The Dissertation Committee for Moises Antonio Bernal Certifies that this is the approved version of the following dissertation:

Hybridization and divergent selection have shaped the evolutionary history of grunts (genus: *Haemulon*)

Committee:

Deana L. Erdner, Supervisor

Luiz A. Rocha

Andrew J. Esbaugh

Peter Thomas

Mikhail V. Matz

Benjamin Walther

Hybridization and divergent selection have shaped the evolutionary history of grunts (genus: *Haemulon*)

by

Moises Antonio Bernal, B.S.

Dissertation

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

Doctor of Philosophy

The University of Texas at Austin

December 2015

Dedication

I would like to dedicate this dissertation to my parents, Aracelly and Juan Bosco, for their everlasting support.

Acknowledgements

This dissertation would have not been possible without the help of many wonderful people. First I need to thank my parents, Juan Bosco and Aracelly, for their continuous support during these 6 years. They have always encouraged me to follow my passion for biology, and their backing is one of the biggest reasons I even decided to apply to a PhD program. Also thanks to my brothers, sisters, nephews and nieces who have been extremely supportive and attentive during all these years.

Many great colleagues and friends helped with the development of the projects. First I would like to thank my advisor, Luiz Rocha, for always providing the conditions for all these projects to take place, for supplying infinite ideas for research projects, for encouraging participation in multiple ventures and always offer solid career advice during these years. I need to thank the staff of the Center of Comparative Genomics of the CalAcademy, especially Brian Simison and Anna Sellas. Without their help, none of this work would have been possible. They were fundamental in all stages of these projects, always offering help with sequencing and data analysis. Also Michelle Gaither, who was always a source of inspiration and advice, both personal and academic. Her impetus was always contagious, and the result of that are the multiple manuscripts we got published. Thanks to Misha Matz for opening the doors of his lab and encouraging me to delve into RNASeq, because of this a big portion of this dissertation could come to fruition. Special thanks to the RNA wizard Galina Aglyamova, for helping me with the laboratory protocols, Groves Dixon for his help with the transcriptome analyses and Marie Strader for her infinite patience in explaining me the gene expression analyses as well as her disposition to read early versions of some of this work. I am in debt with everyone who helped with collection of specimens: Arturo A. Bocos, Fernando Aranceta,

Rubén Sandoval, Hector Reyes Bonilla, Edgardo Ochoa, Ernesto Peña, Irving Bethancourt, Ricardo Porras, José Smith, Arcadio Castillo and the staff of the Smithsonian Tropical Research Institute.

A big thank you to my labmates, Eva Salas and Hudson Pinheiro, for lively discussions about evolution and ecology of coral reef fishes, and especially all the great times we had in Santa Cruz. My good friend Leo Scanley, for all the conversations we had about science, current events and specially, all the witty banter. Thanks to all the friends from the Bernardi lab, for inviting me to all their talks, camping trips and parties, they really made me feel part of the group. I am grateful for all the help provided by the staff of the California Academy of Sciences for making me feel at home and for the help during outreach activities. Also, thanks to the staff of UTMSI, specially Patty Webb and Jamey Pelfrey, for their assistance from the moment I decided to apply to the PhD. I need to acknowledge my funding sources, most notably the National Secretary of Science and Technology of Panama (SENACYT), which financed my PhD throughout a substantial scholarship. Also, the California Academy of Sciences for providing funds via the Kristina Louie Memorial Fund and the Lakeside Foundation. Thanks to the University of Texas at Austin for providing a grant to help finance a portion of the genetic sequencing.

Finally I would like to thank the members of my PhD committee. I am grateful of all the advice and support provided by Benjamin Walther during this time, without his help it would have been impossible to finish the degree. Also, I would like to thank Deana Erdner, for her willingness to become my official advisor after a series of unforeseen events. Also, thanks to Peter Thomas and Andrew Esbaugh for their time and support with the dissertation and the planning of these projects. To all my friends in Panama, Texas and California, thank you. Gracias, totales!

Hybridization and divergent selection have shaped the evolutionary history of grunts (genus: *Haemulon*)

Moises Antonio Bernal, PhD

The University of Texas at Austin, 2015

Supervisor: Deana L. Erdner

Abstract: Speciation of marine organisms remains a contentious topic in evolutionary biology. Divergence is particularly difficult to explain in cases where multiple closely related species have the same geographical range and occupy similar habitats. Further, dispersal of pelagic larvae can facilitate genetic connectivity across broad distances. These represent sizable challenges for traditional ideas that consider geographic isolation the sole driver of divergence in marine systems. In recent years, concepts of ecological speciation, where disruptive selection can lead to divergence in the absence of reproductive isolation, have regained the attention of evolutionary biologists. This project focuses on understanding the process of divergence in species of the genus *Haemulon*, which is composed of multiple sympatric sister species in the Caribbean and Eastern Pacific. The first portion of the dissertation deals with the case of the sister species *Haemulon flaviguttatum* and *H. maculicauda*, where shared haplotypes and reduced divergence in the mitochondria indicate recent hybridization. However, hybridization led to scant introgression of the nuclear genome, which allows the maintenance of phenotypic differences between the two groups. Further, there is strong

evidence for disruptive selection in coding regions of the nuclear genome. The second portion of this project deals with the Caribbean species *H. carbonarium*, *H. flavolineataum* and *H. macrostomum*. The sequencing, annotation and analysis of the transcriptome revealed strong positive selection in genes related with reproductive isolation, metabolism, stress response and development of pharyngeal structures. Genes associated with the pharyngeal apparatus could be important for the divergence of the group, as these structures are strongly associated with diet. Hence, a hypothesis that divergence of the group is partly due to dietary partitions is proposed. Overall, this dissertation provides novel evidence that the radiation of grunts was directly influenced by divergent selection. By doing so, this project becomes part of a growing body of work that suggests ecological speciation is fundamental for explaining the outstanding diversity that characterizes coral reef systems.

Table of Contents

List of Tables

List of Figures

Introduction

The study of speciation remains one of the most debated topics in evolutionary biology. In the publication that laid out the foundations of evolution, the Origin of Species, Darwin (1859) initially suggested natural selection as the main mechanism of differentiation between lineages. In such cases, reproductive isolation is attained as a consequence of divergent selection between lineages (*reviewed by:* Seehausen *et al*. 2014). These views, however, fell out of favor during the evolutionary synthesis, as models of differentiation based on geographic isolation received more support (*reviewed by:* Bird *et al*. 2012). By automatically leading to reproductive isolation, models of geographic isolation provide a practical explanation for the accumulation of differences between lineages (Coyne & Orr 2004). Most of the work resulting from the evolutionary synthesis, however, was based on studies of terrestrial organisms, and there was a recognized dearth of knowledge on the ecology of marine species (Mayr 1963). Ernst Mayr (1954) was perhaps the first to formally address the question on whether speciation followed the same rules in both marine and terrestrial environments. Based on the observation that sister species of echinoids have non-overlapping geographical ranges, the main conclusion of this study was that allopatric speciation constitutes the exclusive mode of divergence for marine species (Mayr 1954).

One of the fundamental differences between marine and terrestrial ecosystems is dispersal, both potential and realized. In the case of terrestrial species, rivers, valleys, mountains and other geographic features can become important barriers to gene flow. Marine systems, however, are characterized by an apparent paucity of strong physical barriers to account for isolation (Palumbi 1994; Rocha *et al*. 2007). Further,

studies in coral reef fishes have demonstrated that dispersal during pelagic larval stages allow for genetic connectivity across distant biogeographic provinces; multiple species of fishes maintain genetic cohesiveness across the thousands of kilometers of open-ocean that separate the Western from the Eastern Pacific (Lessios & Robertson 2006). Further, some of these barriers are considered permeable, as their influence is not homogenous across all species. For example, the Amazon freshwater outflow separates coral reef fishes from the Caribbean and Brazilian coast (Bernal & Rocha 2011) yet some species are able to traverse it (Rocha *et al*. 2002). Also, a handful of species are known to cross the cold waters of the Benguela current in the southern coast of Africa, maintaining connectivity between the Atlantic and Indian Ocean (Gaither *et al*. 2015b). The cost-reduction of DNA sequencing has also led to an expansion of molecular phylogenies over the past 20 years. These studies have brought to our attention groups where multiple sister species have completely overlapping distributions, including: angelfishes (Pomacanthidae; Hodge *et al*. 2013), grunts (Haemulidae; Rocha *et al*. 2008), surgeonfishes (Acanthuridae; Klanten *et al*. 2004), wrasses (Labridae; Bernardi *et al*. 2004), parrotfishes (Labridae; Choat *et al*. 2012), hamlets (Serranidae; Puebla *et al*. 2007) and damselfishes (Pomacentridae; McCafferty *et al*. 2002). These findings suggest that it is highly unlikely that models of geographic isolation can, by themselves, explain the great diversity associated with coral reefs. Considering the spatial complexity, great availability of resources and heterogeneity of environmental conditions in the marine realm, it is likely that selection plays an important role in the speciation of coral reef fishes (Briggs 2006; Rocha & Bowen 2008; Bowen *et al*. 2013).

In recent years, there has been renewed interest in ecological speciation, mostly fueled by advances and cost reduction of next-generation sequencing techniques (Rocha *et al*. 2013). These studies have provided evidence that reproductive isolation can

be attained if divergent selection in a small number of genes leads to a decrease in fitness of intermediate genotypes (*reviewed by:* Nosil 2012). This process of divergence can occur in the presence of gene flow, thus, it has become an attractive alternative for explaining speciation in the absence of geographic barriers (Bird *et al*. 2012). These ideas have been mostly explored in terrestrial organisms, but well-studied examples also include fishes; in sticklebacks, adaptations to freshwater environment (Hohenlohe *et al*. 2010) and differences in body shape (Berner *et al*. 2008) are known to take place in parapatry. Another example is the recent radiation of cichlids, where divergence is maintained by sexual selection (Allender *et al*. 2003; Terai *et al*. 2006) and dietary partitions (Albertson *et al*. 2003; Manousaki *et al*. 2013), despite rampant hybridization.

A small number of examples of ecological speciation have been suggested in coral reef fishes. For instance, divergence in groups of seahorses could arise by preference of females towards males of similar size, which can lead to the development of extreme phenotypes (Jones *et al*. 2003). Also, significant genetic differences have been detected in the slippery dick wrasse, *Halichoeres bivittatus*, between individuals in different habitats, but within adjacent locations (Rocha *et al*. 2005). Another fascinating case is that of the incipient species of hamlets of the genus *Hypoplectrus*, where hybrid individuals with intermediate color patterns have very reduced fitness when compared to pure lines (Puebla *et al*. 2012a, 2014). In addition, adaptations to cold water and low visibility in the Marquesas could be responsible for the differentiation between the closely related *Acanthurus reversus* and *A. olivaceus* (Gaither *et al*. 2015a).

The radiation of the Neotropical genus *Haemulon* is another instance in which ecological speciation could have played an important role. Initially, it was suggested that divergence of the group was caused by the formation of the Isthmus of

Panama, following the "Law of Geminate Species"[1] (Jordan 1908). However, the molecular phylogeny of the group revealed that most of the sister species have completely overlapping ranges (Rocha *et al*. 2008). Hence, this genus is composed by five pairs of sister species in the Tropical Western Atlantic, and two in the Tropical Eastern Pacific (Rocha *et al*. 2008; Tavera *et al*. 2012). These medium bodied fishes are nocturnal, and despite having sympatric distributions, they are known to have dietary and microhabitat partitions (Randall 1967; Pereira *et al*. 2014). Interestingly, studies in *H. flavolineatum* have demonstrated a lack of population structure throughout the Caribbean (Purcell *et al*. 2006; Puebla *et al*. 2012b). Further, grunt species associated with coral reefs have an accelerated rate of morphological diversification compared with non-reef haemulids, especially in traits associated with feeding (Price *et al*. 2013). However, the study of this group also presents a set of challenges, as very little is known about several fundamental aspects of their life history, including reproductive behavior (Lindeman & Toxey 2002).

This study focuses on revealing divergent selection in the genomes of closely related species of grunts, using massively parallel sequencing. The first chapter of the dissertation focuses on a potential case of hybridization between *H. maculicauda* and *H. flaviguttatum* of the Tropical Eastern Pacific. The second chapter describes the sequencing and annotation of the transcriptome of the sister species *H. carbonarium* and *H. macrostomum*, and the closely related *H. flavolineatum* of the Tropical Western Atlantic. Finally, the third chapter deals with finding signatures of positive selection and differential gene expression in the transcriptome of the aforementioned Caribbean species.

[1] The "Law of Geminate Species" dictates that sister species are always found in adjacent regions separated by strong geographical barriers. This law has been disputed by multiple examples of sister species with completely overlapping distributions, and is no longer considered valid.

Chapter 1: Introgression and selection shaped the evolutionary history of sympatric sister-species of coral reef fishes (genus: *Haemulon*)

ABSTRACT

This study focuses on a case of hybridization between the sympatric sister-species *Haemulon maculicauda* and *H. flaviguttatum*, using Sanger sequencing of mitochondrial and nuclear loci as well as 2422 single nucleotide polymorphisms (SNPs) developed via Restriction-site Associated DNA sequencing (RADSeq). Mitochondrial markers revealed a shared haplotype in COI and reduced divergence in CytB and CR between the species. On the other hand, complete lineage sorting was observed at the nuclear loci and most of the SNPs. Under neutral expectations, the smaller effective population size of mtDNA should lead to fixation of mutations faster than nDNA. Thus, these results suggest hybridization in the recent past (0.174 to 0.263Ma) resulted in complete introgression of mtDNA, but had little effect on the nuclear genome. Further, *H. maculicauda* has higher mtDNA diversity than *H. flaviguttatum*, which suggests that the mtDNA of the former species replaced that of the latter. This was most likely caused by positive selection on the mitochondrial genome. Analyses of the SNPs revealed limited introgression of the nDNA, which explains how the two sympatric species have remained distinct. Moreover, strong divergent selection between the sister-species was observed in 28 loci. This study adds to a growing body of research that exemplifies the fundamental role of selection in the origin and maintenance of biodiversity in marine environments.

INTRODUCTION

Hybridization in animals is thought to have a homogenizing effect, halting differentiation between lineages or even leading to complete genetic swamping (Schierenbeck 2011). Once considered rare in animals, the increased use of molecular tools over the past 20 years indicates that hybridization occurs in at least 10% of animal species (Mallet 2005). High rates of hybridization have been recorded among many groups, including ducks (75%), birds of paradise (40%) and passion-flower butterflies (25%; Mallet 2005). It is also fairly common among fishes (Hubbs 1955), where factors such as external fertilization, weak barriers to pre- and post-zygotic isolation, large and dynamic geographical ranges, habitat degradation and human mediated range shifts increase the chances of interspecific mating (Scribner *et al*. 2000). At least 19 families of freshwater fishes are known to hybridize (*reviewed by:* Scribner *et al*. 2000) and well-studied examples abound in groups such as cyprinids (Broughton *et al*. 2011), centrarchids (Avise & Saunders 1984), cichlids (Joyce *et al*. 2011) and poecilids (Rosenthal & García De León 2011).

Historically, hybridization has been thought to be less frequent among marine fishes when compared to freshwater groups (Hubbs 1955). However, recent studies have found evidence of hybridization and introgression in marine fishes (*reviewed by:* Arnold & Fogarty 2009). In the case of species associated with coral reefs, examples include butterflyfishes (Pyle & Randall 1994; Montanari *et al*. 2012), anemonefishes (Litsios & Salamin 2014; Gainsford *et al*. 2015), grunts (Rocha *et al*. 2008; Bernardi *et al*. 2013), hamlets (Puebla *et al*. 2007), pigmy angelfishes (Schultz *et al*. 2006; Gaither *et al*. 2014), surgeonfishes (DiBattista *et al*. 2011), wrasses (Pyle & Randall 1994; Yaakub *et al*. 2007), and damselfishes (Coleman *et al*. 2014). Further, hotspots of hybridization have

been found in areas of overlap between marine biogeographic provinces. For example, hybridization among 11 species pairs across six families has been recorded at the Christmas–Cocos hybrid zone in the eastern Indian Ocean, an area where fish faunas of the Indian and Pacific Oceans come into contact (Hobbs *et al*. 2008). Similarly, around the island of Socotra, species from the Red Sea and Oman intermix and hybridize with closely related taxa from the western Indian Ocean (DiBattista *et al*. 2015). Taken together, these findings demonstrate that hybridization in marine fishes is more common than once thought and raise important questions about how distinct lineages can remain different in the face of gene flow.

Frequently, hybridization occurs when one species is found in low densities in a hybrid zone. Presumably, the difficulty of finding mates drives members of this species to be less selective and more likely to mate with members of a closely related, but more abundant, taxon ("Hubbs Effect"; *e.g.:* Hobbs *et al*. 2008; Puebla *et al*. 2012a). In some cases, admixed lineages are restricted to narrow hybrid zones and do not spread into parental geographic ranges, making the probability of genetic swamping remote (*e.g.: Chaetodon trifasciatus* x *C. lunulatus* at Christmas Island, Montanari *et al*. 2012). In other instances where hybridization is frequent (*e.g.:* genus *Abudefduf*, Coleman *et al*. 2014) and introgressed lineages are widespread (*e.g.:* genus *Centropyge*, DiBattista *et al*. 2012), molecular studies have revealed differing rates of introgression across loci. Evidence suggests that disruptive selection at even a very small number of genes can lead to reduced gene flow and eventual reproductive isolation (Rundle and Nosil 2005). In the face of hybridization, portions of the genome that are selectively neutral or advantageous can be exchanged between lineages, while those under strong divergent selection do not introgress due to selection against hybrids (Nosil and Schluter 2011). This process of selection can lead to "islands" of genomic differentiation in the recombining nuclear

genome (Wu & Ting 2004; Bird *et al*. 2012). Further, discordant signals of introgression between the mitochondrial and nuclear genomes have been reported in numerous groups, including fishes (Bernatchez *et al*. 1995; Mims *et al*. 2010), amphibians (Zieliński *et al*. 2013), birds (Pons *et al*. 2014; Toews *et al*. 2014) and mammals (Roca *et al*. 2005; Berthier *et al*. 2006). While there are documented examples of widespread mitochondrial introgression in terrestrial animals, few detailed studies have involved coral reef fishes.

Here we focus on a sister-species pair of grunts of the genus *Haemulon* (Family Haemulidae). Grunts inhabit shallow reefs of the Americas, having the highest proportion of sympatric sister-species among coral reef fishes (Hodge *et al*. 2013). The sympatric sister-species *Haemulon maculicauda* and *H. flaviguttatum* co-occur in the Tropical Eastern Pacific, ranging from Baja California to Ecuador (Robertson & Allen 2015; Figure 1). These species are usually found in large heterospecific schools with other grunts. While *H. maculicauda* feeds mainly on benthic crustaceans (Raymundo-Huizar 2000), *H. flaviguttatum* feeds primarily on zooplankton (Flores-Ortega *et al*. 2010). Molecular data indicate that interspecific divergence between these two species is considerably lower in the mitochondrial compared to nuclear genome (Rocha *et al*. 2008; Tavera *et al*. 2012). Because the effective population size of mtDNA is four times lower than nDNA (Tavaré 1984) mitochondrial genomes are expected to reach fixation much faster. Thus, cases where divergence of the mtDNA is lower than the nDNA are consistent with a scenario of introgressive hybridization (Funk & Omland 2003).

Here, mitochondrial and nuclear Sanger sequences, as well as genome-wide single nucleotide polymorphisms (SNPs), were combined to determine: 1) What is the level of genetic divergence between the two sister-species? 2) Is there evidence of introgression in the nuclear DNA? 3) Are there signatures of disruptive selection in the nuclear genome between the two sister-species?

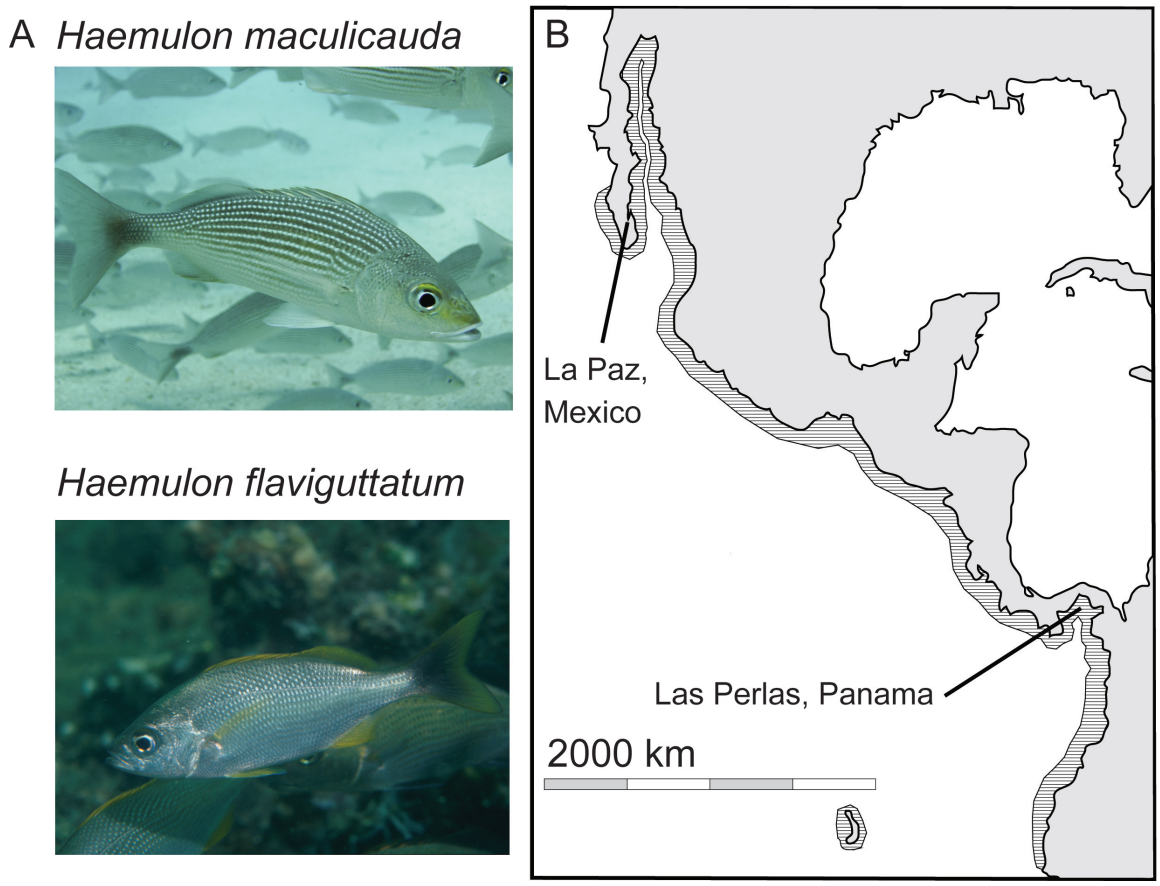


Figure 1. (A) Sister-species *Haemulon maculicauda* and *Haemulon flaviguttatum*. (B) Map of Tropical Eastern Pacific with the geographic range of these sympatric species shaded. Sample locations, La Paz, Mexico and Las Perlas, Panama, are shown. Photo credit: MAB and GR Allen (www.stri.org/sftep).

MATERIALS AND METHODS

Specimen collection and Sanger Sequencing

Samples of *Haemulon maculicauda* and *H. flaviguttatum* were collected from La Paz, Mexico and Las Perlas Archipelago, Panama in the Tropical Eastern Pacific between 2012 and 2013 (Figure 1). Specimens were obtained by SCUBA divers using pole spears

and for each individual gills were preserved in salt-saturated DMSO and stored at -20C at the Ichthyology Collection of the California Academy of Sciences.

DNA was isolated using the DNeasy blood and tissue kits (Qiagen) following the manufacturer's protocol. Three mitochondrial (Cytochrome c oxidase subunit 1, COI; cytochrome b, CytB; and control region, CR) and two nuclear loci (TMO-4C4 and the recombination-activating gene 2, RAG2) were amplified. The primers used for the amplification of mitochondrial markers were: COI, BOL-F1 (5'-TCAAC YAATC AYAAA GATATY GGCAC-3') and BOL-R1 (5'-ACTTC YGGGT GRCCR AARAA TCA-3'; Ward *et al*. 2005); CytB, Cyb7 (5'-AATAG GAAGT ATCAT TCGGG TTTGA TG-3') and Cyb9 (5'-GTGAC TTGAA AAACC ACCGT T-3'; Meyer 1993); CR, CRk (5'-AGCTC AGACT CAGAG CGCCG GTCTT GTAAG C-3') and CRe (5'-cctga agtag gaacc agatg-3'; Lee *et al*. 1995). For amplification of nuclear markers: TMO-4C4-F1 (5'-GAAAA GAGTG TTTGA AAATG A-3') and TMO-4C4-R1 (5'-CATCG TGCTC CTGGG TGACA AAGT-3'; Streelman and Karl 1997; Rocha *et al*. 2008) and RAG2-F1 (5'-GAGGG CCATC TCCTT CTCCA A-3') and RAG2-R3 (5'-GATGG CCTTC CCTCT GTGGG TAC-3'; Lovejoy *et al*. 2001) were used. Polymerase chain reaction for each of the mitochondrial and nuclear markers were conducted in a volume of 15μl, with 10-20ng of DNA, 1.5mM of MgCl2, 1x of buffer, 0.5μM of each primer, 0.3mM of dNTPs, and 0.08 units of PromegaTaq and deionized water to volume. Mitochondrial loci were amplified using the following cycling parameters: initial denaturation at 94°C for 2 min; 32 cycles of 30s at 94°C, 30s at 55°C, and 45s at 72°C; and a final extension at 72°C for 5 min. The nuclear loci were amplified with an initial denaturation at 94°C for 2 min; 10 cycles of 30s at 94°C, 30s at 58°C, and 45s at 72°C; 30 cycles of 30s at 94°C, 30s at 55°C, and 45s at 72°C; and a final extension at 72°C for 5 min. Amplified products were cleaned using 2μl of ExoSAP-IT (Affymetrix) per 5μl of PCR product, with

incubation at 37°C for 30 min, followed by denaturation at 80°C for 15 min. Cycle sequencing reactions were done with the same primers as the PCR, following the STeP cycle (Platt *et al*. 2007). The sequencing reactions were purified via ethanol precipitation. DNA sequencing was performed with fluorescently-labeled dideoxy terminators on an ABI 3130xl Genetic Analyzer (Applied Biosystems) at the Center for Comparative Genomics (CCG) of the California Academy of Sciences. Sequences were trimmed, assembled and aligned using Geneious R7.1.2 (Biomatters).

Analysis of Sanger sequences

Molecular diversity indexes including haplotype (h) and nucleotide diversity (π) were calculated for mitochondrial markers using the program ARLEQUIN 3.5.1.3 (Excoffier & Lischer 2010). An analysis of molecular variance (AMOVA) was performed to estimate divergence between species and populations in ARLEQUIN with 1000 permutations, using the HKY model (Hasegawa *et al*. 1985). The substitution model was chosen as the best fit for all the mitochondrial markers using jModelTest 2.1.4, with the AIC (Darriba *et al*. 2012). Φ_{ST}, an analogue of Wright's F_{ST} for haplotype frequencies, was calculated for each species and between sampling sites. Median-joining haplotype networks were constructed for all markers using the program POPART (http://popart.otago.ac.nz) with the default settings.

Allelic states of nuclear sequences with more than one heterozygous site were estimated using the Bayesian program PHASE V.2.1 (Stephens & Donnelly 2003) as implemented in DNASPv.5.10. Three separate runs, each of 100000 repetitions after a 10000-iteration burn-in, were conducted for each locus. All runs returned consistent allele identities. Observed heterozygosity (H_O) and expected heterozygosity (H_E) were

calculated for both markers. A test of Hardy-Weinberg equilibrium (HWE) using a million steps in a Markov chain was performed using ARLEQUIN. F_{ST} between species and between sample locations for each marker was estimated using ARLEQUIN.

The time to most recent common ancestor (TMRCA) was estimated using the Bayesian MCMC approach as implemented in BEAST 1.7 (Drummond *et al*. 2012). The analysis was conducted using default priors and the HKY mutation model. Mutation rates were estimated using the percent divergence between trans-isthmian populations of the closely related *H. steindachneri*. Rates were calculated by dividing the percent divergence per million years (d), assuming a final closure of the Isthmus of Panama at three million years (Lessios 2008). This value was then divided by two, to obtain the within-lineage rate of divergence ($r= d/2$). Our rates are consistent with observations from other trans-isthmian fish species (Lessios 2008). TMRCA was not calculated for CR, as the mutation rate of this marker is highly variable even between closely related species (Lessios 2008). Despite millions of years of separation, trans-isthmian populations of *Haemulon steindachneri* have no fixed differences in the nuclear markers TMO-4C4 and RAG2. Therefore TMRCA was only calculated for mitochondrial CytB and COI. Pairwise divergence between populations was calculated using Mega 6.06 (Tamura *et al*. 2013), and the Kimura-2-Parameter, with a bootstrap variance estimator of 500 replicates. TMRCA was calculated using three separate chains of 10 million iterations and 10% burn-in, which were combined using the program Tracer v1.4 (Rambaut & Drummond 2007).

RADSeq library preparation

To obtain Single Nucleotide Polymorphisms (SNPs) from the nuclear genome, Restriction Site Associated DNA Sequence (RADSeq) libraries were prepared following

the protocol of Peterson *et al*. (2012), with few modifications. The libraries were prepared for 36 individuals of *H. maculicauda* (Mexico= 18, Panama= 18) and 23 individuals of *H. flaviguttatum* (Mexico= 18, Panama= 5). Following the estimated number of fragments for teleosts in Peterson *et al*. (2012), we selected SphI and MluCl restriction endonucleases (New England Biolabs) to digest DNA for three hours at 37°C. Digested DNA was cleaned using magnetic beads prepared in-house (Rohland & Reich 2012). Samples were divided into four groups of 21 individuals (the library also included 25 individuals of *H. steindachneri*). A P1 adapter with a unique barcode was ligated to each of the 21 individuals. Universal P2 adaptors were added to all samples, after which individuals with the unique barcodes were pooled. In total, four pools of 21 individuals were size selected for fragments between 350-425bp using a Pippin Prep (Sage Science). Following this step, the four pools were amplified using real time PCR, using the Real Time Library Amplification Kit (Kapa Biosystems). During amplification, a unique Illumina index was incorporated into each of the four pools. Finally, the concentration and size of the fragments for each of the pools were quantified using a High Sensitivity DNA Kit on a 2100 Bioanalyzer (Agilent Technologies). Sequencing of the libraries (100bp single end reads) was conducted at the Vincent J. Goats Genomics Sequencing Laboratory at UC Berkeley on an Illumina HiSeq2000.

RADSeq data analysis

Sequences were quality filtered and de-multiplexed according to the unique barcode-index combination of each individual. Only samples with a Phred score higher than 33 were retained. Common loci across individuals were identified by building a de novo map with the software STACKS 0.9 (Catchen *et al*. 2011), which incorporates a maximum likelihood statistical model to call genotypes.

The Populations script of the STACKS suite was used to further filter the RADSeq data. Individuals were divided by species (*H. maculicauda* or *H. flaviguttatum*) and collection locality (Mexico or Panama), giving a total of four groups. Only loci that were present in 70% of the individuals (-r 0.70) of each of the four groups (-p 4), with coverage of 10x or higher (-m 10) were considered in the analyses. With these filtering conditions, a STRUCTURE 2.3.4 (Pritchard *et al*. 2000) input file was exported, using the 'write_single_snp' option. This matrix consisted of 2422 loci, and was converted to various formats using the program PGDSpider 2.0.7.2 (Lischer & Excoffier 2012) for downstream analyses.

A Principal Components Analysis (PCA) of the individual allele frequencies was conducted with Genodive (Meirmans & Van Tienderen 2004). The PCA was calculated using a matrix of covariance, as allele frequencies of all individuals have the same scale (0 to 1). For this analysis, missing data were replaced with the average allele frequencies of its corresponding population. One of the samples of *H. flaviguttatum* from Mexico was not included in the analysis due to a high percentage of missing data. The output of the PCA was graphed with R (R Core Team 2014). An AMOVA using 1000 permutations was performed to test for population structure and estimates of F_{ST} were calculated for divergence between populations and species using Genodive. This program incorporates the Weir & Cockerham (1984) correction for estimation of F_{ST} between populations with different sample sizes

The software STRUCTURE was used to assign individuals to populations. Simulations were done assuming one to six genetic clusters (*K*), with 10 replicates of 500,000 iterations for each *K* value, with 10% burn-in. The most likely number of clusters was assessed implementing the Evanno method with STRUCTURE HARVESTER 0.6.94 (Earl & vonHoldt 2012). The program CLUMPP 1.1.2 (Jakobsson

& Rosenberg 2007) was used to minimize the variance across the 10 replicates of the STRUCTURE runs. The final plot was generated using the software DISTRUCT (Rosenberg 2004).

To test for loci under divergent selection, we used the programs BayeScan 2.1 (Foll & Gaggiotti 2008) and Lositan (Antao *et al*. 2008). BayeScan employs a logistic regression to decompose F_{ST} divergence in a population-specific component and a locus-specific component. A locus is considered under divergent selection if the locus component is necessary to explain the observed diversity (Foll & Gaggiotti 2008). The analysis was conducted using the SNP analysis option, with prior odds of eight and with 5000 iterations. Outlier graphs were constructed using R (R Core Team 2014), following the scripts provided by the developers, with a false discovery rate of 10%. Lositan considers loci to be under either divergent or balancing selection if they demonstrate higher or lower F_{ST} than anticipated based on the expected relationship between F_{ST} and heterozygosity (Antao *et al*. 2008). The analysis consisted of 100000 simulations, and outliers were determined with an estimation of mean neutral F_{ST}. Loci under selection by Lositan also have a false discovery rate of 10%. Only loci identified as outliers by both programs were considered to be under divergent selection. These loci under selection were queried to GenBank using the program BLASTn. Positive identifications were only considered if the sequences had a match of ≥50% and e-values of ≤ e-5.

Bayesian Skyline Plot

Effective population sizes through time were estimated via Bayesian Skyline Plots (BSP), using BEAST 1.7 (Drummond *et al*. 2012). For this analysis, the 'populations' script of the STACKS suite was used to generate a new matrix, in which only loci present in all the individuals of *H. maculicauda* and *H. flaviguttatum* were considered (-r= 1).

The loci with polymorphisms were exported as FASTA files, in which heterozygous individuals are represented by two separate sequences. For these samples, consensus sequences were generated using an in-house script in R, and loci from each individual were concatenated using the program FASconCAT-G (Kück & Longo 2014). This resulted in an alignment of 8835bp for each individual. In order to obtain an estimated mutation rate, the process was repeated for a ddRAD library of the trans-isthmian population of *H. steindachneri*, prepared following the same parameters as described above. The pairwise divergence was estimated using Genodive (d= 0.0003) and the mutation rate was calculated assuming a closure of the Isthmus of Panama at three million years before present (Lessios 2008). This gave a clock rate of 5e-5. Using this rate and a strict clock we constructed a BSP for *H. maculicauda* and *H. flaviguttatum*. Four independent runs of 100 million iterations were run with Beast, combined with logCombiner and viewed using the program Tracer. All the results presented here had estimated sample sizes >3000.

RESULTS

Sanger sequences

For the mitochondrial markers, 563bp of COI were resolved in 59 individuals, 750bp of CytB in 66 individuals, and 395bp of CR in 61 individuals. Molecular diversity indices for these markers are presented in Table 1. Genetic diversity at COI was slightly lower in *H. maculicauda* (h= 0.641, π= 0.0015) compared to *H. flaviguttatum* (h= 0.691, π= 0.0017), but higher at CytB (h= 0.934, π= 0.0052 and h= 0.880, π= 0.0022, respectively) and CR (h= 0.996, π= 0.0348 and h= 0.989, π= 0.0211, respectively). The median joining networks for COI revealed a star pattern, with a common haplotype

shared between species (Figure 2a). The network for CytB revealed two fixed differences between sister-species, with greater genetic divergence among haplotypes of *H. maculicauda* (Figure 2b). For the rapidly mutating CR, almost every individual represented a unique haplotype but haplotypes of the two species are separated by as little as one mutation (Figure 2c). Results of *F*-statistics are presented in the Supplementary Table 1 (Appendix A). Significant genetic structure was detected between *H. maculicauda* and *H. flaviguttatum* with Φ_{ST}= 0.17(p< 0.001) for COI, Φ_{ST}= 0.35(p< 0.001) for CytB and Φ_{ST}= 0.30(p< 0.001) for CR (Table 1). There was also significant population structure between Mexico and Panama in *H. flaviguttatum* at CytB (Φ_{ST}= 0.24, p< 0.001) and CR (Φ_{ST}= 0.258, p< 0.001), but no corresponding structure with COI (Φ_{ST}= 0.19, p= 0.093). On the other hand, there were no significant differences between the two locations for *H. maculicauda*.

For the nuclear markers, 440bp of RAG2 were resolved in 42 individuals and 500bp of the TMO-4C4 in 60 individuals. Molecular diversity indices including number of alleles, observed heterozygosity and expected heterozygosity, are presented in Table 2. Four alleles at each locus were detected in *H. maculicauda* and three in *H. flaviguttatum*, none of which are shared between the two species (Figure 3). Pairwise F_{ST} revealed high levels of significant genetic structure between the two species. For RAG2 F_{ST}= 0.77 (p< 0.001), and divergence was similarly high with F_{ST}= 0.88 (p< 0.001) for TMO-4C4. Significant population structure was observed between the sample locations of Mexico and Panama in both species at RAG2 (*H. maculicauda* F_{ST}= 0.294, p< 0.001; *H. flaviguttatum* F_{ST}= 0.340, p= 0.02). On the other hand, no significant population structure was detected at TMO-4C4 (*H. maculicauda* F_{ST}= 0.294, p= 0.239; *H. flaviguttatum* F_{ST}= 0.107, p= 0.312).

	COI				CytB				CR			
	N	*Nh*	*h*	π	*N*	*Nh*	*h*	π	*N*	*Nh*	*h*	π
Haemulon maculicauda	39	10	0.641 (±0.084)	0.0015 (±0.0012)	39	19	0.934 (±0.023)	0.0052 (±0.0030)	33	31	0.996 (±0.009)	0.0348 (±0.0178)
Mexico	15	6	0.714 (±0.117)	0.0018 (±0.0014)	15	7	0.867 (±0.057)	0.0043 (±0.0026)	13	13	1.000 (±0.030)	0.0379 (±0.0204)
Panama	24	8	0.612 (±0.113)	0.0013 (±0.0011)	24	16	0.946 (±0.029)	0.0055 (±0.0032)	20	18	0.990 (±0.019)	0.0315 (±0.0166)
Haemulon flaviguttatum	20	4	0.690 (±0.060)	0.0017 (±0.0014)	27	13	0.880 (±0.045)	0.0022 (0.0015)	28	25	0.989 (±0.014)	0.0211 (±0.0112)
Mexico	15	4	0.714 (±0.081)	0.0018 (±0.0014)	21	10	0.886 (±0.043)	0.0022 (±0.0015)	22	19	0.983 (±0.021)	0.0203 (±0.0110)
Panama	5	2	0.400 (±0.237)	0.0007 (±0.0009)	6	4	0.800 (±0.172)	0.0013 (±0.0012)	6	6	1.000 (±0.096)	0.0123 (±0.0081)

Table 1 Molecular diversity indices for COI, CytB and CR for *Haemulon maculicauda* and *Haemulon flaviguttatum*. Sample location, number of individuals (*N*), number of haplotypes (*Nh*), haplotype diversity (*h*), and nucleotide diversity (π) are listed for each location and each species.

	TMO					RAG2				
	N	*Na*	H_O	H_E	*P*-value	*N*	*Na*	H_O	H_E	*P*-value
Haemulon maculicauda	33	4	0.36	0.32	1.00	21	4	0.81	0.60	0.22
Mexico	22	4	0.32	0.29	1.00	8	3	0.50	0.43	1.00
Panama	11	3	0.45	0.39	1.00	13	3	1.00	0.59	0.002
Haemulon flaviguttatum	26	3	0.31	0.41	0.47	20	3	0.25	0.22	1.00
Mexico	23	2	0.26	0.35	0.25	16	3	0.13	0.12	1.00
Panama	3	3	0.67	0.73	1.00	4	2	0.75	0.54	1.00

Table 2. Molecular diversity indices for populations of *Haemulon maculicauda* and *Haemulon flaviguttatum*. Sample location, number of individuals (*N*), number of alleles (*Na*), observed heterozygosity (H_O), and expected heterozygosity (H_E) are listed for the nuclear loci. *P*-values are the result of exact tests for Hardy-Weinberg equilibrium using a Markov chain with 100,000 steps in ARLEQUIN 3.5 (Excoffier *et al*. 2005).

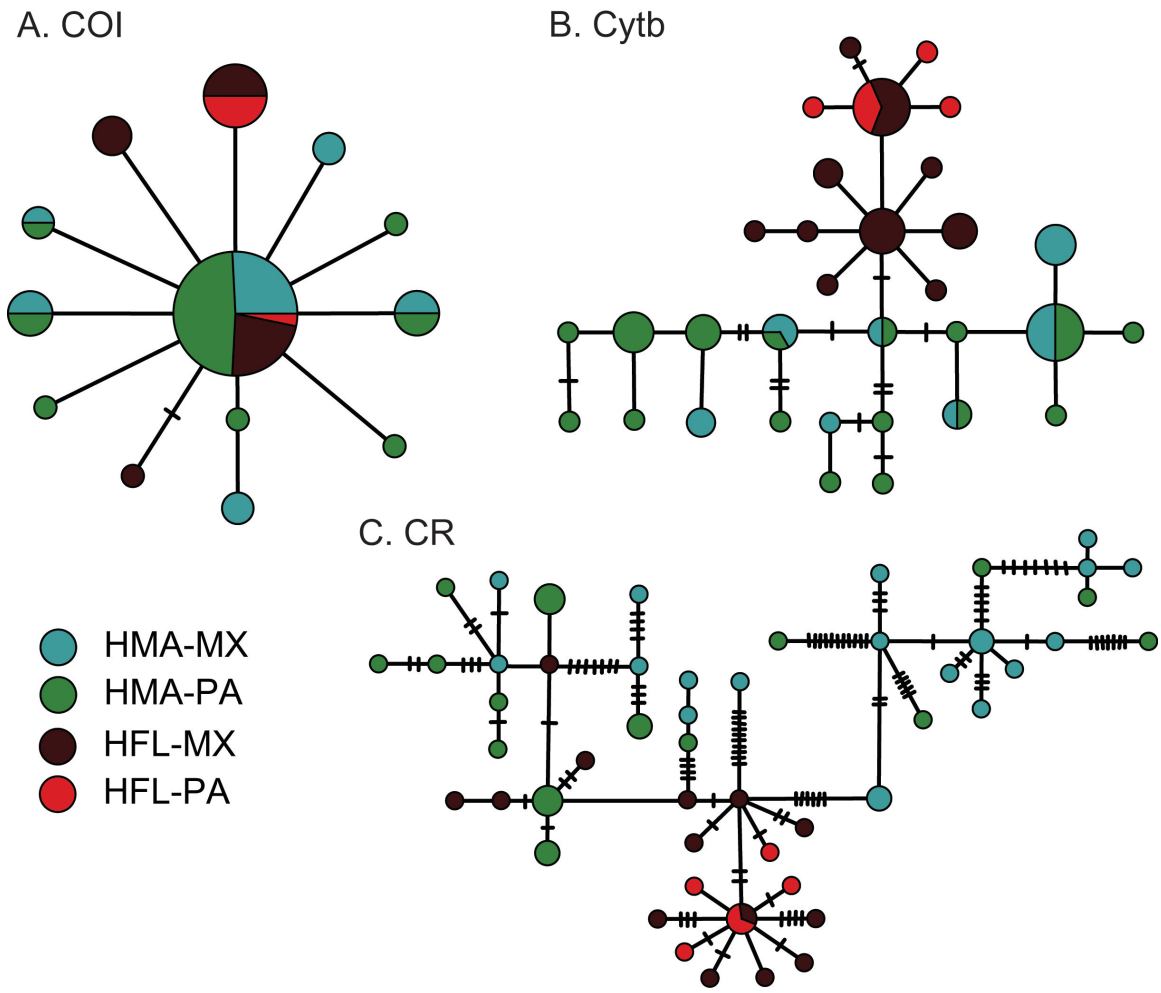


Figure 2. Haplotype networks for mitochondrial markers (A) Cytochrome Oxidase 1, (B) Cytochrome B and (C) Control Region. Each circle represents one haplotype and the size of the circle is proportional to the frequency of that haplotye in the populations. Black bars represent missing haplotypes. Colors represent the species and collection site: HMA-MX is *Haemulon maculicauda*, Mexico; HMA-PA, *Haemulon maculicauda*, Panama; HFL-MX, *Haemulon flaviguttatum*, Mexico; HFL-PA, *Haemulon flaviguttatum*, Panama.

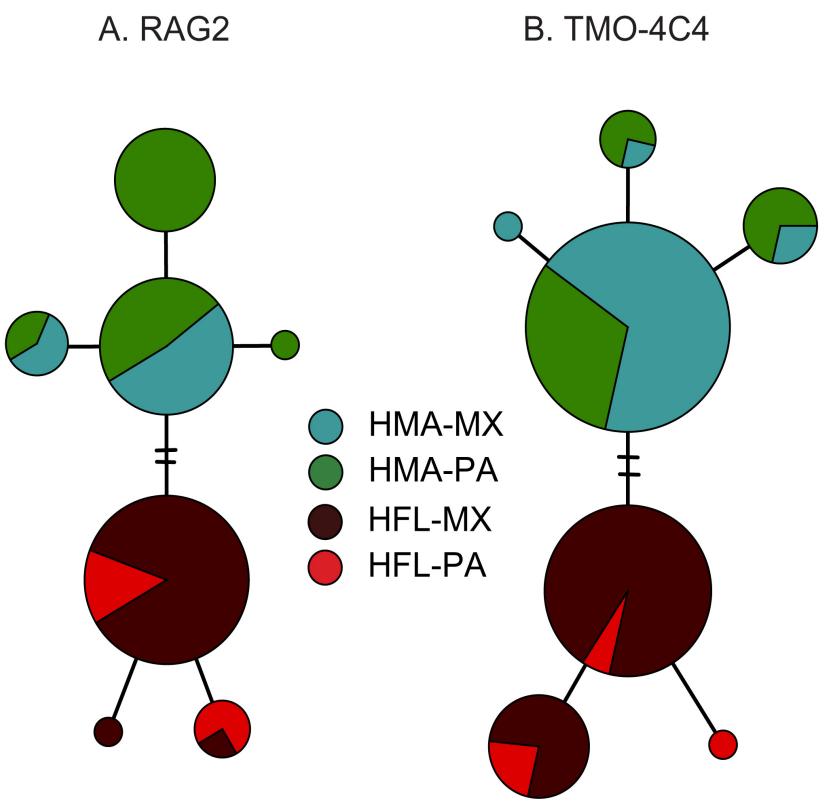


Figure 3. Median-joining networks for nuclear alleles of (A) Recombination-Activating Gene 2 and (B) TMO-4C4. Circles represent one nuclear allele and the area of the circle is proportional to the allele frequency. Lines represent missing haplotypes. Colors represent the species and collection site: HMA-MX is *Haemulon maculicauda*, Mexico; HMA-PA, *Haemulon maculicauda*, Panama; HFL-MX, *Haemulon flaviguttatum*, Mexico; HFL-PA, *Haemulon flaviguttatum*, Panama.

Estimates of TMRCA were 0.174Ma (95% HPD= 0.066-0.309) for COI and 0.263Ma (95% HPD= 0.153-0.388) for CytB (Table 3). For COI, TMRCA within species was similar for *H. maculicauda* (0.160Ma, 95% HPD= 0.056-0.284) and *H. flaviguttatum* (0.161Ma, 95% HPD= 0.055-0.294). CytB, on the other hand, indicated an older coalescence time for *H. maculicauda* (0.255Ma, 95% HPD= 0.142-0.387) than for *H. flaviguttatum* (0.112Ma, 95% HPD= 0.045-0.191).

Marker	*d* of HST	Mutation Rate (per Ma)	TMRCA (95% HPD)	*Haemulon maculicauda*	*Haemulon flaviguttatum*
COI	0.104	1.73%	0.174 (0.066-0.309)	0.160 (0.0559-0.2837)	0.161 (0.0549- 0.2936)
CytB	0.111	1.85%	0.263 (0.153-0.388)	0.255 (0.1423-0.3872)	0.112 (0.0453- 0.1906)

Table 3. Time to most recent common ancestor (TMRCA) in millions of years, with the respective 95% confidence interval (95% HPD) in parenthesis and coalescent times for *Haemulon maculicauda* and *Haemulon flaviguttatum*, for COI and CytB. Results are based on mutation rates calculated using the pairwise divergence (*d*) of Trans-isthmian populations of *Haemulon steindachneri* (HST) and assuming a closure of the Isthmus of Panama of 3Ma.

SNP Data

After filtering and processing the RADSeq dataset, 2422 polymorphic loci were retained. These loci were shared across 18 individuals from Mexico and 18 from Panama of *H. maculicauda*, and 17 individuals from Mexico and four from Panama of *H. flaviguttatum*. One individual of *H. flaviguttatum* from Mexico was eliminated from the

dataset as it had more than 90% missing data. For the remaining individuals, the number of Illumina reads ranged from 324958 to 6761299.

The PCA revealed the major axis of variation (PCA1) is the split between *H. flaviguttatum* and *H. maculicauda* (29.4% of the variation explained, Figure 4). There was also considerable variation between populations of *H. flaviguttatum* from Mexico and Panama (6.47% of the variation, Figure 4). Estimates of pairwise F_{ST} values detected significant differences between the sister-species (F_{ST}= 0.36, p< 0.001; Table 4), and between populations of *H. flaviguttatum* in Mexico and Panama (F_{ST}= 0.163, p= 0.002). These results are consistent with the mitochondrial dataset. The results of the Bayesian clustering analyses of STRUCTURE (Pritchard *et al*. 2000) suggested a *K*= 2 had the highest likelihood (ΔK= 8794.40), corresponding to the differences in allele frequencies between the sister-species (Figure 5a). This *K* value suggests most individuals of *H. flaviguttatum* had less than 1% ancestry corresponding to *H. maculicauda*, with only one individual having ~8% admixture (Figure 5a). The *K*=3 (ΔK=650.07) revealed some differentiation between the two populations of *H. flaviguttatum* and a higher proportion of shared alleles between both species (Figure 5b). With the *K*=3 individuals of *H. flaviguttatum* from Panama had higher percent of shared ancestry with *H. maculicauda*, yet this value was always lower than ~15%.

		1	2	3	4
Haemulon maculicauda	1. Mexico	--	0.025	<0.001	<0.001
	2. Panama	0.004	--	<0.001	<0.001
Haemulon flaviguttatum	3. Mexico	**0.359**	**0.349**	--	0.002
	4. Panama	**0.365**	**0.355**	**0.163**	--

Table 4. Pairwise F_{ST} values between populations of *Haemulon maculicauda* and *Haemulon flaviguttatum* using RADseq (below diagonal). *P*-values are above diagonal, significant comparisons are shown in bold ($p< 0.005$).

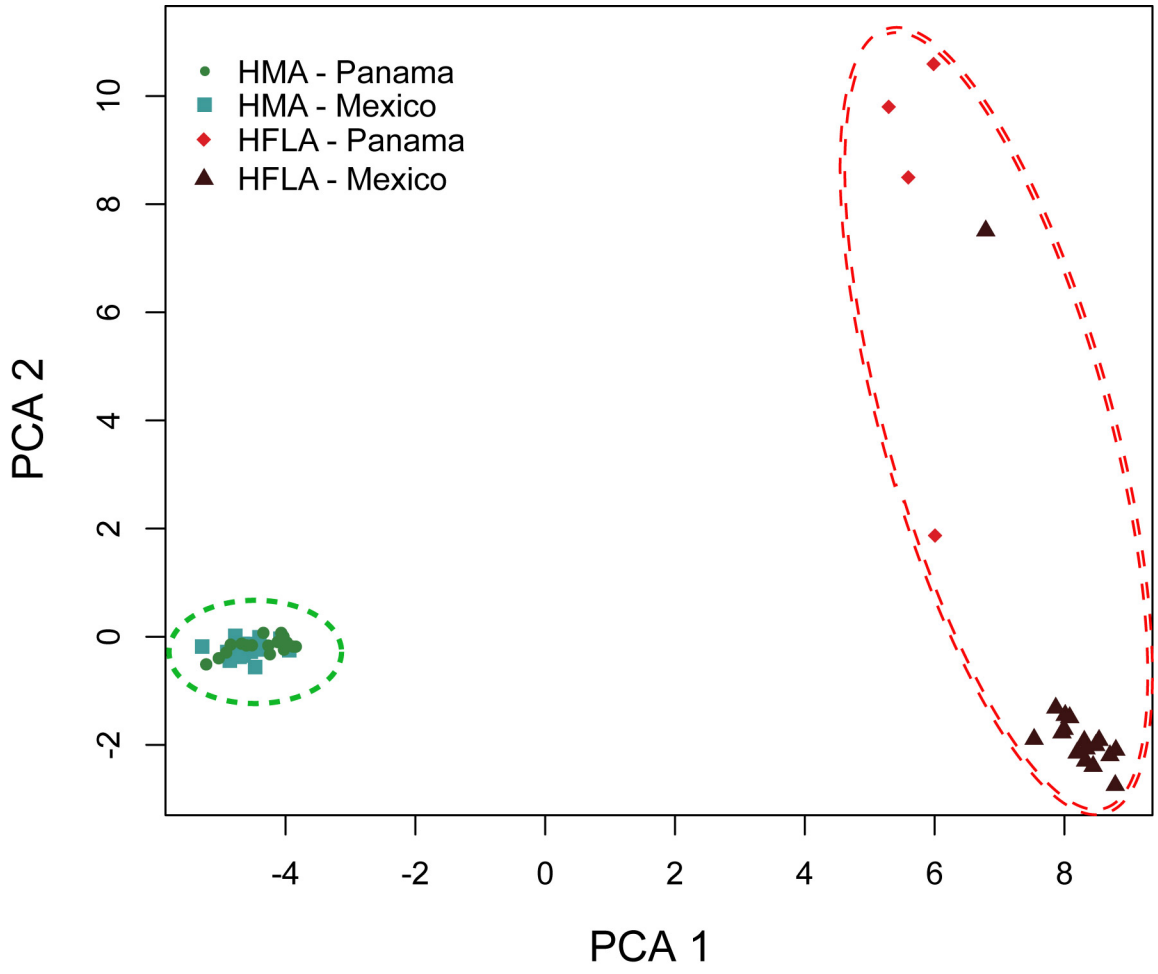


Figure 4. Principal Component Analysis of all loci obtained via RADSeq. The green dotted area represents individuals of *Haemulon maculicauda*, where the circles represent Panama and the squares represent Mexico. The red are in the red dotted line represents *Haemulon flaviguttatum*, where the rhombi represent Panama and the triangles represent Mexico. The largest axis of variation was between the sister-species (PCA 1: 29.4%), followed by differences between populations of *H. flaviguttatum* (PCA 2: 6.47%).

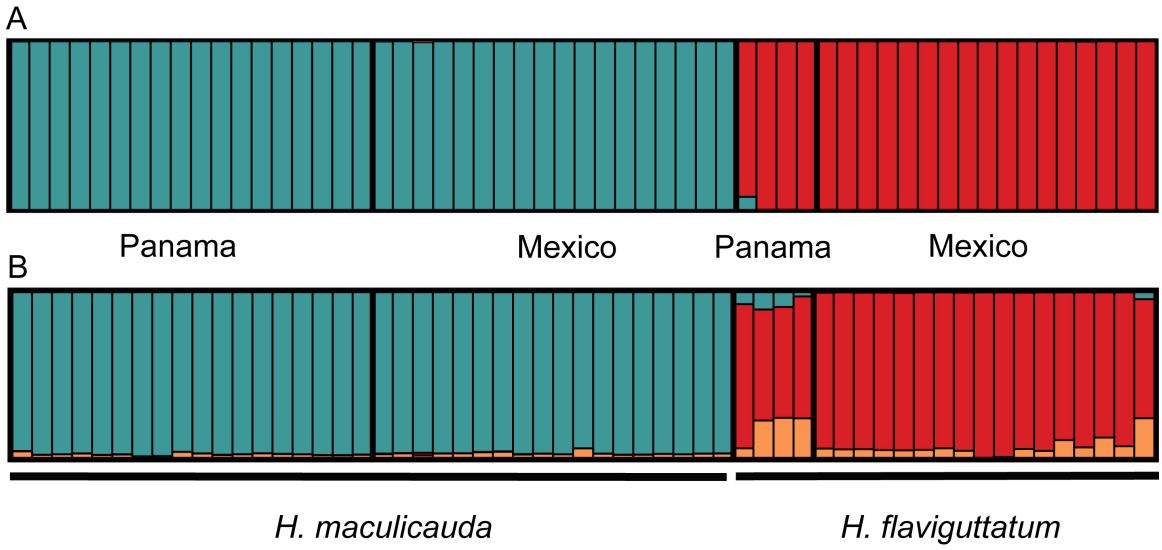


Figure 5. Structure plot for *Haemulon maculicauda* and *Haemulon flaviguttatum*, from populations of Panama and Mexico. (A) The most likely number of partitions was K=2 (ΔK=8794.40), and (B) represents the second most likely partition of K=3 (ΔK=650.07).

Signals of divergent selection between the sister-species were identified with two different programs (Appendix A, Supplementary Figure 1). Using Lositan, 81 loci under disruptive selection were detected, whereas 33 were found using BayeScan. The latter were largely a subset of the loci found using Lositan (85%, or 28 of 33). Of the 28 loci detected using both analyses, 12 had matches higher than e-5 using BLASTn (Table 5).

Bayesian Skyline Plot reconstruction based on SNP loci showed two distinct patterns of effective population size over time. *H. maculicauda* appears to have undergone a significant population expansion during the Pleistocene (0.75 – 0.50Ma), as the effective population size went from 1e1 to 1e2 (Appendix A, Supplementary Figure 2). Effective population sizes of *H. flaviguttatum* show much less change over the past 2

Ma (Appendix A, Supplementary Figure 2). We suggest caution when interpreting the time estimates reported here as these analyses were conducted using the divergence rate estimated using the congeneric *H. steindachneri*.

Loci ID	Blastn Match	% Overlap	E-value
3448	*Erpetoichthys calabaricus* non-LTR retrotransposon Rex-3 pseudogene- partial sequence	82	2.00E-12
7434	*Rhabdosargus sarba* prolactin-releasing peptide mRNA- complete cds	100	5.00E-33
41505	*Takifugu rubripes* F-box/LRR-repeat protein 16-like (LOC101066113)- mRNA	100	8.00E-24
44476	*Oreochromis niloticus* tyrosine-protein kinase BAZ1B-like (LOC100702342)- transcript variant X2- mRNA	85	1.00E-20
47315	*Haplochromis burtoni* serine/threonine-protein phosphatase 2B catalytic subunit alpha isoform-like (LOC102312674)- transcript variant X3- mRNA	79	2.00E-06
58070	*Neolamprologus brichardi* A disintegrin and metalloproteinase with thrombospondin motifs 17-like	90	2.00E-17
70678	*Astyanax mexicanus* calcium channel - voltage-dependent- alpha 2/delta subunit 3 (cacna2d3)- mRNA	90	1.00E-17
79116	*Fugu rubripes* cosmid 151J19 covering the WT1-reticulocalbin and PAX6 genes	98	1.00E-20
89060	*Haplochromis burtoni* src kinase-associated phosphoprotein 1-like (LOC102313695)- transcript variant X3- mRNA	81	2.00E-18
89758	*Labrax* chromosome sequence corresponding to linkage group 18- complete sequence	92	3.00E-29
101771	*Maylandia zebra* nestin-like (LOC101479744)- mRNA	100	3.00E-16
109351	*Dicentrarchus labrax* chromosome sequence corresponding to linkage group 1- bottom part-complete sequence	97	3.00E-36

Table 5. List of 12 loci under selection with their corresponding GenBank annotation, percentage of overlap and E-value. Matches were identified using the BlastN algorithm of NCBI and successful matches have ≥50% overlap and ≤e-5.

DISCUSSION

Molecular studies have revealed that closely related species frequently hybridize in marine ecosystems, especially in areas where biogeographic provinces overlap (Hobbs *et al*. 2008; DiBattista *et al*. 2015). This raises important questions about how divergent lineages maintain their integrity in the face of gene flow. Our analyses of the sympatric sister-species *H. maculicauda* and *H. flaviguttatum* suggest that hybridization occurred as recently as 0.174Ma, but with little effect on the nDNA. Analyses of targeted loci and 2422 SNPs demonstrated significant distinction between the sister species. We hypothesize that the mitochondrial genome of *H. maculicauda* replaced that of *H. flaviguttatum* following a historic hybridization event, or succession of events, having little effect in the nDNA.

Introgression vs. low mutation rates of mtDNA

The mitochondrial genome has an effective population size four times smaller than that of nDNA (Tavaré 1984). As a result, mutations in the mitochondrial genome are expected to reach fixation faster than in the nuclear genome, which results in higher rates of divergence in the former with respect to the latter (Funk & Omland 2003). Further, the lower population size of mtDNA will lead to the loss of ancestral polymorphisms much faster than in nuclear genes, solely by the effect of genetic drift (Bachtrog *et al*. 2006). However, the opposite is observed between *H. maculicauda* and *H. flaviguttatum*. First, the most common COI haplotype was shared between sister-taxa (Figure 2a), and CytB and CR showed fewer differences between the two species than within *H. maculicauda* (Figure 2b-c). Second, there were no shared haplotypes between species at the two nuclear markers, with three fixed mutations separating the most common alleles (Figure

3). The analysis of 2422 SNPs shows a similar pattern, as only one individual of *H. flaviguttatum* had an observable proportion of ancestry matching *H. maculicauda*. Hence the observations of the sister species here studied are not consistent with neutral expectations (Funk & Omland 2003).

In addition to the contrast between mitochondrial and nuclear markers, only one other group shows a similar pattern in the genus *Haemulon*: *H. parra* and *H. bonariense*. These species show small divergence in their mtDNA (~0.5%, Rocha *et al*. 2008). However, they are distantly related according to nuclear markers, as *H.* parra is sister to *H. squamipinna* and *H. bonariense* is more closely related to *H. sexfasciatum* and *H. scudderi* (Rocha *et al*. 2008). This suggests hybridization is also leading to the homogeneity in the mitochondrial genome. All of these remaining species of the genus, which include both sympatric and allopatric species pairs, follow the neutral expectations of high divergence in the mtDNA with respect to nDNA (Rocha *et al*. 2008).

Based on the mitochondrial evidence, hybridization between the two species is not ongoing. Coalescent analyses suggest that the most recent hybridization event was sometime between 160 and 260 thousand years ago (Table 3).

Positive selection facilitates introgression

Given that mtDNA diversity is higher in *H. maculicauda* than in *H. flaviguttatum* (Table 1), it appears that the mitochondrial genome of the former replaced that of the latter, following a historic hybridization event (or succession of events). This pattern of introgression could be explained by assortative mating if females of *H. maculicauda* preferred male *H. flaviguttatum* and their offspring backcrossed with *H. flaviguttatum*. However, informal observations suggest grunts reproduce in large heterospecific spawning aggregations, where large numbers of individuals simultaneously release

gametes into the water column (Domeier & Colin 1997). This makes the hypothesis of assortative mating unlikely.

An alternative explanation is positive selection. Selective sweeps of mtDNA have been observed in cases where the introgressed genomes provide metabolic advantage. For example, in North American warblers, mtDNA from the migratory Myrtle Warbler (*Setophaga coronata*) has introgressed across most of the range of the Audubon's Warbler (*S. auduboni*; Brelsford *et al*. 2011). However, in areas closer to Mexico, individuals of Audubon's Warbler share mtDNA with the Black-fronted Warbler (*S. nigrifrons*), which is a non-migratory lineage. Despite the fact that both lineages of Audubon's warblers are very similar morphologically, only individuals with mtDNA from the Myrtle lineage migrate (Toews *et al*. 2014). Interestingly, individuals with the Myrtle lineage (migratory) have more efficient metabolism, suggesting that introgression of this strain of mtDNA is adaptive in migrating populations (Toews *et al*. 2014). Similar evidence has been reported in hybrid lineages of arctic hares (Melo-Ferreira *et al*. 2014).

In the case of *H. flaviguttatum* and *H. maculicauda*, all the substitutions in COI and CytB were synonymous. Thus, it is still unclear whether introgression of the mtDNA was facilitated by positive selection. Future studies with the complete mitochondrial genome of both species are required to assess the mechanisms of selection.

Low rates of nuclear introgression and evidence of divergent selection

The analysis of 2422 SNP loci suggests introgression affected only a small portion of the nuclear genome. Previous studies have found similar patterns in hybrid zones of terrestrial organisms (*reviewed by:* Toews & Brelsford, 2012). Perhaps the best-studied example is that of *Drosophila yakuba* and *D. santomea*, where hybridization

leads to introgression of the mitochondrial genome of the former into the latter (Llopart *et al*. 2014). Despite hybridization, there is very little admixture in the nuclear genome of the two species (Llopart *et al*. 2005; Bachtrog *et al*. 2006). However, there is rampant introgression of portions of the nuclear genome associated with oxidative phosphorilation, especially genes that conform the Cytochrome C Oxidase (Beck et al 2015). This suggests that the mtDNA is strongly associated with specific regions of nDNA related with cellular respiration, and breaking of these associations can lead to metabolic deficiencies (*reviewed by*: Burton & Barreto 2012). The study here presented shows a similar pattern, suggesting that hybridization lead to introgression of only small portions of the nDNA, which in turn allows the distinction of the sympatric lineages.

Between *H. maculicauda* and *H. flaviguttatum*, 28 nuclear loci under strong disruptive selection were detected. This is a conservative estimate, as only loci found to be under selection by two separate methods were considered. Out of these 28 loci, there are four potentially important genes (Table 5). Locus 7434 mapped to the prolactin-releasing peptide (5e-33). Prolactin has multiple functions related with reproduction (Whittington & Wilson 2013) and osmoregulation (Sakamoto *et al*. 1997). Locus 47315 matched the serine/threonine-protein phosphatase 2B catalytic subunit alpha isoform-like mRNA (2e-06), which has been suggested to play a role in the activation of calcineurin. In teleosts, calcineurin has multiple functions, including formation of skeletal muscle (Martin & Johnston 2005), tissue regeneration (Kujawski *et al*. 2014) and immunological response (Fric *et al*. 2014). Further, locus 79116 matched to the calcium channel-voltage-dependent-alpha 2/delta subunit 3 mRNA, which is involved in the activation and inactivation of P/Q calcium channels. The P/Q type calcium channels are involved with the release of synaptic vesicles in axon terminals (Nimmrich & Gross 2012). Meanwhile, locus 101771 matched nestin-like mRNA. Nestin is key protein in the formation of the

nervous system in fishes (Mahler & Driever 2007). Hence, disruptive selection of these last two genes suggests differences in the transmission of nervous impulses. Despite the fact that the specific role of these genes is not yet known for grunts, their divergence could be associated with differences in the ecology and behavior of the targeted species.

CONCLUSION

Taken together, these results suggest introgression of the mitochondria between the sister species *H. maculicauda* and *H. flaviguttatum*. Since instances of mitochondrial replacement are usually found in narrow hybrid zones, this study is a rare example in which introgression is seen throughout the entire range of sympatric species. The low levels of nuclear introgression can explain, at least in part, how closely related sympatric species of coral reef fishes maintain distinction despite hybridization. Overall this project is part of a growing body of research that aims to explain the paradox of high diversity and few opportunities for isolation in coral reef systems.

Chapter 2. De novo assembly of transcriptomes of three sympatric species of grunts (genus: *Haemulon*) from the tropical western Atlantic

ABSTRACT

The sequencing and annotation of transcriptomes is now feasible for non-model organisms thanks to the advances in massively parallel sequencing. In this study, the transcriptomes of three species of grunts (genus: *Haemulon*) from the tropical western Atlantic were sequenced and annotated. *Haemulon carbonarium*, *H. flavolineatum* and *H. macrostomum* are closely related, have completely overlapping distributions, and are frequently seen together hiding in coral reefs during the day and feeding over seagrass beds at night. Because of the recent evolutionary origin of this clade (~4Ma), sequencing and annotation of the transcriptomes represents a well-suited tool for the study of speciation in this group. Between 38 and 48 million paired-end sequences were retained per-species after removal of adaptors and quality control. Using Trinity, 155245 contigs were assembled for *H. carbonarium*, 109649 for *H. flavolineatum* and 111768 for *H. carbonarium*. The longest contig for each of the three species was approximately 6000bp, and between 25% and 28% of the assembled transcripts were annotated. Contiguity at 75% for the three transcriptomes was between 12% and 16%, and representativity of conserved eukaryotic proteins was at least 90% for the three species. Annotation of transcripts was also done for GO, KOG and KEGG databases for future analyses of selection and gene expression. This represents the most complete genomic resource available for the Haemulidae family, and is one of the few efforts of transcriptome sequencing and annotation in coral reef fishes.

INTRODUCTION

Genetic studies on broadly distributed marine fishes have been mostly focused on "neutral" markers, with the main goal of understanding aspects of their population genetics (Nielsen *et al*. 2009). This has revolutionized our understanding of connectivity of marine populations, which is fundamental to detect barriers for dispersal, as well as for delineating conservation strategies (Rocha *et al*. 2007). However, focusing on "neutral" markers has created a void in our understanding on how marine organisms adapt to different environments, and the role selection plays on the divergence of these groups. Fortunately, the significant cost-reduction on massively parallel sequencing over the past decade is enabling the capture of millions of sequences to tackle questions of adaptation in non-model organisms (Rocha *et al*. 2013).

Transcriptomes are mostly constituted by protein-coding sequences, and not a random subset of genetic information as in loci associated with restriction sites (RADSeq; Wang *et al*. 2009). This makes them well-suited tools for studying the evolution of orthologous genes (*e.g.:* Gerstein *et al*. 2014), and identifying signatures of selection between populations or closely related lineages (Yang *et al*. 2015). Furthermore their sequencing and annotation is more feasible than genomes, as they 10 to 100 times smaller than the latter.

The study of transcriptomes has been traditionally associated with identifying the physiological response of individuals to a particular stressor. However, the possibility of obtaining a large number of orthologous genes from numerous individuals also makes it a convenient tool for investigating speciation (Gayral *et al*. 2011). Studies in fishes have successfully taken advantage of RNASeq to identify traits likely driving divergence between closely related taxa. The most noteworthy examples can be found in cichlids, where genes related with mating coloration have been found in haplochromines (Santos

et al. 2014), and genes that influence feeding morphology (Manousaki *et al*. 2013) and coloration (Henning *et al*. 2013) have been identified in Midas cichlids. Other studies have found signatures of divergent selection in orthologous genes between closely related cichlids, providing support for hypotheses of ecological speciation in this hyper-diverse group (Baldo *et al*. 2011; Elmer *et al*. 2010). Fewer studies, however, have ventured into using transcriptomics for understanding speciation in coral reef fishes.

The genus *Haemulon* is composed of 21 nominal species that inhabit coral reefs of the New World. Due to their high abundance, they are a key component in subsistence fisheries and predator-prey interactions (Lindeman & Toxey 2002). This group offers interesting opportunities for the study of speciation. First, the molecular phylogeny of *Haemulon* shows that most sister species in the genus have completely overlapping distributions (Rocha *et al*. 2008; Tavera *et al*. 2012). This is especially noticeable in the tropical Western Atlantic, which is occupied by 14 species of the genus. Second, it has been suggested that the rate of morphological diversification is much faster in haemulids that inhabit coral reefs, than in non-coral reef groups (Price *et al*. 2013). These findings suggest that the complexity of coral reef environments have accelerated the morphological changes in *Haemulon*, specially regarding feeding morphology (Price *et al*. 2013). Third, the most studied species in the group, *H. flavolineatum*, shows no evidence of deep genetic breaks in the Caribbean, but just modest isolation by distance (Purcell *et al*. 2006; Puebla *et al*. 2012b). Considering their distribution, the disparity of feeding morphology between closely related species, and the low probability of micro-allopatry in the Western Atlantic, this group is an excellent candidate to study questions of ecological speciation in coral reef fishes.

This study focuses on three closely related species with completely overlapping distributions: *H. carbonarium*, *H. flavolineatum* and *H. macrostomum*. The phylogenetic

hypothesis of this group recovers *H. carbonarium* as sister to *H. macrostomum*, and *H. flavolineatum* as sister to this branch (Rocha *et al*. 2008; Tavera *et al*. 2012). Coalescent estimates suggest the three species shared a most recent common ancestor approximately 4Ma, while the split between *H. carbonarium* and *H. macrostomum* occurred approximately 1.5Ma (Tavera *et al*. 2012). These three species inhabit shallow reefs of the Caribbean, but move towards seagrass beds at night to feed on benthic invertebrates (Lindeman & Toxey 2002). As a result, they are considered important species in nutrient cycling between reefs and seagrass beds (Meyer & Schultz 1985).

The main focus of this study is to sequence and annotate the transcriptome of the three sympatric species of grunts *H. carbonarium*, *H. flavolineatum* and *H. macrostomum*. These annotated transcriptomes will be the baseline for analyses of gene expression and divergent selection in the third chapter of the dissertation.

METHODS

Specimen collections:

Specimens of *Haemulon carbonarium*, *H. flavolineatum* and *H. macrostomum* were collected in Bocas del Toro Archipelago in the Caribbean coast of Panama, using pole spears while SCUBA diving, on March of 2012. For each specimen, both liver and muscle were extracted. These tissue samples were stored in RNAlater (Qiagen) in ice for the first 12 hours after collection, then at -20C for two weeks and finally stored at -80C in the Center for Comparative Genomics (CCG) of the California Academy of Sciences (CAS).

Library preparation:

Total RNA was extracted from liver of one individual of each species using the RNAqueous Kit (Life Technologies), following the manufacturer's instructions. Final extractions were re-suspended in 40μl of elution buffer, and treated with RNA free DNAase. The quantity and quality of the extractions was assessed with a Nanodrop spectrophotometer (Thermo Scientific) and via electrophoresis in an agarose gel. Extracted samples of the three species contained between 400 and 600ng/μl of RNA.

The library preparation for the three species was done following the protocol of (Meyer *et al*. 2009) available at: http://www.bio.utexas.edu/research/matz_lab/matzlab/Methods_files), with some modifications for Illumina sequencing, which are detailed bellow. First-strand cDNA synthesis was done with 500ng of RNA, using the SMARTer cDNA Synthesis Kit (Clontech). The cDNA synthesis primer was replaced with the primer T-tr (5'-TCAGACGTGTGCTCTTCCGATCTAA CGGAC TTTTTTTTTTTTTV-3'). The cDNA amplification was done in 12 PCR reactions for each of the individuals, using the cDNA amplification primer 3ILL-tr (5'-AGTTCAGACGTGTGCTCTTCCGATCT-3') for 15 cycles. The 12 PCR reactions were purified and later normalized with 1μl of Duplex-Specific Nuclease enzyme. The cDNA amplification was done in eight separate reactions for 15 cycles, using the amplification primer 3ILL-tr. Amplified and purified cDNA was randomly sheered via sonication for three minutes. Barcoded primers were incorporated during the adaptor ligation. The final sample was purified in an agarose gel, selecting bands between 500 and 800bp. The libraries of the three individuals were sequenced as 100bp paired end reads using an Illumina HiSeq 2000, at the Genomics Sequencing and Analysis Facility at the University of Texas at Austin (GSAF). Due to the overabundance of sequenced adaptors in the first run, a second run was prepared to sequence more of the

middle portions of the transcriptome. This second library was sequenced at the GSAF in one lane of Illumina MySeq.

Data Analysis:

Sequences for the three species were de-multiplexed according to their specific barcodes. The processing of sequences and annotation to difference references was done following the pipeline of the Matz lab, available at: *https://github.com/z0on/annotatingTranscriptomes*. Removal of adaptors and low quality reads was done following the mentioned protocol, and only sequences with a fret score higher than Q33 were retained.

The de novo assembly for each species was performed using Trinity V2.0.2 (Grabherr *et al*. 2011), with default settings. This program partitions the data into multiple independent de Brujin graphs, which allows extracting any splicing isomorfs and paralogous genes (Grabherr *et al*. 2011). The assembly was performed using the sequences from both Illumina runs, and only contigs larger than 200bp were generated. The assembled contigs were screened for contamination by vectors using the NCBI UniVec Database (accessed on August of 2014), using the stand-alone version of BLASTn (BLAST 2.2.28+). Contigs that matched the vector database by an e-value of 1e-10 or lower were removed from the assembly.

The identity of the remaining contigs was assessed with the UniProt protein database (downloaded on August 10, 2014; The UniProt Consortium, 2014), using a stand-alone version of BLASTx (BLAST 2.2.28+). Successful hits were considered to be 1e-5 or lower, and a maximum of five matches was considered for each contig. The program CD-HIT (Fu *et al*. 2012) was used to identify potential isogroups. Transcripts

were considered isomorfs of a particular isogroup if they had more than 99% identity and at least 30% overlap.

The annotation of the clean and filtered transcripts with the UniProt database was done following the pipeline developed by the Matz lab, available at: *https://github.com/z0on/annotatingTranscriptomes*. In order to categorize the three transcriptomes, gene names and Gene Ontology (GO) were extracted from the matches to the UniProt database. The number of unique and shared genes between species was determined. The database REVIGO (Supek *et al*. 2011) was used to determine the number of GO terms in each of the categories: Biological Processes (BP), Cellular Components (CC) and Molecular Functions (MF). After extracting GO annotations and coding sequences, contiguity at a threshold of 75% was calculated. Contiguity represents the percentage of a particular transcript that is covered by one of the denovo assembled contigs (Martin & Wang, 2011). The corresponding graphs of CDS coverage Vs Frequency of Contigs were generated with R (R Core Team 2013). To estimate the percentage of representativity, the presence of 458 proteins corresponding to Core Eukaryotic Genes Dataset (Parra *et al*. 2007) was assessed in the annotated contigs. Annotation of euKaryotic Clusters of Orthologous Groups (KOGs) of NCBI was done with the Web Server of Metagenomic Analysis (WebMGA, Wu *et al*. 2011). For this sequences were converted to the corresponding amino acids as a Protein FASTA file. Finally, to understand the biological pathways associated with sequenced genes, annotation with the Kyoto Encyclopedia of Genes and Genomes (KEGG, Kanehisa *et al*. 2002) was conducted using the KEGG Automatic Annotation Server (KAAS, Moriya *et al*. 2007).

RESULTS

After removal of adaptors, quality control and concatenating the two Illumina libraries for each species, 38137060 paired end reads were recovered for *H. carbonarium*, 38999126 for *H. flaviguttatum*, and 48779031 for *H. macrostomum* (Table 6). The number of contigs assembled with Trinity, once suspected contamination by vectors was removed, are presented for each species in Table 7. *Haemulon carbonarium* had the largest number of contigs, with 155245, the longest contig was 6471bp and the average length of contigs was 425bp. For *H. flavolineatum*, 109649 contigs were assembled, the largest of which was 6613bp, and the average contig size was 461bp. Lastly, *H. macrostomum* is represented by 111768 contigs, the longest of which was 6079bp and the average length was 465bp. For the three species, approximately 38% of the contigs were between 200 and 300bp, while only ~5% were longer than 1000bp (Figure 6). The number of isogroups detected for each species using CD-HIT were 111822 for *H. carbonarium*, 80962 for *H. flavolineatum* and 81132 for *H. macrostomum* (Table 7).

Species	R1-1strun	R2-1st run	R1-2nd run	R2-2nd run	Total
H. flavolineatum	17565501	17355489	2380447	1697689	38999126
H. carbonarium	20834024	12573314	2755724	1973998	38137060
H. macrostomum	21863012	21698343	2997502	2220174	48779031

Table 6. Number of paired-end sequences obtained for each species after the removal of adaptors.

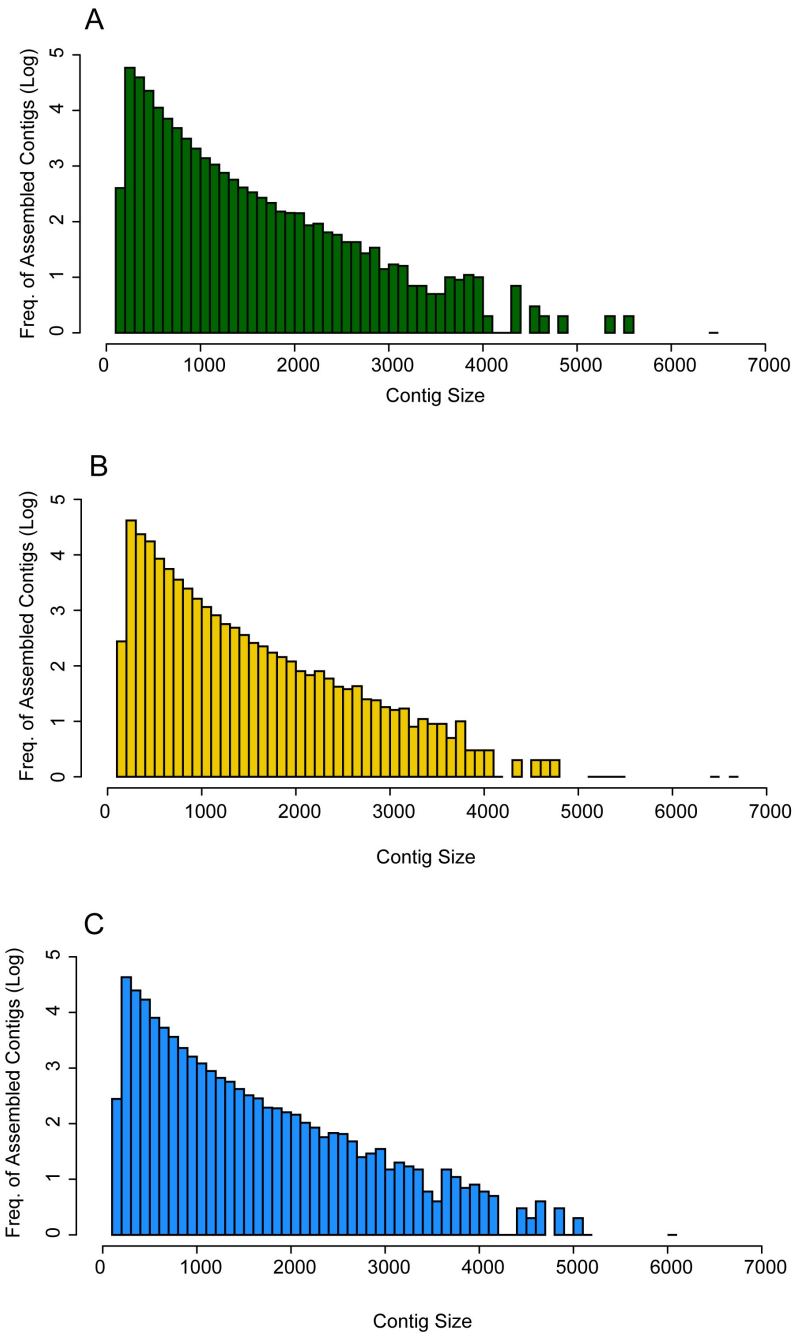


Figure 6. Size distribution of the assembled contigs for the sympatric species *Haemulon carbonarium* (A), *Haemulon flavolineatum* (B) and *Haemulon macrostomum* (C). Due to the overabundance of short transcripts (200-300bp), the Frequency of Assembled Contigs is given in logarithmic scale. The largest contigs for the three species were ~6000bp.

Species	Number of Contigs	Maximum Length	Average	Total Length	N50	Isogroups
Haemulon flavolineatum	109649	6613	438	48026915	461	80963
Haemulon carbonarium	155245	6471	425	66008214	440	111822
Haemulon macrostomum	111768	6079	443	49556767	465	81132

Table 7. Number of contigs for each of the species, maximum length of contigs, average length, total length, N50 and number of isogroups for each species, after removal of potential contaminants. The minimum length of the contigs was set to 200bp for the three species.

Species	Contig size	Total contigs	Percentage of contigs by size	Contigs with BLAST hits	Percentage with hits
H. carbonarium	All	155245	-	40342	25.99%
	200-299	58596	37.74%	11758	20.07%
	>300	96649	62.26%	28584	29.58%
	>1000	6045	3.89%	3508	58.03%
H. flavolineatum	All	109649	-	31348	28.59%
	200-299	41733	38.06%	8979	21.52%
	>300	67916	61.94%	22369	32.94%
	>1000	4889	4.46%	3060	62.59%
	All	111768	-	27532	24.63%
H. macrostomum	200-299	43024	38.49%	7589	17.64%
	>300	68744	61.51%	19943	29.01%
	>1000	5707	5.11%	3541	62.05%

Table 8. Number of BLAST hits per species and per contig size, and percentage of reads with matches for each of these categories.

Using the UniProt database, 40342 contigs of *H. carbonarium* (25.99% of the contigs) were successfully annotated, 31348 were annotated for *H. flavolineatum* (28.59% of contigs) and 27532 for *H. macrostomum* (24.63% of contigs, Table 8). As expected, the probability of annotation is strongly correlated with contig length. Only 17 to 20% of the contigs between 200 and 300bp had successful matches (Table 8). On the other hand, 58% of the reads above 1000bp had significant matches in *H. carbonarium*, 62.59% in *H. flavolineatum* and 62.05% in *H. macrostomum* (Table 8). Contiguity at 75% overlap with reference sequences was 12% for *H. carbonarium*, 13% for *H. flavolineatum* and 16% for *H. macrostomum* (Figure 7). Despite having relatively short contigs, the transcriptomes contain a high percentage of the Core Eukaryotic Genes (Parra *et al*. 2007), as representativity was 94.76% for *H. carbonarium*, 89.92% for *H. flavolineatum*, and 91.94% for *H. macrostomum*.

Based on the annotation to UniProt, there were 904 genes shared between the three species (Figure 8). The sister species *H. carbonarium* and *H. macrostomum* had 1742 genes in common, with 6882 and 6224 unique genes, respectively. The closely related *H. flavolineatum* shared 2818 genes with *H. carbonarium*, and 1711 genes with *H. macrostomum*, having 5534 private genes (Figure 8).

The number of isogroups annotated with GO, KOG and KEGG categories is presented in Table 9. With respect to GO categories, the 15 most common GO terms were the same in the three transcriptomes. Eight of these correspond to the category Cellular Components (Cytosol, Membrane, Integral Component of Membrane, Extracellular Exosome, Cytoplasm, Plasma Membrane, Nucleus and Mitochondrion), five to Molecular Function (RNA Binding, Zinc Ion Binding, ATP Binding, DNA Binding and Metal Ion Binding) and two belong to BP (Transcription and Regulation of

Transcription,). The most common GO term in the three species was GO:0005634, Cellular Component involved in the formation of cell nucleus. For the three species, the category of Biological Processes had approximately three times more GO terms than Cellular Component and Molecular Function (Figure 9).

Species	Gene	GO terms	KOG	KEGG Pathways
H. flavolineatum	22243	21635	22140	14337
H. carbonarium	28025	27248	27969	19176
H. macrostomum	19467	18909	19132	13208

Table 9. Number of isogroups with gene annotation, GO annotation, KOG annotation and matching KEGG Pathways in three sympatric species of grunts.

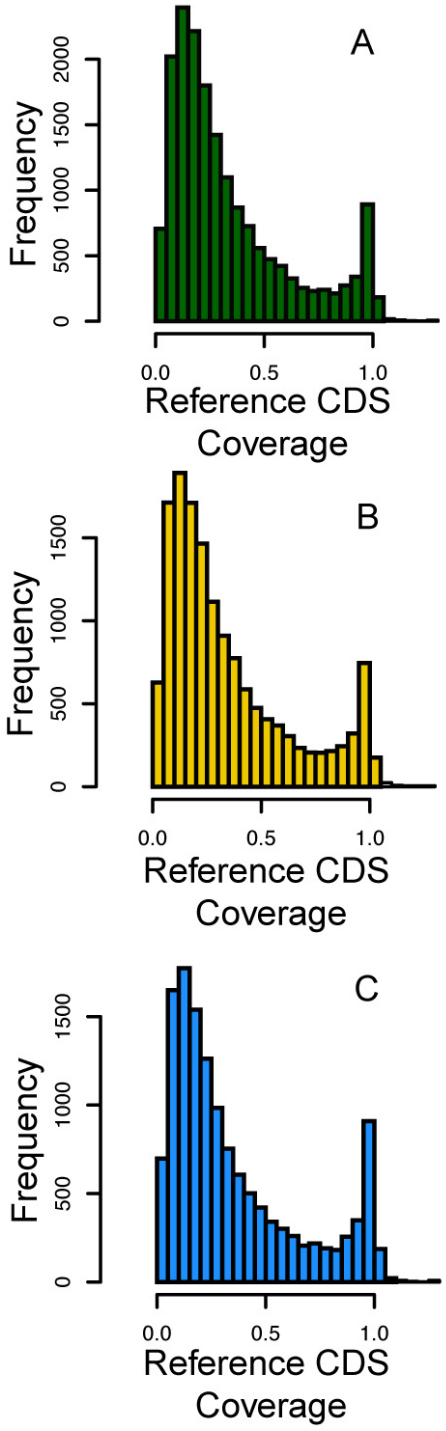


Figure 7. Contiguity at 75% for *Haemulon carbonarium* (A), *Haemulon flavolineatum* (B) and *Haemulon macrostomum* (C). Most of the assembled contigs covered less than 50% than the reference coding sequences.

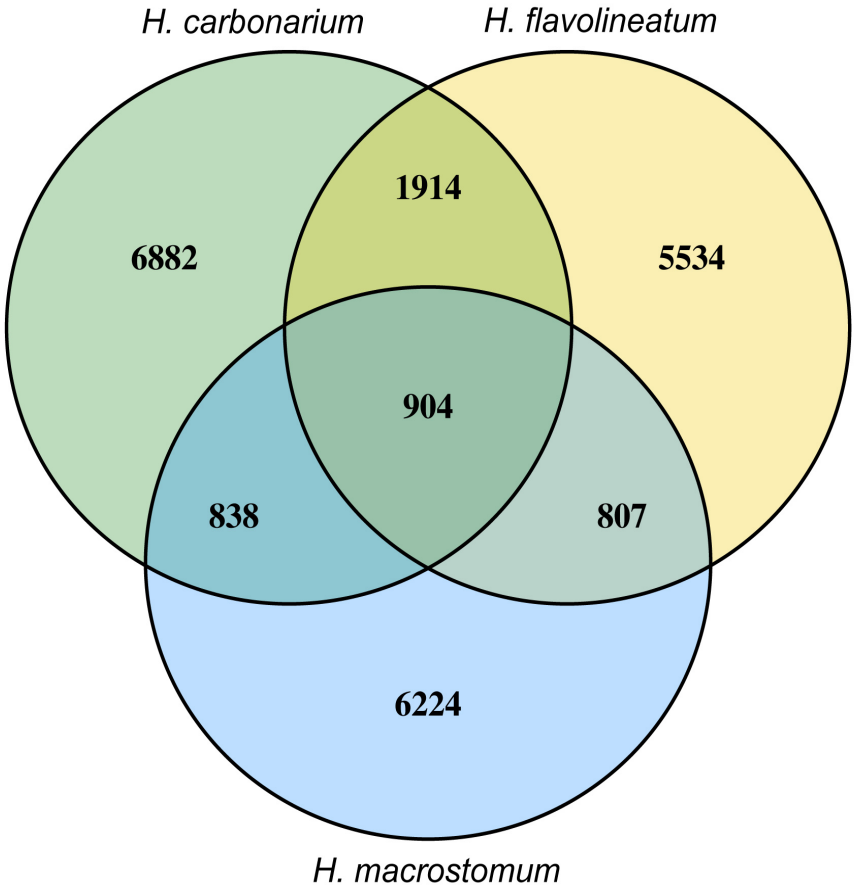


Figure 8. Venn diagram displaying the number of shared and unique genes found in the annotation of the transcriptomes of *Haemulon carbonarium*, *H. flavolineatum* and *H. macrostomum*. There are 904 annotated genes that are shared between the three species, and *H. carbonarium* had the highest number of unique genes.

In the case of the KOG, the most abundant conserved gene for *H. carbonarium* was a G protein-coupled receptor, part of the Signal Transduction Mechanisms category (abbreviated as T). In *H. flavolineatum* and *H. macrostomum*, the most abundant was a C2H2-type Zn-finger protein, which belongs to the Transcription category (K). As with the GO annotation, the five most common categories were also shared across the three transcriptomes, and include: Signal Transduction Mechanisms (T); General Function Prediction (R); Posttranslational Modification, Protein Turnover, Chaperones (O); Transcription (K); and Intracellular Trafficking, Secretion, and Vesicular Transport (U, Figure 10). There was also a considerable proportion of KOG terms considered "uncategorized conserved proteins" (S, Figure 10).

The most common KEGG pathway categories for the three species were related to Environmental Information Processing, and the most represented hierarchy was Signal Transduction (Figure 11). The second and third most represented hierarchies were Immune system and Endocrine system, and are both part of the Organismal Systems category. The pathways related to human diseases were not taken into account in this analysis.

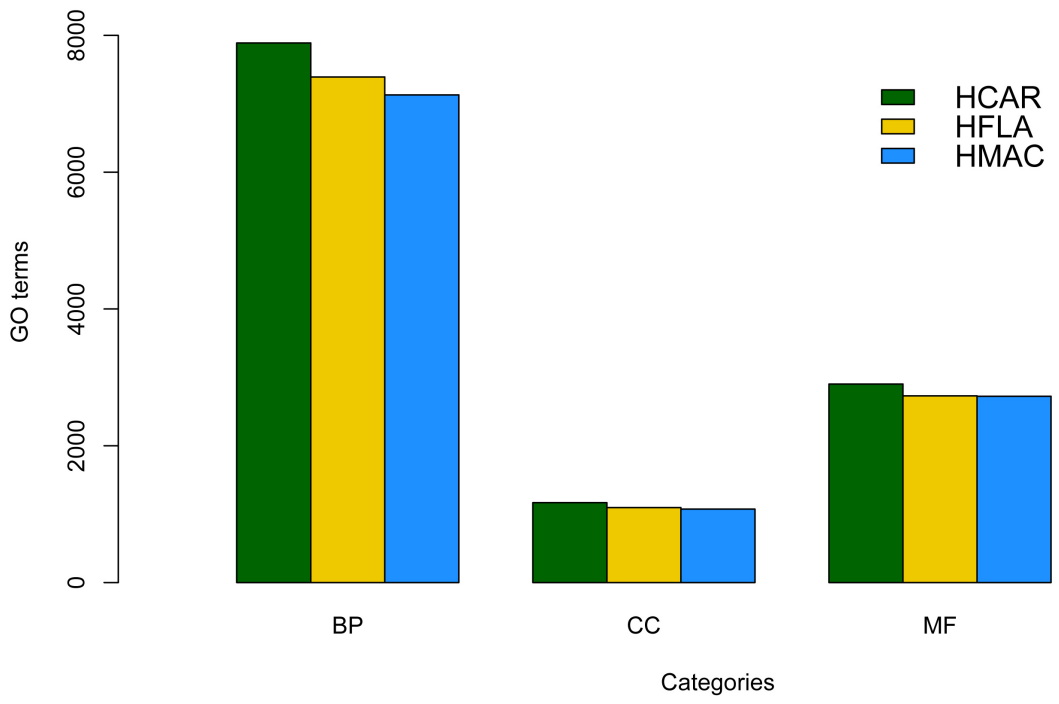


Figure 9. Gene Ontology (GO) terms that belong to each of the three GO categories: Biological Processes (BP), Cellular Components (CC) and Molecular Function (MF), for *Haemulon carbonarium* (HCAR), *H. flavolineatum* (HFLA) and *H. macrostomum* (HMAC).

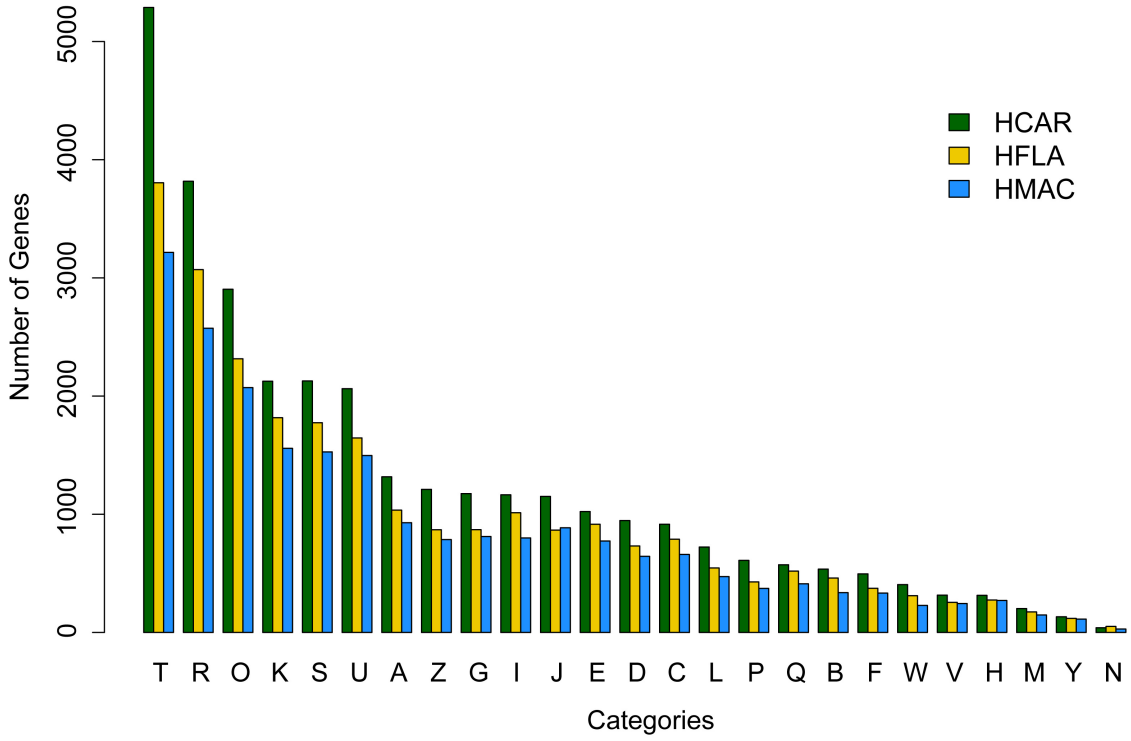


Figure 10. Number of unique Eukaryotic Clusters of Orthologous Groups (KOG) that belong to each of database categories: cell motility (N); nuclear structure (Y); cell wall/membrane/envelope biogenesis (M); coenzyme transport and metabolism (H); defense mechanisms (V); extracellular structures (W); nucleotide transport and metabolism (F); chromatin structure and dynamics (B); secondary metabolites biosynthesis, transport and catabolism (Q); inorganic ion transport and metabolism (P); replication, recombination and repair (L); energy production and conversion (C); cell cycle control, cell division, chromosome partitioning (D); amino acid transport and metabolism (E); translation, ribosomal structure and biogenesis (J); translation, ribosomal structure and biogenesis (I); carbohydrate transport and metabolism (G); cytoskeleton (Z); RNA processing and modification (A); intracellular trafficking, secretion, and vesicular transport (U); function unknown (S); transcription (K); posttranslational modification, protein turnover, chaperones (O); general function prediction only (R); and signal transduction mechanisms (T) for *Haemulon carbonarium* (HCAR), *H. flavolineatum* (HFLA) and *H. macrostomum* (HMAC).

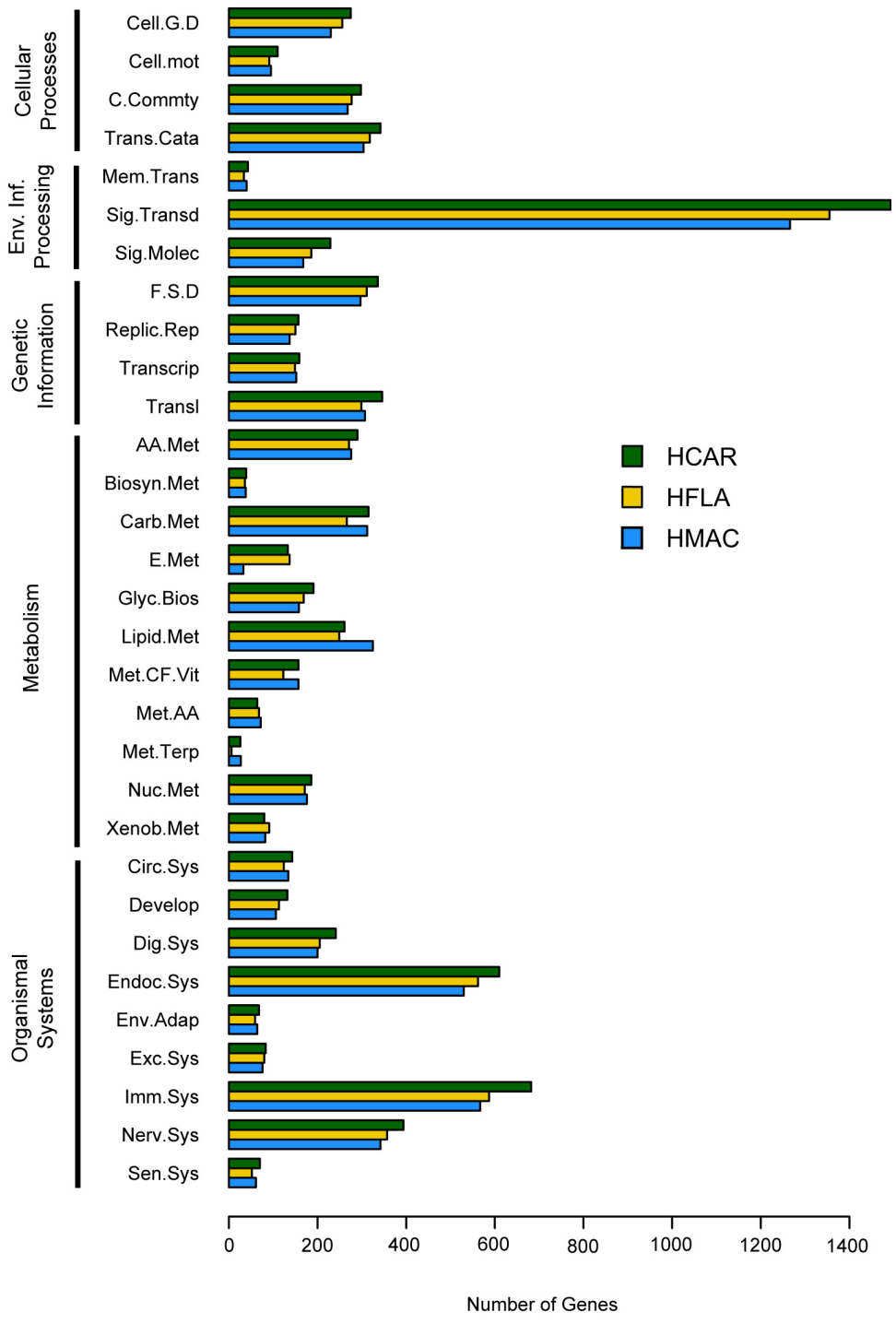


Figure 11. Kegg Pathways annotated for the sympatric species *Haemulon carbonarium* (HCAR), *H. flavolineatum* (HFLA), *H. macrostomum* (HMAC), and their corresponding categories.

DISCUSSION

The study of the complex evolutionary history of non-model organisms has received a significant boost with the advancement of massively parallel sequencing techniques. However, even with the rapid development and improvement of these methods, it is still relatively cumbersome and expensive to sequence and annotate full genomes for a large number of individuals. In this study, the annotation and sequencing of the transcriptome of three sympatric species of grunts, *H. carbonarium*, *H. flavolineatum* and *H. macrostomum* is described.

For each of the three transcriptomes, between 38 million and 48 million sequences were processed. The assembly of these short reads via Trinity lead to an average contig size of between 425 and 443bp. This is comparable to what other studies have reported when sequencing and annotating fish transcriptomes using Illumina reads (*e.g.:* Fraser *et al*. 2011; Schunter *et al*. 2014). A considerable percentage of contigs did not have significant hits with the UniProt database (Table 8). This could be related to the fact that most of the contigs obtained after assembly with Trinity were between 200 and 300bp long (~40%, Table 8), and the longest contig for the three transcriptomes was ~6,000bp. The resulting short fragments also had an effect in the overall contiguity, where only 12% to 16% of the annotated genes covered 75% of their genes or more.

Previous studies have revealed that short contigs have a lower probability of pairing with known genes, as the maximum achievable statistical significance of the match depends on the length of the query sequence (e.g. Meyer *et al*. 2009). The results from the present study are comparable to other fish transcriptomes that have used the UniProt databases for annotation. Examples include: annotation of 22.41% of the flounder genome (*Paralichthys olivaceus*, Huang *et al*. 2015), 21.4% of the black-faced blenny (*Tripterygion delaisi*, Schunter *et al*. 2014) and 26.9% of the silver carp

(*Hypophthalmichthys molitrix*, Fu & He 2012). Other examples have increased the percent of annotation by using both the UniProt datase and the Non-Redundant NCBI databases (*e.g.: Poecillia reticulata*, Fraser *et al*. 2011; *Scophthalmus maximus*, Pereiro *et al*. 2012). An interesting example is that of *Sparus aurata*, in which 51% of the 125263 sequences were successfully annotated (Calduch-Giner *et al*. 2013). For this, however, the authors queried 24 different databases, including the UniGenes of fugu, humans, killifish, medaka, salmon, trout and zebrafish (Calduch-Giner *et al*. 2013). Even though this approach could lead to optimal annotation, such an endeavor is computationally very intensive. Not only does this process involve a massive number of sequences that must be queried, but there should also be an optimal bioinformatics pipeline in order to avoid having multiple different matches for a particular transcript.

Another factor that influences the success of annotation, as mentioned before, is the size of the assembled contigs, which in turn can also be influenced by the number of sequences obtained. In the case of the Atlantic molly (*Poecilia mexicana*), sequencing of the cDNA library resulted in more than 70 million reads, which were assembled into ~80,000 contigs with a mean length of 932bp, and a max contig size of 15623bp (Kelley *et al*. 2012). As a result, 49.7% of the transcripts had a successful match using only the SwissProt database (Kelley *et al*. 2012). Thus, successful annotation is not only limited to the number of databases queried, but also to the size of the assembled contigs. Two other factors to take into account are that not all the annotations are done with the same confidence for e-values (this ranges from e-3 to e-10), and that those species closely related to model organisms will be better annotated (*e.g.:* Cichlids, Baldo *et al*. 2011).

Still, for the three species of interest between 27,000 and 40,000 transcripts matched known genes. This is by far the most complete genomic dataset of its kind for the Haemulidae family, and one of the few available transcriptomes for coral reef fishes.

Further, despite the fact that contiguity is low, the transcriptomes of the three species had at least 90% of representativity. This indicates that highly conserved proteins that are encoded in a considerable portion of eukaryotes were present in the denovo transcriptomes of the three grunts (Parra *et al*. 2007). This furthers the opportunities to compare the proteins of coral reef fishes with those of distantly related eukaryotes.

The relative difference in the percent of genes that belong to a particular GO, KOG and KEGG category varies significantly according to the tissue selected for building the transcriptome (*e.g.:* Ramsköld *et al*. 2009). These categories should not be seen as a way of characterizing different species, but rather, as a guide to determine whether a particular process is occurring at a distinct rate when comparing groups or treatments. Therefore, they will be of great use when identifying differentially expressed genes and targets of positive selection in future analyses.

CONCLUSION

This study significantly increases the genomic resources for the study of speciation of the genus *Haemulon*, the genus of coral reef fish with the highest proportion of sister species with sympatric distribution. No other study to date has done such extensive sequencing in this group of fishes, and very few other such datasets are available for marine fishes. The generation, annotation and description of these transcriptomes provides a critical first step for further analysis of orthologous genes and differential gene expression between recently diverged species, which could help identify molecular mechanisms driving the speciation of groups associated with coral reefs.

Chapter 3: Positive selection and differential gene expression support ecological speciation between three sympatric species of grunts (genus: *Haemulon*)

ABSTRACT

Understanding the genomic basis of adaptation is essential for discerning how the initial stages of divergence take place, especially in cases where ecological speciation is suspected. This could be particularly relevant in marine systems, which feature a dearth of geographical barriers and where organisms have the potential for long distance dispersal. This project aims to find signatures of positive selection and measure differential gene expression on three sympatric species of grunts: *Haemulon carbonarium*, *H. flavolineatum*, *H. macrostomum*. For the analysis of positive selection, 4120 orthologous genes were identified. The branch-site model was applied to the phylogeny of the group, resulting in significant evidence of positive selection in 87 annotated genes for *H. flavolineatum*, 150 for the branch of *H. carbonarium*/*H. macrostomum*, 79 for *H. carbonarium* and 88 for *H. macrostomum*. RNASeq analysis identified 1258 differentially expressed genes when comparing the three species. Further, despite the fact that the three species occupy the same geographic area, individuals grouped more closely with conspecifics based on gene expression. The genes pointed out by these analyses suggest there are significant differences in gene regulation, signal transduction, evolution of pharyngeal morphology, metabolic pathways and even reproductive traits, possibly driving the divergence in this group. This study hence supports the case for ecological speciation in the genus *Haemulon*.

INTRODUCTION

The accretion of molecular phylogenies over the past decades has revealed multiple cases of broadly distributed sister-species with completely overlapping ranges. Traditionally, reproductive isolation has been considered an essential pre-requisite for the onset of divergence between lineages (Wu & Ting 2004). This line of thought derives from geographic models of speciation, in which paramount importance is given to allopatry as a mechanism for differentiation (*e.g.:* Mayr 1947). Hence, such groups with completely overlapping distributions raise important questions about the origin and maintenance of biodiversity, offering opportunities to better our understanding of speciation.

In recent years, interest towards understanding the role natural selection plays in divergence has increased significantly, reinvigorating the concept of ecological speciation (*reviewed by*: Nosil 2012). Ecological views of speciation derive from the premise that strong divergent selection can ultimately lead to reproductive isolation, whether in allopatry or sympatry, and even when gene-flow is still persistent (Bird *et al*. 2012). Recent genetic studies have found direct evidence that adaptive advantage and reproductive isolation can be attained by disruptive selection in a small number of genes (Rundle & Nosil 2005; Nosil & Schluter 2011). While most examples to date correspond to terrestrial organisms, this "genic view" of ecological speciation could be especially important for explaining the outstanding diversity associated with coral reefs (Rocha & Bowen 2008; Puebla 2009). First, most marine species have sedentary lifestyles as adults, yet they begin life as highly mobile larvae that can maintain genetic connectivity across thousands of kilometers (*e.g.:* Lessios & Robertson 2006; Horne *et al*. 2008; Reece *et al*. 2011). Second, there is an apparent paucity of strong geographic barriers to account for

reproductive isolation (Rocha *et al*. 2007). Both of these points challenge the premise that allopatry is the predominant mode of speciation in marine ecosystems (Bowen *et al*. 2013).

Detecting genes under strong divergent selection between closely related lineages may enable us to understand the genomic basis of adaptations and speciation. For example, adaptations from riverine to lacustrine environments in cichlids are associated with strong positive selection of genes related to body shape, such as BMP4 (Terai *et al*. 2002) and EPCAM (Fan *et al*. 2011). Likewise, the gene Ectodysplasin has been associated with reduction of armor in sticklebacks that invade freshwater ecosystems (Barrett *et al*. 2008). Strong selection of opsin genes has also been found in lineages of fishes that inhabit environments with different water turbidity, with examples in cichlids (Sugawara *et al*. 2002; Seehausen *et al*. 2008), gobies (Larmuseau *et al*. 2010) and surgeonfishes (Gaither *et al*. 2015a).

However, limiting studies to a single gene could lead us to disregard other portions of the genome relevant in the process of differentiation (Berner & Salzburger 2015). Further, studies have demonstrated that evolution of differential gene expression can lead to large differences in phenotypes of closely related species (Jeukens *et al*. 2010; Kitano *et al*. 2013), and even to reproductive isolation (Pavey *et al*. 2010; Dion-Côté *et al*. 2014). Thus, the analysis of transcriptomes, with their corresponding levels of expression, could reveal a more complete picture of the genes associated with specific adaptations and reproductive isolation in coral reef fishes.

The species *H. carbonarium*, *H. flavolineatum* and *H. macrostomum* have completely overlapping distributions, and are commonly seen together with other grunts in shallow reefs and seagrass beds of the Caribbean (Lindeman & Toxey 2002). It has been suggested that due to their recent origin and dispersal ability (*H. flavolineatum*,

Purcell *et al*. 2006; Puebla *et al*. 2012b), it is highly unlikely that the group originated via allopatric speciation followed by range expansion (Rocha *et al*. 2008). The phylogeny of the group shows *H. carbonarium* and *H. macrostomum* are each other closest relatives, while *H. flavolineatum* is sister to this group (Rocha *et al*. 2008; Tavera *et al*. 2012). Despite the lack of information on reproductive behavior of grunts, the phylogeny of the group suggests there is reproductive isolation between the three species. The group is also characterized by an interesting dietary partition, as *H. flavolineatum* feeds mostly on polychetes and crabs, *H. carbonarium* feeds on small crabs and gastropods, while *H. carbonarium* feeds mostly on sea urchins (Randall 1967). In addition, there is evidence that coral reef ecosystems have accelerated the rate of morphological diversification of grunts, leading to a wide diversity of feeding morphologies (Price *et al*. 2013).

Taking into account their overlapping distribution and the ecological differences, this group is a well-suited candidate for the study of ecological speciation. Hence, the main objective of this study is to search for signatures of positive selection in orthologs of the focal species, as well as contrast patterns of gene expression of multiple individuals of the aforementioned groups.

METHODS

Assessing selection in grunt transcriptomes

Defining orthologs between the three species

In order to determine whether natural selection is acting on a particular gene or set of genes, the ratio of non-synonymous to synonymous substitutions ($\omega=dN/dS$) was estimated for the three transcriptomes assembled and annotated in Chapter 2. Gene orthology was assessed conducting a three way Reciprocal Best BLAST hits (RBB), a

fast and reliable method suggested to be as accurate as tree-based methods of orthology (Jeffares *et al*. 2015). First amino acid and coding sequences were extracted from the transcriptomes, using BLASTx to contrast the sequences with the Uniprot database (scripts available at: *https://github.com/z0on/annotatingTranscriptomes*). Extracted protein sequences from each pair of species were then reciprocally blasted using BLASTp. Orthologous sequences were identified as reciprocal best hits between all three species using custom python scripts (available at: *https://github.com/grovesdixon/metaTranscriptomes*). A sequence was considered an ortholog only if the reciprocal match had an e-value <e-5, >75% similarity and if sequence overlap was >50%. The orthologous protein sequences were then aligned with MAFFT v7 (Katoh & Standley 2013), and PAL2NAL (Suyama *et al*. 2006) was used to convert the coding sequences into codon alignments.

Ratios of non-synonymous to synonymous mutations

The *dN/dS* ratios between orthologous genes of the three species were estimated with Codeml (Yang 2007). Tests were conducted following the branch-sites models, which determines if positive selection is influencing sites of a particular gene in branches of a phylogenetic tree. This model was implemented using the "Test2", in which the "Null Model A" is contrasted against an "Alternative model A" (Yang *et al*. 2005), following the known phylogenetic hypothesis of the group (Rocha *et al*. 2008). In this case, significance was assessed comparing the null and alternate models with a likelihood ratio test in R, with one degree of freedom (available at: *https://github.com/grovesdixon/metaTranscriptomes*).

A Gene Ontology (GO) enrichment analysis conducted via a Mann-Whitney U (MWU) test was carried out to determine if any of the gene ontology

categories (Biological Processes, Cellular Components and Molecular Functions) were significantly enriched in any of the comparisons (Dixon *et al*. 2015). The main advantage of this particular method is that there is no need to establish a significance cut-off, estimating enrichment from the logarithm of the uncorrected *p*-values. The test was conducted following the pipeline developed by the Matz lab (available at: *https://github.com/z0on/GO_MWU*), and it was applied to the four comparisons, with a false discovery rate (FDR) ≤0.10.

Gene Expression

Sample collections

Specimens of the three sympatric species of grunts were collected by SCUBA divers using pole spears in the Bocas del Toro Archipelago, Panama in March of 2012. In total, eight specimens of *H. carbonarium*, 17 of *H. flavolineatum* and 11 of *H. macrostomum* were collected in a period of two days. Samples of liver were preserved in RNAlater (ThermoFisher) after each dive, then frozen at -20C at the Smithsonian Tropical Research Institute Bocas del Toro Station, and ultimately stored in -80C at the Center of Comparative Genomics of the California Academy of Sciences (CAS).

Tag-based RNASeq Library Preparation

Liver samples from all the individuals were extracted using the RNAqueous Kit (Life Technologies), following the manufacturer's instructions. The final product was eluted in 40μl of buffer, and treated with DNAase. Extractions were quantified with a Nanodrop spectrophotometer (Thermo Scientific), and all samples had concentrations between 75 and 600ng/μl of RNA. Preparation of RNAseq libraries was done over three days, following the protocol of (Meyer *et al*. 2011) with few modifications for Illumina

sequencing (available at: *https://github.com/z0on/tag-based_RNAseq*). This protocol for library preparation produces a single tag per mRNA molecule at a random position near the 3' end (Meyer *et al*. 2011). RNA extractions were fragmented using heat incubation to eliminate any bias by transcript length. The first-strand cDNA synthesis was done with 300 to 500ng of RNA product, with a primer that specifically targets the 3' ends. cDNA amplifications were individually labeled with specific barcodes. After barcoding, samples were size selected via gel-extraction, targeting fragments between 400bp and 500bp. Because all the samples were sequenced in a single lane, purified fragments were quantified via qPCR before pooling. Single end reads of 50bp were sequenced in a single lane of Illumina HiSeq HiSeq 2000, at the Genomics Sequencing and Analysis Facility at the University of Texas at Austin (GSAF). Once sequences were obtained they were de-mutliplexed, according to their specific barcodes. Removal of low quality reads and PCR duplicates, as well as adaptor trimming, was done following the bioinformatics pipeline of the Matz lab (available at: *https://github.com/z0on/tag-based_RNAseq*).

Mapping of reads to the reference transcriptome was done using Bowtie2 (Langmead & Salzberg 2012), reporting only the best match for each sample. The annotated transcriptome of *H. carbonarium* was used as reference, as it had the highest number of transcripts with successful Blast hits (40342). The number of reads that mapped to a particular isogroup was extracted into a single text file using samcount.pl and expression_compiler.pl (available at: *https://github.com/z0on/tag-based_RNAseq*). The counts file was used for all the gene expression analyses, which were conducted with R 3.2.1 (R Core Team 2015).

Differential Gene Expression

Differentially Expressed Genes (DEGs) were identified with DESeq2 (Love *et al*. 2014), using negative binomial generalized linear models. In order to avoid any skews by unequal sequencing efforts, size factors were estimated for all the counts, using the three different species as conditions. A likelihood ratio test (LRT) was used to determine if there was significant contrast in gene expression between the three species (Love *et al*. 2014). An empirical FDR was estimated using the R package *empiricalFDR.DESeq2* (Wright *et al*. 2015), which offers the advantage of not having to discard low abundant genes. A table with adjusted and un-adjusted p-values of the log2 fold changes was exported for further analyses. Pair-wise comparisons between the three species (*H. carbonarium* x *H. flavolineatum*, *H. carbonarium* x *H. macrostomum*, *H. macrostomum* x *H. flavolineatum*) were conducted using Wald tests, which determine whether the standard error of the log2 fold change is different from zero (Love *et al*. 2014).

A Principal Coordinate Analysis (PCoA) was carried out to visualize if individuals of the three species clustered together based on gene expression. For the PCoA, a matrix of nearly homoscedastic values was obtained via variance stabilization of the data with DESeq2. The principal coordinate decomposition of the variance-stabilized values was estimated with the R package ape (Paradis *et al*. 2004). To visualize the significant comparisons of gene expression between the three species of grunts, a heat map with two-way hierarchical clustering was generated using the R package pheatmap (Kolde 2015).

As in the tests of positive selection, a GO-MWU test was carried out to determine if any of the gene ontology categories were significantly enriched in the contrasts of differential gene expression (Dixon *et al*. 2015). The test was conducted following the pipeline developed by the Matz lab (available at

https://github.com/z0on/GO_MWU), and it was applied to the three species and the pair-wise comparisons, with a FDR ≤0.10.

RESULTS

Positive selection in sympatric grunt species

The RBB resulted in 4120 orthologous genes between the three sympatric species. The largest ortholog was 3216bp and the smallest one 150bp. When implementing the branch-site models in Codeml, genes were considered to be under positive selection if the likelihood ratio test significantly favored the alternate models, at a FDR ≤0.05. This resulted in 190 positively selected genes for the branch of *H. flavolineatum*, 87 of which were annotated (Appendix B, Supplementary Table 1). Examples of genes under positive selection in this branch include: Mediator of RNA polymerase II transcription subunit 12, Poly [ADP-ribose] polymerase 4 and Zinc finger proteins. For the branch of *H. carbonarium*/ *H. macrostomum* 311 genes showed evidence of positive selection, 150 of them with protein annotation (Appendix B, Supplementary Table 2). Examples include: Amphiphysin, Glomulin, Leucine-rich repeat-containing protein C10orf11 homolog, Stonustoxin and the Zona pellucida sperm-binding protein 3. For the branch of *H. carbonarium*, 157 genes were under positive selection (79 annotated; Appendix B, Supplementary Table 3), whereas 194 were for *H. macrostomum* (88 annotated; Appendix B, Supplementary Table 4). Out of these, 26 were exclusive of the branch of *H. carbonarium*, while 19 were exclusive of *H. macrostomum*. Hence, these represent transcripts under selection between the two sister-species. The number of coding regions with annotation that overlapped between the three species is presented in Figure 12.

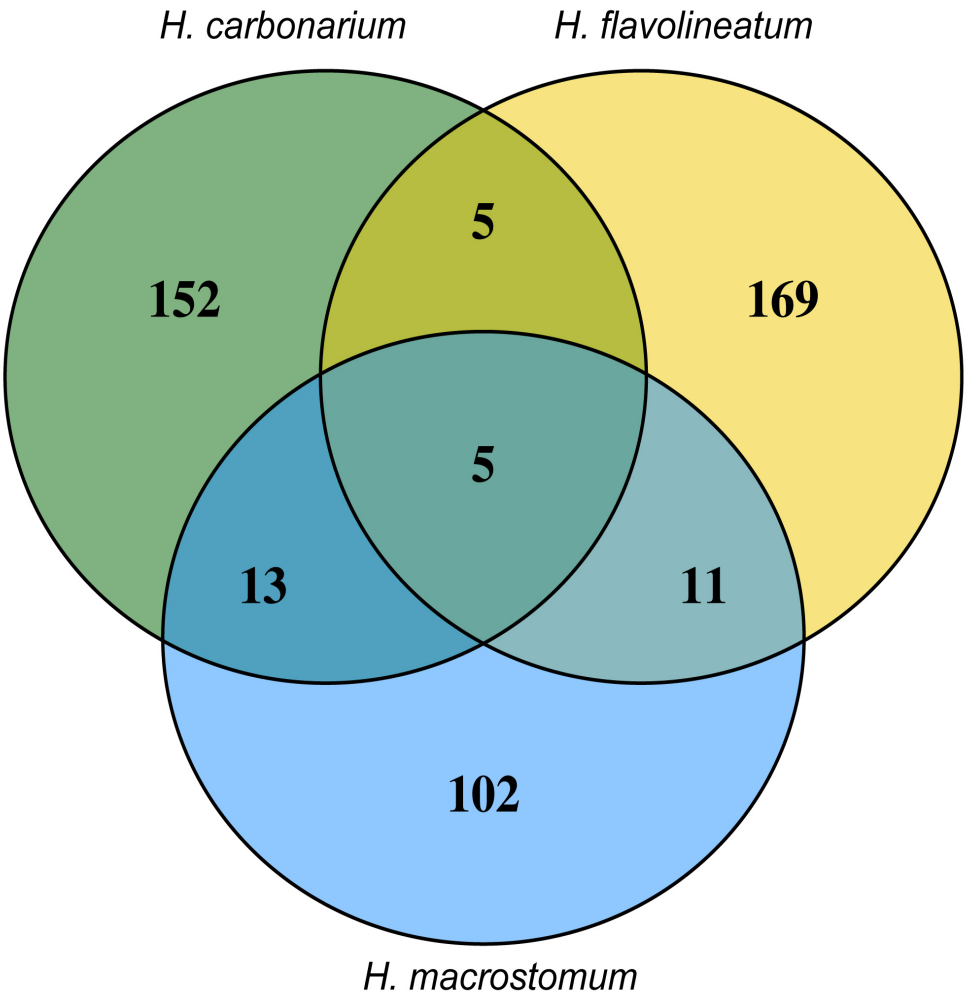


Figure 12. Venn diagram with the number of genes with *dN*/*dS*>1 between the branches of *H. carbonarium*, *H. flavolineatum* and *H. macrostomum*. Numbers in the middle represents the number of genes that overlapped between the three comparisons.

The GO enrichment tests for branch-sites comparisons revealed that the category Biological Processes had enriched terms on the individual branches of the three species. In these cases, the most significant term was Adaptive immune response based on somatic recombination of immune receptors built from immunoglobulin superfamily domains (Appendix B, Supplementary Figure 1). Meanwhile, for the branch of *H. carbonarium*/ *H. macrostomum*, 29 GO terms belonging to the Biological Processes were significantly enriched, the most significant was Mismatch Repair. Meanwhile, the category Cellular Component had five enriched GO terms, MHC protein complex was the most significant. Finally, the categories Molecular Functions had eight enriched GO terms, the most significant being Cytokine Binding (Supplementary Figure 2).

Gene Expression

Sequenced samples were represented by a minimum of 894891 to a maximum of 4693033 reads, with an average of 2320732 reads per individual, after removal of adaptors and low quality fragments. The average number of reads per sample per species was: 2255449 for *H. carbonarium*, 2200265 for *H. flavolineatum* and 2554387 for *H. macrostomum* (Table 10). Mapping to the transcriptome of *H. carbonarium* lead to an overall efficiency of 81% for *H. carbonarium*, 79% for *H. flavolineatum* and 80% for *H. macrostomum* (Table 10). Reads of the three species matched a total of 76869 isogroups. After filtering, for the analysis of DE of the three species, the minimum number of counts for an individual was 616969 and the maximum was 3343800, with a mean of 1603110.

Species	Number of individuals	Total Number of Reads	Average Reads (per individual)	Efficiency of mapping
H. carbonarium	8	18043592	2255449	81%
H. flavolineatum	17	37404508	2200265	79%
H. macrostomum	11	28098252	2554387	80%

Table 10. Number of individuals, total number of reads, average number of reads and efficiency of mapping to the transcriptome of *H. carbonarium* for *H. carbonarium*, *H. flavolineatum* and *H. macrostomum*.

The analysis of differential expression between the three species with the LRT (Love *et al*. 2014), and the subsequent calculation of empirical FDR (Wright *et al*. 2015), yielded 1258 significant transcripts, with an estimated FDR of 2.09%. The pair-wise comparisons with the Wald test resulted in 8232 significant genes between *H. carbonarium* and *H. flavolineatum* (FDR= 2.33%), 5399 up-regulated and 2833 down-regulated. The comparison between *H. carbonarium* and *H. macrostomum* yielded 8508 DEGs (FDR= 2.47%), 5540 up-regulated and 2968 down-regulated. Finally, between *H. flavolineatum* and *H. macrostomum* 5705 genes were up-regulated and 5565 down-regulated, out of 11270 DEGs (FDR= 3.07%).

The PCoA shows that individuals cluster together by species based on differential gene expression (Figure 13). In this case, the first principal coordinate explained 13.14% of the variation, while the second principal coordinate 8.92%. Dispersion of points for *H. carbonarium* was 19, 15 for *H. flavolineatum* and 17 for *H. macrostomum*.

The heatmap for the three species was elaborated first with the 1258 genes that had significant differential expression, where 150 genes had annotation (Appendix B, Supplementary Figure 3). However, due to the large number of significant genes, a second heatmap was elaborated only with the lowest 300 unadjusted *p*-values, 34 of which were annotated (Figure 14). Both cases revealed major differences between the three species (Figure 14).

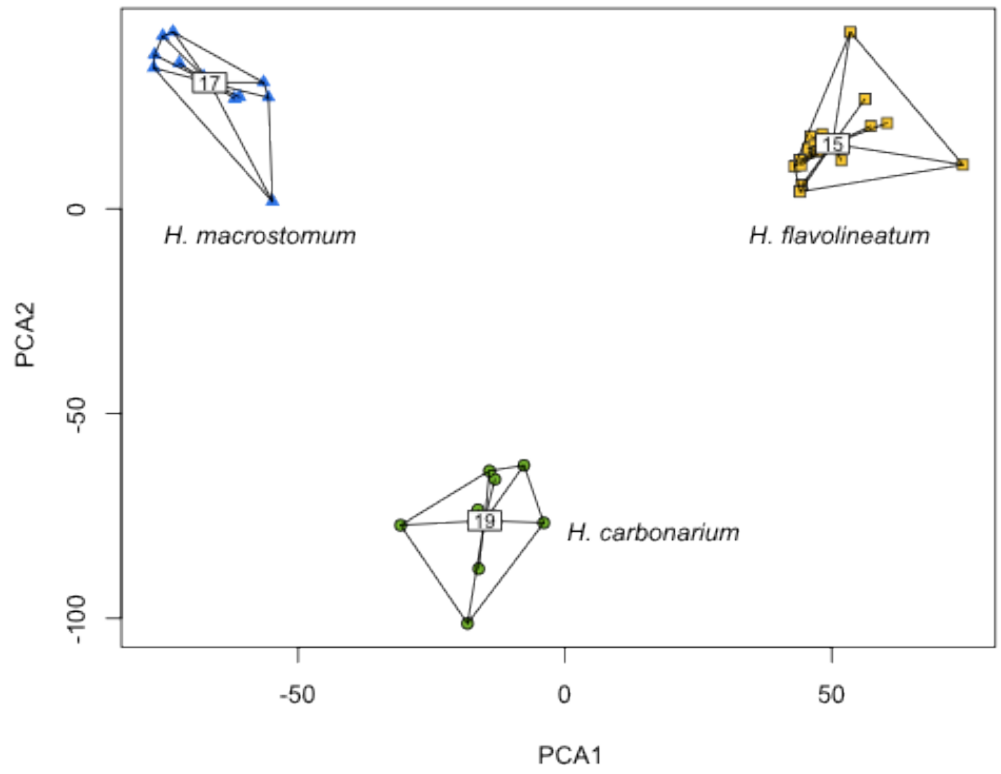


Figure 13. Principal Coordinate analysis using the differentially expressed genes found between the three species of grunts. PC1 explains 13.14% of the variance, while PC2 8.92%. Numbers in the plot indicate the dispersion of individuals for each of the species.

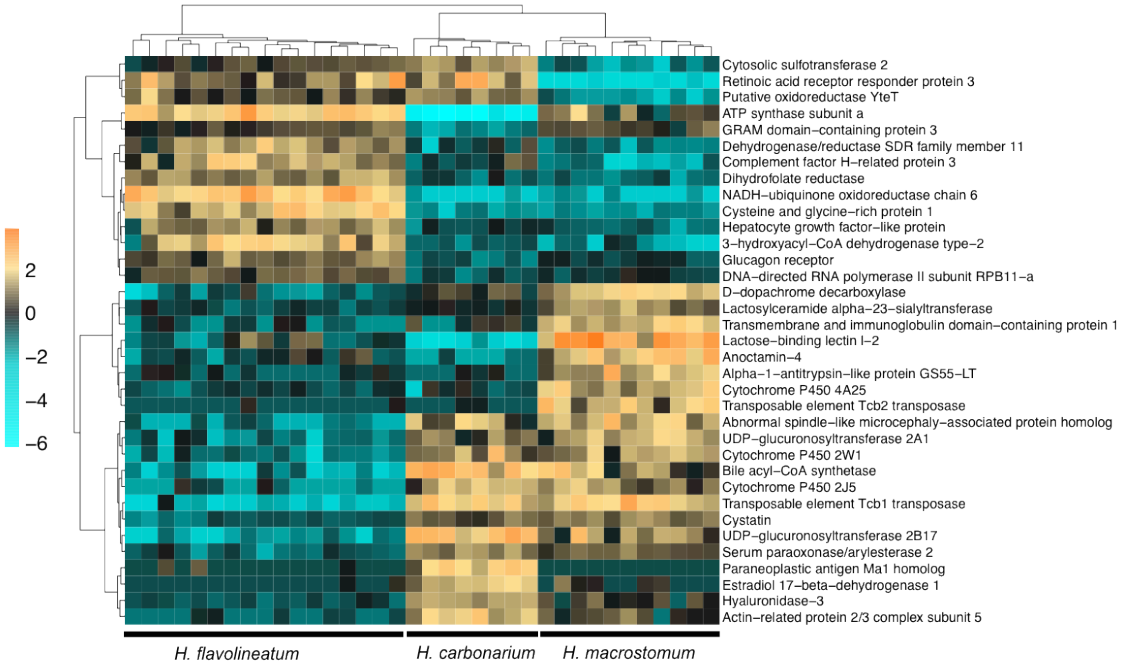


Figure 14. Heat map for the top 36 differentially expressed genes with successful annotation for three sympatric species of grunts of the Western Atlantic.

There were 37 genes under positive selection that also showed differential expression between the three species (Appendix B, Supplementary Table 5). A Fisher's exact test suggested this number is significantly lower than what is expected by chance (P<0.0001). These include genes related with development (Fibroblast growth factor receptor 1-A and Angiopoietin-related protein 3), signal transduction (A-kinase anchor protein 2) and structures of the nervous system (Complement C1q-like protein 4).

The analysis of GO enrichment for the DEG of the three species based on a LRT revealed multiple significant terms with a 10% FDR (Figure 15). For Biological Processes, 22 terms were significantly enriched, the most up-regulated (15 terms in total) was Indol-alkylamine catabolic process, the most down-regulated (7 terms in total) was FC receptor mediated stimulatory signaling pathway. For Cellular Components, 25 GO

terms were enriched, the most up-regulated was Fibrinogen Complex (13 terms), and the most down-regulated was Pronucleus (12 terms). For Molecular Function 46 GO terms were significant, the most up-regulated was Nutrient reservoir activity (28 terms), while the most down-regulated was Myosin binding (18 terms).

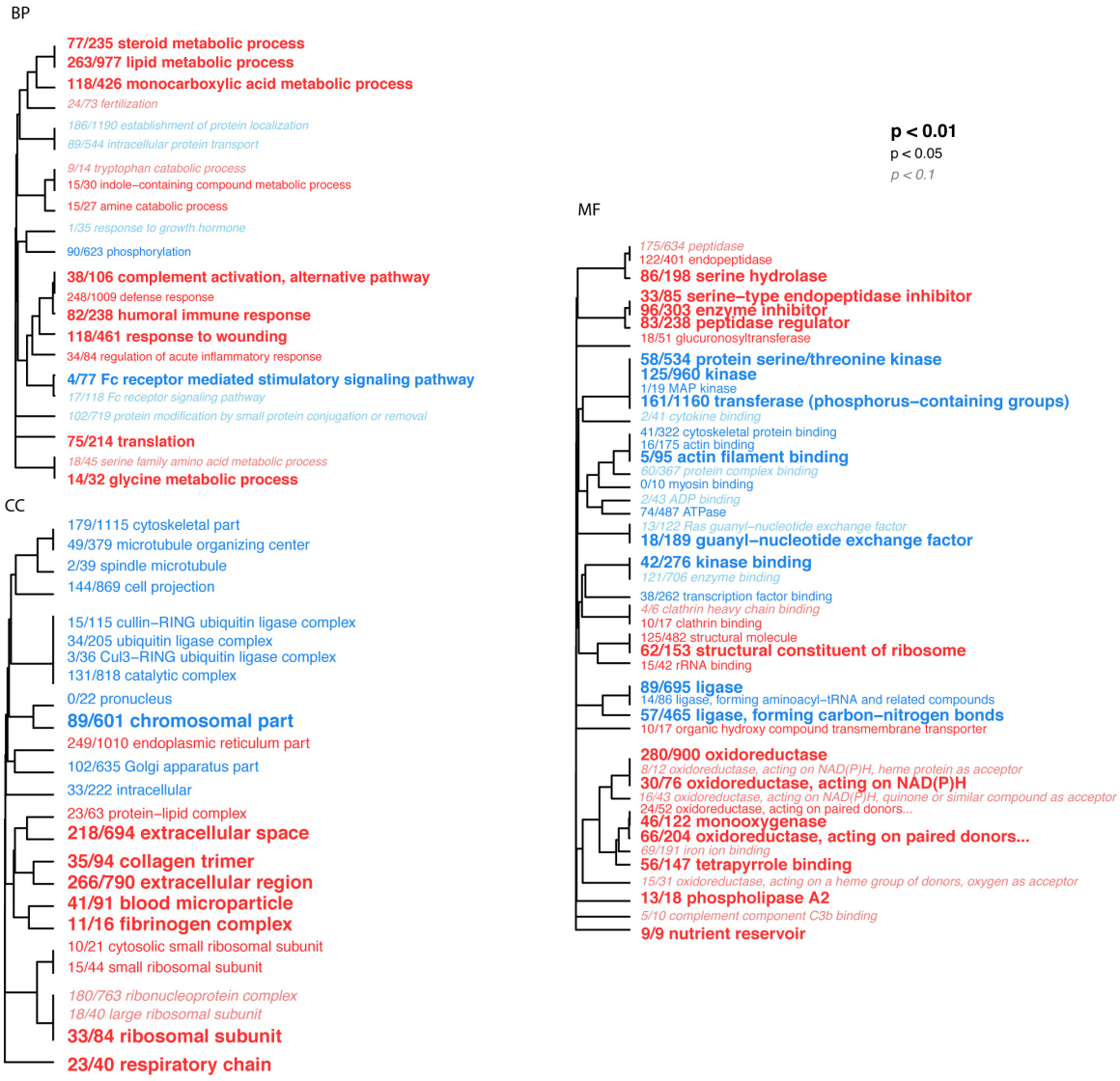


Figure 15. Enriched GO terms corresponding to the differentially expressed genes between the three target species of grunts.

Species	Direction	Significant Genes	Significant GO terms, and most up- or down-regulated		
			BP	CC	MF
HCAR vs HFLA FDR=2.33%	Up	5,399	2 (Steroid Metabolic Process)	8 (Fibrinogen Complex	12 (Nutrient reservoir activity)
	Down	2833	-	1 (ISWI-type complex)	7 (Intracellular ligand-gated ion channel activity)
HCAR vs HMAC FDR=2.47%	Up	5,540	9 (Complement activation, alternative pathway)	8 (Fibrinogen complex)	10 (Phospholipase A2 activity)
	Down	2,968	3 (FC receptor mediated stimulatory signailing pathway)	3 (Spindle microtubule)	8 (Chondroitin sulfotransferase activity)
HFLA vs HMAC FDR=3.07%	Up	5,705	17 (Triptophan catabolic process)	11 (Fibrinogen Complex)	20 (Nutrient Reservoir Activity)
	Down	5,565	5 (Reciprocal Meiotic recombination)	7 (Cul3-RING ubiquitin ligase complex)	11 (Antigen Binding)

Table 11. Number of significant genes that were up- and down-regulated in pairwise comparisons of *H. carbonarium* (HCAR), *H. flavolineatum* (HFLA) and *H. macrostomum* (HMAC), along with the number of significant GO terms enriched in the comparisons. Names in parenthesis represent the most significantly enriched GO term for that particular category.

Results of the GO enrichment analyses of the pairwise comparisons are presented in Table 11 (Appendix B, Supplementary Figures 4-6). Interestingly, two terms were the most significantly enriched in multiple pairwise comparisons. The most significantly enriched GO term in the category of Cellular Component for the three comparisons was

Fibrinogen Complex (Table 11). For the Molecular Function category, Nutrient Reservoir Activity was the most significantly up-regulated term in the comparison of *H. flavolineatum* and *H. macrostomum*, and *H. carbonarium* vs *H. flavolineatum* (Table 11). The comparison with the highest number of significant GO terms was between *H. flavolineatum* and *H. macrostomum* (Table 11; Appendix B, Supplementary Figure 6).

DISCUSSION

Cases of ecological speciation are particularly appealing opportunites to understanding the genomic basis of divergence, especially in cases where closely related lineages are found in sympatry. This is especially relevant for the fauna associated with coral reef ecosystems, where long distance dispersal and the dearth of physical boundaries challenge the efficacy of allopatric speciation (Bowen *et al*. 2013). With this in mind, signatures of positive selection and differential expression were assessed in three closely related species of grunts with completely overlapping distributions. These large bodied fishes show very little population structure through the Caribbean (Purcell *et al*. 2006; Puebla *et al*. 2012b), have stark differences in diet despite occupying the same habitat (Randall 1967) and have accelerated rates of morphological diversification (Price *et al*. 2013), making them well-suited candidates for the study of ecological speciation.

Positive selection influenced large-effect genes, metabolism and reproduction of grunts

The analyses with orthologous sequences yielded high numbers of genes under positive selection with the branch-sites model. The branch of *H. flavolineatum* had 87 annotated genes showing evidence of positive selection, some of which are of large-effect. For example, Poly [ADP-ribose] polymerase 4, is responsible of adding long

ADP-ribose units, influencing vital tasks such as DNA repair, chromosomal and transcription regulation, mitosis and cell apoptosis (Ahel *et al*. 2008; Kienzler *et al*. 2013). Further, multiple genes involved in transcriptional regulation, including Mediator of RNA Polymerase II, Transcriptional regulator ATRX and multiple Zync finger proteins, were also under selection. Recent studies have highlighted the importance of transcription regulation in speciation, as these proteins can influence multiple molecular pathways, modulating the expression of many genes (*reviewed by*: Romero *et al*. 2012).

The branch with the highest number of genes under positive selection was that of *H. carbonarium*/*H. macrostomum*, with 150 annotated genes. This included genes associated with the development and arrangement of the vascular system, such as Glomulin and the Atrial natriuretic peptide receptor 2 (Appendix B, Supplementary Table 2). Since the circulatory system is associated with transporting oxygen to tissues, differences in its arrangement have been associated with contrasts in metabolic efficiency (*e.g.:* Dion-Côté *et al*. 2014). This is also supported by selection of genes related with stress response, such as Stress-70 protein mitochondrial, which are involved in tolerance of food deprivation and temperature changes (Cara *et al*. 2005). These metabolic differences could be related to contrast in diet between the three species (Randall 1967).

Positive selection in *H. carbonarium*/*H. macrostomum* also influenced genes associated with reproductive isolation. In humans, Leucine-rich repeat-containing protein C10orf11 homolog is known to play an important role in the differentiation of color in melanocytes (Grønskov *et al*. 2013). The three species show marked difference in color pattern, and there are several examples in reef fishes in which color patterns play a fundamental role in mating (Puebla *et al*. 2012a), crypsis (Losey Jr 2003) and mimicry (Eagle & Jones 2004). Although differences in color pattern could be associated with speciation, this hypothesis has to be tested in grunts. The other gene is the Zona pellucida

sperm-binding protein 3, which is an important component of oocyte envelopes in several animal groups. This particular protein serves as a sperm receptor during fertilization in mammals (Swanson & Vacquier 2002) and mollusks (Aagaard *et al*. 2006). In teleosts, however, this protein is lacking the sperm recognition site that is present in other vertebrates (Wang & Gong 1999; Hyllner *et al*. 2001). Thus, it has been suggested that Zona pellucida proteins have an important function guiding the sperm to the egg's micropyle during fertilization (Iwamatsu *et al*. 1997), suggesting its involvement in the evolution of prezygotic barriers of the group. Another gene that could be associated with reproduction is Stonustoxin. This protein is well known as a component of the venom produced by stonefishes (Ghadessy *et al*. 1996). However, studies have also found it to be associated with the proper development and maturation of ovaries in salmon (von Schalburg *et al*. 2004).

Differential gene expression offers insights into the origins of divergence:

Over a thousand differentially expressed genes were found between the three species, as well as with the pair-wise comparisons. However only 37 genes were both significantly positively selected and differentially expressed. A relevant gene in this category is the Fibroblast growth factor receptor 1-A (FGFR1), which plays a fundamental role in the pharyngeal development of vertebrates, especially with relation to the development of gill arches (Trokovic *et al*. 2005). A knock out experiment in zebrafish found that FGFR1 is expressed in the pharyngeal endoderm, and that its absence causes disruption in cranial cartilage formation (Larbuisson *et al*. 2013). Fibroblast growth receptors could be very important in the diversification of this group, as grunts possess special adaptations in branchial arches for crushing hard prey (Wainwright 1989), and have high divergence of traits associated with feeding

morphology (Price *et al*. 2013, Appendix B, Supplementary Figure 7). In fishes, pharyngeal morphology correlates well with diet, as fishes that consume hard-bodied preys have well-developed pharyngeal teeth as well as larger jaw muscles, when compared with fishes that feed on soft prey (Grubich 2003). These differences can be observed in the three focal species (Appendix B, Supplementary Figure 7), as *H. flavolineatum* feeds mostly on soft-bodied organisms, *H. carbonarium* feeds on both soft and hard-bodied prey, and *H. macrostomum* consumes large organisms with hard shells almost exclusively (Randall 1967). Dietary partitioning has been suggested as one of the multiple mechanisms that can lead to ecological speciation in fishes (*reviewed by:* Bernardi 2013), and the available information suggests it played a role in the differentiation of grunts. Further, it is well known that grunts produce sounds via stridulation of pharyngeal teeth (Bertucci *et al*. 2014), thus differences in pharyngeal morphology could also be associated with sound production and recognition of conspecifics.

A small number of significant GO terms overlapped in the pairwise comparisons of DEG, when performing the enrichment analyses (Appendix B, Supplementary Figures 4-6). Terms like Respiratory Chain, Ribosomal subunit and Nutrient Reservoir were significantly enriched in the three pairwise comparisons, suggesting fundamental differences in the metabolism of the three species. Meanwhile, terms such as Protein serine/threonine kinases, Transferase, Transferring phosphorus-containing groups and Oxido-reducatase could also be responsible for large differences, as they are important in signal transduction and gene regulation.

The largest number of enriched GO terms was observed between *H. flavolineatum* and *H. macrostomum*. This is not surprising, considering both of these species have the highest divergence in morphology and ecology (Lindeman & Toxey

2002). As in other comparisons, categories related with regulation of gene expression, such as Signal transducer, Transcription factor binding and Regulatory region nucleic acid binding, were up-regulated in *H. macrostomum*.

Finally, one of the GO terms significantly up-regulated in *H. flavolineatum* with respect to *H. macrostomum* was Female Pregnancy. This was corroborated by the up-regulation of Zona pellucida genes and Vitellogenin in half of the individuals of *H. flavolineatum* (Appendix B, Supplementary Figure 3). This suggests that some individuals of this species were mature and ready for reproduction, while the other individuals were not. Considering that all samples from this study were collected in the wild, it was impossible to account for differences in the developmental stage and life history of each of the individuals analyzed. Hence, the results here presented should be interpreted with caution.

CONCLUSION

The analyses of divergent selection and differential gene expression revealed important differences in portions of the genome related with gene regulation and signal transduction, which have been associated with phenotypic diversity in teleosts. Further, genes related with reproduction give clues on how these species can maintain isolation despite their overlapping distribution. By finding significant differences in genes related with metabolic processes and pharyngeal morphology, this study provides a novel hypothesis on how ecological speciation in grunts could be related with dietary partitions. By suggesting a source of initial divergence, as well as mechanisms for reproductive isolation, this study adds further evidence that ecological speciation is an important mechanism for explaining the high levels of diversity associated with coral reefs.

Conclusions

The genus *Haemulon* offers incredible opportunities for the study of speciation. The group is composed of species pairs with completely overlapping distributions, and stark ecological differences between them. Despite this, the group has been relatively neglected from evolutionary studies until recently, mainly because grunts have little commercial value and are not part of the aquarium trade. Hence, this study represents one of the few efforts to determine the mechanisms that lead to the diversification of a fascinating group associated with coral reefs.

This research project provides novel evidence that indicates the evolutionary history of grunts has been influenced by hybridization and divergent selection, supporting the case for ecological speciation. Shared haplotypes and reduced divergence in the mitochondria indicate recent hybridization between the sister species *H. flaviguttatum* and *H. maculicauda* in the recent past. The presence of a higher number of haplotypes in *H. maculicauda* with respect to *H. flaviguttatum* suggests directionality of the introgression, from the former species to the latter. This pattern could be explained by hybridization, followed by a selective sweep of the mitochondrial genome in *H. flaviguttatum*. Furthermore, it appears that hybridization has had little effect in the nuclear genome, possibly because nuclear admixture leads to individuals of reduced fitness. This is further supported by divergent selection found in genes of important functions in the nuclear DNA.

Meanwhile, in the case of *H. carbonarium*, *H. flavolinetaum* and *H. macrostomum*, there was strong positive selection in genes related with sperm guiding during fertilization, which give clues on the prezygotic barriers between the three species. Further, a gene related with the development of pharyngeal structures was both positively

selected and differentially expressed. In fishes, distinction in the pharyngeal apparatus is related to contrasting feeding habits. Species that feed on large organisms with shells have larger pharyngeal plates and well-developed muscles, when compared to fishes that feed on soft prey. Selection and differential expression of coding regions associated with the circulatory system, immunology, stress response and metabolism could also be correlated with differences in diet between the three species. This provides evidence that the divergence of grunts could have originated by dietary partitions that characterize these sympatric species.

In studies where ecological speciation is suggested, many more questions arise from the proposed mechanisms of differentiation. Thus, in the case of *H. maculicauda* and *H. flaviguttatum*, the role of selection played in the differentiation of the nuclear genome needs to be evaluated. Moreover, in the case of *H. carbonarium*, *H. flavolineatum* and *H. macrostomum*, future studies need to assess the link between divergent selection in genes related to pharyngeal morphology and their corresponding effects on phenotype. In addition, studying the potential role of sounds in recognition of conspecifics, as well as a detailed characterization of dietary partitions of grunts, would help provide a clear picture of the mechanisms of isolation in this fascinating group. By finding direct evidence of divergent selection in grunts of the genus *Haemulon*, this study opens new avenues for future projects that will help elucidate the mechanisms of diversification of coral reef fishes.

Appendices

APPENDIX A

Supplementary Table 1. Pairwise F_{ST} between populations of *Haemulon maculicauda* (HMA) and *Haemulon flaviguttatum*, in Mexico and Panama, with (A) COI, (B) CytB and CR. *P*-values are presented above the diagonal and significant values are displayed in bold.

A – COI: Between species F_{ST} = 0.1726 (p= 0.00)

		1	2	3	4
Haemulon maculicauda	1. Mexico	-	0.968	0.011	<0.001
	2. Panama	-0.031	-	<0.001	0.002
Haemulon flaviguttatum	3. Mexico	**0.192**	**0.103**	-	0.093
	4. Panama	**0.392**	**0.092**	0.195	-

B – CytB: Between species F_{ST} = 0.3592(p=0.00)

		1	2	3	4
Haemulon maculicauda	1. Mexico	--	0.064	<0.001	<0.001
	2. Panama	0.060	--	<0.001	<0.001
Haemulon flaviguttatum	3. Mexico	**0.434**	**0.346**	--	<0.001
	4. Panama	**0.500**	**0.393**	**0.245**	--

C - CR: Between species F_{ST} = 0.299 (p= 0.00)

		1	2	3	4
Haemulon maculicauda	1. Mexico	--	0.07129	<0.001	<0.001
	2. Panama	0.044	--	<0.001	<0.001
Haemulon flaviguttatum	3. Mexico	**0.275**	**0.320**	--	0.004
	4. Panama	**0.408**	**0.449**	**0.258**	--

D - RAG2: Between species F_{ST} = 0.771 (p= 0.00)

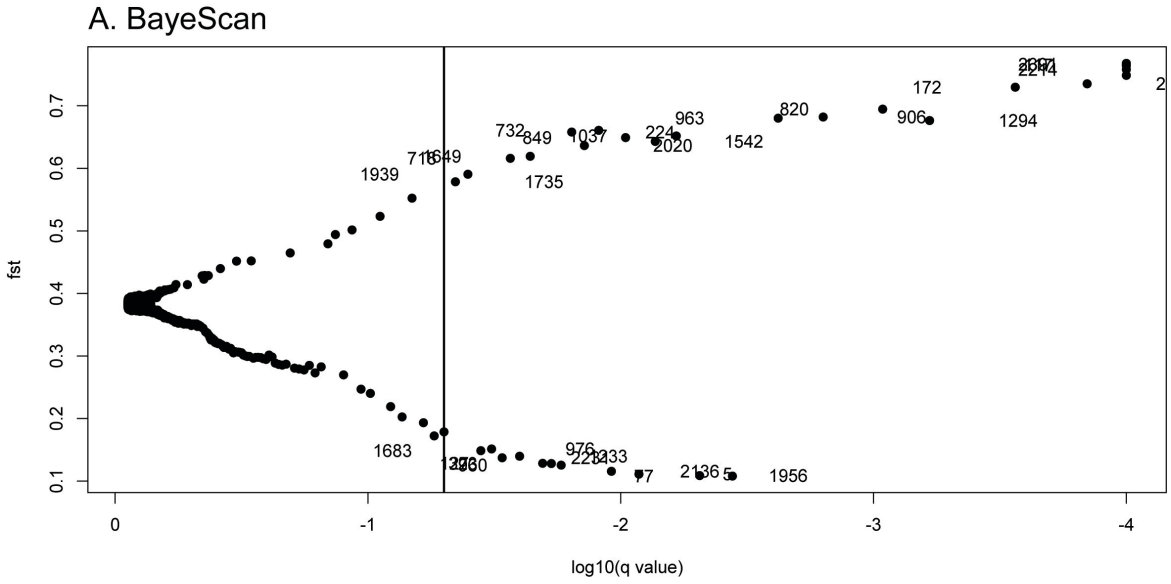
		1	2	3	4
Haemulon maculicauda	1. Mexico	--	<0.001	<0.001	<0.001
	2. Panama	**0.294**	--	<0.001	<0.001
Haemulon flaviguttatum	3. Mexico	**0.924**	**0.8973**	--	0.020
	4. Panama	**0.864**	**0.823**	**0.340**	--

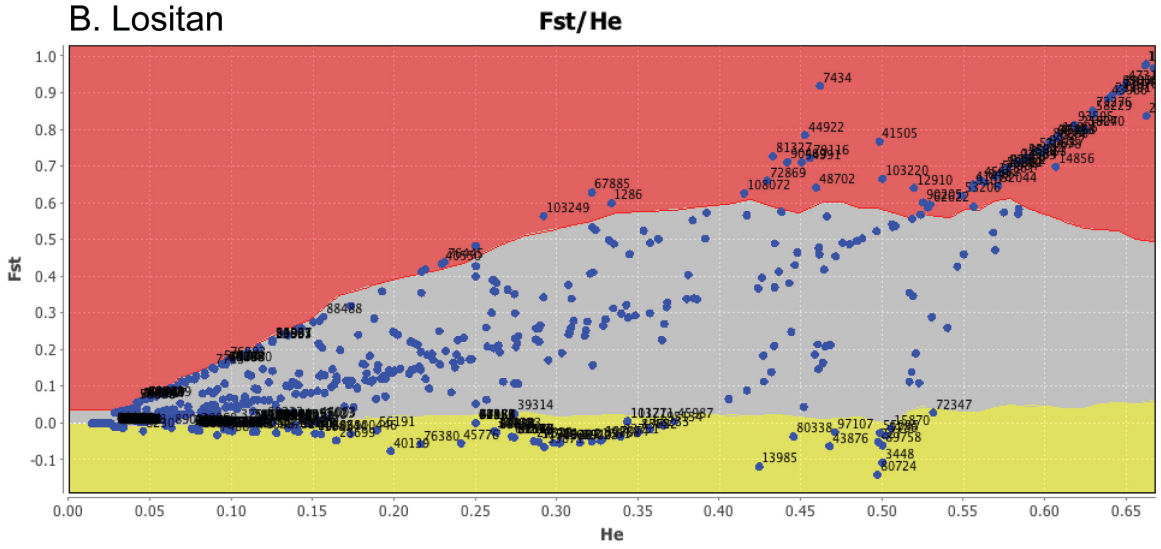
(Supplementary Table 1 Continued)

E – TMO-4C4: Between species F_{ST} = 0.883 (p=0.00)

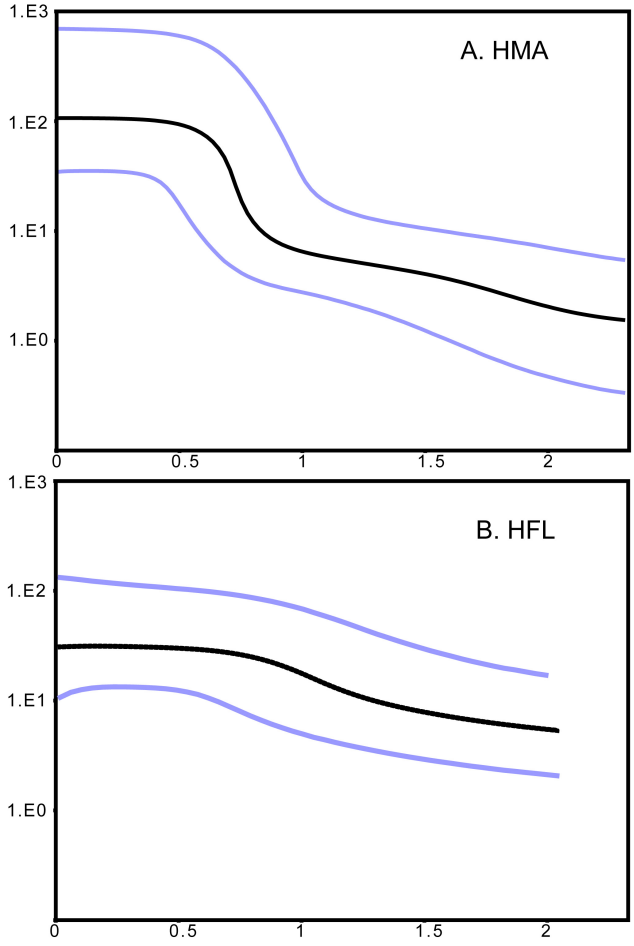
		1	2	3	4
Haemulon maculicauda	1. Mexico	--	0.239	<0.001	<0.001
	2. Panama	0.012	--	<0.001	<0.001
Haemulon flaviguttatum	3. Mexico	**0.867**	**0.904**	--	0.312
	4. Panama	**0.830**	**0.908**	0.107	--

Supplementary Figure 1. Results from the analysis of selection with 2422 polymorphic loci. (A) BayeScan revealed 33 loci under disruptive selection, while (B) Lositan revealed 81 loci under disruptive selection (red shaded area). Interestingly 85% (28) of the loci detected by BayeScan coincided with the loci detected by Lositan.





Supplementary Figure 2. Bayesian skyline plot of (A) *Haemulon maculicauda* (HMA) and (B) *Haemulon flaviguttagum* (HFL). The *X* axis represents time in million years and the *Y* axis represents effective population size.



APPENDIX B

Supplementary Table 1. Genes under positive selection in *Haemulon flavolineataum*, according to the branch-site model.

Isogroup	P values	Protein Name
isogroup4063	1.24E-12	Protein-L-isoaspartate O-methyltransferase domain-containing protein 1
isogroup13070	1.73E-12	Trans-2-enoyl-CoA reductase mitochondrial
isogroup9142	1.59E-11	High affinity cGMP-specific 35-cyclic phosphodiesterase 9A
isogroup1498	5.74E-11	SPRY domain-containing protein 3
isogroup15210	1.19E-10	SWI/SNF-related matrix-associated actin-dependent regulator of chromatin subfamily E member 1
isogroup35066	6.26E-10	WD repeat-containing protein 61
isogroup47632	6.50E-10	Histone deacetylase 1
isogroup55097	1.37E-09	UDP-glucuronosyltransferase 1-2
isogroup31000	1.56E-09	Autophagy-related protein 2 homolog B
isogroup15853	3.67E-09	Mitochondrial import inner membrane translocase subunit tim16
isogroup6646	3.79E-09	UPF0461 protein C5orf24 homolog
isogroup72849	1.12E-08	Inositol-trisphosphate 3-kinase B
isogroup43989	1.65E-08	Cyclin-related protein FAM58A
isogroup7971	1.95E-08	Angiopoietin-related protein 3
isogroup10529	3.09E-08	Ubiquitin-like protein 3
isogroup43558	3.57E-08	Transcriptional regulator ATRX
isogroup9083	4.48E-08	Solute carrier family 52 riboflavin transporter member 3
isogroup18954	4.61E-08	Pyruvate dehydrogenase E1 component subunit beta mitochondrial
isogroup30583	7.40E-08	Cell division cycle protein 16 homolog
isogroup28134	3.08E-07	Lysosomal acid lipase/cholesteryl ester hydrolase
isogroup55	5.04E-07	Dihydropyrimidine dehydrogenase [NADP(+)]
isogroup3524	6.51E-07	Protein SMG9
isogroup15198	6.95E-07	Small glutamine-rich tetratricopeptide repeat-containing protein alpha
isogroup53340	1.04E-06	Calcineurin B homologous protein 2

isogroup29495	1.96E-06	Ras-related protein Rab-31
isogroup28427	2.55E-06	Pre-mRNA-splicing factor 38A
isogroup51727	2.72E-06	L-seryl-tRNA(Sec) kinase
isogroup22500	2.75E-06	Golgi reassembly-stacking protein 1
isogroup67062	3.35E-06	Solute carrier family 52 riboflavin transporter member 3-A
isogroup76462	5.07E-06	Pre-mRNA-splicing factor SLU7
isogroup5060	7.05E-06	Sialidase-3
isogroup12879	1.31E-05	Probable arginine-tRNA ligase mitochondrial
isogroup36724	1.59E-05	Cullin-2
isogroup13806	2.21E-05	Nef-associated protein 1
isogroup63428	2.33E-05	Vesicular integral-membrane protein VIP36
isogroup62387	2.98E-05	Tripartite motif-containing protein 35
isogroup4211	3.10E-05	Probable ATP-dependent RNA helicase DHX58
isogroup26343	3.42E-05	Low-density lipoprotein receptor-related protein 1
isogroup12373	4.69E-05	Caprin-1
isogroup40295	5.13E-05	A-kinase anchor protein 2
isogroup14823	5.66E-05	Neuroepithelial cell-transforming gene 1 protein
isogroup51149	7.28E-05	Galectin-related protein
isogroup79362	8.94E-05	Cytochrome b-c1 complex subunit 1 mitochondrial
isogroup15643	0.000101392	Unconventional myosin-Ib
isogroup66876	0.00015506	Protein strawberry notch homolog 2
isogroup30227	0.000182672	Isoamyl acetate-hydrolyzing esterase 1 homolog
isogroup29697	0.000208542	Putative helicase mov-10-B.1
isogroup39129	0.000233565	Acyl-coenzyme A thioesterase 1
isogroup41269	0.000296487	Myelin expression factor 2
isogroup17572	0.000313172	Lipid phosphate phosphohydrolase 3
isogroup13411	0.000322067	Zinc finger protein ubi-d4
isogroup15325	0.000410803	Golgin subfamily B member 1
isogroup19016	0.00041948	Rho GTPase-activating protein 5
isogroup62629	0.000430303	Putative homeodomain transcription factor 1
isogroup23877	0.000526613	Phosphorylase b kinase regulatory subunit alpha skeletal muscle isoform
isogroup60440	0.0006585	Mediator of RNA polymerase II transcription subunit 12
isogroup34349	0.000683372	Toll-like receptor 1

isogroup959	0.000765464	ADP-ribosylation factor 1
isogroup43105	0.001061146	Protein RRP5 homolog
isogroup21306	0.001899907	RNA demethylase ALKBH5
isogroup2130	0.002824995	Chromodomain-helicase-DNA-binding protein 1
isogroup5405	0.003000283	Gastrula zinc finger protein XlCGF8.2DB (Fragment)
isogroup6331	0.003045672	5-hydroxyisourate hydrolase
isogroup17426	0.003917044	Poly [ADP-ribose] polymerase 4
isogroup372	0.004204496	Complement C1q-like protein 4
isogroup2043	0.004363505	Fibroblast growth factor receptor 1-A
isogroup40854	0.005563073	Protein kinase C eta type
isogroup85826	0.006036267	TBC1 domain family member 25
isogroup63940	0.010661991	Tyrosine-protein kinase BTK
isogroup57537	0.011204924	Cysteine desulfurase mitochondrial
isogroup8904	0.012262089	Large subunit GTPase 1 homolog
isogroup1194	0.013331028	HLA class II histocompatibility antigen DRB1-3 chain
isogroup41291	0.013675452	Adseverin
isogroup19300	0.015067176	Cytochrome c oxidase subunit 7A2 mitochondrial
isogroup24856	0.016723665	Tripartite motif-containing protein 7
isogroup57681	0.016964449	Archaemetzincin-2
isogroup1501	0.022700047	C-type lectin domain family 4 member E
isogroup3676	0.024915293	Myelin protein zero-like protein 1
isogroup45272	0.026818	Zinc finger protein 582
isogroup37406	0.02722111	Stonustoxin subunit beta
isogroup4081	0.027705309	Peptidyl-prolyl cis-trans isomerase
isogroup33281	0.029653675	Kinesin light chain 1
isogroup27684	0.032744467	Protein furry homolog
isogroup125	0.032949946	Amine sulfotransferase
isogroup58137	0.033393629	Histone-lysine N-methyltransferase SETD1B-A
isogroup4505	0.042837432	Tryptase-2
isogroup45953	0.045891626	LysM and putative peptidoglycan-binding domain-containing protein 4

Supplementary Table 2. Genes under positive selection for the branch of the sister species *Haemulon carbonarium* and *H. macrostomum*, according to the branch-site model.

Isogroup	P values	Protein name
isogroup19198	2.23E-11	Cyclin-C
isogroup21091	3.10E-11	FGFR1 oncogene partner 2 homolog
isogroup54282	4.03E-10	Spermine oxidase
isogroup36454	9.81E-10	Histone-lysine N-methyltransferase setd3
isogroup817	2.24E-09	N-acetylated-alpha-linked acidic dipeptidase-like protein
isogroup14880	4.47E-09	Sulfatase-modifying factor 1
isogroup1194	9.49E-09	HLA class II histocompatibility antigen DRB1-3 chain
isogroup3959	1.03E-08	Prosaposin
isogroup62425	1.50E-08	Protein FAM208B
isogroup15463	1.78E-08	KATNB1-like protein 1
isogroup5917	2.29E-08	2-oxo-4-hydroxy-4-carboxy-5-ureidoimidazoline decarboxylase
isogroup2602	2.83E-08	Mediator of RNA polymerase II transcription subunit 10
isogroup15025	3.30E-08	Calcineurin B homologous protein 3
isogroup19138	4.85E-08	Rhophilin-2
isogroup81363	4.92E-08	Transmembrane 9 superfamily member 4
isogroup31626	6.29E-08	C-X-C chemokine receptor type 3
isogroup40197	6.34E-08	Hydroxylysine kinase
isogroup12006	8.72E-08	ATP synthase subunit s-like protein
isogroup71567	1.03E-07	Peroxisomal N(1)-acetyl-spermine/spermidine oxidase
isogroup19190	2.46E-07	Nesprin-1
isogroup25800	2.85E-07	Gastrula zinc finger protein XlCGF49.1 (Fragment)
isogroup21230	4.06E-07	Insulin-like growth factor II
isogroup17585	4.42E-07	Tetratricopeptide repeat protein 27
isogroup58137	4.73E-07	Histone-lysine N-methyltransferase SETD1B-A
isogroup10877	7.98E-07	Mitochondrial ribonuclease P protein 3
isogroup24359	1.15E-06	Toll/interleukin-1 receptor domain-containing adapter protein
isogroup47458	1.31E-06	Succinate dehydrogenase [ubiquinone] flavoprotein subunit mitochondrial
isogroup47036	2.69E-06	RAC-alpha serine/threonine-protein kinase
isogroup6711	2.84E-06	Synaptobrevin homolog YKT6
isogroup8754	2.95E-06	Apoptosis-resistant E3 ubiquitin protein ligase 1
isogroup3487	6.42E-06	Glycogen phosphorylase liver form
isogroup61756	1.11E-05	Tubulin-specific chaperone E

isogroup60413	1.64E-05	Arfaptin-2
isogroup497	2.17E-05	Induced myeloid leukemia cell differentiation protein Mcl-1 homolog
isogroup7881	2.70E-05	F-box/WD repeat-containing protein 9
isogroup2373	2.98E-05	Integrin alpha-X
isogroup11585	3.13E-05	Zinc finger protein 706
isogroup11055	3.28E-05	Probable palmitoyltransferase ZDHHC24
isogroup12181	3.98E-05	UPF0661 TPR repeat-containing protein C16D10.01c
isogroup65122	4.06E-05	Importin subunit alpha-6
isogroup28158	4.08E-05	Transcription regulator protein BACH1
isogroup26056	5.18E-05	Kelch-like protein 25
isogroup16179	5.43E-05	UPF0462 protein C4orf33 homolog
isogroup14969	5.84E-05	Cornifelin homolog B
isogroup62208	6.06E-05	Bromodomain-containing protein 7
isogroup19002	6.26E-05	Sarcospan
isogroup73249	6.82E-05	Hepatocyte nuclear factor 4-beta
isogroup7750	7.92E-05	GDP-mannose 46 dehydratase
isogroup66876	9.04E-05	Protein strawberry notch homolog 2
isogroup10385	0.000103103	Vacuolar protein sorting-associated protein 72 homolog
isogroup57911	0.000103882	DNA repair protein RAD50
isogroup3913	0.000112698	Complement factor H
isogroup43558	0.000112736	Transcriptional regulator ATRX
isogroup4506	0.000121105	Far upstream element-binding protein 3
isogroup28244	0.000134486	TRMT1-like protein
isogroup36629	0.000158779	Cysteine and histidine-rich domain-containing protein 1
isogroup2328	0.000162043	Complement factor I
isogroup95769	0.000177275	Serine/threonine-protein kinase PLK2
isogroup29605	0.000177712	Ribonuclease P protein subunit p38
isogroup96319	0.000178273	Serine/threonine-protein kinase Sgk3
isogroup68120	0.000218052	Zinc-binding protein A33
isogroup13106	0.000231016	Double-strand-break repair protein rad21 homolog
isogroup71269	0.000250744	Structural maintenance of chromosomes protein 1A
isogroup9326	0.000335395	DnaJ homolog subfamily C member 2
isogroup48144	0.000453362	Ig kappa chain V region Mem5 (Fragment)
isogroup51967	0.000479034	Endothelial PAS domain-containing protein 1
isogroup24856	0.000504508	Tripartite motif-containing protein 7
isogroup8056	0.000574277	Alpha-ketoglutarate-dependent dioxygenase alkB homolog 6
isogroup37406	0.000670523	Stonustoxin subunit beta
isogroup4713	0.000705934	Coagulation factor VIII

isogroup50939	0.000727877	Tripartite motif-containing protein 29
isogroup4505	0.000847103	Tryptase-2
isogroup9406	0.000858928	Glomulin
isogroup29425	0.000965261	Golgi-associated plant pathogenesis-related protein 1
isogroup20430	0.001086027	Histone-lysine N-methyltransferase EZH1
isogroup7395	0.001107775	Ribonuclease H2 subunit A
isogroup15406	0.001110966	Potassium channel subfamily K member 5
isogroup12802	0.001215785	NAD(P) transhydrogenase mitochondrial
isogroup6096	0.001248926	Butyrophilin subfamily 2 member A1
isogroup27643	0.001328261	H-2 class II histocompatibility antigen A-B alpha chain
isogroup62102	0.001414281	Structural maintenance of chromosomes protein 1A
isogroup30490	0.001482993	Ribosome-recycling factor mitochondrial
isogroup5905	0.001678721	Probable cytosolic iron-sulfur protein assembly protein ciao1-B
isogroup61083	0.001696583	Zinc finger protein 778
isogroup27026	0.001806829	C4b-binding protein alpha chain
isogroup15152	0.001908635	Meckel syndrome type 1 protein
isogroup14446	0.001912433	Galectin-3
isogroup11902	0.001946166	Nucleolar protein 11-like
isogroup2043	0.001960478	Fibroblast growth factor receptor 1-A
isogroup36363	0.001986318	Protein spinster homolog 2
isogroup9945	0.002031897	Proliferating cell nuclear antigen
isogroup19832	0.002099704	EH domain-binding protein 1
isogroup7864	0.002440652	Leucine-rich repeat-containing protein C10orf11 homolog
isogroup12213	0.002611234	Serine/threonine-protein kinase Sgk3
isogroup11665	0.00299479	Furin
isogroup78602	0.003089481	Small integral membrane protein 12
isogroup56484	0.003114736	Microtubule-actin cross-linking factor 1
isogroup22064	0.003119954	Ubiquitin-conjugating enzyme E2 Q2
isogroup68409	0.00365546	Transcription initiation factor TFIID subunit 11
isogroup70095	0.003760004	Paired amphipathic helix protein Sin3a
isogroup13494	0.00381359	FAD-linked sulfhydryl oxidase ALR
isogroup3216	0.004262494	Glycoprotein-N-acetylgalactosamine 3-beta-galactosyltransferase 1-B
isogroup1217	0.004572125	Transmembrane protein 26
isogroup84606	0.005013718	Cyclin-I
isogroup30253	0.005047698	Acid trehalase-like protein 1
isogroup13290	0.005515533	2-deoxynucleoside 5-phosphate N-hydrolase 1
isogroup108664	0.006135001	Glomulin
isogroup2156	0.00673689	Nucleolar protein 8

isogroup4211	0.007310927	Probable ATP-dependent RNA helicase DHX58
isogroup786	0.007569556	Glycoprotein integral membrane protein 1
isogroup33906	0.007625397	Acyl-coenzyme A thioesterase 11
isogroup100139	0.007862362	Macrophage-expressed gene 1 protein
isogroup19106	0.00794555	39S ribosomal protein L43 mitochondrial
isogroup8625	0.009047593	Atrial natriuretic peptide receptor 2
isogroup2180	0.009107673	Transmembrane protein 50A
isogroup32919	0.009117656	Adenosine deaminase-like protein
isogroup21887	0.009153434	Eukaryotic elongation factor 2 kinase
isogroup27627	0.009415021	Solute carrier family 35 member F5
isogroup31077	0.009950874	DNA mismatch repair protein Msh3
isogroup10524	0.010784222	Amphiphysin
isogroup8056	0.011603394	Alpha-ketoglutarate-dependent dioxygenase alkB homolog 6
isogroup8452	0.012653662	Hexosaminidase D
isogroup1950	0.013736298	Mitochondrial import inner membrane translocase subunit Tim21
isogroup48121	0.014068861	Hepcidin
isogroup48769	0.014487933	Protein FAM115
isogroup20764	0.016065216	Tissue factor
isogroup21906	0.01707088	Pumilio homolog 2
isogroup1073	0.018009714	Hemolytic toxin Avt-1
isogroup51931	0.019211879	Tyrosine-protein kinase ITK/TSK
isogroup1408	0.019879992	NADH-cytochrome b5 reductase 2
isogroup5544	0.019924356	Zona pellucida sperm-binding protein 3
isogroup489	0.020874689	Cytochrome P450 2W1
isogroup9673	0.02245935	Transcription termination factor 3 mitochondrial
isogroup16891	0.023697321	Stress-70 protein mitochondrial
isogroup21317	0.023977378	Rho GTPase-activating protein 29
isogroup3698	0.025465643	Acyl-coenzyme A thioesterase 2 mitochondrial
isogroup5060	0.025549373	Sialidase-3
isogroup25750	0.027938571	WD repeat-containing protein 5
isogroup89718	0.028744544	Transforming growth factor beta regulator 1
isogroup1170	0.030057651	Stonustoxin subunit beta
isogroup1697	0.035001868	Trans-12-dihydrobenzene-12-diol dehydrogenase
isogroup83713	0.037245601	N-alpha-acetyltransferase 15 NatA auxiliary subunit
isogroup68086	0.039498124	RNA-binding protein 28
isogroup391	0.042166312	3-oxo-5-beta-steroid 4-dehydrogenase
isogroup41237	0.042694892	Regulating synaptic membrane exocytosis protein 2
isogroup1167	0.04294144	Leukocyte elastase inhibitor
isogroup1386	0.044798474	Leucine-rich alpha-2-glycoprotein

isogroup35836	0.046809687	DDB1- and CUL4-associated factor 6
isogroup4266	0.04687354	Transcription factor E2F6
isogroup47632	0.047663821	Histone deacetylase 1

Supplementary Table 3. Genes under positive selection for the branch of *Haemulon carbonarium*, according to the branch-site model.

Isogroup	P values	Protein name
isogroup5917	2.34E-12	2-oxo-4-hydroxy-4-carboxy-5-ureidoimidazoline decarboxylase
isogroup817	2.78E-11	N-acetylated-alpha-linked acidic dipeptidase-like protein
isogroup19138	1.98E-10	Rhophilin-2
isogroup54282	4.03E-10	Spermine oxidase
isogroup73249	2.31E-08	Hepatocyte nuclear factor 4-beta
isogroup2328	3.73E-08	Complement factor I
isogroup71567	6.31E-08	Peroxisomal N(1)-acetyl-spermine/spermidine oxidase
isogroup40197	6.34E-08	Hydroxylysine kinase
isogroup9406	1.47E-07	Glomulin
isogroup12181	2.33E-07	UPF0661 TPR repeat-containing protein C16D10.01c
isogroup62425	2.58E-07	Protein FAM208B
isogroup61083	2.75E-07	Zinc finger protein 778
isogroup1194	4.37E-07	HLA class II histocompatibility antigen DRB1-3 chain
isogroup21230	5.36E-07	Insulin-like growth factor II
isogroup57911	5.77E-07	DNA repair protein RAD50
isogroup11055	6.66E-07	Probable palmitoyltransferase ZDHHC24
isogroup58924	1.15E-06	Endothelial zinc finger protein induced by tumor necrosis factor alpha
isogroup4506	1.25E-06	Far upstream element-binding protein 3
isogroup36629	2.20E-06	Cysteine and histidine-rich domain-containing protein 1
isogroup65122	2.97E-06	Importin subunit alpha-6
isogroup62208	3.51E-06	Bromodomain-containing protein 7
isogroup2156	4.48E-06	Nucleolar protein 8
isogroup10385	8.92E-06	Vacuolar protein sorting-associated protein 72 homolog
isogroup61756	1.12E-05	Tubulin-specific chaperone E
isogroup60413	1.64E-05	Arfaptin-2
isogroup30253	2.49E-05	Acid trehalase-like protein 1
isogroup29605	2.78E-05	Ribonuclease P protein subunit p38
isogroup27627	3.06E-05	Solute carrier family 35 member F5
isogroup16179	5.43E-05	UPF0462 protein C4orf33 homolog
isogroup28158	5.66E-05	Transcription regulator protein BACH1
isogroup125	6.00E-05	Amine sulfotransferase
isogroup19002	6.26E-05	Sarcospan
isogroup2373	6.64E-05	Integrin alpha-X

isogroup13290	7.78E-05	2-deoxynucleoside 5-phosphate N-hydrolase 1
isogroup7750	7.92E-05	GDP-mannose 46 dehydratase
isogroup20430	0.000112112	Histone-lysine N-methyltransferase EZH1
isogroup9326	0.000114821	DnaJ homolog subfamily C member 2
isogroup21887	0.000116431	Eukaryotic elongation factor 2 kinase
isogroup21091	0.000150259	FGFR1 oncogene partner 2 homolog
isogroup51931	0.000174474	Tyrosine-protein kinase ITK/TSK
isogroup29425	0.000177238	Golgi-associated plant pathogenesis-related protein 1
isogroup1217	0.000252674	Transmembrane protein 26
isogroup2043	0.000327966	Fibroblast growth factor receptor 1-A
isogroup100139	0.000482426	Macrophage-expressed gene 1 protein
isogroup78602	0.000515938	Small integral membrane protein 12
isogroup8056	0.000574321	Alpha-ketoglutarate-dependent dioxygenase alkB homolog 6
isogroup786	0.000679622	Glycoprotein integral membrane protein 1
isogroup8625	0.000836039	Atrial natriuretic peptide receptor 2
isogroup11665	0.000934321	Furin
isogroup15406	0.001111012	Potassium channel subfamily K member 5
isogroup89718	0.001158354	Transforming growth factor beta regulator 1
isogroup27026	0.001314157	C4b-binding protein alpha chain
isogroup40792	0.001692138	UMP-CMP kinase
isogroup16891	0.001812549	Stress-70 protein mitochondrial
isogroup27643	0.001895118	H-2 class II histocompatibility antigen A-B alpha chain
isogroup24856	0.00198238	Tripartite motif-containing protein 7
isogroup9945	0.002032007	Proliferating cell nuclear antigen
isogroup917	0.002808711	Properdin
isogroup35836	0.002891757	DDB1- and CUL4-associated factor 6
isogroup2180	0.004229186	Transmembrane protein 50A
isogroup985	0.005215033	Charged multivesicular body protein 3
isogroup4094	0.005365939	Cytosolic sulfotransferase 3
isogroup11447	0.005833678	Methionyl-tRNA formyltransferase mitochondrial
isogroup14567	0.008648621	Prostaglandin reductase 1
isogroup32919	0.009118361	Adenosine deaminase-like protein
isogroup8754	0.009165563	Apoptosis-resistant E3 ubiquitin protein ligase 1
isogroup31077	0.011035647	DNA mismatch repair protein Msh3
isogroup77262	0.011067916	E3 ubiquitin-protein ligase TRIM39
isogroup3698	0.011375649	Acyl-coenzyme A thioesterase 2 mitochondrial
isogroup57705	0.01315309	Xaa-Pro aminopeptidase 1
isogroup9110	0.013264208	Leukocyte elastase inhibitor
isogroup1170	0.017371944	Stonustoxin subunit beta

isogroup45817	0.020072915	Plastin-2
isogroup20764	0.020652154	Tissue factor
isogroup489	0.024270444	Cytochrome P450 2W1
isogroup3247	0.025366922	Acyl-CoA synthetase family member 2 mitochondrial
isogroup13649	0.02605963	Alpha-1-macroglobulin
isogroup50939	0.026519535	Tripartite motif-containing protein 29
isogroup5544	0.037687426	Zona pellucida sperm-binding protein 3

Supplementary Table 4. Genes under positive selection for the branch of *Haemulon macrostomum*, according to the branch-site model.

Isogroup	P values	Protein name
isogroup19198	2.23E-11	Cyclin-C
isogroup12006	3.34E-10	ATP synthase subunit s-like protein
isogroup21091	4.48E-10	FGFR1 oncogene partner 2 homolog
isogroup31626	3.17E-09	C-X-C chemokine receptor type 3
isogroup2602	3.47E-09	Mediator of RNA polymerase II transcription subunit 10
isogroup14880	4.47E-09	Sulfatase-modifying factor 1
isogroup15025	6.58E-09	Calcineurin B homologous protein 3
isogroup3959	1.03E-08	Prosaposin
isogroup15463	1.78E-08	KATNB1-like protein 1
isogroup36454	1.91E-08	Histone-lysine N-methyltransferase setd3
isogroup81363	4.92E-08	Transmembrane 9 superfamily member 4
isogroup3487	6.25E-08	Glycogen phosphorylase liver form
isogroup10877	7.07E-08	Mitochondrial ribonuclease P protein 3
isogroup24359	1.17E-07	Toll/interleukin-1 receptor domain-containing adapter protein
isogroup19190	2.46E-07	Nesprin-1
isogroup47458	2.61E-07	Succinate dehydrogenase [ubiquinone] flavoprotein subunit mitochondrial
isogroup25800	2.85E-07	Gastrula zinc finger protein XlCGF49.1 (Fragment)
isogroup17585	4.42E-07	Tetratricopeptide repeat protein 27
isogroup6711	4.81E-07	Synaptobrevin homolog YKT6
isogroup58924	6.51E-07	Endothelial zinc finger protein induced by tumor necrosis factor alpha
isogroup6096	1.70E-06	Butyrophilin subfamily 2 member A1
isogroup7864	2.15E-06	Leucine-rich repeat-containing protein C10orf11 homolog
isogroup47036	2.69E-06	RAC-alpha serine/threonine-protein kinase
isogroup8754	2.95E-06	Apoptosis-resistant E3 ubiquitin protein ligase 1
isogroup497	3.99E-06	Induced myeloid leukemia cell differentiation protein Mcl-1 homolog
isogroup14969	4.84E-06	Cornifelin homolog B
isogroup13106	1.52E-05	Double-strand-break repair protein rad21 homolog
isogroup7881	1.59E-05	F-box/WD repeat-containing protein 9
isogroup28244	2.52E-05	TRMT1-like protein
isogroup11585	3.13E-05	Zinc finger protein 706
isogroup3913	4.69E-05	Complement factor H
isogroup26056	5.18E-05	Kelch-like protein 25

isogroup66876	9.04E-05	Protein strawberry notch homolog 2
isogroup4713	9.73E-05	Coagulation factor VIII
isogroup62425	0.000104468	Protein FAM208B
isogroup43558	0.000112746	Transcriptional regulator ATRX
isogroup95769	0.000177275	Serine/threonine-protein kinase PLK2
isogroup71269	0.000184021	Structural maintenance of chromosomes protein 1A
isogroup11902	0.000187046	Nucleolar protein 11-like
isogroup7395	0.000245485	Ribonuclease H2 subunit A
isogroup36363	0.000311952	Protein spinster homolog 2
isogroup12802	0.00031333	NAD(P) transhydrogenase mitochondrial
isogroup5905	0.000355955	Probable cytosolic iron-sulfur protein assembly protein ciao1-B
isogroup1167	0.000405244	Leukocyte elastase inhibitor
isogroup33906	0.000416691	Acyl-coenzyme A thioesterase 11
isogroup15152	0.000442329	Meckel syndrome type 1 protein
isogroup51967	0.000479002	Endothelial PAS domain-containing protein 1
isogroup96319	0.00049938	Serine/threonine-protein kinase Sgk3
isogroup9475	0.00080378	TGF-beta receptor type-2
isogroup41237	0.001148742	Regulating synaptic membrane exocytosis protein 2
isogroup108664	0.00126995	Glomulin
isogroup62102	0.001414444	Structural maintenance of chromosomes protein 1A
isogroup30490	0.001483126	Ribosome-recycling factor mitochondrial
isogroup8056	0.00168493	Alpha-ketoglutarate-dependent dioxygenase alkB homolog 6
isogroup84606	0.001818042	Cyclin-I
isogroup19832	0.002099537	EH domain-binding protein 1
isogroup4211	0.002334485	Probable ATP-dependent RNA helicase DHX58
isogroup12213	0.002610634	Serine/threonine-protein kinase Sgk3
isogroup1073	0.002623967	Hemolytic toxin Avt-1
isogroup9165	0.002875571	PC4 and SFRS1-interacting protein
isogroup56484	0.003114985	Microtubule-actin cross-linking factor 1
isogroup22064	0.003119991	Ubiquitin-conjugating enzyme E2 Q2
isogroup68409	0.003654853	Transcription initiation factor TFIID subunit 11
isogroup13494	0.003814283	FAD-linked sulfhydryl oxidase ALR
isogroup68120	0.004163697	Zinc-binding protein A33
isogroup1980	0.004253606	High mobility group protein B2
isogroup3216	0.004263491	Glycoprotein-N-acetylgalactosamine 3-beta-galactosyltransferase 1-B
isogroup27026	0.005259415	C4b-binding protein alpha chain
isogroup70095	0.005374717	Paired amphipathic helix protein Sin3a
isogroup83713	0.005828322	N-alpha-acetyltransferase 15 NatA auxiliary subunit

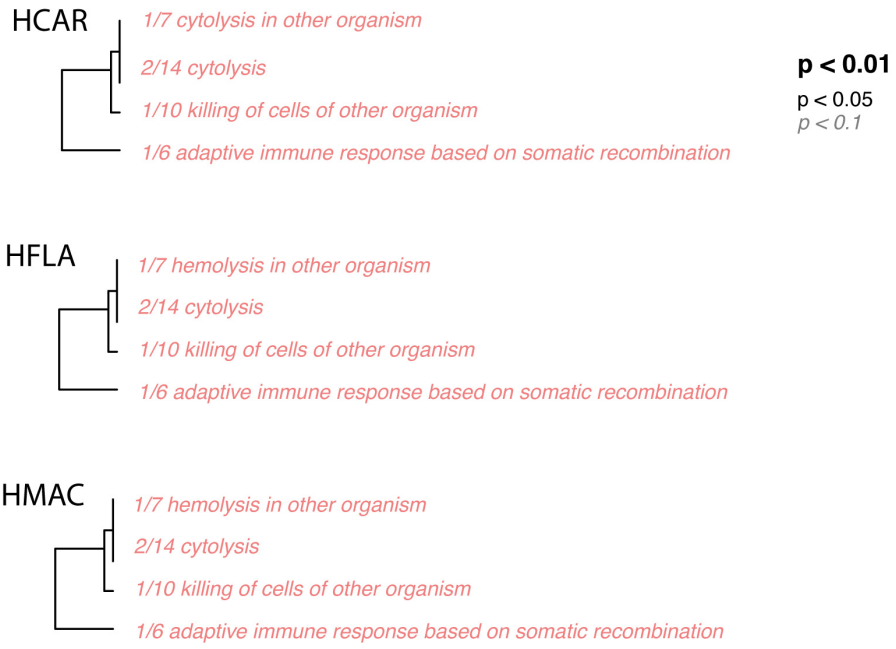
isogroup817	0.006207633	N-acetylated-alpha-linked acidic dipeptidase-like protein
isogroup68086	0.007128404	RNA-binding protein 28
isogroup19106	0.007944778	39S ribosomal protein L43 mitochondrial
isogroup5164	0.008928307	TIR domain-containing adapter molecule 1
isogroup9097	0.009217518	Growth arrest and DNA damage-inducible protein GADD45 gamma
isogroup10524	0.010782245	Amphiphysin
isogroup14446	0.010794538	Galectin-3
isogroup1697	0.012073661	Trans-12-dihydrobenzene-12-diol dehydrogenase
isogroup21906	0.017072534	Pumilio homolog 2
isogroup71567	0.017249883	Peroxisomal N(1)-acetyl-spermine/spermidine oxidase
isogroup10805	0.019751674	DNA primase small subunit
isogroup10	0.022161893	Complement factor B
isogroup5060	0.02554825	Sialidase-3
isogroup25750	0.027938668	WD repeat-containing protein 5
isogroup1408	0.035587924	NADH-cytochrome b5 reductase 2
isogroup1194	0.042084962	HLA class II histocompatibility antigen DRB1-3 chain
isogroup48121	0.046205468	Hepcidin
isogroup47632	0.048500789	Histone deacetylase 1

Supplementary Table 5. List of genes under both positive selection with the *dN/dS* analysis and differential gene expression between the three focal species.

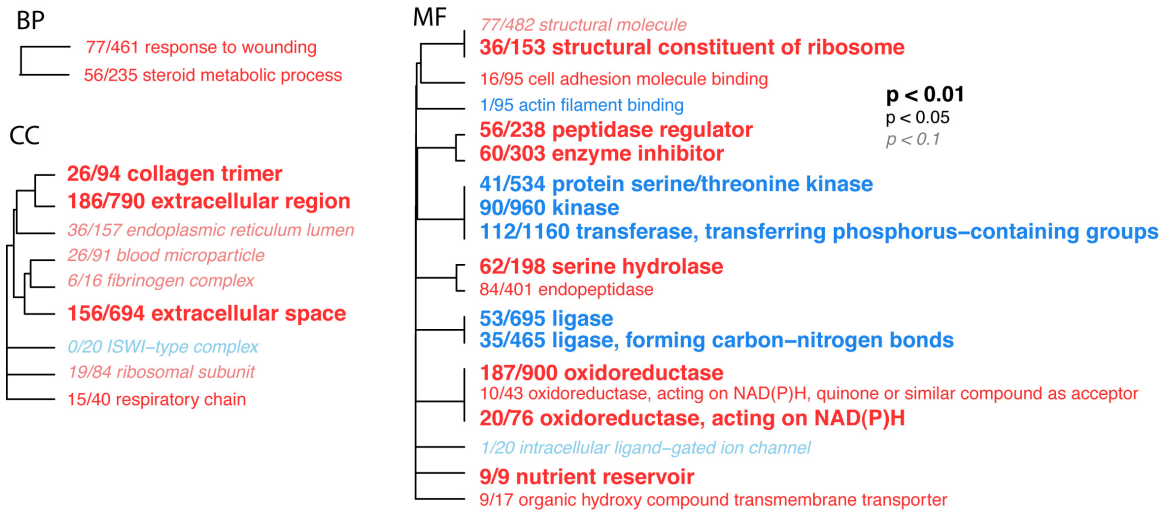
Isogroup	P value (dN/dS)	P value (DGE)	Protein
isogroup5917	2.29E-08	0.002246216	2-oxo-4-hydroxy-4-carboxy-5-ureidoimidazoline decarboxylase
isogroup391	0.042166312	5.57E-18	3-oxo-5-beta-steroid 4-dehydrogenase
isogroup6331	0.003045672	4.69E-05	5-hydroxyisourate hydrolase
isogroup40295	5.13E-05	1.04E-10	A-kinase anchor protein 2
isogroup3698	0.011375649	0.016492978	Acyl-coenzyme A thioesterase 2 mitochondrial
isogroup32919	0.009117656	0.005122631	Adenosine deaminase-like protein
isogroup7971	1.95E-08	2.88E-06	Angiopoietin-related protein 3
isogroup6096	0.001248926	0.005313713	Butyrophilin subfamily 2 member A1
isogroup1501	0.022700047	2.31E-05	C-type lectin domain family 4 member E
isogroup27026	0.001806829	3.26E-05	C4b-binding protein alpha chain
isogroup372	0.004204496	0.000159037	Complement C1q-like protein 4
isogroup10	0.022161893	6.69E-16	Complement factor B
isogroup3913	0.000112698	0.003804668	Complement factor H
isogroup2328	3.73E-08	0.000354231	Complement factor I
isogroup489	0.024270444	1.24E-24	Cytochrome P450 2W1
isogroup4094	0.005365939	1.15E-11	Cytosolic sulfotransferase 3
isogroup55	5.04E-07	7.01E-05	Dihydropyrimidine dehydrogenase [NADP(+)]
isogroup2043	0.001960478	7.21E-08	Fibroblast growth factor receptor 1-A
isogroup14446	0.001912433	1.13E-05	Galectin-3
isogroup1073	0.018009714	8.16E-05	Hemolytic toxin Avt-1
isogroup2373	2.98E-05	2.19E-09	Integrin alpha-X
isogroup1386	0.044798474	2.56E-05	Leucine-rich alpha-2-glycoprotein
isogroup1167	0.000405244	2.12E-05	Leukocyte elastase inhibitor
isogroup12802	0.001215785	8.85E-07	NAD(P) transhydrogenase mitochondrial
isogroup1408	0.035587924	0.005341632	NADH-cytochrome b5 reductase 2
isogroup15406	0.001110966	5.14E-13	Potassium channel subfamily K member 5
isogroup4211	0.002334485	0.000830053	Probable ATP-dependent RNA helicase DHX58
isogroup11055	6.66E-07	0.001424265	Probable palmitoyltransferase

			ZDHHC24
isogroup9945	0.002031897	1.73E-10	Proliferating cell nuclear antigen
isogroup62425	1.50E-08	0.005722992	Protein FAM208B
isogroup19138	4.85E-08	1.30E-07	Rhophilin-2
isogroup9083	4.48E-08	0.006071212	Solute carrier family 52 riboflavin transporter 3
isogroup54282	4.03E-10	0.011830198	Spermine oxidase
isogroup1498	5.74E-11	2.51E-05	SPRY domain-containing protein 3
isogroup2180	0.004229186	0.000437625	Transmembrane protein 50A
isogroup24856	0.016723665	2.86E-13	Tripartite motif-containing protein 7
isogroup4505	0.042837432	0.000534776	Tryptase-2

Supplementary Figure 1. GO enrichment analysis of genes under positive selection (FDR≤ 0.10) corresponding to the category Biological Processes, for the individual branches of *Haemulon carbonarium* (HCAR), *H. flavolineatum* (HFLA) and *H. macrostomum* (HMAC).

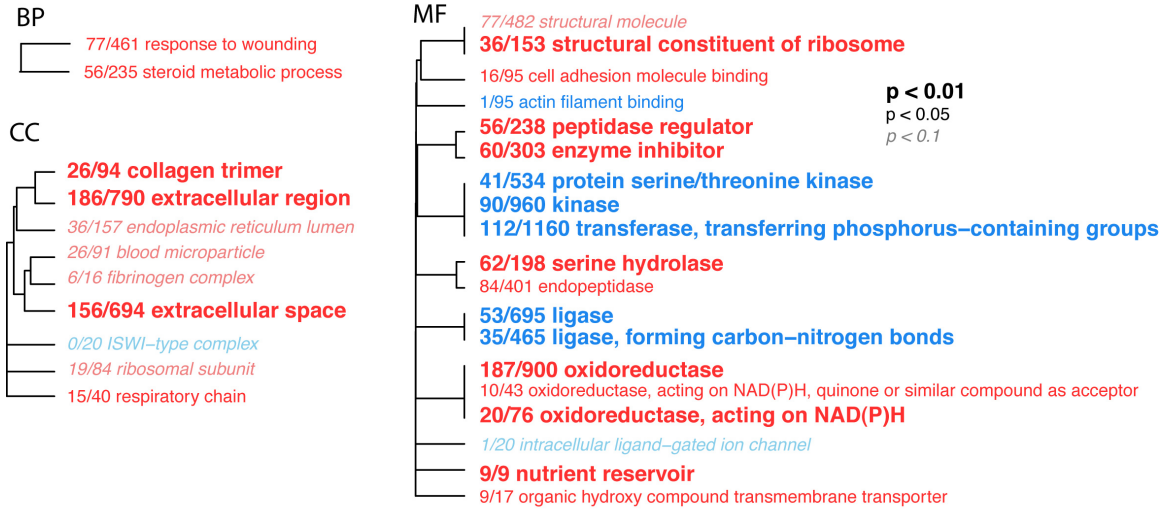


Supplementary Figure 2. GO enrichment analysis for genes under positive selection (FDR≤ 0.10) for the categories Biological Processes (BP), Cellular Components (CC) and Molecular Functions (MF), for the branch of the sister species *Haemulon carbonarium* and *H. macrostomum*. The size of the fonts represents the significance of the *P*-value.

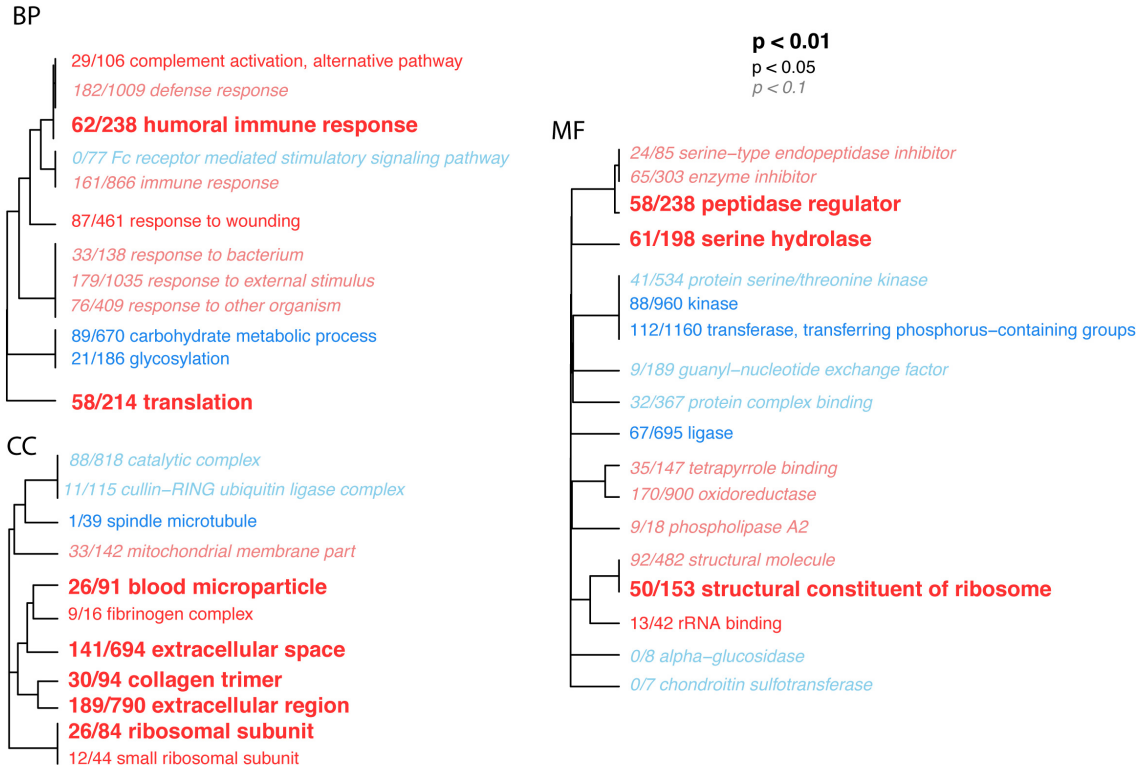


Supplementary Figure 3. Heatmap of annotated genes significantly differentiated between *Haemulon carbonarium*, *H. flavolineatum* and *H. macrostomum*.

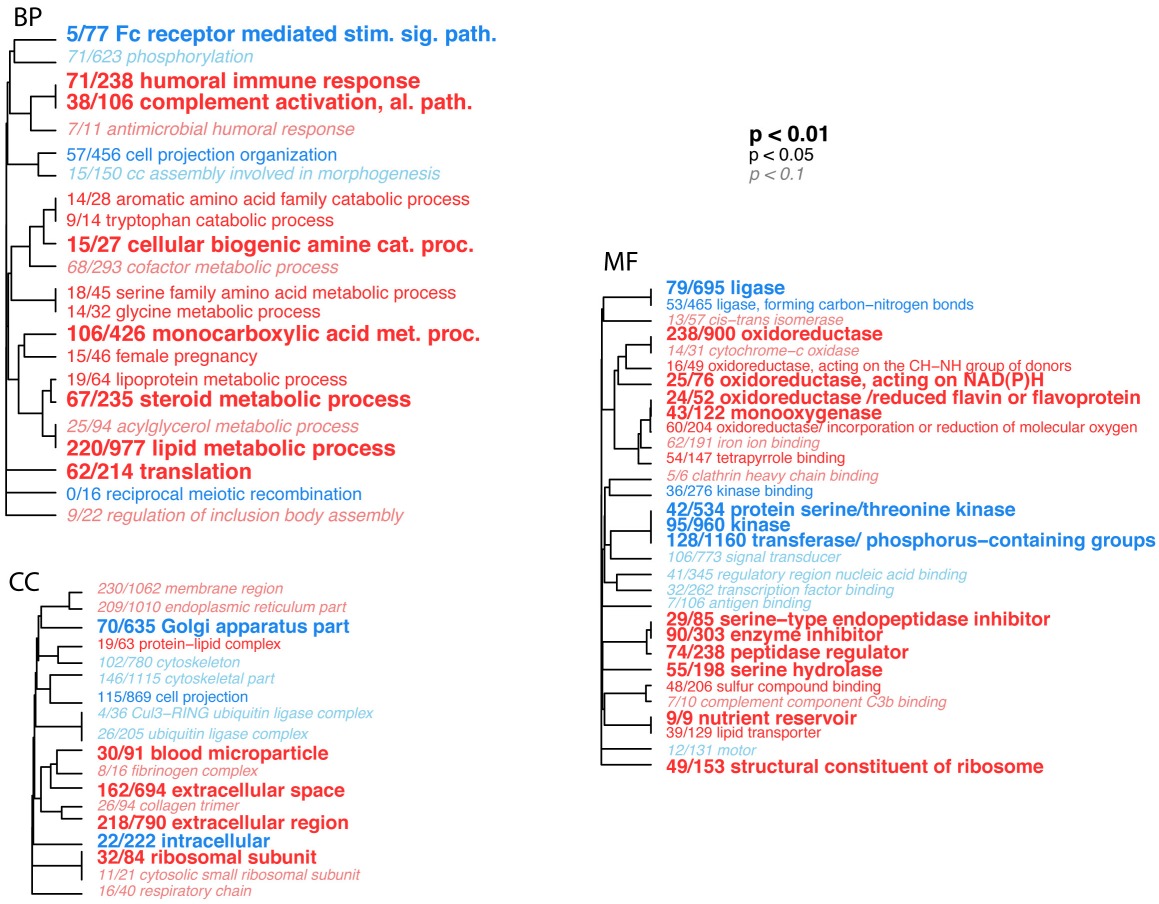
Supplementary Figure 4. GO enrichment analysis for differentially expressed genes between *Haemulon carbonarium* and *H. flavolineatum* (FDR≤ 0.10) for the three categories Biological Processes (BP), Cellular Components (CC) and Molecular Functions (MF). Names in red correspond to up-regulated genes, while blue represents down-regulated ones. The size of the fonts represents the significance of the *P*-value.



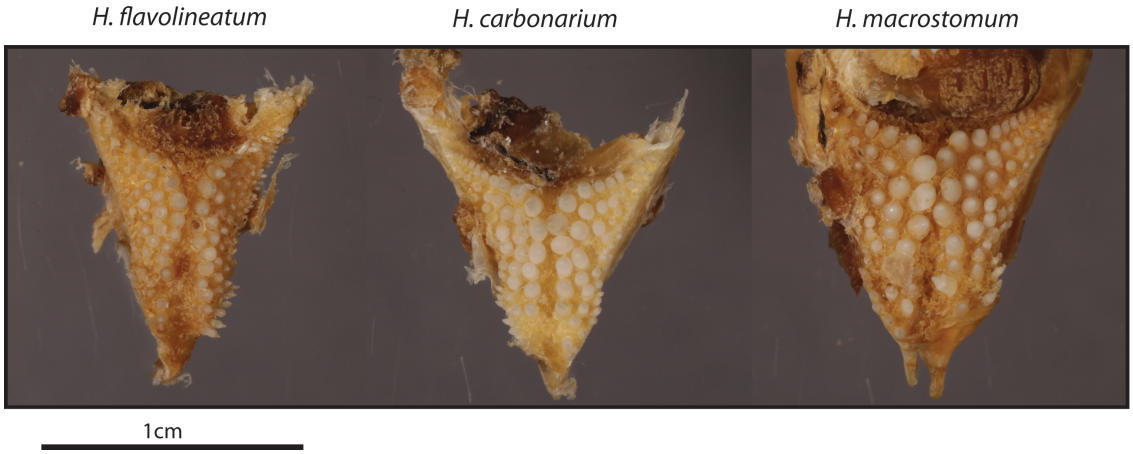
Supplementary Figure 5. GO enrichment analysis for differentially expressed genes between sister species *Haemulon carbonarium* and *H. macrostomum* (FDR ≤0.10) for the three categories Biological Processes (BP), Cellular Components (CC) and Molecular Functions (MF). Names in red correspond to up-regulated genes, while blue represents down-regulated ones. The size of the fonts represents the significance of the *P*-value.



Supplementary Figure 6. GO enrichment analysis for differentially expressed genes between the species with the highest number of significant differentially expressed genes, *Haemulon flavolineatum* and *H. macrostomum* (FDR ≤0.10). Figures represent the three GO categories: Biological Processes (BP), Cellular Components (CC) and Molecular Functions (MF). Names in red correspond to up-regulated genes, while blue represents down-regulated ones. The size of the fonts represents the significance of the *P*-value.



Supplementary Figure 7. Lower pharyngeal teeth of *Haemulon* flavolineatum, *H. carbonarium* and *H. macrostomum* exemplify differences in the pharyngeal apparatus of the three species.



APPENDIX C

Manuscripts authored and co-authored by Moises Antonio Bernal during the span of his PhD, relevant to the evolution and ecology of marine fishes:

1. Gaither MR, **MA Bernal**, I Fernandez-Silva, M Mwale, SA Jones, C Rocha and LA Rocha (2015) Two deep evolutionary lineages in the circumglobal Glasseye, *Heteropriacanthus cruentatus* (Teleostei, Priacanthidae) with admixture in the Western Indian Ocean. Journal of Fish Biology, 87, 715-727. doi:10.1111/jfb.12754
2. Sellas AB, K Bassos-Hull, JC Perez-Jimenez, JA Angulo-Valdes, **MA Bernal** and RE Hurt (2015) Population structure and seasonal migration of the spotted eagle ray, *Aetobatus narinari*. Journal of Heredity, 106, 266-275. doi: 10.1093/jhered/esv011
3. Gaither MR, **MA Bernal**, R Coleman, B Bowen, S Jones, WB Simison and LA Rocha (2015) Genomic signatures of geographic isolation and natural selection in coral reef fishes. Molecular Ecology. doi: 10.1111/mec.13129.
4. **Bernal MA**, SR Floeter, MR Gaither, GO Longo, R Morais, CEL Ferreira, MJA Vermeij and LA Rocha (2015) High prevalence of dermal parasites among coral reef fishes of Curacao. Marine Biodiversity. doi: 10.1007/s12526-015-0322-z
5. **Bernal MA**, NL Sinai, C Rocha, MR Gaither, F Dunker and LA Rocha (2014) Long-term sperm storage in the Brownbanded Bamboo Shark *Chiloscyllium punctatum* Müller & Henle, 1838 (Chondrichthyes: Hesmiscyliidae). Journal of Fish Biology 86(3): 1171-1176. doi: 10.1111/jfb.12606.
6. Selkoe KA, OE Gaggiotti, K Andrews, **MA Bernal**, *et al*. (2014) Emergent Patterns of Population Genetic Structure for a Coral Reef Community. Molecular Ecology. doi: 10.1111/mec.12804
7. Rocha LA, **MA Bernal**, MR Gaither and ME Alfaro (2013) Massively parallel DNA sequencing: the new frontier in biogeography. Frontiers of Biogeography, 5.1.
8. Ludt W, **MA Bernal**, BW Bowen and LA Rocha (2012) Living in the Past: Phylogeography and Population Histories of Indo-Pacific Wrasses (genus *Halichoeres*) in Shallow Lagoons versus Outer Reef Slopes. PLoS One, 7, e38042. doi: 10.1371/journal.pone.0038042
9. **Bernal MA** and LA Rocha (2011) *Acanthurus tractus* Poey, 1860, a valid Western Atlantic species of surgeonfish (Teleostei, Acanthuridae), distinct from Acanthurus bahianus Castelnau, 1855. Zootaxa, 2905, 63–68.

References

Ahel I, Ahel D, Matsusaka T et al. (2008) Poly(ADP-ribose)-binding zinc finger motifs in DNA repair/checkpoint proteins. Nature, 451, 81–85.

Albertson RC, Streelman JT, Kocher TD (2003) Directional selection has shaped the oral jaws of Lake Malawi cichlid fishes. Proceedings of the National Academy of Sciences, 100, 5252–5257.

Allender CJ, Seehausen O, Knight ME, Turner GF, Maclean N (2003) Divergent selection during speciation of Lake Malawi cichlid fishes inferred from parallel radiations in nuptial coloration. Proceedings of the National Academy of Sciences, 100, 14074–14079.

Antao T, Lopes A, Lopes RJ, Beja-Pereira A, Luikart G (2008) LOSITAN: A workbench to detect molecular adaptation based on a F_{ST} -outlier method. BMC Bioinformatics, 9, 323.

Arnold ML, Fogarty ND (2009) Reticulate evolution and marine organisms: the final frontier? International journal of molecular sciences, 10, 3836–3860.

Avise JC, Saunders NC (1984) Hybridization and Introgression among Species of Sunfish (*Lepomis*): Analysis by Mitochondrial DNA and Allozyme Markers. Genetics, 108, 237–255.

Bachtrog D, Thornton K, Clark A, Andolfatto P (2006) Extensive Introgression of Mitochondrial Dna Relative to Nuclear Genes in the *Drosophila yakuba* Species Group. Evolution, 60, 292–302.

Baldo L, Santos ME, Salzburger W (2011) Comparative Transcriptomics of Eastern African Cichlid Fishes Shows Signs of Positive Selection and a Large Contribution of Un-translated Regions to Genetic Diversity. Genome Biology and Evolution, 3, 443–455.

Ballard JWO, Whitlock MC (2004) The incomplete natural history of mitochondria. Molecular Ecology, 13, 729–744.

Barrett RDH, Rogers SM, Schluter D (2008) Natural Selection on a Major Armor Gene in Threespine Stickleback. Science, 322, 255–257.

Bernal MA, Rocha LA (2011) Acanthurus tractus Poey, 1860, a valid western Atlantic species of surgeonfish (Teleostei, Acanthuridae), distinct from *Acanthurus bahianus* Castelnau, 1855. Zootaxa, 2905, 63–68.

Bernardi G, Bucciarelli G, Costagliola D, Robertson DR, Heiser JB (2004) Evolution of coral reef fish *Thalassoma spp*. (Labridae). 1. Molecular phylogeny and biogeography. Marine Biology, 144, 369–375.

Bernardi G (2013) Speciation in fishes. Molecular ecology, 22, 5487–5502.

Bernardi G, Noguchi R, Anderson AB, Floeter SR, Ferreira CEL (2013) Sargo Amarelo, a traditionally recognized hybrid between two species of Brazilian reef fishes. Marine Biodiversity, 43, 255–256.

Bernatchez L, Glémet H, Wilson CC, Danzmann RG (1995) Introgression and fixation of Arctic char (*Salvelinus alpinus*) mitochondrial genome in an allopatric population of brook trout (*Salvelinus fontinalis*). Canadian Journal of Fisheries and Aquatic Sciences, 52, 179–185.

Berner D, Adams DC, Grandchamp A-C, Hendry AP (2008) Natural selection drives patterns of lake–stream divergence in stickleback foraging morphology. Journal of Evolutionary Biology, 21, 1653–1665.

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

Berthier P, Excoffier L, Ruedi M (2006) Recurrent replacement of mtDNA and cryptic hybridization between two sibling bat species *Myotis myotis* and *Myotis blythii*. Proceedings of the Royal Society of London B: Biological Sciences, 273, 3101–3123.

Bertucci F, Ruppé L, Wassenbergh SV, Compère P, Parmentier E (2014) New insights into the role of the pharyngeal jaw apparatus in the sound-producing mechanism of *Haemulon flavolineatum* (Haemulidae). The Journal of Experimental Biology, 217, 3862–3869.

Bird CE, Fernandez-Silva I, Skillings DJ, Toonen RJ (2012) Sympatric speciation in the post "modern synthesis" era of evolutionary biology. Evolutionary Biology, 39, 158–180.

Bowen BW, Rocha LA, Toonen RJ, Karl SA (2013) The origins of tropical marine biodiversity. Trends in ecology & evolution, 28, 359–366.

Brelsford A, Milá B, Irwin DE (2011) Hybrid origin of Audubon's warbler. Molecular Ecology, 20, 2380–2389.

Briggs JC (2006) Proximate sources of marine biodiversity. Journal of Biogeography, 33, 1–10.

Broughton RE, Vedala KC, Crowl TM, Ritterhouse LL (2011) Current and historical hybridization with differential introgression among three species of cyprinid fishes (genus *Cyprinella*). Genetica, 139, 699–707.

Burton RS, Barreto FS. 2012. A disproportionate role for mtDNA in Dobzhansky–Muller incompatibilities? Molecular Ecology, 21, 4942–4957.

Calduch-Giner JA, Bermejo-Nogales A, Benedito-Palos L et al. (2013) Deep sequencing for de novo construction of a marine fish (*Sparus aurata*) transcriptome database with a large coverage of protein-coding transcripts. BMC Genomics, 14, 178.

Cara JB, Aluru N, Moyano FJ, Vijayan MM (2005) Food-deprivation induces HSP70 and HSP90 protein expression in larval gilthead sea bream and rainbow trout. Comparative Biochemistry and Physiology Part B: Biochemistry and Molecular Biology, 142, 426–431.

Catchen JM, Amores A, Hohenlohe P, Cresko W, Postlethwait JH (2011) Stacks: building and genotyping loci de novo from short-read sequences. G3: Genes, Genomes, Genetics, 1, 171–182.

Choat JH, Klanten OS, Van Herwerden L, Robertson DR, Clements KD (2012) Patterns and processes in the evolutionary history of parrotfishes (Family Labridae). Biological Journal of the Linnean Society, 107, 529–557.

Coleman RR, Gaither MR, Kimokeo B et al. (2014) Large-scale introduction of the Indo-Pacific damselfish *Abudefduf vaigiensis* into Hawai'i promotes genetic swamping of the endemic congener *A. abdominalis*. Molecular Ecology, 23, 5552–5565.

Consortium TU (2014) UniProt: a hub for protein information. Nucleic Acids Research, gku989.

Coyne JA, Orr HA (2004) Speciation. Sinauer Associates, Sunderland, MA.

Darriba D, Taboada GL, Doallo R, Posada D (2012) jModelTest 2: more models, new heuristics and parallel computing. Nature methods, 9, 772–772.

Darwin CR (1859) On the origin of species by means of natural selection, or the preservation of favoured races in the struggle for life. John Murray, London, UK.

DiBattista JD, Feldheim KA, Bowen BW (2011) Microsatellite DNA markers to resolve population structure and hybridization of two closely related surgeonfish species, *Acanthurus nigricans* and *Acanthurus leucosternon*. Conservation Genetics Resources, 3, 159–162.

DiBattista JD, Waldrop E, Bowen BW et al. (2012) Twisted sister species of pygmy angelfishes: discordance between taxonomy, coloration, and phylogenetics. Coral Reefs, 31, 839–851.

Dibattista JD, Rocha LA, Hobbs JPA, He S, Priest MA, Sinclair-Taylor TH, Bowen BW, Berumen ML (2015) When biogeographical provinces collide: hybridization of reef fishes at the crossroads of marine biogeographical provinces in the Arabian Sea. Journal of Biogeography, doi:10.1111/jbi.12526

Dion-Côté A-M, Renaut S, Normandeau E, Bernatchez L (2014) RNA-seq Reveals Transcriptomic Shock Involving Transposable Elements Reactivation in Hybrids of Young Lake Whitefish Species. Molecular Biology and Evolution, 31, 1188–1199.

Dixon GB, Davies SW, Aglyamova GV et al. (2015) Genomic determinants of coral heat tolerance across latitudes. Science, 348, 1460–1462.

Domeier ML, Colin PL (1997) Tropical reef fish spawning aggregations: defined and reviewed. Bulletin of Marine Science, 60, 698–726.

Drummond AJ, Suchard MA, Xie D, Rambaut A (2012) Bayesian phylogenetics with BEAUti and the BEAST 1.7. Molecular Biology and Evolution, 29, 1969–1973.

Eagle JV, Jones GP (2004) Mimicry in coral reef fishes: ecological and behavioural responses of a mimic to its model. Journal of Zoology, 264, 33–43.

Earl DA, vonHoldt BM (2012) STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. Conservation genetics resources, 4, 359–361.

Elmer KR, Fan S, Gunter HM et al. (2010) Rapid evolution and selection inferred from the transcriptomes of sympatric crater lake cichlid fishes. Molecular ecology, 19, 197–211.

Excoffier L, Lischer HEL (2010) Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. Molecular Ecology Resources, 10, 564–567.

Fan S, Elmer KR, Meyer A (2011) Positive Darwinian selection drives the evolution of the morphology-related gene, EPCAM, in particularly species-rich lineages of African cichlid fishes. Journal of Molecular Evolution, 73, 1–9.

Flores-Ortega JR, Godínez-Domínguez E, González-Sansón G, Rojo-Vázquez JA, Corgos A, Galván Piña VH, González Sansón G (2010). Interacciones tróficas de las seis especies de peces más abundantes en la pesquería artesanal en dos bahías del Pacífico Central Mexicano. International Journal of Tropical Biology and Conservation, 58, 383–397.

Foll M, Gaggiotti O (2008) A genome-scan method to identify selected loci appropriate for both dominant and codominant markers: a Bayesian perspective. Genetics, 180, 977–993.

Fraser BA, Weadick CJ, Janowitz I, Rodd FH, Hughes KA (2011) Sequencing and characterization of the guppy (*Poecilia reticulata*) transcriptome. BMC Genomics, 12, 202.

Fric J, Zelante T, Ricciardi-Castagnoli P (2014) Phagocytosis of particulate antigens – all roads lead to Calcineurin/NFAT Signaling Pathway. Frontiers in Immunology, 4, 513. doi:10.3389/fimmu.2013.00513.

Fu B, He S (2012) Transcriptome Analysis of Silver Carp (*Hypophthalmichthys molitrix*) by paired-end RNA sequencing. DNA Research, dsr046.

Fu L, Niu B, Zhu Z, Wu S, Li W (2012) CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics (Oxford, England), 28, 3150–3152.

Funk DJ, Omland KE (2003) Species-Level Paraphyly and Polyphyly: Frequency, causes, and consequences, with insights from animal mitochondrial DNA. Annual Review of Ecology, Evolution, and Systematics, 34, 397–423.

Gainsford A, van Herwerden L, Jones GP (2015) Hierarchical behaviour, habitat use and species size differences shape evolutionary outcomes of hybridization in a coral reef fish. Journal of Evolutionary Biology, 28, 205–222.

Gaither MR, Schultz JK, Bellwood DR et al. (2014) Evolution of pygmy angelfishes: Recent divergences, introgression, and the usefulness of color in taxonomy. Molecular Phylogenetics and Evolution, 74, 38–47.

Gaither MR, Bernal MA, Coleman RR et al. (2015a) Genomic signatures of geographic isolation and natural selection in coral reef fishes. Molecular Ecology, 24, 1543–1557.

Gaither MR, Bernal MA, Fernandez-Silva I et al. (2015b) Two deep evolutionary lineages in the circumtropical glasseye *Heteropriacanthus cruentatus* (Teleostei, Priacanthidae) with admixture in the south-western Indian Ocean. Journal of fish biology, 87, 715–727.

Gayral P, Weinert L, Chiari Y et al. (2011) Next-generation sequencing of transcriptomes: a guide to RNA isolation in nonmodel animals. Molecular Ecology Resources, 11, 650–661.

Gerstein MB, Rozowsky J, Yan K-K et al. (2014) Comparative analysis of the transcriptome across distant species. Nature, 512, 445–448.

Ghadessy FJ, Chen D, Kini RM et al. (1996) Stonustoxin is a novel lethal factor from Stonefish (*Synanceja horrida*) venom cDNA cloning and characterization. Journal of Biological Chemistry, 271, 25575–25581.

Grabherr MG, Haas BJ, Yassour M et al. (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nature Biotechnology, 29, 644–652.

Grønskov K, Dooley CM, Østergaard E et al. (2013) Mutations in c10orf11, a melanocyte-differentiation gene, cause autosomal-recessive albinism. The American Journal of Human Genetics, 92, 415–421.

Grubich J (2003) Morphological convergence of pharyngeal jaw structure in durophagous perciform fish. Biological Journal of the Linnean Society, 80, 147–165.

Hasegawa M, Kishino H, Yano T (1985) Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. Journal of Molecular Evolution, 22, 160–174.

Henning F, Jones JC, Franchini P, Meyer A (2013) Transcriptomics of morphological color change in polychromatic Midas cichlids. BMC Genomics, 14, 171.

Hobbs J-PA, Frisch AJ, Allen GR, Herwerden LV (2008) Marine hybrid hotspot at Indo-Pacific biogeographic border. Biology Letters, rsbl.2008.0561.

Hodge JR, Read CI, Bellwood DR, Herwerden L (2013) Evolution of sympatric species: a case study of the coral reef fish genus *Pomacanthus* (Pomacanthidae). Journal of Biogeography, 40, 1676–1687.

Hohenlohe PA, Bassham S, Etter PD et al. (2010) Population genomics of parallel adaptation in Threespine Stickleback using sequenced RAD tags. PLoS Genet, 6, e1000862.

Horne JB, van Herwerden L, Choat JH, Robertson DR (2008) High population connectivity across the Indo-Pacific: congruent lack of phylogeographic structure in three reef fish congeners. Molecular Phylogenetics and Evolution, 49, 629–638.

Huang L, Li G, Mo Z et al. (2015) De Novo Assembly of the Japanese Flounder (*Paralichthys olivaceus*) Spleen Transcriptome to Identify Putative Genes Involved in Immunity. PLoS ONE, 10, e0117642.

Hubbs CL (1955) Hybridization between Fish Species in Nature. Systematic Zoology, 4, 1–20.

Hyllner SJ, Westerlund L, Olsson P-E, Schopen A (2001) Cloning of Rainbow Trout egg envelope proteins: members of a unique group of structural proteins. Biology of Reproduction, 64, 805–811.

Iwamatsu T, Yoshizaki N, Shibata Y (1997) Changes in the chorion and sperm entry into the micropyle during fertilization in the teleostean fish, *Oryzias latipes*. Development, Growth & Differentiation, 39, 33–41.

Jakobsson M, Rosenberg NA (2007) CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. Bioinformatics, 23, 1801–1806.

Jeffares DC, Tomiczek B, Sojo V, dos Reis M (2015) A beginners guide to estimating the non-synonymous to synonymous rate ratio of all protein-coding genes in a genome. In: Parasite Genomics Protocols, pp. 65–90. Springer.

Jeukens J, Renaut S, St-Cyr J, Nolte AW, Bernatchez L (2010) The transcriptomics of sympatric dwarf and normal lake whitefish (*Coregonus clupeaformis spp*., Salmonidae) divergence as revealed by next-generation sequencing. Molecular Ecology, 19, 5389–5403.

Jones AG, Moore GI, Kvarnemo C, Walker D, Avise JC (2003) Sympatric speciation as a consequence of male pregnancy in seahorses. Proceedings of the National Academy of Sciences, 100, 6598–6603.

Jordan DS (1908) The Law of Geminate Species. The American Naturalist, 42, 73–80.

Joyce DA, Lunt DH, Genner MJ et al. (2011) Repeated colonization and hybridization in Lake Malawi cichlids. Current Biology, 21, R108–R109.

Kanehisa M, Goto S, Kawashima S, Nakaya A (2002) The KEGG databases at GenomeNet. Nucleic Acids Research, 30, 42–46.

Katoh K, Standley DM (2013) MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. Molecular Biology and Evolution, 30, 772–780.

Kelley JL, Passow CN, Plath M et al. (2012) Genomic resources for a model in adaptation and speciation research: characterization of the *Poecilia mexicana* transcriptome. BMC Genomics, 13, 652.

Kienzler A, Bony S, Devaux A (2013) DNA repair activity in fish and interest in ecotoxicology: a review. Aquatic toxicology, 134, 47–56.

Kitano J, Yoshida K, Suzuki Y (2013) RNA sequencing reveals small RNAs differentially expressed between incipient Japanese threespine sticklebacks. BMC Genomics, 14, 214.

Klanten SO, Herwerden L van, Choat JH, Blair D (2004) Patterns of lineage diversification in the genus *Naso* (Acanthuridae). Molecular phylogenetics and evolution, 32, 221–235.

Kolde, Raivo (2015). pheatmap: Pretty Heatmaps. R package version 1.0.7. http://CRAN.R-project.org/package=pheatmap

Kück P, Longo GC (2014) FASconCAT-G: extensive functions for multiple sequence alignment preparations concerning phylogenetic studies. Frontiers in Zoology, 11, 81.

Kujawski S, Lin W, Kitte F, Börmel M, Fuchs S, Arulmozhivarman G, Vogt S, Theil D, Zhang Y, Antos CL. 2014. Calcineurin regulates coordinated outgrowth of zebrafish regenerating fins. Developmental Cell, 28, 573–587.

Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. Nature Methods, 9, 357–359.

Larbuisson A, Dalcq J, Martial JA, Muller M (2013) Fgf receptors Fgfr1a and Fgfr2 control the function of pharyngeal endoderm in late cranial cartilage development. Differentiation, 86, 192–206.

Larmuseau MHD, Vancampenhout K, Raeymaekers J a. M, Van Houdt JKJ, Volckaert F a. M (2010) Differential modes of selection on the rhodopsin gene in coastal Baltic and North Sea populations of the sand goby, *Pomatoschistus minutus*. Molecular Ecology, 19, 2256–2268.

Lee W-J, Conroy J, Howell WH, Kocher TD (1995) Structure and evolution of teleost mitochondrial control regions. Journal of Molecular Evolution, 41, 54–66.

Lessios HA (2008) The great American schism: divergence of marine organisms after the rise of the Central American Isthmus. Annual Review of Ecology, Evolution, and Systematics, 39, 63–91.

Lessios HA, Robertson DR (2006) Crossing the impassable: genetic connections in 20 reef fishes across the eastern Pacific barrier. Proceedings of the Royal Society B: Biological Sciences, 273, 2201–2208.

Lindeman KC, Toxey CS (2002) Haemulidae: grunts. The living marine resources of the Western Central Atlantic, 3, 1522–1529.

Lischer HEL, Excoffier L (2012) PGDSpider: an automated data conversion tool for connecting population genetics and genomics programs. Bioinformatics, 28, 298–299.

Litsios G, Salamin N (2014) Hybridisation and diversification in the adaptive radiation of clownfishes. BMC Evolutionary Biology, 14, 245.

Llopart A, Lachaise D, Coyne JA (2005) Multilocus analysis of introgression between two sympatric sister species of *Drosophila*: *Drosophila yakuba* and *D. santomea*. Genetics 171:197–210.

Llopart A, Herrig D, Brud E, Stecklein Z (2014) Sequential adaptive introgression of the mitochondrial genome in *Drosophila yakuba* and *Drosophila santomea*. Molecular Ecology, 23, 1124–1136.

Losey Jr GS (2003) Crypsis and communication functions of UV-visible coloration in two coral reef damselfish, *Dascyllus aruanus* and *D. reticulatus*. Animal Behaviour, 66, 299–307.

Love MI, Huber W, Anders S (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol, 15, 550.

Lovejoy NR, Collette BB, McEachran JD (2001) Phylogenetic relationships of New World Needlefishes (Teleostei: Belonidae) and the biogeography of transitions between marine and freshwater habitats. Copeia, 2001, 324–338.

Mahler J, Driever W (2007) Expression of the zebrafish intermediate neurofilament Nestin in the developing nervous system and in neural proliferation zones at postembryonic stages. BMC Developmental Biolology, 7, 89.

Mallet J (2005) Hybridization as an invasion of the genome. Trends in Ecology & Evolution, 20, 229–237.

Manousaki T, Hull PM, Kusche H et al. (2013) Parsing parallel evolution: ecological divergence and differential gene expression in the adaptive radiations of thick-lipped Midas cichlid fishes from Nicaragua. Molecular ecology, 22, 650–669.

Martin CI, Johnston IA (2005) The role of myostatin and the calcineurin-signalling pathway in regulating muscle mass in response to exercise training in the rainbow trout *Oncorhynchus mykiss* Walbaum. Journal of Experimental Biology, 208, 2083–2090.

Martin JA, Wang Z (2011) Next-generation transcriptome assembly. Nature Reviews Genetics, 12, 671–682.

Matz, Mikhail V. (2015). empiricalFDR.DESeq2: Simulation-Based False Discovery Rate in RNA-Seq. R package version 1.0.3. http://CRAN.R-project.org/package=empiricalFDR.DESeq2

Mayr E (1947) Ecological Factors in Speciation. Evolution, 1, 263–288.

Mayr E (1954) Geographic speciation in tropical echinoids. Evolution, 1–18.

Mayr E (1963) Animal species and evolution. Belknap Press of Harvard University Press. Cambridge, MA.

McCafferty S, Bermingham E, Quenouille B et al. (2002) Historical biogeography and molecular systematics of the Indo-Pacific genus *Dascyllus* (Teleostei: Pomacentridae). Molecular Ecology, 11, 1377–1392.

Meirmans PG, Van Tienderen PH (2004) genotype and genodive: two programs for the analysis of genetic diversity of asexual organisms. Molecular Ecology Notes, 4, 792–794.

Melo-Ferreira J, Vilela J, Fonseca MM, da Fonseca RR, Boursot, P, Alves PC (2014) The elusive nature of adaptive mitochondrial DNA evolution of an arctic lineage prone to frequent introgression. Genome Biology and Evolution, 6, 886–896.

Meyer A (1993) Evolution of mitochondrial DNA in fishes. In: Biochemistry and Molecular Biology of Fishes (eds. Hochanchka PW, Mommsen TP), pp. 1–38. Elsevier, New York.

Meyer E, Aglyamova GV, Wang S et al. (2009) Sequencing and de novo analysis of a coral larval transcriptome using 454 GSFlx. BMC Genomics, 10, 219.

Meyer E, Aglyamova GV, Matz MV (2011) Profiling gene expression responses of coral larvae (*Acropora millepora*) to elevated temperature and settlement inducers using a novel RNA-Seq procedure. Molecular Ecology, 20, 3599–3616.

Meyer JL, Schultz ET (1985) Migrating haemulid fishes as a source of nutrients and organic matter on coral reefs1. Limnology and Oceanography, 30, 146–156.

Mims MC, Darrin Hulsey C, Fitzpatrick BM, Todd Streelman J (2010) Geography disentangles introgression from ancestral polymorphism in lake Malawi cichlids. Molecular Ecology, 19, 940–951.

Montanari SR, Van Herwerden L, Pratchett MS, Hobbs J-PA, Fugedi A (2012) Reef fish hybridization: lessons learnt from butterflyfishes (genus *Chaetodon*). Ecology and evolution, 2, 310–328.

Moriya Y, Itoh M, Okuda S, Yoshizawa AC, Kanehisa M (2007) KAAS: an automatic genome annotation and pathway reconstruction server. Nucleic Acids Research, 35, W182–W185.

Nielsen EE, Hemmer-Hansen J, Larsen PF, Bekkevold D (2009) Population genomics of marine fishes: identifying adaptive variation in space and time. Molecular Ecology, 18, 3128–3150.

Nimmrich V, Gross G (2012) P/Q-type calcium channel modulators. British Journal of Pharmacology, 167, 741–759.

Nosil P (2012) Ecological Speciation. Oxford University Press, Oxford, NY.

Nosil P, Schluter D (2011) The genes underlying the process of speciation. Trends in Ecology & Evolution, 26, 160–167.

Palumbi SR (1994) Genetic divergence, reproductive isolation, and marine speciation. Annual Review of Ecology and Systematics, 25, 547–572.

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

Parra G, Bradnam K, Korf I (2007) CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. Bioinformatics, 23, 1061–1067.

Pavey SA, Collin H, Nosil P, Rogers SM (2010) The role of gene expression in ecological speciation. Annals of the New York Academy of Sciences, 1206, 110–129.

Pereira PHC, Barros B, Zemoi R, Ferreira BP (2014) Ontogenetic diet changes and food partitioning of *Haemulon* spp. coral reef fishes, with a review of the genus diet. Reviews in Fish Biology and Fisheries, 25, 245–260.

Pereiro P, Balseiro P, Romero A et al. (2012) High-throughput sequence analysis of turbot (*Scophthalmus maximus*) transcriptome using 454-pyrosequencing for the discovery of antiviral immune genes. PLoS ONE, 7, e35369.

Peterson BK, Weber JN, Kay EH, Fisher HS, Hoekstra HE (2012) Double digest RADseq: an inexpensive method for de novo SNP discovery and genotyping in model and non-model species. PLoS ONE, 7, e37135.

Platt AR, Woodhall RW, George AL (2007) Improved DNA sequencing quality and efficiency using an optimized fast cycle sequencing protocol. BioTechniques, 43, 58.

Pons J-M, Sonsthagen S, Dove C, Crochet P-A (2014) Extensive mitochondrial introgression in North American great black-backed gulls (*Larus marinus*) from

the American Herring Gull (*Larus smithsonianus*) with little nuclear DNA impact. Heredity, 112, 226–239.

Price SA, Tavera JJ, Near TJ, Wainwright P, others (2013) Elevated rates of morphological and functional diversification in reef-dwelling haemulid fishes. Evolution, 67, 417–428.

Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. Genetics, 155, 945–959.

Puebla O, Bermingham E, Guichard F, Whiteman E (2007) Colour pattern as a single trait driving speciation in *Hypoplectrus* coral reef fishes? Proceedings of the Royal Society B: Biological Sciences, 274, 1265–1271.

Puebla O (2009) Ecological speciation in marine *v*. freshwater fishes. Journal of Fish Biology, 75, 960–996.

Puebla O, Bermingham E, Guichard F (2012a) Pairing dynamics and the origin of species. Proceedings of the Royal Society of London B: Biological Sciences, 279, 1085–1092.

Puebla O, Bermingham E, McMillan WO (2012b) On the spatial scale of dispersal in coral reef fishes. Molecular Ecology, 21, 5675–5688.

Puebla O, Bermingham E, McMillan WO (2014) Genomic atolls of differentiation in coral reef fishes (*Hypoplectrus* spp., Serranidae). Molecular Ecology, 23, 5291–5303.

Purcell JFH, Cowen RK, Hughes CR, Williams DA (2006) Weak genetic structure indicates strong dispersal limits: a tale of two coral reef fish. Proceedings of the Royal Society of London B: Biological Sciences, 273, 1483–1490.

Pyle RL, Randall JE (1994) A review of hybridization in marine angelfishes (Perciformes: Pomacanthidae). Environmental Biology of Fishes, 41, 127–145.

R Core Team (2015) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/

Rambaut A, Drummond AJ (2007) Tracer version 1.4.

Ramsköld D, Wang ET, Burge CB, Sandberg R (2009) An abundance of ubiquitously expressed genes revealed by tissue transcriptome sequence data. PLoS Computatitional Biology, 5, e1000598.

Randall JE (1967) Food habits of reef fishes of the West Indies. Institute of Marine Sciences, University of Miami.

Raymundo-Huizar AR (2000) Análisis de la dieta de los peces demersales de fondos blandos en la plataforma continental de Jalisco y Colima, México. Unpublished Master's thesis, Universidad de Colima, México.

Reece JS, Bowen BW, Smith DG, Larson A (2011) Comparative phylogeography of four Indo-Pacific moray eel species (Muraenidae) reveals comparable ocean-wide genetic connectivity despite five-fold differences in available adult habitat. Marine Ecology Progress Series, 437, 269–277.

Robertson DR, Allen GR. 2015. Shorefishes of the Tropical Eastern Pacific: online information system, Version 2.0. http://biogeodb.stri.si.edu/sftep/. Smithsonian Tropical Research Institute, Panama.

Roca AL, Georgiadis N, O'Brien SJ (2005) Cytonuclear genomic dissociation in African elephant species. Nature Genetics, 37, 96–100.

Rocha LA, Bass AL, Robertson DR, Bowen BW (2002) Adult habitat preferences, larval dispersal, and the comparative phylogeography of three Atlantic surgeonfishes (Teleostei: Acanthuridae). Molecular Ecology, 11, 243–251.

Rocha LA, Robertson DR, Roman J, Bowen BW (2005) Ecological speciation in tropical reef fishes. Proceedings of the royal society B: biological sciences, 272, 573–579.

Rocha LA, Craig MT, Bowen BW (2007) Phylogeography and the conservation of coral reef fishes. Coral Reefs, 26, 501–512.

Rocha LA, Bowen BW (2008) Speciation in coral-reef fishes. Journal of Fish Biology, 72, 1101–1121.

Rocha LA, Lindeman KC, Rocha CR, Lessios HA (2008) Historical biogeography and speciation in the reef fish genus *Haemulon* (Teleostei: Haemulidae). Molecular Phylogenetics and Evolution, 48, 918–928.

Rocha LA, Bernal MA, Gaither MR, Alfaro ME (2013) Massively parallel DNA sequencing: the new frontier in biogeography. Frontiers of Biogeography, 5.

Rohland N, Reich D (2012) Cost-effective, high-throughput DNA sequencing libraries for multiplexed target capture. Genome Research, 22, 939–946.

Romero IG, Ruvinsky I, Gilad Y (2012) Comparative studies of gene expression and the evolution of gene regulation. Nature Reviews Genetics, 13, 505–516.

Rosenberg NA (2004) distruct: a program for the graphical display of population structure. Molecular Ecology Notes, 4, 137–138.

Rosenthal GG, García de León FJ (2011) Speciation and hybridization. Ecology and evolution of Poeciliid Fishes. University of Chicago Press, Chicago, 109–119.

Rundle HD, Nosil P (2005) Ecological speciation. Ecology Letters, 8, 336–352.

Sakamoto T, Shepherd BS, Madsen SS, Nishioka RS, Siharath K, Richman III NH, Bern HA, Grau EG (1997) Osmoregulatory actions of growth hormone and prolactin in an advanced teleost. General and Comparative Endocrinology, 106, 95–101.

Santos ME, Braasch I, Boileau N et al. (2014) The evolution of cichlid fish egg-spots is linked with a *cis*-regulatory change. Nature Communications, 5.

Schierenbeck K. 2011. Hybridization and introgression. In: Encyclopedia of biological invasions (eds: Simberloff D, Rejmanek M), pp. 342–346. University of California Press, California.

Schultz JK, Pyle RL, DeMartini E, Bowen BW (2006) Genetic connectivity among color morphs and Pacific archipelagos for the flame angelfish, *Centropyge loriculus*. Marine Biology, 151, 167–175.

Schunter C, Vollmer SV, Macpherson E, Pascual M (2014) Transcriptome analyses and differential gene expression in a non-model fish species with alternative mating tactics. BMC Genomics, 15, 167.

Scribner KT, Page KS, Bartron ML (2000) Hybridization in freshwater fishes: a review of case studies and cytonuclear methods of biological inference. Reviews in Fish Biology and Fisheries, 10, 293–323.

Seehausen O, Terai Y, Magalhaes IS et al. (2008) Speciation through sensory drive in cichlid fish. Nature, 455, 620–626.

Seehausen O, Butlin RK, Keller I et al. (2014) Genomics and the origin of species. Nature Reviews Genetics, 15, 176–192.

Stephens M, Donnelly P (2003) A comparison of bayesian methods for haplotype reconstruction from population genotype data. The American Journal of Human Genetics, 73, 1162–1169.

Streelman JT, Karl SA (1997) Reconstructing labroid evolution with single–copy nuclear DNA. Proceedings of the Royal Society of London B: Biological Sciences, 264, 1011–1020.

Sugawara T, Terai Y, Okada N (2002) Natural selection of the Rhodopsin gene during the adaptive radiation of east African Great Lakes Cichlid Fishes. Molecular Biology and Evolution, 19, 1807–1811.

Supek F, Bošnjak M, Škunca N, Šmuc T (2011) REVIGO summarizes and visualizes long lists of Gene Ontology terms. PLoS ONE, 6, e21800.

Suyama M, Torrents D, Bork P (2006) PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. Nucleic Acids Research, 34, W609–W612.

Swanson WJ, Vacquier VD (2002) The rapid evolution of reproductive proteins. Nature Reviews Genetics, 3, 137–144.

Tamura K, Stecher G, Peterson D, Filipski A, Kumar S (2013) MEGA6: Molecular evolutionary genetics analysis Version 6.0. Molecular Biology and Evolution, 30, 2725–2729.

Tavaré S (1984) Line-of-descent and genealogical processes, and their applications in population genetics models. Theoretical Population Biology, 26, 119–164.

Tavera JJ, Arturo AP, Balart EF, Bernardi G (2012) Molecular phylogeny of grunts (Teleostei, Haemulidae), with an emphasis on the ecology, evolution, and speciation history of New World species. BMC evolutionary biology, 12, 57.

Terai Y, Morikawa N, Okada N (2002) The evolution of the pro-domain of bone morphogenetic protein 4 (Bmp4) in an explosively speciated lineage of east African cichlid fishes. Molecular Biology and Evolution, 19, 1628–1632.

Terai Y, Seehausen O, Sasaki T et al. (2006) Divergent selection on opsins drives incipient speciation in Lake Victoria cichlids. PLoS Biol, 4, e433.

Toews DPL, Brelsford A (2012) The biogeography of mitochondrial and nuclear discordance in animals. Molecular Ecology, 21, 3907–3930.

Toews DPL, Mandic M, Richards JG, Irwin DE (2014) Migration, mitochondria, and the Yellow-Rumped warbler. Evolution, 68, 241–255.

Trokovic N, Trokovic R, Partanen J (2005) Fibroblast growth factor signalling and regional specification of the pharyngeal ectoderm. International journal of developmental biology, 49, 797.

Via S (2012) Divergence hitchhiking and the spread of genomic isolation during ecological speciation-with-gene-flow. Philosophical Transactions of the Royal Society B: Biological Sciences, 367, 451–460.

von Schalburg KR, Rise ML, Brown GD, Davidson WS, Koop BF (2004) A comprehensive survey of the genes involved in maturation and development of the rainbow trout ovary. Biology of reproduction.

Wainwright PC (1989) Functional morphology of the pharyngeal jaw apparatus in perciform fishes: An experimental analysis of the haemulidae. Journal of Morphology, 200, 231–245.

Wang H, Gong Z (1999) Characterization of two zebrafish cDNA clones encoding egg envelope proteins ZP2 and ZP3. Biochimica et Biophysica Acta (BBA) - Gene Structure and Expression, 1446, 156–160.

Wang Z, Gerstein M, Snyder M (2009) RNA-Seq: a revolutionary tool for transcriptomics. Nature Reviews Genetics, 10, 57–63.

Ward RD, Zemlak TS, Innes BH, Last PR, Hebert PDN (2005) DNA barcoding Australia's fish species. Philosophical Transactions of the Royal Society B: Biological Sciences, 360, 1847–1857.

Weir BS, Cockerham CC (1984) Estimating F-Statistics for the analysis of population structure. Evolution, 38, 1358–1370.

Whittington CM, Wilson AB (2013) The role of prolactin in fish reproduction. General and Comparative Endocrinology, 191, 123–136.

Wright RM, Aglyamova GV, Meyer E, Matz MV (2015) Gene expression associated with white syndromes in a reef building coral, *Acropora hyacinthus*. BMC Genomics, 16, 371.

Wu C-I, Ting C-T (2004) Genes and speciation. Nature Reviews Genetics, 5, 114–122.

Wu S, Zhu Z, Fu L, Niu B, Li W (2011) WebMGA: a customizable web server for fast metagenomic sequence analysis. BMC Genomics, 12, 444.

Yaakub SM, Bellwood DR, Van Herwerden L (2007) A rare hybridization event in two common Caribbean wrasses (genus *Halichoeres*; family Labridae). Coral Reefs, 26, 597–602.

Yang L, Wang Y, Zhang Z, He S (2015) Comprehensive transcriptome analysis reveals accelerated genic evolution in a Tibet fish, *Gymnodiptychus pachycheilus*. Genome Biology and Evolution, 7, 251–261.

Yang Z (2007) PAML 4: Phylogenetic Analysis by Maximum Likelihood. Molecular Biology and Evolution, 24, 1586–1591.

Yang Z, Wong WS, Nielsen R (2005) Bayes empirical Bayes inference of amino acid sites under positive selection. Molecular biology and evolution, 22, 1107–1118.

Zieliński P, Nadachowska-Brzyska K, Wielstra B et al. (2013) No evidence for nuclear introgression despite complete mtDNA replacement in the Carpathian newt (*Lissotriton montandoni*). Molecular Ecology, 22, 1884–1903.