UC Santa Cruz
UC Santa Cruz Electronic Theses and Dissertations

Title
Population Genomics of Holacanthus Angelfishes in the Tropical Eastern Pacific

Permalink
https://escholarship.org/uc/item/9w60675m

Author
Gatins, Remy

Publication Date
2021

Peer reviewed|Thesis/dissertation

eScholarship.org Powered by the California Digital Library
University of California

UNIVERSITY OF CALIFORNIA
SANTA CRUZ

POPULATION GENOMICS OF *HOLACANTHUS* ANGELFISHES IN THE TROPICAL EASTERN PACIFIC

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

DOCTOR OF PHILOSOPHY

in

ECOLOGY AND EVOLUTIONARY BIOLOGY

by

Remy A. Gatins

December 2021

The Dissertation of Remy A. Gatins is approved:

_____
Professor Giacomo Bernardi, chair

_____
Professor Peter Raimondi

_____
Assistant Professor Michelle Gaither

_____
Professor Carlos A. Sánchez Órtiz

_____
Peter F. Biehl
Vice Provost and Dean of Graduate Studies

Copyright © by

Remy A. Gatins

2021

Table of Contents

LIST OF FIGURES ..... **V**

LIST OF TABLES ..... **IX**

ABSTRACT ..... **X**

DEDICATION ..... **XIII**

ACKNOWLEDGMENTS ..... **XIV**

INTRODUCTION **1**

REFERENCES 5

CHAPTER 1 WHOLE GENOME ASSEMBLY AND ANNOTATION OF THE KING ANGELFISH (*HOLACANTHUS PASSER*) GIVES INSIGHT INTO THE EVOLUTION OF MARINE FISHES OF THE TROPICAL EASTERN PACIFIC **8**

ABSTRACT 8

INTRODUCTION 9

MATERIALS AND METHODS 13

RESULTS & DISCUSSION 19

CONCLUSION 32

SUPPLEMENTARY MATERIALS 34

REFERENCES 41

CHAPTER 2 PREDICTABLE AND STOCHASTIC POPULATION GENOMIC PATTERNS OF THE WIDESPREAD KING ANGELFISH (*HOLACANTHUS PASSER*) IN THE TROPICAL EASTERN PACIFIC **47**

ABSTRACT 47

INTRODUCTION 48

MATERIALS AND METHODS 52

RESULTS 60

DISCUSSION 67

SUPPLEMENTARY MATERIALS 77

REFERENCES 82

CHAPTER 3 INCOMPLETE LINEAGE SORTING DESPITE HYBRIDIZATION IN HOLACANTHUS ANGELFISHES IN THE TROPICAL EASTERN PACIFIC 88

ABSTRACT 88

INTRODUCTION 89

MATERIALS AND METHODS 93

RESULTS 99

DISCUSSION 107

SUPPLEMENTARY MATERIALS 115

REFERENCES 119

List of Figures

Figure 0.1. *Geographic distribution of Holacanthus passer (blue), H. clarionensis (orange), and H. limbaughi (green). Black lines indicate the Sinaloan Gap (SG) and Central American Gap (CAG)..... 5*

Figure 1.1 *The King angelfish, Holacanthus passer. (A) Adult Male – white pelvic fin; (B) Adult female – yellow pelvic fin; (C) H. passer harem; (D) juvenile. Photo credits: (A,D) Remy Gatins, (B,C) Carlos A. Sánchez-Órtiz..... 13*

Figure 1.2 *Whole-genome assembly pipeline using Oxford Nanopore and Illumina sequencing. Dashed orange lines indicate quality assessment checkpoints carried out during the assembly pipeline 22*

Figure 1.3. *BUSCO completeness of the Holacanthus passer genome assembly (first row) assessed by the 4,584 orthologous actynopterygii (odb9) dataset. For comparison, we also assessed BUSCO scores for two closely related species (Pomacanthidae family: C. vrolikii, C. austriacus) and eight not closely related fish genomes to compare assemblies across fish biodiversity. 26*

Figure 1.4. *Genome completeness assessment using BUSCO v.3.0.2 of the genome assembly subsets of Holacanthus passer sequences generated with 20X, 45X and 75X coverage of long Oxford Nanopore sequences combined with 36X and 145X coverage of short Illumina sequencing reads. BUSCO completeness is based on detecting complete sequences of single-copy orthologs in the Actinopterygii (n=4,584) and Eukaryota (n=303) specific dataset. The assessment was carried out at multiple steps of the genome assembly pipeline (see Fig. 2) after the initial assembly with Wtdbg2 and consequently after each polishing event which consists of two rounds of polishing with long Nanopore reads using Racon and two rounds of polishing with short Illumina reads using Pilon. 27*

Figure 1.5. *PSMC analysis showing the demographic history (red line) of Holacanthus passer using a generation time of 5 years and a mutation rate (μ) of 10^{-8} (A) and 10^{-9} (B). Global sea level model fluctuations over the past 5 million years are shown in the background (grey) (data from de Boer et al 2014). Vertical blue bars refer to the last glacial maximum (LGM) period (~19- 26.5 kya) and the orange bar represents the closure of the Isthmus of Panama (~3.2-2.8 Mya). Triangles represent marine population expansion events previously recorded in the Tropical Eastern Pacific (see text)..... 29*

Figure 2.1. *(A) Photographs of Holacanthus passer as an adult (top) and juvenile (bottom). (B) Geographic distribution of H. passer (dark blue shade). Although not a part of H. passer's range, few vagrants have been reported at the Revillagigedo Archipelago and Clipperton Island (light blue shade). Black lines indicate the Sinaloan and Central American Gap, from North to South, respectively. Sampling locations are color coded by region: North Sea of Cortez (red), South Sea of Cortez (orange), Baja California Pacific (yellow), Mainland Mexico (green), Clipperton (purple), Panama (light blue), and Galapagos (dark blue). Sampling site key: IAG, Isla Ángel de la Guarda; SPM, San Pedro Mar; LPA, La Paz; CPU, Cabo Pulmo; LCA, Los Cabos; BMA, Bahía Magdalena; ZIH, Zihuatanejo; CLI, Clipperton; ICO_07, Isla Contadora 2007; ICO_17, Isla Contadora 2017; GAL- Galapagos. 53*

Figure 2.2. *Bayesian clustering analysis of Holacanthus passer for neutral and outlier loci, assuming no priori. Plots show K = 2 and K = 3 using 19,635 neutral loci (top) and 28 outlier loci (bottom). The most likely number of clusters based on ΔK was K = 3 for both neutral and outlier loci. Sampling sites and regions are arranged from North to South and from West to East of the TEP (See Figure 2.1 and Figure 2.1 for site details). 65*

Figure 2.3. *Discriminant analysis of principal components (DAPC) for H. passer RADseq markers showing: only neutral loci (top: a,b), only outlier loci (bottom: c,d), and with all samples*

(a,c) and after excluding Galapagos and Clipperton populations due to small sample size (b and d). Analyses retained 30 PCs and two DAs which explained 40% of the variance. 66

Figure 2.4. *Isolation by distance of 9 mainland populations using a total of 88 individuals and 19,809 SNPs. Distance represents the shortest aquatic distance between populations measured on GoogleEarth. Negative F_{ST} pairwise population comparisons were set to zero. The shaded area represents 95% confidence intervals. Reported R^2 and p-value were calculated with a Mantel test with 10,000 permutations.* 70

Figure 2.5. *Linear regression models comparing distance from the population point of origin (Panama) with (a) observed heterozygosity, (b) expected heterozysity, (c) nucleotide diversity, and (d) number of alleles. Shaded area represents 95% confidence intervals. (p-values = 0.16 < p < 0.66)* 71

Figure 2.6. *Linear regression models comparing environmental conditions to nucleotide diversity (Pi; left) and number of alleles (Na; right). Environmental conditions: SST range, sea surface temperature range (SST range); SST max, sea surface temperature max; Chlorophyll mean, chlorophyll max. Shaded area represents 95% confidence intervals. (p-values = 0.18 < p < 0.97)* 72

Figure 3.1. *Geographic distribution of Holacanthus passer (blue), H. clarionensis (orange), and H. limbaughi (green) showing sampling sites across the Tropical Eastern Pacific. Site and Region ID correspond to numbered sampling sites from the map. IAG, Isla Ángel de la Guarda; SPM, San Pedro Mar; LPA, La Paz; CPU, Cabo Pulmo; LCA, Los Cabos; BMA, Bahía Magdalena; ZIH, Zihuatanejo; ICO, Isla Contadora; SOC, Socorro Island; SBE, San Benedicto Island; RPA, Roca Partida; CLA, Clarion Island; CLI, Clipperton; GAL- Galapagos; NSC, North Sea of Cortez; SSC, South Sea of Cortez; BCP, Baja California Pacific; MEX, Mainland Mexico; PAN, Panama.* 94

Figure 3.2. *Photograph of a putative Holacanthus clarionensis-passer hybrid (left) swimming with a Holacanthus passer (right) taken off the coast in Los Cabos, Baja California Sur, Mexico. Photo credit: Remy Gatins* *95*

Figure 3.3. *Principal components analysis (PCA) of Holacanthus passer (blue), H. clarionensis (orange), H. limbaughi (green), and putative H. passer – H. clarionensis hybrids (purple) from the Tropical Eastern Pacific using 19,471 RADseq loci. Each point represents one individual fish. Percent variation explained is indicated in parenthesis for PC1 and PC2.* *103*

Figure 3.4. *Results of Bayesian clustering analysis for K = 3 using 19,471 SNPs. Each bar represents one individual fish and colors in each bar represent estimates of admixture proportion. Individuals are arranged per species and sampling region, separated by white solid bars and dotted lines, respectively. (for sampling region information see Figure 1 and Table S1).* *105*

List of Tables

Table 1.1. *Genome assembly and annotation statistics of the King angelfish (Holacanthus passer). .. 23*

Table 2.1. *Population genomic summary statistics of Holacanthus passer populations based on 19,809 RADseq loci, generated using Genodive and the Stacks. 63*

Table 2.2. *Pairwise F_{ST} values (above the diagonal) and Nei's G'_{ST} (below the diagonal) between populations based on 19,809 RADseq loci. Bold values indicate significant differentiation. Ley: IAG, Isla Ángel de la Guarda; SPM, San Pedro Martir; LPA, La Paz; CPU, Cabo Pulmo; LCA, Los Cabos; BMA, Bahía Magdalena; ZIH; Zihuatanejo; ICO_07, Isla Contadora 2007; ICO_17, Isla Contadora 2017. ..... 67*

Table 3.1. *Population genomic summary statistics of Holacanthus populations based on 20,281 RADseq loci, generated using GENODIVE and 'populations' from STACKS. Summary statistics per species were carried out using 21,020 loci with all individuals pooled by species. 101*

Table 3.2. *Effective population size (Ne) per species calculated with NeEstimator and Tajima's π ($\pi = 4 N_e \mu$). Range of Ne from Tajima's π corresponds to a mutation rate, μ, of $10^{-8} – 10^{-9}$. 102*

Table 3.3. *Pairwise F_{ST} values between sampling regions per species based on 20,248 RADseq loci. F_{ST} values are shown below the diagonal and p-values above the diagonal. Bold values indicate significant differentiation. 106*

Abstract

Population Genomics of *Holacanthus* Angelfishes

in the Tropical Eastern Pacific

By

Remy Gatins

Connectivity refers to the amount of gene flow present between populations of the same species. The transfer of genetic material between populations allows species to increase their genetic diversity, thus allowing advantageous mutations to spread. When gene flow between populations becomes restricted, each population may evolve independently, diverging into what could eventually become two different species. In reef fishes, speciation events such as these are common particularly in peripheral populations that disperse to remote islands. Species that occupy small geographic ranges (e.g., endemics) tend to have low genetic diversity, thus being more vulnerable to strong environmental changes. A region where connectivity has been relatively understudied is the Tropical Eastern Pacific (TEP), which exhibits high level of endemism of shore fishes and is biogeographically isolated from other provinces. Despite having had multiple speciation events of marine organisms in the TEP, most connectivity studies of this region show high gene flow across long distances. However, the majority of these studies rely on single mitochondrial DNA markers. In this study I use Restriction Site Associated DNA sequencing (RADseq) to obtain 1000s of loci per individual to compare intra- and inter-specific populations of

Holacanthus angelfishes in the TEP. For my first chapter, I assembled the whole genome of *Holacanthus passer* using a combination of high coverage long- (Oxford Nanopore) and short-read (Illumina) technology to further our understanding of the evolutionary history of *H. passer*. The draft genome resulted in a final assembly of 583 Mb contained in 476 contigs with a contig N50 length of 5.7 Mb. The genome contained 97.5% complete conserved actinopterygian orthologs, making it comparable, if not superior, to many chromosome-level genome assemblies of fishes. Using whole-genome sequence information, the demographic history of *H. passer* indicates a population expansion in the TEP preceded the last glacial maximum, as supported by other studies. For my second chapter, I used RADseq markers to detect genetic breaks and genetic diversity hotspots for intra-specific populations of *H. passer* across the TEP. I obtained a total of 19,809 polymorphic loci that revealed high gene flow along the TEP coastline ($F_{ST} = 0.00$) as predicted by the literature. However, pairwise differentiation detected weak but significant structure between Panama and the Sea of Cortez ($0.002 < F_{ST} < 0.005$; $0.007 < p < 0.043$), driven principally by isolation by distance. Interestingly, we detected a temporal discord between individuals collected in Panama 10 years apart and did not show this same genetic signal. In addition, we detected 28 outlier loci that revealed subtle genetic signatures that differentiated populations from the mainland and oceanic islands. My third chapter took a broader approach to assess the inter-specific genomic signatures of *Holacanthus* angelfishes in the TEP. *H. passer* is mainly found on the mainland in the TEP, while its sister species *H. clarionensis* is endemic to the Revillagigedo

Archipelago, and *H. limbaughi* is endemic to Clipperton Island. RADseq markers detected three hybrid individuals between *H. passer* and *H. clarionensis* but none with the third sister species, *H. limbaughi*. Moreover, equal amounts of ancestral variation of *H. passer* among *H. clarionensis* individuals and the lack of F2 or back-cross hybrids suggests that hybrids are sterile and provides evidence of incomplete lineage sorting. Although *H. limbaughi* and *H. clarionensis* are presumed to have diverged around the same time from *H. passer*, *H. limbaughi'*s smaller effective population size may have led to a faster rate of lineage sorting. Overall, this study highlights the power of using genome-wide markers (e.g., RADseq) to deliver a higher resolution perspective on the population dynamics within the TEP while giving insight into the evolutionary mechanisms that drove divergence of *Holacanthus* species.

Dedication

To my grandparents, Moncho and Bea

"La Vida es Bella"

Acknowledgments

The land on which I completed my Phd "is the unceded territory of the Awaswas-speaking Uypi Tribe. The Amah Mutsun Tribal Band, comprised of the descendants of indigenous people taken to missions Santa Cruz and San Juan Bautista during Spanish colonization of the Central Coast, is today working hard to restore traditional stewardship practices on these lands and heal from historical trauma (UCSC land acknowledgement statement)."

First and foremost, I would like to express my deep gratitude to my wife and family for always being by my side and supporting me throughout my academic career. You have always been my greatest fans and constantly keep me motivated. To May Roberts, for being the best house mate, office mate, travelling companion, and overall, my sister from another mother. I am so grateful to have been able to navigate graduate school together. Thank you to my PI and friend, Giacomo Bernardi, for helping me throughout this PhD and being an excellent role model for work-life balance in academia (it is possible!). To Carlos Sanchez for opening the opportunity for me to collaborate with you in Mexico and setting up such incredible fieldwork opportunities. Thank you to my committee members for all your support, patience, and guidance during this dissertation. I would like to thank all EEB professors, staff, and fellow graduate students for your lessons and discussions, which made me grow as a person and a scientist. Finally, I want to thank all my friends from Santa Cruz who really made my PhD experience unforgettable. I look forward to more laughs and adventures in the future.

This work was financially supported by the Consejo Nacional de Ciencia y Tecnología (CONACYT) and the University of California Institute for Mexico and the United States (UC-MEXUS), STARS scholarship, Formech Inspire, Quino el Guardian Liveaboard, the Graduate Student Association, and the department of Ecology and Evolutionary Biology at UCSC. I would also like to thank Dr. Luis De Leon and his lab at UMASS Boston for opening their doors to me and allowing me to join them and learn alongside them.

Introduction

Speciation, genetic diversity, and gene flow

The concept and enigmas behind speciation have been around since the time of Charles Darwin and his publication "On the Origin of Species" (Darwin 1859). To date, understanding the processes that drive the divergence of species remains of particular interest to scientists. Although, the emergence and accessibility of genetic tools has advanced our knowledge in this area, many questions remain unanswered (Schluter 2009). Speciation, the process by which new species arise, is driven by natural selection, limited gene flow, and spatial isolation. The different forms of isolation can be categorized as; complete isolation (i.e. allopatric speciation), partial isolation (i.e. peri- or parapatric speciation), or complete overlap (i.e. sympatric speciation) (Rocha and Bowen 2008a). Allopatric speciation is widely accepted as the principal driver by eliminating the effect of gene flow (Coyne and Orr 2004). However, in marine ecosystems, the lack of biogeographic barriers and the extensive dispersal potential of fishes provides an ideal study system to understand mechanisms behind speciation when gene flow is still possible (Rocha and Bowen 2008a; Bernardi 2013).

Population genetic studies are used to estimate effective population sizes, migration rates, understand population dynamics, quantify genetic diversity, identify cryptic species, detect inbreeding, study local adaptation, and identify hybridization events. Herein, a population is defined as a group of inter-breeding individuals of the same species found in a specific location or region. Population genetic signatures,

such as genetic diversity, tend to be related to population sizes, which are themselves correlated with distribution ranges (Bernardi *et al.* 2014). However, assessing population size is traditionally done using survey data. Recent developments in genomic analyses and new computational tools have allowed us to estimate population size based on genomic markers (Waples 2016).

The Tropical Eastern Pacific

The Tropical Eastern Pacific (TEP) consists of a continental corridor that runs from Baja California Sur in Mexico to the northern tip of Peru and includes seven oceanic islands or archipelagos (the Revillagigedo Archipelago, Clipperton Atoll, Cocos Island, Malpelo Island, the Galápagos Archipelago, and Easter Island) (Figure 0.1). This region is physically isolated to the East by the Isthmus of Panama (closing ~ 3 Mya) (Bellwood *et al.* 2004; O'Dea *et al.* 2016), and to the west from the Indo-Pacific by the Eastern Pacific Barrier (EPB). The EPB consists of 4000 to 7000 km of deep water that prevents most dispersers from traveling between the Central- and Eastern Pacific due to the lack of reefs to use as a stepping-stones (Lessios and Robertson 2006). It is considered the widest marine biogeographical barrier on the planet with only a few species being known to successfully cross (Lessios and Robertson 2006; Duda and Lessios 2009).

According to Robertson and Cramer (2009), the TEP is divided into three main biogeographic regions: the oceanic islands/archipelagos, and within the continental coast, the Cortez and Panamic Province. The Cortez province

encompasses the Sea of Cortez and lower Pacific Baja, while the Panamic province covers the entire southward continental coast. These distinct biogeographic provinces were defined using (i) the number of endemic fish species and (ii) species richness per area (Robertson and Cramer 2009). The continental provinces are hypothesized to be separated by the Sinaloan Gap – a long stretch with rocky reef habitat that may act as a barrier to dispersal (Figure 0.1) (Hastings 2000). However, it has also been hypothesized that the south-westward eddy found at the entrance of the Sea of Cortez, may act as a barrier separating the Cortez and Panamic province (Kurczyn *et al.* 2012). More recent studies suggest that environmental differences between the subtropical and equatorial regions may be responsible for the differences seen in species composition between the northern and southern TEP (Rocha and Bowen 2008b; Robertson and Cramer 2009; Briggs and Bowen 2012). Overall, these results suggest that the continental barrier between the Cortez and Panamic province may be driven by multiple factors.

Study species

Holacanthus angelfishes are comprised of only seven species and are particularly interesting to study evolutionary processes that drive speciation because they are thought to have diverged by all three modes of divergence described above (i.e., allopatric, parapatric, and sympatric speciation) (Alva-Campbell *et al.* 2010; Tariel *et al.* 2016). Following the closure of the Isthmus of Panama around 3.2 to 2.8 Mya (O'Dea *et al.* 2016) two clades of *Holacanthus* were separated in the Atlantic

and Pacific Oceans. These geminate clades are estimated to have diverged allopatrically approximately 1.7 to 1.4 Mya (Bellwood *et al.* 2004; Alva-Campbell *et al.* 2010; Tariel *et al.* 2016). Approximately 1.5 Mya, additional *Holacanthus* species diverged within each ocean basin. Within the Tropical Eastern Pacific (TEP), the genus *Holacanthus* is a monophyletic clade comprised of three species: *Holacanthus passer, H. clarionensis,* and *H. limbaughi. Holacanthus passer* is widely distributed along the TEP coastline, including the southern oceanic islands of Cocos, Malpelo, and Galapagos. Its sister species, *H. clarionensis* and *H. limbaughi*, in contrast are endemic to the Revillagigedo Archipelago and Clipperton Island, respectively (Figure 0.1). On the other hand, the Tropical Western Atlantic (TWA) clade is believed to have diverged in sympatry and is comprised of *H. bermudensis* and *H. ciliaris*. Finally, the last two *Holacanthus* species, *H. tricolor* and *H. africanus*, are considered the sister taxon of the TEP-TWA clade, and the most ancestral *Holacanthus* taxon, respectively (Bellwood *et al.* 2004; Alva-Campbell *et al.* 2010; O'Dea *et al.* 2016; Tariel *et al.* 2016). Moreover, putative hybrids have been observed showing mixed phenotypic marking between *H. passer* and *H. clarionensis* off the tip of Baja California (Sala *et al.* 1999, RG *personal observation*), however, none has been reported with *H. limbaughi*.

Holacanthus angelfishes are protogynous sequential hermaphrodites, changing sex from female to male as they grow. Their pelagic larval duration (PLD) is estimated to be between 23 –26 days based on data from the closest relative to *Holacanthus*, *Pygoplites diacanthus* (Thresher and Brothers 1985; Alva-Campbell *et*

al. 2010). *Holacanthus* exhibit sexual dimorphism (pelvic fin coloration) (Moyer *et al.* 1983) and can partition their habitat by sex and size classes (Aburto-Oropeza *et al.* 2000). They are important sponge feeders and herbivores, but have been observed feeding in the water column on fish feces (Aburto-Oropeza *et al.* 2000; Sánchez-Alcántara *et al.* 2006) and interacting as fish cleaners (Quimbayo *et al.* 2017). Additionally, their social organization can vary from solitary individuals to harems (Moyer *et al.* 1983). Little information is available regarding observations of *H. clarionensis* and *H. limbaughi* due to the difficulty of accessing the Revillagigedo and Clipperton Islands.

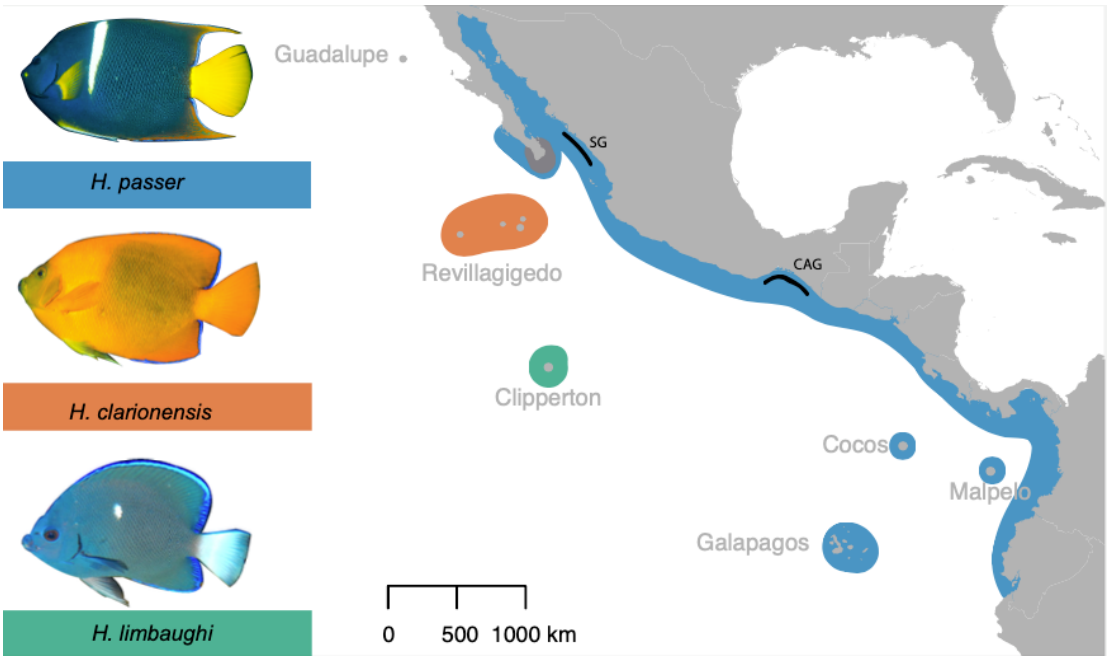


Figure 0.1. Geographic distribution of *Holacanthus passer* (blue), *H. clarionensis* (orange), and *H. limbaughi* (green). Black lines indicate the Sinaloan Gap (SG) and Central American Gap (CAG).

References

Aburto-Oropeza, O., E. Sala, and C. Sánchez-Ortiz, 2000 Feeding behavior, habitat use, and abundance of the angel fish *Holacanthus passer*. Environmental Biology of Fishes 57:.

Alva-Campbell, Y., S. R. Floeter, D. R. Robertson, D. R. Bellwood, and G. Bernardi, 2010 Molecular phylogenetics and evolution of Holacanthus angelfishes (Pomacanthidae). Mol Phylogenet Evol 56: 456–461.

Bellwood, D. R., L. van Herwerden, and N. Konow, 2004 Evolution and biogeography of marine angelfishes (Pisces: Pomacanthidae). Mol Phylogenet Evol 33: 140–155.

Bernardi, G., 2013 Speciation in fishes. Mol Ecol 22: 5487–5502.

Bernardi, G., M. L. Ramon, Y. Alva-Campbell, J. E. McCosker, G. Bucciarelli *et al.*, 2014 Darwin's fishes: phylogeography of Galápagos Islands reef fishes. B Mar Sci 90: 533–549.

Briggs, J. C., and B. W. Bowen, 2012 A realignment of marine biogeographic provinces with particular reference to fish distributions. J Biogeogr 39: 12–30.

Coyne, J. A., and H. A. Orr, 2004 *Speciation*. Sinauer Associates, Inc, Sunderland, MA.

Darwin, C., 1859 *On the Origin of Species by Means of Natural Selection or Preservation of Favoured Races in the Struggle for Life*. Murray, London.

Duda, T. F., and H. A. Lessios, 2009 Connectivity of populations within and between major biogeographic regions of the tropical Pacific in Conus ebraeus, a widespread marine gastropod. Coral Reefs 28: 651–659.

Hastings, P. A., 2000 Biogeography of the Tropical Eastern Pacific: distribution and phylogeny of chaenopsid fishes. Zool J Linn Soc-lond 128: 319–335.

Kurczyn, J. A., E. Beier, M. F. Lavín, and A. Chaigneau, 2012 Mesoscale eddies in the northeastern Pacific tropical-subtropical transition zone: Statistical characterization from satellite altimetry. J Geophys Res Oceans 1978 2012 117: n/a-n/a.

Lessios, H. A., and D. R. Robertson, 2006 Crossing the impassable: genetic connections in 20 reef fishes across the eastern Pacific barrier. Proc Royal Soc B Biological Sci 273: 2201–2208.

Moyer, J. T., R. E. Thresher, and P. L. Colin, 1983 Courtship, spawning and inferred social organization of American angelfishes (Genera *Pomacanthus*, *Holacanthus* and *Centropyge*; pomacanthidae). Environ Biol Fish 9: 25–39.

O'Dea, A., H. A. Lessios, A. G. Coates, R. I. Eytan, S. A. Restrepo-Moreno *et al.*, 2016 Formation of the Isthmus of Panama. Sci Adv 2: e1600883.

Quimbayo, J. P., M. S. Dias, O. R. C. Schlickmann, and T. C. Mendes, 2017 Fish cleaning interactions on a remote island in the Tropical Eastern Pacific. Mar Biodivers 47: 603–608.

Robertson, D., and K. Cramer, 2009 Shore fishes and biogeographic subdivisions of the Tropical Eastern Pacific. Mar Ecol Prog Ser 380: 1–17.

Rocha, L. A., and B. W. Bowen, 2008a Speciation in coral-reef fishes. J Fish Biol 72: 1101–1121.

Rocha, L. A., and B. W. Bowen, 2008b Speciation in coral-reef fishes. J Fish Biol 72: 1101–1121.

Sala, E., O. Aburto-Oropeza, and J. L. Arreola-Robles, 1999 Observations of a Probable Hybrid Angelfish of the Genus Holacanthus from the Sea of Cortez, México. Pacific Science 53: 181–184.

Sánchez-Alcántara, I., O. Aburto-Oropeza, E. F. Balart, A. L. Cupul-Magaña, H. Reyes-Bonilla *et al.*, 2006 Threatened Fishes of the World: *Holacanthus passer* Valenciennes, 1846 (Pomacanthidae). Environ Biol Fish 77: 97–99.

Schluter, D., 2009 Evidence for Ecological Speciation and Its Alternative. Science 323: 737–741.

Tariel, J., G. C. Longo, and G. Bernardi, 2016 Tempo and mode of speciation in *Holacanthus* angelfishes based on RADseq markers. Mol Phylogenet Evol 98: 84–88.

Thresher, R. E., and E. B. Brothers, 1985 Reproductive Ecology and Biogeography of Indo-West Pacific Angelfishes (Pisces: Pomacanthidae). Evolution 39: 878.

Waples, R. S., 2016 Tiny estimates of the Ne/N ratio in marine fishes: Are they real? J Fish Biol 89: 2479–2504.

Chapter 1 Whole genome assembly and annotation of the King angelfish (*Holacanthus passer*) gives insight into the evolution of marine fishes of the Tropical Eastern Pacific

Abstract

Holacanthus angelfishes are some of the most iconic marine fishes of the Tropical Eastern Pacific (TEP). However, very limited genomic resources currently exist for the genus. In this study we: i) assembled and annotated the genome of the King angelfish (*Holacanthus passer)*; ii) assessed the optimal combination of long- Oxford Nanopore (ONT) and short- Illumina reads in genome quality and completeness; and iii) examined the demographic history of *H. passer* in the TEP. We generated 43.8 Gb of ONT and 97.3 Gb Illumina reads representing 75X and 167X coverage, respectively. The final genome assembly size was 583 Mb with contig N50 of 5.7 Mb, which captured 97.5% complete Actinoterygii Benchmarking Universal Single-Copy Orthologs (BUSCO's). Repetitive elements account for 5.09% of the genome, and 33,889 protein-coding genes were predicted, of which 22,984 have been functionally annotated. Our coverage comparisons show that high ONT coverage improved overall assembly contiguity, from 804 to 486 contigs, representing 20X and 75X, respectively. However, although short-read Illumina sequences are crucial to improve genome completeness, coverage variability between 36X and 145X showed no significant difference in genome quality. Our demographic model suggests that population expansions of *H. passer* occurred prior to the last glacial maximum

(LGM) and were more likely shaped by events associated with the closure of the Isthmus of Panama. Overall, this annotated genome assembly will serve as a resource to improve our understanding of the evolution of *Holacanthus* angelfishes while facilitating novel research into local adaptation, speciation, and introgression in marine fishes.

Keywords: Pomacanthidae; genome assembly; whole genome; long reads; hybrid assembly; coverage; tropical eastern pacific; demographic history; Nanopore; Illumina

Introduction

The King angelfish, *Holacanthus passer*, is one of the most iconic fish species of the Tropical Eastern Pacific (TEP) (Figure 1.1). Its distribution ranges from the northern Gulf of California (Sea of Cortez) to Peru, including the Revillagigedos, Cocos, Malpelo, and the Galápagos Islands (Allen and Robertson 1994; Sánchez-Alcántara *et al.* 2006). Due to their conspicuous coloration, the King angelfish have become a target for the aquarium trade (Sánchez-Alcántara *et al.* 2006), with individuals costing between $150 and $900, while individuals of the sister species, *H. clarionensis*, endemic to the Revillagigedos, can be sold for up to $15,000 (https://www.bluezooaquatics.com). *Holacanthus passer* is currently protected from harvest in Mexico (Norma Official Mexicana) (Sánchez-Alcántara *et al.* 2006), but is identified as having a stable population under the IUCN red list (Pyle *et al.* 2010).

Holacanthus angelfishes are protogynous sequential hermaphrodites, changing sex from female to male as they grow. They exhibit sexual dimorphism (pelvic fin coloration) (Figure 1.1) (Moyer *et al.* 1983), and can partition their habitat by sex and size classes (Aburto-Oropeza *et al.* 2000). They are important sponge feeders and herbivores, but have been observed feeding in the water column on fish feces (Aburto-Oropeza *et al.* 2000; Sánchez-Alcántara *et al.* 2006) and interacting as fish cleaners (Quimbayo *et al.* 2017). Additionally, their social organization can vary from solitary individuals to harems (Moyer *et al.* 1983).

The genus *Holacanthus* is an interesting model system for assessing the drivers of diversification in marine fishes. Although it is comprised of only seven species, the genus presents a complex history of diversification, which includes three modes of speciation: allopatric, peripatric, and sympatric (Alva-Campbell *et al.* 2010; Tariel *et al.* 2016). Following the closure of the Isthmus of Panama around 3.2 to 2.8 Mya (O'Dea *et al.* 2016), two clades of *Holacanthus* were separated in the Atlantic and Pacific Oceans by the newly formed Isthmus. These so-called geminate species (Jordan 1908) are estimated to have diverged allopatrically approximately 1.7 to 1.4 Mya (Bellwood *et al.* 2004; Alva-Campbell *et al.* 2010; Tariel *et al.* 2016) along with about 40 other marine fishes (Jordan 1908; Thacker 2017) and many invertebrates (Lessios 1981; Miura *et al.* 2010). Additional *Holacanthus* species diverged, within each ocean basin, approximately 1.5 Mya. The Tropical Eastern Pacific (TEP) clade, which consists of *H.* passer, *H. limbaughi*, and *H. clarionensis*, is thought to have diverged via peripatry. In contrast, the Tropical Western Atlantic (TWA) clade,

comprised by *H. bermudensis* and *H. ciliaris*, is thought have diverged in sympatry (Alva-Campbell *et al.* 2010; Tariel *et al.* 2016). The last two *Holacanthus* species, *H. tricolor* and *H. africanus*, are considered the sister taxon of the TEP-TWA clade, and the most ancestral *Holacanthus* taxon, respectively. To facilitate the study of the history of diversification in *Holacanthus*, here we assemble a reference genome for King angelfish (*H. passer*), one of the most widespread species of the genus.

The increased accessibility of novel genomic tools has led to a rapid proliferation of whole-genome assemblies for non-model species. In particular, recent genome assemblies have used of a combination of short and accurate (~99%) Illumina data with long, but less accurate reads (~95%) generated by Oxford Nanopore (ONT) or PacBio sequencing (Fernandez-Silva *et al.* 2018; Tan *et al.* 2018; Lehmann *et al.* 2019; Shafin *et al.* 2019; Johnson *et al.* 2020). Although the ONT long-read sequencing has an error rate between 5-15% (Jain *et al.* 2016; Rang *et al.* 2018), it can deliver real-time targeted sequencing, while improving genome assembly contiguity and completeness (Austin *et al.* 2017; Tan *et al.* 2018; Shafin *et al.* 2019; Johnson *et al.* 2020). Thus, the combination of both sequencing technologies can be an effective method to generate highly accurate reference genomes for non-model organisms. However, deciding the optimal combination of coverage from both technologies needed to obtain a high quality genome on a limited budget has received little attention to date. As a general rule, the higher the coverage the higher the confidence in nucleotide sequencing accuracy, which can improve overall genome assembly, given that higher coverage can lead to more overlapping

sequences. However, this general rule of thumb may ignore relevant questions regarding the cost-efficiency of using both technologies for *de novo* genome assemblies. For instance, how much can coverage of long vs. short reads affect the quality of genome assembly? Is there a minimum amount of coverage needed from each technology for an accurate genome assembly? Furthermore, at what point does an assembly stop benefiting from the addition of reads? We assess this issue by assembling six fish genomes and comparing the quality and completeness of those assemblies with subsets of variable coverage of raw ONT and Illumina data obtained for the King angelfish genome.

The main goals of this study are: i) to deliver a high-coverage whole genome assembly of the King angelfish, *Holacanthus passer*; ii) to assess the optimal combination of coverage of both long- and short-reads in the quality and completeness of hybrid genome assembly in non-model organisms; and iii) to examine the demographic history of *H. passer* in the TEP using *de novo* genome sequence data. Overall, these genomic resources will facilitate future studies of the evolutionary history of the genus *Holacanthus* and its population dynamics in the TEP. In addition, they will add to the growing knowledge of long-read technology, hybrid assemblies, and fish genomics using Oxford Nanopore and Illumina technologies.

Figure 1.1 The King angelfish, *Holacanthus passer*. (A) Adult Male – white pelvic fin; (B) Adult female – yellow pelvic fin; (C) *H. passer* harem; (D) juvenile. Photo credits: (A,D) Remy Gatins, (B,C) Carlos A. Sánchez-Órtiz.

Materials and Methods

Sample collection and DNA extraction

Fin and gill clips were collected from 13 individuals of *Holacanthus passer* in La Paz, Baja California Sur, Mexico. Collections were made with pole spears while SCUBA diving, abiding by UCSC IACUC protocols. Tissue samples were immediately placed in 95% ethanol and stored at -20°C. DNA was extracted using a DNeasy Blood and Tissue kit according to manufacturer's protocol (Qiagen).

DNA quality and concentration of the 13 samples were assessed using a Nanodrop 2000c and Qubit 4.0 Fluorometer. The sample with the highest quality was

further evaluated on an Agilent 2200 TapeStation DNA ScreenTape to check high molecular weight. The sample chosen to carry out the genome assembly of *Holacanthus passer* had a final DNA concentration of 205 ng/µl, a 260/280 and 260/230 ratio of 2.02 and 2.26, respectively, and an average fragment length of 38 Kb (Additional File: Figure S1A). This sample came from an adult *H. passer* female with a total length size of 20.4 cm. Before beginning with our library prep, DNA was transferred from the AE buffer, provided in the Qiagen kit, to EB to remove traces of EDTA, as recommended by Nanopore library prep, using a 3X KAPA Pure Bead clean up (Roche Molecular Systems). DNA was eluted in 90 µl of EB, reaching a final concentration of 128 ng/µl. This sample was sequenced using ONT and Illumina (HiSeq4000; 150 bp paired-end) sequencing.

Whole-genome library construction and sequencing

DNA was first sheared using the Covaris g-TUBE following the manufacturer's protocol for 10 Kb fragments (Additional File: Figure S1B). Four individual ONT libraries were prepared with 1.5 µg of DNA using the SQK-LSK109 library prep protocol according to manufacturer's protocol (Oxford Nanopore Technologies, Oxford, UK). Each library was sequenced on a R9.4 flow cell using the MinION DNA sequencer. Maximum run time ranged between 48 to 72 hours. Raw data was basecalled separately using Guppy 3.3 basecaller on a GPU-based high-performance computer cluster server of the University of Massachusetts at Boston. A total of 43.8 Gb (N_{50}: 6626 bp, longest read: 474 205 bp) were generated on the Oxford Nanopore

MinION device. Individual MinION sequencing statistics can be found in Additional File: Table S1.

The Illumina library was prepared with 250 ng of the same DNA as above using the Kapa Hyperplus Library Preparation Kit by modifying all volume reactions to use only one third of the volumes described in the manufacturer's protocol (Kapa Biosystems, Wilmington, MA). The total fragmentation volume was 16.66 μl and was incubated at 37°C for 7:45 min. The incubation parameters were previously optimized to target fragments of ~500 bp. Post-ligation purification was done using a 0.8X KAPA Pure bead cleanup. Library amplification was carried out with a total PCR reaction volume of 16.6 μl for 8 PCR thermal cycles. Finally, we did a double size-selection post-amplification cleanup with SPRIselect beads using a 0.56X upper and 0.72X lower selection ratio (Beckman Coulter, Inc). The final Illumina library was sequenced in a pool of three individuals with a HiSeq4000 (150 bp paired-end) (Novogene Corporation Inc.), which generated a total of 97.3 Gb of sequence data with an average cleaned read length of 149 bp.

GenomeScope (Vurture *et al.* 2017) was used to estimate genome size, repeat content, and heterozygosity across all k-mers (k = 21) previously detected using Jellyfish v2.2.10 (Marçais and Kingsford 2011) to help choose parameters for downstream analysis. Using only raw Illumina data, the genome size of *H. passer* was estimated to have a length of 579 Mb with approximately 95.1% of unique content and a heterozygosity level of 0.43% (Additional File: Figure S3). Additionally, k-mers with 110X coverage showed the highest frequency. Considering a genome size

of 579 Mb, the output of 43.8 Gb of ONT and 97.3 Gb of Illumina reads represented a total of 75X and 167X coverage respectively, based on the size of our final genome assembly.

Genome assembly

Long reads obtained from the ONT were concatenated into one large fastq file and trimmed with Porechop v. 0.2.3 (https://github.com/rrwick/Porechop). Nanofilt v. 2.5.0 (https://github.com/wdecoster/nanofilt) was used to create two different filtered datasets to help the contiguity of the final assembly. The first filtered dataset was used to keep the longest reads and to obtain an initial more contiguous assembly (Nanofilt parameters -q 3; -l 1000). The second filtered dataset was explicitly used for downstream assembly polishing (-q 5 and -l 500). The former sequences were assembled using Wtdbg2 v2.5 (Ruan and Li 2019), setting a minimum sequence length of 1000 bp (-L 1000). In order to improve the draft assembly, two rounds of consensus correction were performed using the -q 5 filtered ONT reads by mapping reads to the draft genome with Minimap2 v. 2.17 and polishing with Racon v. 1.4.7.

The shorter but more accurate Illumina reads were used to further polish the ONT genome. Raw sequences were adapter-trimmed with Trimmomatic v. 0.39 (Bolger *et al.* 2014) and quality checked before and after trimming using FastQC v 0.11.8 (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Two rounds of polishing were carried out by mapping the trimmed Illumina reads to the assembly

using BWA v 0.7.17 (Li and Durbin 2009), sorted and indexed with Samtools v 1.9 (Li *et al.* 2009), and consensus corrected using Pilon v 1.23 (Walker *et al.* 2014).

Finally, given that the DNA used for the genome assembly was extracted from gill tissue, which could be more exposed to microorganisms, the final assembly was screened for sequences of bacteria, viruses, and plasmids using Kraken 2.0.9 (Wood and Salzberg 2014), resulting in the removal of 2% of the assembly. Genome completeness was assessed using Benchmarking Universal Single-Copy Orthologs (BUSCO v3.0.2) (Simao *et al.* 2015; Waterhouse *et al.* 2017) by comparing the *H. passer* genome to the Actinopterygii (n = 4,584) and Eukaryota (n = 303) ortholog gene datasets. Assembly statistics and BUSCO completeness were assessed after the initial draft assembly, and subsequently, after each polishing iteration (Figure 1.1). The complete flow chart of the full genome assembly pipeline is shown in Figure 1.2.

Genome annotation

To annotate our genome, we used the homology-based gene prediction pipeline GeMoMa (v1.6.4). GeMoMa uses protein-coding genes models and intron position conservation from reference genomes to predict possible protein-coding genes in a target genome (Keilwagen et al., 2018). Here, we run GeMoMa pipeline using annotations from three fish species: *Amphiprion ocellaris, Oreocromis niloticus, Electrophorus electricus* (downloaded from NCBI see links in Additional File: Table S2). In our particular case, the pipeline performed four main steps: 1) Extractor or external search, using the search algorithm tbalstn with cds parts as queries from our

reference genomes, 2) Gene Model Mapper (GeMoMa), which builds gene models from the extractor results , 3) GeMoMa Annotation Filter (GAF) that filters and combines common gene predictions and 4) AnnotationFinalizer, which predicts UTRs for annotated coding sequences and generates gene and transcript names (Keilwagen et al., 2018). Additionally, Repetitive elements were predicted by running RepeatMasker (open-4.0.6, Smit et al. 2013–2015) with the Teleostei database to identify repetitive elements in the genome and soft-mask the assembly. RepeatMasker.out was converted to GFF with RepeatMasker script rmOutToGFF3.pl.

Comparing subsets of ONT and Illumina coverage genome assemblies

To assess the impact that the coverage of long- and short- reads can have on a hybrid whole-genome assemblies, we assembled six draft genomes using different subsets of raw *H. passer* sequence read data for both sequencing technologies. Thus, we randomly sampled sequence data to 25X, 45X and 75X coverage of long ONT reads and combined with 36X and 145X coverage of short-read Illumina data. The same genome assembly pipeline as described above (Figure 1.2) was used to assemble all six subset assemblies. Similarly, genome completeness was assessed as previously described using BUSCO v3.0.2 (Simao *et al.* 2015; Waterhouse *et al.* 2017). In particular, we assessed completeness after each initial assembly with Wtdbg2 and after each subsequent polishing iteration to visualize the effect of coverage throughout the assembly pipeline.

Inferring demography history in H. passer

In order to infer the demographic history of *H. passer* in the TEP, a Pairwise Sequentially Markovian Coalescent (PSMC) model was used to explore temporal changes in effective population size based on genome-wide diploid sequence data (Li and Durbin 2011). The PSMC analysis is particularly powerful to infer demographic histories beyond 20,000 years, which fits well with the known history of the *Holacanthus* genus (Alva-Campbell *et al.* 2010; Tariel *et al.* 2016). The PSMC simulation was run with 30 iterations (-N), a maximum 2N0 coalescent time of 30 (-t), initial theta/rho ratio of 5 (-r), and the pattern parameter (-p) set to "4+30*2+4+6+10" (Li and Durbin 2011; Liu and Hansen 2017). Generation time (g) is defined as the age at which half of the individuals of the population are reproducing. Given that *H. passer* is protogynous, generation time for females is around three years, while for males it is around six years, after they transition from female to male (Hernández 1998; Arellano-Martínez *et al.* 1999; Sánchez-Alcántara *et al.* 2006). Thus, we set the average generation time (-g) for *H. passer* to 5 years. Mutation rate (μ) per site per generation in fishes has been estimated to be between 10^{-8} to 10^{-9} mutations per site (Brumfield *et al.* 2003; Crane *et al.* 2018), thus we ran two simulations to represent the range of the expected mutation rates.

Results & Discussion

Genome assembly

The final assembled and polished genome of *Holacanthus passer* yielded a total size of ~583 Mb gathered in 476 contigs, with the largest contig at 17 Mb and a N50 of 5.7 Mb. The final assembly was slightly larger than the initial ~579 Mb estimated by GenomeScope as well as the 581 Mb assembly before the polishing iterations. Detailed assembly statistics can be found in Figure 1.1. The number of contigs remained at 486 contigs throughout the assembly until the last step when we removed 2% of the assembly due to contamination, leaving a total of 476 contigs. After four iterations of polishing using ONT and Illumina reads, BUSCO completeness improved from 82.4% to 97.5% and 90.1% to 95.4% in the Actinopterygii (n = 4,584) and Eukaryota (n = 303) dataset, respectively. The largest completeness increase (10.6%) in the BUSCO Actinopterygii dataset occurred after the first ONT polishing iteration, while in the Eukaryota dataset for both the first ONT polishing and the second Illumina polishing iteration showed the highest increase (2.3%) (Figure 1.1). Additionally, the N50 contig length increased from 5.6 to 5.7 Mb after polishing. These results indicate that polishing with both ONT and Illumina reads greatly improved the assembly, by correcting assembly bases, fixing misassembles, and filling assembly gaps. Moreover, contiguity did not improve after the initial assembly which was carried out with the Wtdbg2 assembler using long ONT reads. This suggests that the assembler and initial input reads play an important role in how contiguous the assembled genome will be, while multiple polishing iterations will further improve upon the accuracy of the assembly.

The *H. passer* genome assembly presented here is comparable in quality to other recently published fish genomes. To the best of our knowledge, the closest other available genome assemblies, which belong to the same family, Pomacanthidae, is *Centropyge vrolikii* (Fernandez-Silva *et al.* 2018) and its sister family, Chaetodontidae, *Chaetodon austriacus* (DiBattista *et al.* 2018), exhibit slightly larger genome sizes of 696.5 Mb and 712.2 Mb, respectively. Our *H. passer* genome resulted in a much more contiguous assembly (contigs: *H. passer*, 450; *C. vrolikii*, 30,500; *C. austriacus*, *13*,441) and a N50 of 5.7 Mb that is smaller than the N50 of *C. vrolikii* (9 Mb), but larger than that of *C. austriacus* (0.17 Mb) (Additional File: Table S2). Regarding genome completeness, *H. passer* showed a slightly higher number of complete orthologous matches in BUSCO using the Actinopterygii (odb9) dataset than the *C. vrolikii* and *C. austriacus* assemblies (Figure 1.3). When compared with numerous other recently published chromosome level fish genomes, *H. passer* showed comparable, if not higher, BUSCO scores despite not being a chromosome level assembly (Figure 1.3). In general, our assembly is highly contiguous with zero gaps, which could result in less fragmented genes. Overall, this *H. passer* assembly will serve as a high-quality genomic reference assembly for the Pomacanthidae family, and it exemplified how N50 values do not always correlate with the best BUSCO scores as outlined in Jauhal and Newcomb (2021).

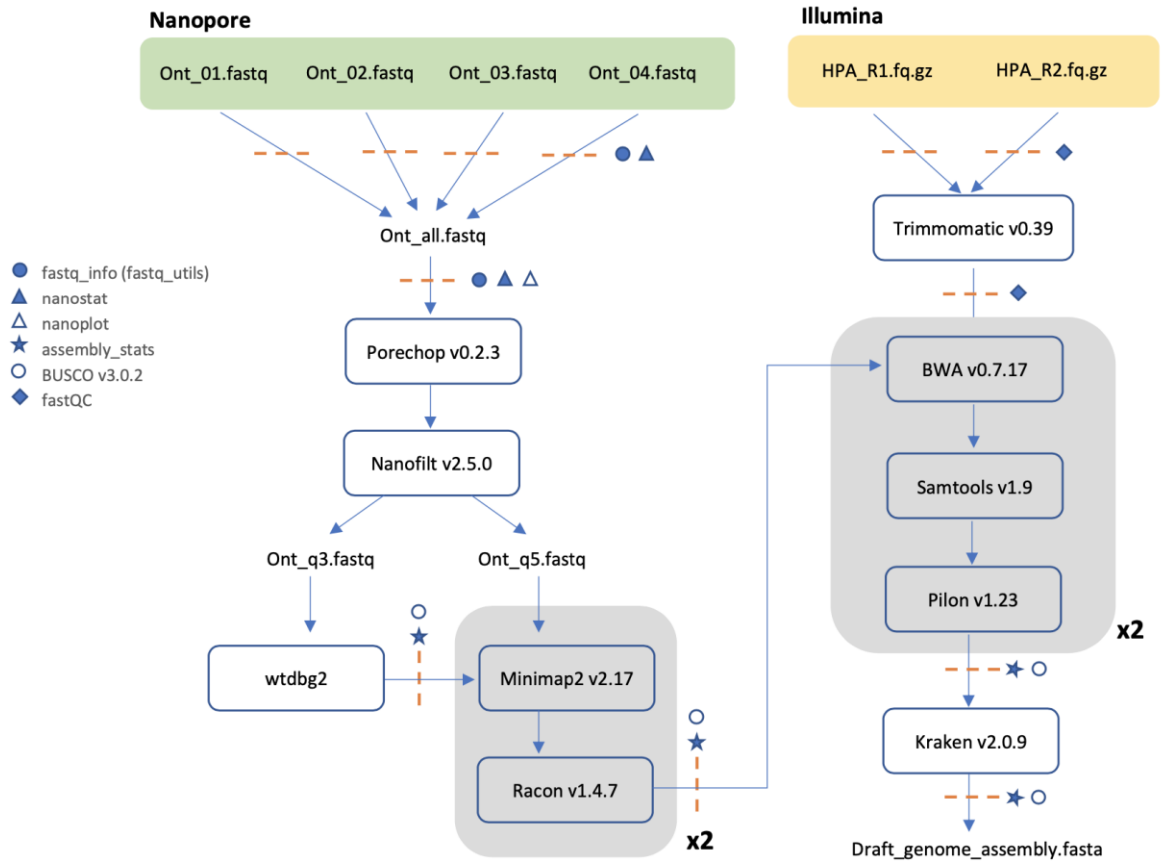


Figure 1.2 Whole-genome assembly pipeline using Oxford Nanopore and Illumina sequencing. Dashed orange lines indicate quality assessment checkpoints carried out during the assembly pipeline

Table 1.1. Genome assembly and annotation statistics of the King angelfish (*Holacanthus passer*).

	Nanopore			Nanopore + Illumina		
Genome assembly	**Wtdbg2**	**Wtdbg2 + 1X Racon**	**Wtdbg2 + 2X Racon**	**Wtdbg2 + 2X Racon + 1X Pilon**	**Wtdbg2 + 2X Racon + 2X Pilon**	**Wtdbg2 + 2X Racon + 2X Pilon + Kraken**
Total assembly size of contigs (bp)	581 422 425	583 574 933	583 552 491	583 601 337	583 528 366	583 428 144
Number of contigs	486	486	486	486	486	476
N50 contig length (bp)	5 681 869	5 707 473	5 709 778	5 708 674	5 708 022	5 708 022
N90 contig length (bp)	997 074	1 000 168	1 000 597	1 000 715	1 000 532	1 000 532
Longest contig (bp)	17 088 287	17 147 963	17 147 963	17 150 647	17 148 928	17 148 928
GC/AT/N, %						
Actinopterygii						
Complete BUSCOs	3779 (82.4%)	4263 (93%)	4296 (93.7%)	4468 (97.5%)	4471 (97.5%)	4471 (97.5%)
Complete and single-copy BUSCOs	3674 (80.1%)	4133 (90.2%)	4163 (90.8%)	4364 (95.2%)	4368 (95.3%)	4368 (95.3%)
Complete and duplicated BUSCOs	105 (2.3%)	130 (2.8%)	133 (2.90%)	104 (2.3%)	103 (2.2%)	103 (2.2%)
Fragmented BUSCOs	374 (8.2%)	176 (3.8%)	155 (3.40%)	38 (0.8%)	37 (0.8%)	37 (0.8%)
Missing BUSCOs	431 (9.4%)	145 (3.2%)	133 (2.9%)	78 (1.7%)	76 (1.7%)	76 (1.7%)
Eukaryota						
Complete BUSCOs	273 (90.1%)	280 (92.4%)	280 (92.4%)	282 (93.10%)	289 (95.4%)	289 (95.4%)
Complete and single-copy BUSCOs	267 (88.1%)	270 (89.10%)	270 (89.10%)	267 (88.1%)	274 (90.4%)	274 (90.4%)
Complete and duplicated BUSCOs	6 (2.0%)	10 (3.3%)	10 (3.3%)	15 (5%)	15 (5%)	15 (5%)
Fragmented BUSCOs	4 (1.3%)	3 (1%)	4 (1.3%)	2 (0.7%)	2 (0.7%)	2 (0.7%)
Missing BUSCOs	26 (8.6%)	20 (6.60%)	19 (6.3%)	19 (6.2%)	12 (3.9%)	12 (3.9%)
Annotation						
Number of protein-coding genes						33889
Mean gene length (bp)						
Number of CDSs						392382
Longest gene (bp)						
Functionally annotated						22984

Genome Annotation

RepeatMasker estimated that 5.09% of the genome consisted of repetitive sequences, primarily LINEs (0.85%), LTR elements (0.31%), DNA transposons (1.36%) and simple repeats (2.14%) (Additional File: Table S3). Repeat content was nearly identical to that estimated by GenomeScope (4.9%). GeMoMa identified 33,889 gene models and 392,382 CDSs, where 67.8% (22,984) of the gene models had a functional annotation (Figure 1.1). The number of coding sequences identified for *H. passer* was within the rage of those found in other closely related fish species genomes (see https://www.ncbi.nlm.nih.gov/genome/annotation_euk/all/; assembled and annotated fish genomes, visited April 28, 2021).

Comparing subsets of ONT and Illumina coverage genome assemblies

Overall, BUSCO scores assessing genome completeness after four polishing iterations were reasonably similar for all six datasets, ranging from 96.4% to 97.6% and 94.1% to 95.4% in Actinopterygii and Eukaryota dataset, respectively (Figure 1.4; Additional File: Table S3). The most significant change in BUSCO scores was seen after the initial assembly step before downstream polishing (Figure 1.4). Assemblies created with long-read 20X coverage had lower BUSCO scores with more than a 10% difference from assemblies carried out with a 45X or 75X coverage. However, after the first polishing step all BUSCO scores increased to comparable values. The subsequent most crucial increase in BUSCO scores occurred after the third polishing step, which was the first polishing step that incorporated the Illumina

short reads using Pilon. Additionally, the most meaningful difference between assemblies was how contiguous each assembly was, dependent on the amount of long-read coverage. Assemblies with a long-read coverage of 20X, 45X, and 75X had 804, 528, and 486 contigs, respectively, suggesting that assemblies with a greater amount of long-read coverage were more contiguous. Illumina short-read coverage did not affect contiguity. Genome size across assemblies ranged from 577 Mb to 583.5 Mb, with the shortest genome being from a medium coverage dataset (long45X_short36X). As previously mentioned, consensus polishing helps correct assembly bases, fix misassembles, and fill assembly gaps; however, this might result in removing sequences that would affect the genome size. N50 values ranged from 1,844,885 to 5,708,366 bp and followed an expected trend from low to high values with increasing long and short-read coverage. Thus, our results indicate that although having a high coverage genome is desirable, especially considering assembly contiguity and N50 values, a relatively low coverage dataset of 20X long-read and 36X short-reads has enough power to assemble a genome with high genome completeness comparable to higher coverage genomes and, perhaps, at a lower cost.

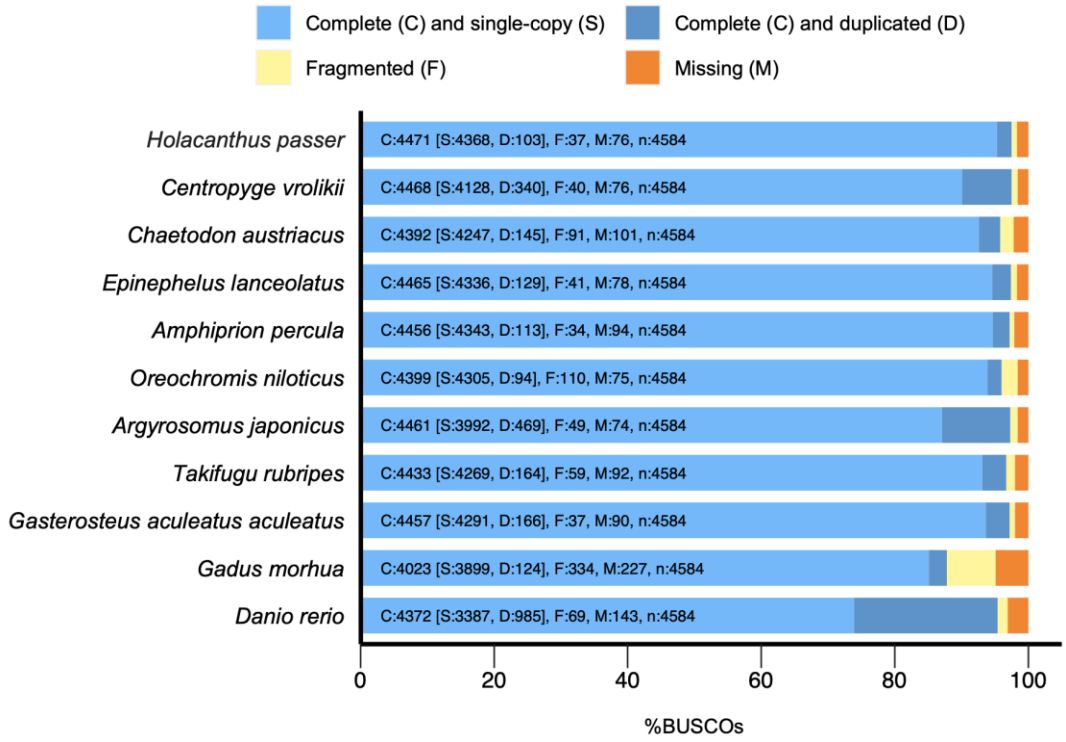


Figure 1.3. BUSCO completeness of the *Holacanthus passer* genome assembly (first row) assessed by the 4,584 orthologous actynopterygii (odb9) dataset. For comparison, we also assessed BUSCO scores for two closely related species (Pomacanthidae family: *C. vrolikii*, *C. austriacus*) and eight not closely related fish genomes to compare assemblies across fish biodiversity.

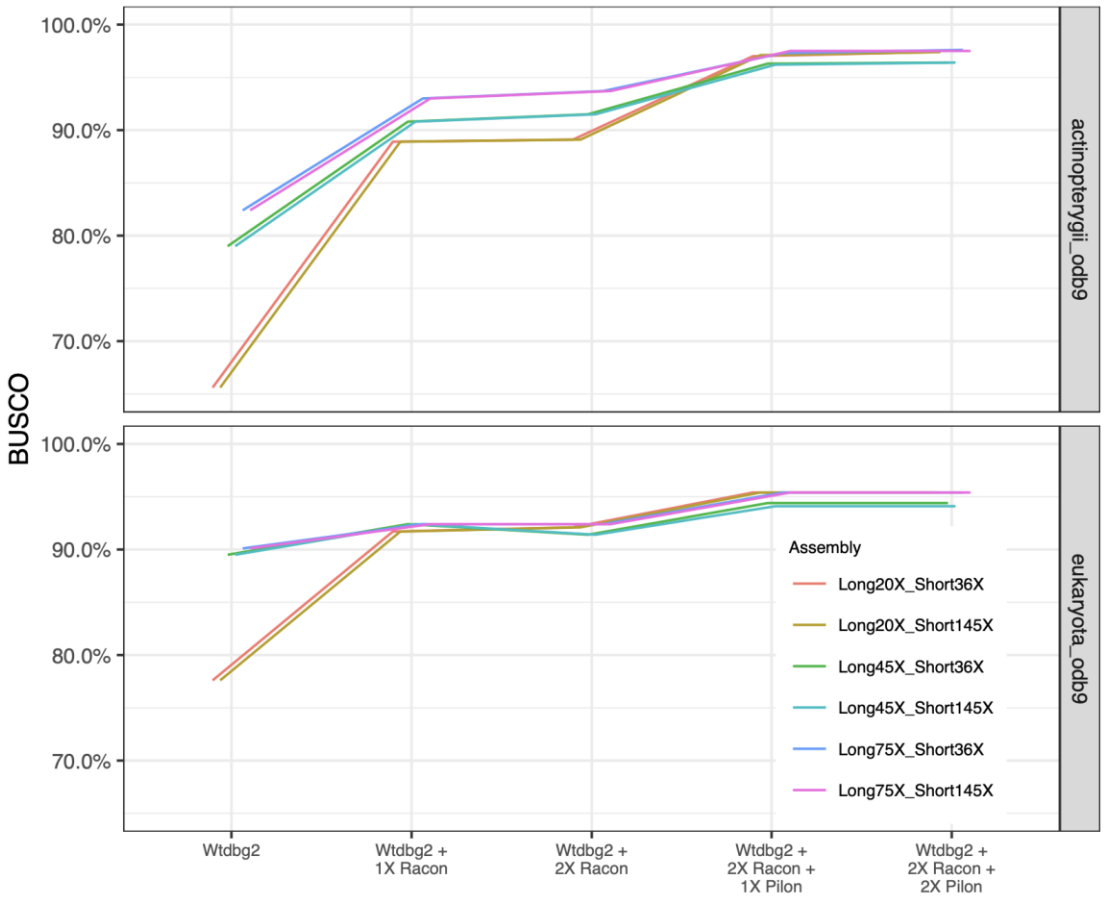


Figure 1.4. Genome completeness assessment using BUSCO v.3.0.2 of the genome assembly subsets of *Holacanthus passer* sequences generated with 20X, 45X and 75X coverage of long Oxford Nanopore sequences combined with 36X and 145X coverage of short Illumina sequencing reads. BUSCO completeness is based on detecting complete sequences of single-copy orthologs in the Actinopterygii (n=4,584) and Eukaryota (n=303) specific dataset. The assessment was carried out at multiple steps of the genome assembly pipeline (see Fig. 2) after the initial assembly with Wtdbg2 and consequently after each polishing event which consists of two rounds of polishing with long Nanopore reads using Racon and two rounds of polishing with short Illumina reads using Pilon.

Inferring demography history in H. passer

The PSMC analysis showed two extreme scenarios for the population evolutionary history. When considering a faster mutation rate (μ) of 10^{-8}, the population showed a

slow expansion ~300 Kya, with a small population decline occurring ~70 Kya, followed by a second rapid expansion 30 Kya that reached a maximum effective population size of ~300,000 individuals (Figure 1.5A). When using a slower mutation rate of 10^{-9}, the population showed an initial expansion around 2.8 Mya, with a small decline ~600 Kya, and the subsequent rapid expansion 300 Kya, reaching a maximum effective population size of ~2,800,000 individuals (Figure 1.5B).

Considering the slower mutation rate scenario, an effective population size in the order of millions of individuals for *H. passer* seems plausible when you consider the vast available habitat it occupies compared to its sister species *H. limbaughi* who's effective population size was estimated ~60,000 individuals (Crane *et al.* 2018). *H. limbaughi* is endemic to Clipperton Island and occupies a fraction of the distribution of *H. passer*, which is found across the entire TEP coastline. However, considering the higher mutation rate scenario may seem likely when observing the first rapid population expansion occurring much after the closure of the Isthmus of Panama once oceanographic conditions in the TEP became more suitable.

H. passer was previously estimated to have diverged from its geminate Atlantic species (*H. ciliaris*) between 1.7 and 1.4 Mya (Alva-Campbell *et al.* 2010; Tariel *et al.* 2016), considering a molecular clock that was calibrated according to the closure of the Isthmus of Panama dated around 3.1 to 3.5 Mya (Bellwood *et al.* 2004). However, recent studies suggest the closure of the Isthmus of Panama might have happened more recently, around 2.8 Mya (O'Dea *et al.* 2016). Therefore, the genetic

divergence between Holacanthus geminates could be more recent than previously believed.

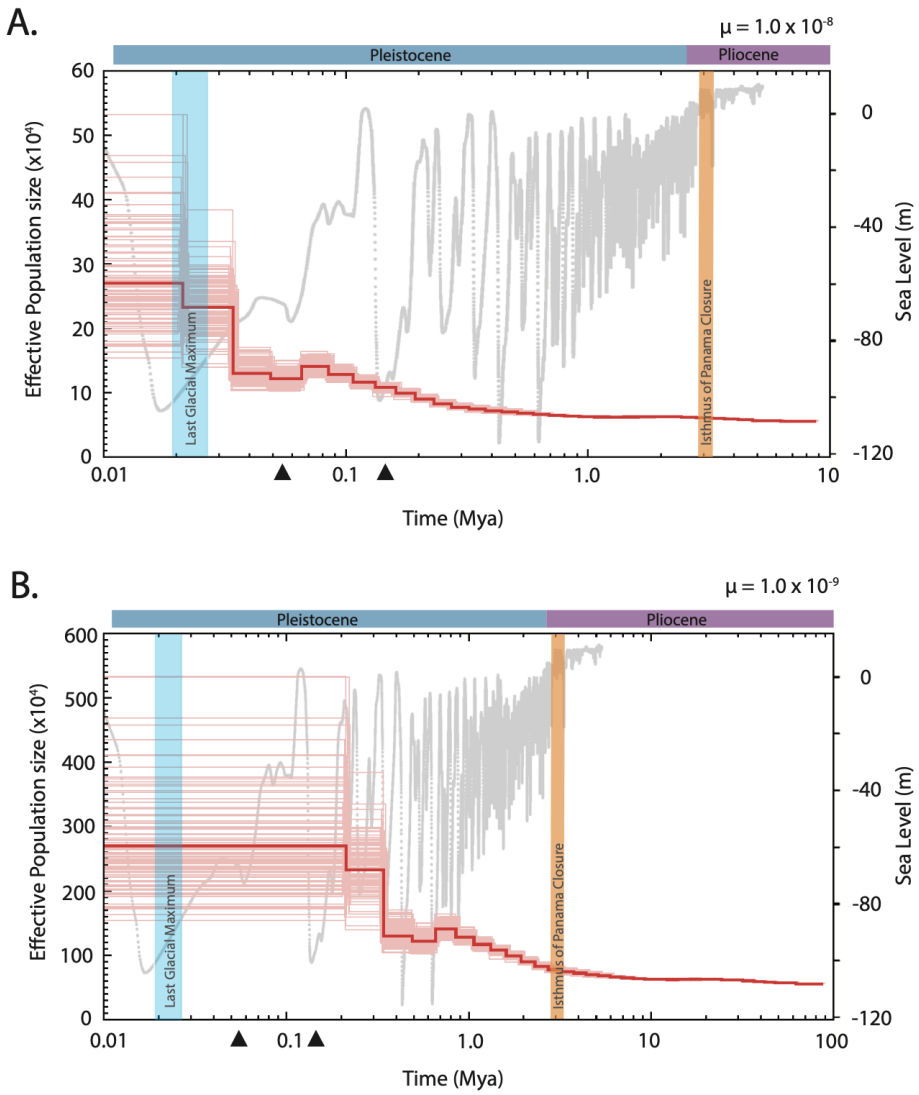


Figure 1.5. PSMC analysis showing the demographic history (red line) of *Holacanthus passer* using a generation time of 5 years and a mutation rate (μ) of 10^{-8} (A) and 10^{-9} (B). Global sea level model fluctuations over the past 5 million years are shown in the background (grey) (data from de Boer et al 2014). Vertical blue bars refer to the last glacial maximum (LGM) period (~19- 26.5 kya) and the orange bar represents the closure of the Isthmus of Panama (~3.2-2.8 Mya). Triangles represent marine population expansion events previously recorded in the Tropical Eastern Pacific (see text).

After the closure of the Isthmus, oceanographic conditions in the TEP varied drastically following sea level changes due to multiple glaciation periods in the Pleistocene (Chadwick-Furman 1996; Lambeck 2004), likely leading to important demographic consequences (Ludt and Rocha 2015). Most rapid population expansions in both freshwater (Aguilar *et al.* 2019) and marine organisms (Jenkins *et al.* 2018) have been reported to occur globally after the last glacial maximum (LGM) that took place from 26.5 to 19 Kya (Clark *et al.* 2009). However, only a few species have reported population expansions prior to the LGM (Jenkins *et al.* 2018). On the contrary, in the TEP, most studies that have assessed the demographic history of marine organisms have found population expansions that precede the LGM (Dawson *et al.* 2011; Sandoval-Huerta *et al.* 2018; Palmerín-Serrano *et al.* 2020; Torres-Hernández *et al.* 2020) and few reporting population expansions in the last 20 Kya (Lessios *et al.* 2001; Palmerín-Serrano *et al.* 2020). For instance, the goby, *Elacatinus puncticulatus*, and clingfish, *Gobieosox adustus*, experienced a population expansion around 170-130 Kya and 200-150 Kya, respectively (Sandoval-Huerta *et al.* 2018; Torres-Hernández *et al.* 2020). While another reef fish, *Anisotremus interruptus*, experienced an expansion in its continental populations after the LGM (~5 kya). Interestingly, *A. interruptus* populations from the oceanic islands of Revillagigedos and the Galapagos Archipelago showed earlier expansions at around 55 kya (Palmerín-Serrano *et al.* 2020). Yet, all demographic history studies in the TEP to date are based on single mitochondrial markers.

To the best of our knowledge, our study is the first to assess the demographic history of a marine fish in the TEP using genome-wide nuclear DNA. Our results support previous findings of marine population expansions in the TEP occurring prior to the LGM (Dawson *et al.* 2011; Sandoval-Huerta *et al.* 2018; Palmerín-Serrano *et al.* 2020; Torres-Hernández *et al.* 2020). This pattern is consistent with our analyses using both slow and fast mutation rates for *H. passer*, which showed population expansions beyond 30 Kya. Overall, drops in sea level decrease the available marine habitat, potentially restricting gene flow between populations, thus resulting in population bottlenecks. This was particularly prominent in areas where shallow marine habitats (<60 m) are abundant, such as the Western Atlantic, Western Pacific, and Eastern Indian Ocean (Ludt and Rocha 2015). Map projections of the TEP during the LGM show relatively small differences of the exposed landmasses at low sea level (-60m) compared to present day (Ludt and Rocha 2015), possibly indicating that glaciation sea level drops might not have changed the overall topology and gene flow in the TEP as much as it did in other ocean basins. Overall, although our demographic estimates varied considerable with our choice of mutation rate, our results are generally consistent with previous studies indicating that population expansions of marine fishes in the TEP may have preceded the LGM (Sandoval-Huerta *et al.* 2018). Furthermore, this also suggests that the demography history in *H. passer* is likely to be shaped by historical events associated with the closure of the Isthmus of Panama, rather than by the more recent LGM.

Conclusion

Here we present the first *Holacanthus* angelfish genome assembly using high-coverage (~75X) Nanopore long-reads and high-coverage (~154X) Illumina short-reads. Hybrid assemblies have become the method of choice to obtain high-quality genome assemblies and often rely on high coverage sequences. However, our data suggests that a comparable high-quality and complete genome may be obtained with a minimum of 20X long-read and 36X short-read coverage. These values indicate the lowest coverage assessed in this study, nonetheless, the real minimum could potentially be lower and should be assessed in a separate study. As accuracy of long-read technology continues to increase, genome assemblies will become rapidly more accessible for non-model species and will eventually remove the need of short-reads for genome polishing. Additionally, this study presents the first demographic history model of a marine fish in the TEP using whole genome data. Our results support the importance of the Isthmus of Panama in shaping the demographic history of marine fish in the TEP.

Data Availability

Raw sequencing reads (Illumina and Nanopore) have been deposited into the NCBI Sequence Read Archive (SRA), while the Whole Genome Shotgun project has been deposited at DDBJ/ENA/GenBank under the accession number JAFREQ000000000, both under BioProject PFJNA713824. Step-by-step code to reproduce the methods

can be found at https://github.com/remygatins/Holacanthus_passer-ONT-Illumina-Genome-Assembly

List of Abbreviations

TEP: Tropical Eastern Pacific; TWA: Tropical Western Atlantic; ONT: Oxford Nanopore; Kb: kilobase; Gb: gigabase; bp: base pair; BUSCO: Benchmarking Universal Single-Copy Orthologs; PSMC: Pairwise Sequentially Markovian Coalescent model; g: Generation time; μ: Mutation rate; Mya: Million years ago; Kya: Thousand years ago; LGM: Last Glacial Maximum

Supplementary Materials

Additional file: Table S1. General summary of individual flow cells of Minion Nanopore sequencing data.

	HPA_Ont_01	HPA_Ont_02	HPA_Ont_03	HPA_Ont_04	Total
Mean read length:	2,750.90	3,104.70	3,018.80	3,260.70	2,992.30
Mean read quality:	11.3	11.2	11.1	10.7	11.1
Median read length:	1,325.00	1,493.00	1,212.00	1,611.00	1,416.00
Median read quality:	12	12	11.8	11.4	11.8
Number of reads:	5,668,864.00	3,705,080.00	1,912,392.00	3,360,912.00	14,647,248.00
Read length N50:	6,031.00	6,865.00	7,390.00	7,044.00	6,626.00
Total bases:	15,594,491,159.00	11,502,976,706.00	5,773,092,305.00	10,959,029,480.00	43,829,589,650.00
Number, percentage and megabases of reads above quality cutoffs					
Q5:	5432664 (95.8%) 15225.6Mb	3538133 (95.5%) 11232.3Mb	1838789 (96.2%) 5662.1Mb	3170908 (94.3%) 10563.3Mb	13980494 (95.4%) 42683.3Mb
Q7:	5064102 (89.3%) 14436.0Mb	3293317 (88.9%) 10635.2Mb	1718005 (89.8%) 5364.3Mb	2922594 (87.0%) 9910.1Mb	12998018 (88.7%) 40345.6Mb
Q10	4189759 (73.9%) 12265.7Mb	2723010 (73.5%) 8990.8Mb	1392232 (72.8%) 4265.7Mb	2275136 (67.7%) 7880.8Mb	10580137 (72.2%) 33402.9Mb
Q12:	2816311 (49.7%) 8763.1Mb	1840829 (49.7%) 6468.1Mb	893583 (46.7%) 2993.0Mb	1372134 (40.8%) 5040.8Mb	6922857 (47.3%) 23265.1Mb
Q15:	176972 (3.1%) 334.2Mb	95115 (2.6%) 185.4Mb	24156 (1.3%) 38.6Mb	33231 (1.0%) 59.0Mb	329474 (2.2%) 617.2Mb
Top 5 highest mean basecall quality scores and their read lengths					
1	22.6 (274)	21.6 (174)	19.6 (330)	20.2 (260)	22.6 (274)
2	21.6 (165)	21.2 (321)	19.5 (184)	19.3 (275)	21.6 (165)
3	21.5 (274)	21.1 (484)	19.4 (256)	19.2 (191)	21.6 (174)
4	21.4 (244)	21.0 (387)	19.3 (425)	19.2 (202)	21.5 (274)
5	21.0 (249)	21.0 (220)	19.2 (196)	19.1 (315)	21.4 (244)
Top 5 longest reads and their mean basecall quality score					
1	474205 (3.3)	154847 (3.5)	207768 (4.5)	81895 (3.4)	474205 (3.3)
2	265487 (4.2)	149825 (4.4)	165102 (4.4)	78643 (3.3)	265487 (4.2)
3	229142 (3.9)	148558 (4.1)	103615 (4.4)	73008 (3.3)	229142 (3.9)
4	176471 (3.9)	140906 (3.6)	77550 (12.5)	67716 (7.9)	207768 (4.5)
5	168737 (4.6)	117098 (4.1)	76398 (4.3)	67096 (4.4)	176471 (3.9)

Additional file: Table S2. Comparison summary statistics for 11 selected fish genome assemblies, including *Holacanthus passer* from this study. *indicates percent masked reported using RepeatMasker program.

Species	***Holacanthus passer***	***Centropyge vrolikii***	***Chaetodon austriacus***	***Epinephelus lanceolatus***	***Amphiprion percula***	***Oreochromis niloticus***	***Argyrosomus japonicus***	***Takifugu rubripes***	***Gasterosteus aculeatus aculeatus***	***Gadus morhua***	***Danio rerio***
Common name	King angelfish	Pearlscale pygmy angelfish	Blacktail butterflyfish	Giant grouper	Orange clownfish	Nile tilapia	Japanese meagre	Fugu	Three-spined stickleback	Atlantic cod	Zebrafish
Family	Pomacanthidae	Pomacanthidae	Pomacanthidae	Serranidae	Pomacentridae	Cichlidae	Sciaenidae	Tetraodontidae	Gasterosteidae	Gadidae	Cyprinidae
Platform											
Shotgun		x	x								x
Illumina (paired-end)	x		x	x		x	x			x	x
Mate-Pairs		x	x								
10 x Genomics								x		x	
Nanopore	x									x	
PacBio					x	x	x	x	x	x	
Hi-C				x	x		x	x	x	x	
Chicago		x									
BioNano										x	
Total Length (Mb)	583.4	696.5	712.2	1087.4	909.0	1005.7	792.0	384.1	471.9	684.3	1679.2
% GC	41.27	41.76	42.48	41.26	39.53	40.73	41.25	45.67	44.66	45.69	36.6
Scaffolds											
Number	476	30,500	13,441	4,200	366	2,459	1984	127	2,911	1,126	1,922
N50 length (Mb)	5.7	9	0.17	46.2	38.4	38.8	13.1	16.7	20.4	27.4	52.2
Longest (Mb)	17.1	31	2	57.7	46.1	87.6	30.3	29.2	34.2	41.8	78.1
Ns (Kb)	0.0	11709.3	48772.7	39254.8	32.4	55.0	0.0	3688.5	3574.2	12.6	4693.6
Gaps	0	30486	105028	23415	682	551	0	402	3125	126	20258

Chromosomes				24	24	23	24	22	22	23	25
% Masked	5.09*	15.94		2.61*	2.92*	5.4*		10.26*		10.17*	57.77*
GenBank Assembly Accession				GCA_005281545.1	GCA_003047355.2	GCA_001858045.3	GCA_015710095.1	GCA_901000725.2	GCA_016920845.1	GCA_010882105.1	GCA_000002035.4
RefSeq Assembly accession				GCF_005281545.1	GCF_002776465.1	GCF_001858045.2		GCF_901000725.2	GCF_016920845.1	GCF_902167405.1	GCF_000002035.6
Reference	This study	Fernandez-Silva *et al* 2018	DiBattista *et al* 2016	Zhou *et al* 2019	Lehman *et al* 2018	Conte *et al* 2017	Zhao *et al* 2020			Kirubakaran *et al* 2020	

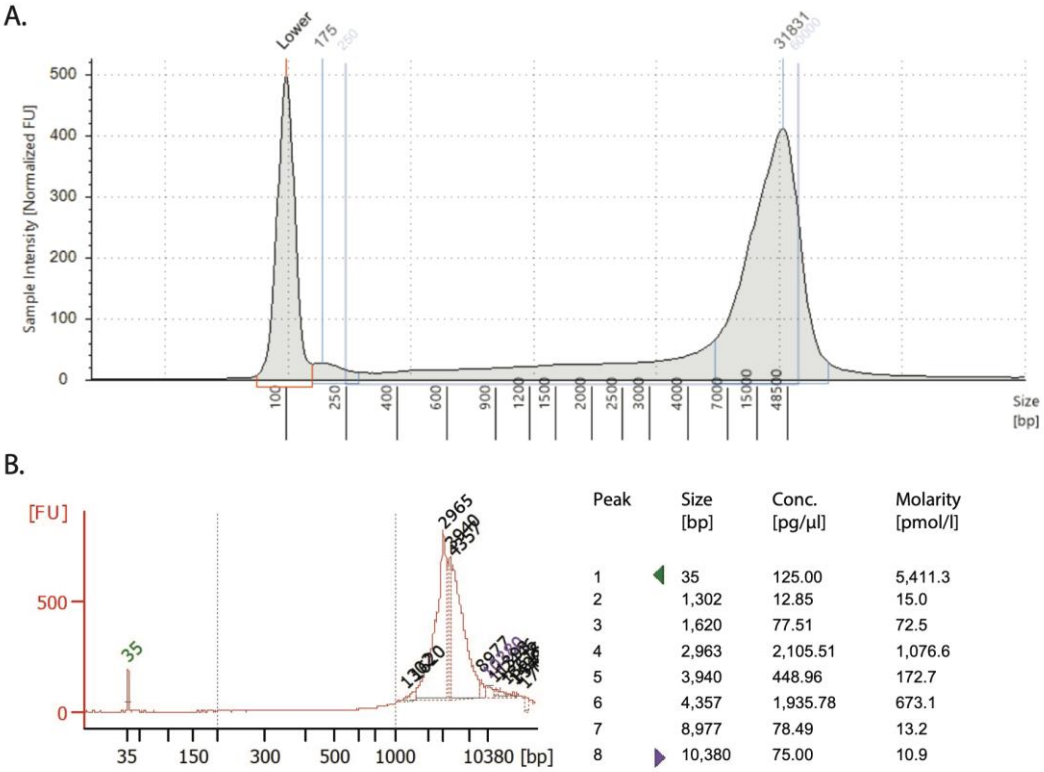
Additional file: Table S3. Summary output of repetitive elements of *H. passer* predicted by RepeatMasker v. 2.9.0+. The query species was assumed to be *Danio rerio*.

Sequences:	476
total length:	583428144 bp (583428144 bp excl N/X-runs)
GC level:	41.27%
bases masked:	29697917 bp (5.09 %)

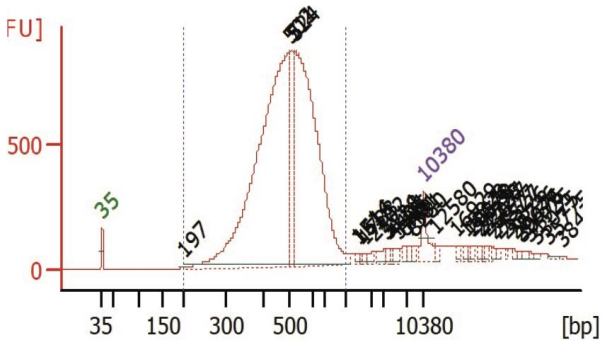
	Number of elements*	**Length occupied (bp)**	**Percentage of sequence**
Retroelements	32155	6929458	1.19%
SINEs:	1264	127776	0.02%
Penelope	303	35535	0.01%
LINEs:	19008	4969875	0.85%
CRE/SLACS	0	0	0.00%
L2/CR1/Rex	13019	3276411	0.56%
R1/LOA/Jockey	644	120329	0.02%
R2/R4/NeSL	298	121376	0.02%
RTE/Bov-B	1555	536109	0.09%
L1/CIN4	2566	719770	0.12%
LTR elements:	11883	1831807	0.31%
BEL/Pao	1084	310271	0.05%
Ty1/Copia	25	16958	0.00%
Gypsy/DIRS1	6189	1075353	0.18%
Retroviral	2370	223106	0.04%
DNA transposons	67082	7952853	1.36%
hobo-Activator	24016	2282989	0.39%
Tc1-IS630-Pogo	10111	2729719	0.47%
En-Spm	0	0	0.00%
MuDR-IS905	0	0	0.00%
PiggyBac	217	31373	0.01%
Tourist/Harbinger	2024	230982	0.04%
Other (Mirage, P-element, Transib)	1538	262794	0.05%
Rolling-circles	421	48093	0.01%
Unclassified:	269	71601	0.01%
Total interspersed repeats:		14953912	2.56%
Small RNA:	1676	161165	0.03%
Satellites:	961	80911	0.01%
Simple repeats:	303667	12478423	2.14%
Low complexity:	38523	2143664	0.37%

Additional file: Table S4. Summary statistics for six genome assemblies of *Holacanthus passer* created using subsets of the original data.

Nanopore Coverage	20X	20X	45X	45X	75X	75X
Illumina Coverage	36X	145X	36X	145X	36X	145X
Number of Contigs	804	804	528	528	486	486
sum	580 745 346	580 850 390	577 019 287	577 141 894	583 382 494	583 528 366
N50	1 844 885	1 844 888	3 800 341	3 800 961	5 706 861	5 708 022
N60	1 354 617	1 354 790	2 529 728	2 529 556	4 056 084	4 056 739
N70	1 094 304	1 094 839	2 012 233	2 012 541	3 091 610	3 091 862
N80	790 952	791 044	1 227 818	1 228 020	1 691 323	1 691 943
N90	472 091	471 885	732 801	732 447	1 000 524	1 000 532
N100	1 872	1 872	1 341	1 341	1 676	1 676
N_count	0	0	0	0	0	0
Gaps	0	0	0	0	0	0
ave	722 320.08	722 450.73	1 092 839.56	1 093 071.77	1 200 375.50	1 200 675.65
largest	14 166 795	14 169 211	15 858 881	15 858 554	17 144 075	17 148 928
Actinopterygii v3.0.2 (n=4584)						
Complete BUSCOs	4463 (97.4%)	4464 (97.4%)	4417 (96.4%)	4417 (96.4%)	4472 (97.6%)	4471 (97.5%)
Complete and single-copy BUSCOs	4355 (95.0%)	4359 (95.1%)	4317 (94.2%)	4318 (94.2%)	4367 (95.3%)	4368 (95.3%)
Complete and duplicated BUSCOs	108 (2.4%)	105 (2.3%)	100 (2.2%)	99 (2.2%)	105 (2.3%)	103 (2.2%)
Fragmented BUSCOs	40 (0.9%)	37 (0.8%)	44 (1.0%)	45 (1.0%)	34 (0.7%)	37 (0.8%)
Missing BUSCOs	81 (1.7%)	83 (3.9%)	123 (2.6%)	122 (2.6%)	78 (1.7%)	76 (1.7%)
Eukaryota v3.0.2 (n=303)						
Complete BUSCOs	289 (95.4%)	289 (95.4%)	286 (94.4%)	285 (94.1%)	289 (95.4%)	289 (95.4%)
Complete and single-copy BUSCOs	273 (90.1%)	273 (90.1%)	271 (89.4%)	270 (89.1%)	274 (90.4%)	274 (90.4%)
Complete and duplicated BUSCOs	16 (5.3%)	16 (5.3%)	15 (5%)	15 (5.0%)	15 (5.0%)	15 (5%)
Fragmented BUSCOs	2 (0.7%)	2 (0.7%)	2 (0.7%)	3 (1.0%)	2 (0.7%)	2 (0.7%)
Missing BUSCOs	12 (3.9%)	12 (3.9%)	15 (4.9%)	15 (4.9%)	12 (3.9%)	12 (3.9%)

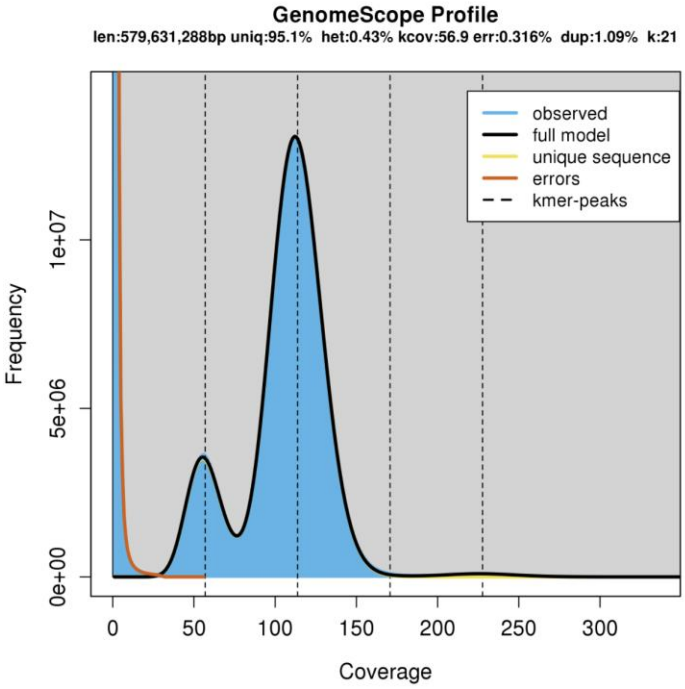


Additional file: Figure S1. *Holacanthus passer* genomic DNA profile used for Nanopore Sequencing. (A) TapeStation analysis using a Genomic DNA ScreenTape (Agilent Technologies, Inc 2017) of DNA sample used pre-fragmentation. Peak molecular weight was found to be at 31831 bp with a calibrated concentration of 19.6 ng/µl. Between 250 and 60000 bp, a region representing 84% of the sequences, the average size was 18931 bp with a concentration of 23.5 ng/µl. (B) Bioanalyzer 2100 profile and statistics using a High Sensitivity DNA Assay (Agilent Technologies, Inc 2009) of genomic DNA post sheared with Covaris g-TUBE following manufacturers protocol for 10 Kb fragments.



Peak	Size [bp]	Conc. [pg/µl]	Molarity [pmol/l]
1	35	125.00	5,411.3
2	197	2.21	17.0
3	502	2,357.89	7,113.0
4	511	238.74	708.3
5	524	1,766.48	5,103.1
6	1,514	8.16	8.2
7	2,202	16.22	11.2
8	3,114	19.31	9.4
9	5,486	18.30	5.1
10	6,401	15.32	3.6
11	7,381	12.49	2.6
12	8,280	10.81	2.0
13	10,380	75.00	10.9

Additional file: Figure S2. Bioanalyzer 2100 profile of *Holacanthus passer* DNA after Kapa Hyperplus library prep followed by a double size-selection cleanup with SPRIselect beads (0.56X and 0.72X). The Bioanalyzer was run on a High Sensitivity DNA Assay (Agilent Technologies, Inc 2009). Our final Illumina library was sequenced on a HiSeq4000 (150PE) at Novogene Corporation Inc.



Additional file: Figure S3. Histogram for the 21 k-mer distribution of Illumina short reads for *Holacanthus passer* plotted in GenomeScope VX. The highest frequency of k-mer coverage was seen around 110X (excluding k-mers with low coverage).

Competing Interest

The authors declare that they have no competing interests.

Funding

Molecular and computational resources were funded by the Department of Biology and the Ronald E. McNair Post-Baccalaureate Achievement Program at the University of Massachusetts Boston. R.G. was financially supported by the Consejo Nacional de Ciencia y Tecnología (CONACYT) and the University of California Institute for Mexico and the United States (UC-MEXUS) under the Contract No. 536570.

Acknowledgements

We would like to thank the lab members from Proyecto de Fauna Arrecifal lab at the Universidad de Baja California Sur, La Paz (UABCS) for helping secure fieldwork logistics and sample collections. We would additionally like to thank the UMass Boston High Performance Computing team for their assistance with our computational research needs.

Ethics statement

The care and handling of all vertebrate animals used in this study was in compliance with the Institutional Animal Care and Use Committee (IACUC/BERNG-1601).

References

Aburto-Oropeza, O., E. Sala, and C. Sánchez-Ortiz, 2000 Feeding behavior, habitat use, and abundance of the angel fish *Holacanthus passer*. Environmental Biology of Fishes 57:.

Aguilar, C., M. J. Miller, J. R. Loaiza, R. González, R. Krahe *et al.*, 2019 Tempo and mode of allopatric divergence in the weakly electric fish Sternopygus dariensis in the Isthmus of Panama. Sci Rep-uk 9: 18828.

Allen, G., and D. Robertson, 1994 *Fishes of the tropical eastern Pacific* (U. of H. Press, Ed.).

Alva-Campbell, Y., S. R. Floeter, D. R. Robertson, D. R. Bellwood, and G. Bernardi, 2010 Molecular phylogenetics and evolution of Holacanthus angelfishes (Pomacanthidae). Mol Phylogenet Evol 56: 456–461.

Arellano-Martínez, M., B. P. Ceballos-Vázquez, B. P. Ceballos-Vázquez, and F. Galván-Magaña, 1999 Reproductive Biology of the King Angelfish *Holacanthus passer* Valenciennes 1846 in the Gulf of California, Mexico. Bulletin of Marine Science 65: 677–685.

Austin, C. M., M. H. Tan, K. A. Harrisson, Y. P. Lee, L. J. Croft *et al.*, 2017 De novo genome assembly and annotation of Australia's largest freshwater fish, the Murray cod (*Maccullochella peelii*), from Illumina and Nanopore sequencing read. Gigascience 6:.

Bellwood, D. R., L. van Herwerden, and N. Konow, 2004 Evolution and biogeography of marine angelfishes (Pisces: Pomacanthidae). Mol Phylogenet Evol 33: 140–155.

Boer, B. de, L. J. Lourens, and R. S. W. van de Wal, 2014 Persistent 400,000-year variability of Antarctic ice volume and the carbon cycle is revealed throughout the Plio-Pleistocene. Nat Commun 5: 2999.

Bolger, A. M., M. Lohse, and B. Usadel, 2014 Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30: 2114–2120.

Brumfield, R. T., P. Beerli, D. A. Nickerson, and S. V. Edwards, 2003 The utility of single nucleotide polymorphisms in inferences of population history. Trends Ecol Evol 18: 249–256.

Chadwick-Furman, N. E., 1996 Reef coral diversity and global change. Global Change Biol 2: 559–568.

Clark, P. U., A. S. Dyke, J. D. Shakun, A. E. Carlson, J. Clark *et al.*, 2009 The Last Glacial Maximum. Science 325: 710–714.

Crane, N. L., J. Tariel, J. E. Caselle, A. M. Friedlander, D. R. Robertson *et al.*, 2018 Clipperton Atoll as a model to study small marine populations: Endemism and the genomic consequences of small population size. Plos One 13: e0198901.

Dawson, M. N., P. H. Barber, L. I. González-Guzmán, R. J. Toonen, J. E. Dugan *et al.*, 2011 Phylogeography of Emerita analoga (Crustacea, Decapoda, Hippidae), an eastern Pacific Ocean sand crab with long-lived pelagic larvae. J Biogeogr 38: 1600–1612.

DiBattista, J. D., X. Wang, P. Saenz-Agudelo, M. J. Piatek, M. Aranda *et al.*, 2018 Draft genome of an iconic Red Sea reef fish, the blacktail butterflyfish (*Chaetodon austriacus*): current status and its characteristics. Mol Ecol Resour 18: 347–355.

Fernandez-Silva, I., J. B. Henderso, L. A. Rocha, and W. B. Simison, 2018 Whole-genome assembly of the coral reef Pearlscale Pygmy Angelfish (*Centropyge vrolikii*). Sci Rep-uk 8: 1--11.

Hernández, M. C., 1998 Estructura de tallas y crecimiento individual del Ángel Rey, *Holacanthus passer*, Valenciennes 1846 (Teleostei: Pomacanthidae), en la Bahía de La Paz, B.C.S. México.

Jain, M., H. E. Olsen, B. Paten, and M. Akeson, 2016 The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. Genome Biol 17: 239.

Jauhal, A. A., and R. D. Newcomb, 2021 Assessing genome assembly quality prior to downstream analysis: N50 versus BUSCO. Mol Ecol Resour.

Jenkins, T. L., R. Castilho, and J. R. Stevens, 2018 Meta-analysis of northeast Atlantic marine taxa shows contrasting phylogeographic patterns following post-LGM expansions. Peerj 6: e5684.

Johnson, L. K., R. Sahasrabudhe, J. A. Gill, J. L. Roach, L. Froenicke *et al.*, 2020 Draft genome assemblies using sequencing reads from Oxford Nanopore Technology and Illumina platforms for four species of North American *Fundulus* killifish. Gigascience 9: giaa067-.

Jordan, D. S., 1908 The law of geminate species. The American naturalist 42:.

Lambeck, K., 2004 Sea-level change through the last glacial cycle: geophysical, glaciological and palaeogeographic consequences. C R Geosci 336: 677–689.

Lehmann, R., D. J. Lightfoot, C. Schunter, C. T. Michell, H. Ohyanagi *et al.*, 2019 Finding Nemo's Genes: A chromosome-scale reference assembly of the genome of the orange clownfish Amphiprion percula. Mol Ecol Resour 19: 570–585.

Lessios, H. A., 1981 Molecular and Morphological Differentiation Between Sea Urchins Separated by the Isthmus of Panama. Evolution 35: 618–634.

Lessios, H. A., M. J. Garrido, and B. D. Kessing, 2001 Demographic history of Diadema antillarum, a keystone herbivore on Caribbean reefs. Proc Royal Soc Lond Ser B Biological Sci 268: 2347–2353.

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

Li, H., and R. Durbin, 2011 Inference of human population history from individual whole-genome sequences. Nature 475: 493–496.

Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan *et al.*, 2009 The Sequence Alignment/Map format and SAMtools. Bioinformatics 25: 2078–2079.

Liu, S., and M. M. Hansen, 2017 PSMC (pairwise sequentially Markovian coalescent) analysis of RAD (restriction site associated DNA) sequencing data. Mol Ecol Resour 17: 631–641.

Ludt, W. B., and L. A. Rocha, 2015 Shifting seas: the impacts of Pleistocene sea-level fluctuations on the evolution of tropical marine taxa. J Biogeogr 42: 25–38.

Marçais, G., and C. Kingsford, 2011 A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. Bioinformatics 27: 764–770.

Miura, O., M. E. Torchin, and E. Bermingham, 2010 Molecular phylogenetics reveals differential divergence of coastal snails separated by the Isthmus of Panama. Mol Phylogenet Evol 56: 40–48.

Moyer, J. T., R. E. Thresher, and P. L. Colin, 1983 Courtship, spawning and inferred social organization of American angelfishes (Genera *Pomacanthus, Holacanthus* and *Centropyge*; pomacanthidae). Environ Biol Fish 9: 25–39.

O'Dea, A., H. A. Lessios, A. G. Coates, R. I. Eytan, S. A. Restrepo-Moreno *et al.*, 2016 Formation of the Isthmus of Panama. Sci Adv 2: e1600883.

Palmerín-Serrano, P. N., J. Tavera, E. Espinoza, A. Angulo, J. E. Martínez-Gómez *et al.*, 2020 Evolutionary history of the reef fish *Anisotremus interruptus*

(Perciformes: Haemulidae) throughout the Tropical Eastern Pacific. J Zool Syst Evol Res.

Pyle, R., A. G., R. Myers, F. Zapata, R. Robertson *et al.*, 2010 *Holacanthus passer*. The IUCN Red List of Threatened Species.

Quimbayo, J. P., M. S. Dias, O. R. C. Schlickmann, and T. C. Mendes, 2017 Fish cleaning interactions on a remote island in the Tropical Eastern Pacific. Mar Biodivers 47: 603–608.

Rang, F. J., W. P. Kloosterman, and J. de Ridder, 2018 From squiggle to basepair: computational approaches for improving nanopore sequencing read accuracy. Genome Biol 19: 90.

Ruan, J., and H. Li, 2019 Fast and accurate long-read assembly with wtdbg2. Nat Methods 17: 155–158.

Sánchez-Alcántara, I., O. Aburto-Oropeza, E. F. Balart, A. L. Cupul-Magaña, H. Reyes-Bonilla *et al.*, 2006 Threatened Fishes of the World: *Holacanthus passer* Valenciennes, 1846 (Pomacanthidae). Environ Biol Fish 77: 97–99.

Sandoval-Huerta, E. R., R. G. Beltrán-López, C. del R. Pedraza-Marrón, M. A. Paz-Velásquez, A. Angulo *et al.*, 2018 The evolutionary history of the goby *Elacatinus puncticulatus* in the tropical eastern pacific: effects of habitat discontinuities and local environmental variability. Mol Phylogenet Evol 130: 269–285.

Shafin, K., T. Pesout, R. Lorig-Roach, M. Haukness, H. E. Olsen *et al.*, 2019 Efficient de novo assembly of eleven human genomes using PromethION sequencing and a novel nanopore toolkit A Preprint. Biorxiv.

Simao, F. A., R. M. Waterhouse, P. Ioannidis, E. V. Kriventseva, and E. M. Zdobnov, 2015 BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics 31: 3210--3212.

Tan, M. H., C. M. Austin, M. P. Hammer, Y. P. Lee, L. J. Croft *et al.*, 2018 Finding Nemo: Hybrid assembly with Oxford Nanopore and Illumina reads greatly improves the Clownfish (*Amphiprion ocellaris*) genome assembly. Gigascience 7: 1--6.

Tariel, J., G. C. Longo, and G. Bernardi, 2016 Tempo and mode of speciation in *Holacanthus* angelfishes based on RADseq markers. Mol Phylogenet Evol 98: 84–88.

Torres-Hernández, E., I. Betancourt-Resendes, P. Díaz-Jaimes, A. Angulo, E. Espinoza *et al.*, 2020 Independent evolutionary lineage of the clingfish Gobiesox adustus (Gobiesocidae) from Isla del Coco, Costa Rica. Revista De Biología Tropical 68: S306–S319.

Vurture, G. W., F. J. Sedlazeck, M. Nattestad, C. J. Underwood, H. Fang *et al.*, 2017 GenomeScope: fast reference-free genome profiling from short reads. Bioinformatics 33: 2202–2204.

Walker, B. J., T. Abeel, T. Shea, M. Priest, A. Abouelliel *et al.*, 2014 Pilon: An Integrated Tool for Comprehensive Microbial Variant Detection and Genome Assembly Improvement. Plos One 9: e112963.

Waterhouse, R. M., M. Seppey, F. A. Simão, M. Manni, P. Ioannidis *et al.*, 2017 BUSCO Applications from Quality Assessments to Gene Prediction and Phylogenomics. Mol Biol Evol 35: 543–548.

Wood, D. E., and S. L. Salzberg, 2014 Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome Biol 15: R46.

Chapter 2 Predictable and stochastic population genomic patterns of the widespread King angelfish (*Holacanthus passer*) in the Tropical Eastern Pacific

Abstract

Understanding the drivers and scale of connectivity is crucial for basic and applied ecology and evolutionary studies. Genetic connectivity refers to the amount of gene flow among populations and varies between two extremes: fixed population structure and panmixia (no population structure). A lack of physical barriers in the ocean predicts panmixia for marine populations, however, a growing amount of evidence has reported weak but statistically significant genetic structure in marine species, inconsistent with true panmixia, but consistent with the fact that, in general, marine populations are very large. This subtle genetic structure is not always correlated with space and may be driven by temporal or ecological factors. A region where connectivity has been relatively understudied, and which is biogeographically isolated from other provinces, is the Tropical Eastern Pacific (TEP). Despite the high occurrence of speciation events of marine organisms in the TEP, potentially due to a lack of gene flow within the region, most TEP connectivity studies show high gene flow across long distances. Yet, the majority of these studies rely on single mitochondrial DNA markers. In this study we used restriction site-associated DNA sequencing (RADseq) to generate 19,809 genome-wide markers to evaluate the population structure and genetic diversity patterns of the widespread TEP King

angelfish, *Holacanthus passer*. We found no significant correlation between genetic diversity and distance from the population origin (here considered to be Panama after the closure of the Isthmus of Panama) or with environmental conditions. Global F_{ST} revealed high gene flow among populations along the TEP coastline ($F_{ST} = 0.00$) as predicted by the literature. However, pairwise comparisons detected weak but significant structure between Panama and the Sea of Cortez ($0.002 < F_{ST} < 0.005$; $0.007 < p < 0.043$), driven principally by isolation by distance. Interestingly, we had two sampling periods in Panama that occurred 10 years apart and did not show this same genetic signal. After the addition of a small number of samples collected from Galapagos and Clipperton Island, we detected 28 outlier loci that were used to run a DAPC and a Bayesian clustering analysis (i.e., STRUCTURE). Here, our results revealed subtle genetic signatures that differentiated the mainland populations, Galapagos, and Clipperton Island. Overall, our results are consistent with previous mtDNA and microsatellite studies showing high gene flow within the TEP with a few potential genetic breaks between the Sea of Cortez and the mainland, as well as between the mainland and the oceanographic islands. Our results add to the growing evidence of weak but significant stochastic genetic structure found in marine species.

Keywords: Genetic structure; population genetics; connectivity; isolation by distance; chaotic genetic patchiness; RADseq; Eastern Pacific; Angelfishes

Introduction

Understanding the scales of connectivity is crucial for basic and applied ecology and evolutionary studies of marine species. In the marine environment, dispersal often occurs during the larval phase of the organism, settling on a reef where it will remain throughout its life. Larval dispersal consequently eliminates the feasibility of tools such as mark-recapture, satellite tagging, and direct observation to study connectivity. Therefore, genetic tools have been the tool of choice to estimate rates, distances, and patterns of dispersal (Selkoe and Toonen, 2011). In theory, population genetic structure reflects long-term rates of gene flow, drift, and selection. Understanding the relationship between dispersal potential and gene flow has been a longstanding aim in the realm of population genetics since dispersal is such a complex trait to study directly and is fundamental to address questions in ecology. Pelagic larval duration (PLD) has been used as a proxy for dispersal potential (Siegel et al., 2003). However, the ability of larvae to survive for an extended period does not necessarily indicate larger range sizes (Bradbury and Bentzen, 2007; Mora et al., 2012). Therefore, the use of genetic tools have been the method of choice to estimate dispersal in marine environments mainly because the complexity of the environment and life histories of marine species make them a hard study system (Selkoe and Toonen, 2011). The most common genetic metric to estimate gene flow is Wright's fixation index (F_{ST}), which measures genetic variation among geographically separated populations (i.e., genetic structure) (Wright, 1931).

The scale and degree of population genetic structure ranges from population structure or no population structure (i.e., panmixia). Population structure is often

driven by spatial or temporal factors which restrict gene flow between populations, causing a clear genetic break (Savolainen et al., 2006; DiBattista et al., 2017; Longo et al., 2020). With increasing geographic distance, we commonly expect an increase in genetic distance between populations (i.e., gene flow between nearby populations is higher while populations further apart result in lower gene flow), otherwise known as isolation by distance (IBD) (Wright, 1943; Slatkin, 1993). In contrast, populations with large effective population sizes and high gene flow can lead to one large panmictic metapopulation. Interestingly, many marine species exhibit weak but statistically significant genetic structuring, thus deviating from true panmixia (e.g., Johnson and Black, 1982; Iacchei et al., 2013; Moody et al., 2015; Thia et al., 2021). More recently, chaotic genetic patchiness (CGP) has been used to describe this weak but significant structure pattern when it is not correlated with either space or time (Johnson and Black, 1982). Although the mechanisms that drive CGP are still not fully understood, four main processes have been hypothesized to generate these unexpected patterns: selection, sweepstakes reproductive success, collective dispersal, and asynchronous local population dynamics (Eldon et al., 2016).

The Tropical Eastern Pacific (TEP) is an ideal study system where marine population genetic studies are lacking. The TEP has a straight coastline from northern Peru to the northern Sea of Cortez, Mexico, allowing us to test for IBD patterns easily. This region is physically isolated to the East by the Isthmus of Panama (closed ~2.8-3.1 Mya) (O'Dea et al., 2016), and to the west by the Eastern Pacific Barrier (EPB). The EPB consists of 4000 to 7000 km of deep water that prevents most

dispersal between the Central- and Eastern Pacific due to the lack of reefs to use as a stepping-stones (Lessios and Robertson, 2006). A recent study by Romero-Torres *et al* (2018) showed evidence that the EPB may be breached, in both directions, by rare dispersal events. More interestingly, this same study found that the TEP may have a stronger role as a larval source to the Central Tropical Pacific than initially believed. Moreover, relatively few population genetic studies have been carried out in the TEP (Lessios and Baums, 2016). Despite its vast extension spanning more than 6,000 km of coastline that ranges from both sides of the equator, most population genetic studies have found high levels of gene flow in TEP marine species (see review by Lessios and Baums, 2016). Nonetheless, some studies have found genetic structure in the TEP (e.g., Lessios et al., 2003; Riginos, 2005; Lessios and Robertson, 2006; Bernardi et al., 2008; Saarman et al., 2010; Baums et al., 2012; Lessios and Baums, 2016; Reguera-Rouzaud et al., 2021). However, these studies were limited to few molecular markers, most of which relied on single mitochondrial markers and a few microsatellites. Therefore, high gene flow tends to be the norm in the TEP. Nevertheless, a study that incorporates more molecular markers is needed to fully evaluate the population genetic dynamics in the area at a higher resolution.

In this study, we use restriction site-associated DNA sequencing (RADseq) to generate 19,809 single nucleotide polymorphisms (SNPs) from 91 genotyped King angelfish (*Holacanthus passer)* individuals captured throughout its range along the TEP coast. Using these data, we aim to evaluate the population structure and genetic diversity patterns of a widespread TEP species using a more comprehensive and

modern genomic approach. More specifically, we will address the following questions: (i) Does *H. passer* show any genetic diversity patterns across its range and are these correlated across space and environment? (ii) Does *H. passer* show evidence of population structure or panmixia? (iii) If genetic structure is found between populations in the TEP, what factors are potentially contributing to this structure?

Materials and Methods

Study species and sample collection

Our study species is the King angelfish, *Holacanthus passer,* which is one of the most iconic fish species of the TEP (Allen and Robertson, 1994). *Holacanthus* angelfishes are known to be protogynous hermaphrodites and to exhibit sexual dimorphism (Moyer et al., 1983). Their pelagic larval duration (PLD) is estimated to be between 23 –26 days based on *Holacanthus'* closest relative, *Pygoplites diacanthus* (Thresher and Brothers, 1985; Alva-Campbell et al., 2010). Within the TEP there are three *Holacanthus* species: *H. passer* (widespread TEP), *H. clarionensis* (Revillagigedo Archipelago endemic), and *H. limbaughi* (Clipperton Island endemic). The three TEP *Holacanthus* species form a monophyletic clade that diverged from its Atlantic geminate species following the closure of the Isthmus of Panama around 3.2 to 2.8 Mya (O'Dea et al., 2016). *Holacanthus passer* ranges from Bahia Magdalena, on the Pacific side of Baja California, and at the northern tip of the Sea of Cortez, Mexico, to northern Peru, and including Cocos, Malpelo, and the Galápagos Islands (Figure 2.1). Bahia Magdalena is a transition zone between temperate northern and Tropical

Eastern Pacific species, and while this is its northern distribution limit, *H. passer* was abundant in this area.

Samples of *Holacanthus passer* were collected across the TEP between 2005 and 2018 (Figure 2.1; Figure 2.1). A total of 102 individuals were collected across the TEP, with one vagrant individual being collected at Clipperton Island by Clua and Planes (2019) where the endemic sister species *H. limbaughi* is present and abundant (Crane et al., 2018). We also observed one vagrant individual in the Revillagigedo Archipelago at Roca Partida (RG, 2017), but that individual was not collected. Neither of those localities are considered within the normal range of the species (Allen and Robertson, 1994; Figure 2.1). Collections were made with pole spears while snorkeling or SCUBA diving. Fin clips or gill were immediately preserved in 95% ethanol after collection and stored at -20°C at the Molecular Ecology and Evolution Lab of the University of California Santa Cruz.

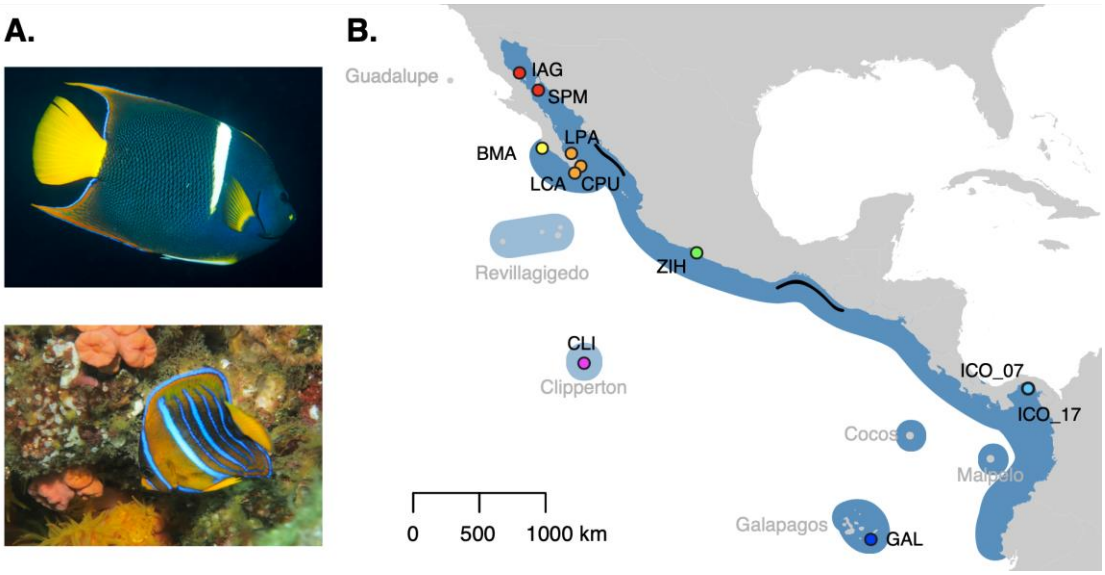


Figure 2.1. (A) Photographs of *Holacanthus passer* as an adult (top) and juvenile (bottom). (B) Geographic distribution of *H. passer* (dark blue shade). Although not a

part of *H.* passer's range, few vagrants have been reported at the Revillagigedo Archipelago and Clipperton Island (light blue shade). Black lines indicate the Sinaloan and Central American Gap, from North to South, respectively. Sampling locations are color coded by region: North Sea of Cortez (red), South Sea of Cortez (orange), Baja California Pacific (yellow), Mainland Mexico (green), Clipperton (purple), Panama (light blue), and Galapagos (dark blue). Sampling site key: IAG, Isla Ángel de la Guarda; SPM, San Pedro Mar; LPA, La Paz; CPU, Cabo Pulmo; LCA, Los Cabos; BMA, Bahía Magdalena; ZIH, Zihuatanejo; CLI, Clipperton; ICO_07, Isla Contadora 2007; ICO_17, Isla Contadora 2017; GAL- Galapagos.

RADseq library preparation

DNA was extracted using the DNeasy blood and tissue kits following the manufacturers protocol (Qiagen 2006). In order to identify single nucleotide polymorphisms (SNPs), two restriction site-associated (RAD) libraries were constructed using a variation of the original protocol (Miller et al. 2007; Baird et al. 2008) with the restriction enzyme SbfI as described in Longo & Bernardi (2015) with few modifications. Starting genomic DNA was standardized to 100ng per sample and sheared using a Covaris S2 sonicator with an intensity of 5, duty cycle of 10%, cycles/burst of 200, and a cycle time of 30 s. A final PCR amplification step of 10 cycles was carried out on a total volume of 50 μl. Ampure XP beads (Agencourt) were used for all size selection and purification steps. Individual samples were ligated to a unique barcode and index combination. Finally, both libraries were sequenced together on one lane of an Illumina HiSeq 4000 (150-bp single end reads) at the Vincent J. Coates Genomics Sequencing laboratory at UC Berkeley.

RAD-seq data processing and SNP calling

STACKS v.2.55 was used to sort, filter, and demultiplex reads using the '*process_radtags'* script with individual barcodes that were ligated during the RADseq library preparation (Catchen et al., 2013; Rochette et al., 2019). RADseq loci were then trimmed to 80bp with TRIM GALORE to compare with previous RADseq samples (Alva-Campbell et al., 2010; Tariel et al., 2016). Reads from all 102 individuals were aligned to the *Holacanthus passer* genome, *H.passer*_genome_1.0 (Gatins et al. *in review*) using BWA v 0.7.17 (Li and Durbin, 2009). Each aligned .sam file was then converted to .bam and sorted using SAMTOOLS v1.9 (Li et al., 2009). Reads aligned at a rate greater than 99%. To build the initial loci catalog we used all reference aligned samples by running the '*gstacks'* script on STACKS with default parameters (per sample coverage: mean=18.1x, stdev=12.4x, min=3.0x, max=69.0x).

Multiple iterations of the '*populations'* script of STACKS were carried out to generate output files for downstream analyses using the option -write_single_snp. In order for a locus to be kept in the analyses it needed to pass the following requirements: be present in 50% of the individuals (-r 0.50) per population, be present in 80% of individuals across all populations (-R 0.80), and have a minor allele frequency of 0.05 (--min-maf 0.05). All '*populations'* scripts were run following these parameters. Individuals were split into 11 populations that corresponded to their collection sites (Figure 2.1). A first '*populations'* script was run and any individuals with more than 30% missing data were posteriorly removed using VCFTOOLS. A second '*populations*' was run with filtered individuals to produce a vcf file for

downstream analysis that resulted in a matrix with a total of 91 individuals, 19 663 loci, and exhibited only 4.628% of missing data. This initial vcf output file was used to identify outlier loci using the program OUTFLANK (described in further detail below), which was then used to run two more '*populations'* using only neutral or outlier loci (--blacklist or –whitelist, respectively) to produce a genepop output file (--genepop). The genepop files for both the neutral and outlier loci were converted to a structure file using the software GENODIVE v 3.03 (Meirmans, 2020). Finally, since two of our populations had very few individuals that could not be grouped into a neighboring population, we ran one final '*populations*' excluding any populations with less than four individuals to produce a vcf file that was used for downstream analyses to calculate genetic diversity and population differentiation metrics.

Genetic diversity statistics

Genomic statistics were calculated using all 19,809 RADseq loci for populations with at least four individuals per site (9 sites). Number of alleles, nucleotide diversity, observed and expected heterozygosity, and inbreeding coefficient, were calculated on GENODIVE v 3.03 (Meirmans, 2020). Nucleotide diversity and number of alleles were both obtained from the STACKS '*populations'* output. To test whether genetic diversity differs with distance from the population point of origin, genetic statistics were regressed against distance from the Panama collection site. Panama was considered the point of origin since *H. passer* diverged from the Atlantic after the rise of the Isthmus of Panama (Alva-Campbell et al., 2010; Tariel et al., 2016). The

shortest distance over water between sites was calculated on Google Earth Pro v7.3.4.8248. In addition, the TEP experiences high environmental variability across its range, thus, genetic diversity was regressed against sea surface temperature (SST) and chlorophyll conditions to test whether these environmental conditions drive genetic diversity signatures. Chlorophyll layers, 'BO_chlomean' and 'BO_chlomax', were extracted in R using Bio-ORACLE (Tyberghein et al., 2012; Assis et al., 2017) and are based on monthly averages from 2000 to 2014 with a resolution of 5 arcmin. Sea surface temperature layers, 'MS_biogeo16_sst_range_5m' and 'MS_biogeo15_sst_max_5m', were extracted in R using MARSPEC data (Sbrocco and Barber, 2013) that spans monthly averages from 2002 to 2010 with a resolution of 30 arcmin (Sbrocco and Barber, 2013). All linear regressions were calculated and plotted in R using ggplot2 (Wickham, 2011).

We estimated the contemporary effective population size (N_e) for all *H. passer* individuals using two approaches. We first used *NeEstimator* v2.1 (Do et al., 2014) using the linkage dis-equilibrium (LD) method (Waples and Do, 2008) under the random mating model and report jackknifed 95% confidence intervals of a critical value of 0.05. Second, we estimated N_e by obtaining the value of Tajima's π (Pi) from STACKS. When in neutral equilibrium π is correlated with N_e and mutation rates ($\pi = 4 N_e \mu$) (Watterson, 1975; Tajima, 1983). Mutation rate (μ) is expressed as mutation rate per site per generation. In fishes μ has been estimated to be between 10^{-8} to 10^{-9} mutations per site (Brumfield et al., 2003; Crane et al., 2018), thus we ran two simulations to represent the range of the expected mutation rates. Furthermore,

generation time (g) is defined as the age at which half of the individuals of the population are reproducing. Given that *H. passer* is protogynous, generation time for females is around three years, while for males it is around six years, after they transition from female to male (Hernández, 1998; Arellano-Martínez et al., 1999; Sánchez-Alcántara et al., 2006). Thus, we estimate the average generation time *H. passer* to be 5 years.

Kinship

We tested for potential relatedness using kinship coefficients (Loiselle et al., 1995) for each pair of individuals using GENODIVE v 3.03 (Meirmans, 2020). Coancestry coefficients (full-sib = 0.25, half-sib = 0.125) of Loiselle *et al* (1995) were used to generate relatedness bins for 'nearly identical' ($0.57 > k > 0.375$), 'full-sib' ($0.374 > k > 0.1875$), 'half-sib' ($0.1874 > k > 0.09375$) and 'quarter-sib' ($0.09374 > k > 0.047$) (e.g., Iacchei et al., 2013; Crane et al., 2018).

Population structure

Pairwise population differentiation F_{ST} (Weir and Cocckerham, 1984) and G'_{ST} (Nei, 1975), as well as global F_{ST} (AMOVA), were calculated using GENODIVE v 3.03 (Meirmans, 2020) with total of 19,809 RADseq loci and 10,000 permutations. Isolation by distance (IBD) was tested using a Mantel test to compare our pairwise population differentiation matrix (F_{ST}) with a geographic distance matrix. For this analysis, negative values of F_{ST} were replaced by zero, because there is no such thing

as negative gene flow (Hudson et al., 1992). Geographic distance was estimated by using the shortest distance over water between sites using Google Earth Pro v7.3.4.8248.

After adding the small number of samples from Galapagos and Clipperton Island (see Figure 2.1), we ran a Bayesian clustering analysis using 19,635 neutral loci and 28 outlier loci (described in further detail below) with the software STRUCTURE v2.3.4 (Pritchard et al., 2000; Falush et al., 2003, 2007; Hubisz et al., 2009). Five replicates were run for each genetic cluster assumed (K = 1-7), each with a burn-in of 10,000 iterations, 100,000 MCMC reps, no admixture (NOADMIX = 0), and no priori location assumptions (LOCPRIOR = 0). The most likely number of clusters (ΔK) was assessed using the Evanno method (Evanno et al., 2005) as implemented with the R package *pophelper* v2.3.1 (Francis, 2017). *Pophelper* was also used to summarize and plot results from replicate STRUCTURE runs. In addition, a discriminant analysis of principal components (DAPC) was performed on neutral and outlier loci to summarize the diversity and variation of RADseq loci using the R package *adegenet* v2.1.4 (Jombart, 2008; Jombart et al., 2010). The DAPC analysis is a multivariate method that allows us to identify genetic structures in large datasets with the absence of any underlying population genetic model assumption, while maximizing on the variability found between groups (Jombart et al., 2010). To avoid over fitting the DAPC analysis, we used the '*xvalDapc'* function to select the most appropriate number of principal components (PC) to retain for the analysis using

1000 replicates. 30 PCs and two DAs were retained for the analyses and explained 40% of the variance.

Outlier analyses

To compare the relative abundance of SNPs that might be under divergent selection, we used OutFlank version 0.2 (Whitlock and Lotterhos, 2015) to perform pairwise outlier scans between all 11 populations. OutFlank is a software that calculates a likelihood on a trimmed distribution of F_{ST} values to infer the distribution of F_{ST} for neutral markers, which is considered to result in far fewer false positives than other programs (e.g., Bayescan, Lositan). We ran OUTFLANK with a 5% trim on the left and right of the F_{ST} null distribution, a minimum heterozygosity of 0.1 (Hmin=0.1), and a 1% false discovery rate (qthreshold = 0.01). Loci found to be under putative directional selection are referred to "outlier loci" and were separated from neutral loci in to run population structure analyses. Outlier loci were compared to GenBank sequences with BLAST and only matches with a probability of 0.001 and lower of obtaining the same result by chance (*E*-values < 0.001) were kept. Matching sequences that identified as protein-coding genes were then classified using KEGG assignments with BlastKOALA.

Results

Single nucleotide polymorphisms

Restriction site-associated DNA (RAD) libraries were created by individually barcoding 102 individuals of *Holacanthus passer* from across its range (Figure 2.1). STACKS filtering parameters resulted in an initial 113,825 catalog loci with a mean coverage of 17.9x (SD 12.4x) per sample. Subsequent filtering for loci with a minimum minor allele frequency of 0.05, loci found in more than 80% of all individuals, and more than 50% of individuals per population, resulted in 19,663 loci. A total of 91 individuals across 11 populations remained after removing any duplicate individuals and individuals with more than 30% missing data (Figure 2.1). However, we recognize that two populations had very few individuals (< 4), and although those individuals are important and interesting for the overall population structure, to not bias our results, we removed them from the genetic diversity and *F*statistics dataset. Here, a total of 88 individuals from nine populations and 19,809 loci remained.

Genetic diversity and effective population size

A summary of the principal genetic diversity statistics (mean number of alleles, observed and expected heterozygosity, nucleotide diversity, and inbreeding coefficient) is presented in Figure 2.1. We tested whether *H. passer* populations showed genetic diversity signatures of population range expansion considering Panama as the point of origin. In general, we would expect greater diversity at the point of origin (Bors et al., 2019), however, evidence showed no significant correlation between distance from the origin and observed and expected heterozygosity, nucleotide diversity, or number of alleles (0.16 < p-value < 0.66)

(Figure 2.5). Additionally, we tested whether nucleotide diversity (*Pi*) or mean allelic richness (*Na*) could be explained by environmental conditions such as: SST range, SST max, mean chlorophyll, and maximum chlorophyll. Neither *Pi* or *Na* were explained by the environmental conditions in the TEP ($0.18 < \text{p-value} < 0.97$) (Figure 2.6).

Effective population size was determined using direct values of N_e based on the LD method and by using values of π (Pi). Using NeEstimator (LD method) N_e was estimated to be between 388.9 – inifinity, while calculations based on π from STACKS (*H.passer* π = 0.28935) gave a narrower estimation N_e between 1.45×10^6 and 14.5×10^6 individuals.

Kinship

Considering the wide range distribution, high abundance, and large effective population size of *H. passer* in the TEP, finding related pairs of individuals is unlikely. Nonetheless, we wanted to test whether we found evidence of kinship within and between populations. Only one 'quarter sib' ($0.09374 > k > 0.047$) pair was identified between individual 'HPA_LFR_030509' from Cabo Pulmo in the Sea of Cortez and 'HPA_RCP_110504' from Panama. If this result indicates true kinship (samples were collected the same year, eight months apart), these individuals sampled approximately ~4,000 km apart underscore the vast dispersal capability of *H. passer*.

Table 2.1. Population genomic summary statistics of Holacanthus passer populations based on 19,809 RADseq loci, generated using Genodive and the Stacks.

Region	Site	Site ID	Lat	Long	*N*	*N* (Stacks)	*Na*	H_O	H_E	*Pi* (Stacks)	F_{IS}
North Sea of Cortez											
	Isla Ángel de la Guarda	IAG	-113.5930	29.5317	6	5.662	1.797	0.281	0.282	0.282	0.005
	San Pedro Martir	SPM	-112.3206	28.3850	6	5.653	1.808	0.288	0.288	0.288	0.003
South Sea of Cortez											
	La Paz	LPA	-110.0745	24.2043	12	11.511	1.935	0.275	0.288	0.288	0.048
	Cabo Pulmo	CPU	-109.4264	23.3567	11	10.833	1.933	0.286	0.292	0.291	0.018
	Los Cabos	LCA	-109.8435	22.9020	4	3.681	1.681	0.279	0.279	0.278	0.000
Baja California- Pacific											
	Bahía Magdalena	BMA	-112.0584	24.5437	7	6.546	1.833	0.279	0.282	0.282	0.009
Mainland Mexico											
	Zihuatanejo	ZIH	-101.5541	17.6222	17	16.237	1.968	0.264	0.291	0.290	0.091
Clipperton											
	Clipperton	CLI	-109.2069	10.3138	1	-	-	-	-	-	-
Panama											
	Isla Contadora- 2007	ICO_07	-79.0423	8.6346	19	18.206	1.979	0.260	0.291	0.292	0.108
	Isla Contadora- 2017	ICO_17	-79.0423	8.6346	6	5.696	1.805	0.296	0.286	0.289	-0.035
Galapagos											
	Galapagos	GAL	-89.7221	-1.3533	2	-	-	-	-	-	-

N: number of individuals; *N* (Stacks): average number of individuals used across all sampled loci; *Na:* Number of alleles; H_O*:* observed heterozygosity: H_E*:* expected heterozygosity; *Pi:* nucleotide diversity; F_{IS}*:* inbreeding coefficient.

Population structure

Global F_{ST} among all populations revealed no significant genetic differentiation while accounting for all loci (F_{ST} = 0.00; p-value=1). Pairwise population differentiation F_{ST} ranged between -0.002 – 0.006 and G'_{ST} between -0.005 – 0.005 (Figure 2.2). Low but significant differentiation was found between Isla Contadora 2017 in Panama and most Sea of Cortez populations (0.007 < p-value < 0.043), with the most significant comparisons between Panama (ICO_17) and North Sea of Cortez populations (IAG, SPM) (Figure 2.2). Interestingly, individuals from Isla Contadora collected in 2007 did not show the same significant differentiation, showing a temporal difference. Yet, sampling size differed between both years, which may be driving this difference (Table 2.1). Finally, we found significant evidence of IBD (Figure 2.4) (R^2=0.304; p-value = 0.02).

The Bayesian STRUCTURE analysis revealed that the most likely number of clusters based on ΔK was K = 3 for both the neutral and outlier loci detected by OUTFLANK (Figure S4; S5). Visually, no distinctive pattern emerges between populations (Figure 2.2). However, within the outlier loci, Galapagos shows the most distinctive genetic differentiation in K = 3. The DAPC on neutral loci revealed Galapagos clustered separately from the rest of the populations (Figure 2.3A). In addition, the DAPC analysis on outlier loci revealed populations from Galapagos and Clipperton both clustered separately (Figure 2.3B). To get a closer look at most of the populations, we removed Galapagos and Clipperton from the DAPC analyses due to their small population size and ran an additional analyses. Here, neutral loci revealed

no distinctive clustering, however, Cabo Pulmo clustered separately in outlier loci. Interestingly, both Panama populations that were sampled in the same location 10 years apart, clustered separately from each other in both neutral and outlier loci (Figure 2.3).

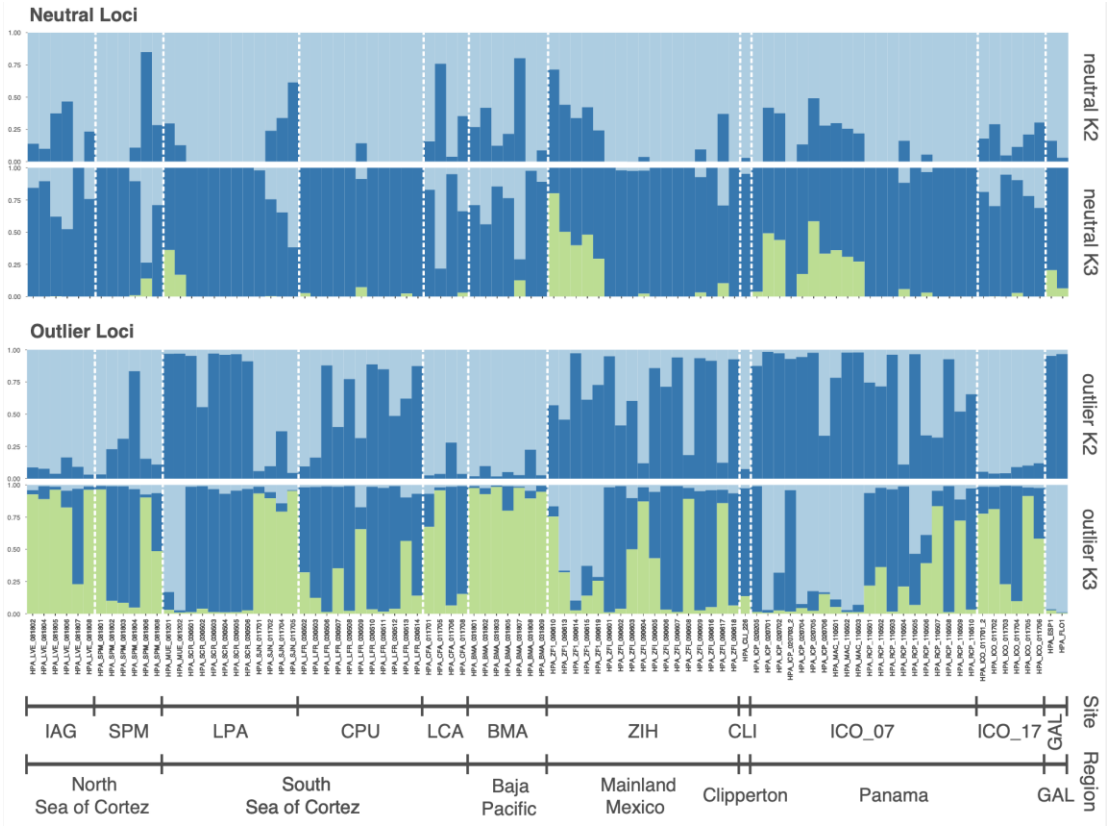


Figure 2.2. Bayesian clustering analysis of *Holacanthus passer* for neutral and outlier loci, assuming no priori. Plots show K = 2 and K = 3 using 19,635 neutral loci (top) and 28 outlier loci (bottom). The most likely number of clusters based on ΔK was K = 3 for both neutral and outlier loci. Sampling sites and regions are arranged from North to South and from West to East of the TEP (See Figure 2.1 and Figure 2.1 for site details).

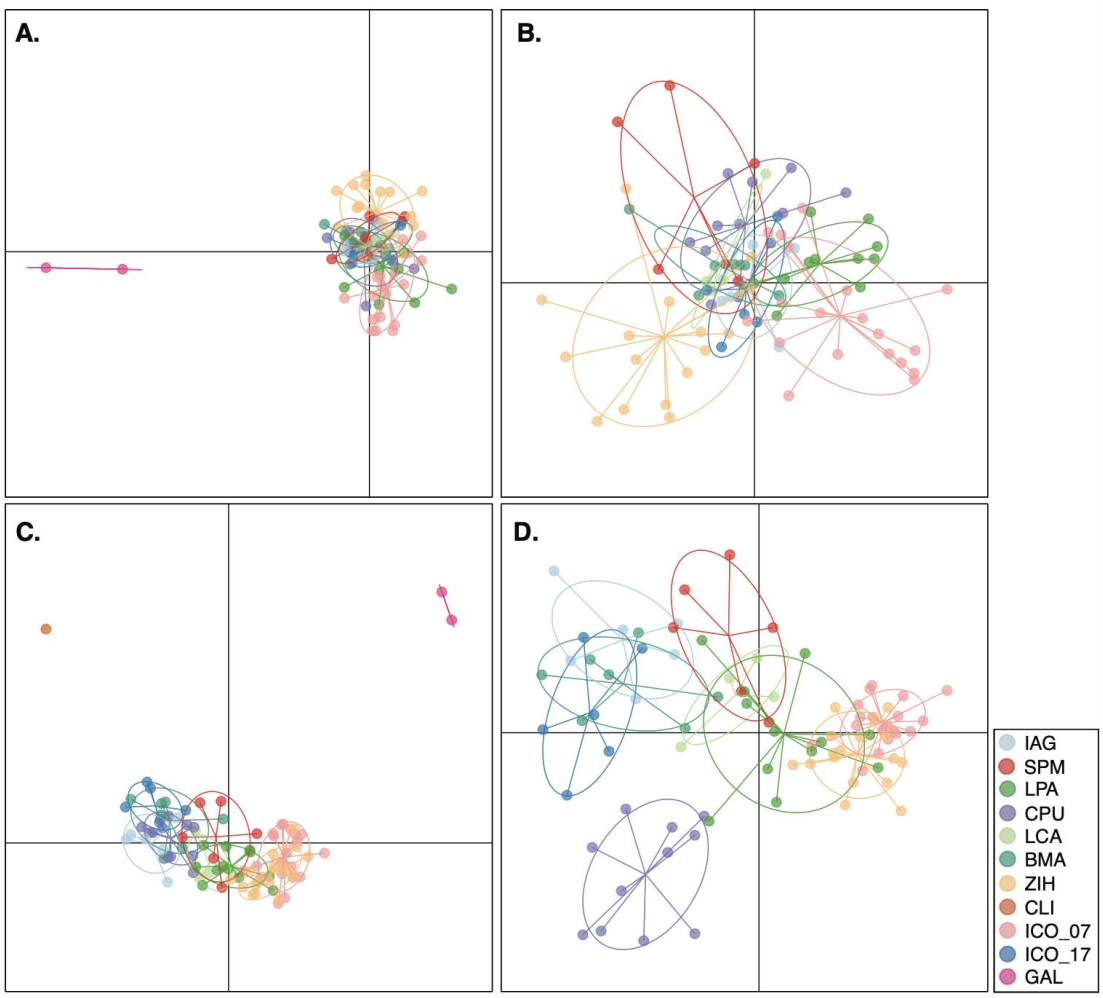


Figure 2.3. Discriminant analysis of principal components (DAPC) for *H. passer* RADseq markers showing: only neutral loci (top: a,b), only outlier loci (bottom: c,d), and with all samples (a,c) and after excluding Galapagos and Clipperton populations due to small sample size (b and d). Analyses retained 30 PCs and two DAs which explained 40% of the variance.

Outlier analyses

OUTFLANK identified 28 loci under putative divergent selection from a total of 19,663 loci (Figure S1). Of the 28 loci, 19 matched GenBank entries and all of them corresponded to fish sequences. From the 19 matched sequences, only six matched

protein-coding regions and the remaining 13 sequences matched unannotated fish genome sequences. Protein-coding regions were analyzed using the program *Kegg Koala* and they clustered to four functional groups: environmental information processing (n=3), nucleotide metabolism (n=1), glycan biosynthesis and metabolism (n=1), and genetic information processing (n=1) (Table S1).

Table 2.2. Pairwise F_{ST} values (above the diagonal) and Nei's G'_{ST} (below the diagonal) between populations based on 19,809 RADseq loci. Bold values indicate significant differentiation. Key: IAG, Isla Ángel de la Guarda; SPM, San Pedro Martir; LPA, La Paz; CPU, Cabo Pulmo; LCA, Los Cabos; BMA, Bahía Magdalena; ZIH; Zihuatanejo; ICO_07, Isla Contadora 2007; ICO_17, Isla Contadora 2017.

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
1. IAG		0.002	-0.001	0.000	0.002	0.001	-0.003	-0.001	**0.004****
2. SPM	0.002		0.000	-0.002	-0.002	-0.001	-0.003	-0.001	**0.005****
3. LPA	0.000	0.000		0.000	-0.002	0.001	0	0.000	0.002
4. CPU	0.000	-0.001	0.000		-0.001	-0.001	0	0.000	0.001
5. LCA	0.002	-0.001	0.000	0.001		-0.001	-0.005	-0.005	**0.005***
6. BMA	0.001	-0.001	0.001	0.000	0.001		-0.003	-0.001	**0.002***
7. ZIH	-0.002	-0.001	0.001	0	-0.002	-0.001		0.002	0.000
8. ICO_07	0.001	0.001	0.000	0.001	-0.002	0.000	0.002		0.002
9. ICO_17	**0.004****	**0.004****	**0.003***	**0.002***	**0.006***	**0.002***	0.002	0.002	

(* < 0.05; ** < 0.005)

Discussion

The lack of physical barriers in the marine environment led to the initial assumption that most marine populations were homogenous. However, a growing body of evidence suggests that dispersal is often highly constrained (Jones et al., 1999; Swearer et al., 1999). Recently, with the advent of more powerful genomic techniques that allowed to obtain thousands of genome wide markers, population genomic

studies have been able to detect more subtle genetic differences. In this study, we found that *H. passer* populations displayed high gene flow and similar genetic diversity along the TEP coastline, as expected from previous studies (see review by Lessios and Baums, 2016). However, we detected 28 outlier loci that drive subtle genetic signatures that differentiate the mainland, Galapagos, and Clipperton Island populations. In addition, pairwise differentiation detected low but significant structure between individuals collected in Panama and most of the Sea of Cortez populations showing evidence of isolation by distance. Interestingly, individuals collected in the same location in Panama 10 years prior showed no evidence of structure.

Large Ne and early population expansion as potential drivers for homogeneous genetic diversity

Genetic diversity values were similar throughout the sampling range of *H. passer* (Figure 2.1) and were comparable to other studies on marine fishes using RADseq (e.g., Saenz-Agudelo et al., 2015; DiBattista et al., 2017; Bors et al., 2019). The homogeneity in the genetic diversity levels may be due to their large effective population size (*Ne:* 388 – infinity; $1.45x10^6$ – $14.5x10^6$), their potentially long PLD (~23-26 days estimated from closest relative *P. diacanthus*; Thresher and Brothers, 1985), and/or their early population expansion dated between 300 Kya and 2.8 Mya (Gatins R et al., *in review*).

Holacanthus passer diverged in the TEP from its Atlantic geminate species (*H. ciliaris*) around 1.7 to 1.4 Mya (Alva-Campbell et al., 2010; Tariel et al., 2016),

after the closure of the Isthmus of Panama that is suggested to have occurred around 3.1 to 3.5 Mya (O'Dea et al., 2016). A recent demographic study of *H. passer* shows a slow population expansion took place more than 300 Kya (Gatins R et al., *in review*), allowing enough time for high gene flow to homogenize allelic frequencies throughout its range. The demographic study by Gatins et al (*in review)* estimated *Ne* to be between 300,000 and 3,000,000 individuals (using a mutation rate, μ, of 10^{-8} to 10^{-9}, respectively). Our study shows *Ne* is most likely in the millions and seems to better support the demographic model showing an earlier population expansion taking place around 2.8 Mya ($\mu = 10^{-9}$), which is consistent with an event happening closer to the closure of the Isthmus of Panama.

Considering the straight coastline distribution of *H. passer*, and our extensive sampling throughout, an IBD test was used to test genetic diversity patterns between the range center and range edge. Due to the evolutionary history of *H. passer*, we assumed the population center to be Panama (Bellwood et al., 2004; Alva-Campbell et al., 2010; Tariel et al., 2016). In theory, populations at the range edge are often characterized by having smaller population sizes and lower genetic diversity (Brow et al., 1995; Vucetich and Waite, 2003; Slatkin and Excoffier, 2012). However, we found no correlation between any genetic diversity metric (i.e., observed and expected heterozygosity, nucleotide diversity, and mean allelic richness) and the distance from the population center (Figure 2.5). The TEP is known to experience extreme temperature range shifts and upwelling systems (Kessler, 2006) that could possibly drive non spatial genetic patterns. Thus, we additionally checked whether

there was a correlation between genetic diversity and the environment (SST, CHLO).

Nonetheless, results showed no significant correlation (Figure 2.6).

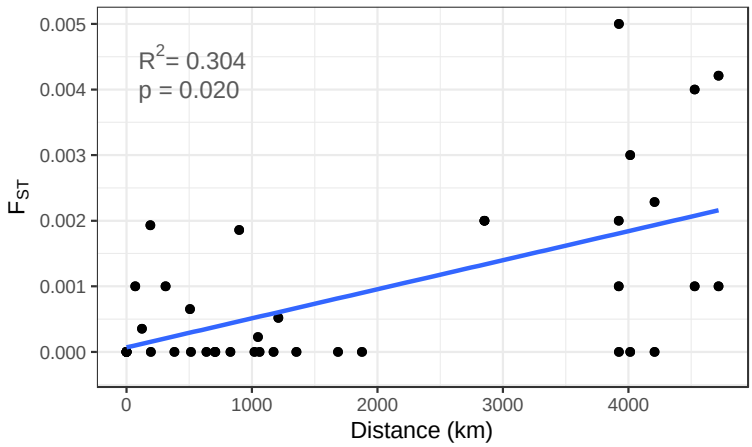


Figure 2.4. Isolation by distance of 9 mainland populations using a total of 88 individuals and 19,809 SNPs. Distance represents the shortest aquatic distance between populations measured on GoogleEarth. Negative F_{ST} pairwise population comparisons were set to zero. The shaded area represents 95% confidence intervals. Reported R^2 and p-value were calculated with a Mantel test with 10,000 permutations.

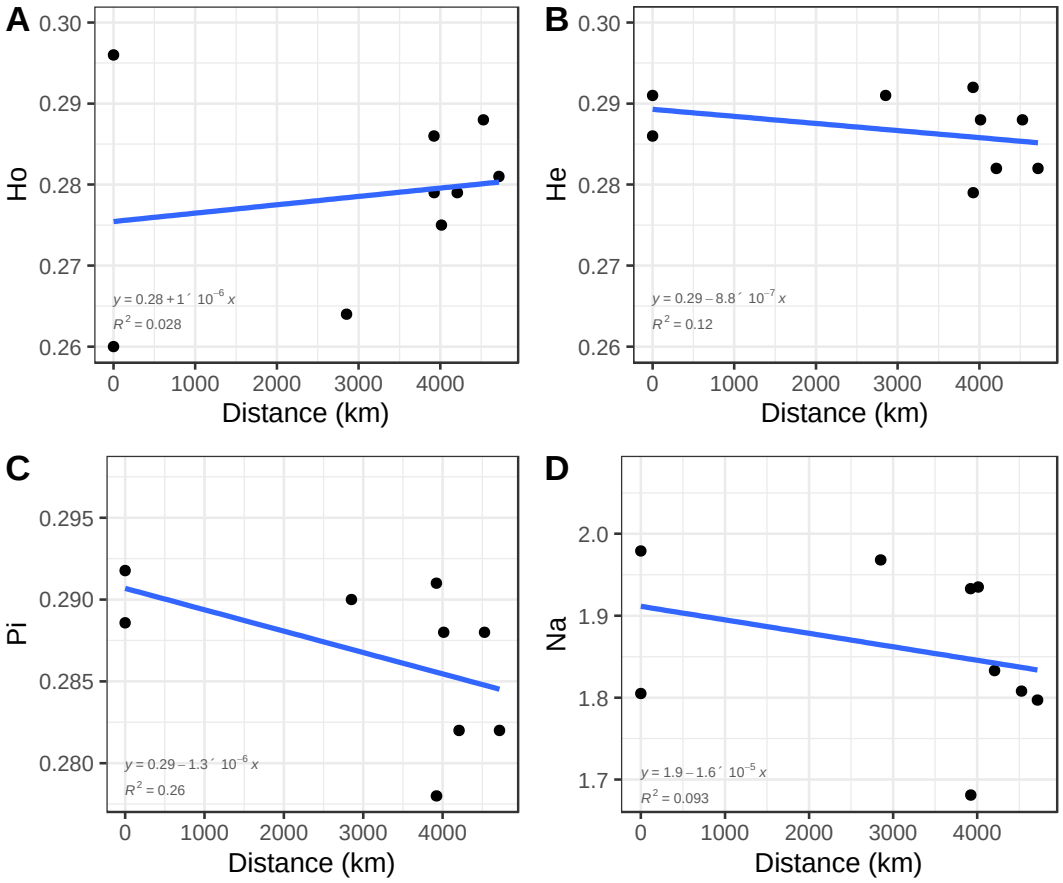


Figure 2.5. Linear regression models comparing distance from the population point of origin (Panama) with (a) observed heterozygosity, (b) expected heterozysity, (c) nucleotide diversity, and (d) number of alleles. Shaded area represents 95% confidence intervals. (p-values = 0.16 < p < 0.66).

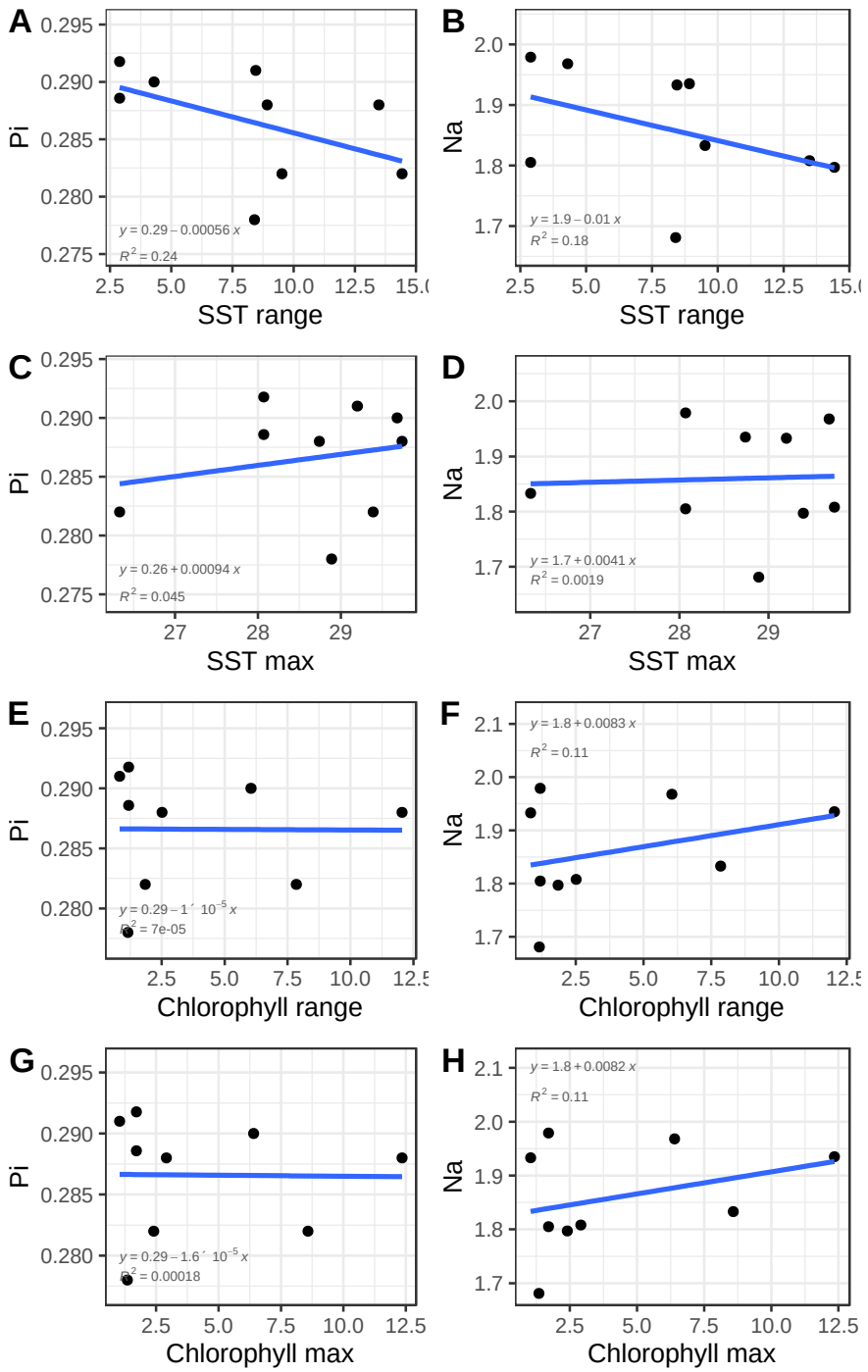


Figure 2.6. Linear regression models comparing environmental conditions to nucleotide diversity (Pi; left) and number of alleles (Na; right). Environmental conditions: SST range, sea surface temperature range (SST range); SST max, sea surface temperature max; Chlorophyll mean, chlorophyll max. Shaded area represents 95% confidence intervals. (p-values = 0.18 < p < 0.97).

Subtle population structure despite high gene flow across the Tropical Eastern Pacific

Global F_{ST} values indicate high gene flow of *H. passer* across the mainland, suggesting panmixia. This result aligns well with most of the literature on genetic connectivity in the TEP (Craig et al., 2006; Lessios and Robertson, 2006; Bernardi et al., 2008; Pinzón and LaJeunesse, 2011; Lessios and Baums, 2016). High gene flow additionally supports our earlier results showing similar genetic diversity across its range. Despite this, low but significant structure was identified using pairwise population differentiation (F_{ST} and G'_{ST}) between the Isla Contadora in Panama (2017 sampling) and the Sea of Cortez populations (Figure 2.2), and followed a positive IBD trend (Figure 2.4). IBD has previously been detected in the area and particularly with species that only occur along the TEP coast (Craig et al., 2006; Baums et al., 2012; Lessios and Baums, 2016; Reguera-Rouzaud et al., 2021). Population differentiation between the Sea of Cortez and the rest of the coastal TEP has also been reported in *Lutjanus peru* (Reguera-Rouzaud et al., 2021), *Haemulon flaviguttatum* (Bernal et al., 2017), and *Hippocammmpus ingens,* (Saarman et al., 2010; Lessios and Baums, 2016).

Interestingly, although two sampling periods were carried out in Panama 10 years apart, only the 2017 sampling showed this weak but significant pairwise differences, suggesting a temporal mismatch consistent with chaotic genetic patchiness. The kinship analyses found no closely related individuals, thus excluding the possibility of sweepstake reproductive success or collective dispersal driving this

pattern. However, the last strong El Niño event was reported in 2015-2016 and lead to a long period of increased ocean temperatures that caused significant coral bleaching and calcification decline(Brainard et al., 2018). Although not conclusive, high larval mortality during this event may possibly explain our temporal mismatch of cryptic genetic patterns.

After the addition of two *H. passer* individuals collected in the Galapagos and one vagrant collected in Clipperton Island (Figure 2.1), we detected 28 outlier loci (Figure S1). Although both oceanic populations suffer from a low sampling size, and were consequently removed from all prior analyses, we believe they were important to keep during our population structure analysis (i.e., STRUCTURE and DAPC). The individual collected in Clipperton was the first record of *H. passer* on the island (Clua and Planes, 2019), thus obtaining more samples from here was highly unlikely. Bayesian clustering analysis suggested K = 3 as the most likely number of clusters with neutral and outlier loci (Figure S4, S5). However, no clear visual pattern seemed to arise in the STRUCTURE plots seen in Figure 2.2 (see Figure S2, S3 for K2 – K7 plots), except for the Galapagos individuals from the outlier loci K3 plot. The DAPC gave a higher resolution result than the STRUCTURE plots and clearly showed Galapagos clustering separately using only neutral loci (Figure 2.3a), and both Galapagos and Clipperton clustering away from the rest of the populations when only using the outlier loci (Figure 2.3c). These results suggest that the outlier loci seem to drive most of the genetic differences between the oceanic islands and the coastal mainland populations, in particular that of Clipperton. However, when we remove the

oceanic islands from the DAPC analysis to get a better look at the mainland populations, we surprisingly do not see Panama (ICO_17) cluster separately with either the neutral or outlier loci (Figure 2.3b, 3d), contrary to what we expected given our pairwise differentiation results. Cabo Pulmo showed the most distinctive clustering when we only use the outlier loci (Figure 2.3d).

According to Robertson and Cramer (2009), the TEP is divided into three main biogeographic regions: the oceanic islands/archipelagos, and within the continental coast, the Cortez and Panamic Province. The Cortez province encompasses the Sea of Cortez and lower Pacific Baja, while the Panamic province covers the entire southward continental coast. These distinct biogeographic provinces were defined using (i) the number of endemic fish species and (ii) species richness per area (Robertson and Cramer, 2009). The continental provinces are hypothesized to be separated by the Sinaloan Gap – a long stretch with rocky reef habitat that may act as a barrier (Figure 2.1) (Hastings, 2000). However, it has also been hypothesized that the south-westward eddy found at the entrance of the Sea of Cortez, may act as a barrier separating the Cortez and Panamic province (Kurczyn et al., 2012). More recent studies suggest that environmental differences between the subtropical and equatorial regions seem to be responsible for the differences we see in species composition between the northern and southern TEP (Rocha and Bowen, 2008; Robertson and Cramer, 2009; Briggs and Bowen, 2012). Overall, these results suggest that the continental barrier between the Cortez and Panamic province may be driven by multiple factors.

In conclusion, this study presents the first exhaustive population genetic study of a reef fish in the Tropical Eastern Pacific using thousands of genome wide markers. The King angelfish, *Holacanthus passer*, comprises one largely panmictic population. However, our findings add to the growing evidence of weak but significant structure in the presence of high gene flow in marine populations. This subtle genetic structure supports the designation of the three main biogeographic provinces in the TEP: the Cortez Province, Panamic Province, and the oceanic islands/archipelagos. However, spatial genetic patterns were not consistent across time, leading to some degree of chaotic genetic patchiness. Finally, we detected genomic signatures of spatially divergent selection at a few select loci, some of which were associated with environmental conditions. Future studies should incorporate local adaptation, as well as spatial and temporal data to provide the best insights into what processes predictably structure biological variation in metapopulations.

Supplementary Materials

Table S1. Functional groups of protein-coding regions identified by the program *Kegg Koala*.

Loci_ID	KO	Definition	Functional Category
5268	K01191	MAN2C1; alpha-mannosidase [EC:3.2.1.24]	Glycan biosynthesis and metabolism
6599	K16753	CCDC34; coiled-coil domain-containing protein 34	Protein families: genetic information processing
32936	K08270	DDIT4, REDD1;DNA-damage-inducible transcript 4	Environmental Information Processing
44027	K20064	GRB10; growth factor receptor-bound protein 10	Environmental Information Processing
57633	K04358	FGF; fibroblast growth factor	Environmental Information Processing
62605	K03783	punA, PNP; purine-nucleoside phosphorylase [EC:2.4.2.1]	Nucleotide metabolism

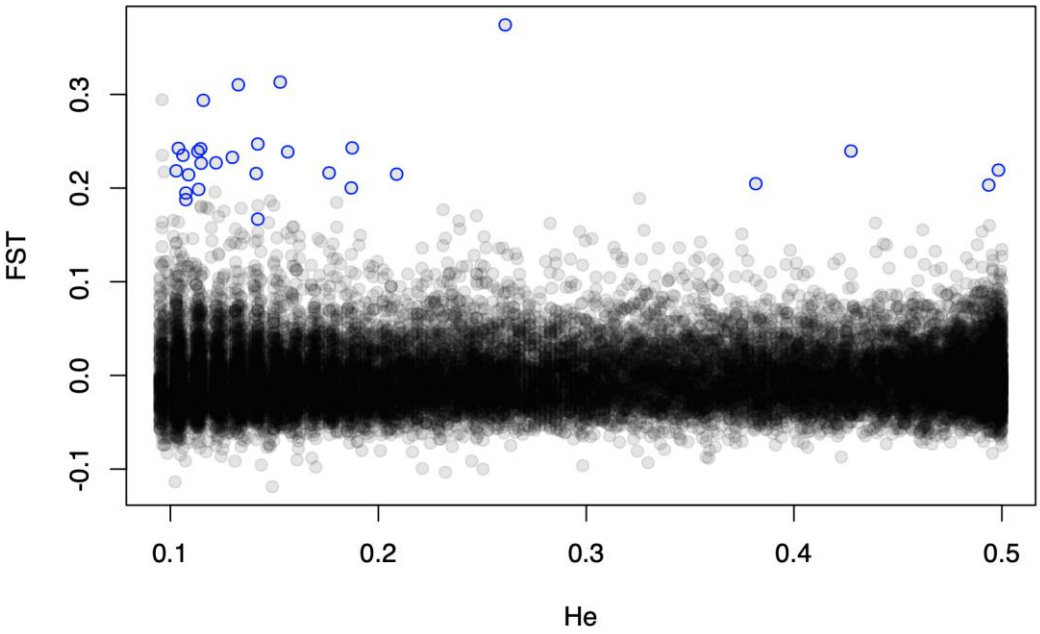


Figure S1. Outlier scan carried out using OUTFLANK for 19,663 loci, which detected 28 loci under putative divergent selection (highlighted in blue).

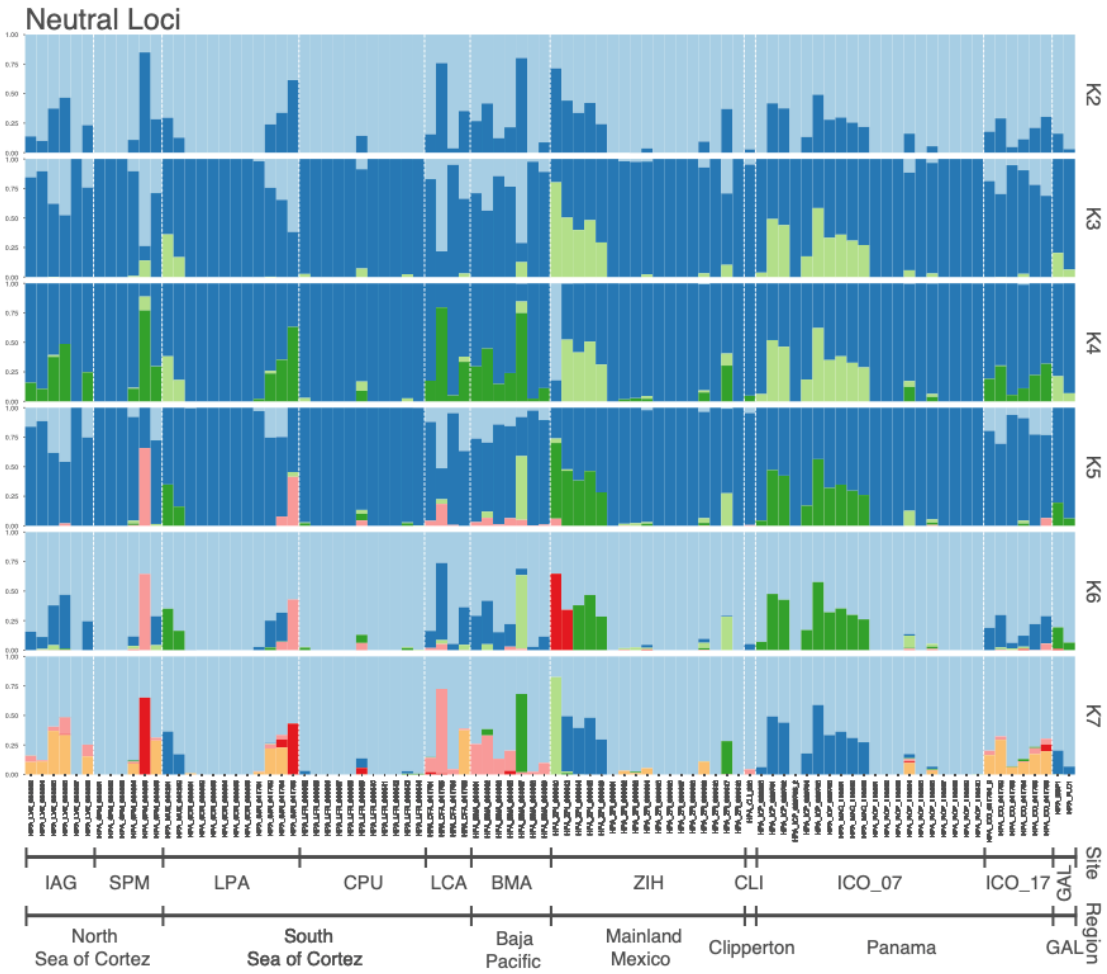


Figure S2. STRUCTURE plots generated from 19,635 neutral loci for *H. passer* for K= 2 - 7. The most likely number of clusters based on ΔK was K = 3. Sampling sites and regions are arranged from North to South and from West to East of the TEP. Sampling site key: IAG- Isla Ángel de la Guarda, SPM- San Pedro Mar, LPA- La Paz, CPU- Cabo Pulmo, LCA- Los Cabos, BMA- Bahía Magdalena, ZIH- Zihuatanejo, CLI- Clipperton, ICO_07- Isla Contadora 2007, ICO_17- Isla Contadora 2017, GAL- Galapagos.

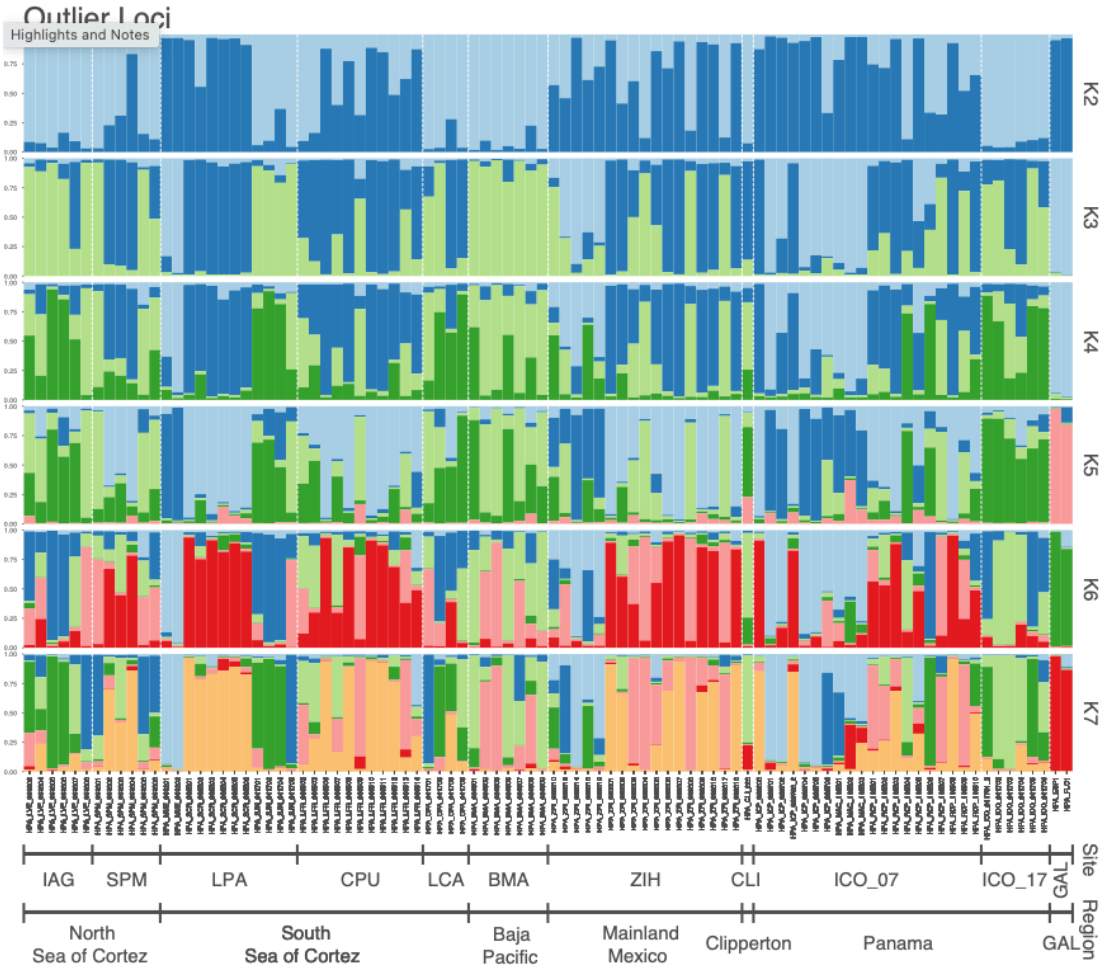


Figure S3. STRUCTURE plots generated from 28 outlier loci for *H. passer* for K= 2 - 7. The most likely number of clusters based on ΔK was K = 3. Sampling sites and regions are arranged from North to South and from West to East of the TEP. Sampling site key: IAG- Isla Ángel de la Guarda, SPM- San Pedro Mar, LPA- La Paz, CPU- Cabo Pulmo, LCA- Los Cabos, BMA- Bahía Magdalena, ZIH- Zihuatanejo, CLI- Clipperton, ICO_07- Isla Contadora 2007, ICO_17- Isla Contadora 2017, GAL- Galapagos.

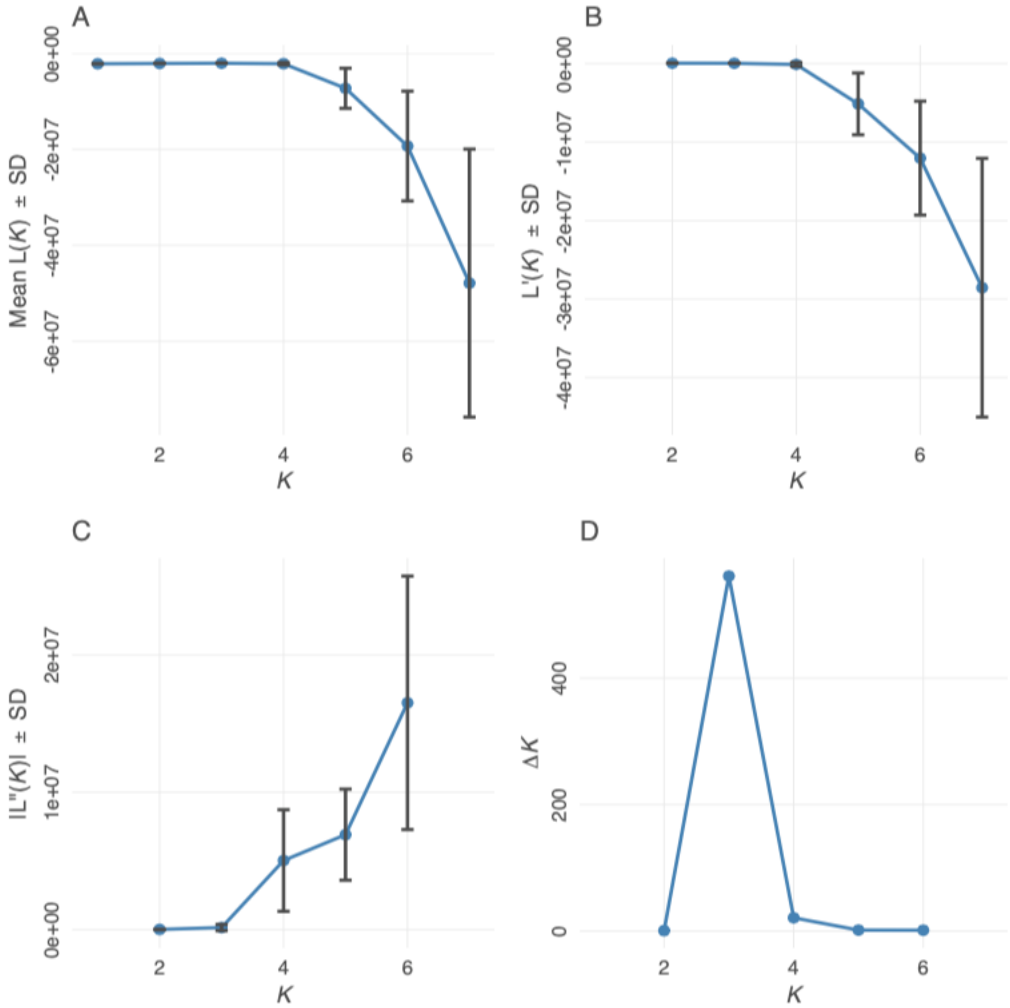


Figure S4. Evanno method plots from 19,635 neutral loci showing (A) the estimated log probability over increasing values of K, (B) first derivative, (C) second derivative, and (D) ΔK. The most likely number of clusters based on (D) ΔK shows K = 3.

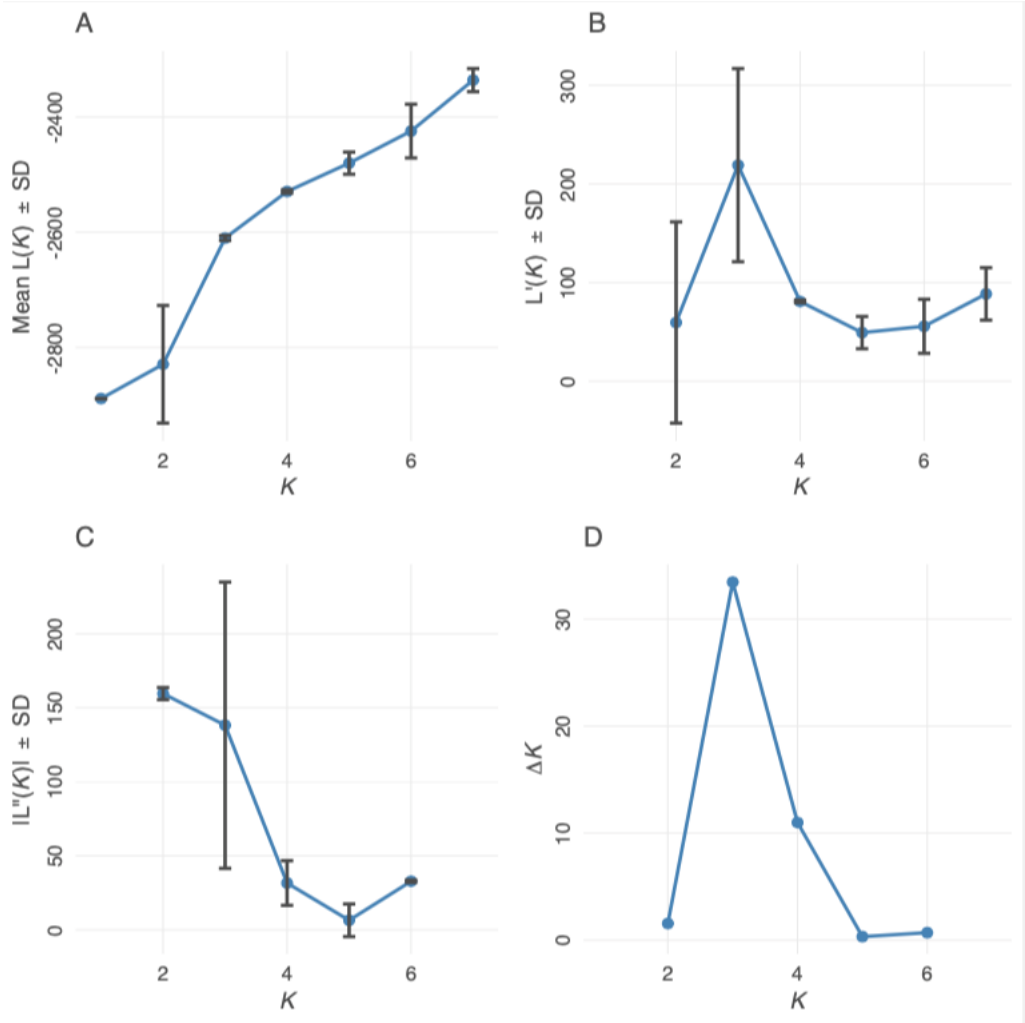


Figure S5. Evanno method plots from 28 outlier loci showing (A) the estimated log probability over increasing values of K, (B) first derivative, (C) second derivative, and (D) ΔK. The most likely number of clusters based on (D) ΔK shows K = 3.

References

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

Falush, D., Stephens, M., and Pritchard, J. K. (2003). Inference of Population Structure Using Multilocus Genotype Data: Linked Loci and Correlated Allele Frequencies. *Genetics*.

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

Francis, R. M. (2017). pophelper: an R package and web app to analyse and visualize population structure. *Mol Ecol Resour* 17, 27–32. doi:10.1111/1755-0998.12509.

Gatins, R., Arias, C. F., Sánchez, C., Bernardi, G., and DeLeón, L. F. Whole genome assembly and annotation of the King Angelfish (*Holacanthus passer*) gives insight into the evolution of marine fishes of the Tropical Eastern Pacific. *in review*.

Hastings, P. A. (2000). Biogeography of the Tropical Eastern Pacific: distribution and phylogeny of chaenopsid fishes. *Zool J Linn Soc-lond* 128, 319–335. doi:10.1111/j.1096-3642.2000.tb00166.x.

Hernández, M. C. (1998). Estructura de tallas y crecimiento individual del Ángel Rey, *Holacanthus passer*, Valenciennes 1846 (Teleostei: Pomacanthidae), en la Bahía de La Paz, B.C.S. México.

Hubisz, M. J., Falush, D., Stephens, M., and Pritchard, J. K. (2009). Inferring weak population structure with the assistance of sample group information. *Mol Ecol Resour* 9, 1322–1332. doi:10.1111/j.1755-0998.2009.02591.x.

Hudson, R. R., Slatkin, M., and Maddison, W. P. (1992). Estimation of levels of gene flow from DNA sequence data. *Genetics* 132, 583–589. doi:10.1093/genetics/132.2.583.

Iacchei, M., Ben-Horin, T., Selkoe, K. A., Bird, C. E., García-Rodríguez, F. J., and Toonen, R. J. (2013). Combined analyses of kinship and FST suggest potential

drivers of chaotic genetic patchiness in high gene-flow populations. *Mol Ecol* 22, 3476–3494. doi:10.1111/mec.12341.

Johnson, M. S., and Black, R. (1982). Chaotic genetic patchiness in an intertidal limpet, *Siphonaria* sp. *Mar Biol* 70, 157–164. doi:10.1007/bf00397680.

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

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

Jones, G. P., Milicich, M. J., Emslie, M. J., and Lunow, C. (1999). Self-recruitment in a coral reef fish population. *Nature* 402, 802–804. doi:10.1038/45538.

Kessler, W. S. (2006). The circulation of the eastern tropical Pacific: A review. *Prog Oceanogr* 69, 181–217. doi:10.1016/j.pocean.2006.03.009.

Kurczyn, J. A., Beier, E., Lavín, M. F., and Chaigneau, A. (2012). Mesoscale eddies in the northeastern Pacific tropical-subtropical transition zone: Statistical characterization from satellite altimetry. *J Geophys Res Oceans 1978 2012* 117, n/a-n/a. doi:10.1029/2012jc007970.

Lessios, H. A., and Baums, I. B. (2016). Coral Reefs of the Eastern Tropical Pacific, Persistence and Loss in a Dynamic Environment. 477–499. doi:10.1007/978-94-017-7499-4_16.

Lessios, H. A., Kane, J., and Robertson, D. R. (2003). Phylogrography of the pantropical sea urchin *Tripneustes*: Contrasting patterns of population structure between oceans. *Evolution* 57, 2026–2036. doi:10.1554/02-681.

Lessios, H. A., and Robertson, D. R. (2006). Crossing the impassable: genetic connections in 20 reef fishes across the eastern Pacific barrier. *Proc Royal Soc B Biological Sci* 273, 2201–2208. doi:10.1098/rspb.2006.3543.

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

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25, 2078–2079. doi:10.1093/bioinformatics/btp352.

Loiselle, B. A., Sork, V. L., Nason, J., and Graham, C. (1995). Spatial Genetic Structure of a Tropical Understory Shrub, *Psychotria officinalis* (Rubiaceae). *Am J Bot* 82, 1420. doi:10.2307/2445869.

Longo, G. C., Lam, L., Basnett, B., Samhouri, J., Hamilton, S., Andrews, K., et al. (2020). Strong population differentiation in lingcod (*Ophiodon elongatus*) is driven by a small portion of the genome. *Evol Appl* 13, 2536–2554. doi:10.1111/eva.13037.

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

Moody, K. N., Hunter, S. N., Childress, M. J., Blob, R. W., Schoenfuss, H. L., Blum, M. J., et al. (2015). Local adaptation despite high gene flow in the waterfall-climbing Hawaiian goby, *Sicyopterus stimpsoni*. *Mol Ecol* 24, 545–563. doi:10.1111/mec.13016.

Mora, C., Treml, E. A., Roberts, J., Crosby, K., Roy, D., and Tittensor, D. P. (2012). High connectivity among habitats precludes the relationship between dispersal and range size in tropical reef fishes. *Ecography* 35, 89–96. doi:10.1111/j.1600-0587.2011.06874.x.

Moyer, J. T., Thresher, R. E., and Colin, P. L. (1983). Courtship, spawning and inferred social organization of American angelfishes (Genera *Pomacanthus*, *Holacanthus* and *Centropyge*; pomacanthidae). *Environ Biol Fish* 9, 25–39. doi:10.1007/bf00001056.

Nei, M. (1975). *Molecular population genetics and evolution*. North Holland, Amsterdam, The Netherlands.

O'Dea, A., Lessios, H. A., Coates, A. G., Eytan, R. I., Restrepo-Moreno, S. A., Cione, A. L., et al. (2016). Formation of the Isthmus of Panama. *Sci Adv* 2, e1600883. doi:10.1126/sciadv.1600883.

Pinzón, J. H., and LaJeunesse, T. C. (2011). Species delimitation of common reef corals in the genus *Pocillopora* using nucleotide sequence phylogenies,

population genetics and symbiosis ecology. *Mol Ecol* 20, 311–325. doi:10.1111/j.1365-294x.2010.04939.x.

Pritchard, J. K., Stephens, M., and Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics* 155, 945–59.

Reguera-Rouzaud, N., Díaz-Viloria, N., Pérez-Enríquez, R., Espino-Barr, E., Rivera-Lucero, M. I., and Munguía-Vega, A. (2021). Drivers for genetic structure at different geographic scales for Pacific red snapper (*Lutjanus peru*) and yellow snapper (*Lutjanus argentiventris*) in the tropical eastern Pacific. *J Fish Biol*. doi:10.1111/jfb.14656.

Riginos, C. (2005). Cryptic vicariance in Gulf of California fishes parallels vicariant patterns found in Baja California mammals and reptiles. *Evolution* 59, 2678–2690. doi:10.1554/05-257.1.

Robertson, D., and Cramer, K. (2009). Shore fishes and biogeographic subdivisions of the Tropical Eastern Pacific. *Mar Ecol Prog Ser* 380, 1–17. doi:10.3354/meps07925.

Rocha, L. A., and Bowen, B. W. (2008). Speciation in coral-reef fishes. *J Fish Biol* 72, 1101–1121. doi:10.1111/j.1095-8649.2007.01770.x.

Rochette, N. C., Rivera-Colón, A. G., and Catchen, J. M. (2019). Stacks 2: Analytical methods for paired-end sequencing improve RADseq-based population genomics. *Mol Ecol* 28, 4737–4754. doi:10.1111/mec.15253.

Romero-Torres, M., Treml, E. A., Acosta, A., and Paz-García, D. A. (2018). The Eastern Tropical Pacific coral population connectivity and the role of the Eastern Pacific Barrier. *Sci Rep-uk* 8, 9354. doi:10.1038/s41598-018-27644-2.

Saarman, N. P., Louie, K. D., and Hamilton, H. (2010). Genetic differentiation across eastern Pacific oceanographic barriers in the threatened seahorse *Hippocampus ingens*. *Conserv Genet* 11, 1989–2000. doi:10.1007/s10592-010-0092-x.

Saenz-Agudelo, P., Dibattista, J. D., Piatek, M. J., Gaither, M. R., Harrison, H. B., Nanninga, G. B., et al. (2015). Seascape genetics along environmental gradients in the Arabian Peninsula: insights from ddRAD sequencing of anemonefishes. *Mol Ecol* 24, 6241–6255. doi:10.1111/mec.13471.

Sánchez-Alcántara, I., Aburto-Oropeza, O., Balart, E. F., Cupul-Magaña, A. L., Reyes-Bonilla, H., and Sánchez-Ortiz, C. (2006). Threatened Fishes of the World:

Holacanthus passer Valenciennes, 1846 (Pomacanthidae). *Environ Biol Fish* 77, 97–99. doi:10.1007/s10641-006-9047-y.

Savolainen, V., Anstett, M.-C., Lexer, C., Hutton, I., Clarkson, J. J., Norup, M. V., et al. (2006). Sympatric speciation in palms on an oceanic island. *Nature* 441, 210–213. doi:10.1038/nature04566.

Sbrocco, E. J., and Barber, P. H. (2013). MARSPEC: ocean climate layers for marine spatial ecology. *Ecology* 94, 979–979. doi:10.1890/12-1358.1.

Selkoe, K., and Toonen, R. (2011). Marine connectivity: a new look at pelagic larval duration and genetic metrics of dispersal. *Mar Ecol Prog Ser* 436, 291–305. doi:10.3354/meps09238.

Siegel, D., Kinlan, B., Gaylord, B., and Gaines, S. (2003). Lagrangian descriptions of marine larval dispersion. *Mar Ecol Prog Ser* 260, 83–96. doi:10.3354/meps260083.

Slatkin, M. (1993). Isolation by distance in equilibrium and non-equlibrium populations. *Evolution* 47, 264–279. doi:10.1111/j.1558-5646.1993.tb01215.x.

Slatkin, M., and Excoffier, L. (2012). Serial Founder Effects During Range Expansion: A Spatial Analog of Genetic Drift. *Genetics* 191, 171–181. doi:10.1534/genetics.112.139022.

Swearer, S. E., Caselle, J. E., Lea, D. W., and Warner, R. R. (1999). Larval retention and recruitment in an island population of a coral-reef fish. *Nature* 402, 799–802. doi:10.1038/45533.

Tajima, F. (1983). Evolutionary relationship of DNA sequences in finite populations. *Genetics*.

Tariel, J., Longo, G. C., and Bernardi, G. (2016). Tempo and mode of speciation in *Holacanthus* angelfishes based on RADseq markers. *Mol Phylogenet Evol* 98, 84–88. doi:10.1016/j.ympev.2016.01.010.

Thia, J. A., McGuigan, K., Liggins, L., Figueira, W. F., Bird, C. E., Mather, A., et al. (2021). Genetic and phenotypic variation exhibit both predictable and stochastic patterns across an intertidal fish metapopulation. *Mol Ecol* 30, 4392–4414. doi:10.1111/mec.15829.

Thresher, R. E., and Brothers, E. B. (1985). Reproductive Ecology and Biogeography of Indo-West Pacific Angelfishes (Pisces: Pomacanthidae). *Evolution* 39, 878. doi:10.2307/2408687.

Tyberghein, L., Verbruggen, H., Pauly, K., Troupin, C., Mineur, F., and Clerck, O. D. (2012). Bio-ORACLE: a global environmental dataset for marine species distribution modelling. *Global Ecol Biogeogr* 21, 272–281. doi:10.1111/j.1466-8238.2011.00656.x.

Vucetich, J. A., and Waite, T. A. (2003). Spatial patterns of demography and genetic processes across the species' range: Null hypotheses for landscape conservation genetics. *Conserv Genet* 4, 639–645. doi:10.1023/a:1025671831349.

Waples, R. S., and Do, C. (2008). LDNE: a program for estimating effective population size from data on linkage disequilibrium. *Mol Ecol Resour* 8, 753–756. doi:10.1111/j.1755-0998.2007.02061.x.

Watterson, G. A. (1975). On the number of segregating sites in genetical models without recombination. *Theor Popul Biol* 7, 256–276. doi:10.1016/0040-5809(75)90020-9.

Weir, B. S., and Cocckerham, C. C. (1984). Estimating F-statistics for the analysis of population structure. *Evolution* 38, 1358–1370.

Whitlock, M. C., and Lotterhos, K. E. (2015). Reliable Detection of Loci Responsible for Local Adaptation: Inference of a Null Model through Trimming the Distribution of FST. *The American Naturalist* 186. doi:10.1086/682949.

Wickham, H. (2011). ggplot2. *WIREs Compuational Statistics* 3. doi:10.1002/wics.147.

Wright, S. (1931). Evolution in Mendelian Populations. *Genetics* 16, 290–290. doi:10.1093/genetics/16.3.290.

Wright, S. (1943). Isolation by distance. *Genetics* 28.

Chapter 3 Incomplete lineage sorting despite hybridization in Holacanthus Angelfishes in the Tropical Eastern Pacific

Abstract

Genetic variation shared between closely related species is caused by introgression after secondary contact and/or retention of ancestral variation because of incomplete lineage sorting (ILS). Both mechanisms produce similar genetic signatures making them hard to differentiate, yet this is fundamental to infer the evolutionary history of parapatric and sympatric species and develop proper conservation and management strategies. Pomacanthid angelfishes have some of the highest reported rates of hybridization in marine fishes. However, whether ancestral variation was due to ILS or hybridization, has not been thoroughly investigated. Here we evaluate the genomic composition of three closely related *Holacanthus* angelfishes across the Tropical Eastern Pacific (TEP): *Holacanthus passer, H. clarionensis,* and *H. limbaughi*. The relatively recent divergence of *Holacanthus* angelfishes from their Atlantic geminate species, following the closure of the Isthmus of Panama (~2.8 to 3.1 Mya), and reports of putative hybrids between *H. passer* and *H. clarionensis* make this an ideal study system to address our questions. PCA and population structure analyses confirm the presence of F1 hybrids of *H. passer- H. clarionensis* individuals. Additionally, evidence of ancestral variation was found in *H. clarionensis* but none in *H. limbaughi.* Equal amounts of ancestral variation of *H. passer* was found among *H. clarionensis* individuals and the lack thereof in F2 or back-cross hybrids suggests that

hybrids are sterile and that introgression is not possible. This indicates ILS as the most likely scenario in this system. Although *H. limbaughi* and *H. clarionensis* are presumed to have diverged around the same time from *H. passer*, *H. limbaughi'*s smaller effective population size may have led to a faster rate of lineage sorting. Our results highlight that differentiating between introgression and incomplete lineage sorting informs our understanding of speciation, and warns against the common assumption of introgression as soon as hybridization is detected.

Keywords: *Incomplete lineage sorting, hybridization, fishes, angelfishes, speciation, reinforcement, effective population size*

Introduction

Mechanisms behind speciation has been one of the most debated topics in evolutionary biology for more than a century (Darwin 1859; Mayr 1942; White 1968; Coyne and Orr 2004). With recent increases in loss of biodiversity, understanding the processes that drive divergence of species remains of particular interest to scientists. Speciation, herein defined as the formation of a new species by the divergence or spliting of one species into two, is driven by natural selection, limited gene flow, and spatial isolation (Rocha and Bowen 2008; Bernardi 2013). Allopatric speciation by eliminating gene flow was historically considered to be the principal mode of speciation (Mayr 1942; Coyne and Orr 2004). However, a growing field of evidence demonstrates that speciation in the presence of gene flow is more common than

initially thought (Nosil 2008). In parapatric and sympatric speciation events, gene flow can continue to occur prior to complete reproductive isolation. In such cases, divergent selection is believed to be the primary mechanism driving divergence between incipient species (Coyne and Orr 2004). However, in other cases, examples thought to be parapatric or sympatric species may have actually occurred as a result of secondary contact following an initial speciation event, where complete reproductive isolation has yet to be achieved (Barton and Hewitt 1985; Harrison and Larson 2014; Bernal *et al.* 2017; Svardal *et al.* 2020). Here, introgression introduces alleles that reduce differentiation between species (Coyne and Orr 2004). Similarly, when rapid speciation occurs, lineages often retain ancestral genetic variation due to incomplete lineage sorting (ILS). Introgression and ILS both produce similar patterns of shared genomic diversity, making it difficult to disentangle the true evolutionary history behind these genetic signatures (Bae *et al.* 2016; Zhou *et al.* 2017; Edelman *et al.* 2019).

ILS is particularly common in species with long generation times and a large effective population size (N_e) and in species that have recently diverged, where selection and recombination have not completely sorted these ancestral alleles (Bae *et al.* 2016; Zhou *et al.* 2017). According to Hudson and Coyne (2002), assuming no gene flow and with genetic drift and mutation the only evolutionary forces as work, ~9-12 N generations (N = historical effective population size of the descendant) are necessary for more than 95% of the nuclear loci of the descendant species to be reciprocally monophyletic. Thus, ILS can cause shared genetic diversity long after

species divergence. One key aspect to differentiate between ILS and introgression is the genetic signatures of geographically close and distant populations of the different species. In ILS, shared ancestral variation is expected to be equally distributed among all individuals regardless of their geographic distribution. In contrast, when introgression occurs, one would expect higher gene flow to occur among neighboring populations than those further apart, resulting in higher intraspecific genetic diversity levels and lower interspecific differentiation (e.g., Muir and Schlötterer 2005). Moreover, hybridization events do not always result in introgression between species. Reproductive isolation may be gradually strengthened in the absence of gene flow as a result from mutation, drift, and the indirect effects of natural selection – a process referred to as reinforcement (Ortiz-Barrientos *et al.* 2004; Hoskin *et al.* 2005). Speciation by reinforcement is driven directly by natural selection acting against maladaptive hybridization (i.e., sterile hybrids) (Dobzhansky 1982; Harrison 1993).

In marine ecosystems, the lack of biogeographic barriers and the extensive dispersal potential of fishes provides an ideal system to understand speciation mechanisms when gene flow is still possible (Rocha and Bowen 2008, Bernardi 2013). In particular, coral reef fishes make up more than one-third of the approximately 15 000 marine species despite coral reefs only covering < 0.1% of the ocean's surface (Sale 2002; Helfman *et al.* 2009). This impressive level of biodiversity is generated despite the paucity of physical barriers which drive allopatric speciation, as in freshwater systems. The Isthmus of Panama, which is estimated to have closed around 3.2 to 2.8 Mya, however, is one of few physical

barriers impeding gene flow between oceans (O'Dea *et al.* 2016). The closure of the Isthmus provides us a key opportunity to calibrate the molecular clock of geminate species to estimate divergence times following (Alva-Campbell *et al.* 2010; Tariel *et al.* 2016).

Holacanthus angelfishes are an example of a genus of geminate species, separated into two main clades of the Atlantic and Pacific Oceans following the closure of the Isthmus. The clades are estimated to have diverged allopatrically approximately 1.7 to 1.4 Mya based on phyogenetic studies incorporating the calibrated molecular clock (Bellwood *et al.* 2004; Alva-Campbell *et al.* 2010; Tariel *et al.* 2016). Within the Tropical Eastern Pacific (TEP), the genus *Holacanthus* is a monophyletic clade comprised of three species: *Holacanthus passer*, *H. clarionensis*, and *H. limbaughi*. *Holacanthus passer* is widely distributed along the TEP coastline, including the southern oceanic islands of Cocos, Malpelo, and Galapagos. Its sister species, *H. clarionensis* and *H. limbaughi*, in contrast are endemic to the Revillagigedo Archipelago and Clipperton Island, respectively (Figure 1). For the most part, there is no overlap in distribution between these three species. However, rare *H. clarionensis* individuals have been found off the southern tip of Baja California (Allen and Robertson 1994; Bonilla 2016) where putative hybrid individuals have been previously reported (Sala *et al.* 1999). Hybridization in the family of angelfishes (Pomacanthidae) has been reported multiple times (Feddern 1968; DiBattista *et al.* 2012; Tea *et al.* 2020), suggesting that introgression between *H. passer and H. clarionensis* is not unlikely. In addition, *H. passer* has also been

observed, though rarely, at Clipperton Atoll (Clua and Planes 2019), however, hybrids between the endemic *H. limbaughi* and *H. passer* have never been reported.

Based on their relatively recent divergence (< ~2 Mya) (Bellwood *et al.* 2004; Alva-Campbell *et al.* 2010; Tariel *et al.* 2016) and evidence of hybridization (Sala *et al.* 1999, study herein), the TEP *Holacanthus* species complex presents a system to differentiate between introgression and incomplete lineage sorting. This study uses genomic data from populations of the three *Holocanthus* species as well as hybrids from the TEP region to further explore these ideas. Overall, we aim to compare patterns of genetic diversity and population admixture between intra- and inter-specific populations to reveal signatures of introgression or ILS in *Holacanthus* from the TEP.

Materials and Methods

Sample collection

Samples of *Holacanthus passer* were collected from eight sites along the Tropical Eastern Pacific coastline, in addition to two individuals from Galapagos, and one vagrant found on Clipperton Island collected by Clua and Planes (2019) (Figure 1). We also observed one vagrant *H. passer* individual in the Revillagigedo Archipelago at Roca Partida (RG, 2017), but that individual was not collected. *Holacanthus clarionensis* samples were collected from the four islands in the Revillagigedo Archipelago: Socorro, San Benedicto, Roca Partida, and Clarion Island. In addition, we collected one individual off the southern tip of Baja California, Mexico, where

vagrants had been previously reported (Allen and Robertson 1994; Bonilla 2016). We combined samples of *Holacanthus limbaughi* collected by Crane *et al* (2018) and those collected by Clua and Planes (2019). Finally, two putative adult *Holacanthus clarionensis-passer* hybrids were collected in Los Cabos, Mexico. Putative hybrids were identified by the conspicuous orange coloration of *H. clarionensis* combined with the distinct long white bar of *H. passer* (Figure 2). Juveniles from both species show different coloration than adults, however juvenile markings are similar between species, making it difficult to detect hybrid juveniles.

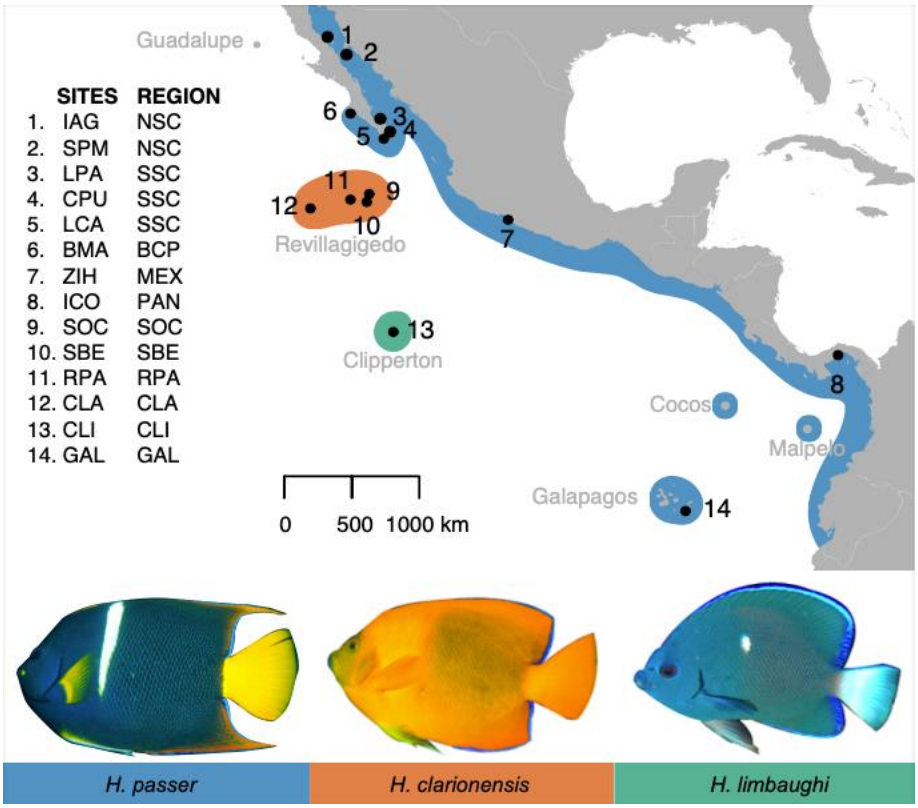


Figure 3.1. Geographic distribution of *Holacanthus passer* (blue), *H. clarionensis* (orange), and *H. limbaughi* (green) showing sampling sites across the Tropical Eastern Pacific. Site and Region ID correspond to numbered sampling sites from the map. IAG, Isla Ángel de la Guarda; SPM, San Pedro Mar; LPA, La Paz; CPU, Cabo Pulmo; LCA, Los Cabos; BMA, Bahía Magdalena; ZIH, Zihuatanejo; ICO, Isla

Contadora; SOC, Socorro Island; SBE, San Benedicto Island; RPA, Roca Partida; CLA, Clarion Island; CLI, Clipperton; GAL- Galapagos; NSC, North Sea of Cortez; SSC, South Sea of Cortez; BCP, Baja California Pacific; MEX, Mainland Mexico; PAN, Panama.

In total, 179 individuals (91 *H. passer*, 40 *H. clarionensis*, 43 *H.limbaughi*, and 3 putative hybrids) were collected across the Tropical Eastern Pacific between 2005 and 2018 (Figure 1; Table S1) with pole spears while on SCUBA or snorkeling and included in this study. Fin or gill tissue were preserved in 95% ethanol immediately after collection and stored at -20°C at the Molecular Ecology and Evolution Lab of the University of California Santa Cruz.

Figure 3.2. Photograph of a putative *Holacanthus clarionensis-passer* hybrid (left) swimming with a *Holacanthus passer* (right) taken off the coast in Los Cabos, Baja California Sur, Mexico. Photo credit: Remy Gatins

RADseq library preparation and sequencing

Genomic DNA was extracted using DNeasy Blood and Tissue kits following manufacturer protocol (Qiagen 2006). Restriction site-associated (RAD) libraries were constructed using a variation of the original protocol (Miller et al. 2007; Baird et al. 2008) with the restriction enzyme SbfI as described in Longo & Bernardi (2015) with additional modifications. Starting genomic DNA was standardized to 100ng per sample and sheared using a Covaris S2 sonicator with an intensity of 5, duty cycle of 10%, cycles/burst of 200, and a cycle time of 30s. Each pool was amplified using 10 PCR cycles in 50 µl reactions. Ampure XP beads (Agencourt) were used for all size selection and purification steps. Individual samples were ligated with a unique barcode and index combination. Finally, libraries were sequenced together on a single lane of an Illumina HiSeq 4000 (150-bp single end reads) at the Vincent J. Coates Genomics Sequencing laboratory at UC Berkeley.

RADseq data processing and SNP calling

STACKS v.2.5 was used to sort, filter, and demultiplex reads with the `process_radtags` command (Catchen *et al.* 2013; Rochette *et al.* 2019). RADseq loci were then trimmed to 80bp with TRIM GALORE to pool with previous RADseq samples (Tariel *et al.* 2016; Crane *et al.* 2018). Reads from all species were aligned with BWA v 0.7.17 to the *Holacanthus passer* genome, H.passer_genome_1.0 (Gatins *et al. in review*) (Li and Durbin 2009). Each aligned .sam file was converted to .bam and sorted with SAMTOOLS v1.9 (Li *et al.* 2009). Reads for all individuals aligned at a rate greater than 98%. To build the initial loci catalog we used all

reference-aligned samples by running '*gstacks'* on STACKS using default parameters. Iterations of the '*populations'* script of STACKS were carried out using multiple popmap files for downstream analyses. Loci that met the following parameters were kept: i) found in at least 50% of individuals within each population (-r); ii) loci shared among all populations (-p); iii) have a minor allele frequency of 0.05 (--min-maf 0.05); (iv) and we only allowed one SNP per locus (--write-single-snp). A summary of STACKS parameters and the total number of SNPs obtained per run can be found in Table S2.

Genetic diversity

Genomic statistics per population were calculated using 20,281 single nucleotide polymorphisms (SNPs) across all populations that contained at least four individuals (see Table S1). Genetic diversity statistics were calculated by species after grouping all individuals of each species together using 21,020 SNPs (see Table S2 for SNP data). Number of alleles, nucleotide diversity, observed and expected heterozygosity, and inbreeding coefficient, were calculated on GENODIVE v 3.03 (Meirmans 2020). Nucleotide diversity and number of alleles were both obtained from the STACKS '*populations'* output.

Effective population size

We estimated the contemporary effective population size (N_e) for all *H. passer*, *H. clarionensis, and H. limbaughi* individuals using two approaches. We used

NeEstimator v2.1 (Do *et al.* 2014) using the linkage dis-equilibrium (LD) method (Waples and Do 2008) under the random mating model and here we report jackknifed 95% confidence intervals with a critical value of 0.05. Secondly, we estimated N_e by obtaining the value of Tajima's π (Pi) from STACKS. When in neutral equilibrium π is correlated with N_e and mutation rates ($\pi = 4 N_e \mu$) (Watterson 1975; Tajima 1983). Mutation rate (μ) is expressed as mutation rate per site per generation. In fishes, μ has been estimated to be between 10^{-8} to 10^{-9} mutations per site (Brumfield *et al.* 2003; Crane *et al.* 2018), thus we ran two simulations to represent the range of the expected mutation rates. Generation time (g) is defined as the age at which half of the individuals of the population are reproducing and has not been studied in detail for any of our study species. All *Holacanthus* are protogynous hermaphrodites, and in *H. passer* generation time for females has been estimated around three years, while for males it is around six years, after they transition from female to male (Hernández 1998; Arellano-Martínez *et al.* 1999; Sánchez-Alcántara *et al.* 2006). Thus, we estimated the average generation time for all *Holacanthus* sp. to be 5 years.

Principal Components Analysis

We performed a principal components analysis (PCA), which does not rely on population genetic models, to summarize the diversity and variation across all RADseq loci using the R package *adegenet* v2.1.4 (Jombart 2008; Jombart *et al.* 2010). Results were color coded by species or putative hybrids.

2.7 Population genetic statistics and population structure

Pairwise population differentiation F_{ST} (Weir and Cocckerham 1984) was calculated for all populations containing more than four individuals using GENODIVE v 3.03 (Meirmans 2020). In addition, to explore genetic structure across sampling sites and species, a Bayesian clustering analysis was performed across all individuals using 19,471 RADseq loci with the software STRUCTURE v2.3.4 (Pritchard *et al.* 2000; Falush *et al.* 2003, 2007; Hubisz *et al.* 2009). The analyses were run under an admixture model (NOADMIX = 0) with no *a priori* location assumptions (LOCPRIOR = 0), and a burn-in of 10,000 iterations followed by 300,000 MCMC repetitions for each run. A total of five replicates were run for each genetic cluster assumed (K = 1-7). The most likely number of clusters (ΔK) was assessed using the Evanno method (Evanno *et al.* 2005) as implemented with the R package *pophelper* v2.3.1 (Francis 2017). Finally, *pophelper* was then used to summarize and plot results from replicate STRUCTURE runs.

Results

Single nucleotide polymorphisms (SNPs)

A total of 193,471,505 reads of 80 bp each were obtained from 179 individual samples from 14 locations across the Tropical Eastern Pacific (Figure 1). Of the total reads, 94.5% passed the quality requirements and were kept to build the catalog loci. The average depth of coverage per individual ranged between 2.1x and 76.0x, with an average of 19.8x across all samples. When we kept all individuals across all 14 populations, a total of 19,471 loci with at least one SNP were recovered and used to

run STRUCTURE and a PCA. This initial assessment showed genomic evidence of a third putative hybrid that had initially been identified as *H. clarionensis* collected at San Benedicto Island in the Revilllagigedo Archipelago. After removing putative hybrids and populations with less than four individuals, a total of 20,281 loci remained and was used downstream to calculate genetic diversity and population genetics statistics. Finally, in the last dataset we removed putative hybrids and grouped the remainder individuals into three groups that corresponded to the putative species. Here, we obtained 21,020 loci and used this dataset to calculate *Ne* and genetic diversity per species (Table S2).

Genetic diversity

A summary of the principal genetic diversity statistics (mean number of alleles, observed and expected heterozygosity, nucleotide diversity, and inbreeding coefficient) are presented in Table 1. *H. passer* showed the highest values for observed and expected heterozygosity, as well as nucleotide diversity. These values were more than two times greater than values reported for *H. limbaughi.* Interestingly, *H. clarionensis'* genetic diversity values were only slightly lower than values of *H. passer,* despite having such different range distributions.

Table 3.1. Population genomic summary statistics of *Holacanthus* populations based on 20,281 RADseq loci, generated using GENODIVE and *'populations'* from STACKS. Summary statistics per species were carried out using 21,020 loci with all individuals pooled by species.

Species	Region	Region ID	N	N (Stacks)	Na	Ho	He	Pi (Stacks)	Fis
***Holacanthus passer* (HPA)**			**93**	**90.04**	**1.891**	**0.256**	**0.275**	**0.0015**	**0.068**
	North Sea of Cortez	NSC	13	12.64	1.821	0.247	0.274	0.00143	0.016
	South Sea of Cortez	SSC	28	27.35	1.863	0.260	0.270	0.00145	0.038
	Baja California- Pacific	BCP	7	6.81	1.745	0.259	0.266	0.00142	0.026
	Mainland Mexico	MEX	17	16.44	1.840	0.247	0.274	0.00146	0.098
	Clipperton	CLI	1	-	-	-	-	-	-
	Panama	PAN	25	24.26	1.858	0.250	0.273	0.00146	0.084
	Galapagos	GAL	2	-	-	-	-	-	-
***Holacanthus clarionensis* (HCL)**			**40**	**39.56**	**1.91**	**0.235**	**0.244**	**0.00133**	**0.035**
	South Sea of Cortez	SSC	1	-	-	-	-	-	-
	Socorro	SOC	6	5.86	1.652	0.228	0.237	0.00126	0.039
	San Benedicto	SBE	18	17.89	1.840	0.236	0.243	0.00130	0.027
	Roca Partida	RPA	1	-	-	-	-	-	-
	Clarion	CLA	14	13.91	1.793	0.229	0.237	0.00127	0.033
***Holacanthus limbaughi* (HLI)**			**43**	**42.29**	**1.442**	**0.116**	**0.136**	**0.00074**	**0.148**
	Clipperton	CLI	43	42.29	1.440	0.116	0.135	0.00074	0.145
***Holacanthus passer-clarionensis* (Hybrid)**			**3**	-	-	-	-	-	-
	South Sea of Cortez	SSC	3	-	-	-	-	-	-

N, number of individuals; *N (Stacks)*, average number of individuals used across all sampled loci; *Na*, Number of alleles; H_O, observed heterozygosity: H_E, expected heterozygosity; *Pi:* nucleotide diversity; F_{IS}: inbreeding coefficient. Numbers in bold reflect genomic statistics adding all individuals per species.

Effective population size

The effective population size estimates were determined using values of Tajima's π (Pi) and direct values of *Ne* based on the linkage disequilibrium method. Both approaches show a similar trend with *H. passer* having the largest *Ne*, followed by *H. clarionensis*, and *H. limbaughi* with the smallest *Ne.* Based on estimates from NeEstimator, effective population size of *H. passer* was found to be 3.5 times greater than *H. clarionensis* and 12.7 times greater than *H. limbaughi.* In addition, *H. clarionensis* was 3.5 times greater than *H. limbaughi.* Meanwhile, when using estimates of *Ne* based on π, *H. passer* was found to be 1.1 times greater than *H. clarionensis* and both *H. passer* and *H. clarionensis* were approximately twice the size of *H. limbaughi* (Table 2).

Effective population size estimates based on NeEstimator resulted in very small population sizes at the lower range estimate, with 915, 256, and 72 individuals for *H. passer*, *H. clarionensis,* and *H. limbaughi*, respectively (Table 2). However, the 95% confidence interval of the higher end reached infinity in all three species, suggesting that considerably larger effective population sizes are possible.

Table 3.2. Effective population size (*Ne*) per species calculated with NeEstimator and Tajima's π ($\pi = 4 N_e \mu$). Range of Ne from Tajima's π corresponds to a mutation rate, μ, of $10^{-8} – 10^{-9}$.

	NeEstimator	**Ne = π / 4μ**
H. passer	915.2 - infinity	$3.75 \times 10^4 – 37.5 \times 10^4$
H. clarionensis	255.9 - infinity	$3.33 \times 10^4 – 33.3 \times 10^4$
H. limbaughi	72.2 - infinity	1.85×10^4 - 18.5×10^4

Principal components analysis

The first principal components (PC1) accounted for 25.6% of the total genotypic variation and was relatively high compared to PC2, which only accounted for 6.4% (Figure 3). PC1 discriminated between all species well, including putative hybrids. Hybrids clustered almost equidistantly between *H. passer* and *H. clarionensis*, as expected from their orange body and white bar phenotype seen in two of the individuals (Figure 2), suggesting they are F1 hybrids. PC2 further distinguished *H. limbaughi* from the hybrid individuals and *H. clarionensis*, and to a lesser extent between *H. limbaughi* and *H. passer*.

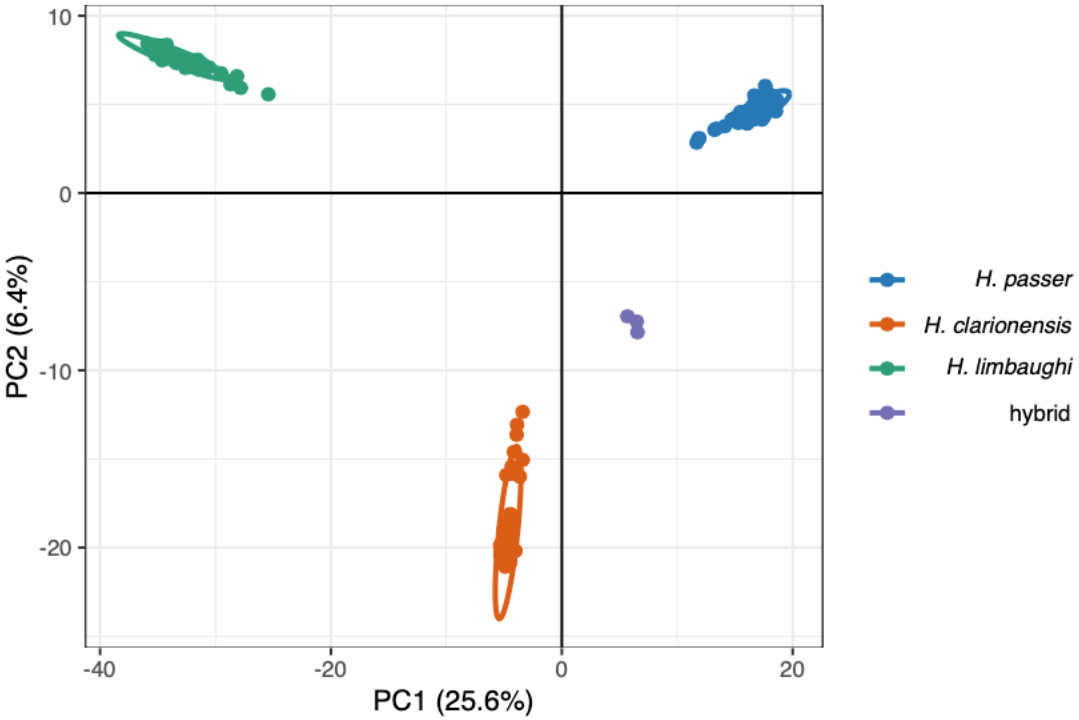


Figure 3.3. Principal components analysis (PCA) of *Holacanthus passer* (blue), *H. clarionensis* (orange), *H. limbaughi* (green), and putative *H. passer* – *H. clarionensis* hybrids (purple) from the Tropical Eastern Pacific using 19,471 RADseq loci. Each

point represents one individual fish. Percent variation explained is indicated in parenthesis for PC1 and PC2.

Population genetics and genetic structure

Pairwise fixation index (F_{ST}), between sampling regions per species showed significant differentiation between species, as expected (Table 3). *H. limbaughi* had the largest F_{ST} with *H. passer* (max F_{ST} =0.496, p-value = 0.00) and *H. clarionensis* (max F_{ST} =0.378, p-value = 0.00), while F_{ST} between *H. passer* and *H. clarionensis* was much lower (max F_{ST} =0.183, p-value = 0.00). Moreover, Clarion Island, the furthest Island of the Revillagigedo Archipelago from mainland, revealed highest differentiation from *H. passer* populations. Interestingly, the only intraspecific populations to reveal low but significant differentiation was between *H. passer* individuals from mainland Mexico and the Southern Sea of Cortez (F_{ST} =0.001, p-value = 0.002).

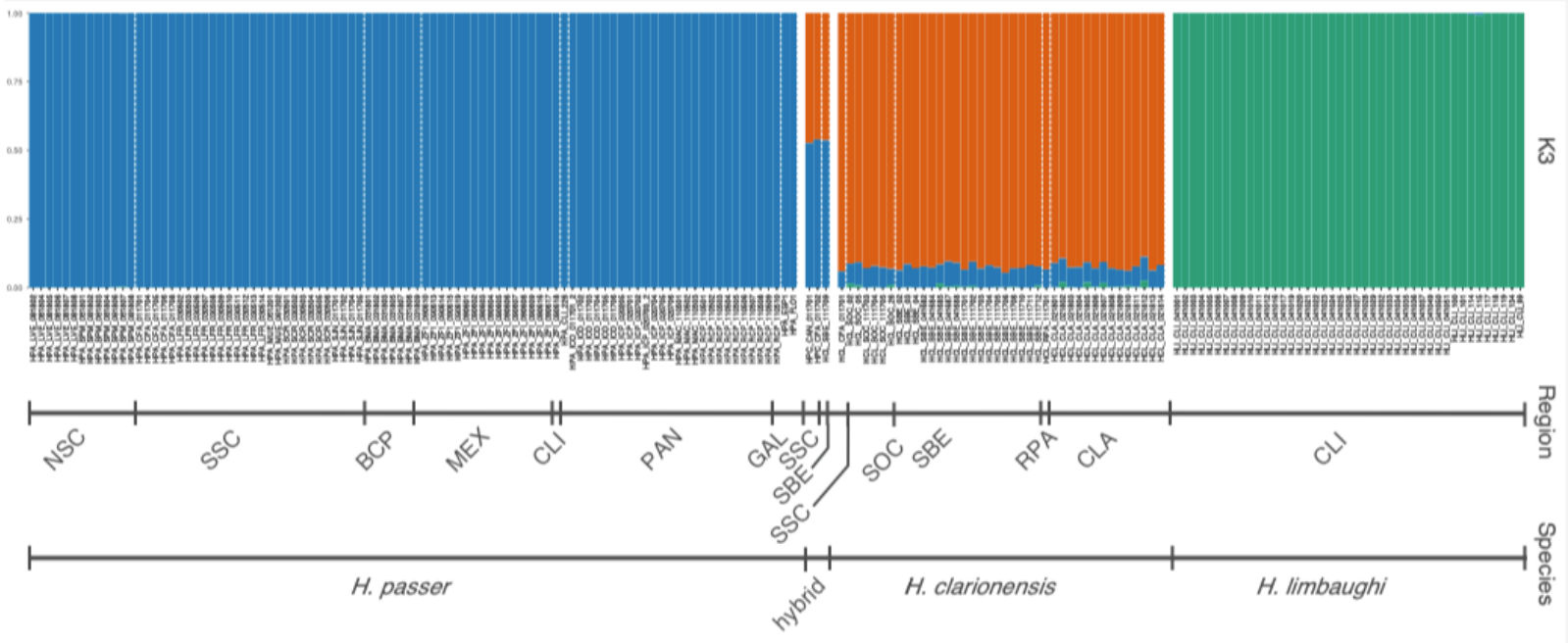


Figure 3.4. Results of Bayesian clustering analysis for K = 3 using 19,471 SNPs. Each bar represents one individual fish and colors in each bar represent estimates of admixture proportion. Individuals are arranged per species and sampling region, separated by white solid bars and dotted lines, respectively. (for sampling region information see Figure 1 and Table S1).

Table 3.3. Pairwise F_{ST} values between sampling regions per species based on 20,248 RADseq loci. F_{ST} values are shown below the diagonal and p-values above the diagonal. Bold values indicate significant differentiation.

Species	Region	NSC	SSC	BCP	MEX	PAN	SOC	SBE	CLA	CLI
HPA	**NSC**		0.826	0.376	0.754	0.223	0	0	0	0
	SSC	0		0.297	0.002	0.093	0	0	0	0
	BCP	0	0		0.765	0.385	0	0	0	0
	MEX	0	**0.001**	-0.001		0.068	0	0	0	0
	PAN	0.001	0.001	0	0.002		0	0	0	0
HCL	**SOC**	**0.172**	**0.17**	**0.178**	**0.172**	**0.171**		0.483	0.592	0
	SBE	**0.171**	**0.17**	**0.175**	**0.174**	**0.172**	0		0.769	0
	CLA	**0.177**	**0.175**	**0.183**	**0.179**	**0.177**	0	0		0
HLI	**CLI**	**0.47**	**0.437**	**0.494**	**0.461**	**0.442**	**0.378**	**0.341**	**0.352**	

HCL: *Holacanthus clarionensis*; HLI: *H. limbaughi*; HPA: *H. passer*; CLA: Clarion Island; SBE: San Benedicto; SOC: Socorro; CLI: Clipperton; NSC: North Sea of Cortez; SSC: South Sea of Cortez; BCP: Baja California Pacific; MEX: Mainland Mexico; PAN: Panama

With the complete dataset of 19,471 SNPs from 179 individuals across 15 sites, the Bayesian clustering analysis of STRUCTURE showed K=2 had the highest likelihood (ΔK = 10,378.22) (Figure S2). However, this division suggests all *H. clarionensis* individuals as being assigned a 50:50 probability of belonging to *H. passer* and *H. limbaughi* (Figure S1). Given what we know about the system, and having sampled across three different species, K=3 reveals more insight into the driver of this genetic signature in *Holacanthus* (Figure 4). Overall, *H. passer* and *H. limbaughi* each belong to one distinct cluster. Interestingly, all *H. clarionensis* individuals show ~10% of shared ancestry with *H. passer,* which could either be a sign of introgression or incomplete lineage sorting. Additionally, the molecular analyses of our putative *H. passer – H. clarionensis* hybrids revealed they were in fact F1 hybrids between *H. passer* and *H. clarionensis.* Two of the hybrids were collected at Los Cabos, Mexico and phenotypically showed signs of hybridization (Figure 2). The F1 hybrid collected on San Benedicto Island of the Revillagigedo Archipelago was not initially cataloged as a putative hybrid based on phenotypic markings, likely because this individual was a juvenile (total length = 13.9 cm), which share similar coloration between *H. passer* and *H. clarionensis*.

Discussion

Understanding the processes that shape genetic diversity and drive speciation are some of the primary objectives of evolutionary biology studies. Here, we shed light on the mechanisms that shaped the divergence of *Holacanthus* angelfishes of the

Tropical Eastern Pacific, and disentangle the genetic signatures of hybridization, introgression, and incomplete lineage sorting in this monophyletic clade. Our genomic analysis identified three F1 hybrids between *H. passer* and *H. clarionensis*, confirming successful hybridization events between both species. In addition, all *H. clarionensis* individuals showed equal amounts of genetic variation from *H. passer*. Yet we detected no genetic evidence of any F2 or back-cross individuals, suggesting that hybrids may be sterile and that introgression is not possible. These genetic signatures are consistent with a scenario of incomplete lineage sorting and speciation by reinforcement. In contrast, *H. limbaughi* showed no evidence of ancestral variation or hybridization events, and F_{ST} differentiation metrics identified this species as being the most genetically differentiated from its sister species complex. *H. limbaughi's* isolated and restricted distribution, as well as its small N_e, may have facilitated a faster divergence time with complete lineage sorting.

Effective population size

In theory, genetic diversity is positively correlated with population size (Kimura 1983; Hague and Routman 2016), and population size is positively correlated with range distribution. The effective population size refers to the number of individuals successfully contributing genes to the next generation, while census population size considers all individuals regardless of reproductive success. In nature, it is common for not all individuals to successfully reproduce, thus N_e is expected to be lower than N (Gasca-Pineda *et al.* 2013; Crane *et al.* 2018). The *Holacanthus* TEP clade have

strikingly different range distributions, ranging from *H. passer* found along more than 7,000 km's of coastline along the mainland, to *H. limbaughi,* the Clipperton Island endemic, spanning less than 14 km's of coastline (Figure 1). Thus we expect *H. passer* to show the highest genetic diversity and largest N_e, while *H. limbaughi* would show the lowest. Therefore, it was to no surprise that N_e of all three species revealed differences of several orders of magnitude. A study by Crane et al (2018) previously estimated the effective population size of *H. limbaughi* using the same two approaches as this study. Using NeEstimator, they found fairly similar results of N_e (Crane et al: 109 – infinity; this study: 72.2 – infinity). However, their estimates using Tajima's π were one order of magnitude smaller (Crane et al: 5.48×10^3 – 54.8×10^3; this study: 1.85×10^4 - 18.5×10^4) which may be attributed to our larger sample size (n = 43 vs 35) and greater number of SNPs (21,021 vs 5,557). In addition, population size (*N*) estimates based on visual counts suggest *H. limbaughi* has approximately 35,000-64,400 individuals (Crane *et al.* 2018), which is within our effective population size range estimate.

Holacanthus clarionensis, is heavily targeted by the aquarium trade due to its bright orange coloration and has been sold for more than $500 an individual (Bonilla 2016). Recent efforts have focused on investigating the status of the Clarion angelfish to develop better management and conservation strategies. The last thorough population size estimate was carried out in 2016 and calculates the total abundance of *H. clarionensis* to be ~60,703 individuals, of which 50,035 are expected to inhabit the Revillagigedo Archipelago and 10,669 the southern tip of Baja California, Mexico

(Bonilla 2016). Our results estimate N_e to have a minimum of 33,300 total individuals, supporting Bonilla's (2016) report (considering $N_e < N$). Interestingly, *H. passer* shows an N_e not much larger than that of *H. clarionensis,* with 37,500 vs 33,300 individuals, respectively. This is surprising given the immense range distribution difference between both species (Figure 1). However, these values of N_e reflect estimates based on genetic diversity calculated from Tajima's π. Shared genetic variation seen between *H. passer* with *H. clarionensis* in our STRUCTURE analyses (Figure 4), may be driving higher nucleotide diversity values seen in *H. clarionensis* (Table 1), which may subsequently be inflating our N_e estimates in this species. In contrast, when we use linkage dis-equilibrium methods applied in NeEstimator, *H. passer* shows an N_e approximately three times greater than values reported for *H. clarionensis* and 12 times greater values for *H. limbaughi* (Table 2), which seems to better align with the size of their distribution ranges (Figure 1).

Evidence of incomplete lineage sorting despite hybridization events

Both the Bayesian clustering analysis and principal components analysis detected three F1 hybrid individuals between *H. passer* and *H. clarionensis* (Figure 3,4). Two of the individuals were found off southern Baja California, Mexico, where both species ranges overlap (Bonilla 2016). Additionally, one hybrid was detected at San Benedicto Island in the Revillagigedo Archipelago. Hybridization between both species had been previously implied after the observation of a hybrid phenotype in the area (Sala *et al.* 1999), but this had yet to be verified genetically. This study

provides the first genetic evidence of hybridization occurring between *H. passer* and *H. clarionensis.* Moreover, *H. passer* vagrants have been observed at Roca Partida in the Revillagigedos (one individual, RG *personal observation)* and one individual collected at Clipperton Island (Clua and Planes 2019), though no *H. passer*- *H. limbaughi* hybrids have been detected. Both observations demonstrate that while rare long-distance dispersal of *H. passer* to the oceanic islands is occuring.

In most cases where hybridization is detected, introgression is often automatically assumed (Montanari *et al.* 2014; Sales *et al.* 2018; Tea *et al.* 2020). However, introgression and incomplete lineage sorting may show similar genetic signatures of shared genetic variation (e.g., in a STRUCTURE analysis), making it hard to distinguish between them (Bae *et al.* 2016; Zhou *et al.* 2017; Edelman *et al.* 2019). The Bayesian clustering analysis showed evidence of shared genetic variation in *H. clarionensis* from *H. passer* (Figure 4). Considering their relatively recent divergence time (< 1.7 Mya) (Bellwood *et al.* 2004; Alva-Campbell *et al.* 2010; Tariel *et al.* 2016), as well as the successful hybridization events herein (Figure 2), either ILS or introgression could be occurring. To differentiate between them, we first looked at whether neighboring inter-specific populations showed lower genetic differentiation and greater genetic diversity than those further apart, as expected under an introgression scenario (Muir and Schlötterer 2005; Petit and Excoffier 2009). The Revillagigedo Archipelago is composed of four islands with increasing distance from the mainland: Socorro, San Benedicto, Roca Partida, and Clarion Island. Socorro and San Benedicto have the closest *H. clarionensis* populations to *H.*

passer's Southern Sea of Cortez (SSC) populations where most hybrids were found, while Clarion Island is the most geographically isolated. Pairwise differentiation between SSC *H. passer* population with *H. clarionensis* populations show a slightly lower F_{ST} than between SSC and Clarion Island (Table 3), as expected. However, genetic diversity of *H. clarionensis* was lower at Socorro Island than at Clarion Island, which is the opposite than expected if introgression were transpiring. Nonetheless, we cannot exclude the possibility that this result may be an artefact of the lower sample size at Socorro Island (n=6).

Secondly, we searched for evidence of F2 hybrids or back-cross individuals using the clustering method STRUCTURE, which would indicate a clear sign of introgression. STRUCTURE assigns a Q-value to define ancestry estimates, where a value of zero and one correspond to 'pure' parent species, and hybrids are represented by intermediate values (e.g., an F1 hybrid would be expected to have a score of 0.5) (Dupuis and Sperling 2016). Assuming K=3, all three of our F1 hybrids had an average inferred ancestry of 53% *H. passer* and 47% *H. clarionensis*. In contrast, all *H. passer* and *H. limbaughi* individuals had > 99.9% inferred ancestry of belonging to their respective species, suggesting true 'pure' individuals. However, *H. clarionensis* individuals had Q-values between 88.8%-94.7% *H. clarionensis* ancestry and 5.3%-9.1% *H. passer* ancestry (Figure 4). In comparison to the other two species, *H. clarionensis* did not have a single 'pure' individual. This result is striking since in the event of introgression we would expect a wider range of Q-values showing both pure and hybrid individuals (F1, F2, and backcrosses). Thus, the fairly equal distribution of

H. passer ancestry present across all *H. clarionensis* individuals, as well as the lack thereof in F2 and backcrosses, strongly suggests a case of incomplete lineage sorting (Hudson and Coyne 2002; Muir and Schlötterer 2005). Moreover, the presence of F1 hybrids and absence of F2 and back-cross individuals may indicate *H. passer*- *H. clarionensis* hybrids are infertile, suggesting speciation by reinforcement. Previous studies have reported hybridization events with infertile offspring in Pomacanthidae and Chaetodontidae (Montanari *et al.* 2014; Tea *et al.* 2020). However, Montanari et al (2014) suggest genetic distance between hybridizing species may influence the fertility of hybrid individuals. Nonetheless, our results show infertile hybrids may occur amongst closely related species (i.e., sister species) as well.

Interestingly, although *H. clarionensis* and *H. limbaughi* are both believed to have diverged approximately 1.7 to 1.4 Mya (Bellwood *et al.* 2004; Alva-Campbell *et al.* 2010; Tariel *et al.* 2016), all *H. limbaughi* individuals show genetic signatures of 'pure' ancestry with no evidence of incomplete lineage sorting. Shared ancestral variation from incomplete lineage sorting assumes a recent speciation event and large effective population sizes (Pamilo and Nei 1988). Phylogenetic studies date the divergence of both *H. clarionensis* and *H. limbaughi* occurred around a relatively similar time frame (Bellwood *et al.* 2004; Alva-Campbell *et al.* 2010; Tariel *et al.* 2016). However, this study estimates *H. clarionensis* has a N_e approximately 2 to 3.5 times greater than *H. limbaughi*, thus these results likely explain the presence and absence of shared variation with *H. passer* seen in both species.

In conclusion, these results highlight that differentiating between introgression and incomplete lineate sorting can better inform us of the mechanisms driving speciation in species, and warns against the common assumption of introgression as soon as hybridization is detected. Moreover, Pomacanthid angelfish have some of the highest reported rates of hybridization in marine fishes (~48%) (Tea *et al.* 2020), thus we recommend caution should be taken to attempt to disentangle the effect of introgression and/or incomplete lineage sorting after speciation events.

Supplementary Materials

Table S1. GPS coordinates and number of samples per sampling sites for Holacanthus passer (HPA), H. clarionensis (HCL), and H. limbaughi (HLI). n = number of individuals.

Region	Site	Site ID	Lat	Long	n HPA	n HCL	n HLI
North Sea of Cortez (NSC)							
	Isla Ángel de la Guarda	IAG	29.5317	-113.5930	6		
	San Pedro Martir	SPM	28.3850	-112.3206	7		
South Sea of Cortez (SSC)							
	La Paz	LPA	24.2043	-110.0745	12		
	Cabo Pulmo	CPU	23.3567	-109.4264	11		
	Los Cabos	LCA	22.9020	-109.8435	5	1	
Baja California- Pacific (BMA)							
	Bahía Magdalena	BMA	24.5437	-112.0584	7		
Mainland Mexico (MEX)							
	Zihuatanejo	ZIH	17.6222	-101.5541	17		
Panama (PAN)							
	Isla Contadora	ICO	8.6346	-79.0423	25		
Revillagigedo (REV)							
	Socorro Island	SOC	18.7633	-110.9119		6	
	San Benedicto	SBE	19.2947	-110.8096		19	
	Roca Partida	RPA	19.1403	-112.2356		1	
	Clarion Island	CLA	18.3420	-114.7086		14	
Clipperton (CLI)							
	Clipperton	CLI	10.3138	-109.2069	1		43
Galapagos (GAL)							
	Galapagos	GAL	-1.3533	-89.7221	2		
Total					**93**	**41**	**43**

Table S2. Summary of parameters used for 'populations' script on STACKS, including total number of single nucleotide polymorphisms (SNPs) obtained per run with their corresponding downstream analyses each output was used for. KEY: n, number of individuals; pop, number of populations individuals were grouped into; -p, minimum number of populations a locus must be present in to process a locus (STACKS); -r, minimum percentage of individuals in a population required to process a locus (STACKS); --min-maf, minimum allele frequency required to process a nucleotide site at a locus; SNPs, single nucleotide polymorphisms detected; Ne, effective population size.

n	pop	-p	-r	--min-maf	SNPs	Downstream Analyses	Notes
171	9	9	0.5	0.05	20,281	Genetic diversity and population genetic statistics by population	Only kept populations with > 4 ind
176	3	3	0.5	0.05	21,020	Genetic diversity and population genetic statistics by species; *Ne*	Removed hybrids and grouped populations per species
179	14	14	0.5	0.05	19,471	STRUCTURE; PCA	Kept all individuals and grouped into 14 separate populations

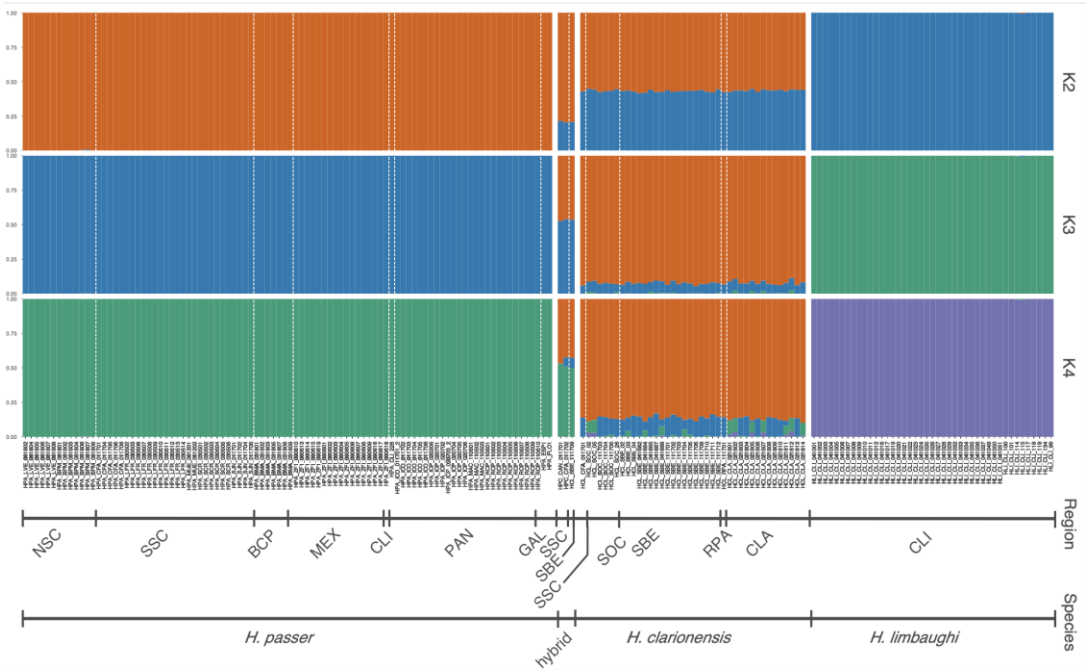


Figure S1. Results of Bayesian clustering analysis for K = 2 to K = 4 using 19,471 SNPs. Each bar represents one individual fish and colors in each bar represent estimates of admixture proportion. Individuals are arranged per species and sampling region, separated by white solid bars and dotted lines, respectively.

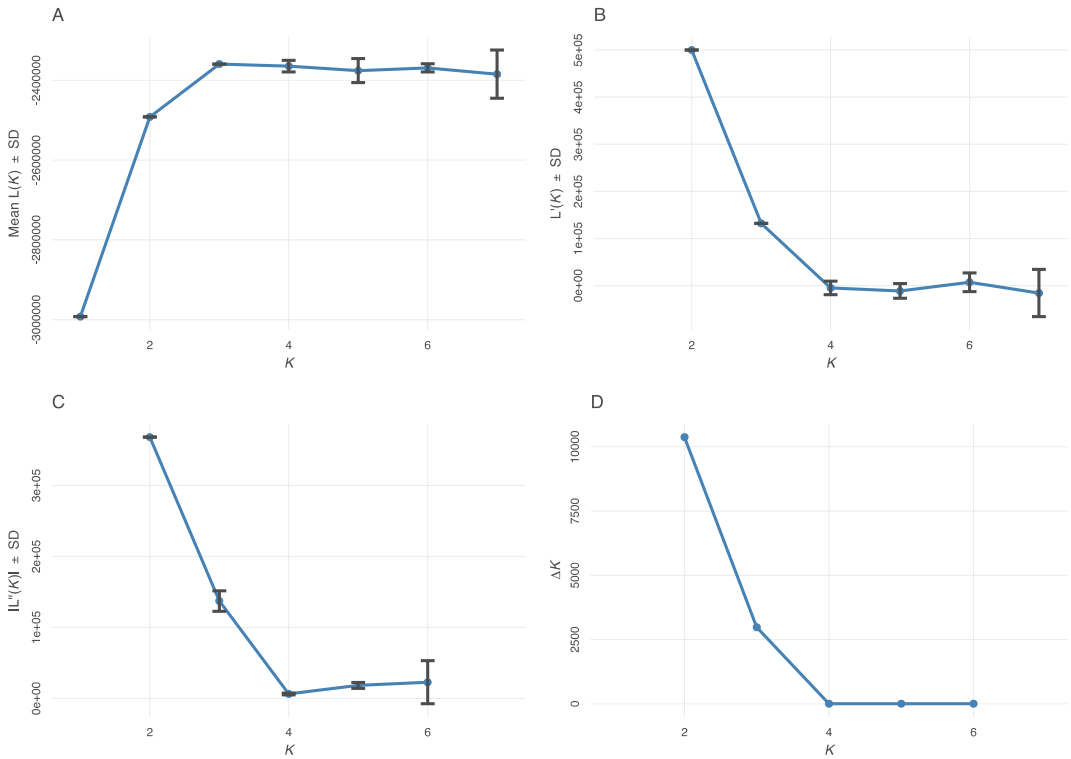


Figure S2. Evanno method plots of 19,471 SNPs showing (A) the estimated log probability over increasing values of K, (B) first derivative, (C) second derivative, and (D) ΔK. The most likely number of clusters based on ΔK shows K = 2.

References

Allen, G., and D. Robertson, 1994 *Fishes of the tropical eastern Pacific* (U. of H. Press, Ed.).

Alva-Campbell, Y., S. R. Floeter, D. R. Robertson, D. R. Bellwood, and G. Bernardi, 2010 Molecular phylogenetics and evolution of Holacanthus angelfishes (Pomacanthidae). Mol Phylogenet Evol 56: 456–461.

Arellano-Martínez, M., B. P. Ceballos-Vázquez, B. P. Ceballos-Vázquez, and F. Galván-Magaña, 1999 Reproductive Biology of the King Angelfish *Holacanthus passer* Valenciennes 1846 in the Gulf of California, Mexico. Bulletin of Marine Science 65: 677–685.

Bae, S. E., J.-K. Kim, and J. H. Kim, 2016 Evidence of incomplete lineage sorting or restricted secondary contact in Lateolabrax japonicus complex (Actinopterygii: Moronidae) based on morphological and molecular traits. Biochem Syst Ecol 66: 98–108.

Barton, N. H., and G. M. Hewitt, 1985 Analysis of Hybrid Zones. Annu Rev Ecol Syst 16: 113–148.

Bellwood, D. R., L. van Herwerden, and N. Konow, 2004 Evolution and biogeography of marine angelfishes (Pisces: Pomacanthidae). Mol Phylogenet Evol 33: 140–155.

Bernal, M. A., M. R. Gaither, W. B. Simison, and L. A. Rocha, 2017 Introgression and selection shaped the evolutionary history of sympatric sister-species of coral reef fishes (genus: *Haemulon*). Mol Ecol 26: 639–652.

Bernardi, G., 2013 Speciation in fishes. Mol Ecol 22: 5487–5502.

Bonilla, H. R., 2016 Situación actual del pez endémico Holacanthus clarionensis (Ángel Clarión), y perspectivas de conservación en México: Ciudad de México Informe final SNIB - CONABIO, Proyecto No. MM003.

Brumfield, R. T., P. Beerli, D. A. Nickerson, and S. V. Edwards, 2003 The utility of single nucleotide polymorphisms in inferences of population history. Trends Ecol Evol 18: 249–256.

Catchen, J., P. A. Hohenlohe, S. Bassham, A. Amores, and W. A. Cresko, 2013 Stacks: an analysis tool set for population genomics. Mol Ecol 22: 3124–3140.

Clua, E., and S. Planes, 2019 First record of Carolines parrotfish (*Calotomus carolinus*) and king angelfish (*Holacanthus passer*) around the Clipperton Atoll-La Passion Island (North-Eastern Tropical Pacific). Ichthyological note.

Coyne, J. A., and H. A. Orr, 2004 *Speciation*. Sinauer Associates, Inc, Sunderland, MA.

Crane, N. L., J. Tariel, J. E. Caselle, A. M. Friedlander, D. R. Robertson *et al.*, 2018 Clipperton Atoll as a model to study small marine populations: Endemism and the genomic consequences of small population size. Plos One 13: e0198901.

Darwin, C., 1859 *On the Origin of Species by Means of Natural Selection or Preservation of Favoured Races in the Struggle for Life*. Murray, London.

DiBattista, J. D., E. Waldrop, B. W. Bowen, J. K. Schultz, M. R. Gaither *et al.*, 2012 Twisted sister species of pygmy angelfishes: discordance between taxonomy, coloration, and phylogenetics. Coral Reefs 31: 839–851.

Do, C., R. S. Waples, D. Peel, G. M. Macbeth, B. J. Tillett *et al.*, 2014 NeEstimator v2: re-implementation of software for the estimation of contemporary effective population size (Ne) from genetic data. Mol Ecol Resour 14: 209–214.

Dobzhansky, T., 1982 *Genetics and the Origin of Species*.

Dupuis, J. R., and F. A. H. Sperling, 2016 Hybrid dynamics in a species group of swallowtail butterflies. J Evolution Biol 29: 1932–1951.

Edelman, N. B., P. B. Frandsen, M. Miyagi, B. Clavijo, J. Davey *et al.*, 2019 Genomic architecture and introgression shape a butterfly radiation. Science 366: 594–599.

Evanno, G., S. Regnaut, and J. Goudet, 2005 Detecting the number of clusters of individuals using the software structure: a simulation study. Mol Ecol 14: 2611–2620.

Falush, D., M. Stephens, and J. K. Pritchard, 2007 Inference of population structure using multilocus genotype data: dominant markers and null alleles. Mol Ecol Notes 7: 574–578.

Falush, D., M. Stephens, and J. K. Pritchard, 2003 Inference of Population Structure Using Multilocus Genotype Data: Linked Loci and Correlated Allele Frequencies. Genetics.

Feddern, H. A., 1968 Hybridization between the Western Atlantic Angelfishes, *Holacanthus isabelita* and *H. ciliaris*. Bulletin of Marine Science.

Francis, R. M., 2017 pophelper: an R package and web app to analyse and visualize population structure. Mol Ecol Resour 17: 27–32.

Gasca-Pineda, J., I. Cassaigne, R. A. Alonso, and L. E. Eguiarte, 2013 Effective Population Size, Genetic Variation, and Their Relevance for Conservation: The Bighorn Sheep in Tiburon Island and Comparisons with Managed Artiodactyls. Plos One 8: e78120.

Gatins, R., C. F. Arias, C. Sánchez, G. Bernardi, and L. F. DeLeón Whole genome assembly and annotation of the King angelfish (*Holacanthus passer*) gives insight into the evolution of marine fishes of the Tropical Eastern Pacific. in review.

Hague, M. T. J., and E. J. Routman, 2016 Does population size affect genetic diversity? A test with sympatric lizard species. Heredity 116: 92–98.

Harrison, R., 1993 *Hybrid zones and the evolutionary process*.

Harrison, R. G., and E. L. Larson, 2014 Hybridization, Introgression, and the Nature of Species Boundaries. J Hered 105: 795–809.

Helfman, G. S., B. B. Collette, C. E. Facey, and B. W. Bowen, 2009 *The Diversity of Fishes: Biology, Evolution, and Ecology*. Wiley-Blackwell.

Hernández, M. C., 1998 Estructura de tallas y crecimiento individual del Ángel Rey, *Holacanthus passer*, Valenciennes 1846 (Teleostei: Pomacanthidae), en la Bahía de La Paz, B.C.S. México.

Hoskin, C. J., M. Higgie, K. R. McDonald, and C. Moritz, 2005 Reinforcement drives rapid allopatric speciation. Nature 437: 1353–1356.

Hubisz, M. J., D. Falush, M. Stephens, and J. K. Pritchard, 2009 Inferring weak population structure with the assistance of sample group information. Mol Ecol Resour 9: 1322–1332.

Hudson, R. R., and J. A. Coyne, 2002 MATHEMATICAL CONSEQUENCES OF THE GENEALOGICAL SPECIES CONCEPT. Evolution 56: 1557–1565.

Jombart, T., 2008 adegenet: a R package for the multivariate analysis of genetic markers. Bioinformatics 24: 1403–1405.

Jombart, T., S. Devillard, and F. Balloux, 2010 Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. Bmc Genet 11: 94.

Kimura, M., 1983 *The neutral theory of molecular evolution*. Cambridge University Press.

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

Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan *et al.*, 2009 The Sequence Alignment/Map format and SAMtools. Bioinformatics 25: 2078–2079.

Mayr, E., 1942 *Systematics and the Origin of Species*. Columbia University Press, New York, NY.

Meirmans, P. G., 2020 genodive version 3.0: Easy-to-use software for the analysis of genetic data of diploids and polyploids. Mol Ecol Resour 20: 1126–1131.

Montanari, S. R., J. A. Hobbs, M. S. Pratchett, L. K. Bay, and L. V. Herwerden, 2014 Does genetic distance between parental species influence outcomes of hybridization among coral reef butterflyfishes? Mol Ecol 23: 2757–2770.

Muir, G., and C. Schlötterer, 2005 Evidence for shared ancestral polymorphism rather than recurrent gene flow at microsatellite loci differentiating two hybridizing oaks (Quercus spp.). Mol Ecol 14: 549–561.

Nosil, P., 2008 Speciation with gene flow could be common. Molecular Ecology 17:.

O'Dea, A., H. A. Lessios, A. G. Coates, R. I. Eytan, S. A. Restrepo-Moreno *et al.*, 2016 Formation of the Isthmus of Panama. Sci Adv 2: e1600883.

Ortiz-Barrientos, D., B. A. Counterman, and M. A. F. Noor, 2004 The Genetics of Speciation by Reinforcement. Plos Biol 2: e416.

Pamilo, P., and M. Nei, 1988 Relationships between gene trees and species trees. Mol Biol Evol 5: 568–83.

Petit, R. J., and L. Excoffier, 2009 Gene flow and species delimitation. Trends Ecol Evol 24: 386–393.

Pritchard, J. K., M. Stephens, and P. Donnelly, 2000 Inference of population structure using multilocus genotype data. Genetics 155: 945–59.

Rocha, L. A., and B. W. Bowen, 2008 Speciation in coral-reef fishes. J Fish Biol 72: 1101–1121.

Rochette, N. C., A. G. Rivera-Colón, and J. M. Catchen, 2019 Stacks 2: Analytical methods for paired-end sequencing improve RADseq-based population genomics. Mol Ecol 28: 4737–4754.

Sala, E., O. Aburto-Oropeza, and J. L. Arreola-Robles, 1999 Observations of a Probable Hybrid Angelfish of the Genus Holacanthus from the Sea of Cortez, México. Pacific Science 53: 181–184.

Sale, P. F., 2002 *Coral Reef Fishes: Dynamics and Diversity in a Complex Ecosystem* (P. F. Sale, Ed.). Elsevier Science.

Sales, N. G., T. C. Pessali, F. R. A. Neto, and D. C. Carvalho, 2018 Introgression from non-native species unveils a hidden threat to the migratory Neotropical fish Prochilodus hartii. Biol Invasions 20: 555–566.

Sánchez-Alcántara, I., O. Aburto-Oropeza, E. F. Balart, A. L. Cupul-Magaña, H. Reyes-Bonilla *et al.*, 2006 Threatened Fishes of the World: *Holacanthus passer* Valenciennes, 1846 (Pomacanthidae). Environ Biol Fish 77: 97–99.

Svardal, H., W. Salzburger, and M. Malinsky, 2020 Genetic Variation and Hybridization in Evolutionary Radiations of Cichlid Fishes. Annu Rev Anim Biosci 9: 1–25.

Tajima, F., 1983 Evolutionary relationship of DNA sequences in finite populations. Genetics.

Tariel, J., G. C. Longo, and G. Bernardi, 2016 Tempo and mode of speciation in *Holacanthus* angelfishes based on RADseq markers. Mol Phylogenet Evol 98: 84–88.

Tea, Y.-K., J.-P. A. Hobbs, F. Vitelli, J. D. DiBattista, S. Y. W. Ho *et al.*, 2020 Angels in disguise: sympatric hybridization in the marine angelfishes is widespread and occurs between deeply divergent lineages. Proc Royal Soc B Biological Sci 287: 20201459.

Waples, R. S., and C. Do, 2008 LDNE: a program for estimating effective population size from data on linkage disequilibrium. Mol Ecol Resour 8: 753–756.

Watterson, G. A., 1975 On the number of segregating sites in genetical models without recombination. Theor Popul Biol 7: 256–276.

Weir, B. S., and C. C. Cocckerham, 1984 Estimating F-statistics for the analysis of population structure. Evolution 38: 1358–1370.

White, M. J. D., 1968 Models of Speciation. Science 159: 1065–1070.

Zhou, Y., L. Duvaux, G. Ren, L. Zhang, O. Savolainen *et al.*, 2017 Importance of incomplete lineage sorting and introgression in the origin of shared genetic variation between two closely related pines with overlapping distributions. Heredity 118: 211–220.