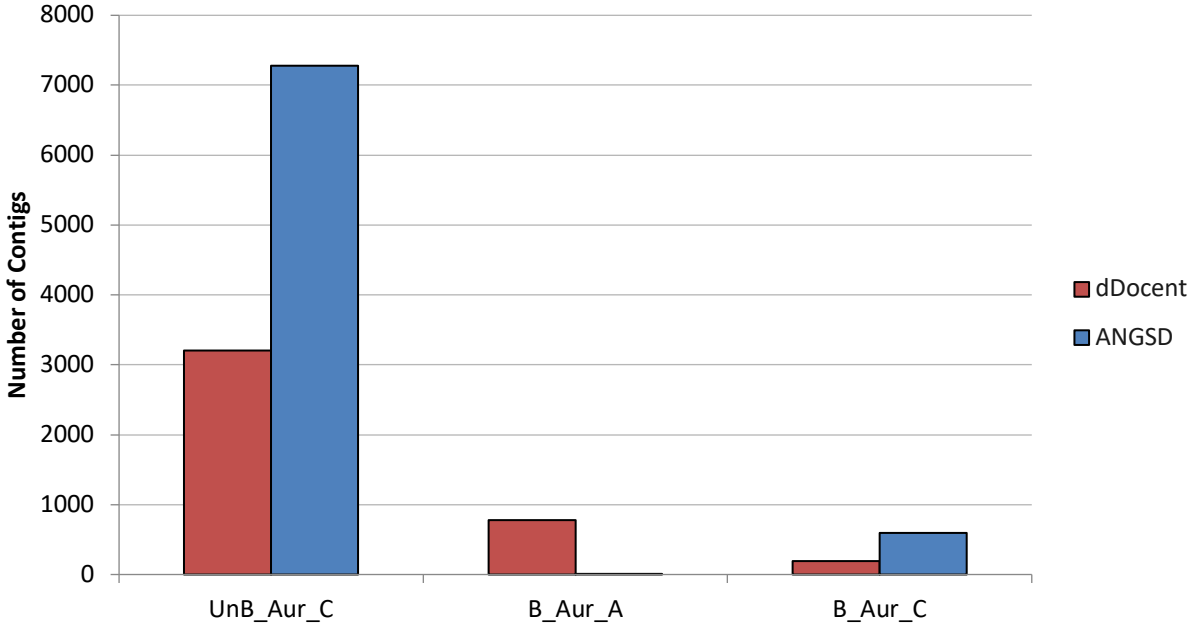

Fig. 6 Final *Ambassis urotaenia* Contigs after Filtering using Two Different Pipelines. (UnB_Aur_C= Unbaited *Ambassis urotaenia* Contemporary, B_Aur_A= Baited *Ambassis urotaenia Albatross*, B_Aur_C=Unbaited *Ambassis urotaenia* Contemporary).

Table 8 Diversity and differentiation estimates for *Siganus spinus* (Ssp) and *Ambasis urotaenina* (Aur) population datasets produced with two separate library preparations (RADseq and Rapture) and two filtering pipelines (dDocentHPC and ANGSD). Ho, observed and Hs, expected heterozygosity; Gis, inbreeding coefficient; TajD, Tajima's D; Pi, nucleotide diversity; Ne, effective population size; Ne 95% CI, effective population size with 95% confidence intervals; FST, fixation index; FST 95% CI; fixation index with 95% confidence intervals. NA refers to Not Applicable and is utilized when no data is available for the given index

Pipeline	Library Prep	Species	Time Period	Ho	Hs (He)	Gis	Pi	TajD	Ne	95% CIs	FST	95%CIs
dDocent	Unbaited	Ssp	Contemporary	0.195	0.238	0.183	0.2367	-1.51726	Infinite	Infinite-Infinite	0.0002	0 - 0.001
			Albatross	0.048	0.048	0	0.0498	-3.97658	Infinite	Infinite-Infinite		
		Aur	Contemporary	0.221	0.214	-0.03	0.2144	-2.346791	Infinite	Infinite-Infinite	NA	NA
	Baited	Ssp	Contemporary	0.189	0.19	0.007	0.1898	-1.258396	598.5	579.4-618.6	NA	NA
			Albatross	NA	NA	NA	NA	NA	Infinite	Infinite-Infinite		
		Aur	Contemporary	0.068	0.07	0.026	0.0698	-2.112697	Infinite	149.1-Infinite	NA	NA
			Albatross	NA	NA	NA	NA	NA	Infinite	Infinite-Infinite		
ANGSD	Unbaited	Ssp	Contemporary	NA	NA	NA	0.0048	-0.0064	914.9	457.5 - 5186.6	0.02	0.01 - 0.02
			Albatross	NA	NA	NA	0.0016	-0.0186				
		Aur	Contemporary	NA	NA	NA	0.0038	-0.0111	NA	NA	NA	NA
	Baited	Ssp	Contemporary	NA	NA	NA	0.0053	0.0306	50.5	22.1 - 840.5	0.24	0.16 - 0.29
			Albatross	NA	NA	NA	0.0015	-0.0443				
		Aur	Contemporary	NA	NA	NA	0.0043	-0.022	328.1	91.3 - 328.1	0.19	0.05 - 0.25
			Albatross	NA	NA	NA	0.0015	0.1096				

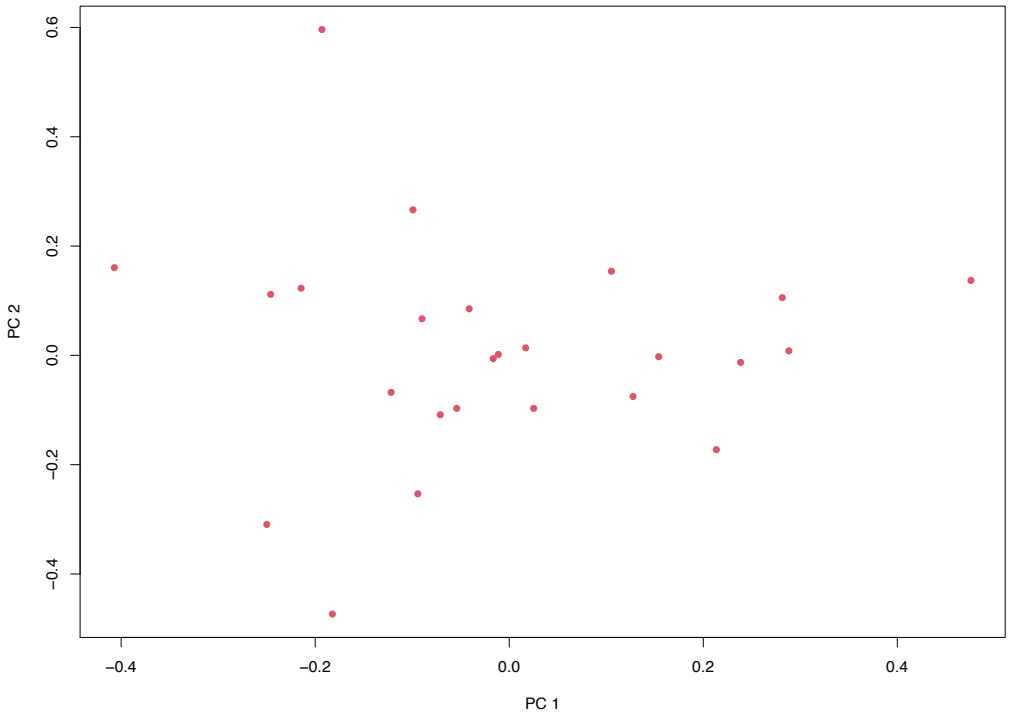

Fig. 7 Principal component analysis from the ANGSD output of unbaited *Ambassis urotaenia* contemporary individuals. Only principal component 1 was significant.

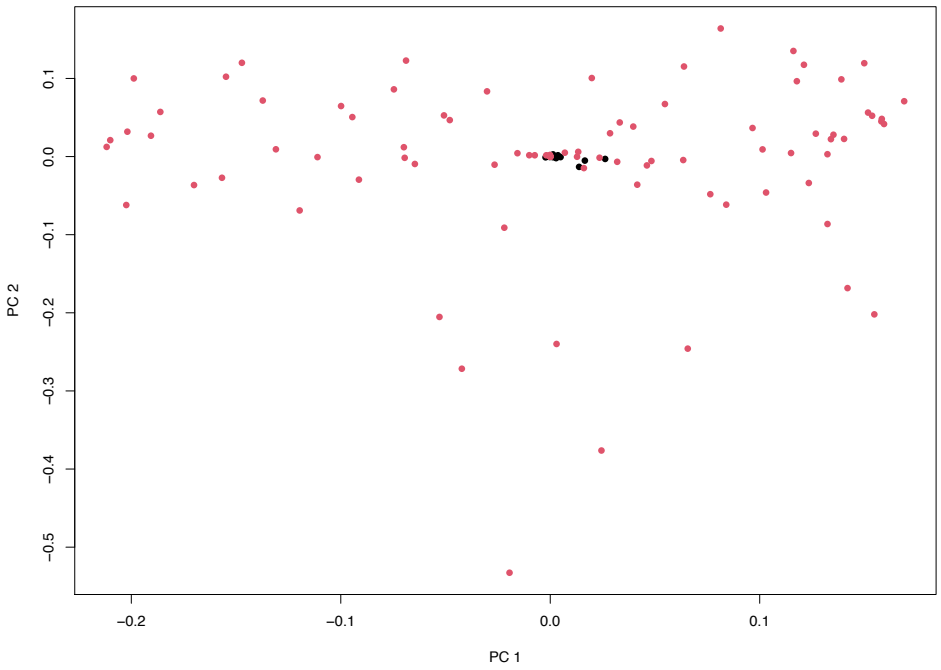

Fig. 8 Principal component analysis from the ANGSD output of unbaited *Siganus spinus Albatross* and contemporary individuals (red = *Albatross,* black = Contemporary). Only principal component 1 was significant.

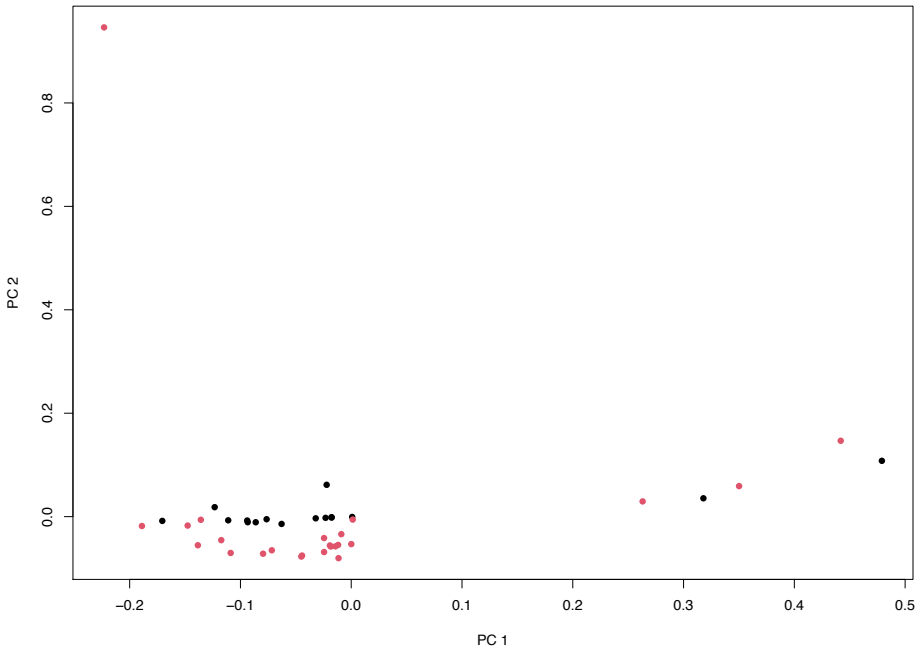

Fig. 9 Principal component analysis from the ANGSD output of baited *Ambassis urotaenia Albatross* and contemporary individuals (red = *Albatross,* black = Contemporary). Only principal component 1 was significant.

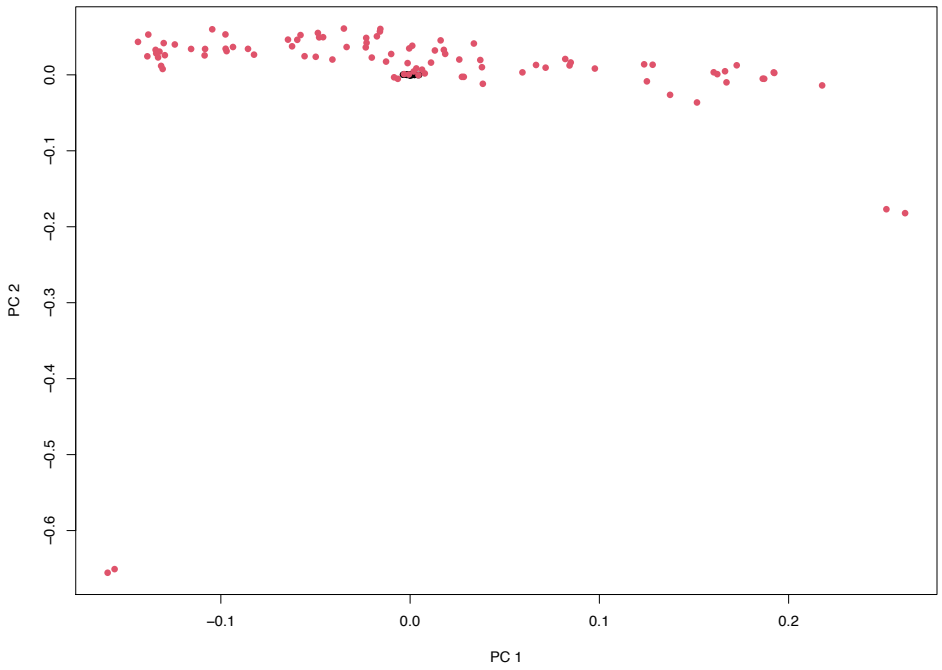

Fig. 10 Principal component analysis from the ANGSD output of baited *Siganus spinus Albatross* and contemporary individuals (red = *Albatross,* black = Contemporary). Only principal component 1 was significant.

DISCUSSION

Study of Genetic Diversity in relation to habitat preference

An objective of this study was to explore variations in population genetic signatures across species with different habitat characteristics to help understand what to expect in the larger PIRE project. The first prediction was that similar patterns of metric values would be observed within the species that share habitat preferences. This was true for values in the effective number of alleles, heterozygosity, and inbreeding coefficient, where a dichotomy of higher or lower values was observed across habitat preference. However, a cohesive picture did not emerge from this dichotomy and there was no correlation between habitat preference and nucleotide diversity or effective population size. For example, while nucleotide diversity showed high variation across species (the pelagic *A. endrachtensis* had the highest value), the demersal species illustrated higher levels of observed heterozygosity when this was expected to be lower than that of pelagic species. Overall, there are no clear patterns in life history characteristic across habitat differences. However, more populations and higher sample sizes might help increase the power in some analysis, such as Ne. Further hypotheses regarding life history characteristics need to be tested in the wider PIRE project in order to better understand this component of variation in population genetic structure.

Other potential sources of noise in the habitat comparison might have been introduced in the filtering process in our efforts to produce a direct comparison between species. The filter settings applied to all the species where the parameters needed to reach a minimum of 500 final contigs (an internal threshold) in the dataset with the lowest quality and depth (*S. delicatulus* was the species that needed the most lenient settings to reach this threshold). This could indicate that

some low-quality data may have been included to salvage the number of useable contigs for each of these species. Additional filtering strategies may also be useful for exploring potential results further.

There are also many potential changes to a molecular protocol to increase coverage in data. For example, including a whole genome amplification step or modifying RAD libraries may increase the amount and quality of SNP data produced. While RADseq has proven to be an effective tool in many studies, alternative methods of optimizing DNA size fragments and quantity will be needed to ensure that RADseq alone can be used effectively on historical DNA.

Comparison of RADseq and Rapture Library Prep Protocols

The Rapture protocol was employed to produce a smaller set of loci but with higher depth of coverage than loci produced by RADseq, and to reduce unwanted fragments that would be expected in historical DNA. Therefore, this study expected to see a higher number of sites, contigs and individuals remaining after dDocentHPC filtering in Rapture datasets. However, results were mixed for both pipelines for both contemporary and historical samples. The RADseq datasets often had more sites and contigs but less individuals than Rapture datasets. The RADseq pipeline also produced substantially more data from the two contemporary populations than from the *Albatross S. spinus,* where only a handful of contigs and individuals remained. The Rapture baits appear to have worked successfully for contemporary *Siganus spinus* but not as effectively for *Ambassis urotaenia*. Using the same filtering scheme as before (optimized for *S. delicatulus)* none of the baited *Albatross* datasets produced any reads at the end of filtering, suggesting that the implementation of Rapture protocols did not increase the effectiveness of our sequencing runs. When filtering was optimized independently, the resulting number of sites and contigs from

these datasets increased but only very few individuals passed all filters. However, the extremely lenient filter settings that were required to salvage baited sequences could have compromised the quality of these datasets. Looking at our results, there may have been problems with capture probe creation or with the sequences used to create these datasets, as the data contradicts the results of previous studies, and did not increase the effectiveness of our data. These observations might also indicate a species or collection effect in the results as *S. spinus* consistently showed higher success across treatments. A variety of library preparation methods will need to be tested in a higher number of populations and species in order to optimize the use of Rapture and see if its success rate changes in future PIRE projects.

Comparison of dDocentHPC and ANGSD Filtering Pipelines

The ANGSD pipeline was generally more effective than dDocentHPC at handling data with very low coverage, consistent with our hypothesis and other studies (Korneliussen et al. 2014). Unlike dDocentHPC, ANGSD maintained some reads, contigs and individuals for all employed filter settings and populations, including all *Albatross* populations. While dDocentHPC calculates allele frequencies from actual genotype calls, ANGSD does this from genotype likelihood scores. This was especially important for the *Albatross* files where low coverage would not have produced as much missing data as with dDocentHPC filtering. Nevertheless, in order to get contigs in our final VCF file from ANGSD for the baited *Albatross* populations, filters had to be very lenient, just as in dDocentHPC. Even with the improved abilities of ANGSD, resulting datasets were still very small and showed a large amount of missing data, indicating a need to explore other methodologies earlier in the protocol to address these limitations.

Historical DNA allows us to examine the evolution of lineages and discover population patterns over time. For the *Albatross* specimens however, many challenges remain that will require extra time and effort to optimize protocols to get adequate good quality data. In addition, alternative questions need to be tested such as whether it is likely that enough change occurred over the past century to account for the observed differences between the *Albatross* sequences and the reference obtained from contemporary specimens. However, the extra effort is fully justified given the promise of unlocking historical population and evolutionary patterns from the over 90,000 fish specimens collected by the *Albatross* from the Philippines and surrounding waters.

CONCLUSIONS

Next generation sequencing and genome reduction techniques have provided much needed capacity and versatility for gaining new insights into ecological, evolutionary and conservation questions. However, researchers should use careful consideration when choosing and applying these methods given intrinsic sources of error and bias. Similarly, optimizing methodologies can profoundly affect all steps of a genomic study, from study design and execution, to the resulting data output (Andrews *et al.,* 2016). Results from our RADseq and Rapture protocol assessment indicate a need for the PIRE project to explore alternative library preparation methods and extraction methodologies to gain higher amounts of DNA with high molecular weight. This study has already prompted the PIRE project to explore the use of Shotgun sequencing (Messing, 2001), whole genome amplification (Borgström *et al.,* 2017), and hybridization RAD (hyRAD) (Suchan *et al.,* 2016) to improve sequencing results from the historical *Albatross* specimens.

REFERENCES

Ackiss, A. S., Pardede, S., Crandall, E. D., Ablan-Lagman, M. C. A., Ambariyanto, Romena, N., ... Carpenter, K. E. (2013). Pronounced genetic structure in a highly mobile coral reef fish, *Caesio cuning*, in the Coral Triangle. *Marine Ecology Progress Series*, 480, 185-197. https://doi.org/10.3354/meps10199.

Alcala, A. C., & Russ, G. R. (2006). No-take Marine Reserves and Reef Fisheries Management in the Philippines: A New People Power Revolution. *AMBIO: A Journal of the Human Environment*, 35(5), 245-254. https://doi.org/10.1579/05-A-054R1.1.

Ali, O. A., O'Rourke, S. M., Amish, S. J., Meek, M. H., Luikart, G., Jeffres, C., & Miller, M. R. (2016). RAD Capture (Rapture): Flexible and Efficient Sequence-Based Genotyping. *Genetics*, 202(2), 389-400. https://doi.org/10.1534/genetics.115.183665.

Allen, G. R., & Werner, T. B. (2002). Coral reef fish assessment in the 'coral triangle' of southeastern Asia. *Environmental Biology of Fishes*, 65, 209-214. https://doi.org/10.1023/A:1020093012502.

Andrews, K. R., Good, J. M., Miller, M. R., Luikart, G., & Hohenlohe, P. A. (2016). Harnessing the power of RADseq for ecological and evolutionary genomics. *Nature Reviews*, 17, 81-92. https://doi.org/10.1038/nrg.2015.28.

Asaad, I., Lundquist, C. J., Erdmann, M. V., & Costello, M. J. (2018). Delineating priority areas for marine biodiversity conservation in the Coral Triangle. *Biological Conservation*, 222, 198-211. https://doi.org/10.1016/j.biocon.2018.03.037.

Baird, N. A., Etter, P. D., Atwood, T. S., Currey, M. C., Shiver, A. L., Lewis, Z. A., … Johnson, E. A. (2008). Rapid SNP Discovery and Genetic Mapping Using Sequenced RAD

Markers. *PLoS One*, 3(10). https://doi.org/10.1371/journal.pone.0003376.

Baloglu, G., Haholu, A., Kucukodaci, Z., Yilmaz, I.,Yildirim, S., & Baloglu, H. (2007). The Effects of Tissue Fixation Alternatives on DNA Content: a Study on Normal Colon Tissue. *Applied Immunohistochemistry & Molecular Morphology*, 16(5), 485-492. doi: 10.1097/PAI.0b013e31815dffa6.

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, 30(15), 2114-2120. https://doi.org/10.1093/bioinformatics/btu170.

Borgström, E., Paterlini, M., Mold, J. E., Frisen, J., & Lundeberg, J. (2017). Comparison of whole genome amplification techniques for human single cell exome sequencing. *PLoS ONE*, 12(2): e0171566. https://doi.org/10.1371/journal.pone.0171566

Carpenter, K. E., & Springer, V. G. (2005). The center of the center of marine shore fish biodiversity: the Philippine Islands. *Environmental Biology of Fishes*, 72, 467-480. https://doi.org/10.1007/s10641-004-3154-4.

Chakraborty, A., Sakai, M., & Iwatsuki, Y. (2006). Museum fish specimens and molecular taxonomy: A comparative study on DNA extraction protocols and preservation techniques. *Journal of Applied Ichthyology*, 22(2), 160-166. https://doi.org/10.1111/j.1439-0426.2006.00718.x.

Chong, Z., Ruan, J., & Wu, C. (2012). Rainbow: an integrated tool for efficient clustering and assembling RADseq reads. *Bioinformatics*, 28(21), 2732-2737. https://doi.org/10.1093/bioinformatics/bts482.

Çiftci, Y., & Okumuş, İ. (2002). Fish Population Genetics and Applications of Molecular Markers to Fisheries and Aquaculture: I- Basic Principles of Fish Population Genetics. *Turkish Journal of Fisheries and Aquatic Sciences*, 2, 145-155.

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., ... Durbin, R. (2011). The variant call format and VCFtools. *Bioinformatics*, 27(15), 2156–2158. https://doi.org/10.1093/bioinformatics/btr330.

Davey, J. W., & Blaxter, M. L. (2011). RADseq: next-generation population genetics. *Breifings in Functional Genomics*, 9(5), 416-423. https://doi.org/10.1093/bfgp/elq031.

DeSalle, R., & Amato, G. (2004). The expansion of conservation genetics. *Nature Reviews Genetics*, 5, 702-712. https://doi.org/10.1038/nrg1425.

Durvasula, A., Hoffman, P. J., Kent, T. V., Liu, C., Kono, T. J. Y., Morrell, P. L., & Ross-Ibarra, J. (2016). ANGSD-wrapper: utilities for analysing next-generation sequencing data. *Molecular Ecology Resources*, 16, 1449-1454. https://doi.org/10.1111/1755-0998.12578.

Do, C., Waples, R. S., Peel, D., Macbeth, G. M., Tillett, B. J., & Ovenden, J. R. (2014). NeEstimator v2: re-implementation of software for the estimation of contemporary effective population size (Ne) from genetic data. *Molecular Ecology Resources*, 14(1), 209-214. https://doi.org/10.1111/1755-0998.12157.

Gaston, K. J. (2000). Global patterns in biodiversity. *Nature*, 405, 220-227. https://doi.org/10.1038/35012228.

Goudet, J. (2005). HEIRFSTAT, a package for R to compute and test hierarchical *F*-statistics. *Molecular Ecology Notes*, 5(1), 184-186. https://doi.org/10.1111/j.1471-8286.2004.00828.x.

Grada, A., & Weinbrecht, K. (2013). Next-Generation Sequencing: Methodology and

Application. *Journal of Investigative Dermatology*, 133, 1-4. https://doi.org/10.1038/jid.2013.248.

Habel, J. C., Husemann, M., Finger, A., Danley, P. D., & Zachos, F. E. (2014). The relevance of time series in molecular ecology and conservation biology. *Biological Reviews*, 89, 484-492. https://doi.org/10.1111/brv.12068.

Harahap, A. P., Takemura, A., Rahman, S., Nakamura, S., & Takano, K. (2002). Lunar synchronization of sperm motility in the spiny rabbitfish *Siganus spinus* (Linnaeus). *Fisheries Science*, 68, 706-708. https://doi.org/10.1046/j.1444-2906.2002.00482.x.

Halpern, B. S., Frazier, M., Potapenko, J., Casey, K. S., Koenig, K., Longo, C., ... Walbridge, S. (2015) Spatial and temporal changes in cumulative human impacts on the world's ocean. *Nature Communications*, 6, 7615. https://doi.org/10.1038/ncomms8615.

Halpern, B. S., Walbridge, S., Selkoe, K. A., Kappel, C. V., Micheli, F., D'Agrosa, C., ... Watson, R. (2008). A Global Map of Human Impact on Marine Ecosystems. *Science*, 319(5865), 948-952. DOI: 10.1126/science.1149345.

Harley, C. D. G., Hughes, A. R., Hultgren, K. M., Miner, B. G., Sorte, C. J. B., Thornber, C. S., ... Williams, S. L. (2006). The impacts of climate change in coastal marine systems. *Ecology Letters*, 9(2), 228-241. https://doi.org/10.1111/j.1461-0248.2005.00871.x.

Hoffberg, S. L., Kieran, T .J., Catchen, J. M., Devault, A., Faircloth, B. C., Mauricio, R., & Glenn, T. C. (2016). RADcap: sequence capture of dual-digest RADseq libraries with identifiable duplicates and reduced missing data. *Molecular Ecology Resources*, 16, 1264-1278. https://doi.org/10.1111/1755-0998.12566.

Iwamoto, K., Takemura, A., Yoshino, T., & Imai, H. (2009). Molecular Ecological Study of *Siganus spinus* and *S. guttatus* from Okinawan Waters Based on Mitochondrial DNA

Control Region Sequences. *Journal of Oceanography*, 65, 103-112. https://doi.org/10.1007/s10872-009-0010-3.

Jombart, T., & Ahmed, I. (2011). adegenet 1.3-1: new tools for the analysis of genome-wide SNP data. *Bioinformatics*, 27, 3070-3071. https://doi.org/10.1093/bioinformatics/btr521.

Jones, S. (1960). Further notes on *Spratelloides delicatulus (Bennett)* as a tuna Live-bait fish with a record of *S. Japonicus (Houttoyn)* from the Laccadive Sea. *Journal of the Marine Biological Association of India*, 2 (2), 267-268.

Kimura, S., Iwatsuki, U., & Yoshino, T. (2001). Redescriptions of the Indo-West Pacific atherinid fishes, *Atherinomorus endrachtensis* (Quoy and Gaimard, 1825) and *A. duodecimalis* (Valenciennes *in* Cuvier and Valenciennes, 1835). *Ichthyological Research*, 48, 167-177. https://doi.org/10.1007/s10228-001-8132-7.

Korneliussen, T. S., Albrechtsen, A., & Nielsen, R. (2014). ANGSD: Analysis of Next Generation Sequencing Data. *BMC Bioinformatics*, 15(356), 1-13. https://doi.org/10.1186/s12859-014-0356-4.

Korneliussen, T. S., Moltke, I., Albrechtsen, A., Nielsen, R. (2013). Calculation of Tajima's D and other neutrality test statistics from low depth next-generation sequencing data. *BMC Bioinformatics*, 14 (289), 1-14. https://doi.org/10.1186/1471-2105-14-289.

Kumar, S., Nei, M., Dudley, J., & Tamura, K. (2008). MEGA: A biologist-centric software for evolutionary analysis of DNA and protein sequences. *Briefings in bioinformatics*, *9*(4), 299–306. https://doi.org/10.1093/bib/bbn017.

Laviña, E. M., & Alcala, A. C. (1974). Ecological Studies on Philippine Siganid Fishes in Southern Negros, Philippines. *Silliman Journal*, 21(2), 191-210.

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics*, 25(14), 1754-1760. https://doi.org/10.1093/bioinformatics/btp324.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., … Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16), 2078-2079. https://doi.org/10.1093/bioinformatics/btp352.

Lischer, H. E. L., & Excoffier, L. (2012). PGDSpider: An automated data conversion tool for connecting population genetics and genomics programs. *Bioinformatics*, 28, 298-299. https://doi.org/10.1093/bioinformatics/btr642.

Martin, T. J., & Blaber, S. J. M. (1983). The feeding ecology of Ambassidae (Osteichthyes: Perciformes) in Natal estuaries. *South African Journal of Zoology*, 18(4), 353-362. https://doi.org/10.1080/02541858.1983.11447838.

Martin, T. J., & Heemstra, P. C. (1988). Identification of *Ambassis* species (Pisces: Perciformes, Ambassidae) from South Africa. *African Zoology*, 23(1), 7-12. https://doi.org/10.1080/02541858.1988.11448070.

McCauley, D. J., Pinsky, M. L., Palumbi, S. R., Estes, J. A., Joyce, F. H., & Warner, R. R. (2015). Marine defaunation: Animal loss in the global ocean. *Science*, 347(6219), 1-7. https://doi.org/10.1126/science.1255641.

Meirmans, P. G., & Van Tienderen, P. H. (2004). GENOTYPE and GENODIVE: two programs for the analysis of genetic diversity of asexual organisms. *Molecular Ecology Notes*, 4, 792-794. https://doi.org/10.1111/j.1471-8286.2004.00770.x.

Messing, J. (2001). The Universal Primers and the Shotgun DNA Sequencing Method. *DNA Sequencing Protocols*, 167, 13-31. https://doi.org/10.1385/1-59259-113-2:013.

Miller, M. R., Dunham, J. P., Amores, A., Cresko, W. A., & Johnson, E. A. (2007). Rapid and cost-effective polymorphism identification and genotyping using restriction site associated DNA (RAD) markers. *Genome Research*, 17, 240-248. https://doi.org/10.1101/gr.5681207.

Milton, D. A., Blaber, S. J. M. & Rawlinson, N. J. F., (1991). Age and growth of three species of tuna baitfish (genus: *Spratelloides*) in the tropical Indo-Pacific. *Journal of Fish Biology*, *39*(6), 849-866. https://doi.org/10.1111/j.1095-8649.1991.tb04414.x.

Mohan, M., & Kunhikoya, K. K. (1985). Biology of the Bait Fishes *Spratelloides delicatulus* (Bennet) and *S. Japonicus* (Houttuyn) from Minicoy Waters. *Central Marine Fisheries Institute*, 36, 155-164.

Nañola, C. L., Aliño, P. M., & Carpenter, K. E. (2011). Exploitation-related reef fish species richness depletion in the epicenter of marine biodiversity. *Environmental Biology of Fishes*, 90, 405–420. https://doi.org/10.1007/s10641-010-9750-6.

Narum, S. R., Buerkle, C. A., Davey, J. W., Miller, M. R., & Hohenlohe, P. A. (2013). Genotyping-by-sequencing in ecological and conservation genomics. *Molecular Ecology*, 22(11), 2841-2847. https://doi.org/10.1111/mec.12350.

O'Leary, S. J., Puritz, J. B., Willis, S. C., Hollenbeck, C. M., & Portnoy, D. S. (2018). These aren't the loci you're looking for: Principles of effective SNP filtering for molecular ecologists. *Molecular Ecology*, 27(16), 3193-3206. https://doi.org/10.1111/mec.14792.

Peñalba, J. V., Smith, L. L., Tonione, M. A., Sass, C., Hykin, S. M., Skipwith, P. L., ... Mortiz, C. (2014). Sequence capture using PCR-generated probes: a cost-effective method of

targeted high-throughput sequencing for nonmodel organisms. *Marine Ecology Resources*, 14, 1000-1010. https://doi.org/10.1111/1755-0998.12249.

Pinheiro, H.T., Shepherd, B., Castillo, C., Abesamis, R.A., Copus, J.M., Pyle, R.L., …Bucol, A.A., 2019. Deep reef fishes in the world's epicenter of marine biodiversity. *Coral Reefs*, 38(5), 985-995. https://doi.org/10.1007/s00338-019-01825-5.

Pinsky, M. L., & Palumbi, S. R. (2014). Meta-analysis reveals lower genetic diversity in overfished populations. *Molecular Ecology*, 23(1), 29-39. https://doi.org/10.1111/mec.12509.

Puritz, J. B., Hollenbeck, C. M., & Gold, J. R. (2014). *dDocent*: a RADseq, variant-calling pipeline designed for population genomics of non-model organisms. *PeerJ*, 2, e431.

R Core Team (2020). R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; Available: https://www.R-project.org

Riede, K. (2004). Global register of migratory species - from global to regional scales. Final Report of the R&D-Projekt 808 05 081. Federal Agency for Nature Conservation, Bonn, Germany.

Roberts, C. M., McClean, C. J., Veron, J. E. N., Hawkins, J. P., Allen, G. R., McAllister, D. E., ...Werner, T. B. (2002). Marine biodiversity hotspots and conservation priorities for tropical reefs. *Science*, 295, 1280–1284. https://doi.org/10.1126/science.1067728.

Shafer, A. B. A., Peart, C. R., Tusso, S., Maayan, I., Brelsford, A., Wheat, C. W., & Wolf, J. B. W. (2017). Bioinformatic processing of RADseq data dramatically impacts downstream population genetic inference. *Methods in Ecology and Evolution*, 8, 907-917. https://doi.org/10.1111/2041-210X.12700.

Shiozawa, D. K., Kudo, J., Evans, R. P., Woodward, S. R., & Williams, R. N. (1992). DNA Extracted from Preserved Trout Tissues. *The Great Basin Naturalist*, 52(1), 29-34. http://www.jstor.org/stable/41712692.

Smith, D. G., & Williams, J. T. (1999). The Great *Albatross* Philippine Expedition and Its Fishes. *Marine Fisheries Review*, 61(4), 31-41.

Soliman, V.S., Yamada, H., & Yamaoka, K. (2010) Early life-history of the spiny siganid *Siganus spinus* (Linnaeus 1758) inferred from otolith microstructure. *Journal of Applied Ichthyology*, 26(4), 540-545. https://doi.org/10.1111/j.1439-0426.2010.01478.x.

Stockwell, B. L., Larson, W. A., Waples, R. K., Abesamis, R. A., Seeb, L. W., & Carpenter, K. E. (2016). The application of genomics to inform conservation of a functionally important reef fish (*Scarus niger*) in the Philippines. *Conservation Genetics*, 17, 239-249. https://doi.org/10.1007/s10592-015-0776-3.

Subade, R. F. (2007). Mechanisms to capture economic values of marine biodiversity: The case of Tubbataha Reefs UNESCO World Heritage Site, Philippines. *Marine Policy*, 31, 135-142. https://doi.org/10.1016/j.marpol.2006.05.012.

Suchan, T., Pitteloud, C., Gerasimova, N. S., Kostikova, A., & Schmid, S., Arrigo, S., …Alvarez, N. (2016). Hybridization Capture Using RAD Probes (hyRAD), a New Tool for Performing Genomic Analyses on Collection Specimens. *PLOS ONE*, 11(3), e0151651. https://doi.org/10.1371/journal.pone.0151651.

Takemura, I., Sado, T., Maekawa, Y., & Kimura, S. (2004). Descriptive morphology of the reared eggs, larvae, and juveniles of the marine atherinid fish *Atherinomorus duodecimalis*. *Ichthyological Research*, 51, 159-164. https://doi.org/10.1007/s10228-004-0212-z.

Tamayo, N. C. A., Anticamara, J. A., & Acosta-Michlik, L. (2018). National estimates of values of Philippine Reefs' ecosystem services. *Ecological Economics*, 146, 633-644. https://doi.org/10.1016/j.ecolecon.2017.12.005.

Therkildsen, N. O., Wilder, A. P., Conover, D. O., Munch, S. B., Baumann, H., & Palumbi, S. R. (2019). Contrasting genomic shifts underlie parallel phenotypic evolution in response to fishing. *Science*, 365, 487-490. https://doi.org/10.1126/science.aaw7271.

Toonen, R. J., Puritz, J. B., Forsman, Z. H., Whitney, J. L., Fernandez-Silva, I., Andrews, K. R., & Bird, C. E. (2013). ezRAD: a simplified method for genomic genotyping in non-model organisms. *PeerJ*, 1, e203 https://doi.org/10.7717/peerj.203.

Tornabene, L., Valdez, S., Erdmann, M., & Pezold, F. (2015). Support for a 'Center of Origin' in the Coral Triangle: Cryptic diversity, recent speciation, and local endemism in a diverse lineage of reef fishes (Gobiidae: Eviota). *Molecular Phylogenetics and Evolution*, 82, 200-210. https://doi.org/10.1016/j.ympev.2014.09.012.

Wandeler, P., Hoeck, P. E. A., & Keller, L. F., (2007). Back to the future: museum specimens in population genetics. *Trends in Evolution and Ecology*, 22(12), 634-642. https://doi.org/10.1016/j.tree.2007.08.017.

White, A. T., Vogt, H. P., & Arin, T. (2000). Philippine Coral Reefs Under Threat: The Economic Losses Caused by Reef Destruction. *Marine Pollution Bulletin*, 40, 598-605. https://doi.org/10.1016/S0025-326X(00)00022-9.

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., … Yutani H (2019). "Welcome to the tidyverse.". *Journal of Open Source Software*, 4(43), 1686. https://doi.org/10.21105/joss.01686.

APPENDIX A

This is a configuration file for fltrVCF to control filters, filter order, and filter thresholds. Each row controls a setting and will be listed by command and argument. Settings here will be overridden by arguments specified at the command line

For all fltrVCF options use the -h argument at the command line.

Notes: These settings are designed to clean a raw VCF file made from individuals and retain as much biological variation as possible.

fltrVCF Settings, run fltrVCF -h for description of settings

fltrVCF -f 01 02 03 04 14 07 05 17 15 06 11 09 08 10 04 13 05 07 18 19 20

fltrVCF -c 3.3

Filters

01 vcftools --min-alleles 2 #Remove sites with less alleles.

01 vcftools –max-alleles 2 #Remove sites with more alleles.

02 vcftools --remove-indels #Remove sites with indels. Not adjustable.

03 vcftools --minQ 100 #Remove sites with lower QUAL.

04 vcftools --min-meanDP 8 #Remove sites with lower mean depth.

05 vcftools --max-missing 0.4 #Remove sites with lower proportion of genotypes present.

06 vcffilter AB min 0.25 #Remove sites with equal or lower allele balance.

06 vcffilter AB max 0.75 #Remove sites with equal or lower allele balance.

06 vcffilter AB nohet 0 #Keep sites with AB=0. Not adjustable.

07 vcffilter AC min 0 #Remove sites with equal or lower MINOR allele count.

08 vcffilter SAF/SAR min 10 #Remove sites where both read1 and 2 overlap. Remove sites with equal or lower (SAF/SAR & SRF/SRR | SAR/SAF & SRR/SRF). These are the number of F and R reads supporting the REF or ALT alleles.

09 vcffilter MQM/MQMR min 0.25 #Remove sites where the difference in the ratio of mean mapping quality between REF and ALT alleles is greater than this proportion from 1. Ex: 0 means the mapping quality must be equal between REF and ALTERNATE. Smaller numbers are more stringent. Keep sites where the following is true: 1-X < MQM/MQMR < 1/(1-X).

10 vcffilter PAIRED #Remove sites where one of the alleles is only supported by reads that are not properly paired (see SAM format specification). Not adjustable.

11 vcffilter QUAL/DP min 0.2 #Remove sites where the ratio of QUAL to DP is deemed to be too low.

12 vcftools QUAL/DP max #Remove sites where the ratio of QUAL to DP is deemed to be too high. Not adjustable.

13 vcftools --max-meanDP 400 #Remove sites with higher mean depth.

14 vcftools --minDP 10 #Code genotypes with lesser depth of coverage as NA.

15 vcftools --maf 0 #Remove sites with lesser minor allele frequency. Adjust based upon sample size.
15 vcftools --max-maf 1 #Remove sites with greater minor allele frequency. Adjust based upon sample size.
16 vcftools --missing-indv 1 #Remove individuals with more missing data.
17 vcftools --missing-sites 0.5 #Remove sites with more data missing in a pop sample.
18 filter_hwe_by_pop_HPC 0.001 #Remove sites with <p in test for HWE by pop sample. Adjust based upon sample size.
19 rad_haplotyper -d 50 #depth of sampling reads for building haplotypes.
19 rad_haplotyper -mp 1 #Remove sites with more paralogous individuals. Adjust according to sample size.
19 rad_haplotyper -u 40 #Remove contigs with more SNPs. Adjust according to sequence length.
19 rad_haplotyper -ml 10 #Remove contigs with more individuals exhibiting low coverage or genotyping errors.
19 rad_haplotyper -h 25 #Remove contigs with greater NumHaplotypes-NumSNPs.
19 rad_haplotyper -z 0.1 #Remove up to this proportion or number of reads when testing for paralogs. The more real variation in your data set, the greater this number will be. (<1) or number (>=1) of reads.
19 rad_haplotyper -m 0.5 #Keep loci with a greater proportion of haplotyped individuals.
20 OneRandSNP #Keep 1 random SNP per contig. Not adjustable. Can't be run after filter 21.
21 MostInformativeSNPs #Keep the most informative SNP per contig. Not adjustable. Can't be run after filter 20.
86 rmContigs #Remove contigs that have had SNPs removed by the previous filter. Intended to be run after filters 05, 06, 13, 14, 17, 18 if desired.

APPENDIX B

Siganus spinus unbaited contemporary individually optimized settings.

This is a configuration file for fltrVCF to control filters, filter order, and filter thresholds. Each row controls a setting and will be listed by command and argument. Settings here will be overridden by arguments specified at the command line

For all fltrVCF options use the -h argument at the command line.

fltrVCF Settings, run fltrVCF -h for description of settings

fltrVCF -f 01 02 03 04 14 07 05 17 15 06 11 09 08 10 04 13 05 07 18 20
fltrVCF -c 5.5

Filters

01 vcftools --min-alleles 2 #Remove sites with less alleles.
01 vcftools --max-alleles 2 #Remove sites with more alleles.
02 vcftools --remove-indels #Remove sites with indels. Not adjustable.
03 vcftools --minQ 100 #Remove sites with lower QUAL.
04 vcftools --min-meanDP 8 #Remove sites with lower mean depth.
05 vcftools --max-missing 0.4 #Remove sites with lower proportion of genotypes present.
06 vcffilter AB min 0.25 #Remove sites with equal or lower allele balance.
06 vcffilter AB max 0.75 #Remove sites with equal or lower allele balance.
06 vcffilter AB nohet 0 #Keep sites with AB=0. Not adjustable.
07 vcffilter AC min 0 #Remove sites with equal or lower MINOR allele count.
08 vcffilter SAF/SAR min 10 #Remove sites where both read1 and 2 overlap. Remove sites with equal or lower (SAF/SAR & SRF/SRR | SAR/SAF & SRR/SRF). These are the number of F and R reads supporting the REF or ALT alleles.
09 vcffilter MQM/MQMR min 0.25 #Remove sites where the difference in the ratio of mean mapping quality between REF and ALT alleles is greater than this proportion from 1. Ex: 0 means the mapping quality must be equal between REF and ALTERNATE. Smaller numbers are more stringent. Keep sites where the following is true: 1-X < MQM/MQMR < 1/(1-X).
10 vcffilter PAIRED #Remove sites where one of the alleles is only supported by reads that are not properly paired (see SAM format specification). Not adjustable.
11 vcffilter QUAL/DP min 0.2 #Remove sites where the ratio of QUAL to DP is deemed to be too low.
12 vcftools QUAL/DP max #Remove sites where the ratio of QUAL to DP is deemed to be too high. Not adjustable.
13 vcftools --max-meanDP 400 #Remove sites with higher mean depth.
14 vcftools --minDP 10 #Code genotypes with lesser depth of coverage as NA.

15 vcftools --maf 0 #Remove sites with lesser minor allele frequency. Adjust based upon sample size.
15 vcftools --max-maf 1 #Remove sites with greater minor allele frequency. Adjust based upon sample size.
16 vcftools --missing-indv 0.6 #Remove individuals with more missing data.
17 vcftools --missing-sites 0.5 #Remove sites with more data missing in a pop sample.
18 filter_hwe_by_pop_HPC 0.001 #Remove sites with <p in test for HWE by pop sample. Adjust based upon sample size.
19 rad_haplotyper -d 50 #depth of sampling reads for building haplotypes.
19 rad_haplotyper -mp 1 #Remove sites with more paralogous individuals. Adjust according to sample size.
19 rad_haplotyper -u 40 #Remove contigs with more SNPs. Adjust according to sequence length.
19 rad_haplotyper -ml 10 #Remove contigs with more individuals exhibiting low coverage or genotyping errors.
19 rad_haplotyper -h 25 #Remove contigs with greater NumHaplotypes-NumSNPs.
19 rad_haplotyper -z 0.1 #Remove up to this proportion or number of reads when testing for paralogs. The more real variation in your data set, the greater this number will be. (<1) or number (>=1) of reads.
19 rad_haplotyper -m 0.5 #Keep loci with a greater proportion of haplotyped individuals.
20 OneRandSNP #Keep 1 random SNP per contig. Not adjustable. Can't be run after filter 21.
21 MostInformativeSNPs #Keep the most informative SNP per contig. Not adjustable. Can't be run after filter 20.
86 rmContigs #Remove contigs that have had SNPs removed by the previous filter. Intended to be run after filters 05, 06, 13, 14, 17, 18 if desired.

APPENDIX C

Siganus spinus unbaited *Albatross* individually optimized settings

This is a configuration file for fltrVCF to control filters, filter order, and filter thresholds. Each row controls a setting and will be listed by command and argument. Settings here will be overridden by arguments specified at the command line

For all fltrVCF options use the -h argument at the command line.

fltrVCF Settings, run fltrVCF -h for description of settings

fltrVCF -f 01 02 03 04 14 05 17 15 06 11 09 08 10 04 13 05 18 20
fltrVCF -c 5.5

Filters

01 vcftools --min-alleles 2 #Remove sites with less alleles.
01 vcftools --max-alleles 2 #Remove sites with more alleles.
02 vcftools --remove-indels #Remove sites with indels. Not adjustable.
03 vcftools --minQ 100 #Remove sites with lower QUAL.
04 vcftools --min-meanDP 1 #Remove sites with lower mean depth.
05 vcftools --max-missing 0.1 #Remove sites with lower proportion of genotypes present.
06 vcffilter AB min 0.25 #Remove sites with equal or lower allele balance.
06 vcffilter AB max 0.75 #Remove sites with equal or lower allele balance.
06 vcffilter AB nohet 0 #Keep sites with AB=0. Not adjustable.
07 vcffilter AC min 0 #Remove sites with equal or lower MINOR allele count.
08 vcffilter SAF/SAR min 10 #Remove sites where both read1 and 2 overlap. Remove sites with equal or lower (SAF/SAR & SRF/SRR | SAR/SAF & SRR/SRF). These are the number of F and R reads supporting the REF or ALT alleles.
09 vcffilter MQM/MQMR min 0.25 #Remove sites where the difference in the ratio of mean mapping quality between REF and ALT alleles is greater than this proportion from 1. Ex: 0 means the mapping quality must be equal between REF and ALTERNATE. Smaller numbers are more stringent. Keep sites where the following is true: 1-X < MQM/MQMR < 1/(1-X).
10 vcffilter PAIRED #Remove sites where one of the alleles is only supported by reads that are not properly paired (see SAM format specification). Not adjustable.
11 vcffilter QUAL/DP min 0.2 #Remove sites where the ratio of QUAL to DP is deemed to be too low.
12 vcftools QUAL/DP max #Remove sites where the ratio of QUAL to DP is deemed to be too high. Not adjustable.
13 vcftools --max-meanDP 400 #Remove sites with higher mean depth.
14 vcftools --minDP 10 #Code genotypes with lesser depth of coverage as NA.

15 vcftools --maf 0 #Remove sites with lesser minor allele frequency. Adjust based upon sample size.
15 vcftools --max-maf 1 #Remove sites with greater minor allele frequency. Adjust based upon sample size.
16 vcftools --missing-indv 0.6 #Remove individuals with more missing data.
17 vcftools --missing-sites 0.9 #Remove sites with more data missing in a pop sample.
18 filter_hwe_by_pop_HPC 0.001 #Remove sites with <p in test for HWE by pop sample. Adjust based upon sample size.
19 rad_haplotyper -d 50 #depth of sampling reads for building haplotypes.
19 rad_haplotyper -mp 1 #Remove sites with more paralogous individuals. Adjust according to sample size.
19 rad_haplotyper -u 40 #Remove contigs with more SNPs. Adjust according to sequence length.
19 rad_haplotyper -ml 10 #Remove contigs with more individuals exhibiting low coverage or genotyping errors.
19 rad_haplotyper -h 25 #Remove contigs with greater NumHaplotypes-NumSNPs.
19 rad_haplotyper -z 0.1 #Remove up to this proportion or number of reads when testing for paralogs. The more real variation in your data set, the greater this number will be. (<1) or number (>=1) of reads.
19 rad_haplotyper -m 0.5 #Keep loci with a greater proportion of haplotyped individuals.
20 OneRandSNP #Keep 1 random SNP per contig. Not adjustable. Can't be run after filter 21.
21 MostInformativeSNPs #Keep the most informative SNP per contig. Not adjustable. Can't be run after filter 20.
86 rmContigs #Remove contigs that have had SNPs removed by the previous filter. Intended to be run after filters 05, 06, 13, 14, 17, 18 if desired.

APPENDIX D

Ambassis urotaenia unbaited contemporary individually optimized settings

This is a configuration file for fltrVCF to control filters, filter order, and filter thresholds. Each row controls a setting and will be listed by command and argument. Settings here will be overridden by arguments specified at the command line

For all fltrVCF options use the -h argument at the command line.

fltrVCF Settings, run fltrVCF -h for description of settings

fltrVCF -f 01 02 03 04 14 07 05 17 15 06 11 09 08 10 04 13 05 07 18 19 20
fltrVCF -c 2.2

Filters

01 vcftools --min-alleles 2 #Remove sites with less alleles.
01 vcftools --max-alleles 2 #Remove sites with more alleles.
02 vcftools --remove-indels #Remove sites with indels. Not adjustable
03 vcftools --minQ 100 #Remove sites with lower QUAL.
04 vcftools --min-meanDP 8 #Remove sites with lower mean depth.
05 vcftools --max-missing 0.4 #Remove sites with lower proportion of genotypes present.
06 vcffilter AB min 0.25 #Remove sites with equal or lower allele balance.
06 vcffilter AB max 0.75 #Remove sites with equal or lower allele balance.
06 vcffilter AB nohet 0 #Keep sites with AB=0. Not adjustable.
07 vcffilter AC min 0 #Remove sites with equal or lower MINOR allele count.
08 vcffilter SAF/SAR min 10 #Remove sites where both read1 and 2 overlap. Remove sites with equal or lower (SAF/SAR & SRF/SRR | SAR/SAF & SRR/SRF). These are the number of F and R reads supporting the REF or ALT alleles.
09 vcffilter MQM/MQMR min 0.25 #Remove sites where the difference in the ratio of mean mapping quality between REF and ALT alleles is greater than this proportion from 1. Ex: 0 means the mapping quality must be equal between REF and ALTERNATE. Smaller numbers are more stringent. Keep sites where the following is true: 1-X < MQM/MQMR < 1/(1-X).
10 vcffilter PAIRED #Remove sites where one of the alleles is only supported by reads that are not properly paired (see SAM format specification). Not adjustable.
11 vcffilter QUAL/DP min 0.2 #Remove sites where the ratio of QUAL to DP is deemed to be too low.
12 vcftools QUAL/DP max #Remove sites where the ratio of QUAL to DP is deemed to be too high. Not adjustable.
13 vcftools --max-meanDP 400 #Remove sites with higher mean depth.
14 vcftools --minDP 10 #Code genotypes with lesser depth of coverage as NA.

15 vcftools --maf 0 #Remove sites with lesser minor allele frequency. Adjust based upon sample size.
15 vcftools --max-maf 1 #Remove sites with greater minor allele frequency. Adjust based upon sample size.
16 vcftools --missing-indv 0.6 #Remove individuals with more missing data.
17 vcftools --missing-sites 0.5 #Remove sites with more data missing in a pop sample.
18 filter_hwe_by_pop_HPC 0.001 #Remove sites with <p in test for HWE by pop sample. Adjust based upon sample size.
19 rad_haplotyper -d 50 #depth of sampling reads for building haplotypes.
19 rad_haplotyper -mp 1 #Remove sites with more paralogous individuals. Adjust according to sample size.
19 rad_haplotyper -u 40 #Remove contigs with more SNPs. Adjust according to sequence length.
19 rad_haplotyper -ml 10 #Remove contigs with more individuals exhibiting low coverage or genotyping errors.
19 rad_haplotyper -h 25 #Remove contigs with greater NumHaplotypes-NumSNPs.
19 rad_haplotyper -z 0.1 #Remove up to this proportion or number of reads when testing for paralogs. The more real variation in your data set, the greater this number will be. (<1) or number (>=1) of reads.
19 rad_haplotyper -m 0.5 #Keep loci with a greater proportion of haplotyped individuals.
20 OneRandSNP #Keep 1 random SNP per contig. Not adjustable. Can't be run after filter 21.
21 MostInformativeSNPs #Keep the most informative SNP per contig. Not adjustable. Can't be run after filter 20.
86 rmContigs #Remove contigs that have had SNPs removed by the previous filter. Intended to be run after filters 05, 06, 13, 14, 17, 18 if desired.

APPENDIX E

Siganus spinus baited contemporary individually optimized settings.

This is a configuration file for fltrVCF to control filters, filter order, and filter thresholds. Each row controls a setting and will be listed by command and argument. Settings here will be overridden by arguments specified at the command line

For all fltrVCF options use the -h argument at the command line.

fltrVCF Settings, run fltrVCF -h for description of settings

fltrVCF -f 01 02 03 04 14 07 05 17 15 06 11 09 08 10 04 13 05 07 18 20
fltrVCF -c 5.5

Filters

01 vcftools --min-alleles 2 #Remove sites with less alleles.
01 vcftools --max-alleles 2 #Remove sites with more alleles.
02 vcftools --remove-indels #Remove sites with indels. Not adjustable.
03 vcftools --minQ 100 #Remove sites with lower QUAL.
04 vcftools --min-meanDP 8 #Remove sites with lower mean depth.
05 vcftools --max-missing 0.4 #Remove sites with lower proportion of genotypes present.
06 vcffilter AB min 0.25 #Remove sites with equal or lower allele balance.
06 vcffilter AB max 0.75 #Remove sites with equal or lower allele balance.
06 vcffilter AB nohet 0 #Keep sites with AB=0. Not adjustable.
07 vcffilter AC min 0 #Remove sites with equal or lower MINOR allele count.
08 vcffilter SAF/SAR min 10 #Remove sites where both read1 and 2 overlap. Remove sites with equal or lower (SAF/SAR & SRF/SRR | SAR/SAF & SRR/SRF). These are the number of F and R reads supporting the REF or ALT alleles.
09 vcffilter MQM/MQMR min 0.25 #Remove sites where the difference in the ratio of mean mapping quality between REF and ALT alleles is greater than this proportion from 1. Ex: 0 means the mapping quality must be equal between REF and ALTERNATE. Smaller numbers are more stringent. Keep sites where the following is true: 1-X < MQM/MQMR < 1/(1-X).
10 vcffilter PAIRED #Remove sites where one of the alleles is only supported by reads that are not properly paired (see SAM format specification). Not adjustable.
11 vcffilter QUAL/DP min 0.2 #Remove sites where the ratio of QUAL to DP is deemed to be too low.
12 vcftools QUAL/DP max #Remove sites where the ratio of QUAL to DP is deemed to be too high. Not adjustable.
13 vcftools --max-meanDP 400 #Remove sites with higher mean depth.
14 vcftools --minDP 10 #Code genotypes with lesser depth of coverage as NA.

15 vcftools --maf 0 #Remove sites with lesser minor allele frequency. Adjust based upon sample size.
15 vcftools --max-maf 1 #Remove sites with greater minor allele frequency. Adjust based upon sample size.
16 vcftools --missing-indv 0.6 #Remove individuals with more missing data.
17 vcftools --missing-sites 0.5 #Remove sites with more data missing in a pop sample.
18 filter_hwe_by_pop_HPC 0.001 #Remove sites with <p in test for HWE by pop sample. Adjust based upon sample size.
19 rad_haplotyper -d 50 #depth of sampling reads for building haplotypes.
19 rad_haplotyper -mp 1 #Remove sites with more paralogous individuals. Adjust according to sample size.
19 rad_haplotyper -u 40 #Remove contigs with more SNPs. Adjust according to sequence length.
19 rad_haplotyper -ml 10 #Remove contigs with more individuals exhibiting low coverage or genotyping errors.
19 rad_haplotyper -h 25 #Remove contigs with greater NumHaplotypes-NumSNPs.
19 rad_haplotyper -z 0.1 #Remove up to this proportion or number of reads when testing for paralogs. The more real variation in your data set, the greater this number will be. (<1) or number (>=1) of reads.
19 rad_haplotyper -m 0.5 #Keep loci with a greater proportion of haplotyped individuals.
20 OneRandSNP #Keep 1 random SNP per contig. Not adjustable. Can't be run after filter 21.
21 MostInformativeSNPs #Keep the most informative SNP per contig. Not adjustable. Can't be run after filter 20.
86 rmContigs #Remove contigs that have had SNPs removed by the previous filter. Intended to be run after filters 05, 06, 13, 14, 17, 18 if desired.

APPENDIX F

Siganus spinus baited *Albatross* individually optimized settings

This is a configuration file for fltrVCF to control filters, filter order, and filter thresholds. Each row controls a setting and will be listed by command and argument. Settings here will be overridden by arguments specified at the command line

For all fltrVCF options use the -h argument at the command line.

fltrVCF Settings, run fltrVCF -h for description of settings

```
fltrVCF -f 01 02 03 14 15 06 11 09 08 10 13 18 20
fltrVCF -c 5.5
```

Filters

```
01 vcftools --min-alleles       2       #Remove sites with less alleles.
01 vcftools --max-alleles       2       #Remove sites with more alleles.
02 vcftools --remove-indels             #Remove sites with indels.  Not adjustable.
03 vcftools --minQ              100     #Remove sites with lower QUAL.
04 vcftools --min-meanDP        1       #Remove sites with lower mean depth.
05 vcftools --max-missing       0.1     #Remove sites with lower proportion of genotypes
present.
06 vcffilter AB min             0.25    #Remove sites with equal or lower allele balance.
06 vcffilter AB max             0.75    #Remove sites with equal or lower allele balance.
06 vcffilter AB nohet           0       #Keep sites with AB=0. Not adjustable.
07 vcffilter AC min             0       #Remove sites with equal or lower MINOR allele
count.
08 vcffilter SAF/SAR min        10      #Remove sites where both read1 and 2 overlap.
Remove sites with equal or lower (SAF/SAR & SRF/SRR | SAR/SAF & SRR/SRF).
These are the number of F and R reads supporting the REF or ALT alleles.
09 vcffilter MQM/MQMR min 0.25  #Remove sites where the difference in the ratio of
mean mapping quality between REF and ALT alleles is greater than this proportion from
1. Ex: 0 means the mapping quality must be equal between REF and ALTERNATE.
Smaller numbers are more stringent. Keep sites where the following is true: 1-X <
MQM/MQMR < 1/(1-X).
10 vcffilter PAIRED                     #Remove sites where one of the alleles is only
supported by reads that are not properly paired (see SAM format specification). Not
adjustable.
11 vcffilter QUAL/DP min        0.2     #Remove sites where the ratio of QUAL to DP is
deemed to be too low.
12 vcftools QUAL/DP max                 #Remove sites where the ratio of QUAL to DP is
deemed to be too high. Not adjustable.
13 vcftools --max-meanDP        400     #Remove sites with higher mean depth.
14 vcftools --minDP             10      #Code genotypes with lesser depth of coverage as
NA.
```

15 vcftools --maf 0 #Remove sites with lesser minor allele frequency. Adjust based upon sample size.
15 vcftools --max-maf 1 #Remove sites with greater minor allele frequency. Adjust based upon sample size.
16 vcftools --missing-indv 0.6 #Remove individuals with more missing data.
17 vcftools --missing-sites 0.9 #Remove sites with more data missing in a pop sample.
18 filter_hwe_by_pop_HPC 0.001 #Remove sites with <p in test for HWE by pop sample. Adjust based upon sample size.
19 rad_haplotyper -d 50 #depth of sampling reads for building haplotypes.
19 rad_haplotyper -mp 1 #Remove sites with more paralogous individuals. Adjust according to sample size.
19 rad_haplotyper -u 40 #Remove contigs with more SNPs. Adjust according to sequence length.
19 rad_haplotyper -ml 10 #Remove contigs with more individuals exhibiting low coverage or genotyping errors.
19 rad_haplotyper -h 25 #Remove contigs with greater NumHaplotypes-NumSNPs.
19 rad_haplotyper -z 0.1 #Remove up to this proportion or number of reads when testing for paralogs. The more real variation in your data set, the greater this number will be. (<1) or number (>=1) of reads.
19 rad_haplotyper -m 0.5 #Keep loci with a greater proportion of haplotyped individuals.
20 OneRandSNP #Keep 1 random SNP per contig. Not adjustable. Can't be run after filter 21.
21 MostInformativeSNPs #Keep the most informative SNP per contig. Not adjustable. Can't be run after filter 20.
86 rmContigs #Remove contigs that have had SNPs removed by the previous filter. Intended to be run after filters 05, 06, 13, 14, 17, 18 if desired.

APPENDIX G

Ambassis urotaenia baited contemporary individually optimized settings

This is a configuration file for fltrVCF to control filters, filter order, and filter thresholds. Each row controls a setting and will be listed by command and argument. Settings here will be overridden by arguments specified at the command line

For all fltrVCF options use the -h argument at the command line.

fltrVCF Settings, run fltrVCF -h for description of settings

fltrVCF -f 01 02 03 04 14 05 17 15 06 11 09 08 10 04 13 05 18 20

fltrVCF -c 2.2

Filters

01 vcftools --min-alleles 2 #Remove sites with less alleles.

01 vcftools --max-alleles 2 #Remove sites with more alleles.

02 vcftools --remove-indels #Remove sites with indels. Not adjustable.

03 vcftools --minQ 100 #Remove sites with lower QUAL.

04 vcftools --min-meanDP 2 #Remove sites with lower mean depth.

05 vcftools --max-missing 0.1 #Remove sites with lower proportion of genotypes present.

06 vcffilter AB min 0.25 #Remove sites with equal or lower allele balance.

06 vcffilter AB max 0.75 #Remove sites with equal or lower allele balance.

06 vcffilter AB nohet 0 #Keep sites with AB=0. Not adjustable

07 vcffilter AC min 0 #Remove sites with equal or lower MINOR allele count.

08 vcffilter SAF/SAR min 10 #Remove sites where both read1 and 2 overlap. Remove sites with equal or lower (SAF/SAR & SRF/SRR | SAR/SAF & SRR/SRF). These are the number of F and R reads supporting the REF or ALT alleles.

09 vcffilter MQM/MQMR min 0.25 #Remove sites where the difference in the ratio of mean mapping quality between REF and ALT alleles is greater than this proportion from 1. Ex: 0 means the mapping quality must be equal between REF and ALTERNATE. Smaller numbers are more stringent. Keep sites where the following is true: 1-X < MQM/MQMR < 1/(1-X).

10 vcffilter PAIRED #Remove sites where one of the alleles is only supported by reads that are not properly paired (see SAM format specification). Not adjustable.

11 vcffilter QUAL/DP min 0.2 #Remove sites where the ratio of QUAL to DP is deemed to be too low.

12 vcftools QUAL/DP max #Remove sites where the ratio of QUAL to DP is deemed to be too high. Not adjustable.

13 vcftools --max-meanDP 400 #Remove sites with higher mean depth.

14 vcftools --minDP 10 #Code genotypes with lesser depth of coverage as NA.

15 vcftools --maf 0 #Remove sites with lesser minor allele frequency. Adjust based upon sample size.
15 vcftools --max-maf 1 #Remove sites with greater minor allele frequency. Adjust based upon sample size.
16 vcftools --missing-indv 0.6 #Remove individuals with more missing data.
17 vcftools --missing-sites 0.5 #Remove sites with more data missing in a pop sample.
18 filter_hwe_by_pop_HPC 0.001 #Remove sites with <p in test for HWE by pop sample. Adjust based upon sample size.
19 rad_haplotyper -d 50 #depth of sampling reads for building haplotypes.
19 rad_haplotyper -mp 1 #Remove sites with more paralogous individuals. Adjust according to sample size.
19 rad_haplotyper -u 40 #Remove contigs with more SNPs. Adjust according to sequence length.
19 rad_haplotyper -ml 10 #Remove contigs with more individuals exhibiting low coverage or genotyping errors.
19 rad_haplotyper -h 25 #Remove contigs with greater NumHaplotypes-NumSNPs.
19 rad_haplotyper -z 0.1 #Remove up to this proportion or number of reads when testing for paralogs. The more real variation in your data set, the greater this number will be. (<1) or number (>=1) of reads.
19 rad_haplotyper -m 0.5 #Keep loci with a greater proportion of haplotyped individuals.
20 OneRandSNP #Keep 1 random SNP per contig. Not adjustable. Can't be run after filter 21.
21 MostInformativeSNPs #Keep the most informative SNP per contig. Not adjustable. Can't be run after filter 20.
86 rmContigs #Remove contigs that have had SNPs removed by the previous filter. Intended to be run after filters 05, 06, 13, 14, 17, 18 if desired.

APPENDIX H

Ambassis urotaenia baited *Albatross* individually optimized settings

This is a configuration file for fltrVCF to control filters, filter order, and filter thresholds. Each row controls a setting and will be listed by command and argument. Settings here will be overridden by arguments specified at the command line

For all fltrVCF options use the -h argument at the command line.

fltrVCF Settings, run fltrVCF -h for description of settings

fltrVCF -f 01 02 03 14 15 06 11 09 08 10 13 18 20

fltrVCF -c 2.2

Filters

01 vcftools --min-alleles	2	#Remove sites with less alleles.
01 vcftools --max-alleles	2	#Remove sites with more alleles.
02 vcftools --remove-indels		#Remove sites with indels. Not adjustable.
03 vcftools --minQ	100	#Remove sites with lower QUAL.
04 vcftools --min-meanDP	1	#Remove sites with lower mean depth.
05 vcftools --max-missing	0.1	#Remove sites with lower proportion of genotypes present.
06 vcffilter AB min	0.25	#Remove sites with equal or lower allele balance.
06 vcffilter AB max	0.75	#Remove sites with equal or lower allele balance.
06 vcffilter AB nohet	0	#Keep sites with AB=0. Not adjustable
07 vcffilter AC min	0	#Remove sites with equal or lower MINOR allele count.
08 vcffilter SAF/SAR min	10	#Remove sites where both read1 and 2 overlap. Remove sites with equal or lower (SAF/SAR & SRF/SRR | SAR/SAF & SRR/SRF). These are the number of F and R reads supporting the REF or ALT alleles.
09 vcffilter MQM/MQMR min	0.25	#Remove sites where the difference in the ratio of mean mapping quality between REF and ALT alleles is greater than this proportion from 1. Ex: 0 means the mapping quality must be equal between REF and ALTERNATE. Smaller numbers are more stringent. Keep sites where the following is true: 1-X < MQM/MQMR < 1/(1-X).
10 vcffilter PAIRED		#Remove sites where one of the alleles is only supported by reads that are not properly paired (see SAM format specification). Not adjustable.
11 vcffilter QUAL/DP min	0.2	#Remove sites where the ratio of QUAL to DP is deemed to be too low.
12 vcftools QUAL/DP max		#Remove sites where the ratio of QUAL to DP is deemed to be too high. Not adjustable.
13 vcftools --max-meanDP	400	#Remove sites with higher mean depth.
14 vcftools --minDP	10	#Code genotypes with lesser depth of coverage as NA.

15 vcftools --maf 0 #Remove sites with lesser minor allele frequency. Adjust based upon sample size.
15 vcftools --max-maf 1 #Remove sites with greater minor allele frequency. Adjust based upon sample size.
16 vcftools --missing-indv 0.6 #Remove individuals with more missing data.
17 vcftools --missing-sites 0.9 #Remove sites with more data missing in a pop sample.
18 filter_hwe_by_pop_HPC 0.001 #Remove sites with <p in test for HWE by pop sample. Adjust based upon sample size.
19 rad_haplotyper -d 50 #depth of sampling reads for building haplotypes.
19 rad_haplotyper -mp 1 #Remove sites with more paralogous individuals. Adjust according to sample size.
19 rad_haplotyper -u 40 #Remove contigs with more SNPs. Adjust according to sequence length.
19 rad_haplotyper -ml 10 #Remove contigs with more individuals exhibiting low coverage or genotyping errors.
19 rad_haplotyper -h 25 #Remove contigs with greater NumHaplotypes-NumSNPs.
19 rad_haplotyper -z 0.1 #Remove up to this proportion or number of reads when testing for paralogs. The more real variation in your data set, the greater this number will be. (<1) or number (>=1) of reads.
19 rad_haplotyper -m 0.5 #Keep loci with a greater proportion of haplotyped individuals.
20 OneRandSNP #Keep 1 random SNP per contig. Not adjustable. Can't be run after filter 21.
21 MostInformativeSNPs #Keep the most informative SNP per contig. Not adjustable. Can't be run after filter 20.
86 rmContigs #Remove contigs that have had SNPs removed by the previous filter. Intended to be run after filters 05, 06, 13, 14, 17, 18 if desired.

APPENDIX I

Siganus spinus ANGSD settings

SNP Calls							Frequencies Per Population			
Bam files	Number of Files	Minimum Depth	Minimum Individuals	Minimum Quality	p-value	Minimum MAF	Era	Number of Files	Minimum Individuals	Minimum Quality
All Unbaited	40	20	10	20	1.00E-06	0.01	*Albatross*	16	6	20
							Contemporary	24	20	20
All Baited	139	2	2	2	1.00E-06	0.01	*Albatross*	46	2	20
							Contemporary	93	30	20

APPENDIX J

Ambassis urotaenia ANGSD settings

SNP Calls							Frequencies Per Population			
Bam files	Number of Files	Minimum Depth	Minimum Individuals	Minimum Quality	p-value	Minimum MAF	Era	Number of Files	Minimum Individuals	Minimum Quality
All Unbaited	25	20	10	20	1.00E-06	0.01	Contemporary	25	20	20
All Baited	133	2	2	2	1.00E-06	0.01	*Albatross* Contemporary	37 96	2 30	20 20

VITA

Madeleine I. Kenton
Department of Biological Sciences, Old Dominion University, Norfolk, Virginia 23529

Education

Bachelor of Science in Marine Science (Biology), University of Tampa, May 2016

Employment

Graduate Research Assistant, Old Dominion University Department of Biological Sciences, Oct. 2017 – 2021

Division of Fishes Collection Volunteer, Smithsonian Institution Museum Support Center, Aug. 2016 – Apr. 2017

Sea Turtle Nesting Survey Research Intern, Conservancy of Southwest Florida, May 2017 – Aug. 2017

Summer Research and Education Intern, Blue Ocean Society for Marine Conservation, Jun. 2014 – Aug. 2014

Relevant Presentations

Philippine National Symposium on Marine Science July 2019
Philippine Association of Marine Science, Kalibo, Aklan (PH)

- Co-authored and Presented a poster entitled "The USS Albatross Philippines Expedition and the PIRE Project: A Historical Genomic Assessment of Fish Populations"

Presentation: Undergraduate Research Symposium April 2015
University of Tampa College of Natural and Health Sciences

- Conducted a scientific experiment, which analyzed the effects of different concentrations of vitamin C on two different strains of bacteria, with the intent to validate whether or not vitamin C is a legitimate antimicrobial agent.
- Designed and presented a poster display summarizing research and findings.
- Awarded *Best Course-Related Project*.

Publications

Suslovitch, V, **Kenton, M**, Freundt, E.C. (2015). Analysis of *Staphylococcus aureus* and *Escherichia coli* inhibition from varying concentrations of Vitamin C. *Acta Spartae*, 1(1), 1-3.