UNIVERSITY OF OKLAHOMA

GRADUATE COLLEGE

BRIDGING MICRO- AND MACRO- EVOLUTION IN TROPICAL FISHES

A DISSERTATION SUBMITTED TO THE GRADUATE FACULTY

in partial fulfillment of the requirements for the

Degree of

DOCTOR OF PHILOSOPHY

By

CARMEN DEL ROCIO PEDRAZA MARRON

Norman, Oklahoma

2023

Downloaded from SHAREOK. Persistent Handle URL: https://hdl.handle.net/11244/340088

BRIDGING MICRO- AND MACRO- EVOLUTION IN TROPICAL FISHES

A DISSERTATION APPROVED FOR THE

SCHOOL OF BIOLOGICAL SCIENCES

BY THE COMMITTEE CONSISTING OF

Chair: Dr. Katharine Marske

Dr. Ricardo Betancur

Dr. Cameron Siler

Dr. Brian Kemp

Dr. Philip Hastings

© Copyright by Carmen del Rocío Pedraza Marrón 2023

All Rights Reserved.

"To my parents, Hugo Pedraza. and Carmen del R. Marrón, and my brother, Hugo Alejandro: you are the reason I have reached this dream"

Abstract

In marine environments, barriers to dispersal can be challenging to identify because they are often cryptic. Unlike terrestrial environments, where a mountain chain that is visible can physically separate two populations of animals, vast masses of water in the ocean make it challenging to pinpoint these barriers. Therefore, the impact of these barriers on the formation of new species in the ocean is still not well understood. While most marine populations have long been considered to be well connected via long-distance dispersal, molecular ecology studies are increasingly unveiling inconspicuous barriers that promote population divergence and ultimately speciation. The advent of genomic techniques that allow the generation of data for thousands of genes has provided an unprecedented opportunity to uncover marine barriers that were previously invisible using more rudimentary tools. This, in turn, has opened new avenues for understanding of how barriers to dispersal affect population connectivity in the marine environment. The overarching goal of my dissertation is to use genome-wide data to look for genetic patterns that correspond to such barriers, and to test for their effect at short-, intermediate- and long-term evolutionary scales, going through a continuum from micro- to macro-evolution, in a time span from thousands to millions of years.

At the short-scale, I examined two controversial cases of species delimitation. Species delimitation is a major question in biology and is essential for adequate management of organismal diversity. The first challenging case involves the red snappers in the Western Atlantic. Red snappers have been traditionally recognized as two separate species based on morphology: *Lutjanus campechanus* (northern red snapper) and *L. purpureus* (southern red snapper). However, recent genetic studies using few molecular markers failed to delineate these nominal species, lumping the northern and southern populations into a single species (*L. campechanus*). To evaluate if the populations of these fish represent one or two species, my project applied *ca.* 40,000 genome-wide markers of 178 individuals collected throughout the range of the two species and population and species delimitation analyses. Overall, my results supported the isolation and differentiation of these species, a result that confirmed the morphology-based delimitation scenario, highlighting the benefits of using genome-wide data in complex cases of species delimitation (Chapter I, published in *Proc. Roy. Soc. B* in 2019).

The second study case involves a species complex of silverside fishes (*Chirostoma humboltianum* group: Atherinidae) in the Central Mexico plateau. The *humboltianum* group represents a taxonomically-controversial species complex where previous morphological and molecular studies based on a few genes produced conflicting species delineation scenarios. I applied an integrative approach that considered multiple lines of evidence to investigate the species numbers and boundaries comprising this contentious group. I used *ca.* 33,000 molecular markers for 77 individuals representing the nine nominal species in the group, spanning their distribution range in the central Mexico plateau, in combination with morphologic and ecologic information. My findings are inconsistent with the morphospecies and ecological delimitation scenarios, identifying three to four species. This study provides an atypical example in which genome-wide analyses delineate fewer species than previously recognized on the basis of morphological data alone. It also highlights the influence of geologic history as a main driver of speciation in the group (Chapter II, published in *BMC Eco. Evol. B* in 2022).

At the intermediate- scale, I evaluated the influence of historical (e.g., geophysical events) and contemporary barriers (e.g., habitat gaps) hindering genetic flow among populations by studying the spatio-temporal phylogenetic concordance of co-distributed lineages. For this study, I investigated the comparative phylogeography of labrisomid blennies in the genus *Malacoctenus*. I generated data for *ca.* 28K genome-

wide markers that were sequenced from over 500 individuals collected from 38 locations, representing 23 (out of 25) species of *Malacoctenus*. With this dataset, I assessed the effect of recognized historical (e.g., the rising of the Isthmus of Panama) and contemporary barriers (e.g., sandy gaps) in the Tropical Eastern Pacific (TEP) and the Tropical Atlantic (TA) biogeographic realms. These blennies represent an ideal system to test the effect of such barriers as they are strongly associated with rocky habitats and coral reefs. Therefore, subtle habitat disruptions may lead to genetic isolation. At the micro-evolutionary scale, the observed population structure patterns identified the Sinaloan and Central American breaks as the major breaks in the TEP; and the Bahamas and Eastern Caribbean breaks as key barriers disrupting connectivity in the TA. All in all, the effect of these breaks varies across species, suggesting that species-specific traits (e.g., habitat preference), also greatly influence their dispersal capabilities. My study identified five instances where marine barriers promoted the diversification of independent evolutionary lineages that could potentially represent species complexes. Some of them supported by evidence of population differentiation from previous morphological analyses as well as by my geometric morphometric analyses. Major environmental variables driving population differentiation in the TEP are depth, temperature, chlorophyll α altogether with spatial components, while in the TA suspended particle matter also influences diversification.

At the long-term scale, my results suggest that depth is a primary driver of speciation in the TEP, leading to niche divergence between tide pool- and reef-associated clades. In contrast, in the TA, patterns of environmental association appeared more intricate, where depth, temperature, chlorophyll α and physical features significantly contributing to speciation in this region. Finally, our time-calibrated analyses at macroevolutionary scales elucidated an Eastern Atlantic origin of the clade followed by an east-to-west dispersal. Although the historical break attributed to the rise of the Isthmus of Panama had a substantial influence on the evolutionary history of the genus, our analyses demonstrate that it did not triggered synchronous cladogenetic events. In summary, by using a combination of population genomics, comparative phylogeography, phylogenomics, seascape genomics, and geometric morphometric approaches, this study highlights major contemporary and historical barriers hindering population connectivity in the TEP and TA biogeographic regions, enhancing our understanding of the forces and processes generating new species in marine systems (Chapter III, to be submitted for publication).

All in all, my thesis highlights that the use of genome-wide data provides unprecedented resolution to unveil patterns of genetic structure, commonly unraveling cryptic diversity, and the opportunity to address species delimitation problems. By uncovering the spatio-temporal genetic patterns of fishes along the evolutionary continuum, my dissertation provides novel insights into the evolutionary and biogeographic history of marine and freshwater Neotropical fishes. Overall, my dissertation not only helps to understand the evolutionary history of the species under study, but more generally, elucidate factors driving evolutionary process in the marine realm, ranging from population-level scales, to speciation, to higher level relationships among groups.

Contents

1	**Genomics overrules mitochondrial DNA, siding with morphology on a controversial case of species delimitation**	**1**
1.1	Abstract……...	1
1.2	Introduction…….	2
1.3	Materials and methods…….	3
1.4	Results…….	6
1.5	Discussion……...	9
1.6	Conclusion……...	11
1.7	References……...	12
2	**Genome-wide species delimitation analyses of a silverside fish species complex in central Mexico indicate taxonomic over-splitting**	**19**
2.1	Abstract……...	19
2.2	Introduction…….	20
2.3	Materials and methods…….	22
2.4	Results…….	26
2.5	Discussion……...	34
2.6	Conclusion……...	39
2.7	References……...	40
3	**Beneath the waves: depth, temperature, and spatial components driving genetic differentiation at micro and macroevolutionary scales in tropical blennies**	**50**
3.1	Abstract……...	50
3.2	Introduction…….	51
3.3	Materials and methods…….	53
3.4	Results…….	55
3.5	Discussion……...	62
3.6	Conclusion……...	66
3.7	References……...	67
A	**Appendix: Supplementary Material for Genomics overrules mitochondrial DNA, siding with morphology on a controversial case of species delimitation**	**74**
B	**Appendix: Supplementary Material for Genome-wide species delimitation analyses of a silverside fish species complex in central Mexico indicate taxonomic over-splitting**	**93**
C	**Appendix: Supplementary Material for Beneath the waves: depth, temperature, and spatial components driving genetic differentiation at micro and macroevolutionary scales in tropical blennies**	**110**

List of figures

1.1 Genetic structure of WA red snappers…… 7
1.2 Phylogenetic and principal component analyses (PCAs)…… 8
1.3 Timeline of the controversial species delimitation of WA red snappers and present resolution using genomic data…… 12
2.1 Sampling localities of the humboldtianum group across the central Mexico plateau…….. 21
2.2 Multivariate and admixture results based on all, neutral, and outlier SNP loci from the matrix D (3564-snps loci), which explain the highest percentage of explained variation in the analyses……... 28
2.3 ML phylogenetic tree based on 3564 SNP loci, multi-coalescent species tree considering circa 3400 neutral SNPs, and phylogenetic inference based on *mt*DNA……. 31
3.1 Population structure at the microevolutionary scales in the Tropical Eastern Pacific and the Tropical Atlantic regions…… 57
3.2 Landmarks scheme and geometric morphometric analyses at micro and macroevolutionary scales……. 59
3.3 Seascape genomic analyses of *Malacoctenus*, illustrating main environmental features driving divergence at both intraspecific and interspecific levels in the Tropical Eastern Pacific and Tropical Atlantic…… 60
3.4 A time-calibrated species tree, based on 28K SNPs with inferred paths of dispersal and colonization events within the genus…….. 61

List of tables

2.1 Results of bayes factor delimitation (BFD*) analyses for the *humboldtianum* group using three SNPs subsets (ranging from 39 to 59 individuals, and 411–1102 SNP loci)…….. 32

2.2 Genomic diversity of 3482 neutral SNPs loci estimates for the *humboldtianum* group, under each species delimitation hypothesis examined in this study (i–iv)……..... 33

Chapter 1

Genomics overrules mitochondrial DNA, siding with morphology on a controversial case of species delimitation

Published in *Proceedings of the Royal Society B* (http://dx.doi.org/10.1098/rspb.2018.2924)

Carmen del R. Pedraza-Marrón*, Raimundo Silva, Jonathan Deeds, Steven M. Van Belleghem, Alicia Mastretta-Yanes, Omar Domínguez-Domínguez, Rafael A. Rivero-Vega, Loretta Lutackas, Debra Murie, Daryl Parkyn, Lew Bullock, Kristin Foss, Humberto Ortiz-Zuazaga, Juan Narváez-Barandica, Arturo Acero, Grazielle Gomes, Ricardo Betancur-R.*

(*) *equal contribution*

1.1 Abstract

Species delimitation is a major quest in biology and is essential for adequate management of the organismal diversity. A challenging example comprises the fish species of red snappers in the Western Atlantic. Red snappers have been traditionally recognized as two separate species based on morphology: *Lutjanus campechanus* (northern red snapper) and *L. purpureus* (southern red snapper). Recent genetic studies using mitochondrial markers, however, failed to delineate these nominal species, leading to the current lumping of the northern and southern populations into a single species (*L. campechanus*). This decision carries broad implications for conservation and management as red snappers have been commercially over-exploited across the Western Atlantic and are currently listed as vulnerable. To address this conflict, we examine genome-wide data collected throughout the range of the two species. Population genomics, phylogenetic and coalescent analyses favor the existence of two independent evolutionary lineages, a result that con-firms the morphology-based delimitation scenario in agreement with conventional taxonomy. Despite finding evidence of introgression in geo- graphically neighboring populations in northern South America, our genomic analyses strongly support isolation and differentiation of these

species, suggesting that the northern and southern red snappers should be treated as distinct taxonomic entities.

1.2 Introduction

Delimitation of species—the basic unit of biological diversity—is of great interest across many fields in biology. The adoption of molecular information for species delimitation analyses has unveiled cryptic diversity across several taxa [1,2]. Initial approximations that integrated genetic markers, such as mitochondrial DNA (mtDNA) or scant nuclear DNA (nDNA) fragments, into traditional taxonomy provided greater resolution for a broad array of groups [1], from marine corals and fishes [3,4] to terrestrial fungi and mammals [5,6]. Mitochondrial and single nuclear markers, however, are not always efficient tools [7 – 9], and can at times fail to discriminate species correctly [10]. This is exemplified by the often incongruent genealogies inferred from different genetic loci that identify conflicting histories [11], which can ultimately arise from incomplete lineage sorting (ILS) or introgression [7,9,12] and reveal the history of the genes examined rather than that of the species [13]. Although mtDNA markers, widely used in molecular barcoding, have proven powerful at detecting cryptic species (e.g. fishes [3,14]), there are few examples in natural populations where mitochondrial-based approaches conflict with both conventional taxonomy and genomic inferences (e.g. sharks [15], lampreys [16], caddisflies [17]; see [18] for a review). Recently, the advent of high-throughput sequencing technologies has facilitated the generation of large-scale datasets with thousands of markers for high resolution of shallow evolutionary inferences [19], further allowing the elucidation of complex speciation scenarios (e.g. [17,20,21]). Uncovering signals of population and species differentiation with genome- wide molecular information is now becoming mainstream [2,22] and permits the rigorous validation of relationships that were previously inferred from single or few genetic loci.

Here, we address a controversial case of species delimitation of red snappers (Teleostei: Lutjanidae) in the Western Atlantic (WA) where mtDNA has delimited fewer species than initially documented. For over a century, two allopatric species of red snappers have been recognized on the basis of morphological and meristic traits, including number of scales in the lateral line (or scale counts in rows above and below the lateral line) and modal differences in anal fin ray counts [23]. The northern red snapper, *Lutjanus campechanus* (Poey, 1860), is distributed along the US East coast and the Gulf of Mexico; the southern red snapper, *Lutjanus purpureus* (Poey, 1866), occurs in the Caribbean Sea and southwards through northeastern Brazil. Recent attempts to investigate population genetic structure and to evaluate the degree of similarity of red snappers using mtDNA sequences [24,25] failed, however, to discriminate the nominal species as independent evolutionary groups. These studies have ultimately suggested that the northern and southern red snappers constitute a single species (*L. campechanus*) that exhibits phenotypic variability throughout the WA. This decision has been recently adopted by several taxonomic authorities [26,27], carrying down- stream repercussions for conservation and fisheries management. The conflicting morphological and mitochondrial evidence has raised a controversial case of species delimitation where an accurate taxonomic demarcation is of particular concern, as

red snappers have been widely overfished and are currently listed as vulnerable by the IUCN Red List of Threatened Species [28].

Using genome-wide markers generated via RAD sequencing approaches, we test the discordance between the mtDNA- and morphology-based hypotheses that has led to a continuing conflict of species delimitation in WA red snappers. We show that southern and northern red snappers represent two independent evolutionary lineages that should be recognized as distinct species. We highlight the importance of using genomic approaches to reconcile complex species delimitation scenarios where different lines of evidence conflict.

1.3 Materials and methods

Sampling

We examined a total of 178 red snapper individuals (105 of *L. campechanus* and 73 of *L. purpureus*) collected from 15 locations across the WA (figure 1.1; Appendix A, table S1). Georeferenced data are available for most sampling sites; for others, the approximate location was inferred by interpreting collecting site descriptions. We also attempted to acquire specimens that would fill the sampling gap through the Caribbean Islands or intermediate populations in Central America. Although we actively searched for over 2 years in two key Caribbean locations (Puerto Rico and San Andre´s Island, Colombia), all surveys were unsuccessful (see details in Appendix A, figure S1). Given the apparent scarcity of the species in the region, other Caribbean locations were also ineffectively probed for samples through networking efforts. To emphasize the low abundance of red snappers in many Caribbean localities, we generated a map with total records for both species using reports available from FishNet (www.fishnet2.net) and the Ocean Biogeographic Information System (OBIS) (www.iobis.org) (Appendix A, figure S1).

Molecular protocols, mitochondrial data and SNP genotyping

All individuals examined were sequenced using restriction-digest-associated DNA sequencing (RADseq) approaches by applying the double-digest (ddRADseq) protocol of Peterson *et al.* [29]. This technique allows for the low cost discovery and genotyping of thousands of genetic markers and is particularly useful for non-model organisms [30]. In order to compare the population structure using genome-wide RADseq markers to that obtained with mtDNA (e.g. as in previous studies [24,25]), a subset of 83 samples was barcoded using the mtDNA gene cytochrome-c oxidase subunit I (*COI*) following standard protocols [31] (Appendix A, table S2). Additional mtDNA sequences for *COI* and *D-loop* were downloaded from available data on NCBI (Appendix A, table S3). ddRADseq data were processed using several packages, including STACKS v1.49 [32], FASTQC v0.11.5 (www.bioinformatics.babraham.ac.uk/pro-jects/fastqc/), TASSEL v5.2.43 [33], and VCFTOOLS v0.1.15 [34]. Different combinations of assembly parameters were first tested on a subset of 30 samples (following [35]) in STACKS. Final locus assembly was performed using a minimum of five raw reads required to form a stack, and allowing a maximum of two mismatches between stacks and three mismatches between loci of different individuals. Loci with a minimum allele frequency of 0.05 and a maximum observed heterozygosity of 0.70 were further

excluded as potential paralogues. The sensitivity of results to number of individuals and missing data was also evaluated by applying a variety of filters. Four datasets that contained between 21 431 and 55 795 loci were first selected based on loci present in multiple predetermined numbers of populations (p) and percentage of individuals (r) (p11r50, p12r50, p9r60 and p8r60). A second filter ('min. sites') was applied after removing individuals with different thresholds for missing sites (0.75, 0.50, 0.25 and 0.05). These filters resulted in 20 datasets (Appendix A, table S4), of which six were further selected according to the amount of missing data (9 – 46%), number of individuals present (44 – 155), number of SNPs (15 112 – 42 406), and number of populations (8 – 15). Additional details on molecular protocols for *de novo* assembly of RAD loci are given in the Appendix A.

Phylogenetic and coalescent analyses

A phylogenetic network was computed based on 15 112 SNPs with the Neighbor-Net algorithm in SPLITSTREE v4.14.6 (www.splitstree.org). Phylogenetic reconstruction was performed in a maximum-likelihood (ML) framework using the software RAxML v8 [36]. Trees were inferred for the six SNP datasets selected in the previous step. All invariant sites were removed from the matrices using the R package phrynomics (https://github.com/bbanbury/phrynomics/). Two alternative ML ana- lyses were performed: one in which heterozygous alleles were collapsed using ambiguity (IUPAC) codes, and another using concatenated variants (extracted in order of appearance from the VCF file). No major differences were found between trees reconstructed from these variants and only the latter trees are reported. To account for acquisition biases inherent to SNP datasets [37], we used the GTR + I model with ascertainment bias correction (ASC) in RAxML. Nodal support was assessed in RAxML using 100 rapid bootstrap replicates. For the mitochondrial matrix, haplotype networks of 654 bp *COI* and 858 bp *D-loop* were constructed using the TCS Network available in POPART [38]. The *COI* sequences were aligned using available references of *L. campechanus* and *L. purpureus* from GenBank (accession no. EU752115 and EU752118).

We also assessed the fit of the two alternative scenarios of species delimitation in WA red snappers in a coalescent frame- work. We used the Bayes factor delimitation (BFD*) method implemented for genome-wide SNP data [39] in the programs SNAPP v1.3.0 [40] and BEAST2 v2.4.1 [41]. To reduce computational burden, we first applied additional filters to the p12r50 dataset (with 15 112 SNPs from 15 populations; see above) by retaining both loci and individuals from each population with the lowest proportions of missing data. Three subsets with 58 – 108 individuals and 149 – 957 loci were assembled (see Appendix A). To set up priors and MCMC runs, we carefully followed the guidelines outlined in the BFD* tutorial by A. Leaché (http://www.beast2.org/bfd/). Because the scenarios tested contained fewer than three species-tree nodes (e.g. for one species the leaf node is also the root node), we removed all tree operators from the analyses (R. Bouckaert 2019, personal communication). Finally, we compared the marginal likelihood estimates for the alternative scenarios using Bayes factors.

Population structure analyses

Principal component analyses (PCAs) were first computed on allele frequencies using TASSEL v5.2.43 [33]. The analyses for p12r50 and p8r60 matrices were performed with three different proportions of 'min. sites' (0.75, 0.25 and 0.5), and figures were plotted using the R [42] package ggplot [43]. The p12r50 matrix was selected for downstream analyses, as this dataset captures the genetic information of 155 individuals from all 15 populations while maximizing population discrimination. Next, the fastStructure v1.0 [44] package, a Bayesian clustering method, was used for inferring population structure. The number of population clusters was evaluated by running multiple values of K (1 – 18) using a logistic prior. The best-fitting model complexity was selected with the chooseK.py routine and the resulting K value was re-run through fastStructure 25 times with multiple random seeds to identify the five highest values of the log-marginal likelihood (LLBO). Final plots were constructed using disruct.py, available from fastStructure. Lastly, Weir and Cockerham F_{ST} values were estimated using the R package hierfstat [45] using 100 bootstrap replicates. Because large amounts of loci with missing data can deviate true values of summary statistics [46], we calculated F_{ST} values from two datasets: (i) p12r50-155, as previously selected; and (ii) p12r50-89, excluding three populations with few represented loci. F_{ST} values were plotted as heatmaps using the R package gplots [47].

Isolation by distance

Limited dispersal capabilities in panmictic populations often result in a correlation between geographical distance and genetic differentiation among populations—a process termed isolation-by-distance (IBD) [48]. To test whether the red snapper populations follow a pattern of IBD, we performed a Mantel test using a correlation and a major axis correction [49,50] between Weir and Cockerham F_{ST} values among populations and their corresponding geographical distances (including 15 populations). Geographical distances were calculated via the least cost path (LCP) distance over seawater using the R package marmap [51]. We constrained the LCP to depth values between 10 and 190 m, which constitute the depth range of suitable habitat for red snappers [26]. Because these species can also disperse through oceanic currents during their pelagic larval phase, additional mantel tests were conducted using Euclidean geo- graphical distances (computed with the R package adegenet [52]). Results of Mantel tests were not affected by the use of LCP or Euclidean distances; therefore, only the former results are reported (figure 1.1*f*).

Hybridization

In order to test for ongoing hybridization between the two species, we used the R package gghybrid to estimate the hybrid index (HI)—a measure of genetic admixture within individuals [53]. The gghybrid package uses a Bayesian algorithm on bi-allelic genomic data to calculate the proportion of allele copies coming from parental reference sets [54] while applying a logit- logistic model for the genomic cline curve [54,55]. We ran HI estimations using 10 000 MCMC iterations and estimated posterior probability values after a 5000 iteration burnin. We selected the northernmost populations of *L. campechanus* (Florida and Apalachicola) and the southernmost populations of *L. purpureus* (Fortaleza and Salvador) as parental references in order to reduce the probability of gene exchange between major lineages. By selecting populations with a low probability of contact, the analysis focuses on loci that are highly differentiated in the parental reference populations.

1.4 Results

In agreement with previous studies [24,25], our mtDNA haplotype networks fail to delimit the nominal species as distinct haplogroups (figure 1.1*a*,*b*). The *COI* network shows an intermingling of *L. campechanus* and *L. purpureus* (figure 1.1*a*), whereas the *D-loop* network identifies one haplogroup formed solely by *L. purpureus* individuals and another where haplotypes of *L. campechanus* are nested within the *L. purpureus* populations (figure 1.1*b*). Similarly, the mtDNA COI tree lacks resolution and reveals no geographical segregation of individuals based on unique haplotypes (figure 1.2*a*). By contrast, trees inferred with genome-wide RADseq data (15 112–42 406 loci) consistently resolve two divergent and well supported reciprocally monophyletic groups (bootstrap support = 100%) that match the established species boundaries for *L. campechanus* and *L. purpureus* (figures 1.1*e* and 1.2*a*). There is no apparent pattern of geographical segregation within each clade, as individuals are not clustered in the SNP-based trees by populations/locations. Coalescent-based analyses using the BFD* method also provide overwhelming support in favor of the two species delimitation scenario (Bayes fac tors for two versus one species 2310.28 – 22 356.22; see details in Appendix A, table S5). A list of diagnostic SNPs differentiating populations of *L. campechanus* from *L. purpureus*, which can be used for barcoding purposes, is given in the Appendix A, table S6.

Population structure results based on fastStructure analyses of SNP data delimit the northern and southern lineages as separate units (figure 1.1*c*,*d*), with a best-fitting model supporting two meta-populations ($K = 2$). Although there have been recent concerns that structure analyses tend to be biased in favor of $K = 2$ [56], we note that our K scheme is consistent with the results inferred using multiple lines of evidence (figures 1.1*e* and 2; Appendix A, table S5). In the PCAs of RAD loci, the first principal component accounts for 21 – 27% of the variation and is congruent with the separation of *L. campechanus* from *L. purpureus* (figure 2*b*; Appendix A, figure S3). The second principal component represents 1 – 6% of the genetic variation, resulting in scattered populations on a cline that unveils fine-scale patterns of population structure according to geography (e.g. Veracruz-Tuxpan and Alabama-Louisiana define slope extremes of *L. campechanus*; the same is true for Guajira and Fortaleza in *L. purpureus*). While results using the 42406 SNPs matrix (figure 1.2*b*; Appendix A, figure S3d–f) show a much clearer species demarcation relative to the 15 112 SNPs matrix (figure 1.2*b*; Appendix A, figure S3a–c), PCAs overall identify the same clustering patterns regardless of the number of SNPs analyzed. Main variations on the observed genetic groups were influenced by the number of individuals contained ('min. sites' filter) in each PCA analysis, where PCAs generated with 'min. site 0.05' superimpose populations from both species that contained the highest amount of missing data (e.g. Yucatán; figure 1.2*b*; Appendix A, figure S3a). These results emphasize that missing data can bias the results obtained with large RADseq datasets [57].

Weir and Cockerham F_{ST} values are substantially lower at intra- versus inter-specific levels (Appendix A, figure S4). *Lutjanus campechanus* shows genetic differences between 0.0010 and 0.0119 in Tabasco- Veracruz and Veracruz-Apalachicola respectively, whereas *L. purpureus* presents a range of F_{ST} values from 0 in Fortaleza-Salvador and 0.0588 in Fortaleza-La Guajira.

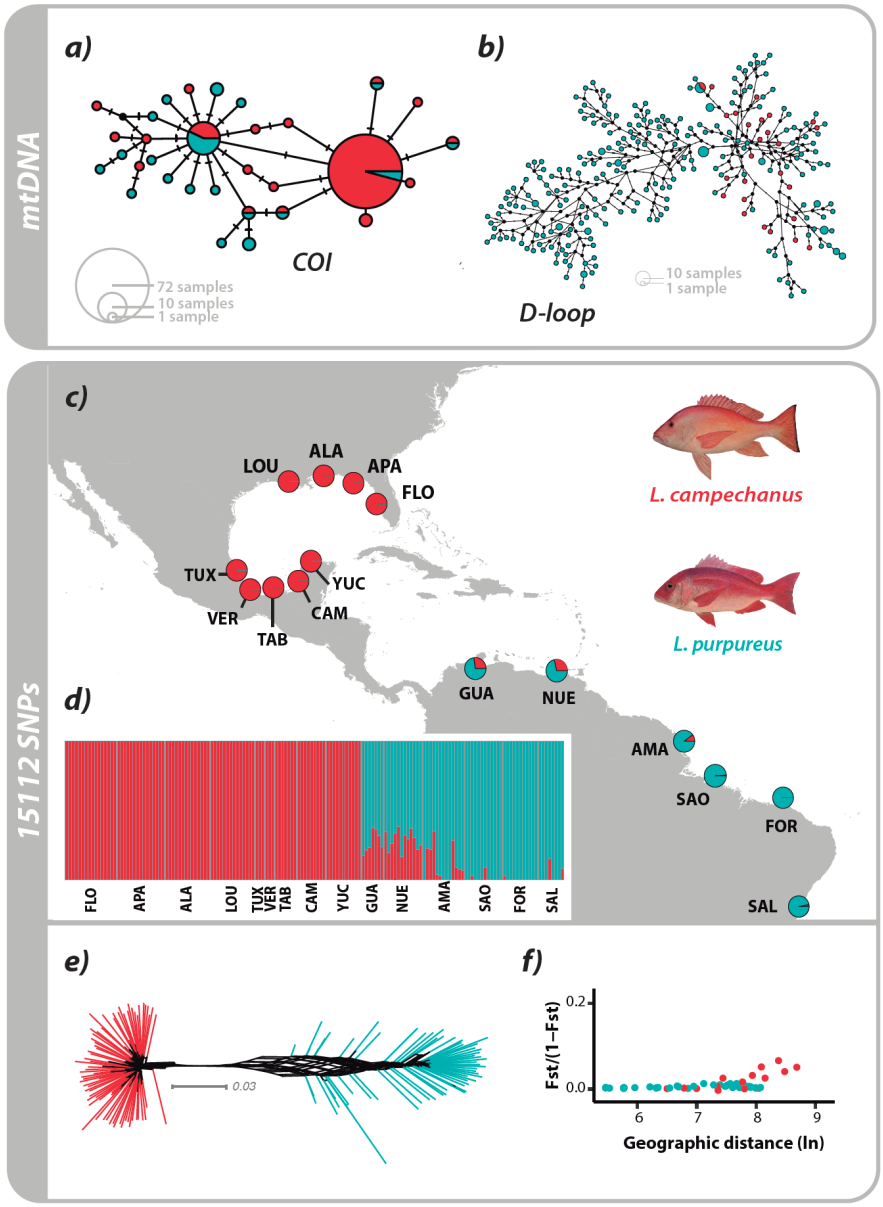


Figure 1.1 Genetic structure of WA red snappers. The northern red snapper (*Lutjanus campechanus*; red) and the southern red snapper (*Lutjanus purpureus*; green) are recognized as two separate species by conventional taxonomy on the basis of morphological characters. Consistent with previous studies, haplotype networks based on mitochondrial DNA sequences lack discriminatory power at the species level (a, *COI*; b, *D-loop*). However, a Bayesian structure analysis using 15,112 genome-wide SNPs identifies two main genetic clusters (K = 2) that are concordant with the traditional taxonomic delineations. Average admixture proportions were calculated for either (c) populations or (d) individuals (each structure bar representing the probability of assignment to each cluster). These results are congruent with (e) the estimated phylogenetic network (see additional trees in Fig. 1.2). (f) Correction of Mantel correlogram between Weir and Cockerham F_{ST} values versus least cost path geographic distances provide little support for a model of isolation by distance for intraspecific comparisons. Population information, descriptions, and abbreviations are given in Table S1 (Appendix A).

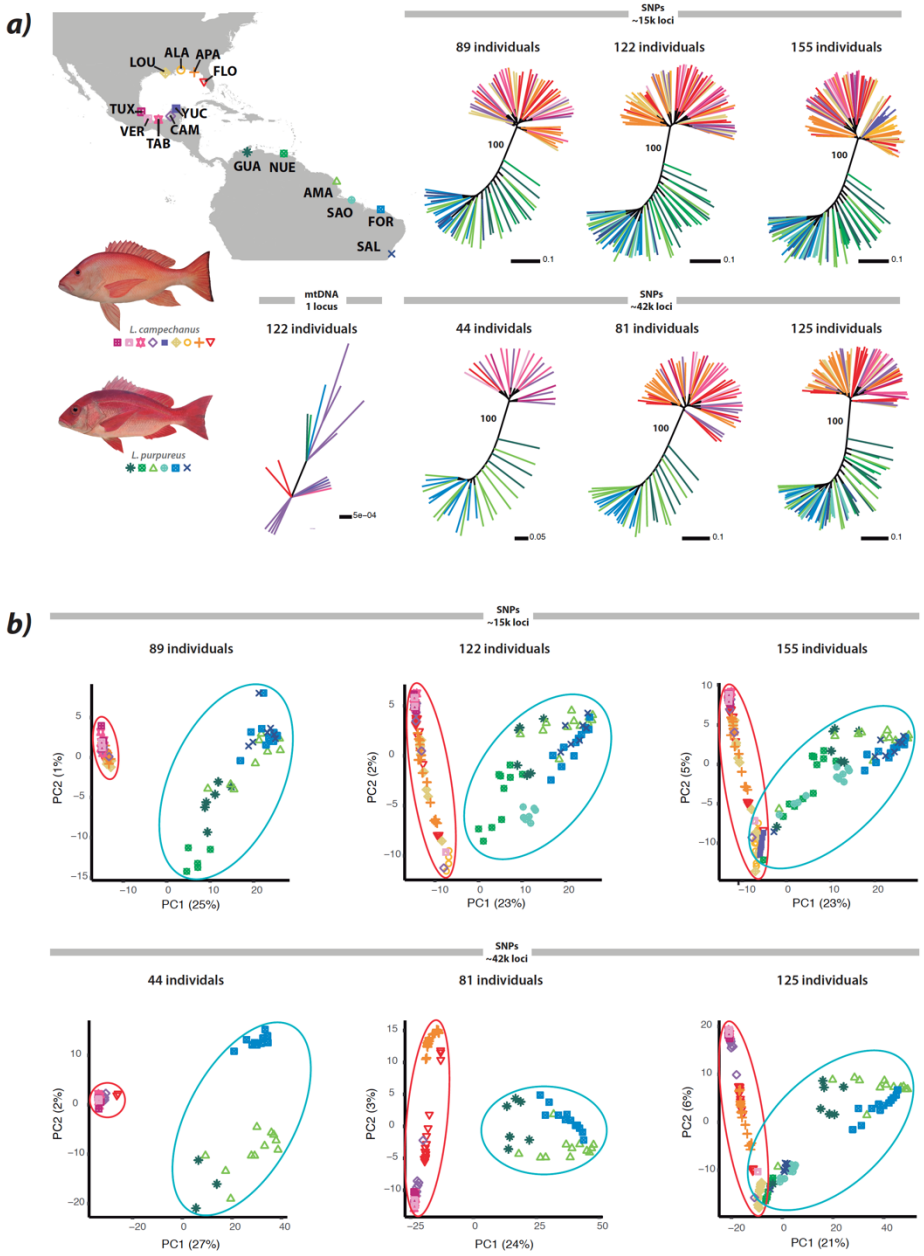


Figure 1.2 (a) Phylogenetic and (b) principal component analyses (PCAs). Phylogenetic trees and PCAs based on ~15,000-42,000 SNPs resolve two well-differentiated clusters that are congruent with the morphology-based hypothesis. These results are largely consistent regardless of the number of individuals or loci analyzed. As in previous studies, a phylogenetic tree based on mtDNA (a; *COI* sequences) fails to identify genetic structure that aligns with the recognized species boundaries. Trees and PCAs plots are color-coded according to their geographic location in map. Abbreviations in map correspond to locality information given in Appendix A Table S1; see also Appendix A Fig. S3 for additional details on PCAs.

Negative and zero F_{ST} values were common across adjacent populations for both species, suggesting higher genetic differences at intra- rather than inter-population scales (i.e. individuals from adjacent locations may form a panmictic population [58,59]). By contrast, interspecific comparisons show substantially higher F_{ST} values, varying from 0.1240 among geographically closer populations (Veracruz-Nueva Esparta) to 0.3406 among the more distant comparisons (Fortaleza-Apalachicola).

Results of Mantel tests are marginally significant when the northern and southern lineages are each analyzed in isolation ($p = 0.06$ for *L. campechanus*; $p = 0.02$ for *L. purpureus*; Appendix A, figure S5a,b). No substantial association between genetic and geographical distances (IBD), however, is detected for Euclidean nor LCP distances, as the points do not form continuous linear plots (figure 1.1*f*; Appendix A, figure S5c,d). Admixture plots from fastStructure reveal introgression (figure 1.1*c*,*d*) at geographically neighboring populations between the two species in northern South America. This result is further confirmed with the hybrid index estimated with gghybrid (Appendix A, figure S6), which identifies admixture in geographically inter- mediate populations in Colombia (La Guajira), Venezuela (Nueva Esparta), and to some extent Brazil (Amapá). By contrast, none of the populations of *L. campechanus* reveal signs of ongoing introgression. Taken together, these results indicate a pattern of unidirectional interspecific introgression in *L. purpureus* from *L. campechanus*.

1.5 Discussion

Species delimitation

Western Atlantic red snappers represent two commercially important species whose taxonomic status has been recently challenged. Described on the basis of morphological and meristic traits [23], *Lutjanus campechanus* and *L. purpureus* were recently considered to be conspecific based on assessments of genetic structure that examined mitochondrial DNA sequences and failed to delineate the formerly recognized species boundaries [24,25]. Despite finding concordant results with previous studies based on expanded mitochondrial *COI* and *D-loop* sequences (figures 1.1*a*,*b* and 1.2*a*), genome-wide analyses implementing SNP data (approx. 15 000 – 42 000) identify remarkable genetic divergences between the northern and southern red snappers. These results are supported by structure analyses (figure 1.1*c*,*d*,*f*), phylogenetic inferences (figures 1*e* and 2*a*), coalescent tests (Appendix A, table S5), PCAs (figure 1.2*b*), and geographical patterns of population abundances (from FishNet and OBIS; Appendix A, figure S1b), all of which are in agreement with the morphospecies delimitation and are largely robust to the number of individuals, SNPs, or missing data included in each of the data matrices analyzed. The mito-nuclear discordance observed for WA red snappers can be the result of mitochondrial introgression or incipient sorting of mitochondrial haplotypes—the most likely biological sources of genealogical incongruence among recently diverged species [9]. Notably, a recent unpublished study that compared the otolith shape among different populations and species of WA red snappers using geometric morphometric approaches also identified well-differentiated and non-overlapping clusters that are consistent with the evolutionary units delineated here using genomic data [60].

While Mantel tests and correlogram analyses indicate that intraspecific populations separated by vast geographical distances have a smaller likelihood of gene flow, our analyses find tenuous support for a pattern of intraspecific IBD (figure 1.1*f*; Appendix A, figure S5). We find evidence, however, of ongoing interspecific hybridization (figure 1.1*c,d*) through circulating gene flow across geographically neighboring locations in northern South America (figure 1.1*c*,*d*; Appendix A, figure S6). Interspecific hybridization is not rare across sister species of marine fishes (e.g. *Haemulon maculicauda* and *H. flaviguttatum* [61]) and could also lead to mito-nuclear discordance via genetic introgression [18]. In the face of introgression, genetic structure is expected to reflect geographical patterns, particularly when sister species pairs become geographically isolated and subsequently come into secondary contact [17,61]. In this case, geographical patterns appear to support secondary admixture over incomplete lineage sorting (ILS), where ongoing hybridization leads to nuclear introgression [17,18].

Remarkably, the apparent direction of the aforementioned introgression (north to south) runs counter to the progression of marine currents in the Greater Caribbean (south to north). Although a northward route seems more plausible than the reverse, this is not entirely unexpected in light of the complex patterns of connectivity in the Greater Caribbean [62,63]. For instance, the progression of the lionfish invasion in the area has taken place southwards, from Florida to South America [64]. An alternative explanation to the observed pattern is that genetic structure reflects the maintenance of ancestral polymorphisms (ILS), possibly as a result of the recent divergence of the species. This interpretation, however, seems unfeasible considering that cline analyses (gghybrid) account for ancestral shared polymorphisms by focusing on loci that are highly differentiated in the parental reference populations. Another possibility involves secondary contact after divergence between the two species, at a time when *L. campechanus* co-occurred in the south, and either southern *L. campechanus* populations are now extinct or they have been diluted into a dominant *L. purpureus* genetic and demographic background. All these possibilities remain to be explored in greater depth using demographic and migration tests.

Although we were unable to examine samples from the western and northern Caribbean region (Appendix A, figure S1), this does not necessarily represent a caveat in our study. The scarcity of records over time in the Bahamian, Eastern Caribbean, Greater Antilles, and Southwestern Caribbean marine ecoregions (regionalization according to [65]; Appendix A, figure S1) suggest that the populations of red snappers are not completely established, possibly formed by vagrant individuals. This, in fact, reflects and actual gap in the connectivity of the two species that reinforces our observations and emphasizes a regional discontinuity pattern. Notwithstanding a worst-case scenario with well-established intermediate populations in the Caribbean and a smooth cline of admixture between the northern and southern lineages, the vast genomic divergences observed between these lineages provide strong evidence for the delimitation of two discrete taxonomic units. For instance, while low gen- etic differentiation values were estimated intraspecifically despite great geographical distances, the closest interspecific locations sampled feature high genetic divergences. Populations of *L. purpureus* from Nueva Esparta and São Salvador da Bahia are separated by an F_{ST} of 0.04 and an LCP of 4819 km; *L. campechanus* from Campeche and Florida have an F_{ST} of 0.003 and an LCP of 2953 km. These results are congruent with observed values proposing panmictic within- species populations [59,66]. Conversely, the

corresponding interspecific values between Puerto de Tuxpan (northern red snapper) and La Guajira (southern red snapper) are 0.18 and 4847 km, respectively (Appendix A, figure S4a).

It is important to note that we are not splitting species here based on genetic differences alone (e.g. [6]). Instead, we are testing the morphological and mitochondrial hypotheses in light of analyses based on thousands of genetic markers, with the former setting a century-long precedent on the validation of two species. In our present scheme, finding the cline between the two lineages would be difficult, as the lack of samples from intermediate locations precludes the determination of accurate geographical boundaries and the extent of the hybrid zone. Novel approaches allow the delimitation of species in the presence of gene flow [67]; however, these require gene trees as input, which is unfeasible using SNP data. Therefore, although we cannot confidently assert that these populations represent two valid species under the Biological Species Concept (BSC), they do represent two well-defined entities that match the Phylogenetic Species Concept (PSC)—the most commonly used criterion to delimit species in ichthyology [3]. Given that similar controversies exist about the specific taxonomic status of other living and extinct organisms (e.g. Neanderthal and Denisovan hominids [68]), debating whether these lineages fail to match particular aspects of species concepts can be a difficult and possibly futile endeavour. As Darwin notes: '.. .to discuss if they are rightly called species or varieties, before any definition of these terms has been generally accepted, is vainly to beat the air' [69, p. 49].

Conservation Implications

Red snappers represent some of the most economically important commercial and recreational fisheries in the Western Atlantic (WA), generating estimated annual revenues of over USD $27 million in the US alone [70]. Such fishing pressures have had an adverse effect on their populations, leading to heavily overfished stocks [71]. Delimitation of their species boundaries is imperative as the IUCN only lists the northern red snapper as Vulnerable (the southern red snapper has not been evaluated) [28]. Generally, an accurate evaluation of species—in particular commercially important and threatened species—represents the basic scientific knowledge required to determine conservation status that is assessed by international conservation organizations including the IUCN and the Food and Agriculture Organization of the United Nations (FAO), as natural species do not follow political delimitations. Even though the northern and southern stocks are managed by different legal entities, it is crucial to include genetic information as a baseline for planned stock enhancement [72] in multiple countries. Finally, correct delimitation of species also has implications for the enforcement of seafood mislabeling given that only *L. campechanus* has been traditionally allowed to use the US market name 'red snapper'; other species labelled as red snapper, including *L. purpureus*, are considered misbranded [73].

1.6 Conclusion

Our genome-wide analyses provided a strong signal of genetic differentiation among the northern and southern red snappers in the Western Atlantic, reconciling a long-standing conflict of species delimitation between mtDNA and morphology (figure 1.3). These results highlight the importance

of using powerful markers for addressing complex species delineation problems, particularly with organisms that rely on accurate recognition of species boundaries to inform their conservation status. We conclude that the two red snappers should be managed as separate taxonomic units. Our findings further emphasize the importance of implementing genomic approaches to settle species delimitation disagreements, where more conventional methods that lack discerning power may lead to the underestimation of biological diversity. These results ultimately align with observations from recent studies [15,16], which recommend that taxonomic decisions should strive to be conservative when based on single locus inferences, as those can be affected by incomplete lineage sorting or introgression in recent speciation events.

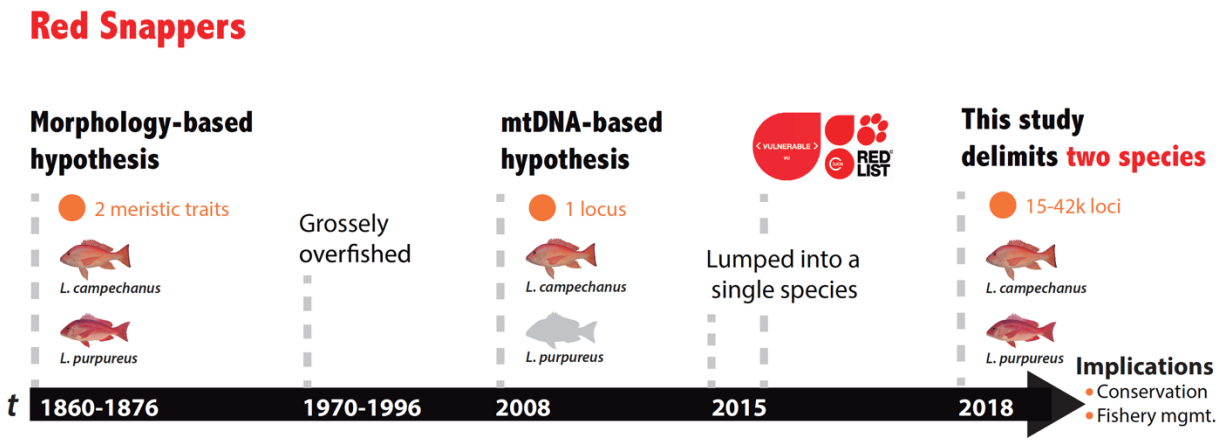


Figure 1.3 Timeline of the controversial species delimitation of WA red snappers and present resolution using genomic data.

1.7 References

1. Thielsch A, Knell A, Mohammadyari A, Petrusek A, Schwenk K. 2017 Divergent clades or cryptic species ? Mito-nuclear discordance in a *Daphnia* species complex. *BMC Evol. Biol.* **17**, 1–9. (doi:10.1186/s12862-017-1070-4)
2. Bickford D, Lohman DJ, Sodhi NS, Ng PKL, Meier R, Winker K, Ingram KK, Das I. 2007 Cryptic species as a window on diversity and conservation. *Trends Ecol. Evol.* **22**, 148–155. (doi:10.1016/j.tree.2006.11.004)
3. Victor BC. 2015 How many coral reef fish species are there? Cryptic diversity and the new molecular taxonomy. In *Ecology of Fishes on Coral Reefs: The Functioning of an Ecosystem in a Changing World. Cambridge University Press, Cambridge, United Kingdom*, pp. 76–87.
4. Pinzo JH. 2011 Species delimitation of common reef corals in the genus *Pocillopora* using nucleotide sequence phylogenies , population genetics and symbiosis ecology. *Mol. Ecol.* **20**, 311–325. (doi:10.1111/j.1365-294X.2010.04939.x)
5. Balasundaram S V, Engh IB, Skrede I, Kauserud H. 2015 How many DNA markers are needed to reveal cryptic fungal species? *Fungal Biol.* **119**, 940–945.

(doi:10.1016/j.funbio.2015.07.006)

6. Fennessy J, Bidon T, Reuss F, Kumar V, Elkan P, Nilsson MA, Vamberger M, Fritz U, Janke A. 2016 Multi-locus Analyses Reveal Four Giraffe Species Instead of One. *Curr. Biol.* **26**, 2543–2549. (doi:10.1016/j.cub.2016.07.036)

7. Suchan T, Espíndola A, Rutschmann S, Emerson BC, Gori K, Dessimoz C, Arrigo N, Ronikier M, Alvarez N. 2017 Assessing the potential of RAD-sequencing to resolve phylogenetic relationships within species radiations: The fly genus *Chiastocheta* (Diptera: Anthomyiidae) as a case study. *Mol. Phylogenet. Evol.* **114**, 189–198. (doi:10.1016/j.ympev.2017.06.012)

8. Funk DJ, Omland KE. 2003 Species-Level Paraphyly and Polyphyly: Frequency, Causes, and Consequences, with Insights from Animal Mitochondrial DNA. *Annu. Rev. Ecol. Evol. Syst.* **34**, 397–423. (doi:10.1146/annurev.ecolsys.34.011802.132421)

9. Mutanen M *et al.* 2016 Species-level para- and polyphyly in DNA barcode gene trees: Strong operational bias in European Lepidoptera. *Syst. Biol.* **65**, 1024–1040. (doi:10.1093/sysbio/syw044)

10. Spinks PQ, Thomson RC, Shaffer BH. 2014 The advantages of going large : genome-wide SNPs clarify the complex population history and systematics of the threatened western pond turtle. *Mol. Ecol.* **23**, 2228–2241. (doi:10.1111/mec.12736)

11. Martin SH, Belleghem SM Van. 2017 Exploring Evolutionary Relationships Across the. *Genetics* **206**, 429–438. (doi:10.1534/genetics.116.194720/-/DC1.1)

12. Toffoli D, Hrbek T, de Araújo MLG, de Almeida MP, Charvet-Almeida P, Farias IP. 2008 A test of the utility of DNA barcoding in the radiation of the freshwater stingray genus *Potamotrygon* (Potamotrygonidae, Myliobatiformes). *Genet. Mol. Biol.* **31**, 324–336. (doi:10.1590/S1415-47572008000200028)

13. Roberts MA, Schwartz TS, Karl SA, August S. 2004 Global Population Genetic Structure and Male-Mediated Gene Flow in the Green Sea Turtle (*Chelonia mydas*): Analysis of Microsatellite Loci. *Genetics* **166**, 1857–1870.

14. Milá B, Tassell JL Van, Calderón JA, Rüber L, Zardoya R. 2017 Cryptic lineage divergence in marine environments : genetic differentiation at multiple spatial and temporal scales in the widespread intertidal goby *Gobiosoma bosc*. *Ecol. Evol.* **7**, 5514–5523. (doi:10.1002/ece3.3161)

15. Corrigan S, Maisano P, Eddy C, Duffy C, Yang L, Li C, Bazinet AL, Mona S, Naylor GJP. 2017 Molecular Phylogenetics and Evolution Historical introgression drives pervasive mitochondrial admixture between two species of pelagic sharks. *Mol. Phylogenet. Evol.* **110**, 122–126. (doi:10.1016/j.ympev.2017.03.011)

16. Mateus CS, Stange M, Berner D, Roesti M, Quintella BR, Alves MJ, Almeida PR, Salzburger W. 2013 Strong genome-wide divergence between sympatric European river and brook lampreys. *Curr. Biol.* **23**, R649–R650. (doi:10.1016/j.cub.2013.06.026)

17. Weigand H, Weiss M, Cai H, Li Y, Yu L, Zhang C, Leese F. 2017 Deciphering the origin of mito-nuclear discordance in two sibling caddisfly species. *Mol. Ecol.* **26**, 5705–5715. (doi:10.1111/ijlh.12426)

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

19. Pie MR, Bornschein MR, Ribeiro LF, Faircloth BC, McCormack JE. 2017 Phylogenomic species delimitation in microendemic frogs of the Brazilian Atlantic Forest. *bioRxiv* **55**.

20. Martin SH *et al.* 2013 Genome-wide evidence for speciation with gene flow in *Heliconius* butterflies. *Genome Res.* **23**, 1817–1828. (doi:10.1101/gr.159426.113.)

21. Van Belleghem SM, Vangestel C, De Wolf K, De Corte Z, Möst M, Rastas P, De Meester L, Hendrickx F. 2018 Evolution at two time frames : Polymorphisms from an ancient singular divergence event fuel contemporary parallel evolution. *PLoS Comput. Biol.* **14**, 1–26. (doi:10.5061/dryad.77r93d5.Funding)

22. Karanth KP. 2017 Species complex , species concepts and characterization of cryptic diversity : vignettes from Indian systems. *Curr. Sci.* **112**, 1320–1324.

23. Anderson WD. 2002 Lutjanidae. In *FAO Species Identification Guides for Fishery Purposes: The Living Marine Resources of the Western Central Atlantic* (ed KE Carpenter), pp. 1479–1504. Rome: FAO.

24. Gomes G, Sampaio I, Schneider H. 2012 Population structure of *Lutjanus purpureus* (Lutjanidae-Perciformes) on the Brazilian coast: Further existence evidence of a single species of red snapper in the western Atlantic. *An. Acad. Bras. Cienc.* **84**, 979–999. (doi:10.1590/S0001-37652012000400013)

25. Gomes G, Schneider H, Vallinoto M, Santos S, Ortí G, Sampaio I. 2008 Can *Lutjanus purpureus* (South red snapper) be "legally" considered a red snapper (*Lutjanus campechanus*)? *Genet. Mol. Biol.* **31**, 372–376. (doi:10.1590/S1415-47572008000200035)

26. Robertson DR, Van Tassell JL. 2015 Shorefishes of the Greater Caribbean online information system. See http://biogeodb.stri.si.edu/caribbean/en/pages (accessed on 20 August 2005).

27. Eschmeyer WN. 2003 The catalog of fishes on-line. *Calif. Acad. Sci. San Fr. California. As viewed online http//www. calacademy. org/research/ichthyology/catalog/fishcatsearch. html* (Accessed 2 October 2019).

28. Anderson W, Claro R, Cowan J, Lindeman K, Padovani-Ferreira B, Rocha LA. 2015 *Lutjanus campechanus* (errata version published in 2017). The IUCN Red List of Threatened Species 2015: e.T194365A115334224.

29. 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**. (doi:10.1371/journal.pone.0037135)

30. Andrews KR, Good JM, Miller MR, Luikart G, Hohenlohe PA. 2016 Harnessing the power of RADseq for ecological and evolutionary genomics. *Nat. Rev. Genet.* **17**, 81–92. (doi:10.1038/nrg.2015.28)

31. Weigt LA, Driskell AC, Baldwin CC, Ormos A. 2012 DNA Barcoding Fishes. In *Methods in Molecular Biology*, pp. 109–126. Humana Press. (doi:10.1007/978-1-61779-591-6)

32. Catchen JM, Hohenlohe PA, Bassham S, Amores A, Cresko WA. 2013 Stacks: an analysis tool set for population genomics. *Mol. Ecol.* **22**, 3124–3140. (doi:10.1111/mec.12354.Stacks)

33. Bradbury PJ, Zhang Z, Kroon DE, Casstevens TM, Ramdoss Y, Buckler ES. 2007 TASSEL: Software for association mapping of complex traits in diverse samples. *Bioinformatics* **23**, 2633–2635. (doi:10.1093/bioinformatics/btm308)

34. Danecek P *et al.* 2011 The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158. (doi:10.1093/bioinformatics/btr330)

35. Mastretta-Yanes A, Arrigo N, Alvarez N, Jorgensen TH, Piñeros D, Emerson BC. 2015 Restriction site-associated DNA sequencing, genotyping error estimation and *de novo* assembly optimization for population genetic inference. *Mol. Ecol.* , 28–41. (doi:10.1111/1755-0998.12291)

36. Stamatakis A. 2014 RAxML Version 8 : A Tool for Phylogenetic Analysis and Post-Analysis of Large Phylogenies post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313. (doi:10.1093/bioinformatics/btu033)

37. Leaché AD, Oaks JR. 2017 The Utility of Single Nucleotide Polymorphism (SNP) Data in Phylogenetics. *Annu. Rev. Ecol. Evol. Syst.* **48**, 69–84. (doi:10.1146/annurev-ecolsys-110316-022645)

38. Clement M, Snell Q, Walker P, Posada D, Crandall K. 2002 TCS : Estimating Gene Genealogies. *Proc 16th Int Parallel Distrib Process Symp* , 2:184.

- 39. Leaché A, Fujita M, Minin V, Bouckaert R. 2014 Species Delimitation using Genome-Wide SNP Data. *Syst. Biol.* **63**, 534–542. (doi:10.1101/001172)

40. Bryant D, Bouckaert R, Felsenstein J, Rosenberg NA, RoyChoudhury A. 2012 Inferring Species Trees Directly from Biallelic Genetic Markers : Bypassing Gene Trees in a Full Coalescent Analysis. *Mol. Biol. Evol.* **29**, 1917–1932. (doi:10.1093/molbev/mss086)

41. Bouckaert R, Heled J, Kühnert D, Vaughan T, Wu CH, Xie D, Suchard MA, Rambaut A, Drummond AJ. 2014 BEAST 2: A Software Platform for Bayesian Evolutionary Analysis. *PLoS Comput. Biol.* **10**, 1–6. (doi:10.1371/journal.pcbi.1003537)

42. RStudio Team. 2015 RStudio: Integrated Development Environment for R.

43. Wickham H. 2009 *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. See http://ggplot2.org.

44. Raj A, Stephens M, Pritchard JK. 2014 FastSTRUCTURE: Variational inference of

population structure in large SNP data sets. *Genetics* **197**, 573–589. (doi:10.1534/genetics.114.164350)

45. Goudet J. 2005 HIERFSTAT , a package for R to compute and test hierarchical F -statistics. *Mol. Ecol. Notes* **2**, 184–186. (doi:10.1111/j.1471-8278)

46. Arnold B, Corbett-Detig RB, Hartl D, Bomblies K. 2013 RADseq underestimates diversity and introduces genealogical biases due to nonrandom haplotype sampling. *Mol. Ecol.* 22, 3179–3190. (doi:10.1111/mec.12276)

47. Warnes GR, Bolker B, Lumley T. In press. gplots: Various R programming tools for plotting data. R package version 2.6.0.

48. Aguillon SM, Fitzpatrick JW, Bowman R, Schoech SJ, Clark AG, Coop G, Chen N. 2017 Deconstructing isolation-by-distance: The genomic consequences of limited dispersal. *PLoS Genet.* **13**, 1–27. (doi:10.1371/journal.pgen.1006911)

49. Diniz-filho JAF, Soares TN, Lima JS, Dobrovolski R, Landeiro VL, Pires M, Telles DC, Rangel TF, Bini LM. 2013 Mantel test in population genetics. *Genet. Mol. Biol.* **485**, 475–485.

50. Rousset F. 1997 Genetic Differentiation and Estimation of Gene Flow from F-Statistics Under Isolation by Distance. *Genetics* **145**, 1219–1228.

51. Pante E, Simon-Bouhet B. 2013 marmap: A Package for Importing, Plotting and Analyzing Bathymetric and Topographic Data in R. *PLoS One* **8**, 6–9. (doi:10.1371/journal.pone.0073051)

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

53. Buerkle CA. 2005 Maximum-likelihood estimation of a hybrid index based on molecular markers. *Mol. Ecol. Notes* **5**, 684–687. (doi:10.1111/j.1471-8286.2005.01011.x)

54. Bailey RI. 2018 gghybrid: Evolutionary Analysis of Hybrids and Hybrid Zones.

55. Fitzpatrick BM. 2013 Alternative forms for genomic clines. *Ecol. Evol.* **3**, 1951–1966. (doi:10.1002/ece3.609)

56. Janes JK, Malenfant M, Andrew RL, Miller JM, Dupuis JR, Gorrell JC, Cullingham CI. 2017 The K = 2 conundrum. *Mol. Ecol.* **26**, 3594–3602. (doi:10.1111/mec.14187)

57. Leaché AD, Banbury BL, Felsenstein J, De Oca ANM, Stamatakis A. 2015 Short tree, long tree, right tree, wrong tree: New acquisition bias corrections for inferring SNP phylogenies. *Syst. Biol.* **64**, 1032–1047. (doi:10.1093/sysbio/syv053)

58. Wells RJD, Cowan JH, Fry B. 2008 Feeding ecology of red snapper *Lutjanus campechanus* in the northern Gulf of Mexico. *Mar. Ecol. Prog. Ser.* **361**, 213–225. (doi:10.3354/meps07425)

59. Silva R, Sampaio I, Schneider H, Gomes G. 2016 Lack of Spatial Subdivision for the Snapper *Lutjanus purpureus* (Lutjanidae–Perciformes) from Southwest Atlantic Based on

Multi-Locus Analyses. *PLoS One* **11**, e0161617. (doi:10.1371/journal.pone.0161617)

60. Marval-Rodríguez A, Renán X, Montero-Muñoz J, Galindo-Cortés G, Jiménez-Badillo M de L, Brulé T. 2018 Inter and intraspecific differences of *Lutjanus campechanus* and *Lutjanus purpureus* in otolith shape. Presented at 71th annual meeting of the Gulf and Caribbean Fisheries Institute Conference, The Royal Decameron Isleño Hotel, San Andres, Colombia, 5–9 November.

61. Bernal MA, Gaither MR, Simison WB, Rocha LA. 2017 Introgression and selection shaped the evolutionary history of sympatric sister-species of coral reef fishes (genus: *Haemulon*). *Mol. Ecol.* **26**, 639–652. (doi:10.1111/mec.13937)

62. Cowen RK. 2006 Scaling of Connectivity in Marine Populations. *Science* **311**, 522–527. (doi:10.1126/science.1122039)

63. Cowen RK, Sponaugle S. 2009 Larval Dispersal and Marine Population Connectivity. *Ann. Rev. Mar. Sci.* **1**, 443–466. (doi:10.1146/annurev.marine.010908.163757)

64. Betancur-r R, Hines A, P AA, Ortı G, Wilbur AE, Freshwater DW. 2011 Reconstructing the lionfish invasion: insights into Greater Caribbean biogeography. *J. Biogeogr.* **38**, 1281–1293. (doi:10.1111/j.1365-2699.2011.02496.x)

65. Spalding MD *et al.* 2007 Marine Ecoregions of the World : A Bioregionalization of Coastal and Shelf Areas. **57**, 573–583.

66. Alva-campbell Y, Floeter SR, Robertson DR, Bellwood DR, Bernardi G. 2010 Molecular Phylogenetics and Evolution Molecular phylogenetics and evolution of *Holacanthus* angelfishes (Pomacanthidae). *Mol. Phylogenet. Evol.* **56**, 456–461. (doi:10.1016/j.ympev.2010.02.014)

67. Jackson ND, Carstens BC, Morales AE, O'Meara BC. 2017 Species Delimitation with Gene Flow. *Syst. Biol.* **66**, 799–812. (doi:10.1093/sysbio/syw117)

68. Gibbons A. 2011 A New View Of the Birth of *Homo sapiens*. *Science* **331**, 392–394. (doi:10.1126/science.331.6016.392)

69. Darwin C. 1859 *The Origin of Species*. London, UK: John Murray.

70. NOAA. 2016 Annual commercial landings by group. See https://www.st.nmfs.noaa.gov/commercial-fisheries/commercial-landings/annual-landings.html.

71. Marko PB, Lee SC, Rice AM, Gramling JM, Fitzhenry TM, McAlister JS, Harper GR, Moran AL. 2004 Mislabelling of a depleted reef fish. *Nature* **430**, 309–310. (doi:10.1038/nature02689)

72. Garber AF, Tringali MD, Stuck KC. 2004 Population Structure and Variation in Red Snapper (*Lutjanus campechanus*) from the Gulf of Mexico and Atlantic Coast of Florida as Determined from Mitochondrial DNA Control Region Sequence. *Mar. Biotechnol.* **6**, 175–185. (doi:10.1007/s10126-003-0023-7)

73. US Food and Drug Administration. 1980 Compliance Policy Guide 540.475 Snapper - Labeling. See https://www.fda.gov/ICECI/ComplianceManuals/CompliancePolicyGuidanceManual/ucm074504.htm.

Chapter 2

Genome-wide species delimitation analyses of a silverside fish species complex in central Mexico indicate taxonomic over-splitting

Published in *BMC Ecology and Evolution* (https://doi.org/10.1186/s12862-022-02063-0)

Victor Julio Piñeros*, Carmen del R. Pedraza-Marrón*, Isaí Betancourt-Resendes, Nancy Calderón-Cortés, Ricardo Betancur-R and Omar Domínguez-Domínguez

(*) *equal contribution*

2.1 Abstract

Delimitation of species in the speciation continuum is a complex task, as the process of species origination is not generally an instantaneous event. The use of genome-wide data provides unprecedented resolution to address convoluted species delimitation cases, commonly unraveling cryptic diversity. However, because genome-wide approaches based on the multispecies coalescent model are known to confound population structure with species boundaries, often resulting in taxonomic over-splitting, it has become increasingly evident that species delimitation research must consider multiple lines of evidence. In this study, we used phylogenomic, population genomic, and coalescent-based species delimitation approaches, in the light of morphologic and ecologic information, to investigate species numbers and boundaries comprising the "*humboltianum* group" (Atherinidae)—a taxonomically controversial species complex where previous morphological and mitochondrial studies produced conflicting species delineation scenarios. We generated ddRADseq data for 77 individuals representing the nine nominal species in the group, spanning their distribution range in the central Mexico plateau. Our results conflict with the morphospecies and ecological delimitation scenarios, identifying four independently evolving lineages organized in three geographically cohesive clades: (i) *chapalae* and *sphyraena* groups in Lake Chapala, (ii) *estor* group in Lakes Pátzcuaro and Zirahuén, and (iii) *humboltianum sensu stricto* group in Lake Zacapu and Lerma river system. All in all, our study provides an

atypical example where genome-wide analyses delineate fewer species than previously recognized on the basis of morphology. It also highlights the influence of the geologic history of the Chapala-Lerma hydrological system in driving allopatric speciation in the *humboltianum* group.

2.2 Introduction

Species delimitation, the task of delineating species boundaries, is a core aim in biology, fundamental not only for understanding the extent of the organismal diversity but also for conservation and management planning of threatened or overexploited species. Delimiting species is difficult as speciation is usually not an instantaneous event, where a diffuse zone between populations and species exists—the speciation continuum [1]. Species delimitation inference then becomes particularly challenging in recently diverged or closely related species where differentiating population-level structure from distinct species is challenging [2].

The use of large-scale genotyping techniques provides an opportunity to generate robust datasets to address species delimitation conundrums with unprecedented resolution, allowing the detection of fine-scale genetic structure [3, 4]. Coupled with species delimitation approaches under the multispecies coalescent (MSC) model, genome-wide data provide considerable statistical power to identify recently differentiated boundaries [4]. However, the MSC model has been recently questioned as it can confound population structure with putative species, often leading to overestimating species diversity [4–6]. Thus, the implementation of MSC analyses should be combined with other approaches, such as phylogenomic and population genomic methods (*e.g.,* multivariate, assignment, and genetic differentiation analyses), in conjunction with other lines of evidence (*e.g.,* biogeographic, ecological, or life-history), confers a robust framework to study species delimitation [7].

New World silverside fishes in the *Chirostoma humboldtianum* species complex (Atherinopsidae; hereafter referred to as the *humboldtianum* group), distributed in central Mexico, are of high economic and cultural importance since pre-Hispanic times [8]. Currently, *Chirostoma* species are considered one of the most important fishery resources in the region, where they have been severely overexploited [8, 9].

The *humboldtianum* complex represents an interesting system to address problematic species boundary delimitation as their nominal species have been subjected to taxonomic controversies, with varying numbers of species recognized based on morphologic [10] or molecular [11] data. This group is geographically restricted to the lacustrine environments of the central Mexico plateau occurring mainly in Lakes Chapala, Pátzcuaro, Zirahuén, and Zacapu [10] (Fig. 2.1)–an area comprising roughly 1716.5 km^2 [12, 13]–making it feasible to address species delimitation by examining the extent of species diversity based on samples collected across their distribution range. Nine nominal species have been described based on morphological, osteological, meristic, and allozyme data: *C. chapalae*, *C. lucius*, *C. promelas*, *C. sphyraena*, and *C. consocium* from Lake Chapala; *C. grandocule*, and *C. patzcuaro* from Lake Pátzcuaro; *C. humboldtianum sensu stricto* from Lake Zacapu and the Lerma River System; and two subspecies for *C. estor*, *C. e. estor*, and *C. e. copandaro* from Lakes Pátzcuaro and Zirahuén respectively [10,

14–17] (Fig. 2.1). While a molecular study using one mitochondrial marker (NADH dehydrogenase subunit 2; *ND2*) failed to identify support for the monophyly of each of the nine nominal species, and also to resolve the phylogenetic relationships within the species complex [18], a recent phylogeographic study based on two mitochondrial (cytochrome b; *Cytb* and a fragment of the hypervariable control region; *D-loop*) and one nuclear (first intron of the S7 ribosomal protein; *S7*) markers identified five haplogroups, indicating that the diversity in this species complex is overestimated [11]. The genetic structure recovered by such haplogroups is largely clustered by lakes, suggesting that diversification in this group is the result of historical geological and hydrographic processes.

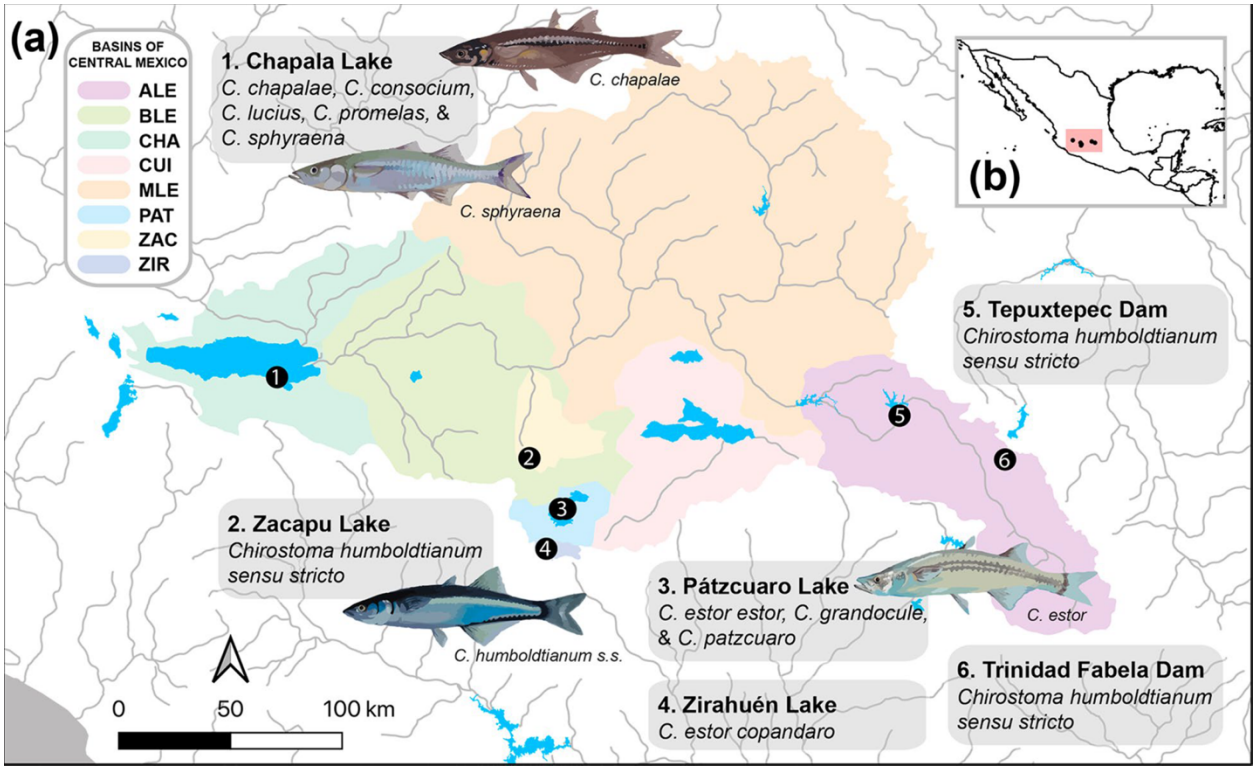


Figure 2.1 (a) Sampling localities of the *humboldtianum* group across the central Mexico plateau: Lake Chapala, Lake Zacapu, Lake Zirahuén, Lake Pátzcuaro, Tepuxtepec Dam, Trinidad Fabela Dam Basins CHA, Chapala; ZIR, Zirahuén; PAT, Pátzcuaro; ZAC, Zacapu; ALE, Alto Lerma; BLE, Bajo Lerma; MLE, Medio Lerma [126]. (b) Location of central Mexico plateau in Mexico.

Ecological divergence has also been suggested as a speciation driver in this group given the high morphological disparity across traits related to different habitat regimes [18, 19]. For instance, ecotypes ('peces blancos' and 'charales'; 117–300 mm and 70–142 mm standard length SL respectively) based on the size of the species could coexist by feeding on preys of different sizes [10]. Betancourt-Resendes et al. [11] observed intra-lacustrine patterns of evolution in Lakes Chapala and Pátzcuaro (two well-differentiated genetic groups within the Lake Chapala over five morphospecies, and two genetic groups in Lake Pátzcuaro over three morphospecies), suggesting

that these lineages probably evolved through ecological speciation. Although no strong evidence of segregation of the nominal species by trophic specialization has been reported, these patterns remain to be evaluated [14, 20, 21]. The discrepancy in the recognition of nine morphological *versus* five mitochondrial species, together with the economic importance and critical conservation status of these species, highlight the necessity of an accurate estimation of species boundaries in this group.

Here, we used double-digest restriction site-associated DNA (ddRAD) sequencing to generate genome-wide single nucleotide polymorphism (SNP) data from all nine nominal species in the *humboldtianum* group, sampled throughout their distribution range. We also considered previous studies to interpret the observed genetic structure in the light of multiple lines of evidence. By generating the most comprehensive molecular dataset for this group, coupled with morphological and ecological lines of evidence, this study aims to: (i) test whether the morphospecies or ecotypes are concordant with the genomic groupings, (ii) investigate the number and boundaries of species in the *humboldtianum* group, (iii) examine the evolutionary processes driving the divergence in this species complex, and (iv) discuss conservation implications in light of the proposed species boundaries and their observed genetic structure. Ultimately, our study adds to the growing body of work addressing complex species delimitation scenarios with genomic data, while providing critical information to guide conservation and management efforts in the *humboldtianum* group.

2.3 Materials and Methods

Sample collections

Our research is a follow-up study of a recently published phylogeographic analysis of the *humboldtianum* group based on two mitochondrial genes, cytochrome b (*Cytb*); and a fragment of the hypervariable control region (*D-loop*); and one nuclear locus, the first intron of the S7 ribosomal protein gene (*S7*) [11]. We carefully selected 77 individuals representing the nine nominal species of the *humboldtianum* group and the genetic diversity observed by Betancourt-Resendes et al. [11] to build a genomic dataset. We added two individuals of *Chirostoma jordani*, and one of *Chirostoma attenuatum* as outgroups. We collected the samples in 2014–2018 with the help of local fishermen, followed the ethical capture methods and regulations approved by the Official Mexican Norm NOM-032-SAG/PESC-2015 and NOM-036-SAG/PESC-2015 for fishing in the lakes of central Mexico. Voucher specimens were preserved in 70% ethanol. Our sampling spans six localities, which together cover the range of distribution of the *humboldtianum* group in the central Mexico plateau (Table S1, Fig. 2.1). We sampled fin clips, preserved them in 95% ethanol, and stored them at -76°C. We deposited tissue and voucher at the fish collection of the Universidad Michoacana de San Nicolás de Hidalgo (UMSNH), Mexico.

Molecular protocols, *de novo* assembly, and SNP genotyping

We extracted high-molecular-weight DNA using the Qiagen DNeasy Blood and Tissue kit (Qiagen, Inc.) following the manufacturer's protocol. We prepared the ddRAD sequencing

libraries at the Sequencing and Genotyping Facility (SGF) at the University of Puerto Rico-Río Piedras (UPR-RP), following the protocol of Peterson et al. [91]. We used the *PstI* and *MseI* restriction enzymes with a size selection window of 300–600 bp. We sequenced ddRADseq libraries in two lanes of Illumina HiSeq 4000 PE 100 bp at the Knapp Center of Biomedical Discovery (KCBD) Genomics Facility, University of Chicago.

We verified the quality of the raw reads using the software FastQC v0.11.5 (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). We demultiplexed the sequenced libraries and removed the restriction sites using the process_radtags.pl script implemented in Stacks v2.4 [92, 93]. We trimmed all demultiplexed reads to 86 bp after removal of the restriction site overhangs. We then applied a second quality filter using a Phred score of 33, producing a total of $3.5x10^8$ reads that we retained. We deposited all demultiplexed raw reads in GenBank (NCBI) (accession no. SAMN26725252–SAMN26725331, SRA BioProject PRJNA816865).

As there is no *Chirostoma* genome that we could use as a reference, we conducted a *de novo* assembly of putative loci using Stacks. To select the assembly parameters that best-fit the data, we first performed an exploratory analysis using default parameters. We then used a subset of 15 samples, formed by the individuals of each nominal species with the highest coverage to conduct a second assembly by applying different combinations of parameters as reported by Mastretta-Yanes et al. [94], Paris et al. [95], and Pedraza-Marrón et al. [96]. Details of the protocol for the *de novo* assembly are given in the Appendix B (SM, Fig. S11-S13). We conducted the final locus assembly using a minimum of five raw reads required to form a stack (m = 5), with a maximum of four mismatches between stacks (M = 4), and five mismatches between loci of different individuals (n = 5).

SNP filtering and database selection

SNP filtering parameters play an important role in the number of recovered loci and the inferred degree of genetic differentiation [97, 98]. We thus implemented a series of exhaustive steps to select the final SNP loci that were included in further analyses (Fig. S14). (*Step 1*) We filtered biallelic loci according to the number of individuals, populations, and nominal species (Table S2). We selected four databases that ranged between ~1000 and ~105000 SNPs to apply further filters. (*Step 2*) We removed low-frequency alleles and potential paralogous loci using a minor allele frequency (*maf*) of 0.01 and 0.05. (*Step 3*) We selected sites with different tolerance thresholds for missing data (0.05, 0.25, 0.50, and 0.75). (*Step 4*) We removed individuals with various cutoffs of missing data (0.05–0.99). Collectively, these filters produced 24 additional datasets, with 150–33800 SNP loci, 0.3–49% of missing data, and 37–80 individuals representing 4–9 morphospecies. (*Step 5*, Table S3). We then selected 19 (out of the 24) matrices to assess the robustness of our analyses to differences in numbers of individuals (37–72), nominal species (4–9), SNPs (~2k–~33k), and proportions of missing data (0.3–49%). The matrices generated during this step were further used to evaluate the sensitivity of the analyses to the exclusion of nominal species with high levels of missing data (see SM). For more exhaustive analyses, we also selected five (out of the 24) datasets with a final set of 72 individuals representing the nine morphospecies: A-33716snps (33716 SNPs, 15.7% missing data), B-10517snps (10517 SNPs, 7.7% missing data), C-4821snps (4821 SNPs, 12.3% missing data), D-3564snps (3564 SNPs, 10.4% missing data), and

E-1887snps (1887 SNPs, 12.4% missing data). (*Step 6*) Finally, to conduct F_{ST} outlier analyses (see section 2.7), the five matrices (A–E) from previous step were partitioned into all, neutral, and outlier loci, for a total of 15 databases. For details refer to the SM (see also Fig. S14).

Multivariate analyses

To determine the number of genetic clusters within the *humboldtianum* group, we conducted a principal component analysis (PCA)—designed to identify genetic groups through eigenvector decomposition of allele frequencies [99]—using matrices A–E with all, neutral, and outlier SNP loci. In addition, to identify *de novo* structure, we conducted an assessment of *a priori* designations using a discriminant analysis of principal components (DAPC). DAPC combines discriminant (DA) with principal component (PC) analyses to maximize genetic variance among groups while minimizing within-group variance [100, 101]. We selected the *a priori* groups by configuring the nine nominal *Chirostoma* morphospecies. We assessed the most likely number of clusters (k = 1–9) using the *find.clusters* function within the *adegenet* v2.1.1. package [102] by selecting the k model with the lowest Bayesian Information Criterion (BIC) score. We calculated the number of PCs to be retained using the *dapcCross* validation function, which is based on the highest successful assignment that presented the lowest mean squared error [100].

Genomic structure and genomic differentiation analyses

To evaluate the number of genetic clusters within the *humboldtianum* group, we analyzed the final matrices (A–E) with all, neutral, and outlier loci under a maximum-likelihood framework using the software ADMIXTURE v1.3.0 [103]. This program estimates the proportion of the ancestral population (Q estimates) for each individual to calculate the number of genetic clusters (k) under a cross-validation procedure, where lower values represent the optimal number of populations [104]. All analyses were run with k values ranging from 1 to 12, which represent the nine nominal morphospecies described for the group, the sub-species from Zirahuen Lake, *C. e. copandaro*, and two additional genomic groups that could be possible detected in the hypothetical case of have a higher number of population clusters. To further assess population clustering, we conducted admixture analyses within Chapala and Pátzcuaro-Zirahuén lakes, corresponding to clade I and clade III, respectively (see Results). The analysis for Chapala lake sample (n=41) was run with of k values ranging from 1 to 5, which represent the five morphospecies of *humboltianum* group described in this lake. The analysis for Pátzcuaro-Zirahuén lakes sample (n=24) was run with k values ranging from 1 to 4, which represent the three morphospecies of *humboltianum* group present described in Pátzcuaro lake, plus the subspecies from Zirahuen lake *C. e. copandaro*. We plotted all cross-validation and Q estimates using the *ggplot2* [105] and *pophelper* [106] packages in R studio.

We computed analyses of pairwise F_{ST} comparisons to test the genetic differentiation among groups using the D-3564snps-all_loci matrix. We selected this matrix as it showed the general patterns observed with other matrices, but also as it had the highest percentage of variation explained in the multivariate analyses (see Results). We examined four alternative species delimitation schemes: i) nine morphospecies (morphological-based hypothesis; *sensu* Barbour, [10]), ii) five *mt*DNA haplogroups (*mt*DNA-based hypothesis; *sensu* Betancourt-Resendes et al.

[11]), iii) four genomic clusters (k = 4) observed with the DAPC analyses (see Results), and iv) three clusters detected by the admixture analyses (k = 3) (see Results). To further test whether an agreement between species' body size and trophic speciation exists [10], we calculated pairwise F_{ST} comparisons among ecotypes by lake ('peces blancos' *vs.* 'charales'). Details on ecotype discrimination are provided in the SM. We calculated these analyses in ARLEQUIN v3.5.1.2 [107], and significance was determined using 10000 permutations and a Bonferroni correction [108] to adjust p values for these calculations (significant levels of α are provided in Tables S4–S8).

Phylogenetic analyses

We estimated phylogenetic trees under a maximum likelihood (ML) framework using the software IQ-TREE v2.0.6 [109]. We conducted phylogenetic inference for the five matrices (A–E) including all, neutral, and outlier SNP loci, and using *C. jordani* as the outgroup. All analyses used the GTR model with gamma distribution and the ascertainment bias correction (ASC) to account for acquisition discrepancies related to SNP datasets. We estimated branch support using IQ-TREE's ultrafast bootstrap algorithm (UFBoot) with 500 replicates [110].

To evaluate patterns of incongruent evolutionary histories between mitochondrial and nuclear loci—mitonuclear discordance—we used the sequence data of two mitochondrial genes (cytochrome b or *Cytb*, and a fragment of the hypervariable control region, *D-loop*) from Betancourt-Resendes et al. [11]. Using the GTR model in IQtree as explained above, we estimated ML *mt*DNA trees for each gene separately and also for the concatenated *mt*DNA matrix. We extracted tips shared between the *mt*DNA trees and our ddRADseq (D-3482snps-neutral_loci matrix) using the *keep.tip* function implemented in the R package *phytools* [111].

Multispecies coalescent analyses

We inferred a multi-coalescent species tree for the D-3482snps-neutral_loci matrix using the SNAPP v1.3.0 plug-in [112] implemented in BEAST2 v2.6.3 [113]. SNAPP allows the inference of a species tree from unlinked SNP data while bypassing gene tree inference [112]. For this analysis we used two chains of 30 million steps, sampling every 500 trees, with a burn-in of 10%. We set default priors for coalescent and mutation rates, as well as ancestral population size parameters. We visualized all results in the software Tracer v1.7 [114] to confirm that the analyses had converged, reached stationary, and that all effective sample sizes (ESS) were higher than 200. Lastly, we summarized the conflict of posterior distribution trees into the species tree using DENSITREE v2.01 [115].

To test the previously outlined species delimitation scenarios (i–iv) in a coalescent framework, we applied the Bayes Factor Delimitation (BFD*) approach [116] using SNAPP and BEAST2. To overcome computational burden, we generated three subsets encompassing the nine morphospecies with 39–59 individuals and 411–1102 SNP loci that were generated from the D-3482snps-neutral_loci matrix (see SM). With BFD* candidate species delimitation scenarios are evaluated according to the marginal likelihood estimates (MLE). The different scenarios are then compared using Bayes Factors (BF) [117], estimated by subtracting MLE values for two models and multiplying the difference by two ($BF=2*(MLE_1-MLE_2)$). As only two models are compared

at the time, we evaluated all possible combinations among the species delimitation scenarios (see Table 1 and SM). To set up the priors and MCMC runs, we followed the recommendations provided in the BFD* manual [118]. The speciation rate (λ), which represents the birth-rate on the Yule tree prior, was set to follow a gamma distribution with α = 2 and β = 200.

Genetic diversity, effective population size (*Ne*), and F_{ST} outlier analyses

We estimated the genomic diversity of the *humboldtianum* group considering the four species delimitation scenarios outlined above. We did not test a scheme that included ecotypes as we did not find any support for this hypothesis (see Results). We used GenAlEx v6.5 software [119] to calculate the observed (*Ho*) and expected (*He*) heterozygosity for each cluster using D-3482snps-neutral_loci matrix. Additionally, we evaluated the effective population size (*Ne*) of each species delimitation scenario by estimating the linkage disequilibrium [83], heterozygote-excess [120], and molecular co-ancestry values [121] with the software NeEstimator v2 [122].

To detect loci with high levels of genetic differentiation, we examined the A–E matrices formed by all SNP loci under a Bayesian framework using the program BayeScan [123]. In this analysis, F_{ST} coefficients are decomposed into a population-specific component (β) shared by all loci, and a locus-specific component (α) shared by all populations. When a locus-specific component is necessary to explain the data, selection is assumed to play a role at that locus [123]. To reduce the risk of false positives without reducing the power to detect loci evolving under selection, we set the default parameters and priors for a neutral model according to the total number of SNP loci in each matrix (100 for matrices with > 1000 loci, and 1000 for matrices with > 10000 loci) [124]. Statistical significance for outlier loci was assumed if *q* values £ 0.05. To determine the strength and direction of selection, we estimated the a parameter, where a positive value suggests diversifying selection and a negative value indicates balancing or purifying selection [124]. Lastly, to identify the approximate genomic region within which the outlier loci occur, we searched the consensus stack sequences on the NCBI server (https://blast.ncbi.nlm.nih.gov/Blast.cgi) using the Blastn algorithm. To choose the most similar candidate sequence, we used an e-value threshold lower than 0.05, a maximum target sequence of 5000 to retain all the hits possible, and a sequence similarity above 75% to filter among the retained sequences. We ran all R analyses using R studio v1.2.1335 and R v4.1.1 [125].

2.4 Results

ddRADseq assembly and datasets

We obtained ~$3.6X10^8$ sequence raw reads, of which ~$3.5X10^8$ (97.6%) passed quality filters. On average, there were *ca*. 2 million reads per sample. The *de novo* assembly using a combination of *m*5*M*4*n*5 parameters generated 275533 putative loci with an average coverage of 16.1 per sample. After filtering for missing data and different *maf* thresholds, we generated five final matrices (A–E), which varied from 1887-33716 SNPs and 7.7–15.77% of missing data. These final matrices were also filtered to contain neutral-only and outlier loci, ranging from 1795–33346 and 82–370 SNP loci respectively. For more information, refer to the Methods section and the Appendix B.

Multivariate analyses

The PCAs performed to evaluate the robustness of the genomic results consistently identified four genomic groups regardless of the number of SNPs (1887–33716), individuals (37–72), morphospecies (4–9), or missing data (0.3–52.1%) included in each of the matrices (Fig. S1). The PCAs including either all or neutral-only SNP loci (matrices A–E) also recovered four genomic groups, which align with the geography of the central Mexico plateau, but not with a delineation according to the previously recognized morphospecies. The first group (*humboldtianum sensu stricto* group hereafter) clustered samples from Lake Zacapu, Tepuxtepec, and Trinidad Fabela dams. The second group (*estor* group hereafter) included individuals of *C. e. estor, C. e. copandaro*, *C. grandocule*, and *C. patzcuaro* from Lakes Pátzcuaro and Zirahuén. The third group (*chapalae* group hereafter) is composed of samples of *C. chapalae*, *C. consocium*, *C. lucius,* and *C. promelas*, from Lake Chapala. Finally, the fourth group (*sphyraena* group hereafter) is formed by *C. sphyraena*, also from Lake Chapala. The first principal component accounted for 24.76–33.51% of the genetic variation and is congruent with the separation of *estor* and *humboldtianum sensu stricto* from *chapalae* and *sphyraena* groups. The second principal component represented 6.37–7.47% of the genetic variation, resulting in the clear segregation of the *humboldtianum sensu stricto* group (Fig. 2.2a and S2).

While the results of the PCAs estimated using outlier SNP loci did not produce a clear demarcation relative to matrices considering all and neutral-only SNPs, these analyses mostly resulted in the same clustering patterns except in two instances (matrices D and E) where the *chapalae* and *sphyraena* groups clustered together. The first two principal components of the PCAs calculated with outlier SNPs accounted for 20.32–86.58% of the genomic variation explained. Overall, PCAs generated with just outlier (82–370) SNP loci delineated between three and four genomic clusters, superimposing populations from *chapalae* with *sphyraena* groups (Fig. 2.2a and S2).

Scatter plots of the DAPC analyses produced results similar those based on PCA. In general, the first two discriminant axes identified the same four genomic clusters (BIC: k=4) (Fig. 2.2b, S3 and S4), retaining 9–12 PCs that represent 48.5–57.7% of the cumulative variance. The DAPC analyses estimated using the outlier loci (matrices C, D and E) also identified three genomic clusters (BIC: k=3) where *chapalae* and *sphyraena* form a single group (Fig. 2.2b, S3 and S4).

Genomic structure and genomic differentiation analyses

Admixture assignment results of the SNP loci (matrices A–E) considering all, neutral-only, and outlier loci identified three genomic clusters (k = 3) as the best-fit model (Fig. S5), each corresponding to the major groups (*humboldtianum sensu stricto*, *estor*, and *chapalae-sphyraena*) obtained with PCA and DAPC analyses using the outlier SNP loci (Fig. 2.2c and S6). As in previous analyses, the admixture results did not delineate morphospecies boundaries.

All individuals of the *humboldtianum sensu stricto* group collected from artificial dams (three from Tepuxtepec and one from Trinidad Fabela) had strong admixture with the *chapalae* group (0.11–0.45), whereas *chapalae* and *estor* groups produced rather low values of shared genetic information (<0.10). While there were higher levels of individual admixture between

humboldtianum sensu stricto and *chapalae* groups when fewer SNP loci were analyzed, a higher number of outlier SNP loci (99–370) identified more admixture between the three groups than those considering fewer outlier SNPs (82–92) (Fig. 2.2c and S6).

Intralake admixture analyses within Lake Chapala found no evidence of finer genetic structure, as the best-fit model supported a scenario represented by a single metapopulation (k = 1, Fig. S7), clustering all individuals of the *sphyraena* and *chapalae* groups. Genetic structure was not observed within the Pátzcuaro-Zirahuén lakes either (k=1). In general, the observed genetic structure based on intra-lake admixture analyses disagreed with both morphospecies and ecotypes ("peces blancos" vs. "charales") (Fig. S7).

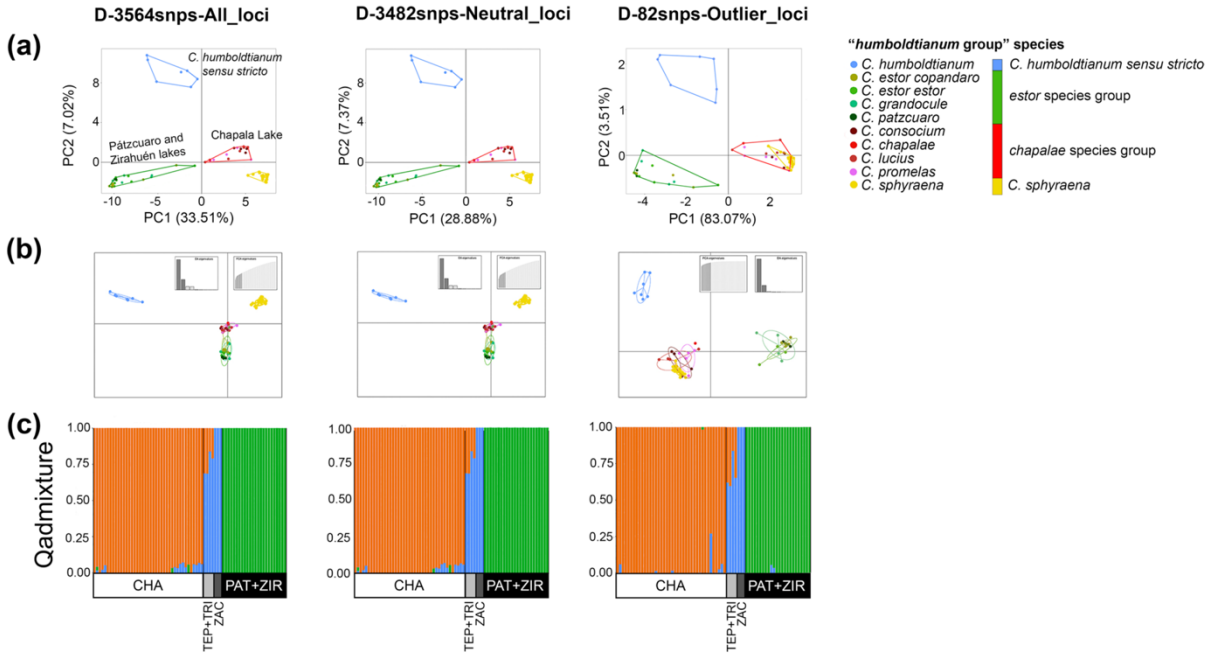


Figure 2.2 Multivariate and admixture results based on all, neutral, and outlier SNP loci from the matrix D (3564-snps loci), which explain the highest percentage of explained variation in the analyzes. (a) Principal component analyses (PCAs), and (b) Discriminant analyses of principal components (DAPCs) consistently recovered four genomic groups (k=4) with all and neutral loci that are in agreement with geographic patterns but not with the previously recognized morphospecies: *humboldtianum sensu stricto* group (blue), from Lake Zacapu; *estor* group (green) from Lakes Patzcuaro and Zirahuén; *chapalae* group (red), from Lake Chapala; and *C. sphyraena* group (yellow), also from Lake Chapala. Outlier loci resolved three groups (k=3) where *chapalae* and *sphyraena* groups clustered together. Morphospecies are color-coded according to the genomic groups observed. In PCAs scatterplots the point clustering groups are delimited by convex hulls. In DAPC scatterplots, the point clustering groups are inside their 95% inertia ellipses, and the lines connect points to the mean value for each group. The eigenvalue bar plots are showing in the upper right of each figure. (c) Admixture assignment analyses estimated using all, neutral, and outlier SNPs consistently identified three well-differentiated clusters (k = 3). Each bar represents the probability of assignment to each cluster. Genomic clusters are color-coded as blue, *humboldtianum sensu stricto*, green, *estor* group; orange, *chapalae-sphyraena* group. CHA, Lake Chapala; TEP, Tepuxtepec Dam; TRI, Trinidad Fabela Dam; PAT, Lake Pátzcuaro; ZIR, Lake Zirahuén.

Pairwise F_{ST} comparisons were much higher and statistically more significant at inter-lacustrine than at intra-lacustrine levels, reflecting a strong correlation between genetic structure and geography (Tables S4–S7). Pairwise F_{ST} values between the nine morphospecies varied from 0.023–0.256 and were mostly significant (p values < 0.0012, α = 0.0014) at inter-lacustrine comparisons, with the exception of *C. chapalae* which showed no significant differentiation with the rest of the morphospecies. By contrast, comparisons among sympatric morphospecies from Lakes Chapala and Pátzcuaro yield negative non-significant values (F_{ST} = -0.317–-0.012, p values > 0.029), except for *C. sphyraena*, which was significantly different than the rest (Table S4). A greater genetic differentiation was found between the five *mt*DNA haplogroups (0.050–0.246), where the majority of the pairwise comparisons were significantly different (p values < 0.0001, α = 0.005), except for comparisons between subspecies *C. e. estor* and *C. e. copandaro* (p value = 0.05; Table S5). F_{ST} values between DAPC clusters were significant and varied from 0.04–0.248 (p values < 0.0001, α = 0.0083; Table S6); the three genomic groups detected by the admixture analyses also resulted in significant F_{ST} values (0.120–0.217, p values < 0.0001, α = 0.017; Table S7). Finally, the genetic structure between 'peces blancos' and 'charales' ecotypes (-0.005–0.001) was not significant in any of the intra-lacustrine comparisons (p values = 0.01–0.027, α = 0.005; Table S8).

Phylogenetic analyses

Phylogenetic inferences in a ML framework based on matrices A, C, D and E of all and neutral SNP loci resolved three highly supported clades (ultrafast bootstrap or UFBoot = 90–99%; Figs. 2.3a and S8), which are largely in agreement with the results from admixture analyses. Matrix B resolve well Clades I and III, but Clade II was paraphyletic (Fig. S8). Clade I clustered species within *chapalae* and *sphyraena* groups from Lake Chapala (UFBoot = 83–95%); clade II is formed by the *humboldtianum sensu stricto* group from Lake Zacapu and dams (UFBoot = 73–100%); and clade III is represented by the *estor* group from Lakes Pátzcuaro and Zirahuén (UFBoot = 100%). Samples of *C. sphyraena* and the subspecies *C. e. copandaro* were resolved as monophyletic within clade I (UFBoot = 80–100%) and clade III (UFBoot = 94–97%), respectively. While the phylogenetic trees obtained with outlier loci datasets showed less resolution than those using the complete matrix, these different analyses resolved the reciprocal monophyly between Clade I and Clades II-III, except in matrix A, where Clade I was paraphyletic. Also, the relationship between Clade II and III was more problematic: in some cases, these clades recovered the reciprocal monophyly between both clades (matrices A and C), but in the other cases, they were paraphyletic (databases B, D, and E) (Fig. S8).

Tree inference based on mitochondrial data failed to resolve the groups identified with our genome-wide RADseq analyses, a case of mitonuclear discordance between both datasets results (Figs. 2.3c and S9). However, similar to the ddRADseq phylogenies, analyses of the mitochondrial genealogies and a phylogeny estimated using the concatenated matrix did not delineate any clear monophyletic groups according to morphospecies boundaries. Although some groupings comprised of neighboring localities were resolved with the *mt*DNA trees, resulting clades from these trees are not cohesively clustered by geography. The *Cytb* genealogy resolved the monophyly of *C. sphyraena*, a result that is in agreement with the RADseq phylogeny and previous studies

[11]. However, these relationships were not resolved with confidence (UFBoot < 95%). Overall, none of the phylogenetic hypotheses were concordant with the morphospecies or ecotypes proposed within the *humboldtianum* group (Figs. 2.3 and S9).

Multispecies coalescent analyses

The species tree estimated with SNAPP also evidenced the three geographically cohesive clades (I, II, and III; Fig. 2.3b). Coalescent-based species delimitation analyses using the BFD* method selected species delimitation scenarios of four and five species as the top-ranked models with the highest MLE values (MLE = -30568.17 and -21025.86–-11910.25 respectively; Table 1). The main difference of the models of four and five species relies on the subspecies *C. e. copandaro* from Lake Zirahuén forming a different group from *C. e. estor* and the sympatric species in Lake Pátzcuaro. Overall, the fact that the Bayes factors analyses supported models of four (subset 1 and 3) and five species (subset 2) is decisive compared with nine species that reflects the current taxonomy (BF = 295.56, 507.42, and 850.28 respectively; Table 1) and when compared to the alternative models of three species (BF = 158.53, 504.88 and 660.41 respectively; Table 1).

Genetic diversity and effective population size (*Ne*) analyses

The genomic diversity of the *humboldtianum* group is summarized in Table 2. The observed and expected heterozygosity (*Ho* and *He*) across different lineages was 0.097–0.136 and 0.084–0.139, respectively. All lineages distributed within Lake Chapala revealed higher genetic diversity (*Ho* = 0.123–0.138) than those in the Lakes Zacapu and Pátzcuaro-Zirahuén (*Ho* = 0.100–0.126) (Table 2). The values of *Ne* estimator (*HetExcessNe* and *CoancentryNe*) across different lineages was 7.2–317 and 2.2–31 respectively (Table 2). The higher *Ne* values varied of location among the species delimitation from Chapala lake lineages to Patzcuaro-Zirahuen lake lineages (Table 2). Values of *LDNe* estimator were infinite with exception of *chapalae-sphyraena* group (86.9), while *C. humboldtianum sensu stricto* obtained infinite values in *LDNe* and *HetExcessNe* estimators (Table 2).

F_{ST} loci outlier analysis

A total of 410 outlier loci (1.1% of the total analyzed, Fig. S10) were compared to GenBank entries using BLAST-n. In summary, 116 sequences did not have a match, while the remaining 294 (71.7%) loci matched fish sequences (E-value = 4E-42–2E-02), including protein-coding genes (26%) (Table S9). Out of these coding regions, we identified genes related to a wide array of biological functions such as genes implicated in immune responses (*kpna3*: *Nothobranchius furzeri*, GenBank accession number XM_015966110.1; *bcl11b*: *Cheilinus* undulates, XM_041804421.1; *trim59*: *Megalops cyprinoides*, XM_036537431.1; *NLR family CARD domain-containing protein 3-like*: *Acanthophagus latus*, XM_037091148.1), sensory systems (*or132-1*: *Danio rerio*, DQ306116.1*; sws2a* and *rh2-1*: *Lucania goodei*, MT850055.1; and *lws2*: *Monopterus albus*, XM_020592984.1), growth (*cand1*: *Poecilia formosa*, XM_007569621.2; and *sgta*: *Archocentrus centrarchus*, XM_030727116.1), and skeletal-muscle system (*neb*: *Gymnodraco acuticeps*, XM_034221710.1; and *ttn*: *Fundulus heteroclitusi*, XM_036135404.1) (Table S9).

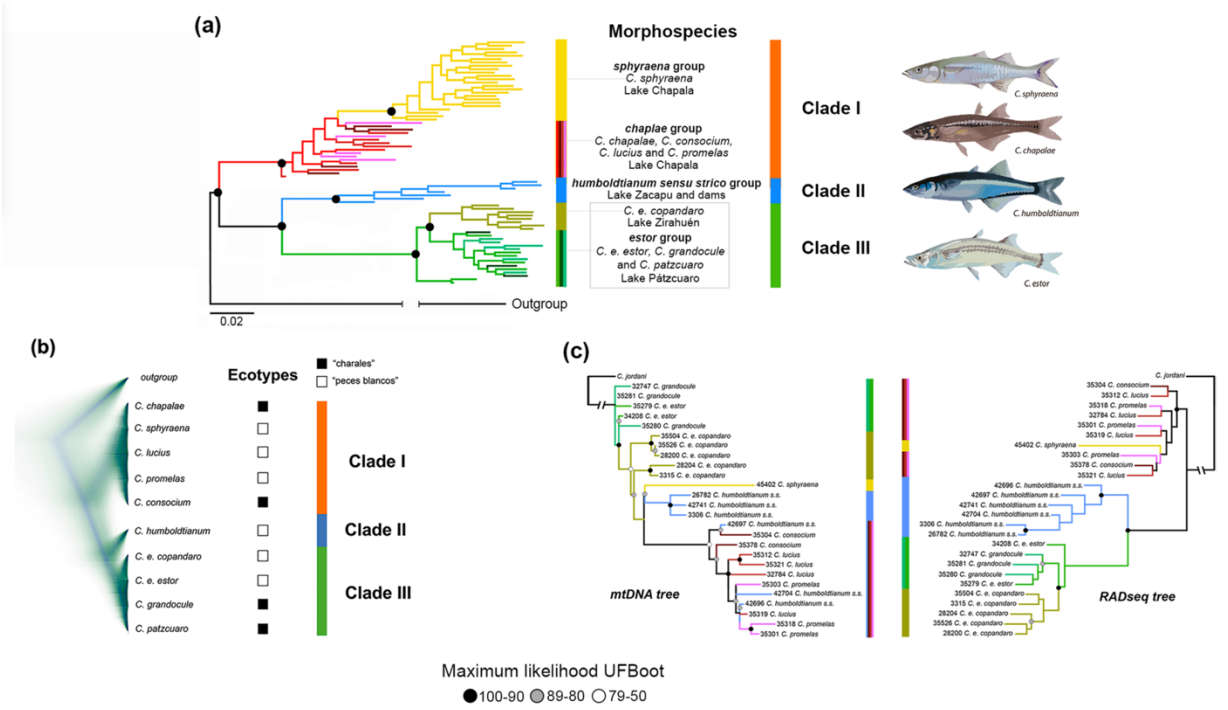


Fig. 2.3 The ML phylogenetic tree based on 3564 SNP loci (a) recovered three well-differentiated clades: clade I formed by species in Lake Chapala, clade II represented by species from Lake Zacapu and dams, and clade III composed by species in Lakes Pátzcuaro and Zirahuén. We observed *C. sphyraena* as a monophyletic group within clade I, and the subspecies *C. e. copandaro* from Lake Zirahuén as a differentiated cluster within clade III. Although the multi-coalescent species tree considering circa 3400 neutral SNPs (b) recovered the main genomic clusters as the rest of the genetic structure analyses and the ML phylogenetic inferences, it did not present any intra-lake divergences. (c) The phylogenetic inference based on *mt*DNA (*Cytb* and *D-loop*) failed to delineate reciprocal monophyletic groups as the ones estimated using the ddRADseq data (3564 SNP loci). The analyses of the genome-wide datasets clearly recovered well-differentiated clusters that are in agreement with the geography of the central Mexico plateau. However, none of our phylogenetic inferences showed concordance with the morphospecies nor ecotypes recognized within the *humboltianum* group. Numbers on branches of the main clades indicate bootstrap values.

Table 2.1 Results of bayes factor delimitation (BFD*) analyses for the *humboldtianum* group using three SNPs subsets (ranging from 39–59 individuals, and 411–1102 SNP loci). Bayes factor (BF) calculations are estimated against the model with the best marginal-likelihood estimate (Model 1). Positive BF values indicate support for model 1 (the best model). Overall, these analyses support a species delimitation scenario between four and five species, rejecting the nine morphospecies scheme.

Model	**Species number**	**Subset 1 39 ind./411 SNPs**		**Subset 2 39 ind./1102 SNPs**		**Subset 3 59 ind./458 SNPs**	
		MLE	**BF**	**MLE**	**BF**	**MLE**	**BF**
current morpho-species	9	-12058.03	295.56	-30993.31	850.28	-21279.57	507.42
*mt*DNA haplogroups	5	**-11910.25**	NA	-30600.16	63.98	**-21025.86**	NA
DAPC groups	4	-11927.68	34.85	**-30568.17**	NA	-21161.43	271.14
Admixture groups	3	-11989.52	158.53	-30898.37	660.41	-21278.3	504.88

MLE: marginal likelihood estimates. Bold values correspond to the best model selected by the analysis. NA correspond to not applied, and its associated marginal likelihood is in bold. 0 < BF < 2, non-significant; 2 <BF < 6, positive evidence; 6 < BF < 10, strong support; BF > 10, decisive support.

Table 2.2 Genomic diversity of 3482 neutral SNPs loci estimates for the *humboldtianum* group, under each species delimitation hypothesis examined in this study (i–iv).

	N	*HO* (SD)	*HE* (SD)	*F* (SD)	*LDNe* 95% CIs	*HetExcessNe* 95% CIs	*CoancestryNe* 95% CIs
Morphospecies							
C. chapalae	1	-	-	-	-	-	-
C. consocium	4	0.14 4*e*-4	0.11 3*e*-3	-0.23 8*e*-3	infinite	8.9	10.9
C. lucius	7	0.13 4*e*-4	0.11 3*e*-3	-0.14 7*e*-3	infinite	12.6	6
C. promelas	5	0.12 4*e*-4	0.10 3*e*-3	-0.22 9*e*-3	infinite	infinite	31
C. sphyraena	24	0.14 4*e*-4	0.12 3*e*-3	-0.08 5*e*-3	infinite	8.9	2.2
C. humboldtianum	7	0.15 4*e*-4	0.12 3*e*-3	-0.06 7*e*-3	infinite	infinite	3
C. estor	15	0.10 4*e*-4	0.10 3*e*-3	-0.04 7*e*-3	infinite	244	3.6
C. grandocule	6	0.10 4*e*-4	0.08 3*e*-3	-0.17 8*e*-3	infinite	10.6	7.2
C. patzcuaro	3	0.11 4*e*-4	0.09 3*e*-3	-0.30 9*e*-3	infinite	7.2	infinite
5 *mt*DNA haplogroups							
C. chapalae	17	0.13 3*e*-3	0.14 3*e*-3	0.08 7*e*-3	infinite	24.7	4.5
C. sphyraena	24	0.14 4*e*-3	0.12 3*e*-3	-0.08 5*e*-3	infinite	8.9	2.2
C. humboldtianum	7	0.13 3*e*-3	0.12 3*e*-3	-0.06 7*e*-3	infinite	infinite	3
C. e. estor	16	0.10 4*e*-3	0.10 3*e*-3	-0.04 7*e*-3	infinite	10.8	4.4
C. e. copandaro	8	0.10 4*e*-3	0.11 3*e*-3	0.08 1*e*-2	infinite	8.9	5.4
4 multivariate DAPC clusters							
C. chapalae group[a]	17	0.13 3*e*-3	0.14 2*e*-3	0.08 7*e*-3	infinite	24.7	4.5
C. sphyraena	24	0.14 4*e*-3	0.12 3*e*-3	-0.08 5*e*-3	infinite	8.9	2.2
C. humboldtianum sensu stricto	7	0.13 3*e*-3	0.12 3*e*-3	-0.06 7*e*-3	infinite	infinite	3
C. estor group[b]	24	0.10 3*e*-3	0.10 3*e*-3	0.10 8*e*-3	infinite	106.4	4.4
Admixture clusters (K=3)							
C. chapalae-sphyraena groups	41	0.14 4*e*-3	0.14 8*e*-3	0.09 6 *e*-3	86.9	317	4.2
C. humboldtianum sensu stricto group	7	0.13 3*e*-3	0.12 3*e*-3	-0.06 7*e*-3	infinite	infinite	3
C. estor group	24	0.10 3*e*-3	0.10 3*e*-3	0.10 8*e*-3	infinite	106.4	4.4

N number of individuals analyzed per group, observed (H_O) and expected (H_E) heterozygosity, fixation index coefficient (*F*), linkage disequilibrium effective (*LDNe*), heterozygote-excess (*HetExcessNe*), and molecular coancestry (*CoancestryNe*) population size estimators. SD= standard deviation; CI= confidence intervals. The grey shading in the cells indicates the highest values of genetic diversity. [a]*C. chapalae* group conformed by *C. chapalae, C. consocium, C. lucius,* and *C. promelas*. [b]*C. estor* group represented by *C. e.* estor, *C. e. copandaro, C. grandocule*, and *C. patzcuaro*.

2.5 Discussion

In this study, we used ~1,800 to ~33,000 SNPs to test the morphological- (nine nominal species) and *mt*DNA-based (five haplogroups) hypotheses to elucidate the number and boundaries of species in the *humboldtianum* group. Our results consistently identified four independently evolving lineages organized in three well-differentiated clades: *chapalae-sphyraena* in Lake Chapala, *humboldtianum sensu stricto* in Lake Zacapu and dams, and *estor* in Lakes Pátzcuaro and Zirahuén. The genomic clusters obtained were not in agreement with the morphospecies scenario but rather with biogeographic concordance reflecting ancient isolation events in central Mexico. This scenario suggests that the geologic history of the Lerma-Chapala hydrologic system has played a major role in driving divergence in this species complex. Our analyses also revealed an intra-lake cladogenetic event where *C. sphyraena* ('pez blanco') is distinguished from its sympatric counterparts ('peces blancos' and 'charales') in Lake Chapala. Although body size in *Chirostoma* species in Lake Chapala has been suggested as a promoter of ecological niche partitioning for this species complex [10, 21], we did not find evidence of genetic structure related to the 'peces blancos' and 'charales' ecotypes.

Herein, the use of genome-wide data provided an unprecedented resolution that had not been achieved using scant genes to test species delimitation scenarios within the *humboldtianum* group [11, 18]. While species delimitation studies examining thousands of genetic loci often unveil cryptic diversity [22–25], our study represents one of the few cases where the use of genome-wide SNP data and MSC approaches provide evidence of taxonomic over-splitting [2, 5, 26]. Recently, it has been recently recognized that the MSC model can potentially confound population structure with species boundaries—particularly when major sampling gaps near the range of distribution exist—, leading to an over-estimation of the number of species due to the high statistical power of genome-wide data [4–6]. However, in this study by evaluating our analyses in a framework that examines all nominal species in the *humboldtianum* group from across their distribution ranges in the Lerma-Chapala hydrologic system, and under the light of morphological and ecological lines of evidence, we overcome these caveats and provide a robust analysis to assess the number of species within the *humboldtianum* group and also to examine the group's evolutionary history.

Inter-lacustrine divergences

Our results suggest that vicariance events during the Pleistocene influenced the early divergence within the *humboltianum* group. Speciation within this group appear to be strongly related to the complex geologic history of volcanism, tectonism, and climatic events that promoted the connection and disconnection of the Lerma-Chapala hydrological system, and surrounding tributaries—including several lakes and paleolakes such as Chapala, Cuitzeo, Zacapu, Pátzcuaro, and Zirahuén [10].

Although our phylogenetic results are somewhat incongruent with those previously estimated by Betancourt-Resendes et al. [11] among the genetic groups (*chapalae*, *sphyraena*, *estor*, and *humboldtianum sensu stricto* groups), their divergence time analyses provide a rough estimate of the timing of divergence in the group, placing the origin of independent evolutionary lineages within the *humboldtianum* group during the Pleistocene <1 Ma (0.58–0.13 Ma). There are

two main biogeographic processes that are synchronous and congruent with the observed genetic patterns recovered in this study, for which we suggest that these events played an important role as the main drivers of diversification in the *humboldtianum* group. First, the allopatric fragmentation of clade I (Lake Chapala) from II-III (Lake Zacapu-Lakes Pátzcuaro and Zirahuén) may correspond to a biogeographical barrier promoted by the geologic activity of the Penjamillo Graben and the formation of the Tarascan corridor, which started much earlier during the Late Miocene–Early Pliocene [27, 28]. These geological events separated the ancient corridors between the paleo Lerma-Chapala system and the Cuitzeo paleolake plus adjacent tributaries that connected Lakes Zacapu, Pátzcuaro, and Zirahuén [10, 29]. The separation of these hydrologic regions experienced its highest peak during the Early Pleistocene [29], where severe climatic fluctuations promoted a series of connection and disconnection events of small water bodies, with the last dry episode starting *ca.* 0.12 Ma [10, 29]. The second biogeographic event is the fragmentation of the ancestral Villa Morelos and Chucandiro-Huaniqueo corridors, also during the Pleistocene. This episode, promoted by the geologic activity of the Northeast-Southwest fault system *ca.* 0.7 to 0.5 Ma, separated Lake Zacapu from Lakes Cuitzeo-Pátzcuaro-Zirahuén [30, 31], correlating with the genetic patterns of divergence between clade II (Lake Zacapu) and III (Lakes Pátzcuaro and Zirahuén).

The same cladogenetic patterns, in agreement with the aforementioned biogeographic events, have been observed in goodeid freshwater species endemic to central Mexico, including the divergence between the sister species *Skiffia multipunctata* and *S. lermae* and the split of the *Allotoca diazi* complex from *A. zacapuensis*, although the diversification events do not appear to have occurred in synchrony [32–35].

Intra-lacustrine divergences

Our genome-wide data revealed the presence of two genetic clusters in Lake Chapala (*sphyraena* and *chapalae* groups), as suggested by Betancourt-Resendes et al. [11] using mitochondrial sequences. The evolutionary processes segregating these populations seem to be related to ecological speciation—divergent specializations promoted by ecological opportunity following reproductive isolation [36, 37]—which represents one of the major drivers of sympatric evolution in lakes [36, 38, 39]. The most iconic examples of ecological speciation are depicted by South American [40–42] and African cichlids [43, 44], and sticklebacks [45, 46], where patterns of morphological divergence are associated with trophic partitioning. For example, preference for soft mobile (*e.g.,* copepods) compared to hard sessile preys (*e.g.,* gastropods) can lead to disruptive selection on skull morphology and body shape [47], where body size can be crucial to succeed in the related foraging mode (*e.g.,* benthic *vs.* limnetic in *Gasterosteus* spp; [48, 49]).

In the case of species of the *humboldtianum* group occurring within the Lake Chapala (*C. sphyraena*, *C. promelas*, *C. consocium*, and *C. lucius*), several hypotheses of ecological speciation suggest the coexistence of ecotypes in agreement with a differentiation of morphological traits (*e.g.*, jaw shape, head length, oral gape, and gill raker structure) related to feeding habitats [10, 50], or an ecological partition in correlation to the body size of the species (larger ‘peces blancos’ *vs* smaller ‘charales’) [10, 16, 21]. However, no strong evidence showing clear patterns of differentiation between morphospecies or ecotypes has been documented to date [21].

Herein, we evaluated the hypothesis from [10] that differentiation in body length would allow the co-occurrence of different species as they feed on distinct prey sizes. Although signals of trophic specialization separating 'peces blancos' and 'charales' have been documented by Mercado-Silva et al. [21] for three species in Lake Chapala, no analysis conducted here (PCAs, DAPCs, inter- and intra-lake admixture, and phylogenetic trees; Figs. 2.2–2.3, and S1–S9) provide support for a scenario of diversification related to body size. However, our multivariate and F_{ST} analyses (Figs. 2.2a, 2.2b, S1, S2, S4 and Tables S4-S6) clearly demonstrate that the *sphyraena* group—the only entity previously resolved as monophyletic based on a scant number of mitochondrial and nuclear markers [11]—represents a separate genetic cluster with almost no gene flow shared with other members in the *chapalae* group. Based on this evidence, and the fact that *C. sphyraena* is the most taxonomically differentiated nominal species (particularly at characters commonly related to prey preferences; [10]), we hypothesize that *C. sphyraena* is in its early stages of ecological speciation. We note, however, that species-specific habitats are unknown for most of the *Chirostoma* species [21], and thus further ecological studies are necessary to better understand the evolutionary history of the *chapalae* and *sphyraena* groups.

Species delimitation and taxonomic implications

The BFD* species delimitation analyses provided strong support for four- and five-species delimitation scenarios (Table 1), while the coalescent-based species tree identified three major monophyletic groups (Fig. 2.3b). Recent simulation studies that evaluated the efficiency of MSC methods suggest that this model tends to confound population structure with species boundaries [4–6]. This becomes particularly problematic in recently-diverged and closely-related species, such as the *humboldtianum* group, where processes promoting differentiation lie at the intersection between population structure and species divergence [4], generating gene trees with short branches and with multi coalescent histories that make species tree inference challenging [51, 52]. This is not the case in the *humboldtianum* group, a recent species complex that diverged less than 1 Ma [11], where the genetic structure detected fewer species than traditional taxonomy.

The finer genetic structure recovered for the *sphyraena* group and the sub-species *C. e. copandaro* favor the selection of models of four and five species over a three species scheme. In this scenario, under a strict reciprocal monophyly criterion, none of the *C. sphyraena* or the sub-species *C. e. copandaro* lineages would represent a species (phylogenetic species concept, PSC; [1]). However, our intra-lake admixture analyses suggest that there is no gene structure among *sphyraena* and *chapalae* groups (one of the requirements to be considered a species under the biological species concept, BSC; [1]). We estimated the intra-lake admixture analyses with the intention to evaluate if including the rest of the *humboldtianum* group species was covering finer genetic structure within each lake. Our results demonstrated that the best model for each lake is represented by only one population (Fig. S7). These results, combined with the admixture analyses that included all individuals (matrices A-E; Figs. 2.2c and S6), phylogenetic trees (Fig. 2.3a and S8), and the species tree (Fig. 2.3b) suggest that *sphyraena* and *chapalae* represent the same genetic clade where sphyraena is an incipient species, as proposed by the multivariate analyses (Figs. 2.2a, 2b, S2, and S4), and pairwise F_{ST} (Tables S4-S7). We suggest that subject to further evidence (*e.g.*, ecological or ethological), the *sphyraena* group and the sub-species of *C. e.*

copandaro should not be considered species *per se* as their ancestral polymorphism has not been fully sorted by genetic drift. Hence, taking into account multiple lines of evidence (population genomics, phylogenomics, morphology, biogeography, and ecology), we propose that *estor*, *humboldtianum,* and *chapalae* groups constitute three well-differentiated species (*C. estor*, *C. humboldtianum sensu stricto*, and *C. chapalae* respectively).

The observed discordance between our genomic analyses, which resolved individuals from each lake as monophyletic, and previous studies [11, 18] where clades are not so cohesively clustered by geography, suggests that few mitochondrial and nuclear markers do not have sufficient statistical power to resolve the phylogenetic relationships of this recently diverged group. Here, we demonstrate that the use of thousands of SNP loci collectively provide strong power to detect phylogenetic signal while reducing the probability of stochastic error [53], as has also been demonstrated in several species of freshwater [46, 54–57] and marine [58–61] fishes. Additionally, we observed a discordance between the phylogenetic placement of the three main clades—where the phylogeny inferred using *mt*DNA places individuals from the *estor* group as basal, while the ddRADseq analyses as one of the most recent clades. Such disagreement could be related to the characteristics of the type of markers (e.g., matrilineal inheritance) as mitonuclear discordance can lead to an inaccurate estimation of the evolutionary history of the species, ultimately misguiding species delimitation [62–64]. In this case, the ddRADseq results are congruent with the geological events that shaped the lakes of central Mexico. Because of the *humboldtianum* group is restricted to lacustrine environments, our inferences make sense when they are combined with biogeographic information.

Finally, in all cases, the nine-species model was strongly rejected thus refuting the morphological-based hypothesis *sensu* Barbour [10]. Our study represents a rare case where genome-wide data evidences an over estimation of species diversity based on morphological characters. The delineation of such morphospecies was based on several characters, particularly head and body traits, that have been considered as the basis for the current taxonomy in the genus. However, trait measurements such as the jaw length, jaw shape, teeth size, number of gill rakers, and body shape are subjected to great environmental plasticity related to the species' trophic ecology and habitat characteristics, making it difficult to find diagnostic characters among *Chirostoma* species [10, 19, 65].

Selection across lakes

The study of alleles involved in local adaptation can unveil loci responsible for adaptive differences among populations. Our analyses of outlier loci detected 294 putative loci under selection, of which at least 106 are related to important biological processes (Table S4). For example, we detected two SNPs associated with immune response loci (*kpna3* and *bcl11b* genes). Previous studies in which loci related to the immune response of trout fish [67] and stickleback fishes [68–70] highlight the importance of these loci in the response of fishes to different pathogens during cladogenetic events. Thus, it is possible that the selection of loci associated with the immune response could be related to exposure to lake-specific pathogens during the colonization of new habitats, promoting the local adaptation and divergence of the *humboltianum* group once geographic isolation started.

Other genes detected in this analysis were the *or132-1*, which is an odor receptor associated with the detection of food in the aquatic system [71], and *sws2a*, *lws2a*, and *rh2-1* genes associated with the visual sensors. These findings could be related to the photic environment of each lake: whereas Lakes Chapala and Pátzcuaro are shallow eutrophic water bodies with a high sediment charge of turbid water [72, 73], Lake Zacapu is a clear-water lake where the light reaches an average depth of 4.3 m (up to 11.5 m in some places; [13]). Due to a different photic environment, this could affect the planktonic community of each lake, promoting the selection of olfactory and visual sensors, which have been identified as important in the diversification in other fishes [74]. Finally, loci associated with the skeleton muscle apparatus, as nebulin (*neb*) and titin (*ttn*) genes, and genes involved in growth hormone regulation, as cullin-associated NEDD8-dissociated protein 1 (*cand1*), and small glutamine-rich tetratricopeptide repeated containing alpha (*sgta*), could also be involved in the phenotypic plasticity of the size of the species of the *humboltianum* group.

We note that these inferences need to be taken with caution as there is a limited performance of F_{ST} outlier approaches in non-model organisms to identify candidate genes without a reference genome [75], particularly if the demographic history of the species is not modeled accurately [76]. Reduced representation genomic libraries that use restriction enzymes to cut the DNA may fail to identify key loci as these techniques only capture a small portion of the genome [77, 78], while many loci are lost due to low coverage and filtering [66, 79]. Without a reference genome, these results are highly sensitive to false positives such as loci linked to sites experiencing purifying selection that will present greater variation that can be confounded with local adaptation to environmental factors [66]. Thus, the set of candidate genes identified in this study, provides a starting point for further targeted research into the operative mechanisms of selection within the *humboltianum* group.

Genomic diversity, *Ne*, fishery management and conservation

Our genome-wide SNP analyses revealed that genomic diversity within the *C. humboldtianum* species complex is low (Table 2). This genomic diversity pattern has been reported in other lacustrine fish species such as the Nile tilapia [80] and stickleback fishes [69]. The low heterozygosity in freshwater fish species may be related to smaller effective population sizes and varying demographic histories involving bottlenecks during the recent colonization of new freshwater habitats, or to lake environmental history [69]. The *humboldtianum* group diverged and colonized the lakes of the central Mexico plateau during recent evolutionary times, between 0.58 to 0.076 Mya [11], and presented sudden demographic expansions from a smaller number of founders [81], which may explain their low genetic diversity. Also, these fishes have recently decreased demographically due to an over-exploitation of the commercial fishery, which could have also influenced the genomic diversity [82]. Although the majority of the results of *LDNe* and *HetExcessNe* estimators present infinite values indicating the effect of sampling error caused by the small sample size (1-24) [83, 84], the *Ne* values observed for the *humboltianum* group were similar o even smaller to lacustrine fish species such as *Amphilophus labiatus* and *A. citrinellus* [85], or species that experiment population collapse as *Oscorhynchus nerka* [86], suggesting a

genetic population effect of historical bottleneck and current fishery exploitation on *humboltianum* group.

Although in this study we have no evidence of the presence of hybrid individuals in natural lakes, as has been previously reported [10, 87]—despite the frequent translocation events promoted by the aquaculture policies since 1970 [88, 89]—, we detected some hybrid individuals of *C. humboldtianum sensu stricto* that share genomic information with the sub-clade I from Chapala Lake in artificial dams. These hybrids could be the result of introductions of different species of the "*humboltianum* group" into several artificial dams of the region [8, 81, 88], with the purpose to promote artisanal fisheries of this economically important resource and improving the local economy. However, it would be important to evaluate the effectiveness of the introductions in the local fisheries and their impact on natural populations within the *humboltianum* group. Some hybrid individuals may show hybrid vigor and could become better competitors than local species, making this practice another factor in the degradation of natural populations [8].

Chirostoma species represent a highly important economic and cultural resource since pre-Hispanic times [8]. However, species in this genus have been heavily overfished, leading to the collapse of several populations and severe conservation problems where some species are now considered extinct or in danger of extinction across several locations [8, 9]. Currently, the *humboldtianum* group is threatened by several factors, particularly habitat loss, pollution, the introduction of exotic species, and overfishing [50]. The decrease of silverside fish populations has also caused the collapse of their fisheries, negatively impacting the local fishermen's communities [8]. Another consequence of the over-exploitation of these fishes is the decrease in the capture size and the age of maturity size [8, 9]. Thus, the delimitation of operational genomic units is critical for fisheries management and conservation plans [8]. Additionally, information on the genomic structure and genetic diversity within and between natural populations of the *humboldtianum* group are crucial to understanding their evolutionary ability to cope with environmental changes [90]. Herein, our results support the presence of four genomic groups within the *humboltianum* group, distributed in the Lakes Chapala, Pátzcuaro-Zirahuén, and Zacapu. We strongly recommend revising management and conservation plans taking into consideration this new evidence.

2.6 Conclusions

Overall, our genome-wide analyses implementing ~2K–~33K SNP loci, under the light of morphological and ecological lines of evidence, provided a remarkable resolution to address a convoluted case of species delimitation within the *humboldtianum* group that was not previously achieved with the use of fewer markers. We resolved four genomic clusters arranged into three geographically cohesive clades (clade I, *chapalae*, and *sphyraena* groups from Lake Chapala; clade II, *humboldtianum sensu stricto* group from Lake Zacapu and dams; and clade III, *estor* group from Lakes Pátzcuaro and Zirahuén). These groups are not in agreement with the previously described morphospecies nor with the 'peces blancos' and 'charales' ecotypes, for which we reject the morphology-based hypothesis and a scenario of diversification under ecological selection

driven by the size of the species. All in all, our results suggest that the main cladogenetic events that gave rise to the three clades within the *humboldtianum* group resulted from allopatric processes generated by the complex geologic history of the Lerma-Chapala paleo system, while the intra-lake divergence of the *sphyraena* group could be the product of ecological speciation—a hypothesis that needs further investigation. Finally, the low levels of genetic diversity and *Ne* values observed inside each genomic cluster should be considered to address their conservation status. It is critical to highlight that lumping the nine morphospecies into three does not imply reducing conservation efforts but enforcing the inclusion of molecular information to create management strategies and conservation plans. All in all, our study represents a rare case where the use of genome-wide data evidence taxonomic over-splitting based on morphological information, while it emphasizes that the use of integrative approaches is fundamental to address complex species delimitation scenarios.

2.7 References

1. De Queiroz K. Species Concepts and Species Delimitation. Syst Biol. 2007;56:879–86.

2. Parker E, Dornburg A, Struthers CD, Jones CD, Near TJ. Phylogenomic Species Delimitation Dramatically Reduces Species Diversity in an Antarctic Adaptive Radiation. *Syst Biol*. 2021;**0**:1–20.

3. Schunter C, Garza JC, Macpherson E, Pascual M. SNP development from RNA-seq data in a nonmodel fish: How many individuals are needed for accurate allele frequency prediction? *Mol Ecol Resour*. 2014;**14**:157–65.

4. Sukumaran J, Knowles LL. Multispecies coalescent delimits structure, not species. *Proc Natl Acad Sci* USA. 2017;**114**:1607–11.

5. Chambers EA, Hillis DM. The multispecies coalescent over-splits species in the case of geographically widespread taxa. *Syst Biol.* 2020;**69**:184–93.

6. Hillis DM, Chambers EA, Devitt TJ. Contemporary Methods and Evidence for Species Delimitation. *Ichthyol Herpetol.* 2021;**109**:895–903.

7. Wiens JJ. Species delimitation: New approaches for discovering diversity. *Syst Biol.* 2007;**56**:875–8.

8. Rojas-Carrillo P, Sasso-Yada LF. El pescado blanco. *Rev Digit Univ*. 2005;**6**:2–18.

9. Alaye-Rahy N, Meléndez-Galicia C, Romero-Acosta C, Hernández-Montaño D. Estado actual de la pesquería de pescado blanco Chirostoma estor del Lago de Pátzcuaro, Michoacán, México. *Cienc Pesq.* 2017;**25**:5–16.

10. Barbour CD. The systematics and evolution of the genus *Chirostoma* Swainson (Pisces, Atherinidae). *Tulane Stud Zool Bot.* 1973;**18**:97–141. https://www.biodiversitylibrary.org/part/11912.

11. Betancourt-Resendes I, Perez-Rodríguez R, Barriga-Sosa IDLA, Piller KR, Domínguez-Domínguez O. Phylogeographic patterns and species delimitation in the endangered silverside "*humboldtianum*" clade (Pisces: Atherinopsidae) in central Mexico: understanding their evolutionary history. *Org Divers Evol*. 2020;**20**:313–30.

12. Barbour CD. A Biogeographical History of *Chirostoma* (Pisces : Atherinidae): A Species Flock from the Mexican Plateau. *Copeia*. 1973;**0**:533–56.

13. Ayala-Ramírez GL, Ruiz-Sevilla G, Chacón-Torres A. La laguna de Zacapu, Michoacán. In: De la Lanza-Espino G, Hernández-Pulido S, editors. Las Aguas Interiores de México: Conceptos y Casos. México D.F.: AGT EDITOR, S.A.; 2007. p. 268–84.

14. Soria-Barreto M, Paulo-Maya J. Morfometría comparada del aparato mandibular en especies de *Chirostoma* (Atheriniformes : Atherinopsidae) del Lago de Pátzcuaro , Michoacán , México Morphometric comparison of the mandibular region in species of *Chirostoma* (Atheriniformes : Atherinopsidae). *Hidrobiológica*. 2005;**15**:161–8.

15. Barbour CD, Chernoff B. Comparative morphology and morphometric of the pescados blancos (genus *Chirostoma*) from Lake Chapala Mexico. In: Echelle AA, Kornfield I, editors. Evolution of fish species flock. Orono, Me.: Univ. Maine Orono; 1984. pp. 111–27.

16. De Los Angeles Barriga-Sosa I, Ibáñez-Aguirre AL, Arredondo-Figueroa JL. Morphological and genetic variation in seven species of the *endangered Chirostoma "humboldtianum species group"* (Atheriniformes: Atherinopsidae). *Rev Biol Trop*. 2002;**50**:199–216.

17. Echelle AA, Echelle AF. Evolutionary genetics of a "species flock" Atherinid fishes on mesa central of Mexico. In: Echelle AA, Kornfield I, editors. Evolution of fish species flock. Orono, Me.: Univ. Maine Orono; 1984. p. 93–110.

18. Bloom DD, Piller KR, Lyons J, Mercado-Silva N, Medina-Nava M. Systematics and biogeography of the silverside tribe menidiini (Teleostomi: Atherinopsidae) based on the mitochondrial ND2 gene. *Copeia*. 2009;408–17.

19. Foster K, Bower L, Piller K. Getting in shape: Habitat-based morphological divergence for two sympatric fishes. *Biol J Linn Soc.* 2015;**114**:152–62.

20. Rodríguez Ruiz A, Granado Lorencio C. Características del aparato bucal asociadas al régimen alimenticio en cinco especies coexistentes del género *Chirostoma* (lago de Chapala; México). 1988;35–51.

21. Mercado-Silva N, Lyons J, Moncayo-Estrada R, Gesundheit P, Krabbenhoft TJ, Powell DL, et al. Stable isotope evidence for trophic overlap of sympatric Mexican Lake Chapala silversides (Teleostei: Atherinopsidae: *Chirostoma* spp.). *Neotrop Ichthyol.* 2015;**13**:389–400.

22. Chaplin K, Sumner J, Hipsley CA, Melville J. An Integrative Approach Using Phylogenomics and High-Resolution X-Ray Computed Tomography for Species Delimitation in Cryptic Taxa. *Syst Biol.* 2020;**69**:294–307.

23. Hosegood J, Humble E, Ogden R, de Bruyn M, Creer S, Stevens GMW, *et al*. Phylogenomics and species delimitation for effective conservation of manta and devil rays. *Mol Ecol.* 2020;**29**:4783–96.

24. Nieto-Montes de Oca A, Barley AJ, Meza-Lázaro RN, García-Vázquez UO, Zamora-Abrego JG, Thomson RC, et al. Phylogenomics and species delimitation in the knob-scaled lizards of the genus *Xenosaurus* (Squamata: Xenosauridae) using ddRADseq data reveal a substantial underestimation of diversity. *Mol Phylogenet Evol.* 2017;**106**:241–53.

25. Leaché AD, Fujita MK. Bayesian species delimitation in West African forest geckos (*Hemidactylus fasciatus*). *Proc R Soc B Biol Sci.* 2010;**277**:3071–7. doi:10.1098/rspb.2010.0662.

26. Poelstra JW, Salmona J, Tiley GP, Schüßler D, Blanco MB, Andriambeloson JB, et al. Cryptic Patterns of Speciation in Cryptic Primates: Microendemic Mouse Lemurs and the Multispecies Coalescent. *Syst Biol.* 2021;**70**:203–18.

27. Ferrari Pedraglio L. Avances en el conocimiento de la Faja Volcánica Transmexicana durante la última década. *Boletín la Soc Geológica Mex.* 2000;**53**:84–92.

28. Quintero-Legorreta O. Análisis estructural de fallas potencialmente activas. *Boletín la Soc Geológica Mex.* 2002;**55**:12–29.

29. Israde-Alcántara I, Velázquez-Durán R, Lozano-García MS, BischoffJ., Domínguez-Vázquez G, Victor Hugo GM. Evolución Paleolimnológica del Lago Cuitzeo , Michoacán durante el Pleistoceno-Holoceno. *Boletín la Soc Geológica Mex*. 2010;**62**:345–57.

30. Israde-Alcantara I, Garduño-Monroy VH. Lacustrine record in a volcanic intra-arc setting: the evolution of the Late Neogene Cuitzeo basin system (central- western mexico, Michoacán). *Palaeogeogr Palaeoclimatol Palaeoecol*. 1999;**151**:209–27.

31. Israde-Alcántara I. Lagos Volcánicos y Tectónicos de Michoacán. In: Garduño-Monroy VH, Corona-Chávez P, Israde-Alcántara I, Menella L, Arreygue E, B B, et al., editors. Carta Geológica de Michoacán Escala 1:250000. Universidad Michoacana de San Nicolás de Hidalgo; 1999. p. 45–72.

32. Domínguez-Domínguez O, Alda F, De León GPP, García-Garitagoitia JL, Doadrio I. Evolutionary history of the endangered fish *Zoogoneticus quitzeoensis* (Bean, 1898) (Cyprinodontiformes: Goodeidae) using a sequential approach to phylogeography based on mitochondrial and nuclear DNA data. *BMC Evol Biol.* 2008;**8**:1–19.

33. Doadrio I, Domínguez O. Phylogenetic relationships within the fish family Goodeidae based on *cytochrome b* sequence data. *Mol Phylogenet Evol*. 2004;**31**:416–30.

34. Corona-Santiago DK, Doadrio I, Domínguez-Domínguez O. Evolutionary history of the live-bearing endemic *Allotoca diazi* species complex (Actinopterygii, Goodeinae): Evidence of founder effect events in the Mexican pre-Hispanic period. *PLoS One*. 2015;**10**:1–21.

35. Domínguez-Vázquez G, Osuna-Vallejo V, Castro-López V, Israde-Alcántara I, Bischoff JA. Changes in vegetation structure during the Pleistocene–Holocene transition in Guanajuato, central Mexico. *Veg Hist Archaeobot*. 2019;**28**:81–91. doi:10.1007/s00334-018-0685-8.

36. Seehausen O, Wagner CE. Speciation in freshwater fishes. *Annu Rev Ecol Evol Syst*. 2014;**45**:621–51.

37. Mayr E. Systematics and the origin of species. New York: Columbia University Press; 1942.

38. Rundle HD, Nosil P. Ecological speciation. *Ecol Lett*. 2005;**8**:336–52.

39. Bernardi G. Speciation in fishes. *Mol Ecol*. 2013;**22**:5487–502.

40. Barluenga M, Stölting KN, Salzburger W, Muschick M, Meyer A. Sympatric speciation in Nicaraguan crater lake cichlid fish. *Nature*. 2006;**439**:719–23.

41. Elmer KR, Fan S, Kusche H, Luise Spreitzer M, Kautt AF, Franchini P, *et al*. Parallel evolution of Nicaraguan crater lake cichlid fishes via non-parallel routes. *Nat Commun*. 2014;5.

42. Burress ED. Ecological diversification associated with the pharyngeal jaw diversity of Neotropical cichlid fishes. *J Anim Ecol*. 2016;**85**:302–13.

43. Salzburger W. Understanding explosive diversification through cichlid fish genomics. *Nat Rev Genet.* 2018;**19**:705–17. doi:10.1038/s41576-018-0043-9.

44. Takeyama T. Feeding Ecology of Lake Tanganyika Cichlids. In: Abate ME, Noakes DLG, editors. The Behavior, Ecology and Evolution of Cichlid Fishes. Dordrecht: Springer Netherlands; 2021. p. 715–51. doi:10.1007/978-94-024-2080-7_19.

45. Härer A, Bolnick DI, Rennison DJ. The genomic signature of ecological divergence along the benthic-limnetic axis in allopatric and sympatric threespine stickleback. *Mol Ecol.* 2021;**30**:451–63.

46. Jones FC, Grabherr MG, Chan YF, Russell P, Mauceli E, Johnson J, et al. The genomic basis of adaptive evolution in threespine sticklebacks. *Nature*. 2012;**484**:55–61.

47. Puebla O. Ecological speciation in marine *vs.* freshwater fishes. *J Fish Biol.* 2009;**75**:960–96.

48. Nagel L, Schluter D. Body Size, Natural Selection, and Speciation in Sticklebacks. *Evolution* (N Y). 1998;**52**:209–18.

49. Bay RA, Arnegard ME, Conte GL, Best J, Bedford NL, McCann SR, et al. Genetic Coupling of Female Mate Choice with Polygenic Ecological Divergence Facilitates Stickleback Speciation. *Curr Biol.* 2017;**27**:3344-3349.e4. doi:10.1016/j.cub.2017.09.037.

50. Miller RR, Minckley WL, Norris SM, of Michigan. Museum of Zoology U. Freshwater Fishes of México. University of Chicago Press; 2005. https://books.google.com.mx/books?id=MZXG-9jKygQC.

51. Yang Z, Rannala B. Bayesian species delimitation using multilocus sequence data. *Proc Natl Acad Sci*. 2010;**107**:9264–9. doi:10.1073/pnas.0913022107.

52. Rannala B, Edwards SVS V, Leaché A, Yang Z. The Multi-species Coalescent Model and Species Tree Inference. In: Scornavacca C, Delsuc F, Galtier N, editors. Phylogenetics in the Genomic Era. No commercial publisher Authors open access book; 2020. pp. 3.3:1--3.3:21. https://hal.archives-ouvertes.fr/hal-02535622.

53. Betancur R, Naylor GJP, Ortí G. Conserved genes, sampling error, and phylogenomic inference. *Syst Biol.* 2014;63:257–62.

54. Deagle BE, Jones FC, Chan YF, Absher DM, Kingsley DM, Reimchen TE. Population genomics of parallel phenotypic evolution in stickleback across stream-lake ecological transitions. *Proc R Soc B Biol Sci*. 2012;**279**:1277–86.

55. Fagbémi MNA, Pigneur LM, André A, Smitz N, Gennotte V, Michaux JR, et al. Genetic structure of wild and farmed Nile tilapia (*Oreochromis niloticus*) populations in Benin based on genome wide SNP technology. *Aquaculture.* 2021;535 January.

56. Burress ED, Alda F, Duarte A, Loureiro M, Armbruster JW, Chakrabarty P. Phylogenomics of pike cichlids (Cichlidae : Crenicichla): the rapid ecological speciation of an incipient species flock. *Evol Biol.* 2018;**31**:14–30.

57. Torati LS, Taggart JB, Varela ES, Araripe J, Wehner S, Migaud H. Genetic diversity and structure in Arapaima gigas populations from Amazon and Araguaia-Tocantins river basins. *BMC Genet.* 2019;**20**:1–13.

58. Carreras C, Ordóñez V, Zane L, Kruschel C, Nasto I, MacPherson E, *et al*. Population genomics of an endemic Mediterranean fish: Differentiation by fine scale dispersal and adaptation. *Sci Rep.* 2016:1–12.

59. Jackson AM, Semmens BX, De Mitcheson YS, Nemeth RS, Heppell SA, Bush PG, *et al*. Population structure and phylogeography in *Nassau grouper* (*Epinephelus striatus*), a mass-aggregating marine fish. *PLoS One*. 2014;9.

60. Laconcha U, Iriondo M, Arrizabalaga H, Manzano C, Markaide P, Montes I, et al. New nuclear SNP markers unravel the genetic structure and effective population size of albacore tuna (*Thunnus alalunga*). *PLoS One*. 2015;**10**:1–19.

61. Pedraza-Marrón C Del R, Silva R, Deeds J, Van Belleghem SM, Mastretta-Yanes A, Domínguez-Domínguez O, et al. Genomics overrules mitochondrial DNA, siding with morphology on a controversial case of species delimitation. *Proc R Soc B Biol Sci*. 2019;286.

62. Hughes LC, Cardoso YP, Sommer JA, Cifuentes R, Cuello M, Somoza GM, et al. Biogeography, habitat transitions and hybridization in a radiation of South American silverside fishes revealed by mitochondrial and genomic RAD data. *Mol Ecol.* 2020;**29**:738–51.

63. Marshall TL, Chambers EA, Matz M V., Hillis DM. How mitonuclear discordance and geographic variation have confounded species boundaries in a widely studied snake. *Mol Phylogenet Evol.* 2021;162 April:107194. doi:10.1016/j.ympev.2021.107194.

64. Shultz AJ, Baker AJ, Hill GE, Nolan PM, Edwards S V. SNPs across time and space: population genomic signatures of founder events and epizootics in the House Finch (*Haemorhous mexicanus*). *Ecol Evol.* 2016;**6**:7475–89.

65. Alarcón-Durán I, Castillo-Rivera MA, Figueroa-Lucero G, Arroyo-Cabrales J, Barriga-Sosa I de los Á. Diversidad morfológica en 6 poblaciones del pescado blanco *Chirostoma humboldtianum. Rev Mex Biodivers*. 2017;**88**:207–14. doi:10.1016/j.rmb.2017.01.018.

66. Hoban S, Kelley JL, Lotterhos KE, Antolin MF, Bradburd G, Lowry DB, *et al*. Finding the genomic basis of local adaptation: Pitfalls, practical solutions, and future directions. *Am. Nat.* 2016. **188**(4):379-97.

67. Amish SJ, Ali O, Peacock M, Miller M, Robinson M, Smith S, et al. Assessing thermal adaptation using family-based association and FST outlier tests in a threatened trout species. *Mol Ecol.* 2019;**28**:2573–93.

68. Robertson S, Bradley JE, MacColl ADC. Eda haplotypes in three-spined stickleback are associated with variation in immune gene expression. *Sci Rep*. 2016:1–9.

69. Jones FC, Chan YF, Schmutz J, Grimwood J, Brady SD, Southwick AM, *et al.* A genome-wide SNP genotyping array reveals patterns of global and repeated species-pair divergence in sticklebacks. *Curr Biol.* 2012;**22**:83–90. doi:10.1016/j.cub.2011.11.045.

70. Scharsack JP, Kalbe M, Harrod C, Rauch G. Habitat-specific adaptation of immune responses of stickleback (*Gastetosteus aculeatus*) lake and river ecotypes. *Proc R Soc B Biol Sci.* 2007;**274**:1523–32.

71. Leck KJ, Zhang S, Hauser CAE. Study of bioengineered zebra fish olfactory receptor 131-2: Receptor purification and secondary structure analysis. *PLoS One*. 2010;5.

72. Jimenez Baltazar J, Hernández Morales R, Domínguez Domínguez O. Caracterización fisicoquímica y microbiológica del vaso norte de Lago de Pátzcuaro, Michoacán, México. *Rev Latinoam el Ambient y las Ciencias*. 2018;**9**:400–15.

73. Lind OT, Dávalos-Lind L. An Introduction to the Limnology of Lake Chapala, Jalisco, Mexico. In: Hansen AM, van Afferden M, editors. The Lerma-Chapala Watershed: Evaluation and Management. Boston, MA: Springer US; 2001. pp. 139–49. doi:10.1007/978-1-4615-0545-7_6.

74. Hollenbeck CM, Portnoy DS, Gold JR. Evolution of population structure in an estuarine-dependent marine fish. *Ecol Evol*. 2019;**9**:3141–52.

75. Theodorou P, Radzevičiūtė R, Kahnt B, Soro A, Grosse I, Paxton RJ. Genome-wide single nucleotide polymorphism scan suggests adaptation to urbanization in an important

pollinator, the red-tailed bumblebee (*Bombus lapidarius* L.). *Proc R Soc B Biol Sci.* 2018;285.

76. Whitlock MC, Lotterhos KE. Reliable detection of loci responsible for local adaptation: Inference of a null model through trimming the distribution of FST. *Am Nat.* 2015;186 October:S24–36.

77. Andrews KR, Good JM, Miller MR, Luikart G, Hohenlohe PA. Harnessing the power of RADseq for ecological and evolutionary genomics. *Nat Publ Gr.* 2016. doi:10.1038/nrg.2015.28.

78. Schweyen H, Rozenberg A, Leese F. Detection and removal of PCR duplicates in population genomic ddRAD studies by addition of a degenerate base region (DBR) in sequencing adapters. *Biol Bull.* 2014;**227**:146–60.

79. Lowry DB, Hoban S, Kelley JL, Lotterhos KE, Reed LK, Antolin MF, *et al*. Breaking RAD: an evaluation of the utility of restriction site-associated DNA sequencing for genome scans of adaptation. *Mol Ecol Resour.* 2017;**17**:142–52.

80. Kajungiro RA, Palaiokostas C, Pinto FAL, Mmochi AJ, Mtolera M, Houston RD, *et al*. Population Structure and Genetic Diversity of Nile Tilapia (*Oreochromis niloticus*) Strains Cultured in Tanzania. *Front Genet.* 2019;10 December:1–12.

81. García Martínez RM, Mejía O, García-De León FJ, Barriga-Sosa IDLA. Extreme genetic divergence in the endemic fish *Chirostoma humboldtianum*: Implications for its conservation. *Hidrobiologica*. 2015;**25**:95–106.

82. Navarrete Salgado NA. *Chirostoma* (Menidia): Ecología Y Utilización Como Especie De Cultivo En Estanques Rústicos. *BIOCYT Biol Cienc y Tecnol*. 2017;**10**:736–48.

83. Waples RS, Do C. Linkage disequilibrium estimates of contemporary *Ne* using highly variable genetic markers: A largely untapped resource for applied conservation and evolution. *Evol Appl.* 2010;**3**:244–62.

84. England PR, Cornuet JM, Berthier P, Tallmon DA, Luikart G. Estimating effective population size from linkage disequilibrium: Severe bias in small samples. *Conserv Genet.* 2006;**7**:303–8.

85. Sowersby W, Cerca J, Wong BBM, Lehtonen TK, Chapple DG, Leal-Cardín M, *et al*. Pervasive admixture and the spread of a large-lipped form in a cichlid fish radiation. *Mol Ecol.* 2021;**30**:5551–71.

86. Setzke C, Wong C, Russello MA. Genotyping-in-Thousands by sequencing of archival fish scales reveals maintenance of genetic variation following a severe demographic contraction in kokanee salmon. *Sci Rep.* 2021;**11**:1–10. doi:10.1038/s41598-021-01958-0.

87. Alaye-Rahy N. Híbridos entre especies del género *Chirostoma* del Lago de Pátzcuaro, Michoacán, México. *Cienc Pesq.* 1996;**13**:10–7.

88. Medina-Nava M. Ictiofauna de la subcuenca del Río Angulo cuenca del Lerma-Chapala, Michoacán. *Zool Inf.* 1997;**35**:25–52.

89. Rojas-Carrillo PM, Fuentes-Castellanos D. Historia y avances del cultivo del pescado blanco. Instituto Nacional de Pesca, Dirección General de Investigación y Acuacultura; 2003.

90. Reed DH, Frankham R. Correlation between fitness and genetic diversity. *Conserv Biol*. 2003;**17**:230–7.

91. Peterson BK, Weber JN, Kay EH, Fisher HS, Hoekstra HE. Double digest RADseq: An inexpensive method for de novo SNP discovery and genotyping in model and non-model species. *PLoS One*. 2012;7.

92. Catchen JM, Hohenlohe PA, Bassham S, Amores A, Cresko WA. Stacks: an analysis tool set for population genomics. *Mol Ecol*. 2013;**22**:3124–40.

93. Catchen JM, Amores A, Hohenlohe P, Cresko W, Postlethwait JH. Stacks : Building and Genotyping Loci De Novo From Short-Read Sequences. *Genes, Genomes*. 2011;**1** August:171–82.

94. Mastretta-Yanes A, Arrigo N, Alvarez N, Jorgensen TH, Piñeros D, Emerson BC. Restriction site-associated DNA sequencing, genotyping error estimation and *de novo* assembly optimization for population genetic inference. *Mol Ecol*. 2015;28–41.

95. Paris JR, Stevens JR, Catchen JM. Lost in parameter space: A road map for Stacks. *Methods Ecol Evol*. 2017;**8**:1360–73.

96. Pedraza-Marrón C del R, Silva R, Deeds J, Van Belleghem SM, Mastretta-Yanes A, Domínguez-Domínguez O, *et al*. Genomics overrules mitochondrial DNA, siding with morphology on a controversial case of species delimitation. *Proc R Soc B Biol Sci.* 2019;286.

97. Díaz-Arce N, Rodríguez-Ezpeleta N. Selecting RAD-Seq data analysis parameters for population genetics: The more the better? *Front Genet.* 2019;**10** MAY:1–10.

98. Linck E, Battey CJ. Minor allele frequency thresholds strongly affect population structure inference with genomic data sets. *Mol Ecol Resour*. 2019;**19**:639–47.

99. Patterson N, Price AL, Reich D. Population structure and eigen analysis. *PLoS Genet.* 2006;**2**:2074–93.

100. Jombart T, Devillard S, Balloux F. Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. *BMC Genet.* 2010;**11**:1–15.

101. Miller JM, Cullingham CI, Peery RM. The influence of a priori grouping on inference of genetic clusters: simulation study and literature review of the DAPC method. *Heredity* 2020;**125**:269–80. doi:10.1038/s41437-020-0348-2.

102. Jombart T. Adegenet: A R package for the multivariate analysis of genetic markers. *Bioinformatics*. 2008;**24**:1403–5.

103. Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res*. 2009;**19**:1655–64.

104. Alexander DH, Lange K. Enhancements to the ADMIXTURE algorithm for individual ancestry estimation. *BMC Bioinformatics*. 2011;12.

105. Wickham H. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York; 2016. http://ggplot2.org.

106. Francis RM. pophelper: an R package and web app to analyse and visualize population structure. *Mol Ecol Resour*. 2017;**17**:27–32.

107. Excoffier L, Laval G, Schneider S. Arlequin (version 3.0): An integrated software package for population genetics data analysis. *Evol Bioinforma.* 2005;1 May 2016:117693430500100.

108. Rice WR. Analyzing Tables of Statistical Tests. *Evolution* (N Y). 1989;43:223–5.

109. Nguyen LT, Schmidt HA, Von Haeseler A, Minh BQ. IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol Biol Evol.* 2015;**32**:268–74.

110. Hoang DT, Chernomor O, Von Haeseler A, Minh BQ, Vinh LS. UFBoot2: Improving the ultrafast bootstrap approximation. *Mol Biol Evol*. 2018;**35**:518–22.

111. Revell LJ. phytools: An R package for phylogenetic comparative biology (and other things). *Methods Ecol Evol*. 2012;**3**:217–23.

112. Bryant D, Bouckaert R, Felsenstein J, Rosenberg NA, Roy Choudhury A. Inferring Species Trees Directly from Biallelic Genetic Markers : Bypassing Gene Trees in a Full Coalescent Analysis. *Mol Biol Evol.* 2012;**29**:1917–32.

113. Bouckaert R, Heled J, Kühnert D, Vaughan T, Wu CH, Xie D, *et al*. BEAST 2: A Software Platform for Bayesian Evolutionary Analysis. *PLoS Comput Biol*. 2014;**10**:1–6.

114. Rambaut A, Drummond AJ, Xie D, Baele G, Suchard MA. Posterior summarization in Bayesian phylogenetics using Tracer 1.7. *Syst Biol*. 2018;**67**:901–4.

115. Bouckaert RR. DensiTree: Making sense of sets of phylogenetic trees. *Bioinformatics*. 2010;**26**:1372–3.

116. Leaché A, Fujita M, Minin V, Bouckaert R. Species Delimitation using Genome-Wide SNP Data. *Syst Biol*. 2014. doi:10.1101/001172.

117. Kass RE, Raftery AE. Bayes Factors. *J Am Stat Assoc*. 1995;**90**:773–95.

118. Leaché AD, Bouckaert RR. Species trees and species delimitation with SNAPP: a tutorial and worked example. 2018.

119. Peakall R, Smouse PE. GENALEX 6: Genetic analysis in Excel. Population genetic software for teaching and research. *Mol Ecol Notes*. 2006;**6**:288–95.

120. Zhdanova OL, Pudovkin AI. Nb_HetEx: A program to estimate the effective number of breeders. *J Hered*. 2008;**99**:694–5.

121. Nomura T. Estimation of effective number of breeders from molecular coancestry of single cohort sample. *Evol Appl*. 2008;**1**:462–74.

122. Do C, Waples RS, Peel D, Macbeth GM, Tillett BJ, Ovenden JR. NeEstimator v2: Re-implementation of software for the estimation of contemporary effective population size (*Ne*) from genetic data. *Mol Ecol Resour*. 2014;**14**:209–14.

123. Foll M, Gaggiotti O. A genome-scan method to identify selected loci appropriate for both dominant and codominant markers: A Bayesian perspective. *Genetics*. 2008;**180**:977–93.

124. Foll M. BayeScan v2.0 User Manual. 2010;1–9.

125. R Core Team. R: A Language and Environment for Statistical Computing. 2020. https://www.r-project.org/.

126. Del Pedraza-Marrón CR. Cambios en la distribución de los peces de agua dulce del Centro de México y sus posibles causas antropogénicas. Universidad Michoacana de San Nicolás de Hidalgo; 2012.

Chapter 3

Beneath the Waves: Depth, Temperature, and Spatial Components Driving Genetic Differentiation at Micro and Macroevolutionary Scales in Tropical Blennies

3.1 Abstract

The study of spatio-temporal phylogenetic concordance of co-distributed lineages—comparative phylogeography—provides invaluable insights into understanding the influence of historical (e.g., geophysical events) and contemporary barriers (e.g., habitat gaps) promoting population divergence and ultimately speciation. Here, we investigated the labrisomid blennies (genus *Malacoctenus*), sequencing *ca.* 28K SNPs from over 500 individuals representing 23 (92%) species, to assess the effect of historical and contemporary barriers in the Tropical Eastern Pacific (TEP) and the Tropical Atlantic (TA) biogeographic realms. Given their strong association with reefs, subtle habitat disruptions may lead to genetic isolation, providing an ideal system for comparative phylogeography. The observed population structure patterns identified the Sinaloan and Central American breaks in the TEP; and the Bahamas and Eastern Caribbean breaks in the TA, as major barriers. We identified five independent evolutionary lineages, one supported by the geometric morphometric analyses. Major seascape genomic features variables promoting divergence at the population and interspecies levels included depth, temperature, and physical features. Notably, depth appeared as a primary cladogenetic driver in the TEP, leading to niche divergence between tide pool- and reef-associated clades. Finally, our time-calibrated analyses elucidated an Eastern Atlantic origin of the clade followed by an east-to-west dispersal. The Isthmus of Panama influenced the evolutionary history of the genus but cladogenetic events were not synchronous. In summary, through an integrative framework that encompasses genomic, morphological, and environmental data, this study evaluated the relative contributions of barriers in the TEP and TA, as well as other contemporary and historical factors influencing divergence.

3.2 Introduction

Understanding the mechanisms governing diversification and speciation processes in nature is a major goal in evolutionary biology. Because microevolutionary processes operating at population levels ultimately drive macroevolution (1), the study of patterns and processes generating genetic divergences at short-, intermediate-, and long-term evolutionary scales, going through a continuum from micro- to macro-evolution is of primary interest. In this context, phylogeography, which involves the study of spatial patterns of genetic diversity within populations or among closely related species, serves as a bridge between microevolution and broader spatial and temporal phylogenetic scales (2). As current geographic structure can potentially reflect past phylogeographic history (3), the assessment of barriers to dispersal promoting population structure, lineage diversification, and ultimately speciation is likewise crucial to understand the evolution of biodiversity.

Unlike terrestrial environments where visible geographical features like mountain chains or rivers can physically isolate animal populations, the vast expanses of water in the ocean are subtler, posing greater challenges in identifying such barriers (4, 5). While life history traits of marine organisms strongly influences their dispersal capacity (e.g., pelagic larvae can be transported large distances via oceanic currents) (6, 7), recent studies utilizing a seascape genomic framework provide evidence that seascape composition (ecological variables) and seascape configuration (spatial features) may hinder connectivity even in species with high dispersal capacities (8, 9). Such environmental variables (e.g., temperature) or oceanographic features (e.g., oceanic currents) may limit dispersal capacity. In this scenario, the comparison of phylogeographic patterns of co-distributed marine species—comparative phylogeography—provides an opportunity to identify historical (e.g., geophysical events) and contemporary (e.g., habitat gaps) barriers that restrict gene flow, isolate populations, and ultimately promote speciation in the marine realm.

The last emergence of the Isthmus of Panama during the Pliocene (~2.8 Mya) embodies one of the best-known large-scale geophysical events that generated an impassible barrier for marine organisms distributed in the Tropical Eastern Pacific (TEP) and the Tropical Atlantic (TA), unchaining a series of diversification events in the Neotropics (10). This historical barrier is known to have triggered geminate or sister-species pairs on each side of the isthmus (11). At more recent scales, several studies have also highlighted habitat gaps within the TEP and the TA as contemporary barriers to gene flow (12–15). Furthermore, these habitat gaps are believed to delineate the boundaries between biogeographic provinces within the TEP (13). Although the limit of biogeographic provinces in the TA relates more to oceanographic processes, biogeographic provinces within both regions feature major environmental differences (16–18). Thus, mayor connectivity patterns may arise from large-scale environmental variation across biogeographic realms. Molecular research investigating intraspecific population structure have identified several marine breaks, primarily linked to environmental changes driven by seasonal upwellings, oceanographic gyres (14, 15).

In the TEP, a rocky coast that goes from the Peninsula of Baja California through the north of Peru is mainly interrupted in two major geographic stretches characterized by sandy bottoms:

the Sinaloan and the Central American Breaks. The large expanse of open ocean between the Galápagos Archipelago and the continental shelf, and an oceanographic gyre west of Panama, although less effective, have also been suggested to impact connectivity in the region (Fig. 1A). In the TA, a wide variety of environments from rocky, coralline reefs, mangroves, and sandy shorelines that stretch from the coast of Florida through the Gulf of Mexico and northern South America are hypothesized to act as phylogeographic breaks. These include the Bahamas, the Gulf of Mexico, the Yucatán Current, the Mona Passage, the Northeastern Colombian Coast, and the Western and Eastern Caribbean breaks (13, 19–22) (Fig. 1B). Phylogeographic studies examining the influence of these contemporary barriers in the TEP and TA have primarily focused on a limited number of species or relied on a scant number of mitochondrial and nuclear markers (8, 14, 23–26). As a result, many of these studies have failed to identify shared genetic structure patterns, which are crucial for understanding the overall impact of dispersal barriers on population connectivity in marine environments (e.g., 8, 14, 23–26).

Comprising 25 species, primarily New World endemics, labrisomid blennies within the genus *Malacoctenus* offer an excellent system for marine phylogeographic analysis in the Neotropics. Due to their strong association with rocky habitats and coral reefs, even minor disturbances in their habitat can result in genetic isolation. Identifying these species can be quite challenging, as traditional taxonomic characteristics used for differentiation often exhibit overlap among closely related species. There have been suggestions of concealed cryptic diversity and the presence of species complexes within these regions (12, 26). Notwithstanding these challenges, *Malacoctenus* provides an opportunity to investigate the influence of historical factors, such as the Isthmus of Panama, and contemporary barriers like the Sinaloan Break, on population divergence in the Tropical Eastern Pacific (TEP) and the Tropical Atlantic (TA).

This study employs a multifaceted approach that integrates population genetics, comparative phylogeography, phylogenomics, geometric morphometrics, and seascape genomics. It is based on a dataset of thousands of loci obtained through restriction-site-associated DNA sequencing (RADseq). The dataset comprises hundreds of individuals, representing 23 out of the 25 *Malacoctenus* species, collected from numerous locations across their distribution ranges in the TEP and TA regions. With this dataset we aimed to: (i) identify current patterns of population structure of co-distributed species and test whether these patterns are concordant with previously-identified contemporary barriers in the regions; (ii) evaluate whether morphological variation is correlated with genetic breaks and population structure analyses; (iii) elucidate major environmental variables and spatial components influencing genetic differentiation at both intra- and inter-species levels using seascape genomic analyses; and (iv), at broader macroevolutionary scales, investigate the biogeographic history and macroevolutionary ecology of the group and assess the extent to which cladogenetic events in multiple subclades prompted by the rise of the isthmus of Panama have occurred synchronously. Our results ultimately contribute to our understanding of the evolutionary history of reef fishes and the forces driving marine connectivity and speciation in the marine realm more generally and the Neotropics in particular.

3.3 Materials and Methods

Extended information on the methods used in this study and supplementary information are provided in the *Appendix C, Materials and Methods*.

Sample collection and genomic data generation

We generated restriction site associated DNA sequencing (RADseq) data from 506 individuals, representing 23 out of the 25 species of *Malacoctenus* plus an outgroup (*Brockius striatus*), collected across 38 localities in the Tropical Eastern Pacific (TEP) and Tropical Atlantic (TA) (30) (*Appendix C, Materials and Methods,* Fig. S1). High-yield DNA extractions and RADseq libraries were prepared at the University of Wisconsin Biotechnology Center (UWBC) by applying the double-digest (ddRADseq) protocol of (68), using the enzymes *Pstl* and *Bfal* with a size selection window of 350–550 bp. ddRADseq libraries were sequenced at Novogene Company Inc. (CA, USA) using a partial lane of Illumina NovaSeq 6000 PE150.

Matrix assembly for evolutionary analyses at different scales

Because genomic datasets are commonly affected by quality control issues (69), we conducted several steps to tackle this, including corroborating taxonomic identification and performing an extensive quality assessment of the raw sequences. As we were interested in analyzing the genetic structure at both intra- and inter-specific levels, we assembled SNP-loci matrices in two different ways. At the microevolutionary scale, we conducted *de novo* assemblies for each of the 14 species across several populations. We used Stacks v2.59 (70), setting a minimum of three raw reads required to form a stack (*m*), allowing a maximum of five mismatches between loci (*M*), and six mismatches between loci of different individuals (*n*). At the macroevolutionary level, we conducted reference-based assemblies for 41 individuals spanning the 23 species of *Malacoctenus* (plus *B. striatus*) using Stacks and the closest reference genome available (*Salarias fasciatus*, Blenniidae). The resulting matrices at both scales were filtered to contain unlinked, bi-allelic, orthologous SNPs using the R packages *SNPfiltR* (71) and VCFtools v0.1.16 (72). The final SNP matrices at the intra-specific level consisted of 4,000–9,000 SNPs and were used to conduct population differentiation analyses. At the inter-specific level, matrices ranged between approximately 1,700 to 28,000 SNPs and were used for phylogenomic analyses. We also assembled a reference-based matrix including all 447 individuals that passed quality filters at the intra-species level. From this dataset, we assembled inter-specific matrices including only trans-isthmic species pairs to evaluate the synchronicity of cladogenetic events triggered by the Isthmus of Panama. These matrices included all sites with two to four individuals per species, consisting of up to eight individuals and 250K bp as *ecoevolity* preforms better when utilizing the full data set since the model doesn't require information about linkage among sites (31). To maximize the amount of genomic information we also assembled *de novo* matrices using the same individuals selected to represent the trans-isthmic pairs, resulting in over 1000K bp. Finally, seascape genomic analyses were conducted using SNP matrices ranging from *ca.* 4,000 to 6,000 SNPs at the intra-species level, and *ca.* 1,500 SNPs at the inter-species level. For more information, refer to the *Appendix C, Materials and Methods,* Fig. S2.

Microevolutionary analyses for individually sampled species

To characterize the genetic structure among populations, we used different genetic clustering approaches. First, we evaluated the optimal number of populations (*k*) by employing a maximum-likelihood approach with the ADMIXTURE v1.3.0 software (73, 74), and a Bayesian clustering method using fastSTRUCTURE (75). Next, we conducted a Discriminant Analysis of Principal Components (DPCA) with and without prior population assignment, using the R package adegenet (76). We evaluated both scenarios (determining *de novo* structure versus *a priori* grouping designations) as the sensitivity of the method to misspecifications in population assignment is unknown (77). Additionally, we estimated phylogenetic trees using IQ-TREE v2.1.2 (78) to represent the evolutionary relationships among individuals and identify evolutionary lineages corresponding to the potential breaks evaluated. For species where population differentiation approaches yield different hypotheses regarding population structure, we conducted an analysis of molecular variance (AMOVA) using the *poppr* package in R (79). Lastly, to generate a graphical representation of the spatial population structure (80), we applied the Estimate Effective Migration Surfaces (EEMS) method (github.com/dipetkov/eems), and conducted a spatial Analysis of Principal Components (sPCA) using the adegenet R package. We determined the relative strength of each break—effectiveness—by estimating the proportion of species affected by a given break. A potential break was considered effective when two or more analyses identified population structure patterns aligning to that break. For more information on the parameters used to run these analyses, refer to the *Appendix C, Materials and Methods*.

Phenotypic analyses of body shape

To quantify morphological disparity at the micro and macroevolutionary levels, we conducted 2D geometric morphometric analysis using 16 landmarks and 20 semi-landmarks. At the intra-specific level, we analyzed only *M. tetranemus* and *M. triangulatus* using a total of 47 and 35 high-quality photographs, respectively, taken during field expeditions or retrieved from online museum repositories. For *M. triangulatus* we also used 100 x-rays downloaded from the fish collection repository of the Smithsonian National Museum of Natural History (NMNH). At the inter-specific level, we used up to ten individuals per species for a total of 89 photographs covering twenty species that had high-quality photographs. We digitalized the landmarks and semi-landmarks using the package StereoMorph (81), and corrected for distortions, size and position of the different specimen images using a generalized Procrustes analysis (GPA). To account for landmark accuracy reflecting body-shape disparity, we created four different schemes with different subsets of landmarks (*Appendix C, Materials and Methods,* Fig. S12). We conducted a principal components analysis (PCA) using the *geomorph* (82) R package for each scheme.

Seascape genomics

To investigate the relative contribution of spatial distribution, and environmental variables on patterns of genetic variation, we conducted redundancy analyses (dbRDA, (83)) using the R package vegan (84). These analyses were ran at the intra-species level on the most widely distributed species within each biogeographic realm (*M. tetranemus* in the TEP and *M. triangulatus* in the TA), and at the inter-species level for each respective realm. These analyses

were fed with Moran's eigenvector maps (MEMs) and eight environmental variables (*e.g.*, sea surface temperature; *Appendix C, Materials and Methods*, Table S7) as explanatory variables, and principal components representing genomic variation as response variables. MEMs were used to describe the seascape configuration and spatial distance among sampling populations and were calculated by decomposing in-water distances computed using the marmap R package (85) through principal coordinate analysis (PCoA) using the R package adespatial (86). Environmental variables were extracted from the Copernicus online data repository (data.marine.copernicus.edu).

Macroevolutionary analyses, phylogenomic inference and biogeographic history

We estimated a mitochondrial tree using partial sequences of cytochrome oxidase subunit I (*COI*) (645 bp, 70% missing data) from 58 individuals representing 17 *Malacoctenus* species, and *L. striatus* as an outgroup. This dataset was formed from sequences retrieved from the National Center for Biotechnology Information (NCBI) data archive. For phylogenomic analyses, we estimated trees using five concatenated SNP-loci matrices, which were comprised of 41 individuals spanning the most divergent populations (when possible) of the 23 *Malacoctenus* species (plus *B. striatus*), ranging from 1,800 to 28,000 SNPs (29.1–81.1% of missing data). All phylogenetic trees were estimated using IQ-TREE. We also estimated a species tree for each of the five matrices under the multispecies coalescent model using SVDquartets implemented in PAUP* v4.0 (87). As the phylogenomic estimations were mostly consistent across matrices (see Results), we time-calibrated the 28 K species tree under the multispecies coalescent (MSC) model using the SNAPP v1.3.0 (88) plug-in, implemented in BEAST2 v2.6.7 (89). We used a secondary calibration prior at the root, using the minimum and maximum ages estimated by previous fossil-calibrated teleost phylogenies for the most recent common ancestor (MRCA) of *Malacoctenus* and *Brockius striatus* (35–32 Mya; *Appendix C, Materials and Methods*, Table S1). To generate a time-calibrated tree with all 447 individuals that passed quality filters, we used secondary calibrations extracted from the time-calibrated species tree to estimate divergence times at intra-specific levels using RelTime in MEGA v10.1.8 (90, 91). We then grafted all intra-specific level subclades into the time-calibrated species tree using the R package ape (92). To test for synchronous divergent events unchained by the Isthmus of Panama, we used *ecoevolity* (31), a full-likelihood Bayesian approach that tests models of co-divergence by integrating gene trees with the population histories of each species pair from genomic data (31). Finally, to examine the biogeographic history of the genus, we inferred ancestral geographic ranges on the time-calibrated species tree with the R package *BioGeoBEARS* (28). We assessed 12 biogeographic models under a maximum likelihood framework, implementing a seven-province biogeographic scheme (13, 29, 30) and time slices reflecting the final closure of the Isthmus of Panama. For additional information refer to *Appendix C, Materials and Methods*.

3.4 Results

Population genetic structure and contemporary breaks

Extended results are reported in the *Appendix C, Supplementary Results*. To identify genetic patterns corresponding to major contemporary breaks, we conducted population differentiation analyses on 14 co-distributed species. We employed a combination of methods including PCA, DAPC, phylogenetic trees, ADMIXTURE, fastStructure, EEMS, sPCA, and AMOVA. These analyses utilized matrices that ranged between 4,000 and 9,000 biallelic orthologous unlinked SNPs, with 3.7%–11.7% of missing data (see *Appendix C, Materials and Methods,* Tables S1–S2). They enabled us to not only pinpoint marine breaks hindering genetic flow across populations in the genus, but also their relative strength or effectiveness, as measured by the proportion of species affected by each break (Fig. 3.1).

In the TEP, major contemporary barriers corresponded to habitat gaps. The Sinaloan Break (SB) and Central American Break (CB)—two major portions of sandy and muddy environments interrupting the rocky outcroppings in the region—as well as the open ocean between Galápagos Archipelago and continental Ecuador (OOGB) (13), appear to influence connectivity at different levels across the evaluated species. The SB hinders population connectivity for all four assessed species: *M. mexicanus*, *M. tetranemus*, *M. zacae*, and the *M. hubbsi-M. polyporosus* species complex (Fig. 3.1*A*, *Appendix C,* Figs. S14–S15, S17–S18). Likewise, the CB acts as a strong barrier to dispersal with an effectiveness of 100%, splitting the population structure of *M. ebisui* and *M. tetranemus* (Fig. 3.1*A*, *Appendix C,* Figs. S13 and S17). The OOGB also influences *M. tetranemus*, however, only the DAPC's PC2 axis (12–13.3%) captured this particular pattern (*Appendix C,* Fig. S17 *E* and *F*). Our data also unveiled the Panama gyre (PGB), a marine break attributed to environmental conditions unchained by seasonal upwellings (15), disrupting connectivity among the populations of *M. sudensis* (Fig. 3.1*A*, *Appendix C,* Fig. S16). Additionally, we observed five instances of potential new breaks (PNB), though they were mostly idiosyncratic (Fig. 3.1*A*; see details *Appendix C, Supplementary Results,* Figs. S13, S16–17, and S19).

In the TA, the Eastern Caribbean Break (ECB) disrupted connectivity for all seven species assessed, emerging as the predominant barrier in the region. This was followed by the Bahamas Break (BB) and the Gulf of Mexico Break (GMB), which hindered dispersal in five and four out of the six species evaluated, respectively (Fig. 3.1*B*, *Appendix C,* Figs. S20–S25). The BB affected population connectivity in *M. erdmani*, *M. gilli*, *M. triangulatus*, *M. macropus* and *M. versicolor* (Fig. 3.1*B*, *Appendix C,* Figs. S22–26). Similarly, the GMB affected those species, except for *M. triangulatus*. In this region, we also observed instances where the evaluated marine breaks did not have an important effect on population structure. The Western Caribbean Break, which was evaluated for six species, seems to influence only the genetic connectivity among the populations of *M. erdmani*, *M. gilli*, and *M. triangulatus* (Fig. 3.1*B*, *Appendix C,* Figs. S22–S23 and S25). Notably, for *M. triangulatus*, we identified contrasting genetic structure patterns related to this barrier (see Discussion). The Yucatán Current Break (YCB) disrupted connectivity for *M. erdmani*, *M. gilli*, and *M. triangulatus* (Fig. 3.1*B*, *Appendix C,* Figs. S22–S23, S25) among the seven species evaluated. Two other significant breaks impacting *M. triangulatus* (the only species assessed for these breaks due to taxonomic sampling) were the Mona Passage Break (MPB) and the Northeastern Colombian Coast (NECC) (Fig. 3.1*B*, *Appendix C,* Figs. S25). In this region, we

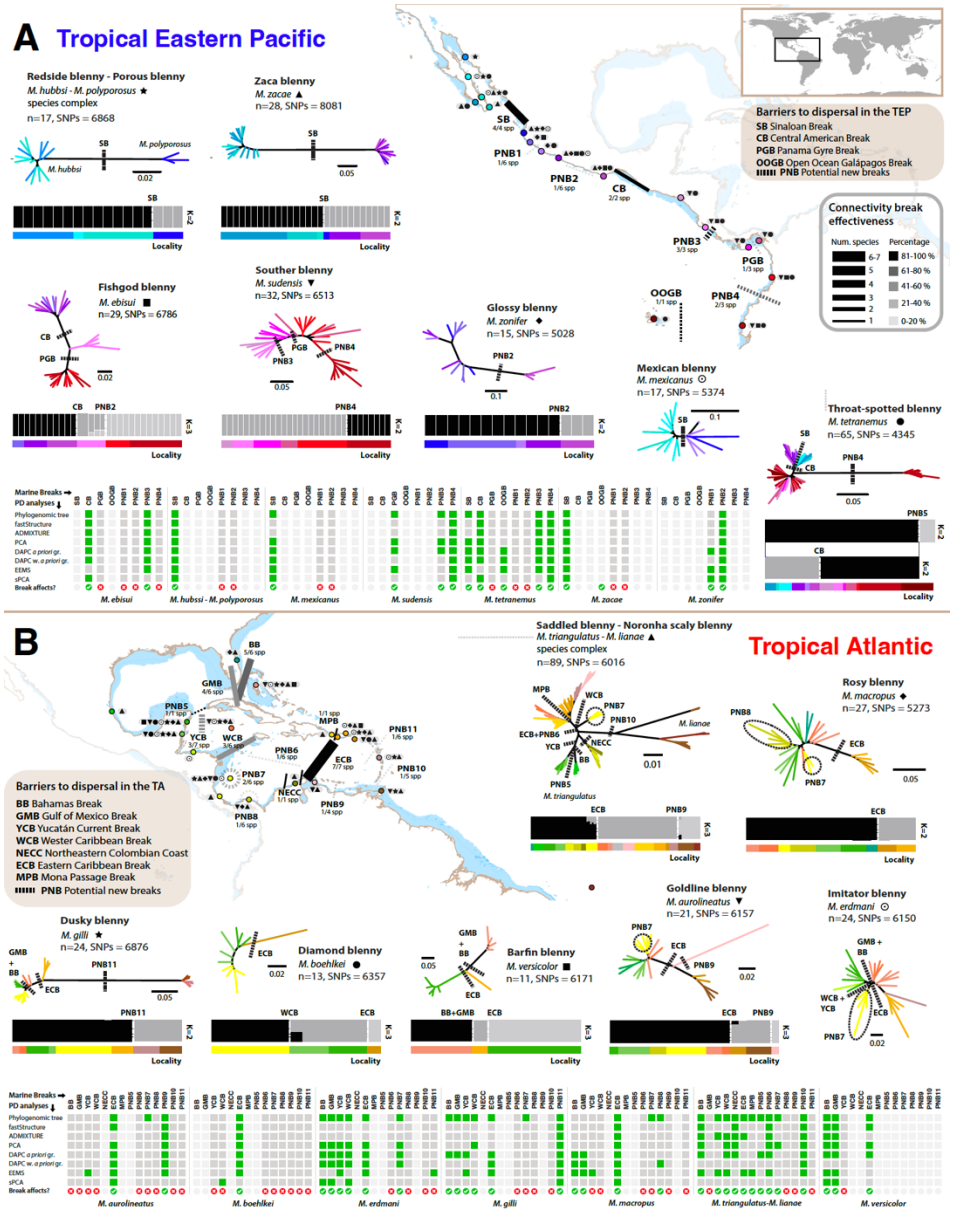


Figure 3.1 Population structure at the microevolutionary scales, represented by phylogenetic trees and fastStructure plots in A) the Tropical Eastern Pacific (TEP) and B) the Tropical Atlantic (TA). The branches of phylogenetic trees and the bars beneath fastStructure plots are color-coded according to sampling locations. Icons adjacent to sampling locations denote the species collected at each site. FastStructure clusters are rendered in grayscale to prevent confusion with sampling localities. Black lines on the maps depict the marine contemporary breaks hindering connectivity across the TEP and TA regions, with line thickness reflecting the number of species affected and color transparency their relative effectiveness. Matrices at the bottom of each panel represent population differentiation (PD) analyses that identified each break for each species. A species was considered affected by a break (denoted in green) when over two analyses identified population structure aligning to such barrier. Some barriers were untested for specific species due to taxonomic sampling or species distribution, marked in light grey on the matrices. Certain breaks, such as the ECB in the TA, and the CB in the TEP, have a strong effect on the population structure of the species. In contrast, others, such as the Western Caribbean Break, impact fewer co-distributed species. PD analyses further revealed potentially new breaks (PNB), mostly idiosyncratic, marked with dashed lines. For instance, PNB11 is evidenced by the population structure of the dusky blenny only.

also observed seven PNBs (Fig. 3.1*B*; see details *Appendix C, Supplementary Results,* Figs. S20, S22–S23, S25).

All in all, we find that in the TEP region, SB, CB, and OOGB marine breaks define boundaries among the Cortez, Mexican, Panamic, and Oceanic Islands biogeographic provinces. In the TA region, BB, GMB, and YCB delineate the border between the Northern and Central Provinces, whereas the PNB9 barrier, having a mere 0.20% effectiveness, is the sole recognized limit; no breaks are found between the Central and Southern provinces

Body-shape disparity across populations and species

To investigate whether morphological disparity aligns with population genetic structure analyses, we performed a geometric morphometric assessment on the most widespread species within each region: *M. tetranemus* in the TEP, and *M. triangulatus* in the TA. We digitized a total of 36 landmarks and semi-landmarks using high quality photographs. For *M. triangulatus*, we also digitized x-rays to maximize the number of sampled localities. Both photographs and x-rays were analyzed separately as the type of data influenced the results (*Appendix C,* Fig. S11). To assess the sensitivity of our landmarks to specimen "bending effects" (i.e., preservation artifacts), we built four alternative schemes: 1) all landmarks and semi-landmarks, 2) landmarks, 3) head-only, and 4) anterior body-only (*Appendix C,* Fig. S12). The results were slightly biased by the bending effect (*Appendix C,* Figs. S28–S30), hence we selected scheme 4 as our final dataset (Fig. 3.2*A*). *Malacoctenus tetranemus* did not display a clear geographic clustering pattern aligning with any marine break (Fig. 3.2*B*). However, within the *M. triangulatus* species complex, the PC2 axis (10.2%) captured differences in snout elongation that, despite extensive overlapping clusters, align with the ECB (Fig. 3.2*C*). Lastly, to probe the extent of morphological variation at the species level, we also conducted these analyses using photographs representing 20 species of *Malacoctenus*. However, species clusters mostly overlapped within the morphospace, offering limited evidence of interspecific differentiation in body shape disparity (Fig. 3.2*D*).

Environmental drivers of speciation and local adaptation

To evaluate the relative contribution of sea composition (ecological variables) and sea configuration (spatial features) to the genetic differentiation in both regions, we employed distance-based redundancy analyses (db-RDA). To this end, we calculated distance-based Moran's eigenvector maps (db-MEMs) using the geographic coordinates of the sampling sites (27). db-MEMs were used to describe spatial genetic structure, while seven environmental variables extracted from Marine Copernicus (http://marine.copernicus.eu/) online data archive were used to represent ecological features. At the microevolutionary level, we evaluated *M. tetranemus* and *M. triangulatus*, while at macroevolutionary scales we incorporated data from 20 species into our analyses. Our results suggest that the population structure of *M. tetranemus* is mainly explained by spatial distribution, temperature and different levels of chlorophyll α, while the population structure of *M. triangulatus* is strongly influenced by temperature, suspended particle matter, depth, salinity, and spatial composition. Both analyses suggest that the environmental features have a great influence (60.9% and 43.6% in *M. tetranemus* and *M. triangulatus*, respectively) on distribution and connectivity across populations in each species (Fig. 3.3 *B* and *E*). At

macroevolutionary scales, our analyses suggest that the complex environment in the TEP has played a major role in the diversification of the genus, where depth is the main environmental variable (Fig. 3.3*C*). Although depth is also an important feature in the TA, the spatial configuration, chlorophyll α , and suspended particle matter have also played an important role in speciation (Fig. 3.3*F*). These analyses suggest that environmental variables are major drivers of divergence in these regions (74.1% and 91.2% in the TEP and TA, respectively).

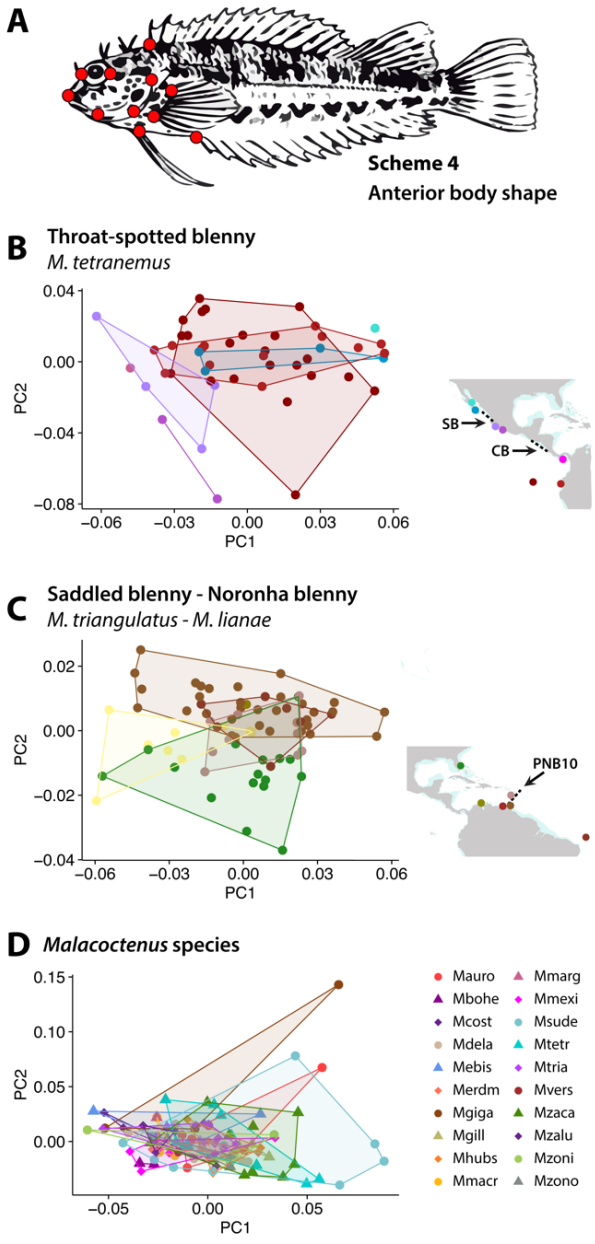


Figure 3.2 A) Landmarks scheme (depicted in red) used to conduct geometric morphometric analyses at microevolutionary scales for B) *M. tetranemus* in the TEP, and C) *M. triangulatus-M. lianae* species complex in the TA; and at macroevolutionary scales for D) 20 species of *Malacoctenus* in both regions.

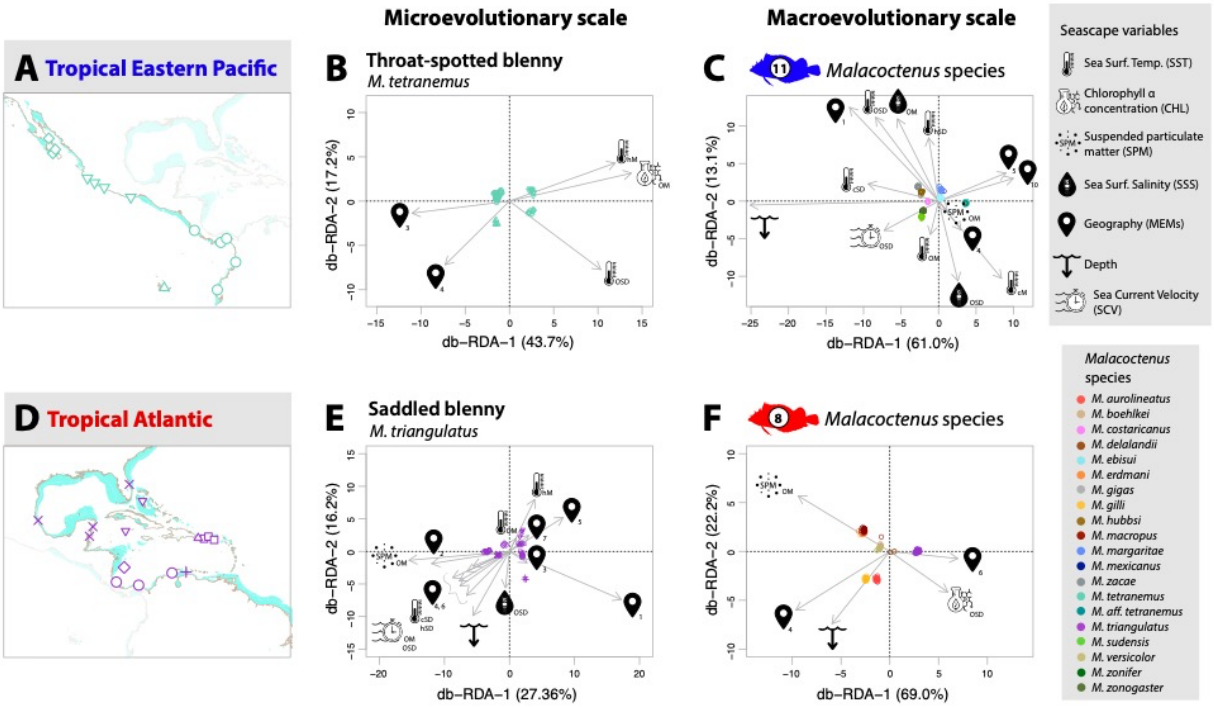


Figure 3.3 Seascape genomic analyses of *Malacoctenus*, illustrating main environmental features driving divergence at both (A and E) intraspecific and (C and F) interspecific levels in the Tropical Eastern Pacific (A–C) and Tropical Atlantic (D–F). Distance-based partial redundancy analysis (db-RDA) show that SST, SSS, SPM, spatial composition (MEMs), and depth, are key environmental variables driving population differentiation and speciation.

Evolutionary and biogeographic history, and the rising of the Panama Isthmus

For macroevolutionary inferences, we generated five matrices consisting of 1,891–28,144 SNPs, selected based on varied levels of missing data (29.1–81.1%), which we used to elucidate the evolutionary and biogeographic history of the genus, with an emphasis on examining the synchronicity of speciation events triggered by the rise of the Isthmus of Panama (*Appendix C, Materials and Methods,* Table S7). To this end, we estimated the evolutionary relationships among the 23 species in the genus, and time-calibrated the 28K SNP tree (Fig. 3.4, *Appendix C,* Fig. S35–S36). We then inferred ancestral geographic ranges by using BioGeoBEARS (28). We fed this analysis with the current distribution of each species across seven biogeographic provinces (13, 29, 30) in the TEP and TA regions.

These analyses place the origin of the genus around 30 Mya in Africa, where *M. carrowi* appears as the earliest-branching lineage (Fig. 3.4*A*). The best-supported biogeographic model for the genus based on seven areas was BayAREA + ω (AICc=189.06; Fig. 3.4). Our analyses suggest that there was an East to West dispersal event, from Africa to an ancient area that included the Greater Caribbean and the Panamic province (event 1; Fig. 3.4*C*). Approximately *ca.* 22 Mya clade-A became established in the Greater Caribbean region. Around *ca.* 18 Mya the genus dispersed from an ancestral area comprising the Panamic and Central provinces to the north in the Pacific (event 2; Fig. 3.4*C*). Clade B1 emerged within the Cortez-Mexican provinces, eventually

leading to the origin of *M. zonogaster* via dispersal into the Galapagos Archipelago (event 3; Fig. 3.4*C*). Clade B2 includes diversification events originated by the final closure of the Isthmus of Panama giving rise to a geminate species—i.e., sister species pairs on each side of the isthmus—, and the trans-isthmic clade—sister lineages on both sides of the Isthmus of Panama. Just before the final closure of the Isthmus, the MRCA of the trans-isthmic clade colonized the Tropical Southwestern Atlantic around 3 Mya (event 4; Fig. 3.4*C*). We used *ecoevolity* (31) to test whether the cladogenetic events between the transisthmic clades occurred synchronously or independently as a result of the rise of the Panama Isthmus, finding that the speciation events triggered by this historical barrier were not temporally synchronic (Fig. 3.4 *A* and *B*). Based on the divergence time estimates, it seems that these speciation events occurred around 5.2 and 7.6 Mya, prior to the final closure of the isthmus of Panama, which is estimated to have occurred between 2.8–3.2 Mya (10).

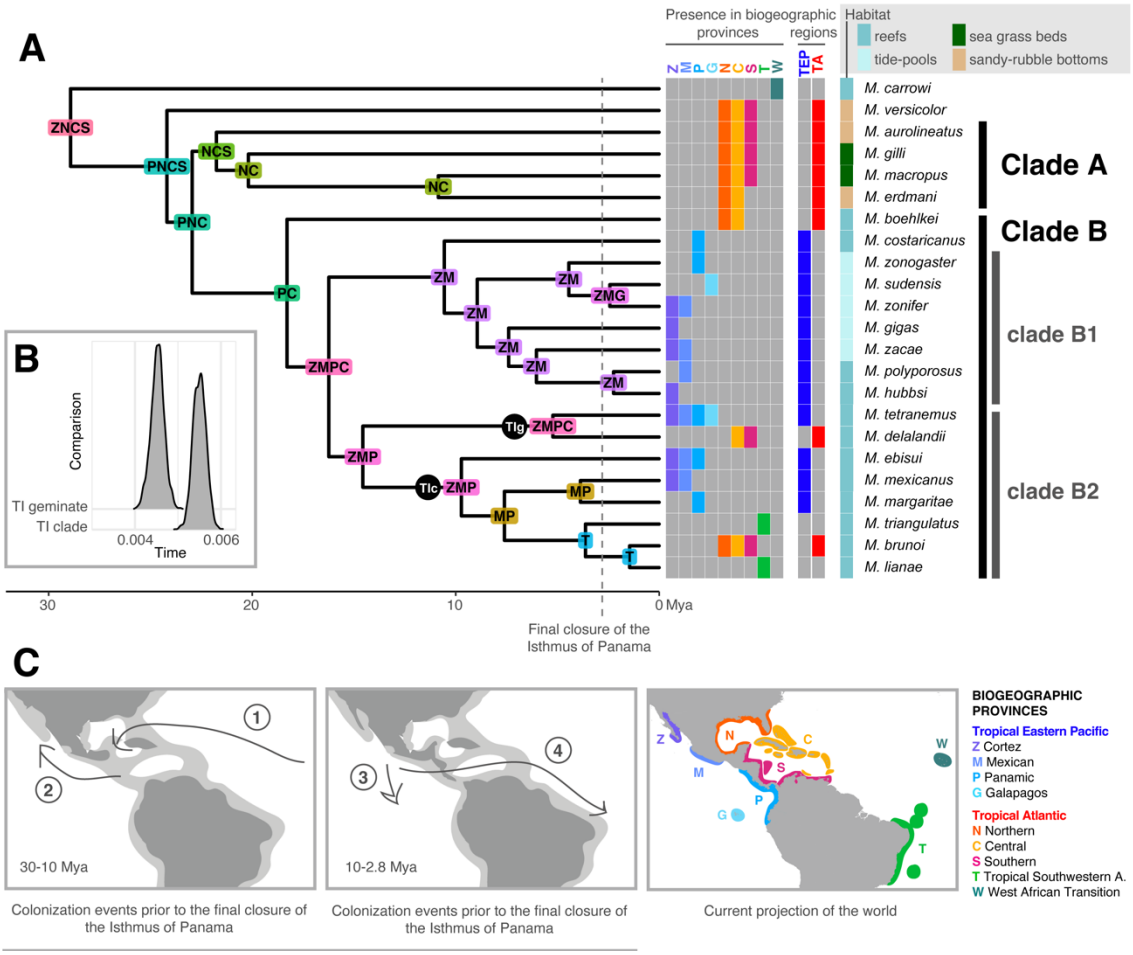


Figure 3.4. A) A time-calibrated species tree, based on 28K SNPs, illustrates the ancestral range reconstruction across the biogeographical provinces of the Tropical Eastern Pacific (TEP) and Tropical Atlantic (TA). Habitat preferences are also illustrated next to species presence across biogeographic provinces and realms. B) The *ecoevolity* results indicate that the two speciation events—the transisthmic clade and transisthmic geminate species-pair—occurred asynchronously. C) Inferred paths of dispersal and colonization events within the genus. D) Biogeographic provinces utilized to reconstruct ancestral range distribution (13, 29, 30).

3.5 Discussion

By integrating genome-wide, morphologic, and ecological data at both micro- and macroevolutionary scales, our study was able to assess the effects of previously proposed contemporary and historical marine breaks on the diversification and speciation of labrisomid blennies in the genus *Malacoctenus*. We detected consistent patterns of genome-wide signatures of genetic structure on codistributed species across the TEP and TA biogeographic regions, which allowed us to assess the effectiveness of each contemporary break evaluated and uncover potential new ones. Our seascape genomic analyses unveiled a complex interplay between seascape composition and seascape configuration features on shaping the distribution of species across both regions. When addressing intraspecies dynamics, major environmental variables influencing local adaptation included temperature, chlorophyll α levels, and physical oceanographic components. At the interspecies level, additional factors identified as speciation drivers included depth and suspended particle matter levels. Furthermore, our time-calibrated analyses at macroevolutionary scales revealed that the sole African species constitutes the earliest-branching genus-level lineage, indicating an east-to-west dispersal of the clade into the Neotropics followed by the origin of asynchronous transisthmian clades.

In the TEP, when considering phylogeographic breaks and provinces, they are mostly arranged in a linear fashion, following the geomorphological configuration of the coastline, although there are exceptions like oceanic islands, such as the Galápagos Archipelago. In contrast, in the TA, there is a greater two-dimensionality that largely stems from the semi-enclosed configuration of the Greater Caribbean. While environmental clusters of sampling locations in the TEP also coincide with the delimitation of biogeographic provinces, this is not the case in the TA (Figs. S31–S32), suggesting a more complex influence of microhabitats and environmental features in the area. Notably, major phylogeographic breaks in the TA do not align with established biogeographic provinces (17), emphasizing the need for broader comparative phylogeographic studies in this region.

In general, major marine barriers that exhibited high effectiveness values are mostly related to habitat gaps, physical oceanographic features and processes (e.g., currents), marked environmental changes (e.g., temperature), or a combination of these. The Sinaloan (SB) and Central American (CB) breaks in the TEP are recognized as habitat gaps for labrisomid blennies, primarily attributed to the absence of significant rocky outcroppings (12, 13). Our study revealed that both the SB and the CB exhibit a 100% effectiveness in influencing the population structure of labrisomid blennies (Fig. 3.1). While the SB has traditionally been considered to primarily impact species with limited dispersal capabilities (29, 32, 33), due to its relatively modest extent compared to the CB (approximately 300 km vs. 1000 km, respectively (13)), our findings challenge this perspective. Here, we identified four distinct instances of population structure aligned with the SB (Fig. 3.1, Figs. S16–17, 19–20). Notably, these results highlight that the SB may play a more significant role than previously assumed in shaping connectivity patterns in the region. For example, the independent evolutionary lineages observed in the zaca blenny on either side of the SB, along with previous meristic differences detected across those populations (12), suggest that they could potentially represent new species. The genomic divergences (Fst = 0.14) observed

among these lineages are higher as those among recognized sister species separated by the SB, such as the species complex including the redside and porous blennies (*Fst* = 0.11) (Fig. 3.1A, Table 10). Population genetic patterns consistent with these habitat gaps have been observed in several cryptobenthic fish species. For instance, chaenopsids (34) exhibit such patterns in the SB, while in the CB similar trends are evident in clingfishes (24, 25), gobies (15), other conspicuous fish species (14), as well as snails (16), oysters (32), and barnacles (35).

The Panama gyre break (PGB) in the Gulf of Panama (TEP) has been proposed as a mechanism for transporting larvae offshore, reducing population connectivity between Las Perlas Archipelago and western Panama (36). This gyre, in combination with seasonal upwelling in the area, triggers a series of environmental changes, including significantly lower temperatures, decreased oxygen levels, and changes in pH, along with an increased nutrient availability (37). These environmental factors can collectively form physical and environmental barriers for cryptobenthic fishes (15). Our results have revealed evidence of population structure in the fishgod and southern blennies (Figs. S13 and S17), which could be attributed to this break. In addition, the open ocean Galápagos break (OOGB) serves as a physical barrier in the region (*ca.* 1000 km from the mainland). Studies on TEP faunas have shown that environmental conditions in oceanic islands differ from those in the mainland (18, 38). We recovered patterns of population structure consistent with this break in the throat-spotted blenny, the most widely distributed species, which inhabits both the mainland and the archipelago (Fig. S17).

In the TA, our study identified multiple marine breaks hindering population connectivity. Major barriers included the Eastern Caribbean (ECB) and the Bahamas (BB) breaks, which show 100% and 83% effectiveness. The ECB, and the western Caribbean break (WCB), with a 50% effectiveness, were proposed based on high-resolution biophysical models and simulations of typical larval dispersal distances of reef fish species (10–100km) (22). This suggest that the oceanographic processes in the region cause these barriers. In contrast, the BB is attributed to different environmental conditions between insular and coastal shelves and the flow of the Gulf stream limiting reef fish dispersal (20, 39, 40). West of the BB, the Gulf of Mexico Break (GMB), with a 67% effectiveness, is also attributed to ecological differences (41). In addition, the Yucatán Current Break (YCB), although less than 50% effective, is influenced by the northward flow of the Caribbean Current, which causes cyclonic gyres and mesoscale eddies that enhance larval retention (42, 43). Our study also highlights Mona Passage (MPB) and northeastern Colombian coast (NECC) breaks as significant barriers for labrisomid blennies. While we only assessed populations of the saddled blenny due to taxonomic sampling limitations, it is likely that both breaks strongly impact other cryptobenthic fishes. As shown previously, strong currents across the deep-sea Mona passage may pose a physical obstacle that some reef fishes cannot overcome (3, 44, 45). On the other hand, previous studies suggest three major components affecting the NECC: the outflow of the Magdalena river, the absence of rocky bottoms east of the Santa Marta massif, and a seasonal upwelling near La Guajira, Colombia (20, 46–48). Although large river deltas are considered habitat breaks for labrisomid blennies (12), the observed clustering patterns in our study do not support the Magdalena River plume hindering population dispersal (Fig. S25). While the absence of rocky bottoms may be a contributing factor, the Guajira upwelling is known to play a significant role triggering speciation in chaenopsids (48).

Our analyses also identified 11 potential new breaks (PNB1–PNB4 in the TEP; PNB5–PNB11 in the TA). However, except for PNB3, which affected three species, and PNB4 and PNB7, each of which affected two species, these breaks were primarily idiosyncratic, impacting in most cases a single species. It is particularly interesting that five of these putative barriers seem to be promoting the generation of independent evolutionary lineages (Figs. S17, S19–S20, S23, S25), suggesting the presence of putative species complexes with high genomic divergences (*Fst* = 0.06–0.19, Table 10). While sampling localities separated by PNB2 and PNB4 exhibit unique environmental features (Fig. S32) in the TEP, PNB5–PNB11 in the TA seem to be linked to environmental shifts influenced by seasonal upwellings, the direction and intensity of predominant oceanographic currents, and gyres (49–51). In this scenario, while our ecological hypothesis suggests that mayor connectivity patterns arise from large-scale environmental variation in the TA, here we observe that several marine breaks reflect a limited gene flow among populations driven by localized environmental variation (e.g. upwellings).

Zamudio et al. (2) recently emphasized the importance of incorporating phenotypic variation and genetic data into phylogeographic studies to gain a deeper understanding of the origins of biodiversity. In this context, a closer examination of phenotypes can help elucidate the mechanisms influencing responses of co-distributed species to environmental variations. In this study, our geometric morphometric analyses revealed varying degrees of congruence with genomic divergences. While they failed to detect morphologic clusters that aligned with the genomic groups in the throat-spotted blenny in the TEP (Fig. 3.2B), they showed a partial geographic clustering in correspondence with the PNB10 in the TA, separating the saddled and Noronha scaly blennies (Fig. 3.2C). Moreover, when these analyses were conducted across 20 species, there was a lack of species-level clustering in morphospace (Fig. 3.2D). Undetectable geographic variation in phenotypes in the presence of phylogeographic structure is generally attributed to biological processes such as stabilizing selection or cryptic diversification (2). Cryptobenthic reef fishes often lack discernible diagnostic features, as many sympatric species exhibit morphological overlap (12, 52). Consequently, the absence of clear morphological differences within populations of the throat-spotted blenny (Figs. 3.4B and S28) as well as between some species (Figs. 3.4D and S30) may be indicative of cryptic speciation, possibly driven by niche conservatism (53–56). In contrast, the presence of both phylogeographic patterns and slightly detectable phenotypic variation between the saddled and the Noronha scaly blennies (Figs. 3.4C and S29) may arise from biological mechanisms like environmental local adaptation or divergent sexual selection (2). As expected, these analyses provide less resolution than our genomic data, highlighting the limitations of integrating phenotypes into phylogeographic studies.

Our seascape genomic analyses of the throat-spotted blenny align with ecological hypotheses confirming that biogeographical provinces in the TEP exhibit unique seascape compositions (Fig. 3.3A–B). Specifically, the Cortez province undergoes significant fluctuations in sea surface temperatures across the year. In turn, the Mexican province presents variations in average chlorophyll α values and temperature during the hottest months of the year. These data suggest that populations on either side of the SB barrier have undergone local adaptation, exhibiting different ecological requirements. Noteworthy, changes in temperature and chlorophyll α levels are considered to play a pivotal role in ecological diversification within marine ecosystems

(57). In contrast, seascape configuration features were responsible for separating populations across the Panamic and Galápagos provinces. At the interspecific level, depth was the strongest variable promoting speciation in this region (Fig. 3.3C). These results also suggest that labrisomid blennies are habitat specific, represented mainly by tide-pool or reef-associated species (Fig. 3.4). In this scenario, diversification processes may be related to intertidal species having a higher resilience to environmental changes as they are subjected to dramatic fluctuations in temperature, salinity and dissolved oxygen levels during daytime (58).

Genotype association analyses conducted at the intraspecific level on the saddled blenny (Fig. 3.3D–E) differentiate the populations of the Gulf of Mexico based on variations in temperature during the hottest and coldest months of the year. This observation aligns with previous studies suggesting that cooler temperatures during winter affect fish distributions in this region (17). These locations are also affected by factors such as sea current velocity and sea configuration features, while areas along the coasts of Central America and northern South America display distinct levels of suspended particulate matter. The remaining populations are primarily segregated by variations in temperature and sea configuration characteristics. At interspecific levels, major environmental factors promoting speciation include depth, suspended particle matter, chlorophyll α levels, and seascape physical features (Fig. 3.3F). Overall, our seascape genomic analyses suggest that environmental associations in the TA are more complex than those in the TEP. This aligns with the observation that species distributions in the TEP are primarily allopatric, whereas in the TA, there is a significant overlap in species distributions, indicating a more dynamic interplay of environmental factors influencing diversification across large areas in the region. In the TEP, habitat types are predominantly dominated by reef or tide-pool environments (13, 18). However, in the TA, several labrisomid blennies are also associated to rubble-sandy bottoms or sea grass beds (12, 17).

At the macroevolutionary level, our ancestral range reconstruction analyses place the origin of the genus in the eastern Atlantic around 30 Mya (event 1; Fig. 3.4C). The transatlantic dispersal occurred from Cabo Verde, Africa, following an east-to-west trajectory probably linked to the flow of oceanic currents (12). While a higher species richness in the Americas might suggest a center of origin in this hemisphere (12), our analyses invariably place the African species as the oldest lineage in the genus, supporting an east-to-west dispersal. This out-of-Africa pattern of colonization of the Caribbean has also been observed in other blennies (54, 59), as well as other reef fishes such as porgies (60). Approximately *ca.* 20–18 Mya, clade-A colonized the Greater Caribbean region, initially occupying an ancestral area that consisted of the Panamic and Central provinces, before expanding northwards in the Pacific. A similar, albeit more recent, dispersal pattern has been observed in the redhead goby (15). Clade B1 settled within the Cortez-Mexican provinces, later dispersing into the Galapagos Archipelago (event 3; Fig. 3.4C), while Clade B2 experienced cladogenetic events triggered by the rise of the Isthmus of Panama, resulting in sister lineages on both sides of the Isthmus (Fig. 3.4A), represented by a geminate species pair (TIg) and a transisthmian clade (TIc). The formation of the isthmus is considered one of the most significant geological events, acting as a historical marine barrier that unchained species diversification strongly impacting marine organisms (61). The emergence of the land was a gradual process, facilitating the connection and disconnection of the Pacific and Atlantic oceans that culminated

with the final rise of the Isthmus of Panama around 2.8 Mya (62). Teleost fishes experienced a series of transisthmian splits between 1–23 Mya, peaking around the final stages of Isthmus completion *ca.* 5 Mya (62–64). Assumptions regarding the timing of split divergences across marine animal groups suggest that the life history parameters play a pivotal role. For example, high intertidal geminate pairs tend to show less genetic differentiation than lower intertidal ones (65). Herein, our study recovered strong evidence supporting a non-synchronous divergence between the TIc and TIg lineages, even though they possess similar life history traits (Fig. 3.4 A–B). Lastly, before the final closure of the Isthmus, the most recent common ancestor of the transisthmian clade colonized the Tropical Southwestern Atlantic around 3 Mya (event 4; Fig. 3.4C). This event is supported by other studies on blennies, which suggest a greater connectivity between the Caribbean and eastern Atlantic than between the Caribbean and the coasts of Brazil (54). Finally, our macroevolutionary analyses also show two major niche divergences in the genus (Fig. 3.4A). The first occurred in clade A, when *Malacoctenus* lineages colonized the Caribbean, encountering a wide variety of habitats. The second was observed in subclade B1, when the ancestral lineage and its daughter species became restricted to tide-pool environments, triggering niche conservatism (66, 67).

3.6 Conclusion

Our study recovered genomic patterns of population structure in co-distributed species, demonstrating the significant impact of marine breaks on the populations of cryptobenthic fishes in the TEP and TA. In these biogeographic regions, major contemporary breaks result from unsuitable habitat gaps combined with oceanographic processes, including marine currents, seasonal upwellings, and gyres, followed by marked environmental changes. In the TEP, the SB and CB emerge as major barriers, promoting not only population diversification but also speciation processes, representing the distribution limit for several labrisomid blennies. In the TA, the ECB and the BB play a crucial role in shaping population connectivity. The discrepancy between phylogeographic breaks in cryptobenthic fishes and biogeographic provinces highlights the need for further comparative phylogeographic studies, as cryptic species complexes may have been overlooked in the assessment of biogeographic patterns. Overall, the effect of these breaks varies across species, suggesting that species-specific traits, such as habitat preference, also greatly influence their dispersal capabilities. Our study identified five instances where marine breaks led to highly differentiated evolutionary lineages that could potentially represent species complexes. Some of these genetically distinct groups are supported by evidence of population differentiation from previous morphological studies as well as by our geometric morphometric analyses. Our seascape genomic analyses highlighted that temperature, chlorophyll α levels, and physical features played pivotal roles in driving local adaptation of the throat-spotted blenny in the TEP. However, at interspecific levels, depth emerged as the primary driver of speciation within this region, leading to niche divergence between tide pool- and reef-associated clades. In contrast, in the TA, patterns of environmental association appeared more intricate, with suspended particle matter, depth, temperature, and physical features significantly influencing population differentiation, with chlorophyll α further contributing to speciation in this region. Finally, our

time-calibrated analyses at macroevolutionary scales elucidated an Eastern Atlantic origin of the clade followed by an east-to-west dispersal. Although the historical break attributed to the rise of the Isthmus of Panama had a substantial influence on the evolutionary history of the genus, our analyses demonstrate that it did not triggered synchronous cladogenetic events. Altogether, through the integration of approaches from population genomics, comparative phylogeography, phylogenomics, seascape genomics, and geometric morphometric analyses, our study has identified the primary contemporary and historical drivers of lineage diversification and speciation in labrisomid blennies in the TEP and TA, encompassing a spectrum from micro- to macroevolution.

3.7 References

1. J. Li, J. P. Huang, J. Sukumaran, L. L. Knowles, Microevolutionary processes impact macroevolutionary patterns. *BMC Evol. Biol.* **18**, 1–8 (2018).
2. K. R. Zamudio, R. C. Bell, N. A. Mason, Phenotypes in phylogeography: Species' traits, environmental variation, and vertebrate diversification. *Proc. Natl. Acad. Sci. U. S. A.* **113**, 8041–8 (2016).
3. M. S. Taylor, M. E. Hellberg, Comparative phylogeography in a genus of coral reef fishes: Biogeographic and genetic concordance in the Caribbean. *Mol. Ecol.* **15**, 695–707 (2006).
4. M. E. Hellberg, R. S. Burton, J. E. Neigel, S. R. Palumbi, Genetic Assessment Of Connectivity Among Marine Populations. *Bull. Mar. Sci.* **70**, 273–290 (2002).
5. S. R. Palumbi, What can molecular genetics contribute to marine biogeography ? An urchin ' s tale. **203**, 75–92 (1996).
6. C. Riginos, B. C. Victor, Larval spatial distributions and other early life-history characteristics predict genetic differentiation in eastern Pacific blennioid fishes. *Proc. R. Soc. B Biol. Sci.* **268**, 1931–1936 (2001).
7. C. Riginos, K. E. Douglas, Y. Jin, D. F. Shanahan, E. A. Treml, Effects of geography and life history traits on genetic differentiation in benthic marine fishes (2011) https:/doi.org/10.1111/j.1600-0587.2010.06511.x.
8. P. N. Palmerín-Serrano, *et al.*, Comparative phylogeography and demographic histories of five widely distributed tropical eastern Pacific fishes. *Mar. Biol.* **170** (2023).
9. V. Mendoza-Portillo, *et al.*, Genetic diversity and structure of circumtropical almaco jack, Seriola rivoliana: tool for conservation and management. *J. Fish Biol.* **97**, 882–894 (2020).
10. A. O'Dea, *et al.*, Formation of the Isthmus of Panama. *Sci. Adv.* **2**, 1–12 (2016).
11. D. S. Jordan, The Law of Geminate Species. *Am. Nat.* **42**, 73–80 (1908).

12. V. G. Springer, “Systematics and Zoogeography of the Clinid Fishes of the Subtribe Labrisomini Hubbs,” University of Texas. (1959).

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

14. A. F. Mar-Silva, *et al.*, Genomic assessment reveals signal of adaptive selection in populations of the Spotted rose snapper Lutjanus guttatus from the Tropical Eastern Pacific. *PeerJ* **11**, 1–30 (2023).

15. E. R. Sandoval-Huerta, *et al.*, The evolutionary history of the goby Elacatinus puncticulatus in the tropical eastern pacific: Effects of habitat discontinuities and local environmental variability. *Mol. Phylogenet. Evol.* **130**, 269–285 (2019).

16. L. A. Hurtado, M. Frey, P. Gaube, E. Pfeiler, T. A. Markow, Geographical subdivision, demographic history and gene flow in two sympatric species of intertidal snails, Nerita scabricosta and Nerita funiculata, from the tropical eastern Pacific. *Mar. Biol.* **151**, 1863–1873 (2007).

17. D. R. Robertson, K. L. Cramer, Defining and Dividing the Greater Caribbean : Insights from the Biogeography of Shorefishes. *PLoS One* **9** (2014).

18. P. April, O. Pen, D. R. Robertson, K. L. Cramer, Shore fishes and biogeographic subdivisions of the Tropical Eastern Pacific. **380**, 1–17 (2009).

19. R. I. Eytan, M. E. Hellberg, Nuclear and mitochondrial sequence data reveal and conceal different demographic histories and population genetic processes in caribbean reef fishes. *Evolution (N. Y).* **64**, 3380–3397 (2010).

20. R. Betancur-r, *et al.*, Reconstructing the lionfish invasion : insights into Greater Caribbean biogeography. *J. Biogeogr.* **38**, 1281–1293 (2011).

21. R. K. Cowen, S. Sponaugle, Larval Dispersal and Marine Population Connectivity. *Ann. Rev. Mar. Sci.* **1**, 443–466 (2009).

22. R. K. Cowen, Scaling of Connectivity in Marine Populations. *Science (80-.).* **311**, 522–527 (2006).

23. G. Bernardi, Baja California disjunctions and phylogeographic patterns in sympatric California blennies. *Front. Ecol. Evol.* **2**, 1–9 (2014).

24. E. Torres-Hernández, *et al.*, Phylogeography and evolutionary history of the Panamic Clingfish Gobiesox adustus in the Tropical Eastern Pacific. *Mol. Phylogenet. Evol.* **173** (2022).

25. E. Torres-Hernández, *et al.*, A multi-locus approach to elucidating the evolutionary history of the clingfish Tomicodon petersii (Gobiesocidae) in the Tropical Eastern Pacific. *Mol. Phylogenet. Evol.* **166** (2022).

26. B. C. Victor, “How many coral reef fish species are there? Cryptic diversity and the new molecular taxonomy” in *Ecology of Fishes on Coral Reefs: The Functioning of an Ecosystem in a Changing World. Cambridge University Press, Cambridge, United Kingdom*, (2015), pp. 76–87.

27. P. R. Peres-Neto, P. Legendre, S. Dray, D. Borcard, Variation partitioning of species data matrices: Estimation and comparison of fractions. *Ecology* **87**, 2614–2625 (2006).

28. N. J. Matzke, nmatzke/BioGeoBEARS: BioGeoBEARS: BioGeography with Bayesian (and likelihood) Evolutionary Analysis with R Scripts (2018) https:/doi.org/10.5281/zenodo.1463216.

29. D. R. Robertson, K. L. Cramer, Defining and dividing the Greater Caribbean: Insights from the biogeography of shorefishes. *PLoS One* **9** (2014).

30. M. D. Spalding, *et al.*, Marine Ecoregions of the World : A Bioregionalization of Coastal and Shelf Areas. **57**, 573–583 (2007).

31. J. R. Oaks, Full Bayesian comparative phylogeography from genomic data. *Syst. Biol.* **68**, 371–395 (2019).

32. S. Arnaud, M. Monteforte, N. Galtier, F. Bonhomme, F. Blanc, Population structure and genetic variability of pearl oyster Pinctada mazatlanica along Pacific coasts from Mexico to Panama. *Conserv. Genet.* **1**, 299–308 (2000).

33. C. Mora, D. R. Robertson, Factors shaping the range-size frequency distribution of the endemic fish fauna of the Tropical Eastern Pacific. *J. Biogeogr.*, 277–286 (2005).

34. H. Lin, G. R. Gallan, Molecular analysis of Acanthemblemaria macrospilus (Teleostei: Chaenopsidae) with description of a new species from the Gulf of California, Mexico. *Zootaxa*, 51–62 (2010).

35. F. B. Pitombo, R. Burton, Systematics and biogeography of Tropical Eastern Pacific Chthamalus with descriptions of two new species (Cirripedia, Thoracica). *Zootaxa*, 1–30 (2007).

36. E. Rodríguez-Rubio, W. Schneider, R. A. del Río, On the seasonal circulation within the Panama Bight derived from satellite observations of wind, altimetry and sea surface temperature. *Geophys. Res. Lett.* **30** (2003).

37. L. D’Croz, D. R. Robertson, Coastal oceanographic conditions affecting coral reefs on both sides of the isthmus of Panama in *Proc. 8th Int. Coral Reef Symp.*, (1997), pp. 2053–2058.

38. E. A. Acevedo-Álvarez, G. Ruiz-Campos, O. Domínguez-Domínguez, Population-level morphological variation of Anisotremus interruptus (Gill, 1862) (Perciformes: Haemulidae) in the Tropical Eastern Pacific, with the description of two new species. *Zootaxa* **4975** (2021).

39. J. L. Carlin, D. R. Robertson, B. W. Bowen, Ancient divergences and recent connections in two tropical Atlantic reef fishes Epinephelus adscensionis and Rypticus saponaceous (Percoidei : Serranidae). 1057–1069 (2003).

40. D. W. Freshwater, R. M. Hamner, S. Parham, A. E. Wilbur, Molecular evidence that the lionfishes Pterois miles and Pterois volitans are distinct species. *J. North Carolina Acad. Sci.* **125**, 39–46 (2009).

41. K. B. Mobley, C. M. Small, N. K. Jue, A. G. Jones, Population structure of the dusky pipefish (Syngnathus floridae) from the Atlantic and Gulf of Mexico, as revealed by mitochondrial DNA and microsatellite analyses. *J. Biogeogr.* **37**, 1363–1377 (2010).

42. A. M. Jackson, *et al.*, Population structure and phylogeography in Nassau grouper (epinephelus striatus), a mass-aggregating marine fish. *PLoS One* **9** (2014).

43. M. J. Shulman, E. Bermingham, Early life histories, ocean currents, and the population genetics of Caribbean reef fishes. *Evolution (N. Y).* **49**, 897–910 (1995).

44. G. D. Dennis, D. A. Hensley, P. L. Colin, J. J. Kimmel, New Records of Marine Fishes from the Puerto Rican Plateau. *Caribb. J. Sci.*, 70–87 (2004).

45. P. L. Colin, Larvae Retention: Genes or Oceanography? *Science (80-.).* **300**, 1657–1659 (2003).

46. R. Betancur-R, P. Arturo Acero, H. Duque-Caro, S. R. Santos, Phylogenetic and morphologic analyses of a coastal fish reveals a marine biogeographic break of terrestrial origin in the Southern Caribbean. *PLoS One* **5**, 1–10 (2010).

47. J. C. Narváez-Barandica, *et al.*, A Comparative Phylogeography of Three Marine Species with Different PLD Modes Reveals Two Genetic Breaks across the Southern Caribbean Sea. *Animals* **13** (2023).

48. P. A. Hastings, R. I. Eytan, A. P. Summers, Acanthemblemaria aceroi, a new species of tube blenny from the Caribbean coast of South America with notes on Acanthemblemaria johnsoni (Teleostei: Chaenopsidae). *Zootaxa* **4816**, 209–216 (2020).

49. E. Díaz-Ferguson, R. A. Haney, J. P. Wares, B. R. Silliman, Genetic structure and connectivity patterns of two Caribbean rocky-intertidal gastropods. *J. Molluscan Stud.* **78**, 112–118 (2012).

50. N. K. Truelove, *et al.*, Biophysical connectivity explains population genetic structure in a highly dispersive marine species. *Coral Reefs* **36**, 233–244 (2017).

51. M. Merino, Upwelling on the Yucatan Shelf: Hydrographic evidence. *J. Mar. Syst.* **13**, 101–121 (1997).

52. S. J. Brandl, C. H. R. Goatley, D. R. Bellwood, L. Tornabene, The hidden half: Ecology and evolution of cryptobenthic fishes on coral reefs. *Biol. Rev.* (2018) https:/doi.org/10.1111/brv.12423.

53. V. G. Springer, M. F. Gomon, Variation in the Western Atlantic Clinid Fish Malacoctenus triangulatus with a Revised Key to the Atlantic Species of Malacoctenus. *Smithson. Contrib. to Zool.*, 1–11 (1975).

54. G. S. Araujo, *et al.*, Phylogeny of the comb-tooth blenny genus Scartella (Blenniiformes: Blenniidae) reveals several cryptic lineages and a trans-Atlantic relationship. *Zool. J. Linn. Soc.* **190**, 54–64 (2020).

55. E. A. Hadly, P. A. Spaeth, C. Li, Niche conservatism above the species level. *Proc. Natl. Acad. Sci. U. S. A.* **106**, 19707–19714 (2009).

56. J. J. Wiens, C. H. Graham, Niche conservatism: Integrating evolution, ecology, and conservation biology. *Annu. Rev. Ecol. Evol. Syst.* **36**, 519–539 (2005).

57. A. R. Amaral, *et al.*, Seascape genetics of a globally distributed, highly mobile marine mammal: The short-beaked common dolphin (genus delphinus). *PLoS One* **7** (2012).

58. R. M. Macieira, T. Simon, C. R. Pimentel, J. Joyeux, Isolation and speciation of tidepool fishes as a consequence of Quaternary sea-level fluctuations (2014) https:/doi.org/10.1007/s10641-014-0269-0.

59. J. E. Carter, M. A. Sporre, R. I. Eytan, Phylogenetic review of the comb-tooth blenny genus Hypleurochilus in the northwest Atlantic and Gulf of Mexico. *Mol. Phylogenet. Evol.* **189** (2023).

60. M. Summerer, R. Hanel, C. Sturmbauer, Mitochondrial phylogeny and biogeographic affinities of sea breams of the genus Diplodus (Sparidae). *J. Fish Biol.* **59**, 1638–1652 (2001).

61. H. A. Lessios, The Great American Schism : Divergence of Marine Organisms After the Rise of the Central American Isthmus. *Annu. Rev. Ecol. Evol. Syst.* (2008) https:/doi.org/10.1146/annurev.ecolsys.38.091206.095815.

62. A. O'Dea, *et al.*, Formation of the Isthmus of Panama. *Sci. Adv.* **2**, 1–11 (2016).

63. H. A. Lessios, Appearance of an early closure of the Isthmus of Panama is the product of biased inclusion of data in the metaanalysis. *Proc. Natl. Acad. Sci. U. S. A.* **112**, E5765 (2015).

64. D. Silvestro, *et al.*, Biological evidence supports an early and complex emergence of the Isthmus of Panama. *Proc. Natl. Acad. Sci.* **112**, E3153–E3153 (2015).

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

66. J. J. Wiens, *et al.*, Niche conservatism as an emerging principle in ecology and conservation biology. *Ecol. Lett.* **13**, 1310–1324 (2010).

67. S. Vaissi, S. Rezaei, Niche Divergence at Intraspecific Level in the Hyrcanian Wood Frog, Rana pseudodalmatina: A Phylogenetic, Climatic, and Environmental Survey. *Front. Ecol. Evol.* **10**, 1–12 (2022).

68. R. J. Elshire, *et al.*, A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS One* **6**, 1–10 (2011).

69. L. Cornet, D. Baurain, Contamination detection in genomic data: more is not enough. *Genome Biol.* **23**, 1–15 (2022).

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

71. D. A. DeRaad, snpfiltr: An R package for interactive and reproducible SNP filtering. *Mol. Ecol. Resour.* **00**, 1–11 (2022).

72. P. Danecek, *et al.*, The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158 (2011).

73. D. H. Alexander, J. Novembre, K. Lange, Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* **19**, 1655–1664 (2009).

74. D. H. Alexander, K. Lange, Enhancements to the ADMIXTURE algorithm for individual ancestry estimation. *BMC Bioinformatics* **12** (2011).

75. A. Raj, M. Stephens, J. K. Pritchard, FastSTRUCTURE: Variational inference of population structure in large SNP data sets. *Genetics* **197**, 573–589 (2014).

76. T. Jombart, Adegenet: A R package for the multivariate analysis of genetic markers. *Bioinformatics* **24**, 1403–1405 (2008).

77. J. M. Miller, C. I. Cullingham, R. M. Peery, The influence of a priori grouping on inference of genetic clusters: simulation study and literature review of the DAPC method. *Heredity (Edinb).* **125**, 269–280 (2020).

78. L. T. Nguyen, H. A. Schmidt, A. Von Haeseler, B. Q. Minh, IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* **32**, 268–274 (2015).

79. Z. N. Kamvar, J. F. Tabima, N. J. Grünwald, Poppr: an R package for genetic analysis of populations with clonal, partially clonal, and/or sexual reproduction. *PeerJ* **2**, e281 (2014).

80. D. Petkova, J. Novembre, M. Stephens, Visualizing spatial population structure with estimated effective migration surfaces. *Nat. Genet.* **48**, 94–100 (2015).

81. A. M. Olsen, M. W. Westneat, StereoMorph: An R package for the collection of 3D landmarks and curves using a stereo camera set-up. *Methods Ecol. Evol.* **6**, 351–356 (2015).

82. D. C. Adams, E. Otárola-Castillo, Geomorph: An r package for the collection and analysis of geometric morphometric shape data. *Methods Ecol. Evol.* **4**, 393–399 (2013).

83. P. Legendre, M. J. Anderson, Distance-based redundancy analysis: Testing multispecies responses in multifactorial ecological experiments. *Ecol. Monogr.* **69**, 1–24 (1999).

84. D. Philip, VEGAN , a package of R functions for community ecology. *J. Veg. Sci.* **14**, 927–930 (2003).

85. E. Pante, B. Simon-Bouhet, marmap: A Package for Importing, Plotting and Analyzing Bathymetric and Topographic Data in R. *PLoS One* **8**, 6–9 (2013).

86. S. Dray, *et al.*, adespatial: Multivariate Multiscale Spatial Analysis (2023).

87. D. L. Swofford, D. L. Swofford, D. L. Swofford, PAUP*: Phylogenetic analysis using parsimony (*and other methods), Version 4.0b10 in (2002).

88. D. Bryant, R. Bouckaert, J. Felsenstein, N. A. Rosenberg, A. Roychoudhury, Inferring species trees directly from biallelic genetic markers: Bypassing gene trees in a full coalescent analysis. *Mol. Biol. Evol.* **29**, 1917–1932 (2012).

89. R. Bouckaert, *et al.*, BEAST 2: A Software Platform for Bayesian Evolutionary Analysis. *PLoS Comput. Biol.* **10**, 1–6 (2014).

90. K. Tamura, G. Stecher, D. Peterson, A. Filipski, S. Kumar, MEGA6: Molecular evolutionary genetics analysis version 6.0. *Mol. Biol. Evol.* **30**, 2725–2729 (2013).

91. S. Kumar, G. Stecher, M. Li, C. Knyaz, K. Tamura, MEGA X: Molecular evolutionary genetics analysis across computing platforms. *Mol. Biol. Evol.* **35**, 1547–1549 (2018).

92. E. Paradis, K. Schliep, Ape 5.0: An environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics* **35**, 526–528 (2019).

Appendix A

Supplementary Material for

Genomics overrules mitochondrial DNA, siding with morphology on a controversial case of species delimitation

Extended Materials and Methods

Supplemental Figures and Tables

Supplementary References

Data repository:

datasets are available from Dryad digital repository: https://doi.org/10.5061/dryad.sk61618

Extended Materials and Methods: molecular protocols, mitochondrial data, SNP genotyping, and Bayes factor delimitation analyses.

DNA extractions were performed from fin-clip tissues using Qiagen DNeasy Blood and Tissue kit (Qiagen, Inc.), following the manufacturer's protocol. Library preparation was carried out at the Sequencing and Genotyping Facility (SGF) at the University of Puerto Rico – Río Piedras (UPR-RP). We applied the double-digest RADseq (ddRADseq) protocol using initially described by Peterson *et al.* (2012). This protocol provides the advantage of reducing genome-wide sequences to size-selected digested fragments [1]. Restriction enzymes used were PstI and MseI, with a size selection window of 300-600 bp. Each library contained up to 20 individuals indexed by a set of six base-pairs barcodes in combinatorial schemes, and further sequenced in two Illumina HiSeq

4000 lanes using 100 base pair-ended sequencing at the Knapp Center of Biomedical Discovery (KCBD) Genomics Facility at the University of Chicago. For *COI* barcoding, PCR products were checked by electrophoresis on 1.8% agarose gels and sequenced in both directions at the SGF at the UPR-RP. Accession numbers and additional information are available from Table S2.

Sequenced libraries were demultiplexed using the process_radtags.pl script as implemented in Stacks v1.49 [2]. Raw reads were trimmed to 86 bp after removing restriction sites. The quality of raw reads was further verified using FastQC v0.11.5 (www.bioinformatics.babraham.ac.uk/projects/fastqc/) and selected a Phred score threshold of 33 for filtering sequencing reads. The total number of raw reads was 1.68 X 10^9, of which 1.184 X 10^9 (~70%) passed quality filters. *De novo* assembling of putative loci and calling of single nucleotide polymorphisms (SNP) was carried out in Stacks using the denovo_map.pl pipeline. Selection of assembly parameters that adjusted best to our data was performed based on alternative strategies [3] (see below). RAD loci were assembled by applying default settings on all samples for a pilot run, which resulted in 4,326,243 putative loci. Different combinations of assembly parameters were tested on a subset of 30 samples that were selected on the basis of sequence coverage (>40% of all loci represented after a pilot test using the default settings). Combinations of parameters were tested as in Mastretta-Yanes [3], including minimum number of raw reads required to form a stack (m = 2-15), number of mismatches allowed between stacks (M = 2-10), number of mismatches allowed between loci upon catalogue building (n = 0-5), and maximum number of stacks allowed per single locus (*--max_locus_stacks* = 2-6). In all tests, only one parameter was changed at a time, while keeping others at their default value (m=3, M=2, n=1, and *max_locus_stacks* = 0; Fig. S2). Results of *de novo* assembly tests varied from 1 to ~8 million putative loci for 30 individuals (Fig. S2). The number of putative RAD loci stop dropping strongly after m = 5, suggesting that many low coverage loci are discarded (Fig. S2a). The same pattern is observed at M = 2 and n = 3 where values also stop, roughly showing large changes. When only polymorphic loci are present in a minimum of 6 populations (80%) and in at least 75% of individuals within populations included, values start to stabilize at M = 2 and *max_loc_stacks* = 3 (Fig. S2b). Exponential increases are observed from 12,000 to 60,000 SNPs for n = 0. In this case, higher values of n reflect an increase in RAD loci as this parameter allows more stacks and loci to be collapsed. Final selected parameters were m=5, M=2, n=3.

Biallelic loci were filtered in the populations pipeline component of Stacks, varying population coverage constraints (p = 15-8 and r = 1.0-0.50) and selecting only one SNP per tag to avoid linkage between loci. To remove low frequency and paralogous loci, SNPs were further filtered using a minimum allele frequency of 0.05 and a maximum observed heterozygosity of 0.70. Finally, we were interested in evaluating the sensitivity of results to number of individuals and missing data. Therefore, four of the previous datasets that contained between 21,431 and 55,795 loci were selected (p11r50, p12r50, p9r60 and p8r60), and applied a second filter removing individuals based on the amount of missing sites (minimum proportion of sites present of 0.75, 0.50, 0.25 and 0.05) using the program Tassel v5.2.43 [4]. We refer to this threshold as "min. sites." Missing data was measured in the output datasets using VCFtools v0.1.15 [5]. These tests resulted in 20 datasets (Table S4), of which six were selected based on the amount of missing data

(9–46%), number of individuals present (44–155), number of SNPs (15,112–42,406), and number of populations (8–15).

For the Bayes factor delimitation (BFD*) analyses, additional filters were applied to the p12r50 dataset (with 15,112 SNPs from 12 populations) by retaining both loci and individuals from each population with the lowest proportions of missing data (e.g., using Tassel v5.2.43 [4]). The subsets were assembled: subset 1: 58 individuals and 149 loci (each locus is present in at least 55 individuals); subset 2: 58 individuals and 938 loci (each locus is present in at least 51 individuals); subset 3: 108 individuals and 957 loci (each locus is present in at least 85 individuals). The XML files for the BFD analyses performed are available from Dryad.

Supplementary Figures and Tables

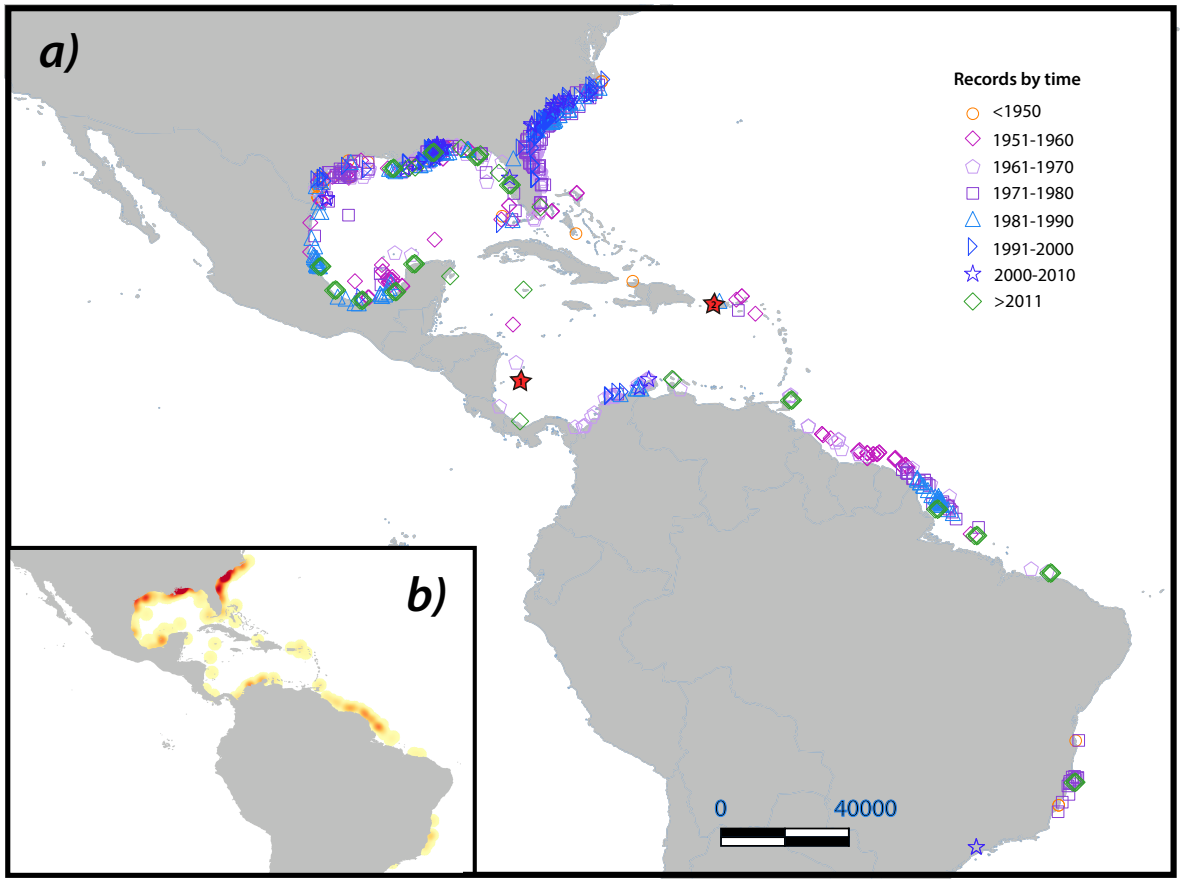


Figure S1. a) map illustrating reports of *Lutjanus campechanus* and *Lutjanus purpureus* throughout their range and over time; b) heatmap of records of both species. Red stars in (a) indicate localities that were exhaustively but unsuccessfully surveyed for samples: 1) San Andrés, Colombia; 2) Puerto Rico.

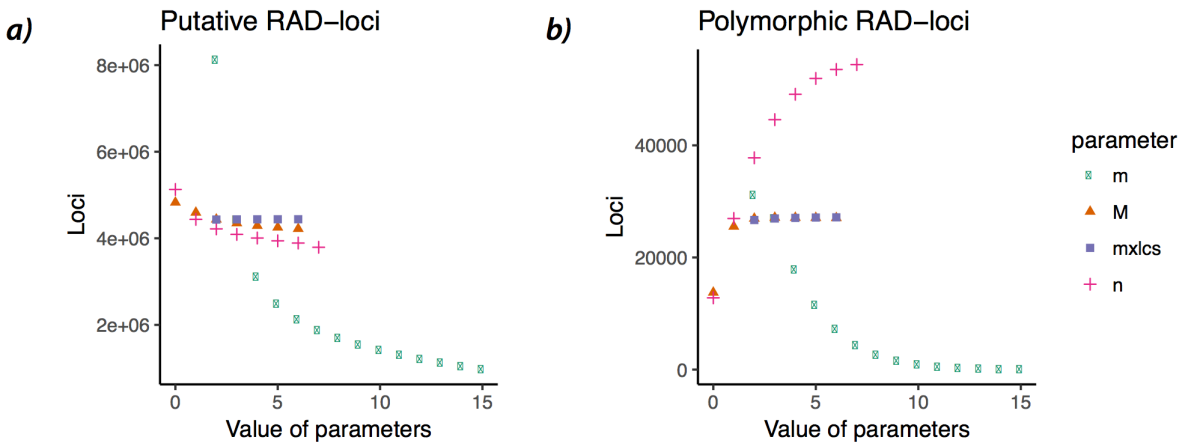


Figure S2. Tests of parameter combinations (following Mastretta-Yanes et al. [3]).

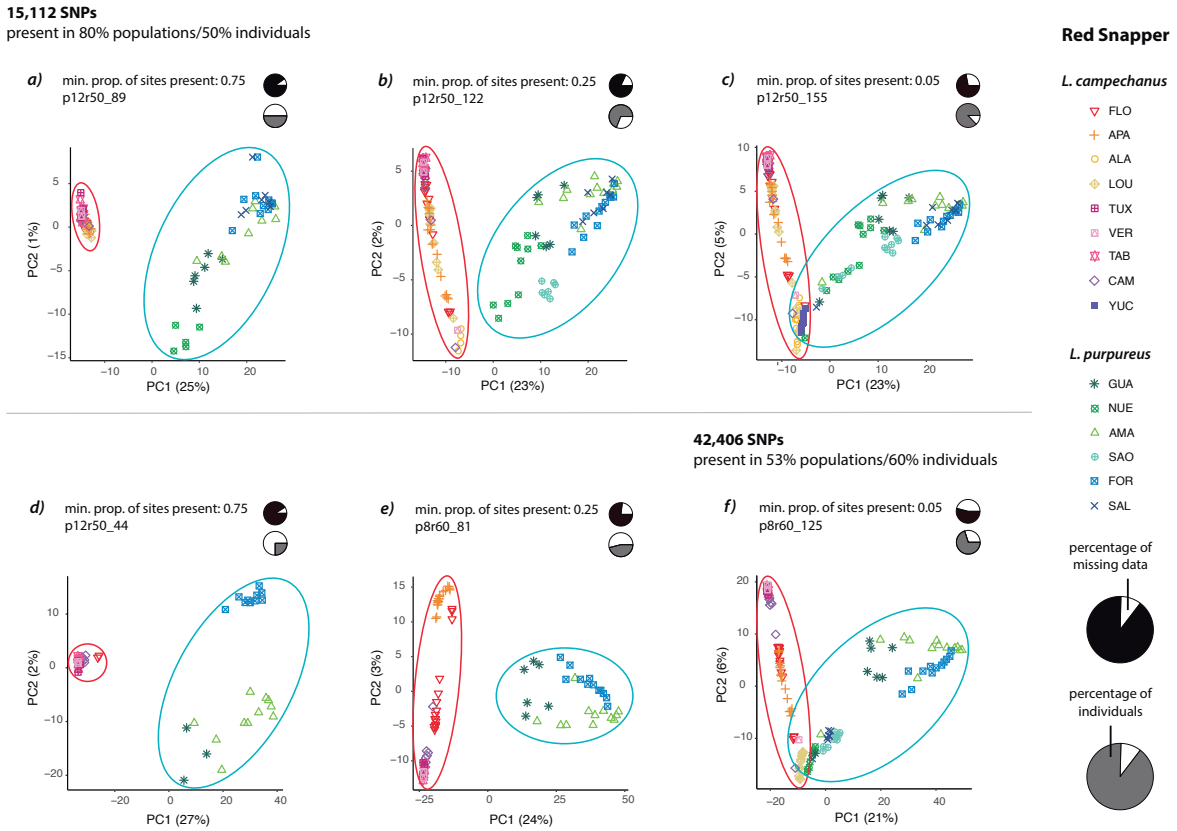


Figure S3. Principal component analyses of allele frequency data from a matrix based on (a-c) 15,112 SNPs and (d-f) 42,406 SNPs.

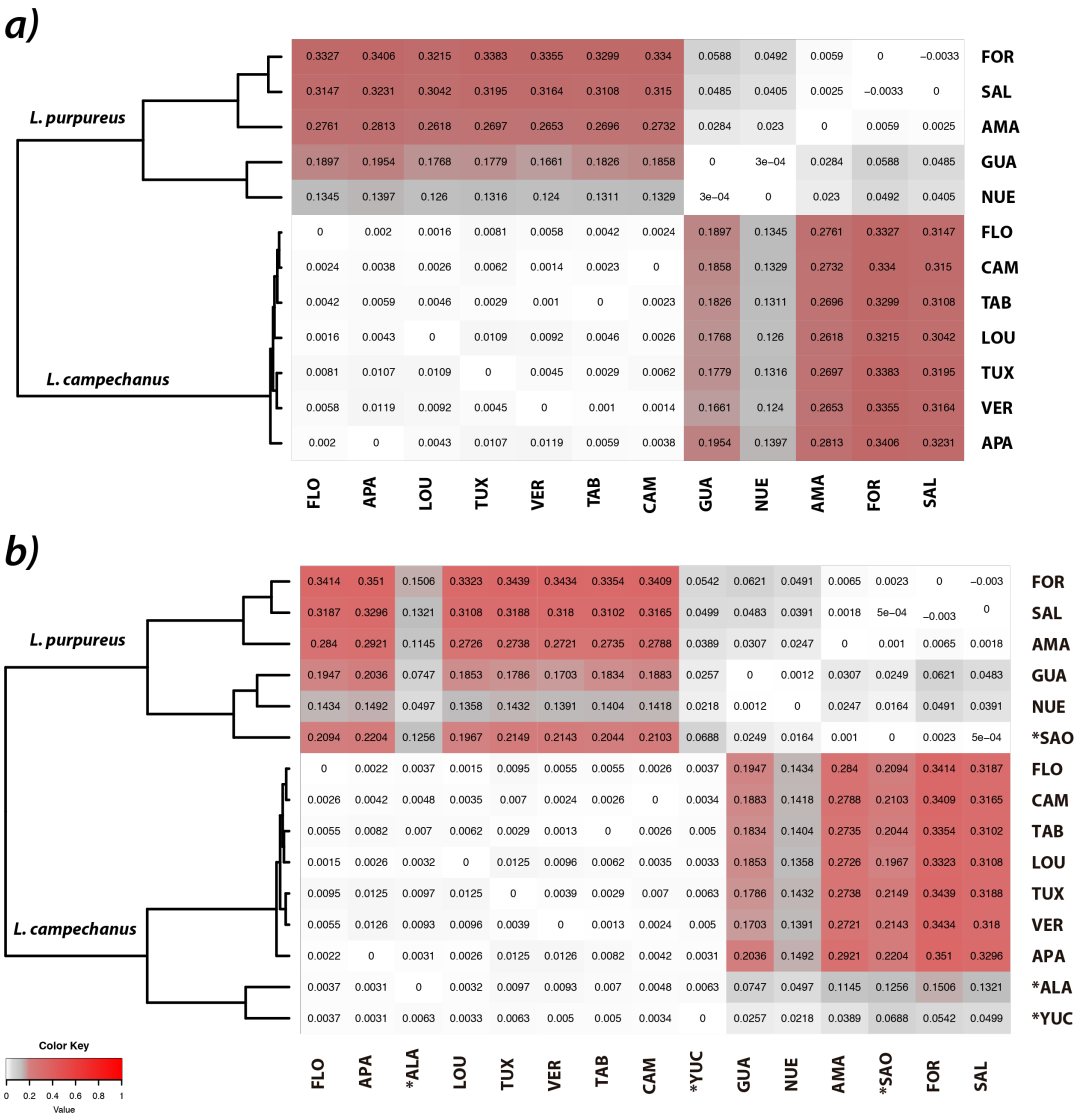


Figure S4. Weir and Cockerham estimates of genetic differentiation among populations of *L. campechanus* and *L. purpureus*, including a) 12 populations (with minimal amount of missing data), and b) 15 populations. Populations removed form final analyses in a) are denoted with an asterisk in b). Some calculations in b) appear underestimated despite their great geographic isolation, possibly as result of missing data: e.g., Yucatán vs. Salvador, F_{ST}=0.0499 (1646 loci); Salvador-Campeche F_{ST}=0.3165 (4282 loci); Salvador-Florida F_{ST}=0.3187 (14102 loci).

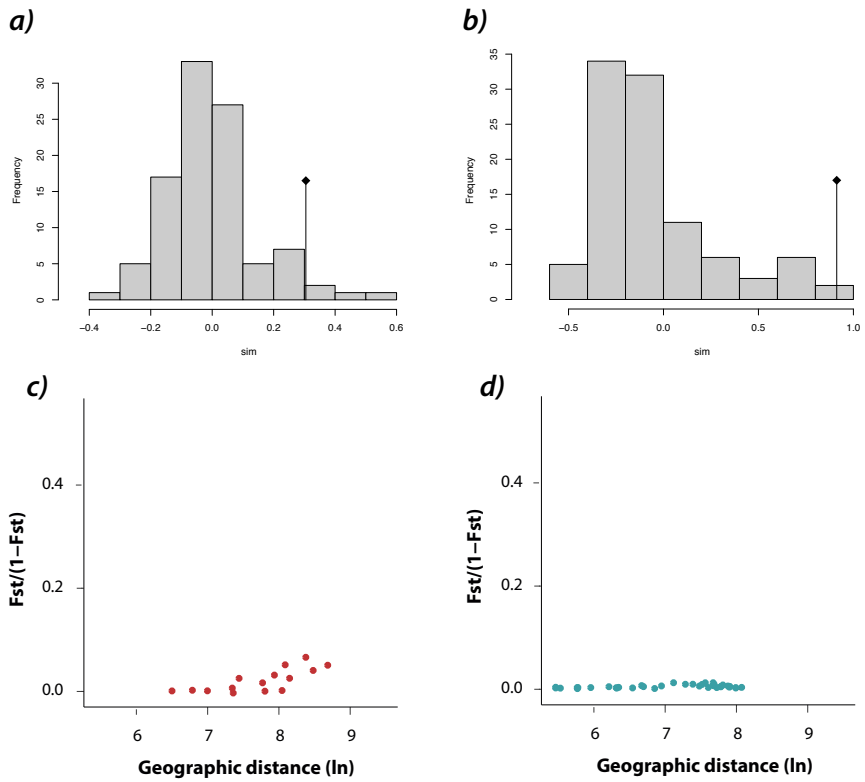
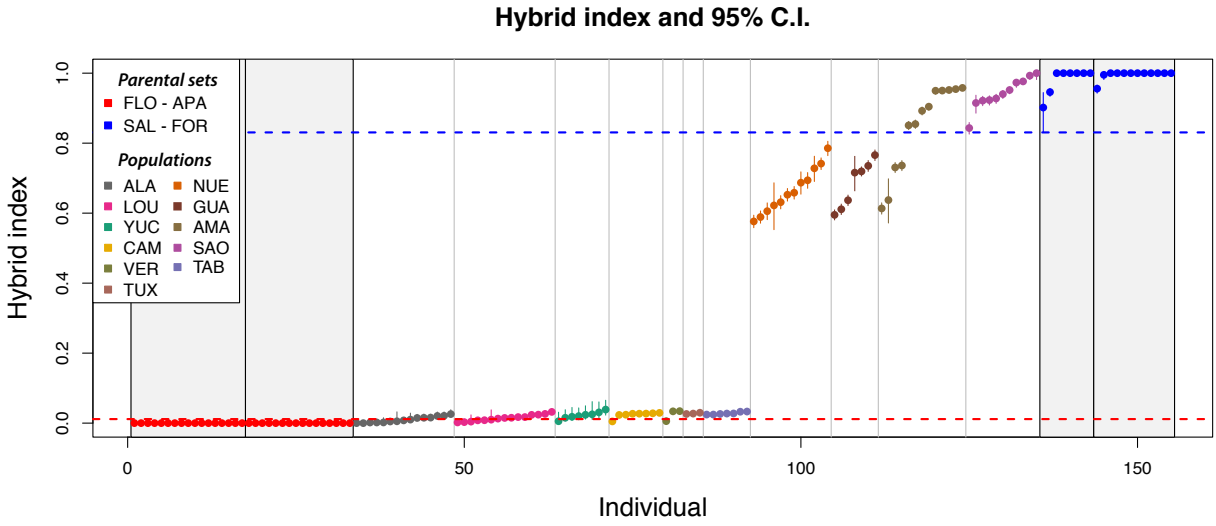


Figure S5. Mantel test simulations (a,b) and correlograms (c,d) for *L. campechanus* (red) and *L. purpureus* (green), respectively. Geographic distances reflect Least Cost Paths (LCPs); similar results were obtained using Euclidean distances (not shown).



studied populations of *L. campechanus*, and *L. purpureus*. Red and blue points represent the parental reference sets of *L. campechanus* (Florida-Apalachicola) and *L. purpureus* (Fortaleza-Salvador), respectively. Individuals are color-coded according to their corresponding population. Population codes on the upper left corner correspond to abbreviations given in Table S1. Points indicate HI estimated values, where 0.0 denotes pure *L. campechanus* individuals and 1.0 denotes pure *L. purpureus* individuals; lines represent 95% credibility interval.

Table S1. Sampling localities

Population abbreviation	Locality	Country	Geographic coordinates	Number of individuals
Lutjanus campechanus				
FLO	West of Florida, Florida	USA	27.1666°N -83.3833°W	18
APA	Apalachicola, Florida	USA	28.7900°N -85.1635°W	18
ALA	Dauphin Island, Alabama	USA	29.7836°N -88.0732°W	18
LOU	Louisiana	USA	28.2443°N -91.4573°W	17
TUX	Puerto de Tuxpan, Veracruz	Mexico	21.0544°N -97.1926°W	3
VER	Puerto de Veracruz, Veracruz	Mexico	19.2802°N -96.0216°W	3
TAB	San Pedro, Tabasco	Mexico	18.7480°N -93.1829°W	7
CAM	Champotón, Campeche	Mexico	19.4576°N -91.1248°W	9
YUC	Puerto Progreso, Yucatán	Mexico	21.6932°N -89.6345°W	12
Lutjanus purpureus				
GUA	Camarones, Guajira	Colombia	11.8624°N -72.8027°W	16
NUE	Nueva Esparta	Venezuela	11.3417°N -63.8558°W	7
AMA	Amapá	Brazil	4.33564°N -50.5537°W	13
SAO	Maranhão, São Luís	Brazil	0.2172°N -46.246°W	13
FOR	Ceará, Fortaleza	Brazil	2.9071°S -39.1452°W	13
SAL	São Salvador da Bahia, Salvador	Brazil	13.226°S -38.6245°W	12

Table S2. GenBank (NCBI) accession numbers for raw data (Bioproject PRJNA524905) and mitochondrial sequences generated. See Table S1 for population abbreviations.

Species	Sample code	Population	COI Accession no.	Raw data (ddRAD) Accession no.
Lutjanus campechanus	30915-1	FLO	-	SAMN11038299
Lutjanus campechanus	30915-10	FLO	-	SAMN11038300
Lutjanus campechanus	30915-3	FLO	-	SAMN11038301
Lutjanus campechanus	30915-4	FLO	-	SAMN11038302
Lutjanus campechanus	30915-5	FLO	-	SAMN11038303
Lutjanus campechanus	30915-6	FLO	-	SAMN11038304
Lutjanus campechanus	30915-7	FLO	-	SAMN11038305
Lutjanus campechanus	30915-8	FLO	-	SAMN11038306
Lutjanus campechanus	30915-9	FLO	-	SAMN11038307
Lutjanus campechanus	31215-10	FLO	-	SAMN11038308
Lutjanus campechanus	31215-11	FLO	-	SAMN11038309
Lutjanus campechanus	31215-12	FLO	-	SAMN11038310
Lutjanus campechanus	31215-13	FLO	-	SAMN11038311
Lutjanus campechanus	31215-14	FLO	-	SAMN11038312
Lutjanus campechanus	31215-15	FLO	-	SAMN11038313
Lutjanus campechanus	31215-16	FLO	-	SAMN11038314
Lutjanus campechanus	31215-17	FLO	-	SAMN11038315
Lutjanus campechanus	31215-9	FLO	-	SAMN11038316
Lutjanus campechanus	ABR169	ALA	-	SAMN11038317
Lutjanus campechanus	ABR170	ALA	MK534317	SAMN11038318
Lutjanus campechanus	ABR34	ALA	-	SAMN11038319
Lutjanus campechanus	ABR35	ALA	-	SAMN11038320
Lutjanus campechanus	ABR55	ALA	-	SAMN11038321
Lutjanus campechanus	ABR59	ALA	-	SAMN11038322
Lutjanus campechanus	ABR60	ALA	-	SAMN11038323
Lutjanus campechanus	ABR82	ALA	-	SAMN11038324
Lutjanus campechanus	ABR84	ALA	-	SAMN11038325
Lutjanus campechanus	ES100614HL-20	ALA	MK534318	SAMN11038326
Lutjanus campechanus	ES100614HL-76	ALA	MK534319	-

Lutjanus campechanus	ES100614HL-78	ALA	-	SAMN11038327
Lutjanus campechanus	ES100614HL-82	ALA	-	SAMN11038328
Lutjanus campechanus	ES100614HL-83	ALA	-	SAMN11038329
Lutjanus campechanus	ES100614HL-86	ALA	-	SAMN11038330
Lutjanus campechanus	ES100614HL-87	ALA	-	SAMN11038331
Lutjanus campechanus	ES100614HL-90	ALA	-	SAMN11038332
Lutjanus campechanus	ES100614HL-91	ALA	-	SAMN11038333
Lutjanus campechanus	ES100614HL-94	ALA	-	SAMN11038334
Lutjanus campechanus	LC424	APA	-	SAMN11038339
Lutjanus campechanus	LC425	APA	-	SAMN11038340
Lutjanus campechanus	LC426	APA	-	SAMN11038341
Lutjanus campechanus	LC427	APA	-	SAMN11038342
Lutjanus campechanus	LC428	APA	-	SAMN11038343
Lutjanus campechanus	LC429	APA	-	SAMN11038344
Lutjanus campechanus	LC430	APA	-	SAMN11038345
Lutjanus campechanus	LC431	APA	-	SAMN11038346
Lutjanus campechanus	LC438	APA	-	SAMN11038347
Lutjanus campechanus	LC447	APA	-	SAMN11038348
Lutjanus campechanus	LC448	APA	-	SAMN11038349
Lutjanus campechanus	LC449	APA	-	SAMN11038350
Lutjanus campechanus	LC450	APA	-	SAMN11038351
Lutjanus campechanus	LC451	APA	-	SAMN11038352
Lutjanus campechanus	LC452	APA	-	SAMN11038353
Lutjanus campechanus	LC453	APA	-	SAMN11038354
Lutjanus campechanus	LC454	APA	-	SAMN11038355
Lutjanus campechanus	LC455	APA	-	SAMN11038356
Lutjanus campechanus	LSU223	LOU	-	SAMN11038367
Lutjanus campechanus	LSU224	LOU	-	SAMN11038368
Lutjanus campechanus	LSU226	LOU	-	SAMN11038369
Lutjanus campechanus	LSU227	LOU	-	SAMN11038370
Lutjanus campechanus	LSU228	LOU	-	SAMN11038371
Lutjanus campechanus	LSU229	LOU	-	SAMN11038372
Lutjanus campechanus	LSU230	LOU	-	SAMN11038373
Lutjanus campechanus	LSU231	LOU	-	SAMN11038374
Lutjanus campechanus	LSU268	LOU	-	SAMN11038375
Lutjanus campechanus	LSU271	LOU	-	SAMN11038376

Lutjanus campechanus	LSU272	LOU	-	SAMN11038377
Lutjanus campechanus	LSU273	LOU	-	SAMN11038378
Lutjanus campechanus	LSU276	LOU	-	SAMN11038379
Lutjanus campechanus	LSU278	LOU	-	SAMN11038380
Lutjanus campechanus	LSU279	LOU	-	SAMN11038381
Lutjanus campechanus	LSU280	LOU	-	SAMN11038382
Lutjanus campechanus	LSU281	LOU	-	SAMN11038383
Lutjanus campechanus	RB1537	CAM	MK534248	-
Lutjanus campechanus	RB1538	CAM	MK534249	-
Lutjanus campechanus	RB1539	CAM	MK534250	-
Lutjanus campechanus	RB1540	CAM	MK534251	-
Lutjanus campechanus	RB1542	CAM	MK534252	-
Lutjanus campechanus	RB1543	CAM	MK534253	-
Lutjanus campechanus	RB1544	CAM	MK534254	-
Lutjanus campechanus	RB1545	CAM	MK534255	-
Lutjanus campechanus	RB1546	CAM	MK534256	-
Lutjanus campechanus	RB1547	CAM	MK534257	-
Lutjanus campechanus	RB1548	CAM	MK534258	-
Lutjanus campechanus	RB1550	CAM	MK534259	-
Lutjanus campechanus	RB1551	CAM	MK534260	-
Lutjanus campechanus	RB1552	CAM	MK534261	-
Lutjanus campechanus	RB1553	CAM	MK534262	-
Lutjanus campechanus	RB1554	CAM	MK534263	-
Lutjanus campechanus	RB1555	CAM	MK534264	SAMN11038434
Lutjanus campechanus	RB1556	CAM	MK534265	SAMN11038435
Lutjanus campechanus	RB1557	CAM	MK534266	-
Lutjanus campechanus	RB1558	CAM	MK534267	SAMN11038436
Lutjanus campechanus	RB1559	CAM	MK534268	-
Lutjanus campechanus	RB1560	CAM	MK534269	-
Lutjanus campechanus	RB1561	CAM	MK534270	SAMN11038437
Lutjanus campechanus	RB1562	CAM	MK534271	SAMN11038438
Lutjanus campechanus	RB1563	CAM	MK534272	-
Lutjanus campechanus	RB1564	CAM	MK534273	-
Lutjanus campechanus	RB1565	CAM	MK534274	-
Lutjanus campechanus	RB1566	CAM	MK534275	-
Lutjanus campechanus	RB1567	CAM	MK534276	SAMN11038439

Lutjanus campechanus	RB1568	CAM	MK534277	-
Lutjanus campechanus	RB1569	CAM	MK534278	-
Lutjanus campechanus	RB1570	CAM	MK534279	-
Lutjanus campechanus	RB1571	CAM	MK534280	SAMN11038440
Lutjanus campechanus	RB1572	CAM	MK534281	-
Lutjanus campechanus	RB1573	CAM	MK534282	-
Lutjanus campechanus	RB1574	CAM	MK534283	-
Lutjanus campechanus	RB1575	CAM	MK534284	-
Lutjanus campechanus	RB1576	CAM	MK534285	-
Lutjanus campechanus	RB1577	CAM	MK534286	-
Lutjanus campechanus	RB1578	CAM	MK534287	-
Lutjanus campechanus	RB1579	CAM	MK534288	-
Lutjanus campechanus	RB1580	CAM	MK534289	-
Lutjanus campechanus	RB1581	CAM	MK534290	-
Lutjanus campechanus	RB1582	CAM	MK534291	-
Lutjanus campechanus	RB1583	CAM	MK534292	-
Lutjanus campechanus	RB1584	CAM	MK534293	-
Lutjanus campechanus	RB1585	CAM	MK534294	-
Lutjanus campechanus	RB1586	TAB	MK534295	-
Lutjanus campechanus	RB1587	TAB	MK534297	-
Lutjanus campechanus	RB1588	TAB	MK534296	-
Lutjanus campechanus	RB1589	TAB	MK534298	-
Lutjanus campechanus	RB1590	TAB	MK534299	-
Lutjanus campechanus	RB1591	TAB	MK534300	-
Lutjanus campechanus	RB1593	CAM	MK534301	-
Lutjanus campechanus	RB1594	CAM	MK534302	-
Lutjanus campechanus	RB1595	CAM	MK534303	-
Lutjanus campechanus	RB1596	CAM	MK534304	-
Lutjanus campechanus	RB1597	CAM	MK534305	-
Lutjanus campechanus	RB1598	CAM	MK534306	-
Lutjanus campechanus	RB1599	TAB	MK534307	SAMN11038441
Lutjanus campechanus	RB1600	TAB	MK534308	SAMN11038442
Lutjanus campechanus	RB1601	TAB	MK534309	SAMN11038443
Lutjanus campechanus	RB1602	TAB	MK534310	SAMN11038444
Lutjanus campechanus	RB1603	TAB	MK534311	SAMN11038445
Lutjanus campechanus	RB1604	TAB	MK534312	SAMN11038446

Lutjanus campechanus	RB1605	TAB	MK534313	SAMN11038447
Lutjanus campechanus	RB1606	TUX	-	SAMN11038448
Lutjanus campechanus	RB1607	TUX	-	SAMN11038449
Lutjanus campechanus	RB1608	TUX	-	SAMN11038450
Lutjanus campechanus	RB1609	VER	-	SAMN11038451
Lutjanus campechanus	RB1610	VER	-	SAMN11038452
Lutjanus campechanus	RB1611	VER	-	SAMN11038453
Lutjanus campechanus	RB1637	YUC	-	SAMN11038454
Lutjanus campechanus	RB1640	YUC	-	SAMN11038455
Lutjanus campechanus	RB1641	YUC	-	SAMN11038456
Lutjanus campechanus	RB1642	YUC	-	SAMN11038457
Lutjanus campechanus	RB1643	YUC	-	SAMN11038458
Lutjanus campechanus	RB1645	YUC	-	SAMN11038459
Lutjanus campechanus	RB1646	YUC	-	SAMN11038460
Lutjanus campechanus	RB1647	YUC	-	SAMN11038461
Lutjanus campechanus	RB1648	YUC	MK534314	SAMN11038462
Lutjanus campechanus	RB1649	YUC	MK534315	-
Lutjanus campechanus	RB1650	YUC	MK534316	SAMN11038463
Lutjanus campechanus	RB1651	YUC	-	SAMN11038464
Lutjanus campechanus	RB1652	YUC	-	SAMN11038465
Lutjanus campechanus	UMSNH41727	CAM	-	SAMN11038469
Lutjanus campechanus	UMSNH41728	CAM	-	SAMN11038470
Lutjanus purpureus	GUA001	GUA	-	SAMN11038335
Lutjanus purpureus	GUA002	GUA	-	SAMN11038336
Lutjanus purpureus	GUA003	GUA	-	SAMN11038337
Lutjanus purpureus	GUA004	GUA	-	SAMN11038338
Lutjanus purpureus	LP1	NUE	MK534320	SAMN11038357
Lutjanus purpureus	LP10	GUA	-	SAMN11038358
Lutjanus purpureus	LP2	VEN	-	SAMN11038359
Lutjanus purpureus	LP3	NUE	MK534321	SAMN11038360
Lutjanus purpureus	LP4	VEN	-	SAMN11038361
Lutjanus purpureus	LP5	VEN	-	SAMN11038362
Lutjanus purpureus	LP6	VEN	-	SAMN11038363
Lutjanus purpureus	LP7	VEN	-	SAMN11038364
Lutjanus purpureus	LP8	GUA	MK534322	SAMN11038365
Lutjanus purpureus	LP9	GUA	MK534323	SAMN11038366

Lutjanus purpureus	RB1432	AMA	-	SAMN11038384
Lutjanus purpureus	RB1433	AMA	-	SAMN11038385
Lutjanus purpureus	RB1434	AMA	-	SAMN11038386
Lutjanus purpureus	RB1436	AMA	-	SAMN11038387
Lutjanus purpureus	RB1438	AMA	-	SAMN11038388
Lutjanus purpureus	RB1439	AMA	-	SAMN11038389
Lutjanus purpureus	RB1442	AMA	-	SAMN11038390
Lutjanus purpureus	RB1444	AMA	-	SAMN11038391
Lutjanus purpureus	RB1445	AMA	-	SAMN11038392
Lutjanus purpureus	RB1446	AMA	-	SAMN11038393
Lutjanus purpureus	RB1458	AMA	-	SAMN11038394
Lutjanus purpureus	RB1459	AMA	-	SAMN11038395
Lutjanus purpureus	RB1460	AMA	-	SAMN11038396
Lutjanus purpureus	RB1470	FOR	-	SAMN11038397
Lutjanus purpureus	RB1471	FOR	-	SAMN11038398
Lutjanus purpureus	RB1473	FOR	-	SAMN11038399
Lutjanus purpureus	RB1474	FOR	-	SAMN11038400
Lutjanus purpureus	RB1475	FOR	-	SAMN11038401
Lutjanus purpureus	RB1476	FOR	-	SAMN11038402
Lutjanus purpureus	RB1477	FOR	-	SAMN11038403
Lutjanus purpureus	RB1478	FOR	-	SAMN11038404
Lutjanus purpureus	RB1479	FOR	-	SAMN11038405
Lutjanus purpureus	RB1482	FOR	-	SAMN11038406
Lutjanus purpureus	RB1484	FOR	-	SAMN11038407
Lutjanus purpureus	RB1490	FOR	-	SAMN11038408
Lutjanus purpureus	RB1496	SAO	-	SAMN11038409
Lutjanus purpureus	RB1497	SAO	-	SAMN11038410
Lutjanus purpureus	RB1498	SAO	-	SAMN11038411
Lutjanus purpureus	RB1499	SAO	-	SAMN11038412
Lutjanus purpureus	RB1502	SAO	-	SAMN11038413
Lutjanus purpureus	RB1503	SAO	-	SAMN11038414
Lutjanus purpureus	RB1504	SAO	-	SAMN11038415
Lutjanus purpureus	RB1505	SAO	-	SAMN11038416
Lutjanus purpureus	RB1507	SAO	-	SAMN11038417
Lutjanus purpureus	RB1510	SAO	-	SAMN11038418
Lutjanus purpureus	RB1514	SAO	-	SAMN11038419

Lutjanus purpureus	RB1515	SAO	-	SAMN11038420
Lutjanus purpureus	RB1516	SAO	-	SAMN11038421
Lutjanus purpureus	RB1520	SAL	-	SAMN11038422
Lutjanus purpureus	RB1522	SAL	-	SAMN11038423
Lutjanus purpureus	RB1523	SAL	-	SAMN11038424
Lutjanus purpureus	RB1524	SAL	-	SAMN11038425
Lutjanus purpureus	RB1525	SAL	-	SAMN11038426
Lutjanus purpureus	RB1526	SAL	-	SAMN11038427
Lutjanus purpureus	RB1527	SAL	-	SAMN11038428
Lutjanus purpureus	RB1528	SAL	-	SAMN11038429
Lutjanus purpureus	RB1531	SAL	-	SAMN11038430
Lutjanus purpureus	RB1532	SAL	-	SAMN11038431
Lutjanus purpureus	RB1533	SAL	-	SAMN11038432
Lutjanus purpureus	RB1534	SAL	-	SAMN11038433
Lutjanus purpureus	UMSNH16756	NUE	MK534325	SAMN11038466
Lutjanus purpureus	UMSNH16996	NUE	MK534324	SAMN11038467
Lutjanus purpureus	UMSNH16997	NUE	MK534326	SAMN11038468
Lutjanus purpureus	VEN001	VEN	-	SAMN11038471
Lutjanus purpureus	VEN002	VEN	-	SAMN11038472
Lutjanus purpureus	VEN003	VEN	-	SAMN11038473
Lutjanus purpureus	VEN004	VEN	-	SAMN11038474
Lutjanus purpureus	VEN005	VEN	-	SAMN11038475
Lutjanus purpureus	VEN006	VEN	-	SAMN11038476

Table S3. Accession numbers for sequences mined from GenBank (NCBI).

Organism	Region/Location information	Genetic marker	GenBank accession no.
Lutjanus campechanus	Gulf of Mexico and Atlantic coast of Florida	D-loop	AF356750-AF356776
Lutjanus campechanus	US coast	COI	EU752115
Lutjanus campechanus	Florida	COI	FJ998466
Lutjanus campechanus	Texas	COI	HQ162371-HQ162373
Lutjanus campechanus	NA	COI	JN021303
Lutjanus campechanus	Alabama	COI	KF461194-KF461195
Lutjanus campechanus	NA	COI	KX119461-KX119464
Lutjanus campechanus	NA	COI	KX119465
Lutjanus campechanus	Gulf of Mexico	COI	MF041054
Lutjanus campechanus	Gulf of Mexico	COI	MF041257
Lutjanus campechanus	Gulf of Mexico	COI	MF041562
Lutjanus campechanus	Gulf of Mexico	COI	MG856504
Lutjanus campechanus	Gulf of Mexico	COI	MF041450
Lutjanus purpureus	Brazilian coast	D-loop	KC122929-KC123167
Lutjanus purpureus	NA	COI	EU752118
Lutjanus purpureus	Brazilian coast	COI	KJ907265
Lutjanus purpureus	Brazilian coast	COI	KU313736-KU313755

Table S4. Selected output datasets based on alternative population parameters. Missing percentage of SNPs ranged between 9 and 46 %, where the inclusion of more individuals was the driving factor affecting this value.

Dataset	Number of Individuals	Min. sites	Number of Populations	Number of loci	SNPs	Missing percentage of SNPs
	178	0	15	21431	15112	0.43
	***155**	**0.05**	**15**	**21431**	**15112**	**0.35**
p12r50	**122**	**0.25**	**14**	**21431**	**15112**	**0.2**
	108	0.5	13	21431	15112	0.15
	89	**0.75**	**12**	**21431**	**15112**	**0.1**
	178	0	15	40210	29798	0.48
	149	0.05	15	40210	29798	0.38
p11r50	115	0.25	13	40210	29798	0.23
	97	0.5	12	40210	29798	0.16
	73	0.75	10	40210	29798	0.1
	178	0	15	30138	21850	0.56
	136	0.05	14	30138	21850	0.43
p9r60	97	0.25	11	30138	21850	0.25
	76	0.5	9	30138	21850	0.13
	67	0.75	9	30138	21850	0.1
	178	0	15	55795	42406	0.61
	125	**0.05**	**13**	**55795**	**42406**	**0.46**
p8r60	**81**	**0.25**	**9**	**55795**	**42406**	**0.23**
	73	0.5	9	55795	42406	0.1
	44	**0.75**	**8**	**55795**	**42406**	**0.09**

p denotes the minimum number of populations where each locus must be present in order to be selected; *r* denotes the minimum percentage of individuals within a population where a locus must be present to be selected. *final selected matrixes used in downstream analyses.

Table S5. Results of Bayes factor delimitation (BFD*) analyses for WA red snappers using three SNP datasets. BF values were estimated with three subsets assembled from 12 populations (filtered from the p12r50 dataset). Positive BF values (calculated as 2 x (MLE of base model - MLE of alternative model)) indicate in all cases overwhelming support in favor of the traditional, two-species delimitation (base) model (BF scale: 0 < BF < 2, non-significant; 2 < BF < 6, positive evidence; 6 < BF < 10, strong support; BF > 10, decisive support; [6]. MLE: marginal likelihood estimates.

	Individuals	SNPs	MLE – 2 species (base – traditional taxonomy)	MLE – 1 species (alternative – mtDNA)	BF
Subset 1	58	149	-6,537.45	-7,692.59	2,310.28
Subset 2	58	938	-40,948.80	-47,287.26	12,676.92
Subset 3	108	957	-67285.99	-78,464.10	22,356.22

Table S6. List of character transformations (SNPs) that differentiate populations of *L. campechanus* from *L. purpureus*. Alignment site indicates the position of the diagnostic SNP in the concatenated nexus file (available from Dryad). Although there are no unambiguous SNPs differentiating all individuals of *L. campechanus* from *L. purpureus*, a combination of SNP sites can be used for diagnostic purposes; for instance, by designing PCR primers from the flanking sites in the nexus and fasta files provided.

Alignment site	Character state transformation	Alignment site	Character state transformation
1059	C --> T	836965	A --> G
41919	G --> A	850906	A --> C
45777	C --> T	886936	G --> C
53139	C --> A	975574	A --> C
68641	C --> T	987321	T --> C
88992	G --> A	1041916	A --> G
207540	A --> G	1144840	T --> C
214748	T --> C	1192851	C --> T
239812	C --> T	1198134	T --> G
248229	G --> A	1202462	C --> T
257027	G --> A	1228440	G --> A
270080	C --> T	1248434	G --> A
270830	G --> A	1274734	C --> T
288972	G --> A	1278851	A --> G
328042	G --> A	1283521	C --> G
459746	C --> T	1317037	G --> A
510028	G --> T	1358528	G --> C
521086	C --> T	1379595	G --> T
532482	C --> T	1387126	C --> A
629127	G --> A	1510047	G --> T
644504	A --> G	1534354	T --> C
648340	T --> G	1600837	A --> C
668691	T --> C	1638177	T --> G
681224	T --> C	1697315	A --> G

690051	G --> A	1772989	C --> T
769005	A --> C	1805758	T --> C
833330	C --> T		

Supplementary References

1. 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**. (doi:10.1371/journal.pone.0037135)

2. Catchen JM, Hohenlohe PA, Bassham S, Amores A, Cresko WA. 2013 Stacks: an analysis tool set for population genomics. *Mol. Ecol.* **22**, 3124–3140. (doi:10.1111/mec.12354.Stacks)

3. Mastretta-Yanes A, Arrigo N, Alvarez N, Jorgensen TH, Piñeros D, Emerson BC. 2015 Restriction site-associated DNA sequencing, genotyping error estimation and de novo assembly optimization for population genetic inference. *Mol. Ecol.* , 28–41. (doi:10.1111/1755-0998.12291)

4. Bradbury PJ, Zhang Z, Kroon DE, Casstevens TM, Ramdoss Y, Buckler ES. 2007 TASSEL: Software for association mapping of complex traits in diverse samples. *Bioinformatics* **23**, 2633–2635. (doi:10.1093/bioinformatics/btm308)

5. Danecek P *et al.* 2011 The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158. (doi:10.1093/bioinformatics/btr330)

6. Kass RE, Raftery AE. 2012 Bayes Factors. *J. Am. Stat. Assoc.* **90**, 773–795.

Appendix B

Supplementary Material for Genome-wide species delimitation analyses of a silverside fish species complex in central Mexico indicate taxonomic over-splitting

Supporting Methods

Supplementary Figures

Supplementary References

Supporting Methods:

Molecular protocols and SNP genotyping.

We extracted DNA from fin clips of 77 individuals of the nine nominal species of the *humboldtianum* group (*sensu* Barbour 1973), using the Qiagen DNeasy Blood and Tissue kit (Qiagen, Inc.) following the manufacturer's protocol. We also included two specimens of *Chirostoma jordani* and one of *Chirostoma attenuatum* as outgroups. We prepared the ddRADseq libraries at the Sequencing and Genotyping Facility (SGF) at the University of Puerto Rico - Río Piedras (UPR-RP) using the protocol of Peterson *et al.* (2012). To this end, we used the restriction enzymes *PstI* and *MseI*, and a size selection window of 300–600 bp. We included 20 individuals per library using a set of six base-pairs barcodes in combinatorial schemes to index each individual. We sequenced the resulting libraries in one Illumina HiSeq 4000 lane using 100 base pair-ended

sequencing at the Knapp Center of Biomedical Discovery (KCBD) Genomics Facility at the University of Chicago.

We checked the quality of a total of 358,591,878 raw reads with FASTQ 0.11.5 (www.bioinformatics.babraham.ac.uk/projects/fastqc). We used the *process_radtags* pipeline available in Stacks v2.4 [3, 4] to demultiplex our ddRADseq libraries. We applied a quality control Phred score of 33 to filter demultiplexed reads and trimmed the sequences to 86 bp after removing the enzyme's overhangs. We used a total of 349,891,799 reads that passed the filters (97.6%) to assemble putative loci.

We conducted a *de novo* assembly pilot run using *denovo.pl* program and default parameters in Stacks on all 80 individuals. As the selection of key parameters for *de novo* assembly (*m* = minimum raw reads required to form a stack, *M* = maximum mismatches between stacks, and *n* = mismatches between loci of different individuals) greatly influences the quality and formation of putative loci [5], we selected a subset of 15 samples (including individuals of each nominal species that presented the highest coverage values) to optimize the assembly parameters that best fit our data. We followed a combination of the protocols used by Mastretta-Yanes *et al.* (2015), Paris *et al.* (2017), and Del Pedraza-Marrón *et al.* (2019), by varying one parameter at the time (*m* = 2–6, *M* = 0–6, and *n* = 0–11). We observed a common pattern of RADseq data in which higher values of *m* increased the average sample coverage (Figure S1) but decreased the number of putative loci (Figure S2). Overall, increasing *m* from three to six produced a lower number of putative loci. Based on these results, we selected a value of five, as higher values could exclude true alleles, underestimating the number of heterozygous loci in the dataset [6]. Finally, we assessed the variation of putative loci while constraining the selection to genetic information available of the individuals in a population (*r* =40, 60, and 80) (Figure S3). After a maximum of four mismatches between stacks (*M* = 4) and five mismatches between loci of different individuals (*n* = 5), the number of putative loci stops dropping drastically. Therefore, we selected a final combination of *m*5*M*4*n*5 to perform the *de novo* assembly on the 80 individuals.

Given the lack of a reference genome that guided the loci assembly, we conducted a series of quality filter steps to form the final datasets used in further analyses (Figure S4).

Step 1. We filtered biallelic loci using Stacks according to the number of populations (*p* = minimum populations), individuals (*r* = minimum percentage of individuals in a population), and samples (*R* = minimum percentage of samples overall), selecting only the first single nucleotide polymorphism (SNP) per tag to avoid linkage between loci. We calculated the percentage of missing data for each dataset (Table S2) with VCFtools v0.1.15 [8], after which we selected four databases (pop_r80, pop10_r80, R80, and R85; Table S2) that included between ~1000 and ~105000 SNPs loci and 9.3 - 48.9% of missing data.

Step 2. To exclude low-frequency alleles and potential paralogous loci we removed sites with a minor allele frequency (*maf*) of 0.01 and 0.05. We created databases with two different *maf* thresholds as the selected cutoff can affect the population structure estimated by model-based (*e.g.*, admixture) or multivariate approaches (*e.g.*, PCA) [9].

Step 3. To remove sites with different tolerance for missing data we applied the 'min. sites' filter (0.05, 0.25, 0.50, and 0.75).

Step 4. We implemented the taxa 'min. sites' filter to remove individuals with different thresholds for missing data (0.05–0.99).

All these filters resulted in 24 datasets ranging among 1,887–33,882 SNP loci, 0.3–48% of missing data, 31–72 individuals, and four to nine species of the *humboldtianum* group (Table S3). All filters described in steps two to four were conducted using the software Tassel 5 v20210210 [10]. We removed five samples (CPUM_35317 and CPUM_35306 of *C. consocium*; CPUM_35300 and CPUM_35305 of *C. lucius*; and CPUM_10617 of *C. estor*) from the analyses because preliminary results showed their genotypes mixed with *C. jordani* (a *Chirostoma* species that does not form part of the *humboldtianum* group), presumably representing hybrid individuals. Additionally, we did not consider the individual of *C. attenuatum* (outgroup) for further analyses due to the bad quality of the genomic data recovered for that individual.

Step 5. To test the robustness of our data we conducted preliminary analyses based on 19 databases (Table S3) varying among 37–72 individuals, 4–9 nominal species, and ~2 k to ~37 k SNP loci (Figure S5). Our preliminary results were consistent regarding the number of nominal species included. Hence, we kept the 72 individuals representing the nine morphospecies of the *humboldtianum* group across the sampled localities.

Then we selected five matrices generated with different combinations of *maf* thresholds, missing data (from 7.7–15.77%), and the number of SNP loci (between 1,887 and 33,716 loci), hereafter referred to as A-33716snps, B-10517snps, C-4821snps, D-3564snps, and E-1887snps matrices.

Step 6. Finally, to estimate F_{ST} outlier analyses (see section 2.7), we separated the five matrices (A–E) by all, neutral-only, and outlier loci for a total of 15 databases that were used in downstream analyses (Figure S4). Neutral-only matrices ranged between 1,795–33,346 SNPs, while the outlier matrices contained 82–370 SNPs.

Characterization of ecotypes.

Chirostoma species in central Mexico have been categorized as 'peces blancos' or 'charales' ecotypes [1]. We considered the nominal species *C. chapalae*, *C. consocium*, *C. grandocule*, and *C. patzcuaro* as 'charales', while *C. sphyraena*, *C. lucius*, *C. promelas*, *C. humboldtianum sensu stricto*, and both subspecies of *C. estor* were categorized as 'peces blancos'. The main character used to discriminate between both ecotypes is the standard-length SL of the individuals during their adult phase (117–300 mm, 70–142 mm SL, respectively; see also Mercado-Silva *et al.,* 2015). Additional morphological (*e.g.*, jaw length, head length, snout shape, anal fin length, snout pigmentation, and size of the teeth) and meristic (number of pre-dorsal scales, number of lateral-line scales, number of gill rakers) traits are also used for the correct taxonomic identification and discrimination of ecotypes [1, 11]. Herein, all individuals were carefully identified using the morphological diagnostic characters suggested by Betancourt-Resendes *et al.* (2020).

Bayes factor delimitation analyses (BFD*).

We applied additional filters to the matrix D-3482snps-neutral_loci by retaining individuals from each of the nine morphospecies and low levels of missing data. The subsets assembled were: subset 1, comprising 39 individuals, 411 SNP loci, and 7.2% of missing data; subset 2, 39 individuals, 1102 SNP loci, and 9.1% of missing data; and subset 3, 59 individuals, 548 SNP loci, and 7.9% of missing data. We calculated Bayes factor (BF) values as 2 x (MLE of model 1 - MLE of model 2) and followed the framework provided by Kass & Raftery (1995) to assess the support of the candidate models. Because of the BF comparisons only address two models at the time (model 1 *vs.* an alternative model represented by model 2), we evaluated all possible combinations among the species delimitation scenarios. First, we considered the current taxonomy that includes nine morphospecies as model 1 (model 1 = nine morphospecies *vs.* the alternative models of three, four, and five species). To set up the priors and MCMC runs, we followed the recommendations provided by Leaché and Bouckaert (2018), where we set the birth-rate on the Yule tree prior (λ) to a gamma distribution with $\alpha = 2$ and $\beta = 200$.

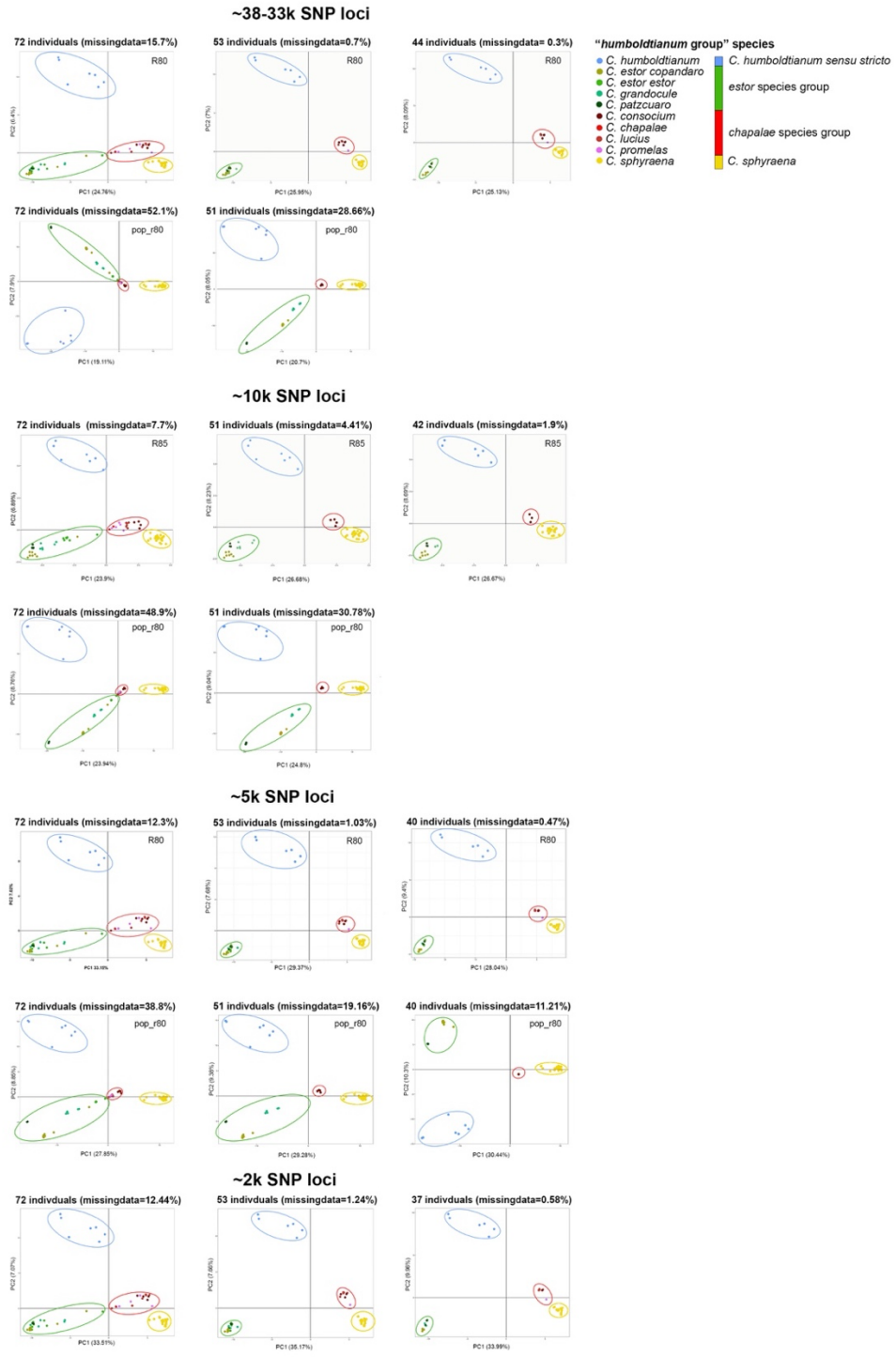


Figure S1. Principal Components Analyses 19 matrices with different numbers of SNPs loci, missing data, individuals, and species of the *humboldtianum* group.

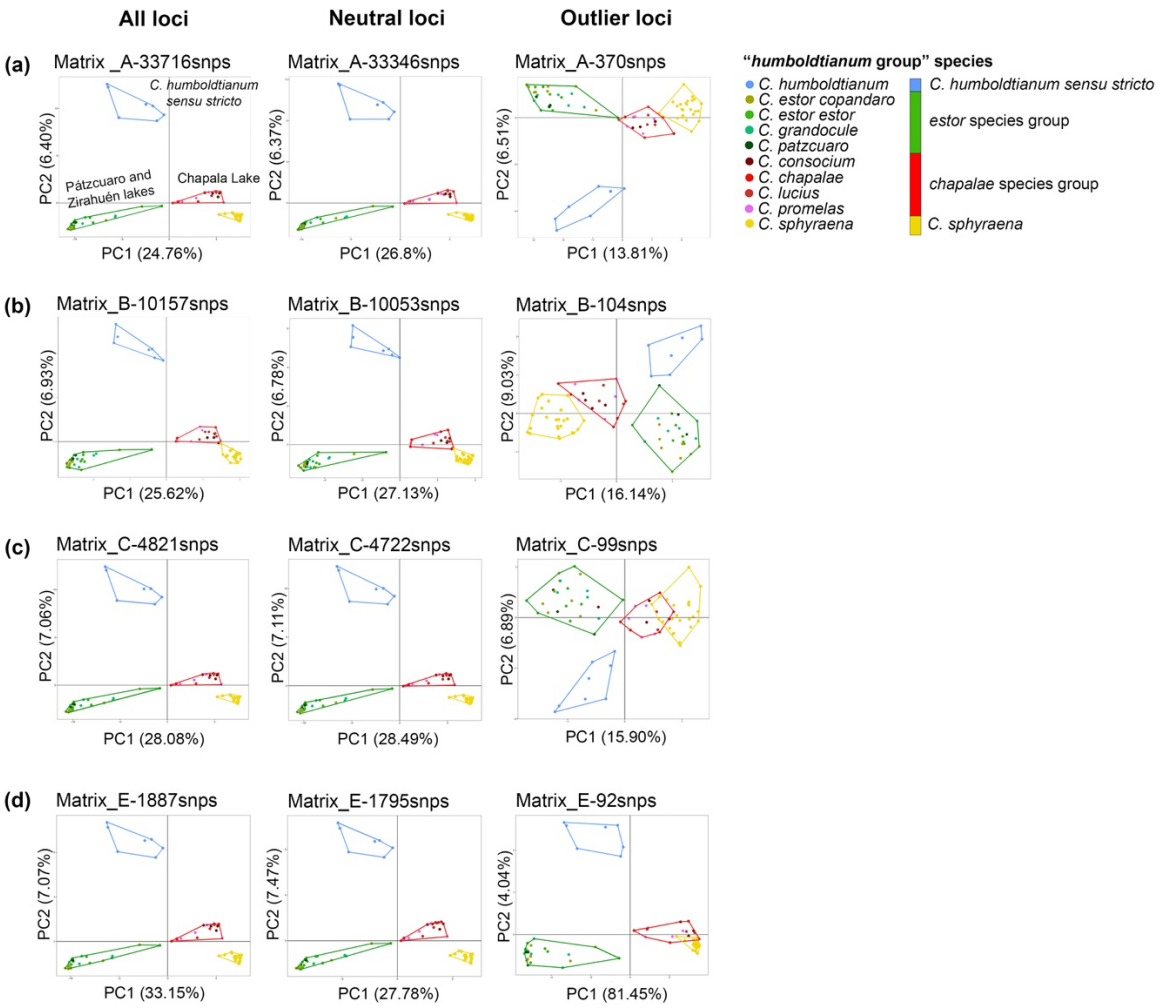


Figure S2. Principal component analyses (PCAs) based on all, neutral, and outlier SNP loci from matrices A, B, C, and E. PCAs estimated using all and neutral SNPs (~2k–33k) consistently recovered four genomic groups that are in agreement with geographic patterns but not with the previously recognized morphospecies; these are delimited with convex hulls: *humboldtianum sensu stricto* group (blue), from Lake Zacapu; *estor* group (green) from Lakes Patzcuaro and Zirahuén; *chapalae* group (red), from Lake Chapala; and *C. sphyraena* group (yellow), also from Lake Chapala. PCA analyses based on ~100–400 SNPs also resolved four genomic groups in matrices A and B, and in matrices C and E, *chapalae* and *sphyraena* groups clustered together. Morphospecies are color-coded according to the genomic groups observed.

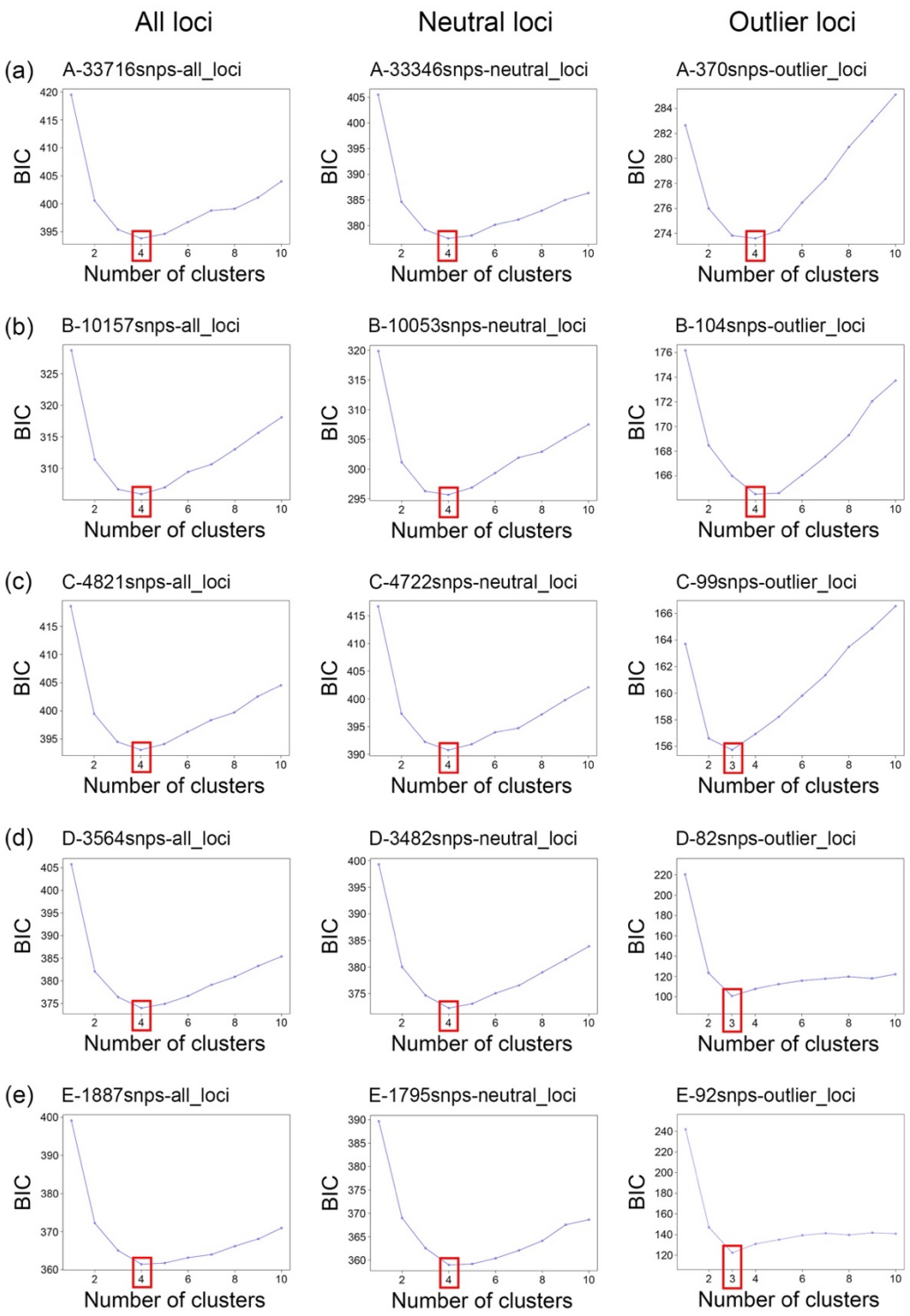


Figure S3. Plots of Bayesian Information Criterion (BIC) vs. number of clusters (*k*) of DAPC analyses for all, neutral, and outlier SNP loci from matrices A-E.

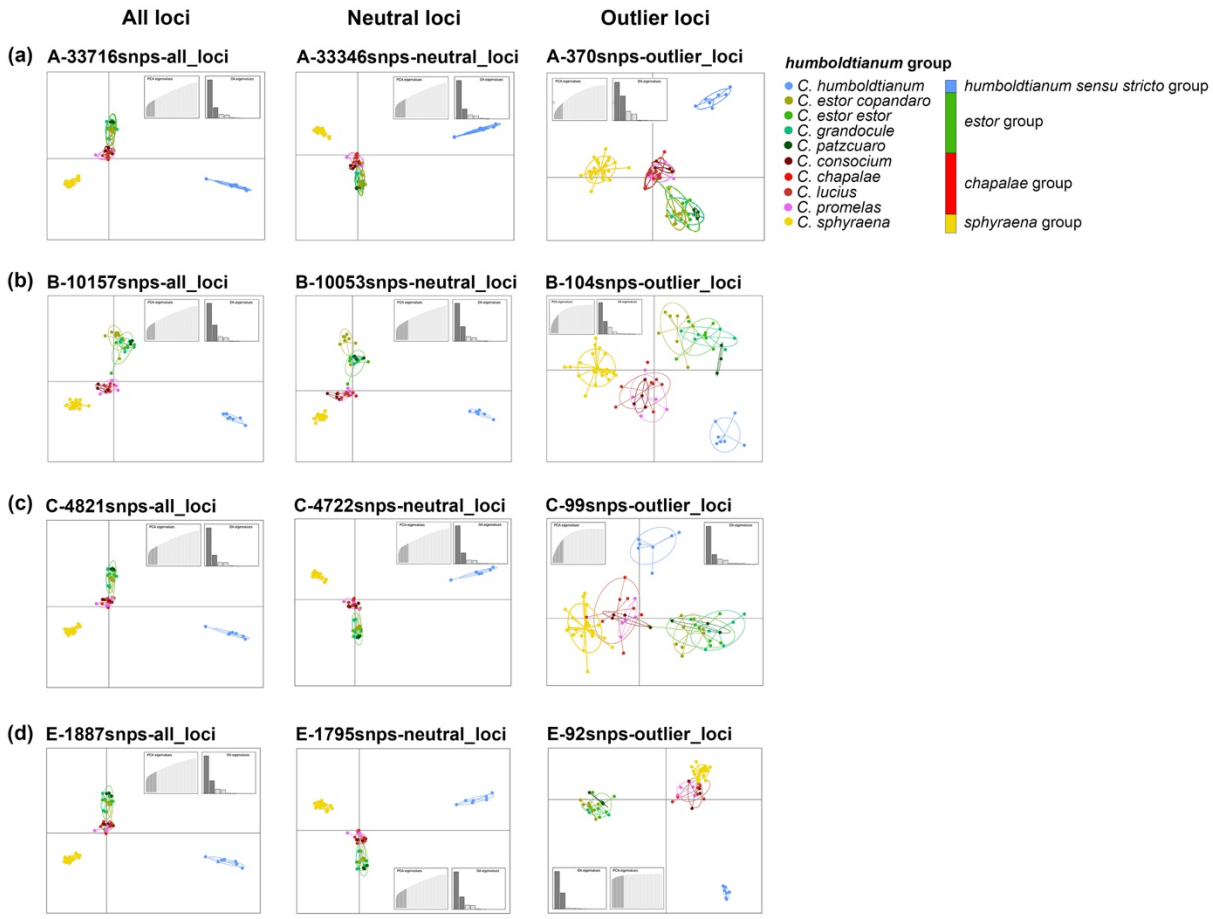


Figure S4. Discriminant analyses of principal components (DAPCs) estimated using *ca.* 1800–33700 SNPs resolve four well-differentiated clusters. DAPCs based on outlier SNPs recover three to four groups. These results are largely consistent across analyses based on matrices A, B, C, and E (a–d), and are also concordant with the PCA analyses. In neither case, the observed genomic clusters do correspond to the morphology-based species delimitation scenario. Morphospecies are color-coded according to the genomic clusters recovered: blue, *humboldtianum sensu stricto*, green, *estor* group; red, *chapalae* group; yellow, *C. sphyraena*.

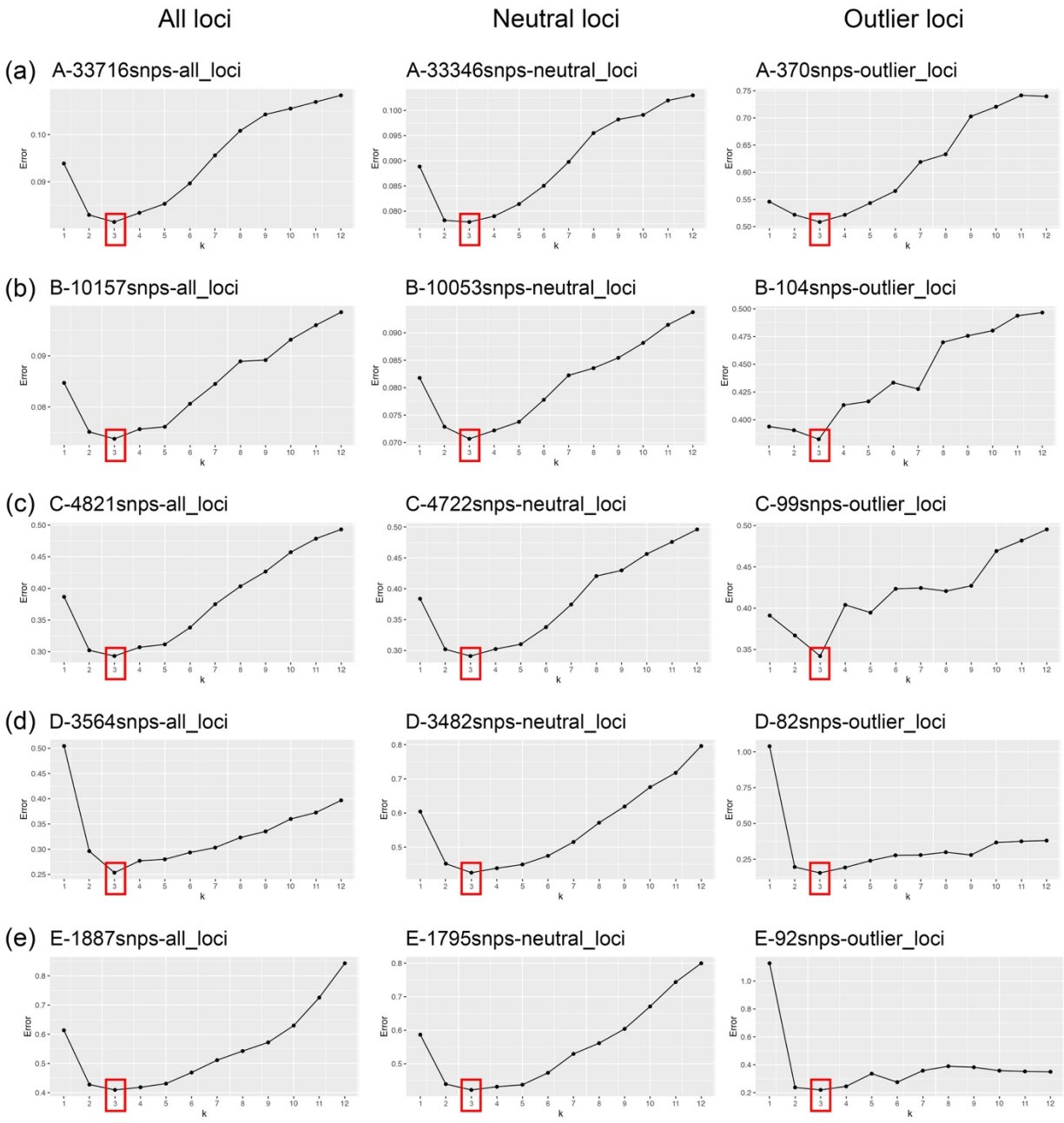


Figure S5. Plots of cross-validation error of each *k* (number of clusters) analyzed in Admixture analyses for all, neutral, and outlier SNP loci from matrices A- E. The cross-validation procedure was performed with the folds value = 5 (the default), a block relaxation algorithm as point estimation method, and the point estimation terminated with the objective function delta < 0.0001.

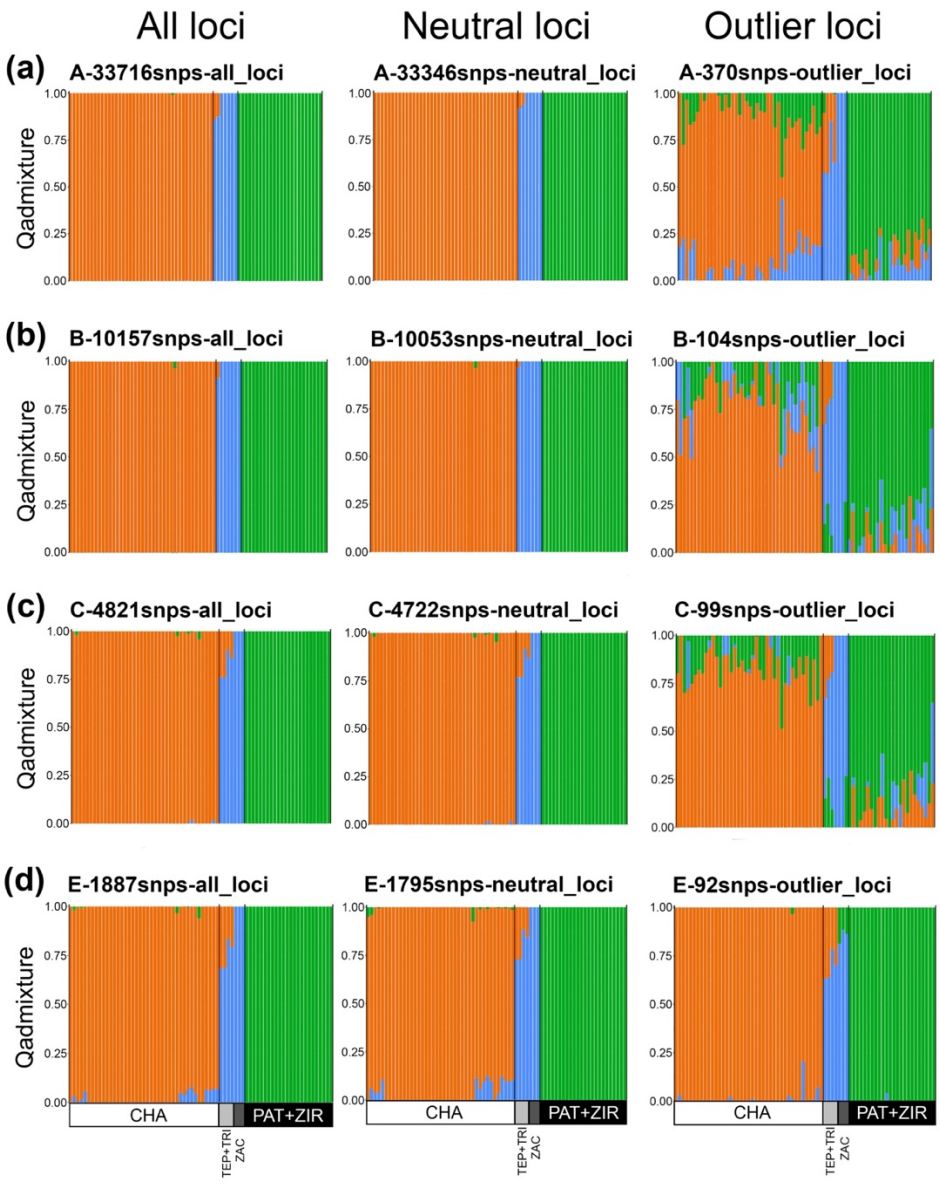


Figure S6. Admixture assignment analyses estimated using *ca.* 80–33700 neutral and outlier SNPs consistently identified three well-differentiated clusters ($k = 3$) using matrices A, B, C, and E (a–d). Outlier SNPs show an increased intermingling of individuals among clusters compared with analyses based on all or neutral loci. Each bar represents the probability of assignment to each cluster. Genomic clusters are color-coded as blue, *humboldtianum sensu stricto*, green, *estor* group; orange, *chapalae-sphyraena* group. CHA, Lake Chapala; TEP, Tepuxtepec Dam; TRI, Trinidad Fabela Dam; PAT, Lake Pátzcuaro; ZIR, Lake Zirahuén.

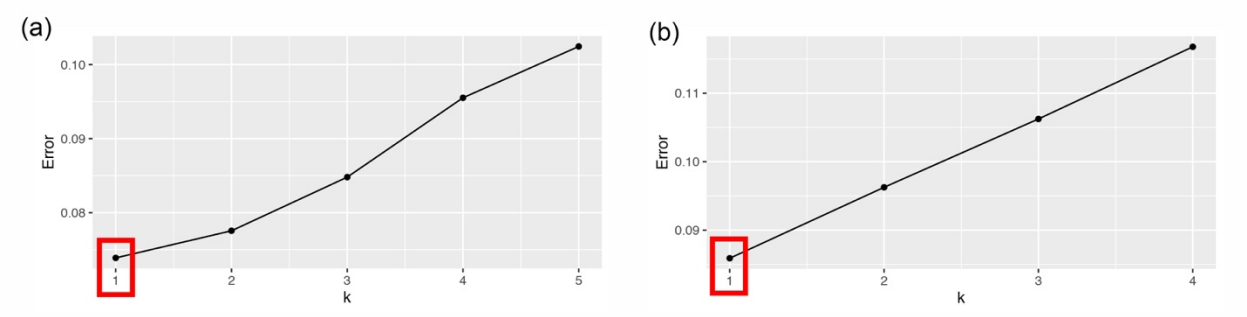


Figure S7. Admixture assignment analyses estimated using *ca.* 33700 SNP loci a) in Chapala Lake, and b) within Lakes Pátzcuaro-Zirahuén.

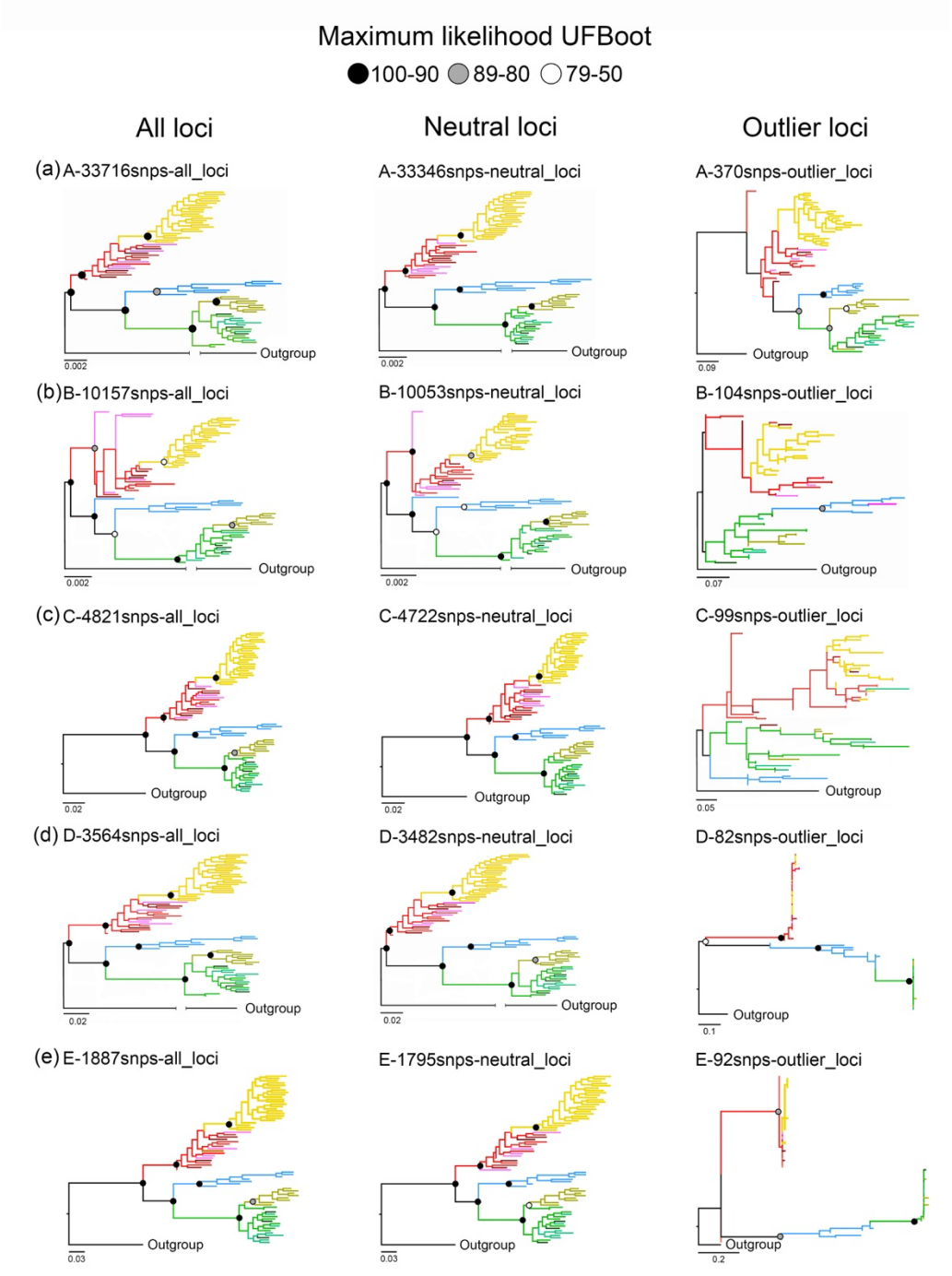


Figure S8. Phylogenetic trees of *ca.* 1800–33700 SNP loci of the *humboldtianum* group estimated under a maximum likelihood framework in IQ-TREE. Phylogenetic trees were estimated using all (~1900–33700), neutral-only (~1800–33300), and outlier (~80–350) SNPs. Individuals are color-coded by genetic clusters according to Fig. 2.

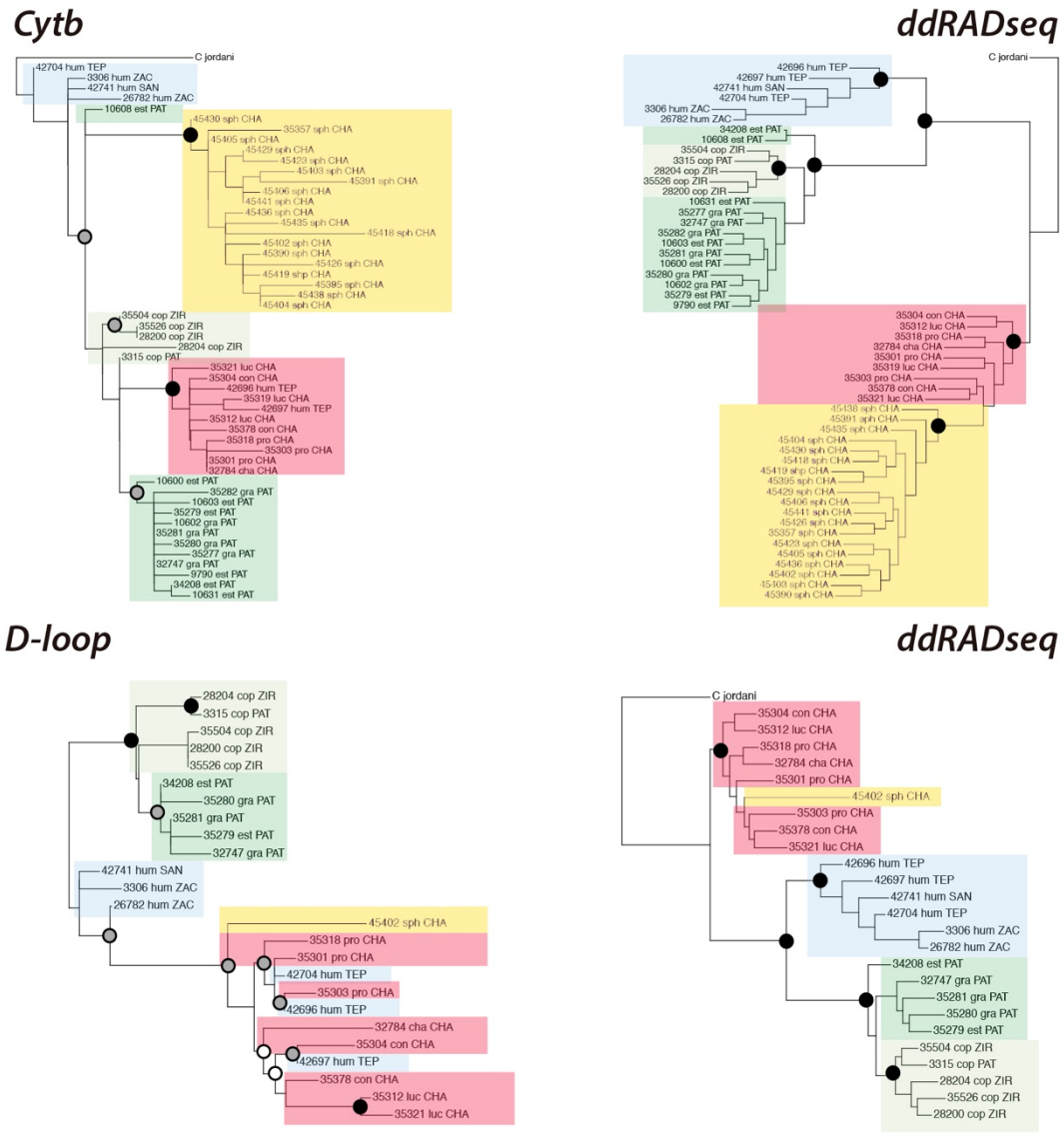


Figure S9. Mitochondrial trees (left) and ddRADseq phylogeny (right) of the *humboldtianum* group. The tips in the RADseq inferences were pruned to include the same individuals present in the mitochondrial trees.

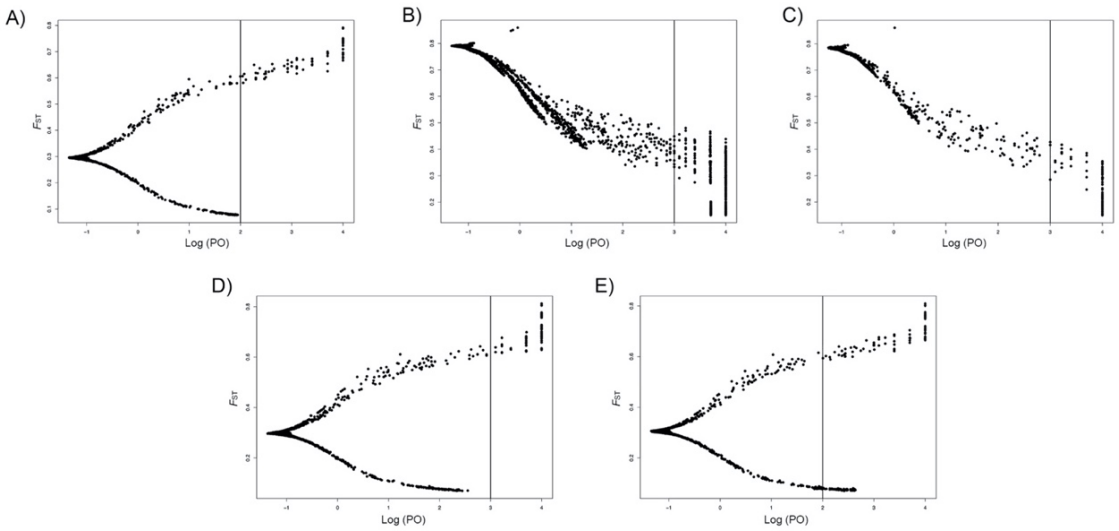


Figure S10. F_{ST} *versus* log10-transformed posterior odds (PO) values for the global outlier detection calculated in BayeScan. The analyses were estimated considering the nine morphospecies and (A) 33716 SNPs, where the vertical line represent the FDR threshold of q=0.037; (B) 10157 SNPs, q=0.040; (C) 4821 SNPs, q=0.025; (D) 3564 SNPs, q=0.034; and (E) 1887 SNPs, q=0.015.

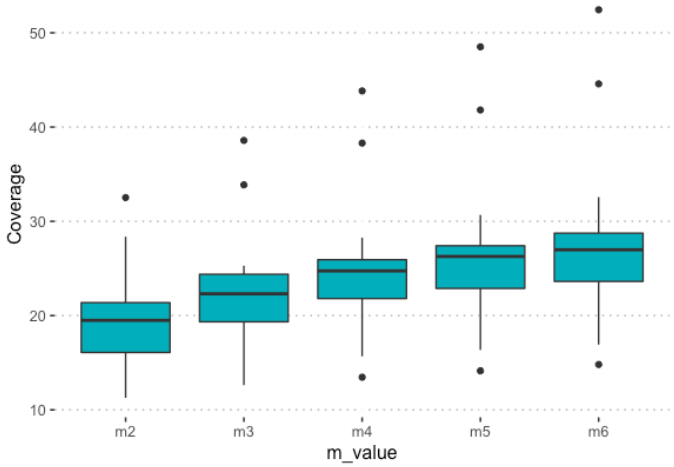


Figure S11. Mean coverage of the subset with 15 samples using different values of the minimum raw reads required to form a stack (*m*1–*m*6).

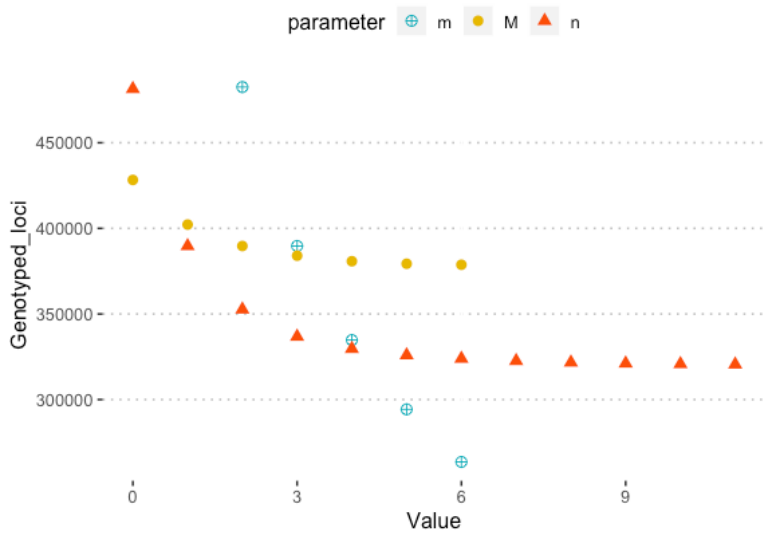


Figure S12. Putative loci at different combinations of *de novo* assembly parameters: *m*, minimum reads required to form a stack; *M*, allowed SNPs in a stack required to form a putative locus in an individual; *n*, allowed SNPs in a stack required to form a locus in the population. Each parameter was changed one at the time (*m* = 2–6, *M* = 0–6, and *n* = 0–11) while keeping the others at default values (*m*3*M*2*n*1).

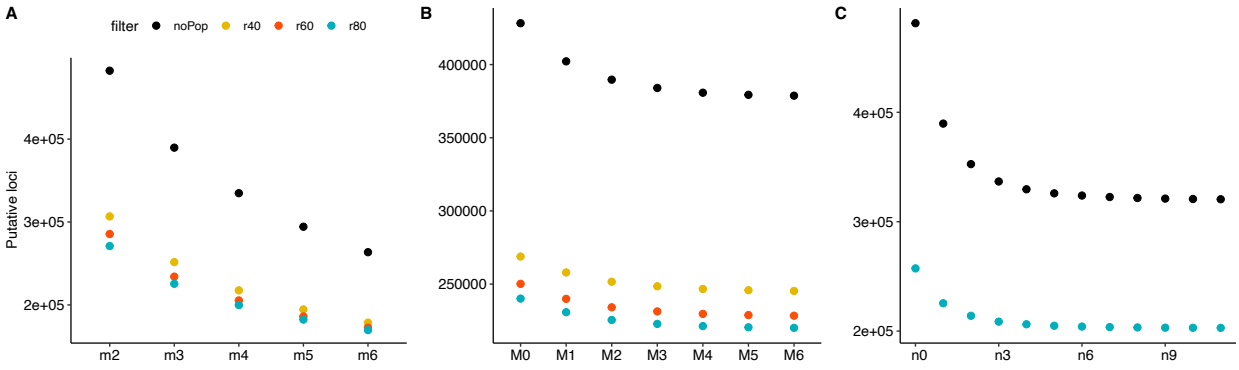


Figure S13. Genotyped loci and variant sites using a constraint on the number of minimum individuals in a population required to have that locus (r = 40, 60, and 80) based on different cutoff, as follow: (A) minimum raw reads required to form a stack (*m* = 2–6), (B) maximum mismatches allowed between stacks of the same individual (*M* = 0–6), and (C) mismatches allowed between loci of different individuals (*n* = 0–11).

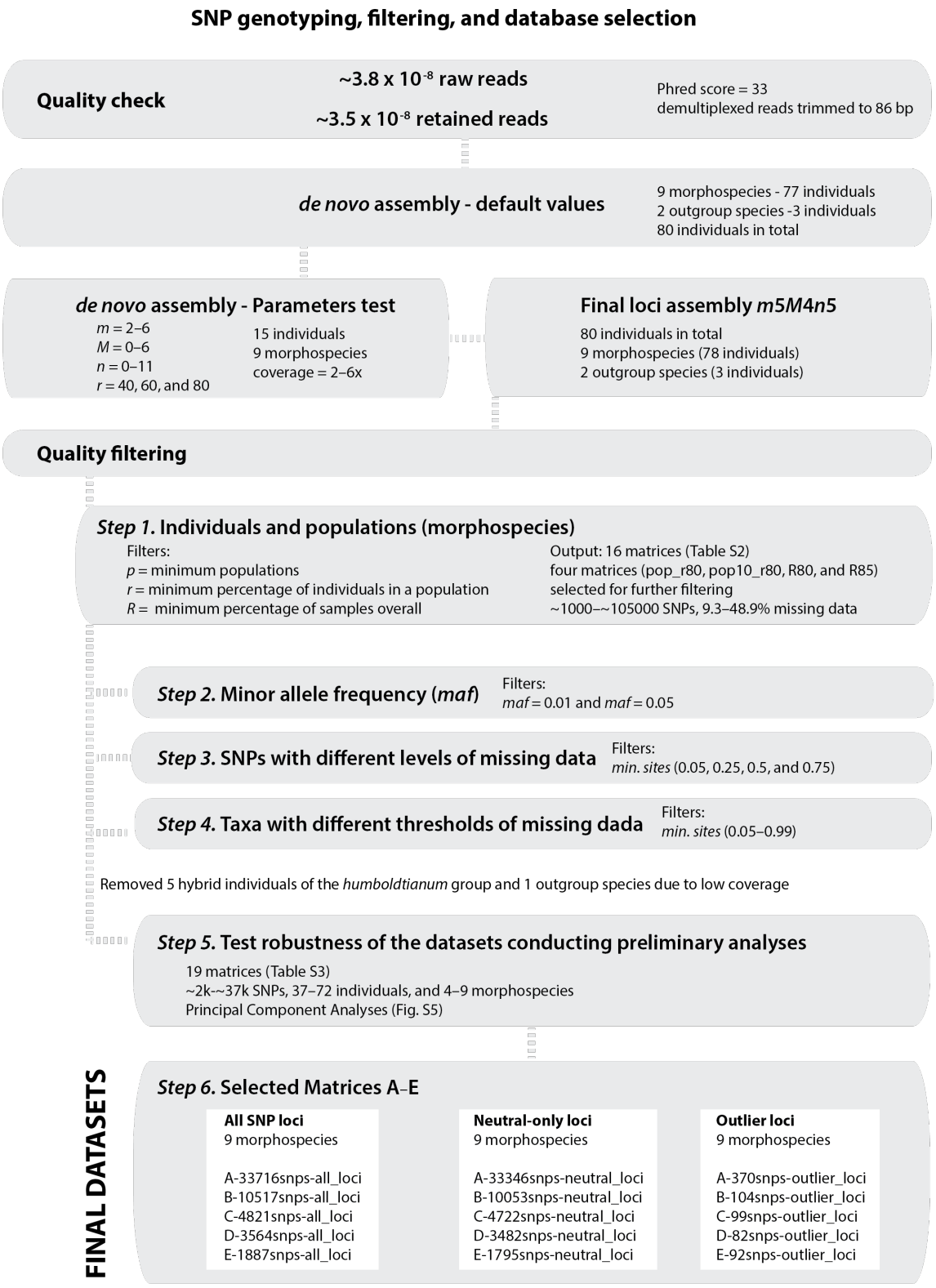


Figure S14. Flow chart of the quality control and SNP filtering steps applied to generate the final datasets.

Supplementary References

1. Barbour CD. The systematics and evolution of the genus Chirostoma Swainson (Pisces, Atherinidae). Tulane Stud Zool Bot. 1973;18:97–141. https://www.biodiversitylibrary.org/part/11912.

2. Peterson BK, Weber JN, Kay EH, Fisher HS, Hoekstra HE. Double digest RADseq: An inexpensive method for de novo SNP discovery and genotyping in model and non-model species. PLoS One. 2012;7.

3. Catchen JM, Hohenlohe PA, Bassham S, Amores A, Cresko WA. Stacks: an analysis tool set for population genomics. Mol Ecol. 2013;22:3124–40.

4. Catchen JM, Amores A, Hohenlohe P, Cresko W, Postlethwait JH. Stacks : Building and Genotyping Loci De Novo From Short-Read Sequences. Genes, Genomes. 2011;1 August:171–82.

5. Paris JR, Stevens JR, Catchen JM. Lost in parameter space: A road map for Stacks. Methods Ecol Evol. 2017;8:1360–73.

6. Mastretta-Yanes A, Arrigo N, Alvarez N, Jorgensen TH, Piñeros D, Emerson BC. Restriction site-associated DNA sequencing, genotyping error estimation and de novo assembly optimization for population genetic inference. Mol Ecol. 2015;:28–41.

7. Del Pedraza-Marrón CR, Silva R, Deeds J, Van Belleghem SM, Mastretta-Yanes A, Domínguez-Domínguez O, et al. Genomics overrules mitochondrial DNA, siding with morphology on a controversial case of species delimitation. Proc R Soc B Biol Sci. 2019;286.

8. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call format and VCFtools. Bioinformatics. 2011;27:2156–8.

9. Linck E, Battey CJ. Minor allele frequency thresholds strongly affect population structure inference with genomic data sets. Mol Ecol Resour. 2019;19:639–47.

10. Bradbury PJ, Zhang Z, Kroon DE, Casstevens TM, Ramdoss Y, Buckler ES. TASSEL: Software for association mapping of complex traits in diverse samples. Bioinformatics. 2007;23:2633–5.

11. Betancourt-Resendes I, Perez-Rodríguez R, Barriga-Sosa IDLA, Piller KR, Domínguez-Domínguez O. Phylogeographic patterns and species delimitation in the endangered silverside “humboldtianum” clade (Pisces: Atherinopsidae) in central Mexico: understanding their evolutionary history. Org Divers Evol. 2020;20:313–30.

12. Kass RE, Raftery AE. Bayes Factors. J Am Stat Assoc. 1995;90:773–95.

Appendix C

Supplementary Material for Beneath the Waves: Depth, Temperature, and Spatial Components Driving Genetic Differentiation at Micro and Macroevolutionary Scales in Tropical Blennies

Extended Methods

Extended Results

Extended Methods

Sample collection and genomic data generation

Taxonomic sampling. We collected 506 individuals from 23 out of the 25 *Malacoctenus* species across 40 localities in the Tropical Eastern Pacific (TEP) and Tropical Atlantic (TA) realms (Fig. S1). We also generated genomic data from a *Brockius striatus* specimen, used as an outgroup for phylogenetic analyses (a closer relative of the genus of *Malacoctenus* (39)). Sampling collections were conducted in various localities across Mexico, Ecuador, Costa Rica, Panama, El Salvador, Puerto Rico, and Colombia between 2012 and 2017 (Table S1). For population differentiation analyses, sampling localities were clustered under the same name using a 1 km ratio, while for seascape genomic analyses, sampling locations were considered independently, as each location can have unique environmental conditions (Appendix S2). The selection of these localities aimed to test a total of eight marine barriers to dispersal, which were proposed as genetic

breaks for marine organisms in these regions by previous studies (Table S2). We collected fin clips and voucher specimens and deposited them into the fish collections of the Universidad Michoacana de San Nicolás de Hidalgo (UMSNH), the University of Puerto Rico - Río Piedras (UPR-RP), and the Sam Noble Oklahoma Museum of Natural History (SNOMNH). We identified each individual to the species level using the taxonomic keys for *Malacoctenus* and *Brockius* (3). We also retrieved samples through collaboration with multiple ichthyological collections from various institutions (see Appendix S2).

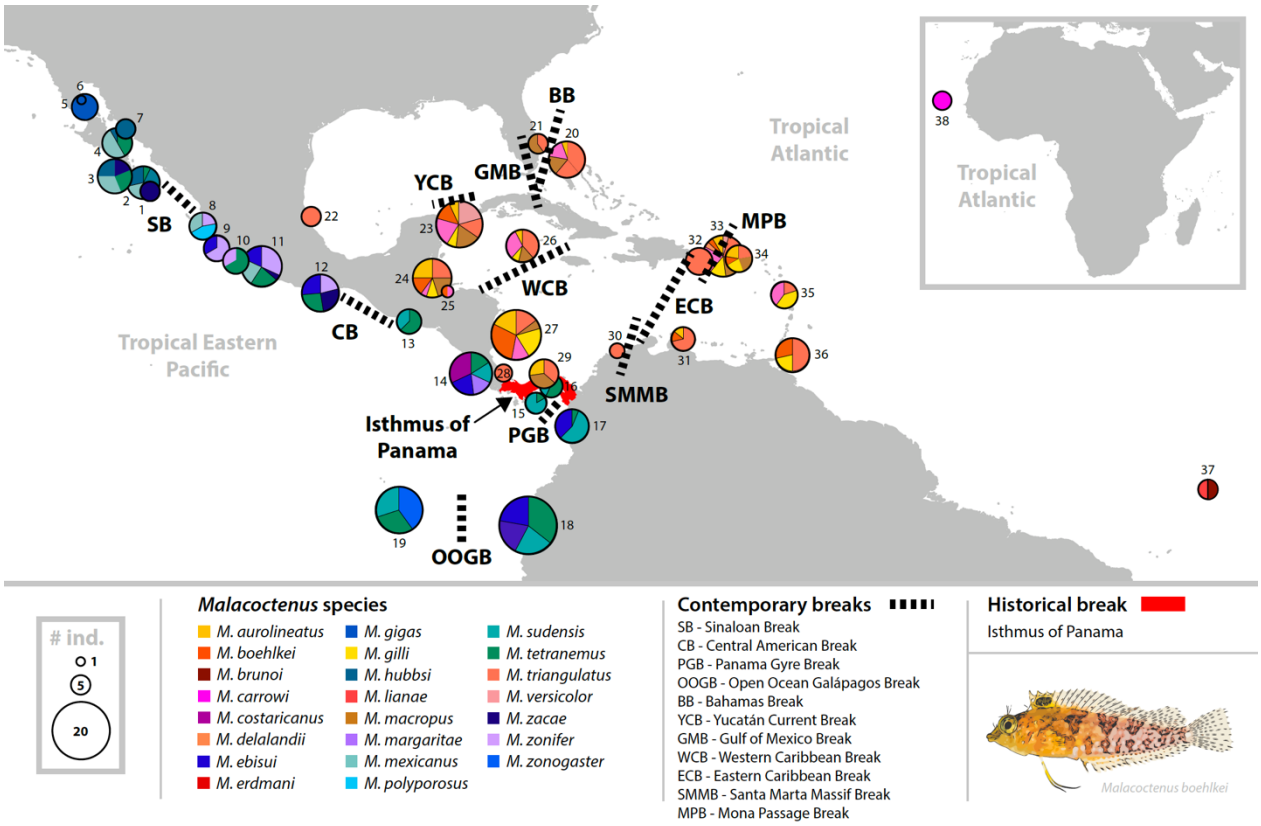


Figure S1. Locations sampled for *Malacoctenus*, including a total of 506 individuals representing 23 species, in the Tropical Eastern Pacific (TEP) and the Tropical Atlantic (TA) biogeographic regions, including Africa (which is part of TA). Numbers next to each pie chart refer to sampling locations in Table S1. Pie size is proportional to the number of individuals per location, where each color represents a species. Black dashed lines indicate contemporary barriers that were evaluated, while the historical barrier, the rising of the Isthmus of Panama, is marked in red.

Table S1. Sampling localities across the Tropical Eastern Pacific (TEP) and Tropical Atlantic (TA)

Number ID	Locality ID	Locality name	Country	Biogeographic region
1	FRA	Los Frailes	Mexico	TEP
2	TPE	El Tepetate	Mexico	TEP
3	CED	Los Cedros	Mexico	TEP
4	STI	Santa Inés	Mexico	TEP
5	SAC	El Sacramento	Mexico	TEP
6	PUE	Puertecitos	Mexico	TEP
7	SER	Seri Muerto	Mexico	TEP
8	VAL	Vallarta	Mexico	TEP
9	SAN	Santiago	Mexico	TEP
10	MIC	Michoacán	Mexico	TEP
11	ZAC	Zacatoso	Mexico	TEP
12	AGU	Agustinillo	Mexico	TEP
13	COB	Los Cóbanos	El Salvador	TEP
14	CBL	Cabo Blanco	Costa Rica	TEP
15	PED	Pedasí	Panama	TEP
16	PER	Las Perlas Archipelago	Panama	TEP
17	NUQ	Nuquí	Colombia	TEP
18	ECC	Continental Ecuador	Ecuador	TEP
19	GAL	Galápagos Archipelago	Ecuador	TEP
20	BAH	The Bahamas	The Bahamas	TA
21	FLO	Florida	USA	TA
22	VER	Veracruz	Mexico	TA
23	YUC	Yucatán	Mexico	TA
24	BLZ	Belize	Belize	TA
25	HON	Utila	Honduras	TA
26	CAY	Cayman Islands	United Kingdom	TA
27	AND	San Andrés	Colombia	TA
28	LIM	Limón	Costa Rica	TA
29	PAN	Panama	Panama	TA
30	TAY	Tayrona	Colombia	TA
31	CUR	Curaçao	Netherlands	TA
32	MON	Mona Island	USA	TA
33	PRI	Puerto Rico	USA	TA
34	VIR	Virgin Islands	USA	TA
35	DOM	Dominica	Dominica	TA
36	TRT	Trinidad and Tobago	Trinidad and Tobago	TA
37	NOR	Fernando da Noronha	Brazil	TA
38	CVE	Cape Verde	Cape Verde	TA

Table S2. Biogeographic breaks tested in Tropical Eastern Pacific (TEP) and Tropical Atlantic (TA) regions

Biogeographic break	Abbreviation	Marine organisms that are separated by the break
Tropical Eastern Pacific		
Sinaloan Break	SB	Pacific Sierra mackerel (*Scomberomorus sierra* (40))
Central American Break	CB	Gobies (*Elacatinus puncticulatus* (7)), snails (*Nerita scabricosta* and *N. funiculata* (41))
Panama Gyre Break	PGB	Goby (*Elacatinus puncticulatus* (7))
Tropical Atlantic		
Bahamas Break	BB	Gobies (*Elacatinus evelynae* (42)*; Elacatinus louisae* (43))
Gulf of Mexico Break	GMB	Lionfish (*Pterois volitans* and *P. miles* (6))
Yucatán Current Break	YCB	*Haemulon aurolineatum* (40)
Western Caribbean Break	WCB	Invasive red lionfish (*Pterois volitans* (44))
Eastern Caribbean Break	ECB	Corals (*Orbicella faveolata* (45), *Madracis auretenra* (46))
Northeastern Colombian Coast	NECC	Sea Catfishes (*Cathorops mapale/C. wayuu* (47))
Mona Passage Break	MPB	Gobies (*Elacatinus* spp (42, 48)), corals (*Acropora palmata* (49))

DNA extraction, library preparation, and sequencing. High-quality DNA extractions were performed from fin-clip tissues using Qiagen DNeasy mericon 96 QIAcube HT Kit (Qiagen, Inc.) at the University of Wisconsin Biotechnology Center (UWBC). We generated data for 506 individuals representing the 23 *Malacoctenus* species, and replicates of eight individuals that were used to measure the assembly-error rate (see below). Replicates were chosen based on the species with fewer individuals to increase the chance of getting a higher final coverage. To identify potential instances of cross-contamination, the well positions for each DNA extraction plate were carefully chosen to avoid placing individuals of the same species or closely related populations consecutively. Double-digest restriction-site associated DNA (ddRADseq) libraries were prepared using the protocol of (13), which is appropriate for species with high levels of genetic diversity. Libraries were prepared with the restriction enzymes *PstI* and *BfaI*, and a size selection window of 350-550 bp. The enzyme assessment phase involved the preparation of four double digested libraries from a pool of two samples representative of the genus. The enzyme combinations used were PstI/MspI, NsiI/MpsI, PstI/BfaI, and NsiI/BfaI. The optimal enzyme combination was selected based on the absence of visible repeated regions within the size selection area and the success of amplification. Individuals were indexed by a set of dual 10-base-pair barcodes in combinatorial schemes. Finally, the RADseq libraries were divided into six pools and sequenced at Novogene Company Inc. (CA, USA) using a partial lane of Illumina NovaSeq 6000 PE150 sequencing (output: 470 GB).

PE-ddRADseq quality control and filtering. Raw data was checked with FastQC v0.11.15 (www.bioinformatics.babraham.ac.uk/projects/fastqc/) to assess the average Phred score of each nucleotide position, and check for the presence of artifacts such as sequencing adaptors (50). Quality filtering and demultiplexing were performed using the *process_radtags* tool, which is part of the Stacks v2.59 (15) pipeline, a package designed to analyze short-reads generated via RADseq. Raw sequencing reads were filtered using a Phred score threshold of 33, discarding reads with low quality scores (*-q*), and removing any read with an uncalled base (*-c*). Demultiplexed raw sequences per individual were deposited at NCBI Sequence Read Archive (SRA) (www.ncbi.nlm.nih.gov/sra).

Species ID quality control. We identified the taxonomic identity of the species using a taxonomic key for the species of *Malacoctenus* (3). We corroborated the taxonomic identification by comparing the mitochondrial markers cytochrome oxidase subunit I (*COI*) and cytochrome b (*Cytb*) of each individual, with reference sequences available at the Barcode of Life Data (BOLD) system (www.boldsystems.org) and the NCBI data archive (www.ncbi.nlm.nih.gov). To this end, we first compiled a reference file that included all available *COI* and *Cytb* information for *Malacoctenus* from these repositories. Then, we mined each molecular marker from the demultiplexed raw data of each individual using custom scripts and the software BWA v0.7.17 (51) and SAMtools v1.11 (52). We retrieved 169 hits (140 bp) for *COI* for 17 species, and 321 hits (141 bp) for *Cytb* for 18 species (Appendix S2). Despite our efforts to extract the nuclear markers rhodopsin (*Rhod*) and recombination activating protein 1 (*Rag1*), the hits we recovered were uninformative and therefore were excluded from further analyses. To provide a final corroboration of the species' identity and to rule out any potential effects of cross contamination in RADseq demultiplexed sequences, we conducted a *de novo* assembly using default parameters per plate, generated matrices that included polymorphic RAD loci found in 25% of the samples and estimated a phylogenetic tree (refer to the Locus assembly, and Phylogenetic Inference sections for further details). Finally, we carefully inspected any individual that clustered with different species in the phylogenetic trees by reviewing field photos and voucher specimens to ensure their correct identification.

MATRIX ASSEMBLY FOR EVOLUTIONARY ANALYSES AT DIFFERENT SCALES

SNP Genotyping. Restriction site-associated DNA sequencing (RADseq) is a next-generation sequencing (NGS) technique that uses restriction enzymes to digest DNA into fragments of a constant size across the genome (53, 54). Recently, the use of RADseq approaches has become increasingly popular in population genomic and phylogeographic studies, as they provide the opportunity to genotype thousands of genome-wide markers across many individuals in non-model systems at a reasonable cost (55–57). Therefore, RADseq data represents an ideal fit for our study, which aims to examine contemporary and historical barriers that have shaped the genetic diversity of multiple species across a shared geographic region, at both micro- and macro-evolutionary scales. However, this technique is prone to several sources of sequencing error that may affect population inferences (58). These sequencing errors may arise from laboratory procedures (e.g., library development issues due to degraded DNA) or species-specific genome properties (e.g., alleles remain unsampled due to a polymorphism at the enzyme's cut-site—allele dropout) (58–61). Additionally, locus assembly from RADseq data is challenging as the selection of main parameters that control *de novo* assembly results in different levels of assembly-related error and can greatly influence genomic variation per locus (58, 62). Nonetheless, the proportion of artifactual loci can be detected due to errors during the PCR amplification steps (63), while parameter optimization and proper filtering (see below) can improve the quality of the recovered loci (61).

In this study, we used a three-step approach to reduce the recovery of false-positive loci resulting from bioinformatic assembly-related errors. First, we conducted a parameter optimization

test on a subset of individuals to select the parameters that best fit the data, identifying "default", "optimal", and "recommended" parameter combinations. Second, we measured the assembly-error rate on the suggested parameter combinations using replicates. Third, we applied several filters to each matrix to reduce missing data, filter paralogs, which are genes related to duplication events in the genome, and identify outlier loci (61) (see below for more details). As we were interested in analyzing the genetic structure at both intra- and inter-specific levels, we assembled SNP matrices in two different ways. The first pipeline aimed to capture the genetic diversity among populations for each of the species analyzed, and the second aimed to identify SNPs common to all 23 *Malacoctenus* species and *L. striatus*. In both pipelines, we applied the three-step approach to generate the final matrices that were used to conduct population differentiation and phylogenomic analyses (Fig. S2). At the macroevolutionary level, we also assembled the data using the closest genome available (see below for more details).

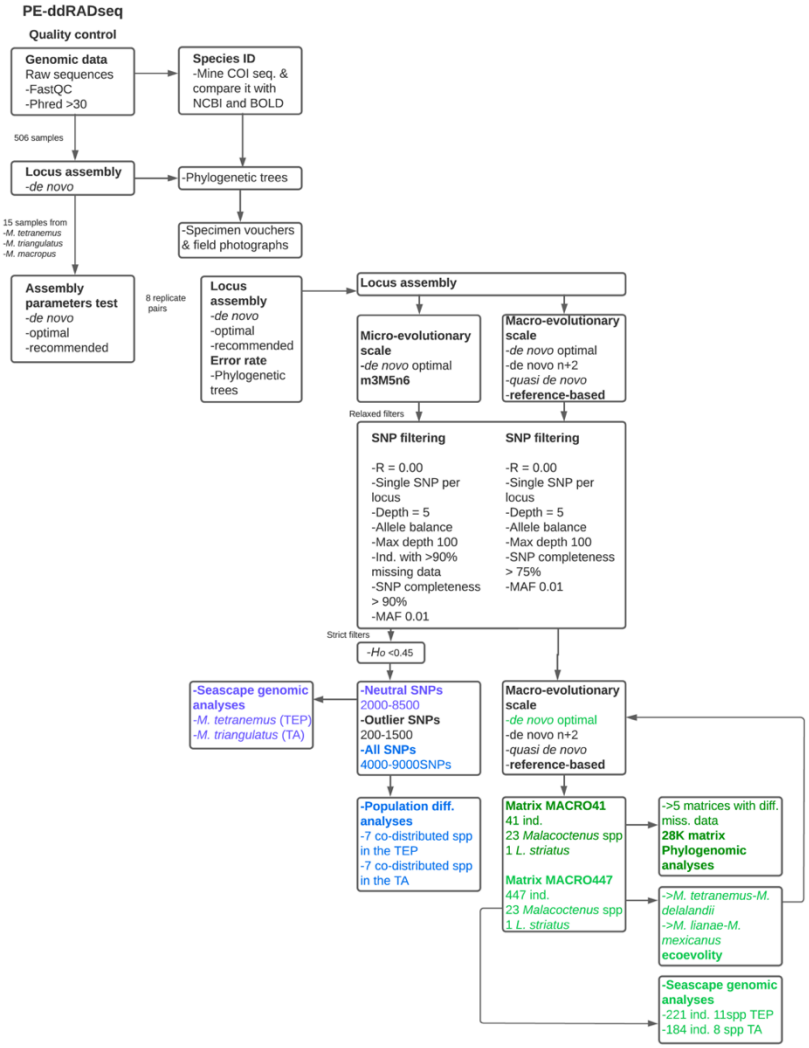


Figure S2. Scheme illustrating the pipeline used to produce genomic matrices for analyses. This pipeline includes quality control, locus assembly, and SNP filtering steps employed in this study, both at the micro- and macroevolutionary scales.

Locus assembly at microevolutionary scales

***1) De novo* assembly optimization.** Assembly of RADseq data in the absence of a reference genome is a daunting task, as the parameter values applied greatly influence the level of locus polymorphism allowed in each dataset (58, 62). The main parameters that control locus assembly and polymorphism within a locus are the minimum depth of coverage (-*m*), the distance allowed between stacks (-*M*), and the distance allowed between catalog loci (-*n*) (61, 64). Selecting these parameters to high values may collapse loci leading to an artificial effect of higher heterozygosity. Conversely, setting them too low may result in a higher proportion of homozygotes being falsely produced (58, 62). Therefore, selecting optimal parameters requires a careful balance between accommodating genetic variation and sequencing errors, while also being stringent enough to reduce false positive or paralogous loci (61). Optimal parameters are study-specific, as the level of genetic diversity present in each raw sequencing data varies across taxa, and the amount of sequencing error depends on the properties of each dataset (61, 64). To optimize our RADseq data assembly, we followed recent recommendations (55, 61, 65) to tackle this using Stacks. First, we conducted an initial *de novo* assembly of putative loci on the 506 individuals using default values. Next, we selected the 15 individuals with the highest coverage after the initial *de novo* test from four *Malacoctenus* species (*M. tetranemus*, *M. zacae* from the TEP, and *M. macropus*, and *M. triangulatus* from the TA) to represented phylogenetic diversity (based on the *mtDNA* phylogenetic tree, Fig. S37). Because this approach was computationally demanding as some individuals had very high coverage values, the assembly of *M. zacae* was excluded from further parametrization tests. For the three remaining species we selected parameter combinations using three different methods:

> ***1.1)*** **Default.** We used default values (*m*3*M*2*n*1) from Stacks which are suitable for population genomic analyses (62, 66).
>
> ***1.2)*** **Optimal.** We conducted parameter optimization tests based on (based on 62, 63, 66, 68). These tests consisted of assembling the data multiple times, varying one parameter at the time (-*m* = 3–6, -*M* = 2–8, -*n* = 0–11), while keeping the other parameters fixed at their default values (*m*3*M*2*n*1). For each run, we recorded the number of assembled loci, the number of polymorphic loci, and the number of SNPs generated (Fig. S3). In addition, we collected data on individual coverage while varying the -*m* parameter (Fig. S4), as setting it too high may result on dropping low-coverage alleles, while extremely low values may falsely validate sequencing errors as putative loci (61). To account for biological differences in polymorphisms and sequencing read depths (68), we varied the minimum percentage of individuals in a population required to process a locus (-*r* = 40, 60, and 80) and recorded the amount of putative loci (Fig. S3). Additionally, since -*M* parameter controls the collapsing of alleles from the same locus, increasing this parameter would be expected to stabilize the number of loci found, with any newly added polymorphic loci representing paralogous loci (62). Therefore, we measured the increment polymorphic loci (with -*r* set to 80) and recorded the corresponding difference in putative loci at each increment. We selected values for -*m*, -*M*, and -*n* that resulted in a stable number of putative loci, as well as values of -*M* that yield low numbers of newly recovered loci. Based on

these analyses, the combination of parameters -*m* = 5, -*M* = 6, and -*n* = 6 were deemed as "optimal".

1.3) **Recommended.** This approach consisted in using the R packages RADstackhelpR (https://devonderaad.github.io/RADstackshelpR/index.html) and *vcfR* (69) to optimize the *de novo* assembly pipeline. These packages offer an easy-to-follow and well-documented pipeline that incorporates the recommendations of (61) for selecting optimal parameter values. We evaluated each parameter using the same range of values as the optimal step (-*m* = 3–6, -*M* = 2–8, -*n* = 0–11). We followed the tutorial available at devonderaad.github.io/RADstackshelpR, which involved testing the parameters -*m*, -*M*, and -*n* in that order. The selected parameter combinations were *m*4*M*6*n*7 for *M. tetranemus* (Fig. S5), *m*4*M*5*n*6 for *M. triangulatus* (Fig. S6), and *m*5*M*8*n*8 for *M. macropus* (Fig. S7), which will be referred to as "recommended" hereafter.

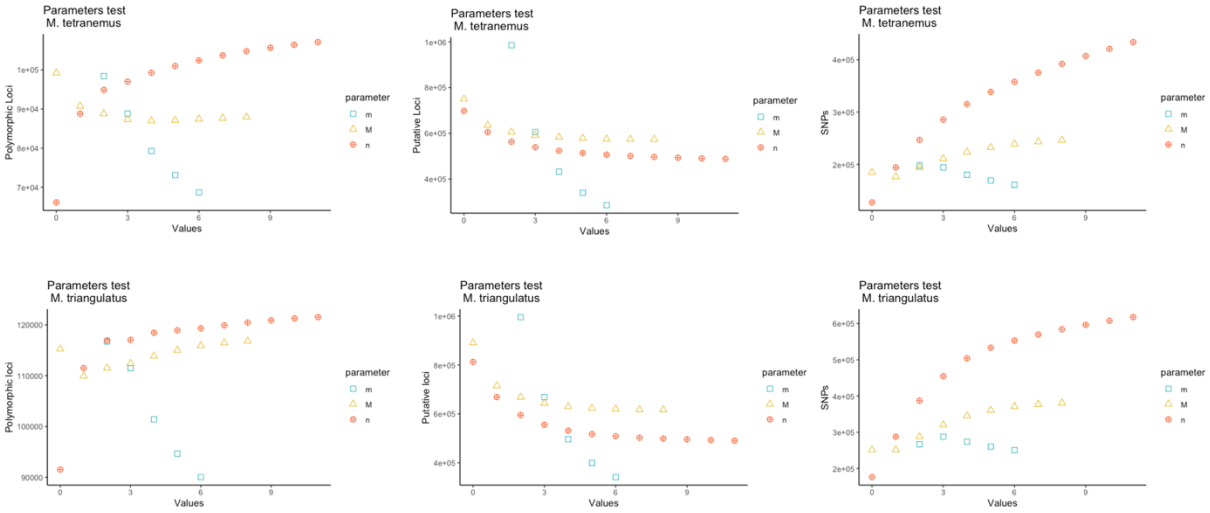


Figure S3. Parameter optimization varying one main parameter at the time for A) *M. tetranemus*, B) *M. triangulatus*, and C) *M. macropus*.

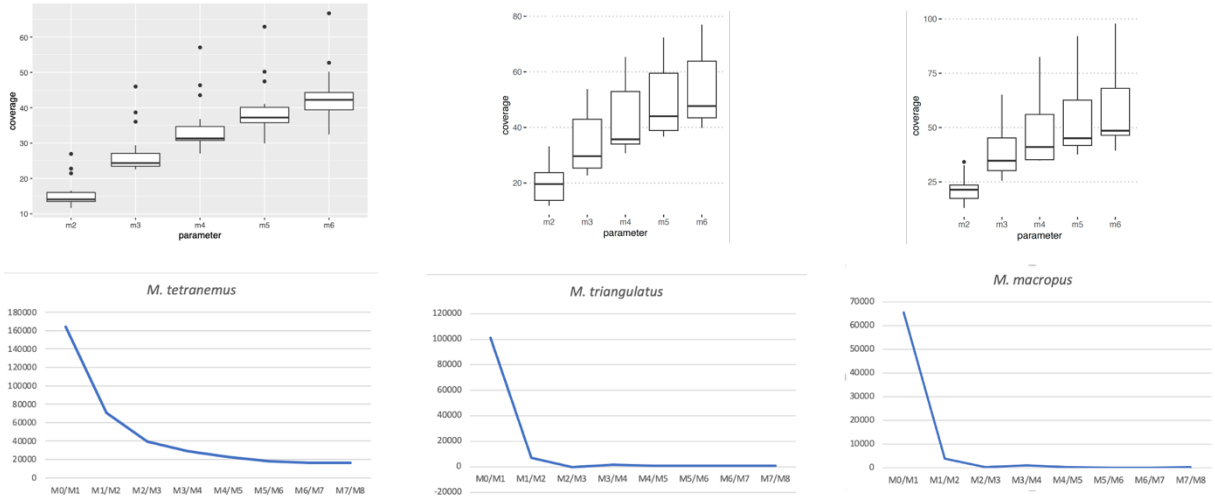


Figure S4. A) coverage of the 15 samples of *M. tetranemus*, *M. triangulatus*, and *M. macropus* during parameter optimization. B) polymorphic loci at each increment of the M parameter.

It is worth noting that gapped alignments (--gapped) were not included in any of the parametrization tests. While this flag can potentially increase the number of loci for analysis, it also enhances the risk of creating complex and putatively error-containing, loci (61).

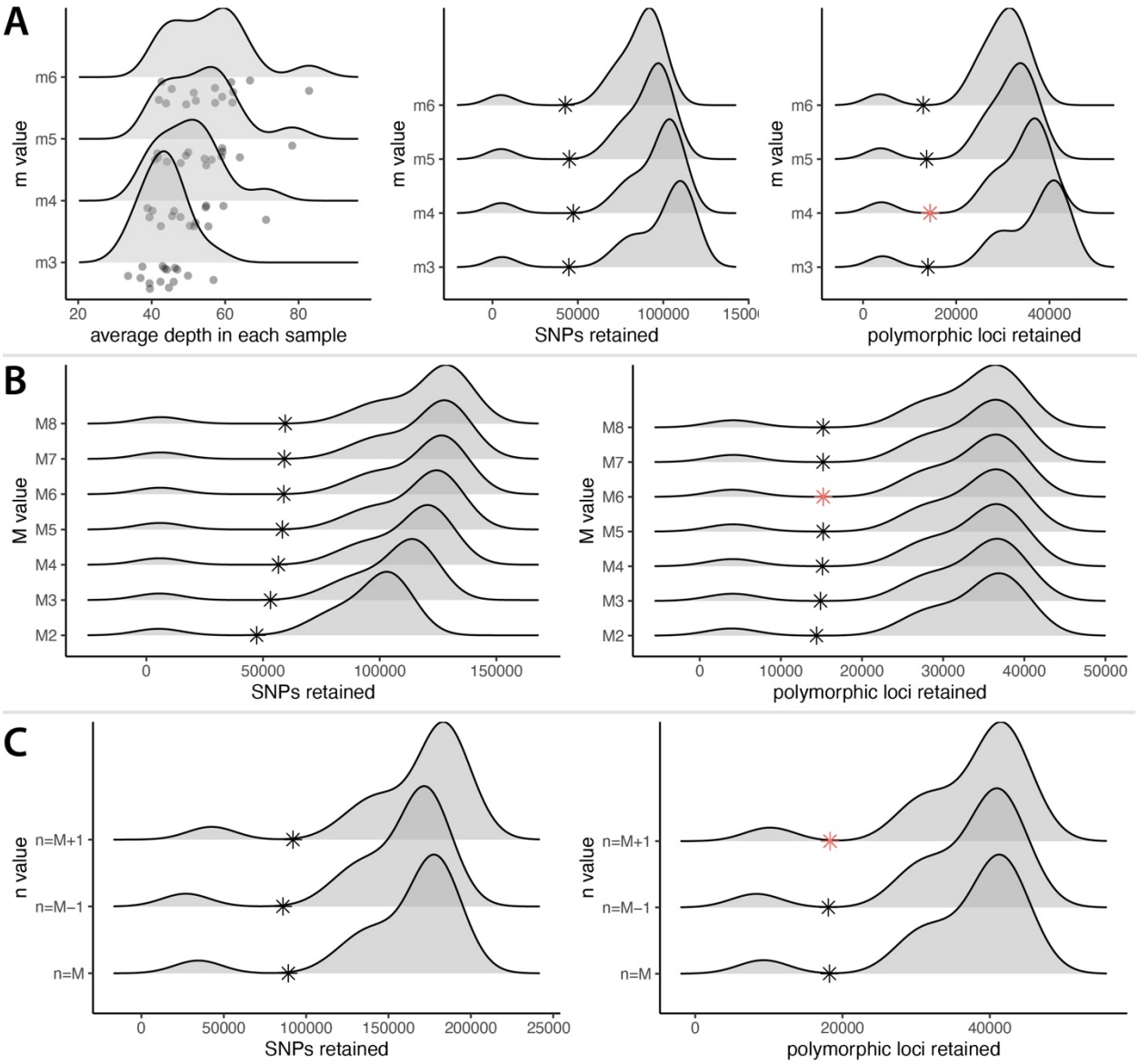


Figure S5. Utilization of RADstackshelpR for parameter optimization in *M. tetranemus*. The recommended assembly parameters include: A) setting a minimum of four raw reads required to form a stack (*m*); B) allowing a maximum of six mismatches between loci (*M*); and C) accommodating seven mismatches between loci of different individuals (*n*).

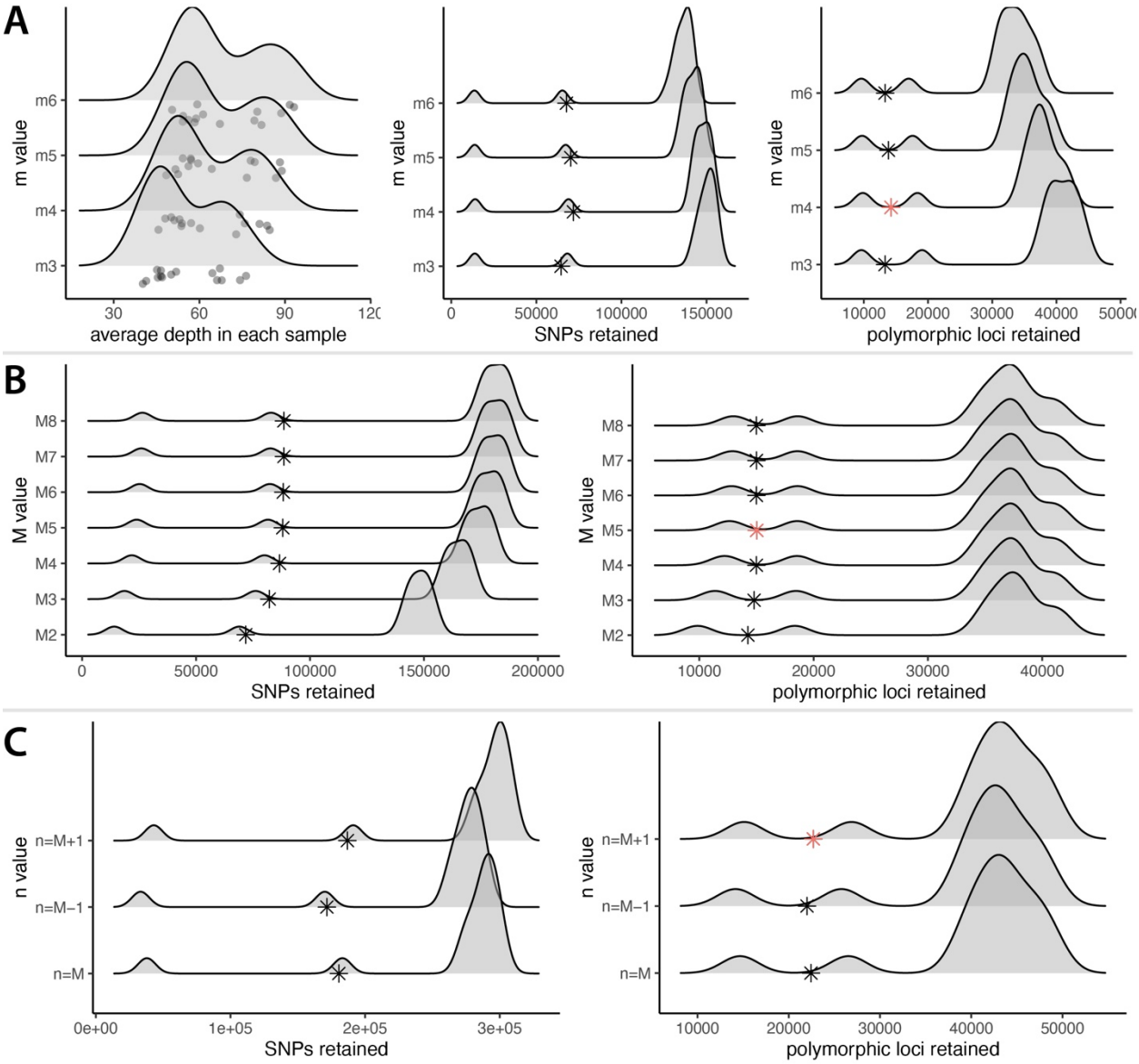


Figure S6. Utilization of RADstackshelpR for parameter optimization in *M. triangulatus*. The recommended assembly parameters include: A) setting a minimum of four raw reads required to form a stack (*m*); B) allowing a maximum of five mismatches between loci (*M*); and C) accommodating six mismatches between loci for different individuals (*n*).

2) **Use of replicates to minimize error rates.** We assembled *de novo* eight replicate-pairs using the "default", "optimal", and "recommended" parameter combinations to generate SNP matrices. We filtered each matrix to include only SNPs present in at least 25% of the individuals, reducing the amount of missing data. We then estimated phylogenetic trees (see below) for each matrix and measured the branch lengths between replicate pairs (Fig. S8A) as a proxy for assembly error. As replicate pairs present the same DNA, it is expected that identical genotypes would result in minimum differences (58, 62). This process allowed us to examine the extent to which genotypes obtained from two replicates of the same individual match, and how the parameters employed in

de novo SNP calling influence this similarity (70). Finally, we selected the "optimal" parameter combination (*m*3*M*5*n*6) as it minimized assembly error (Fig. S8B). Hence, the final *de novo* assemblies of complete datasets per species were performed using a minimum of five raw reads required to form a stack (*m*), allowing a maximum of six mismatches between loci (*M*), and six mismatches between loci of different individuals (*n*).

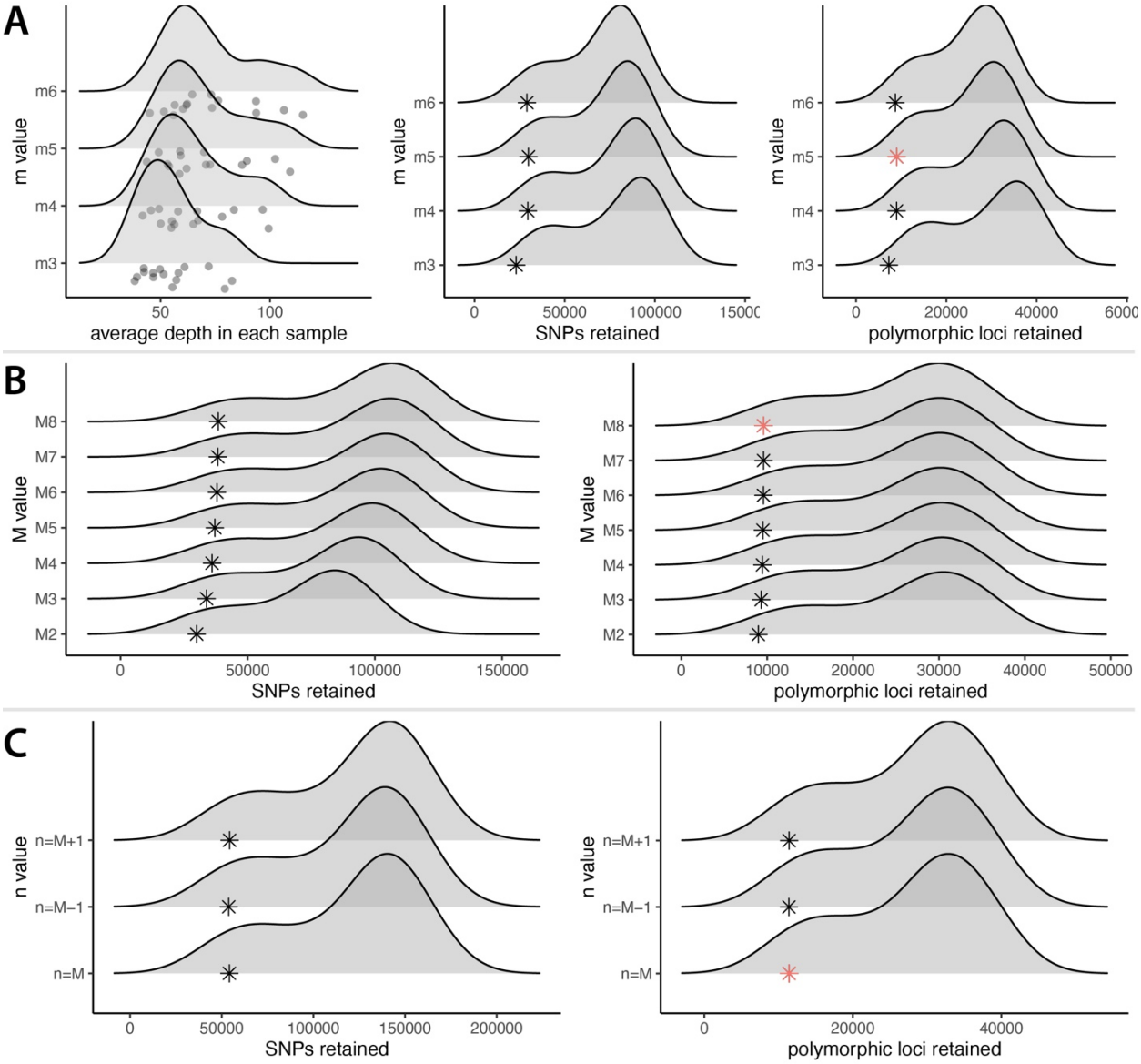


Figure S7. Utilization of RADstackshelpR for parameter optimization in *M. macropus*. The recommended assembly parameters include: A) setting a minimum of five raw reads required to form a stack (*m*); B) allowing a maximum of eight mismatches between loci (*M*); and C) accommodating eight mismatches between loci of different individuals (*n*).

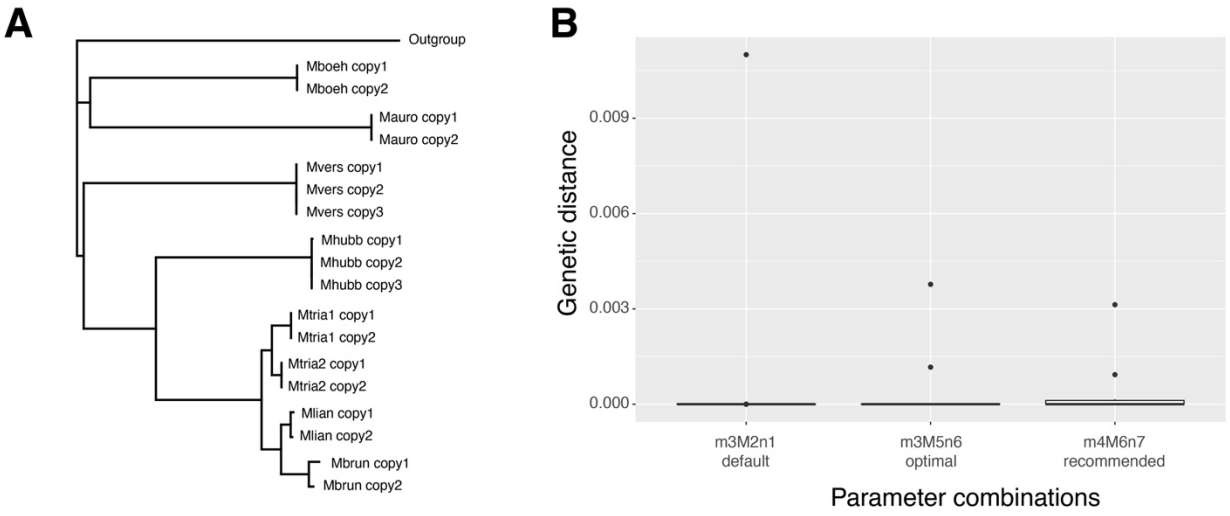


Figure S8. A) Phylogenetic tree depicting replicates. B) Genetic distance among replicates, comparing optimal and recommended assemblies.

3) **SNP Filtering.** Because artefacts introduced during library preparation and bioinformatic processing of RADseq data may influence downstream population genetic inferences (60, 71), we implemented a rigorous SNP filtering pipeline. This involved comparing the effects of different filter combinations (*e.g.*, missing data, minor allele frequency [MAF], and coverage) using the R packages SNPfiltR (16, 69) and VCFtools v0.1.16 (17) to generate SNP matrices. We generated species-specific matrices in Stacks that contained all putative loci (-*R* 0.00) and selected the first SNP of each locus (--*write_single_snp*), to meet the requirement of unlinked SNPs for some population-level analyses, such as Admixture (see below). We used SNPfiltR to retain sites with a minimum depth of five and a minimum quality Phred score of 30. We further applied the allele balance filter, where values from 0 to 1 represent the ratio of reads showing the reference allele to all reads. As RADseq targets specific genomic locations, the data should exhibit an expected allele balance near 0.5, for which we filtered our dataset below 0.25 and above 0.75 (16). Next, we filtered our dataset to contain sites with a maximum depth of 100 and removed individuals with high levels of missing data (>90%), as these represent individuals that were not sequenced well (72). We also applied a SNP completeness threshold of 90% and retained only unlinked biallelic sites. We removed the sites with a low minor allele frequency (MAF < 0.01) using VCFtools to filter out erroneous variant calls and singletons (variants only present in one individual) caused by sequencing or alignment errors (60). To filter homeologs, which are paralogs resulting from whole-genome duplication, we used the R package hierfstat (73) to exclude markers with excessively high observed heterozygosity deviations from Hardy-Weinberg proportions (H_o > 0.45 within samples) (59). We conducted preliminary analyses with and without applying the H_o filter, hereafter referred to as “strict” and “relaxed” filters, respectively. The preliminary results showed similar outcomes; however, there were cases where the relaxed filters recover finer patterns of population structure, allowing us to evaluate the effects of the breaks at a higher resolution (Fig. S9). Therefore, we selected matrices produced using the relaxed filters to conduct further population analyses.

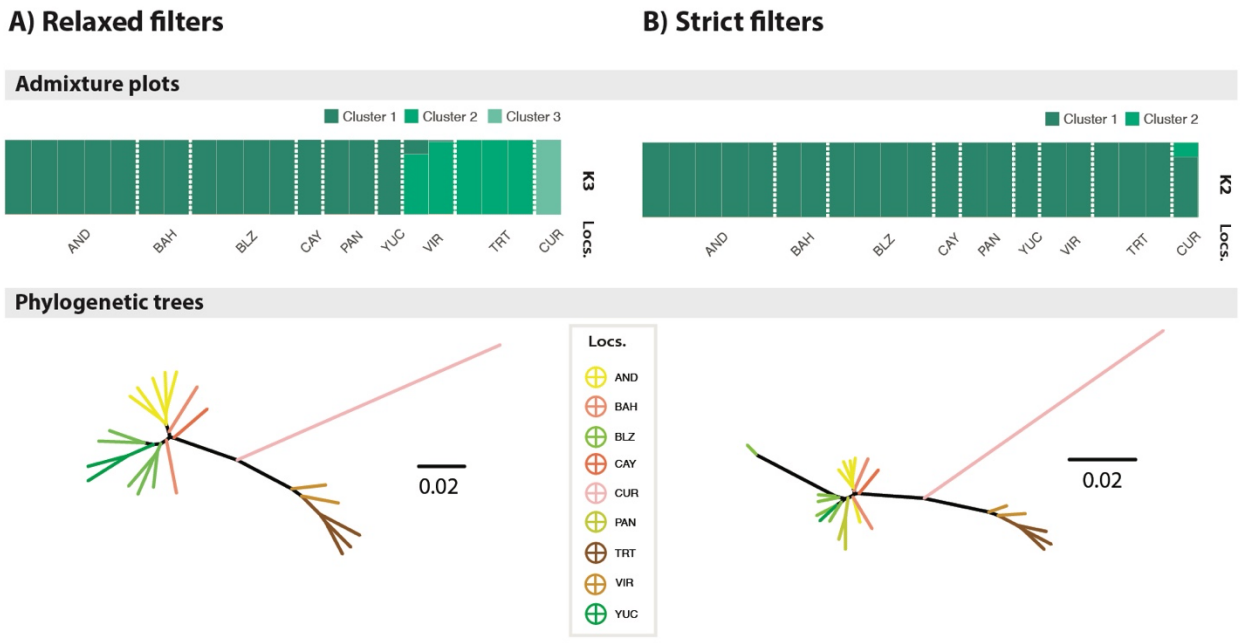


Figure S9. A) Admixture plot and phylogenetic tree of *M. aurolineatus* estimated using matrices applying relaxed filters. B) Admixture plot and phylogenetic tree of *M. aurolineatus* estimated using matrices applying strict filters.

Outlier SNPs detection. To generate matrices containing neutral SNPs, we identified SNPs that were potentially under selection by using two different approaches. First, we used BayeScan v2.1 (74) to apply a Bayesian differentiation-based outlier detection method. This approach calculates population-specific F_{ST} coefficients to identify candidate SNPs under selection using alternative models that consider and exclude selection (74). We conducted the analyses for 5,000 iterations with a burn in of 50,000 steps, following the recommendations in (75) to set the prior odds of neutrality parameter (*pr_odds*) at 10,000 to reduce the possibility of false positives. To determine significance, we specified the false discovery rate (FDR) at a *q-value* of 0.05. This means that SNPs with a posterior probability over 0.95 were considered outliers. The second approach involved using pcadapt (76) which detects outlier loci based on principal component analysis. This method scans the genome-wide data for signatures of positive selection based on patterns of differentiation along the major principal components of genetic variation (76). We used the first principal component identified by pcadapt, as it separated the populations geographically (see Supplementary Results), to detect outlier SNPs that were significantly associated with genetic variation correlating with marine barriers of connectivity and thus potentially involved in local adaptation. Outliers were identified using a false discovery rate of less than 0.1%, with SNPs showing a *q-value* lower than or equal to 0.05 considered outliers. The number of filtered SNPs at each filter-step are presented in Tables S2–S3. We conducted preliminary population genomic analyses on *M. aurolineatus* using matrices containing neutral, outlier, and all SNPs (neutral and outlier SNPs combined) (Fig. S10). We observed that the matrices formed by outlier SNPs did not possess enough statistical power, while the matrices with only neutral SNPs matrices lacked the

variation required to detect subtle genetic structure. Consequently, we performed all population genomic analyses using matrices that included both neutral and outlier SNPs. However, for the seascape genomic analyses of *M. tetranemus* and *M. triangulatus*, we conducted separate analyses using neutral and outlier matrices as we also aimed to identify potential local adaptations (see below).

Table S3. Number of SNPs retained after filtering step for the species in the TEP: *M. ebisui* (Me), *M. hubbsi-M. polyporosus* complex (Mh_spc), *M. mexicanus* (Mm), *M. tetranemus* (Mt), *M. sudensis* (Ms), *M. zacae* (Mza), and *M. zonifer* (Mzo).

Filter	Me	Mh_spc	Mm	Mt	Ms	Mza	Mzo
Stacks catalog --write_single_snp --R 0.00	163,765	103,148	86,250	229,978	131,178	137,579	61,987
Depth = 5, qg = 30	121,197	79,251	72,149	169,757	107,523	103,894	52,473
Allele balance filter	115,436	75,084	66,304	160,818	101,374	99,644	50,275
Max depth = 100	114,437	73,646	64,421	159,213	100,483	97,893	49,986
Remove individuals with > 0.9 % of missing data	113,598	73,646	64,339	159,213	100,303	97,893	49,976
Individuals removed	3	0	2	1	1	0	1
SNP completeness > 0.9 %	7,862	9,580	2,675	11,653	9,863	10,727	6,790
MAF 0.01	7,862	9,580	2,675	6,226	9,863	10,727	6,790
H_O < 0.45	7,843	9,325	2,665	6,088	9,612	10,549	6,310

Table S4. Number of SNPs retained after filtering step for the species in the TA: *M. aurolineatus* (Ma), *M. boehlkei* (Mb), *M. erdmani* (Mer), *M. gilli* (Mg), *M. macropus* (Mc), *M. triangulatus-M. lianae-M. brunoi* complex (Mtr_spc), *M. versicolor* (Mv).

Filter	Ma	Mb	Me	Mg	Mc	Mtr_spc	Mv
Stacks catalog --write_single_snp --R 0.00	133,447	89,098	141,052	149,545	138,123	286,612	67,876
Depth = 5, qg = 30	103,443	71,627	110,522	117,708	109,857	216,132	54,119
Allele balance filter	98,654	68,468	105,033	113,009	103,558	205,163	50,963
Max depth = 100	92,874	64,899	101,994	105,567	99,878	201,594	46,603
Remove individuals with > 0.9 % of missing data	92,874	64,433	101,376	99,012	98,535	197,942	46,603
Individuals removed	1	3	3	1	2		0
SNP completeness > 0.9 %	5,259	7,596	9,959	3,552	3,376	9,533	7,776
MAF 0.01	5,258	7,596	9,959	3,552	3,376	4,608	7,776
H_O < 0.45	5,220	7,393	9,888	3,535	3,358	4,583	7,544

We also assembled SNP matrices for species that were collected only in one location, including several individuals (*M. costaricanus*, *M. gigas*, *M. margaritae*, and *M. zonogaster*). For these species, the SNP completeness filter was more relaxed (65% in *M. gigas*). Final SNP matrices varied between 3,000 and 40,000 SNPs.

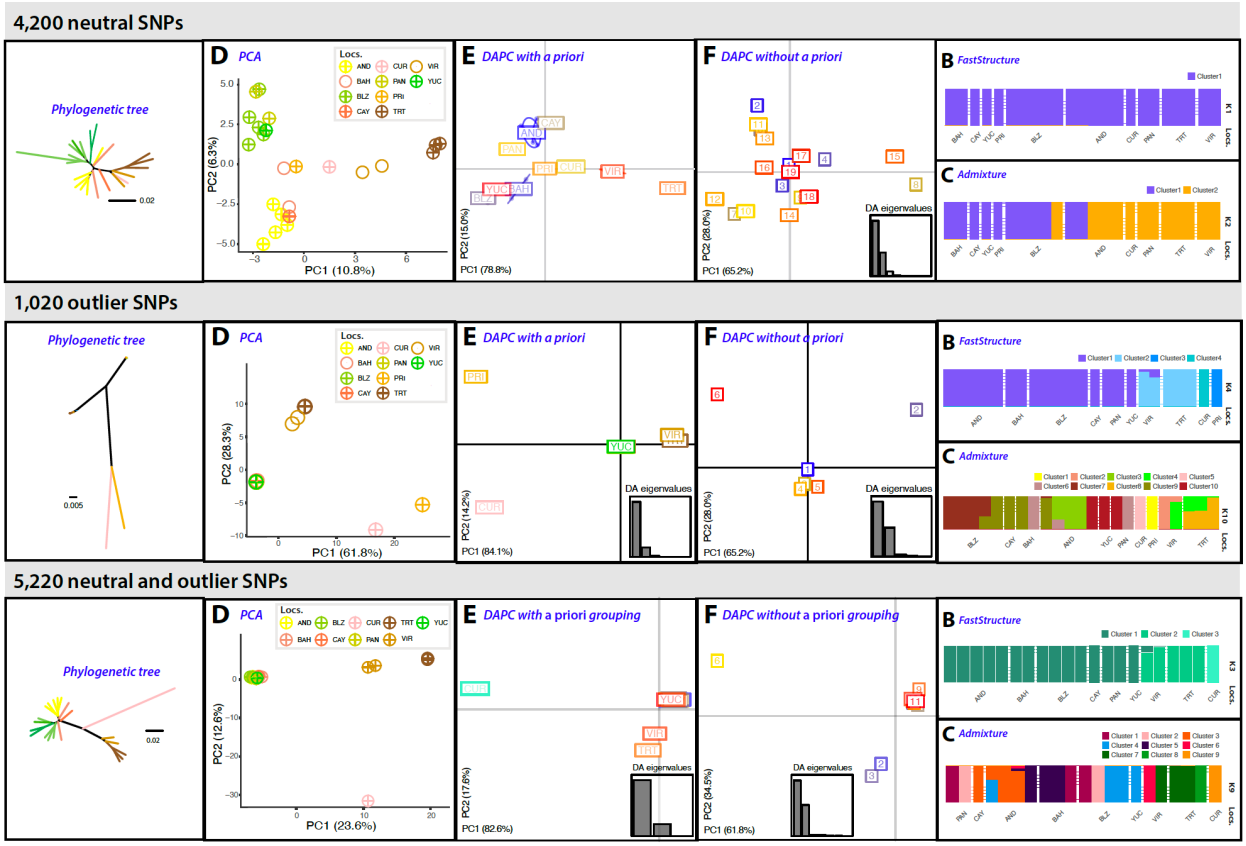


Figure S10. Population clustering analyses using A) neutral, B) outlier, and C) neutral and outlier SNPs of *M. aurolineatus*.

Locus assembly at macroevolutionary scales

At the macroevolutionary scale, we used three different approaches to assemble putative loci:

***1) De novo* assembly optimization:** We used 41 individuals representing the two most divergent individuals (whenever possible) from each of the 23 species of *Malacoctenus* along with the single *Brockius striatus* individual sequenced, which had the highest coverage values after the initial *de novo* exploratory run conducted during the quality control steps. With this dataset, comprising 24 species, we assembled the data *de novo* using the optimal parameter combination (m3M5n6) selected for microevolutionary analyses. Additionally, we tried a combination where we relaxed the *n* parameter to eight (m3M5n8), aiming to account for the inclusion of different species at macroevolutionary scales. However, this latter combination was computationally intensive and failed to produce a final dataset that included all individuals. Consequently, it was discarded from further analyses. We then filtered the SNP matrix to include only bi-allelic, unlinked, and orthologous SNPs, while applying different missing data thresholds using SNPfiltR (Table S5). As this step was also computationally intensive, we were unable to assemble a complete dataset using all 447 individuals that passed quality filters.

2) **Reference Genome assembly:** We assembled the raw reads using the genome of the jewelled blenny (*Salarias fasciatus*, NCBI accession GCA_902148835.1) as reference, as it represents the closest available genome with a divergence time estimated around 65 Ma (77, 78). We utilized the *ref_map.pl* wrapper available in Stacks using default values. We aligned the raw reads to the reference genome using BWA v0.7.17 (51), while allowing for mismatches and gaps (-B and -O parameters, respectively), and controlling for partial alignments (-L). To generate SNP matrices containing only bi-allelic, unlinked, and orthologous SNPs, we applied the same set of filters used in the *de novo* assembly at the macroevolutionary level. The final reference assembled matrices included between 1,891 and 28,144 SNPs, with 29.1–81.1% of missing data (Table S5).

***3*) *Quasi de novo* approach:** Previous studies have demonstrated that generating loci *de novo* and integrating the resulting alignment with a reference genome is more efficient than aligning raw data directly to the reference (61). Therefore, we also aligned the consensus sequences of the "optimal" *de novo* assembly catalog loci to the jewelled blenny (*Salarias fasciatus*) genome using the *stacks-integrate-alignments* option from Stacks. The same SNP filtering steps as before were applied, with a SNP completeness of 25% resulting in a matrix of 1025 SNPs (Table S5).

We conducted preliminary phylogenetic trees for each assembly approach based on SNPs present in at least 25% on the individuals (-R) (Fig. 34). We selected the genome reference-based assembly because it produced similar amounts of genetic information compared to the *de novo* and *quasi de novo* assemblies, but had lower proportions of missing data (28.2%, compared to 52.6% and 49.7% for the *de novo* and *quasi de novo* assemblies, respectively). Although the percentage of raw reads that aligned to the reference genome was less than 3% for both the reference-based and *quasi de novo* approaches, the *de novo* method resulted in significantly more missing data. This method also enabled us to assemble a matrix using the 447 individuals, which was subsequently used to generate matrices for phylogenomic and seascape genomic analyses (see below).

Table S5. Number of SNPs retained using three different approaches to assemble the raw data, and after filtering steps for the data set with 49 individuals, including 23 *Malacoctenus* species and *L. striatus* as an outgroup.

Filter	***De novo***	**Reference-based**	***Quasi De Novo***
Num. Individuals/Species	31/18	41/24	28/24
Stacks catalog --write_single_snp --R 0.00	1184687	151434	73973
Allele balance filter Max. depth 100 SNP completeness > 0.25 %			
SNPs	**1652**	**1891**	**1025**

MICROEVOLUTIONARY ANALYSES OF INDIVIDUALLY SAMPLED SPECIES

Population structure. We evaluated the population structure among sampling locations of 14 species, for which one to 26 individuals were collected across at least two populations (Appendix S1). We performed these analyses at the intra-specific level to detect broader to finer patterns of genomic structure. Additionally, we considered *M. hubbsi* plus *M. polyporosus*; and *M. triangulatus* plus *M. lianae* as species complexes (i.e., *M. triangulatus* complex and *M. hubbsi* complex, respectively) because our analyses identified low genetic divergence within each species pair, suggesting taxonomic over splitting. To investigate population clustering, we first conducted a discriminant analysis of principal components (DAPC) using the R package adegenet (21). We evaluated both *de novo* structures and *a priori* groupings, as the effect of group misspecification is unknown (22). We then used fastSTRUCTURE v1.0 (20) and ADMIXTURE v1.3.0 (18) software to infer population structure. Both methods aim to identify the best value of *k*, the number of populations, based on the same statistical model as STRUCTURE (i.e., independent loci, admixture model). However, they implement faster algorithms. While ADMIXTURE uses a maximum-likelihood approach to determine a single best-fit population scheme, fastSTRUCTURE, a Bayesian clustering method, outputs a range of biologically plausible scenarios based on the observed structure (20). To evaluate potential cryptic diversity, we ran both analyses with *k* values ranging from the number of sampling sites plus an additional genomic grouping ($n + 1$). For the fastSTRUCTURE analyses, we applied a logistic prior and used the script *chooseK.py* from fastSTRUCTURE to select the best-fitting model (*k*). We then re-ran the analyses 25 times with multiple random starting seeds to identify the five highest values of log-marginal likelihood (LLBO). For the ADMIXTURE analyses, we used a cross-validation method to select the optimal number of genetic clusters (*k*), where the lowest rate represents the best-fit model. ADMIXTURE was run with fivefold cross-validation rate (*--cv* = 5) and 2,000 bootstraps (*-B* = 2000), allowing precise calculation of the cross-validation rate and estimation of parameter standard error. To create final plots and visualize the results of fastSTRUCTURE and ADMIXTURE analyses, we used the R package pophelper (79). Finally, we estimated phylogenetic trees using the concatenated SNP-loci matrices under a maximum likelihood framework in IQtree v2 (23). To account for ascertainment bias, we applied the GTR plus ascertainment bias correction model (+ASC).

Spatial patterns of genetic structure. We employed the estimation of effective migration surfaces (EEMS) method (25) as a means to visualize the spatial patterns of population genetic structure and identify barriers to gene flow. EEMS represents genetic differentiation as a function of migration rates, based on an isolation-by-distance model (IBD), and quantifies genetic differentiation among geo-referenced samples. By comparing the observed genetic differentiation among individuals to expectations under an IBD model, this method identifies areas where differentiation exceeds IBD predictions, which correspond to areas of lower migration (barriers to gene flow). Conversely, areas where differentiation is less than predicted represent areas of greater effective migration (corridors) (25, 80). This method requires a genetic distance matrix of geo-referenced samples and a distribution polygon of the geographical area. To estimate the genetic

dissimilarity matrix, we used the *bed2diffs* module from EEMS. To delineate the geographical polygons illustrating the species' distribution range, we visualized the geographic coordinates of museum and scientific collection records available in the Global Biodiversity Information Facility (GBIF, gbif.org) data archive, using QGIS v3.22 software (81). We then used the Google Maps API v3 tool (www.birdtheme.org/useful/v3tool.html) to generate the KML files necessary for the analyses. We ran EEMS using 200, 300, 450, and 600 demes, repeating the process twice for each deme to ensure model convergence using random starting points. We used 10 million MCMC burn-in iterations followed by 50 million sampling iterations. We plotted the surfaces of effective migration rates (*m*) and effective diversity (*q*) resulting from the 600 demes scheme using the R package rEEMSplots (25). We also conducted an individual-based spatial principal component analysis (sPCA) (82) using the *spca* function in the adegenet R package to visualize genetically distinct clusters. This method is suitable to detect complex or cryptic genetic structures as it does not require data to meet Hardy-Weinberg expectations (82). We performed sPCA analyses using the georeferenced SNP matrices, retaining one eigenvalue representing a pattern of global structure, which correspond to strong genetic similarity between neighbors, and one representing local structure, indicating strong genetic dissimilarity between neighbors.

Finally, for some species, population structure analyses yielded different outcomes regarding the number of genomic clusters and patterns of geographical division. In such cases, we used an analysis of molecular variance (AMOVA) to evaluate the relative contributions of each genomic grouping to genetic variation and to determine which hypothesis has the highest support. We conducted AMOVAs using the *poppr.amova* function within poppr R package (24), with variation computed between genomic clusters, between samples within clusters, and within samples. To assess statistical significance, we used a randomization test with 9,999 replications and the pegas (83) package in R.

PHENOTYPIC ANALYSES OF BODY SHAPE

We employed 2D geometric morphometric analysis to quantify intra-species body shape diversity and to evaluate if there are patterns of geographic clustering that correspond to the patterns of genomic structure recovered in this study. We also conducted this analysis at the inter-species level to evaluate the extend of morphological variation across species. A total of 20 fixed-landmarks and 16 semi-landmarks, commonly used in fish morphometric studies at the intra- and inter-specific levels, were placed to represent variation of body shape across localities or species (84–87). These data were obtained from high-quality photographs taken during field expeditions or retrieved from online museum repositories (*Appendix S2*). In total, we examined 101 photographs from 20 species of *Malacoctenus*. At the inter-specific level, we included

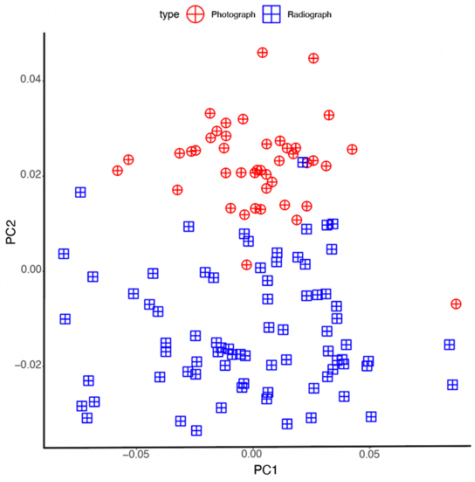


Figure S11. Morphospace of photographs (depicted in red circles) and radiographs (blue squares) of *M. triangulatus*.

one to ten individuals per species while maximizing representation of different sampled localities whenever possible. At the intra-specific level, we analyzed photos of *M. tetranemus* and *M. triangulatus* (47 and 42, respectively), as those were the only species for which we could obtain photos from multiple locations representing their broad geographic ranges. Additionally, we retrieved 75 x-rays of *M. triangulatus* from the fish collection database of the Smithsonian National Museum of Natural History (NMNH) (collections.nmnh.si.edu/search/fishes).

To minimize potential allometric effects, we included only adult specimens in our dataset. We digitized the set of landmarks and semi-landmarks using the R package StereoMorph (26). The semi-landmarks were placed as curves, and represented by ten equally-spaced points that captured the body's deep profile from the anterior base of the dorsal fin to the posterior end, and from the anterior base of the anal fin to the end. To correct for distortions, size, and position of different specimen images, we conducted a generalized Procrustes analysis (GPA). We then performed a principal component analysis (PCA) using the R package geomorph (27). To minimize digitization errors associated with landmarks, all landmarks were consistently positioned by a single individual. Nevertheless, in cases of repeated attempts, it is highly likely that each landmark may be placed slightly differently compared to its previous positioning. Therefore, to measure the error rate of landmark digitization, we digitized five times five individuals from *M. tetranemus* and *M. triangulatus.* Repeating digitization multiple times on the same photographs allowed us to quantify the similarity between digitizations and to calculate measurement error. To this end, we plotted the centroid scores along the morphospace (Csize value) to see if repeated measurements of the same individual clustered together. As replicates are supposed to represent the same underlying value, any variability among them can be attributed to measurement error. Therefore, we calculated the coefficient of variation (CV) among replicates as the ratio of the standard deviation to the mean value, expressed as percentage.

We analyzed photographs and x-rays separately, as preliminary analyses showed that the two datasets formed different, almost non-overlapping clusters separated by the PC2 axis in the morphospace (Fig. S11). This may be related to the differential ability in distinguishing specific characters depending on the image source. For example, the base of the fin rays is easily distinguishable in x-rays. Alternatively, it may be related to the condition of the specimen at the moment of image capture, as it has been suggested that the type of preservation of the specimen influences geometric morphometric analyses, where differences related to the method could be confounded with biologically-relevant disparity (86, 88). For instance, photographs were taken using fresh specimens, while x-rays were taken from fixed and preserved specimens deposited at the NMNH fish collection. As an additional quality control step, we also generated four different schemes, each comprising various combinations of landmarks and semi-landmarks. The first scheme contained all landmarks and semi-landmarks, while the second scheme solely comprised fixed landmarks. In the third scheme, only landmarks located on the head were included, and the fourth scheme comprised landmarks positioned on the anterior body (Fig. S12).

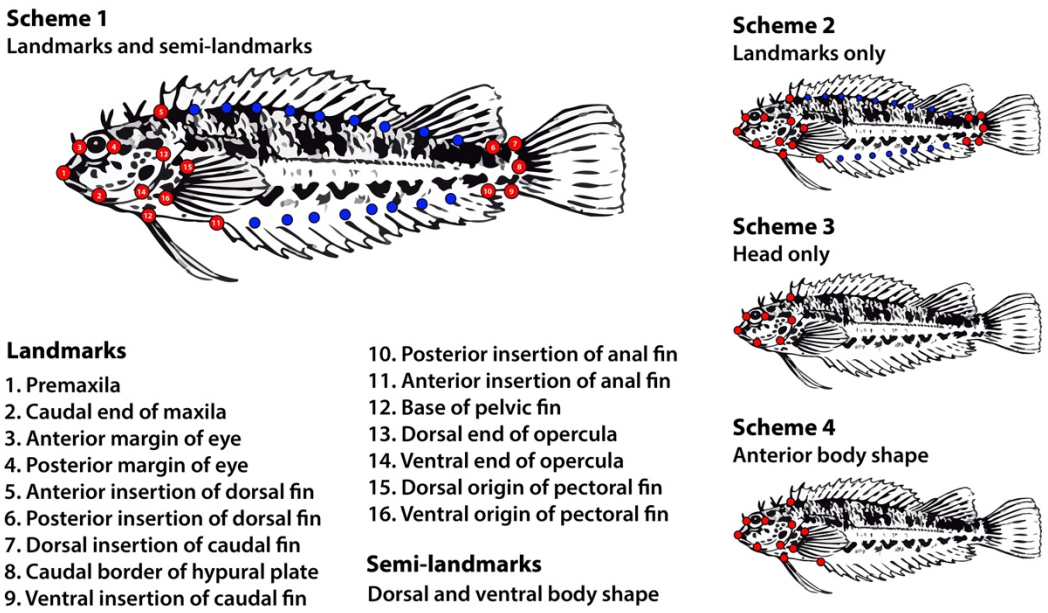


Figure S12. Schemes of landmarks and semi-landmarks used to conduct geometric morphometric analyses.

We selected scheme 4 as landmarks and semi-landmarks placed on the caudal fin showed "bending effects" due to specimen preservation artifacts (Fig. S28–S30).

SEASCAPE GENOMICS

To disentangle the effect of spatial distribution, oceanic currents and environmental variables on the patterns of genomic variation in the TEP and GC, we analyzed *M. tetranemus* and *M. triangulatus* (the most widespread species in the TEP and GC, respectively) under a seascape genomics framework.

Environmental variables. To accurately test the influence of environmental variables and collect most of the variables recognized as drivers of population divergence or speciation among fishes (89–91), we mined environmental information from the Marine Copernicus online data archive (http://marine.copernicus.eu/). We extracted six estimates of sea surface temperature (SST), and the mean and standard deviation values for the mass concentration of chlorophyll α in sea water (CHL), surface partial pressure of carbon dioxide in sea water (CO2), sea water pH reported on total scale (pH), mass concentration of suspended matter in sea water (SPM), sea surface salinity (SSS), and sea current velocity (SCV), sampled at monthly intervals over the sampling years (2015-2017) (Table S6) using the raster (92) R package. We conducted a principal component analysis (PCA) using the environmental matrix containing 18 variables sampled at each sampling location using the *dudi.pca* function in the R package ade4 (93). We retained the first three

principal component (PC) axes to conduct represent environmental seascape variables in the dbRDA analyses (see below).

Table S6. Environmental variables retrieved from Marine Copernicus data archive.

Environmental variable	CMEMS product name	Spatial resolution	Temporal resolution	Temporal extent	Output variables
Sea surface salinity (SSS)	GLOBAL_MULTIYEAR_PHY_001_030	0.083°	monthly	2010-2017	-overall average -overall standard deviation
Sea current velocity (EWV, NWV)	GLOBAL_MULTIYEAR_PHY_001_030	0.083°	monthly	2010-2017	-overall average -overall standard deviation
Chlorophyll a concentration (CHL)	OCEANCOLOUR_GLO_BGC_L4_MY_009_104	4 km	monthly	2010-2017	-overall average -overall standard deviation
Sea water pH (PH)	GLOBAL_MULTIYEAR_BGC_001_029	0.25°	monthly	2010-2017	-overall average -overall standard deviation
CO_2 in sea water (CO2)	GLOBAL_MULTIYEAR_BGC_001_029	0.25°	monthly	2010-2017	-overall average -overall standard deviation
Suspended particulate matter (SPM)	OCEANCOLOUR_GLO_BGC_L4_MY_001_104	4 km	monthly	2010-2017	-overall average -overall standard deviation
Sea surface temperature (SST)	GLOBAL_MULTIYEAR_PHY_001_030	0.083°	daily	2010-2017	-overall average -overall standard deviation -average of the hottest month -average of coldest month -st. dev. hottest month -st. dev. coldest month

To summarize the patterns of environmental variation at the sampling sites, we employed principal component analysis and hierarchical clustering approaches using the *prcomp* and *hclust* functions from the stats v4 R package. Hierarchical clustering was used to construct a similarity tree, which was then visualized to represent the patterns of environmental variation on a map.

Spatial distribution. As environmental features can directly affect the dispersal capabilities or suitable habitat of the species, we evaluated the relative contribution of spatial distribution on patterns of genetic variation at multiple scales (94). To this end, we calculated distance-based Moran's eigenvector maps (dbMEMs) to model symmetric processes (*e.g.*, spatial distance) (95). dbMEMs use a matrix of pairwise distances of spatial locations to measure the autocorrelation of a variable. By computing eigenvectors extracted from the spatial weight matrix, creates eigenvector maps that capture the main patterns of the spatial autocorrelation in the data (96). This technique can detect multiscale spatial patterns while controlling for spatial correlation in tests of y-x relationships and it is suitable to identify complex spatial patterns of autocorrelation (97). To create a spatial pairwise matrix, we first calculated geographical distances using the least cost path (LCP) distance over seawater with the R package *marmap* (30), hereafter referred to as 'in-water distances.' Then, we used the *dbmem* function in the R package adespatial (31) to calculate MEM values. Additionally, we calculated Euclidean geographic distances using the *dist* function from the R package *adegenet* (21) and calculated MEM variables using Euclidean distances between sites. We selected variables contributing to the explained variation in both directions (forward and backward) by using *ordistep* function from the vegan R package, and the model choice is assessed using a permutation method. A total of two significant MEMs were kept for each species. We calculated the adjusted coefficient of determination (adjusted R^2) using the *RsquareAdj* function in vegan to account for the number of observations and degrees of freedom in the model (8, 98).

Finally, maps of the MEMs associated with patterns of genomic variation were plotted to visualize the spatial patterns using ggmaps R package (99).

Multivariate analyses. We used distance-based redundancy analyses (db-RDA; (28)) to evaluate the influence of the spatial distribution, oceanic currents, and environmental variables on the population structure of *M. tetranemus* and *M. triangulatus* at the intraspecific level, as well as on the 23 species analyzed by biogeographic region at the interspecific level. We followed the script by Benestan *et al*., (100). In brief, we calculated Euclidean genomic distances among individuals using the *dist* function from the R package adegenet. The resulting genomic distance matrix was considered the response variable, while the explanatory variables included spatial patterns estimated using Moran's eigenvector maps (MEMs, (101)), and 18 environmental variables. We conducted a principal coordinate analysis (PCoA) on the Euclidean distances and retained the number of PC axes based on the Broken stick model and Kaiser Guttman criterion (eigenvalue of axis should be higher than the average of eigenvalue). To account for the earth curvature we used the *gcd.hf* function available in the codep (102) R package. To eliminate multicollinearity among variables, we used the variance inflation factor with a threshold of 5. When a strong correlation was evidenced, we retained only one of the two variables. We calculated the global db-RDA model using the *dbrda* function available in the 'vegan' package on the binary genotype variation on all explanatory variables, and calculated the model probability and adjusted coefficient of determination (adjusted r^2) using the *RsquareAdj* function in the R package vegan, accounting for the number of observations and degrees of freedom in the best-fit model (8, 98). We selected the variables contributing to the explained variation by applying a stopping criterion with the *ordiR2step* function in the vegan R package. Biplots were generated using *varpart* function available in the vegan package to represent the association between genetic clusters and selected explanatory variables.

MACROEVOLUTIONARY ANALYSES, PHYLOGENOMIC INFERENCE AND BIOGEOGRAPHIC HISTORY

Mitochondrial tree. To estimate a mitochondrial tree, we used the COI sequences previously retrieved from the NCBI archive and the BOLD system during the quality control steps. We aligned sequences from 68 individuals representing 17 species using Geneious v10 (103) for a final alignment of 645 bp. We did not incorporate the COI sequences mined from the raw data from the 169 hits recovered from our dataset, as they were short fragments (140 bp) with insufficient variation for tree estimation. We inferred this tree using IQtree v2 with Model finder using by-codon partitions.

Phylogenomic inference. We estimated phylogenetic trees based on genomic data using five concatenated SNP-loci matrices, each with varying levels of missing data. These matrices encompassed 1,891 to 28,144 SNPs obtained from 40 individuals representing the 23 species of *Malacoctenus*, along with the *B. striatus* outgroup. We used these matrices to estimate phylogenetic trees under a ML framework in IQtree v2, employing the GTR+ASC model. Because

genes can exhibit unique evolutionary histories attributed to biological phenomena like incomplete lineage sorting (ILS), recent criticisms of concatenation methods have emerged. These criticisms stem from the concern that concatenation methods might not truly capture the evolutionary history of the species under investigation (104, 105). To account for this, we also used the multispecies coalescent model (MSC), which conceives the species tree as a collection of gene trees, and has been suggested to outperform concatenation methods, while being able to account for biological processes such as ILS (106, 107). We applied the SVDquartets plug-in (108) implemented in PAUP* v4.0 (32) to estimate species-level phylogenies under the MSC model. SVDquartets uses a quartet-based approach, in which the coalescent histories of all possible quartets of species are compared in the dataset, selecting the species tree that is consistent with the majority of quartet trees. We evaluated 100,000 random quartets, applying a non-parametric bootstrapping of 10,000 replicates which were used to account for uncertainty in the inferred splits. We estimated a majority-rule consensus tree in PAUP* using the reference-based assembled matrix of 28,144 SNP-loci to maximize the amount of genetic information. Trees were visualized in Figtree v4 (tree.bio.ed.ac.uk/software/figtree/).

Time calibration. We time-calibrated the species tree using the multi-species coalescent (MSC) model with the SNAPP plug-in implemented in BEAST2 v2.6.7 (34). We prepared the input files in XML format generating settings that model the MSC using the script *snapp_prep.rb* (109). To reduce time and computational burden, we enforced the monophyly of the phylogenetic groups recovered in our SVDquartets species tree. As this method does not allow for missing data per species, and given the high amount of missing data in our macro-evolutionary dataset, we used the consensus sequences of each species from our matrix, which included 41 individuals across all 23 *Malacoctenus* species and one *B. striatus*, and was generated using the *populations* wrap from Stacks. We used the custom script *select_no_miss_snps.py* to select the first SNP of each locus that did not have missing data, resulting in a dataset composed by 279 bi-allelic SNP loci.

To time-calibrate our species tree, we implemented a secondary calibration prior at the root (crown Labrisomini), using the minimum and maximum ages estimated by previous fossil-calibrated teleost phylogenies for the most recent common ancestor (MRCA) of *Malacoctenus* species and *Brockius striatus* (Table S7). We specified this age constraint as an uniform-distributed calibration, with the upper and lower bounds set to 35 and 32 Ma, respectively. We did not include the age estimate for the closure of the Panama Isthmus (3.1–2.8 Ma (110)) as biogeographic node-age constraint on the transisthmian clade or the geminate species pair to avoid circular inference, since our goal was to infer the timeline for the evolution of *Malacoctenus*, evaluating if the final closure of the Isthmus triggered synchronous divergent events. Also, we did not incorporate any fossil calibration points in our analyses, as there are is fossil record for *Malacoctenus*. Also, the only fossil of Labrisomini, †*Labrisomus pronuchipinnis*, from the Messian deposits of the Oran region (111), is relatively young and thus uninformative, and its phylogenetic placement is unclear. We conducted two replicate analyses, each consisting of 5,000,000 MCMC iterations. We assessed chain convergence and stationarity using Tracer v1.5 (112), checking that ESS values were above 200 for all parameters. After discarding the 10% of each chain as a burn-in, we combined the resulting MCMC chains and selected the maximum clade credibility tree using TreeAnnotator software from the BEAST2 package.

Table S7. Ages estimated by previous fossil-calibrated teleost phylogenies for the most recent common ancestor (MRCA) of *Malacoctenus* species and *Brockius striatus.*

Reference	MRCA of *Malacoctenus* and *L. striatus* (Ma)	Total group age (Ma)	Mean crown age (Ma)
(78)	32.0	24.8	16.3
(77)	35.3	31.0	19.1
(113)	35.0	28.7	20.6
min-max values	32.0–35.0	24.8–31.0	16.3–20.6

Phylogenetic time-calibrated tree with all individuals. To estimate a time-calibrated tree that includes all individuals passing our quality filters at the population level (total 447), we followed several steps. First, we extracted estimated divergence times from common nodes between the previously time-calibrated species tree (which included 23 individuals representing all 23 *Malacoctenus* species) and the ML phylogenetic tree (comprising 49 individuals and the 23 *Malacoctenus* species). We used these divergence times as secondary calibration points to time calibrate our the ML phylogenetic tree with the software RelTime (114) in MEGA X v10.1.8 (36). The resulting time-calibrated tree contained two individuals per species (when possible) and was designated as the backbone tree. We then time-calibrated each species subclade using the resulting divergence time estimates from our time-calibrated backbone tree as secondary calibration points, with all secondary calibration points set to a uniform distribution. Finally, we grafted all intra-species subclades into the time-calibrated backbone tree using the R package ape (37). The final tree portraits the population genetic structure at the micro-evolutionary scale and the macro-evolutionary relationships across the 23 *Malacoctenus* species, including 447 individuals.

Test for synchronous divergent events. We used the software ecoevolity v0.3.2 (12) to test whether the cladogenetic events between the geminate species and the transisthmic clade were synchronous or independently originated by the rise of the Panama Isthmus. Ecoevolity is a full-likelihood Bayesian approach that integrates gene trees with the population histories of each species pair from genomic data to test models of co-divergence (12). This method uses a Dirichlet process prior to assign an unknown number of divergence times to the studied pairs by specifying a concentration parameter (α), which determines the probability of the events to be synchronous, and base distribution, which acts as a prior on idiosyncratic divergence times (115). We assumed that all pairs share the same mutation rate, which we set to 1.0 to scale the effective population sizes and time with the mutation rate, therefore representing time in expected substitutions per site (115).

Although the model implemented in ecoevolity assumes that all genetic characters are orthologous, bi-allelic, and un-linked SNPs, simulations that included linked and constant characters suggest that this method performs better with their inclusion (12). We conducted the analyses using fasta files generated during the genome referenced assembly, including only one to four individuals per species pair. Concatenated alignments used for these analyses comprised 260,297 and 218,442 sites for the species pairs *M. tetranemus–M. delalandii* and *M. mexicanus–*

M. triangulatus, respectively. Additionally, to maximize the amount of genetic information, we assembled the loci *de novo* for each trans-isthmic pair. These data sets comprised 1,022,314 and 1,042,562 bp for the species pairs *M. tetranemus–M. delalandii* and *M. mexicanus–M. triangulatus*, respectively. We set the concentration parameter of the Dirichlet process ($\alpha = 1$), while considering the null hypothesis as the pairs diverged independently (116). We set the prior on divergences times $\tau \sim$ Gamma to 10 in order to simulate a slow divergence scenario. We ran five MCMCs for 2,000,000 generations, sampling every 2,500 generations. We assessed the convergence of the analyses using the *pyco-sumchains* tool in pycoevolity (115).

Biogeographic analyses. We inferred ancestral geographic ranges for the genus with BioGeoBEARS (9), with our time-calibrated species tree as the input phylogeny. We built a presence/absence matrix by coding each species based on the location records retrieved from scientific and museum collections available in GBIF data archive. We implemented a biogeographic scheme based on seven biogeographic provinces that covered the distribution range of the 23 *Malacoctenus* species. Our biogeographic scheme was based on the provinces proposed by (4, 10, 11) to reflect some of the major biogeographical breaks evaluated in this study. Although there is controversy around the extent and number of the biogeographic provinces in the TEP (117), here we considered the four provinces (4) that represent endemism, distribution, and genetic structure of several rocky shore fishes (3). The biogeographic scheme of the TEP was represented by four provinces: Cortez (C), Mexican (M), Panamian (P), and Galapagos (G). The province of Cortez corresponds to the Gulf of California, the Mexican province includes the coast of Mexico from Mazatlán, Sinaloa, to the Isthmus of Tehuantepec in the south, the Panamic province extends from the Gulf of Fonseca in Nicaragua to the south of the Gulf of Guayaquil, Peru, and the Galapagos Archipelago. For the TA, we used the delimitation suggested by (10), which represents the GC and other biogeographic provinces (11) for selected localities outside the GC. The provinces within the GC were suggested based on endemism and diversity patterns of shore reef fishes. This scheme considers three major provincial subdivisions that present a distinctive, primarily tropical fauna: Northern (N), Central (C), and Southern (S) provinces. The Northern province includes the Gulf of Mexico, Florida and southeastern USA; the Central province comprises the West Indies, Bermuda, and Central America; and the Southern province includes the continental shelf of northern South America. For the localities collected outside the GC, the Fernando de Noronha and Atoll das Rocas ecoregion was considered within the Southwestern Atlantic, while the only African species (*M. carrowi*) was sampled from Cape Verde Islands and was classified within the West African Transition province (11).

We evaluated 12 biogeographic models under a maximum likelihood framework that have been used in marine fishes (118–120). Such models included dispersal-extinction-cladogenesis (DEC), dispersal-vicariance analysis (DIVA), and Bayesian estimation of the biogeographical history (BayAREA), each in with a combination two main parameters: the founder-speciation event (*j*), which allows the colonization of an area by a daughter lineage while the splitting-sister lineage stays at the ancestral area (121); and the dispersal matrix power exponential (*w*) parameter, which allows the model to adjust the matrices according to the data (122). We used two time slices (30–2.8 Ma, and 2.8–0 Ma) to reflect the two main geological events that influenced the evolutionary history of the group: the closure of the Panama Isthmus *ca.* 2.8 Ma (110) and the rise

of the Galapagos Archipelago around 0.5-2.5 Ma (123). The connectivity between areas during these time slices was determined by coding three dispersal probability categories: 1.0 for well-connected areas, 0.05 for relatively separated areas, and 0.0001 for widely separated or disconnected areas following (119). The maximum range parameter was four areas. For example, the connectivity between the Mexican and Central provinces was coded as 1.0 in the first time slice (30–2.8 Ma) before the rise of Isthmus of Panama, and as 0.0001 for the second slice (2.8–0 Ma) after the final closure of the isthmus. Finally, the best-fitting model was selected using Akaike Information Criterion scores corrected for small sample size (AICc) values calculated for each biogeographic model.

EXTENDED RESULTS

We generated approximately 3.1 x 10^9 sequence raw reads, of which ~2.8 x 10^9 (91.5%) passed quality filters, including *ca.* 6 million reads per sample. At the microevolutionary scale, we assembled samples *de novo* setting a minimum of three raw reads required to form a stack (*m*), allowing a maximum of six mismatches between loci (*M*), and six mismatches between loci of different individuals (*n*). We generated matrices with 4,000-9,000 biallelic orthologous unlinked SNPs (3.7%–11.7% missing data) that were used to identify genetic patterns corresponding to major contemporary breaks.

At the macroevolutionary level, we assembled the raw data of 41 individuals spanning the 23 *Malacoctenus* species sequenced along with the outgroup (*L. striatus*), using the closest reference genome available (*Salarias fasciatus*, Blenniidae). We generated five matrices consisting between 1,891 to 28,144 SNPs, which were selected based on varied levels of missing data (29.1–81.1%) and were used to conduct phylogenomic analyses. The primary objective at this scale was to elucidate the evolutionary and biogeographic history of the genus with an emphasis on examining the synchronicity of speciation events triggered by historical barriers.

MICROEVOLUTIONARY ANALYSES FOR INDIVIDUALLY SAMPLED SPECIES

Population differentiation analyses

The population clustering approaches yield varying levels of population structure, where PCA and DAPC analyses, and phylogenetic trees unveiled finer patterns of population differentiation. DAPC analyses without *a priori* grouping generally recovered the highest grouping scheme evaluated, such groups clustered mostly in agreement with PCA and DAPC, with *a priori* grouping, analyses. ADMIXTURE and fastStructure analyses consistently detected similar genetic groupings, with few exceptions where ADMIXTURE identified genetic flow among populations (e.g., *M. versicolor*; Fig. S29) or a higher genetic clustering scheme (e.g., *M. aurolineatus*, $k = 9$; Fig. S23). EEMS plots provided a visual description of the population structure highlighting areas of disrupted connectivity, mostly aligning with previously described marine breaks. Contrary,

sPCA plots disagreed with other population differentiation analyses. Detailed genomic patterns per species are described below and depicted on Figs. S15–S29.

Tropical Eastern Pacific

M. ebisui. The phylogenetic tree identifies three main lineages: clade A comprises individuals from Bahía Santiago (SAN), Agustinillo (AGU), and Zacatoso (ZAC); clade B clusters consists of individuals from Cabo Blanco (CBL); and clade C includes individuals from Continental Ecuador (ECC) and Nuquí (NUQ) (Fig. S13A). Both fastStructure and ADMIXTURE analyses depict similar patterns, suggesting the optimal *k* scenario consists of three genetic clusters consistent with the phylogenetic clades. Both analyses indicate genetic flow from ECC and NUQ to CBL (Fig. S13B–C). PCA and DAPC analyses, conducted with and without *a priori* grouping yielded the same genetic clustering pattern. PC1 differentiates the populations of SAN, AGU, and ZAC from the rest, accounting for 61.8 to 82.6% of the variation, while PC2 (11.5–35.4%) identifies the CBL population as a distinct genetic cluster (Fig. S13D–F). These genetic clusters align with a north to south division originated by the Central American Break (CB) and the Panama Gyre Break (PGB). The EEMS plot emphasizes an area restricting genetic flow near CBL (Fig. S13G). sPCA analyses suggest a division into two main clusters: the first encompassing SAN, AGU, and ZAC; and the second comprising ECC, NUQ, and CBL (Fig. S13H).

M. hubbsi-M. polyporosus **species complex.** All population differentiation analyses display consistent genetic structure patterns, which distinctly separate the Vallarta (VAL) population from the others (Fig. S14A–H). The PC1 axis from both PCA and DAPC analyses, estimated with and without *a priori* grouping, delineate these genetic clusters, accounting for 31.4 to 99.5% of the variation (Fig. S14D–F). Notably, the barrier highlighted on the EEMS plot aligns with the Sinaloan Break (SB; Fig. S14G).

M. mexicanus**.** The phylogenetic tree delineates two main clades. Clade A comprises the populations from Santa Inés (STI) and Tepetate (TPE). In contrast, Clade B encompasses the populations of Vallarta (VAL) and Zacatoso (ZAC) (Fig. S15A). The delineation of these clades suggest that the Sinaloan Break (SB) is hindering population connectivity among these lineages. Neither the fastStructure nor the ADMIXTURE analyses discern any population structure (k=1, Fig. S15B–C). The results of the PCA analyses are consistent with the genetic clusters highlighted by the phylogenetic tree. The PC1 axis differentiates STI and TPE from VAL and ZAC, representing 12.1% of the genetic variation (Fig. S15A). DAPC analyses, when conducted with *a priori* grouping, display STI and TPE as a distinct cluster separated by PC1 axis (99.8%) from ZAC and VAL. In contrast, when DAPC analyses were ran without *a priori* grouping, they fail to detect any genetic structure consistent with the previously observed clusters (Fig. S15E–F). The EEMS plot identified an area interrupting genetic flow, coinciding with the Sinaloan Break (Fig. S15G). Additionally, sPCA analyses further distinguish the STI and TPE populations from ZAC and VAL (Fig. S15H).

M. sudensis**.** The phylogenetic tree delineates two major clades: clade A comprises individuals from Continental Ecuador (ECC), while clade B is formed by individuals sampled from Cabo

Blanco (CBL), Los Cóbanos (COB), Nuquí (NUQ), Pedasí (PED), and Las Perlas Archipelago (PER) (Fig. S16A). FastStructure analyses recovered two genetic groups as the primary hypothesis for population structure (k=2). The results from ADMIXTURE were marginal, as the best cross-validation scenario did not distinguish any population structure (k=1); the second-best scenario reflected the same population structure (k=2) as that revealed by fastStructure (Fig. S16B–C). The PC1 axis of the PCA and DAPC, with *a priori* grouping, distinguished ECC from the other populations (11.8–86.4%). The PC2 axis (0.05–13.5%) revealed further population structrure, presenting three genetic groups: CBL and COB, NUQ and PER, and PED (Fig. S16D–E). This analysis suggests a potential new break (PNB3) hindering population connectivity south of CBL. In contrast, DAPC, calculated without *a priori* grouping, did differentiate ECC on PC1 (89.7%), but did not show a population structure associated with geographical locations (Fig. S16F). EEMS plots underscore two primary barriers to genetic flow: the first between PED and PER aligning with the Panama Gyre Break (PGB), and the second near NUQ highlighting a potential new break (PNB4) (Fig. S16G). sPCA analyses also categorized ECC as a distinct genetic cluster (Fig. S16H). AMOVA analyses compared both hypotheses of two and four genetic groupings, finding that the majority of genetic differentiation occurred within individuals (75.18–81.40%) (Table S8). We observed a moderate level of genetic differentiation between groups (13.67–17.02%), while a low percentage of variation was represented among individuals within groups (4.93–7.80%). The division into more groups led to a slight decrease in the percentage of variation between groups, suggesting higher support for a two-group scenario where the ECC population is significantly different from the rest.

***M. tetranemus*.** The phylogenetic tree identified two main clades: clade A, composed of individuals from Continental Ecuador (ECC), and clade B, comprising individuals from the other populations. Within clade B, further population structure is evident: subclade B1 includes individuals from Michoacán (MIC), Zacatoso (ZAC) and Agustinillo (AGU) populations; subclade B2 consists of Cedros (CED), Santa Inés (STI), and Tepetate (TPE) populations; and subclade B3 encompasses Cóbanos (COB), Cabo Blanco (CBL), Pedasí (PED), Las Perlas (PER), Nuquí (NUQ), and Galápagos Archipelago (GAL) populations, along with some individuals from ECC (Fig. S17A). Due to the substantial genetic distance between clade A and clade B, we hypothesize that clade B might represent a new species that is hereafter referred to as *M. aff. tetranemus*. Both fastStructure and ADMIXTURE analyses consistently indicated a best k scenario of two genetic groupings, with some individuals from *M. aff. tetranemus* forming group one. Further fastStructure and ADMIXTURE analyses using only *M. tetranemus* revealed additional population substructure (k=2) that is congruent with the phylogenetic tree, where subclades B1 and B2 constituted a genetic grouping distinct from subclade B3 (Fig. S17B–C). PCA's PC1 consistently recognized *M. aff. tetranemus* as a unique group, accounting for 25% of the variation. PC2, accounting for 5%, identified two groups that correlate with the fastStructure and ADMIXTURE results for *M. tetranemus* (Fig. S17D). DAPC analyses, both with and without *a priori* groupings for *M. tetranemus*, discerned four genetic groups: group 1 comprised of GAL; group 2 consisting of COB, CBL, PED, PER, NUQ, and ECC; group 3 encompassing MIC, ZAC, and AGU; and group 4 made up of CED, STI, and TPE. PC1 differentiated groups 1 and 2 from 3 and 4, explaining 76.1% to 78.0% of the genetic variation. In contrast, PC2 (12.0–13.3%) highlighted a more nuanced genetic structure, differentiating grop 1 from group 2 and group 3

from 4 (Fig. S17E–F). The results observed on the DAPC analyses provide support for the evaluated marine breaks including the Open Ocean Galapagos Break (OOGB). EEMS plots pinpointed three primary areas inhibiting genetic flow, coinciding with the Sinaloan Break (SB), the Central American Break (CB) and a region near ECC that represents a potential new break (PNB4) (Fig. S17G). The sPCA of *M. tetranemus* detected two main genetic groups, one comprising DAPC's groups one and two and another including groups 3 and 4 (Fig. S17H). AMOVA analyses, calculated comparing the different genetic clustering hypotheses (two, three, and four groups in *M. tetranemus*; Table S8), suggest a consistent pattern in which most of the genetic variation is found within individuals, representing a rich genetic diversity (85.55–87.57%). Geographic clusters have undergone some degree of significant genetic differentiation (11.27–11.83%), whereas the genetic differentiation within groups is low indicating that individuals within the same geographic group are relatively similar to one another (0.82–3.17%). In this scenario, although marginal, the genetic clustering pattern that represented the higher proportion of genetic differentiation was the identification of three groups: Gulf of California, Central Mexico, and South America-Galapagos.

***M. zacae*.** The patterns of genetic structure consistently identify two main genetic groups, as depicted in Figure S18A–H. The first group consists of Cedros (CED), Los Frailes (FRA), and Tepetate (TPE), while the second group comprises Vallarta (VAL), Agustinillo (AGU), and Zacatoso (ZAC). The genetic structure represented by these groups aligns with a scenario where the Sinaloan Break hinders population connectivity among these populations. Additionally, both PCA and DAPC analyses—whether conducted with or without *a priori* groupings—identified these genetic groups on the PC1 axis, which accounts for 48.1 to 99.9% of the total variation. Interestingly, the PC2 did not reveal additional genetic patterns, contributing only 0.02–1.1% to the variation (Fig. S18D–F). Although PCA's PC1 only represented 48.1% of the variance, the rest of the variation was contained in 26 PC axes, each of them with less than 0.025 %. Notably, the barrier emphasized in the EEMS plot also coincides with the Sinaloan Break (Fig. S18G). The results from sPCA analyses were also consistent, highlighting the two genetic clusters (Fig. S18H).

***M. zonifer*.** The phylogenetic tree revealed two main clades: Clade A comprises populations of Michoacán (MIC), Santiago (SAN), Vallarta (VAL), and Zacatoso (ZAC), while Clade B is represented by individuals from San Agustinillo (AGU) (Fig. S19A). The fastStructure analyses delineated two major groups (k=2) that align with clades A and B in the phylogenetic tree (Fig. S19B). These patterns of population structure suggest a putative new marine break (PNB2) separating the population of AGU. In contrast, ADMIXTURE analyses failed to detect any population structure (Fig. S19C). PCA and DAPC analyses, with and without *a priori* groupings, identified AGU as a distinct genetic cluster evidenced by PC1 axis (25.0–93.7%) (Fig. S19D–F). The PC2 axis in DAPC analyses, conducted without *a priori* grouping, further differentiates ZAC and MIC from SAN and VAL (Fig. S19E). EEMS plots identified two barriers to genetic flow: one positioned between SAN and MIC, and another near AGU (Fig. S19H). The sPCA analyses recover two main genetic groups: one composed by AGU and ZAC; and another encompassing MIC, SAN, and VAL (Fig. S19H).

Tropical Atlantic

***M. aurolineatus*.** We excluded the Puerto Rican population (PRI) due to an excess of missing data (>95%). The phylogenetic tree identified three main clades: clade A comprises Virgin Islands (VIR) and Trinidad and Tobago (TRT); clade B includes Curaçao (CUR); and clade C encompasses the populations of San Andrés (AND), Belize (BLZ), Bahamas (BAH), Cayman Islands (CAY), Panama (PAN), and Yucatán (YUC) (Fig. S20A). The genetic differentiation observed in clade B suggests that there may be a putative new break east of CUR (PNB9), while within clade C, the monophyletic group containing solely the population of AND also suggests the existence of a possible new break in that area (PNB7). FastStructure analysis suggested an optimal scenario (*k*) of three populations: one cluster comprising populations east of the Eastern Caribbean Break (ECB), a second cluster including the VIR and TRT, and a third distinct cluster representing CUR (Fig. S20B). In contrast, the ADMIXTURE analysis identified $k = 9$ as the optimal scenario, indicating a higher degree of genetic partitioning than FastStructure within the first cluster. This analysis revealed genetic flow among the CAY and AND, and the BAH, and YUC and BLZ. Notably, the ADMIXTURE analysis also suggested connectivity between VIR and TRT, with CUR remaining as an independent group (Fig. S20C). Rather than indicating population structure obstructed by the marine breaks under examination, the genetic partitions observed by the ADMIXTURE analysis suggest cross-population genetic flow in the TA, with the exception of CUR. Despite the DAPC analysis without *a priori* grouping proposing 11 genetic clusters, the DAPC plot aligns with the FastStructure results in revealing population groupings. This genetic clustering hypothesis was also confirmed by DAPC with a priori grouping, PCA analysis, and the phylogenomic tree. In this scenario, PCA analysis revealed that PC1 segregates the populations of TRT, VIR, and CUR (23.6%), while PC2 isolates CUR (12.6%) (Fig. S20D). Both DAPC analyses, with and without *a priori* groupings, emphasize CUR as a distinct cluster on PC1 (61.8–82.6%), with PC2 (17.6–34.5%) further distinguishing VIR and TRT from the rest (Fig. S20E–F). EEMS analysis identified two main areas impeding genetic flow, one corresponding to ECB, and the other to the Yucatán Current Break (YCB) (Fig. S20G). However, due to insufficient sample size, we could not conclusively confirm the influence of the YCB on the population structure of this species. Finally, the sPCA plot supports the delineation of two main clusters, in alignment with the ECB (Fig. S20H).

***M. boehlkei*.** We excluded the populations from Honduras and Puerto Rico due to high levels of missing data (>95%). The phylogenomic tree identified the individuals from Virgin Islands (VIR) and an independent evolutionary lineage (Fig. S21A). FastStructure and ADMIXTURE analyses indicated the best number of populations (*k*) as three, with similar results pointing to San Andrés (AND), and VIR forming two independent clusters each, while Belize (BLZ) and Yucatán (YUC) formed the third cluster. These analyses reveal genetic flow from AND to BLZ, suggesting a minimal influence from the Western Caribbean Break (WCB) and a significant one from the Eastern Caribbean Break (ECB) (Fig. S21B–C). PCA and DAPC analyses corroborated the fastStructure and ADMIXTURE findings. PCA analyses show that PC1 distinguishes VIR from the other the populations (21.3%), and PC2 (15.2%) suggest a pattern of isolation by distance (IBD) among the YUC, BLZ, and AND populations (Fig. S21D). DAPC analyses with *a priori* grouping reflect these patterns, with PC1 (96.3%) separating VIR, and PC2 (3.7%) mirroring the

PCA pattern (Fig. S21E). Without *a priori* grouping, DAPC identified five genetic clusters, with their placement on the PCA plot aligning with other analyses (Fig. S21F). EEMS plots suggest a barrier that is consistent with the ECB, separating VIR from other populations. YUC, BLZ, and AND show a deviation from an IBD pattern (blue shading in Fig. S21G). However, sPCA analyses contrastingly group BLZ and YUC, as well as AND and VIR, as distinct genetic units, supporting an influence by the WCB but not the ECB (Fig. S21H).

***M. erdmani*.** The phylogenomic tree recovered five clades, primarily clustered by geographically proximate populations. The first clade consists of individuals from Belize (BLZ), Yucatán (YUC), and Honduras (HON); the second includes San Andrés (SAN); the third groups individuals from the Bahamas (BAH) and Cayman Islands (CAY); the fourth contains BAH; and the fifth joins Puerto Rico (PRI) and Dominica (DOM). However, the genetic distance between these clades was low (<0.02) (Fig. S22A). Both fastStructure and ADMIXTURE analyses suggest a single metapopulation with an optimal *k* scenario of 1, indicating no apparent population structure (Fig. S22B–C). PCA and DAPC analyses, with and without *a priori* grouping, show similar results with the PC1 axis separating PRI and DOM populations from the rest (34.6–75.1%), and PC2 distinguishing a finer structure between the populations of BLZ, HON, and YUC from AND, BAH, and CAY (18.6–27%) (Fig. S22D–F). These patterns of population suggest that the Bahamas Break (BB), Gulf of Mexico Break (GMB), and Yucatán Current Break (YCB) influence the population connectivity among BAH, and CAY, and HON, BLZ, and YUC. Meanwhile, the Western Caribbean Break (WCB) divides these populations from AND, where a potential new break (PNB7), appears to isolate that population. On the other hand, the Eastern Caribbean Break (ECB) seems to hinder population connectivity between PRI and DOM versus the rest. EESM analyses revealed three barriers hindering genetic flow, the first corresponding to the ECB, YCB and another one representing a previously unidentified barrier (PNB11) (Fig. S22G). In contrast to previous results, sPCA analyses suggest two primary genetic clusters: PRI, DOM, BAH, and AND; and CAY, YUC, BLZ, and HON (Fig. S22H). The results of the AMOVA analyses comparing two, three, and four grouping scenarios find the highest proportion of genetic variance within individuals (73.29–80.88%; Table S8). The variation between the three geographic groups (Eastern Caribbean, Puerto Rico, Dominica) was 17.69%, which was the higher compared to the two- and four-grouping scenarios.

***M. gilli*.** We excluded the Virgin Islands population (VIR) due to an excess of missing data (>95%). The phylogenomic tree identifies two primary clades: clade A, encompassing populations from San Andrés (AND), the Bahamas (BAH), Belize (BLZ), Yucatán (YUC), Cayman Islands (CAY), and Puerto Rico (PRI); and clade B, comprising Dominica (DOM) and Trinidad and Tobago (TRT). These patterns of genetic structure support the influence of a putative new break (PNB11) located between DOM and PRI, which hinders population connectivity between such clades. Additionally, subclades within clades A and B were clustered by geographic locality, following a model of isolation by distance (IBD) (Fig. S23A), where the genetic groups observed align with the Eastern Caribbean Break (ECB), the Bahamas Break (BB) and the Gulf of Mexico Break (GMB), all of which hinder population connectivity. FastStructure and ADMIXTURE analyses echo this finding, suggesting two optimal populations (*k*), one consisting of DOM and TRI, and the second containing the remaining populations, with minimal genetic flow from DOM-

TRI to PRI (Fig. S23B–C). PCA and DAPC analyses, with and without *a priori* groupings, corroborate this genetic clustering pattern. PC1 distinguishes the populations of DOM and TRI, explaining 59.8–78.0% of the variation, while PC2 isolates PRI as a distinct genetic group (0.07–27.3%) (Fig. S23D–F). Despite DAPC analyses without *a priori* grouping identifying nine independent genetic groups, they clustered on the DAPC plot as per PCA analyses. EEMS analyses pinpoint two areas impeding genetic flow across populations, the ECB and an area near to the CAY close to the Yucatán Current Break (YCB) (Fig. S23G). sPCA analyses also distinguish the populations of DOM and TRI as a unique genetic cluster apart from other populations (Fig. S23H). Finally, the AMOVA analyses of the two- and three-grouping scenarios (Table S8) revealed that the highest proportion of significant genetic variation (76.57–76.82%) was observed between groups. Even though these results were marginal, the highest level of genetic differentiation was observed on the three-grouping. On the other hand, the genetic variation between individuals within groups was relatively low (4.28–7.93%), indicating that the groups are quite distinct from each other.

***M. macropus*.** The phylogenomic tree revealed two main clades: clade A, encompassing populations of Puerto Rico (PRI) and Virgin Islands (VIR), and clade B containing Florida (FLO), San Andrés (AND), the Bahamas (BAH), Belize (BLZ), Cayman Islands (CAY), Yucatán (YUC), and Panama (PAN) as a distinct subclade (Fig. S24A). FastStructure and ADMIXTURE analyses discerned two metapopulations in alignment with the phylogenomic tree, with ADMIXTURE pinpointing limited genetic flow from PRI-VIR to BLZ and BAH (Fig. S24B–C). PCA and DAPC analyses, both with and without *a priori* groupings, separated PRI and VIR as a distinct genetic cluster on the PC1 axis, accounting for 14.7–71.2% of the variation. However, these analyses highlighted nuanced distinctions in the finer structure of populations (Fig. S24D–F). Specifically, the PCA's PC2 axis (0.05%) marked PAN as a unique genetic cluster (Fig. S24D), while the DAPC's PC2 axis (21.0%), with *a priori* groupings, grouped PAN with the rest of the populations but set FLO apart as a unique genetic entity (Fig. S24E). Without *a priori* groupings, DAPC analyses suggested that PAN and FLO were both slightly distinct from the remaining localities (Fig. S24F). The genetic structure observed across these analyses corresponds to the influence of the Eastern Caribbean Break (ECB), which separates the populations of PRI-VIR from the other populations, while the Bahamas and Gulf of Mexico Breaks (BB and GMB respectively) appear to isolate the FLO population, and a putative new barrier (PNB8) appears to segregate PAN. EEMS plots indicate two barriers to genetic flow, corresponding to the BB and the ECB. sPCA plots echo prior findings, depicting PRI and VIR as an independent genetic group (Fig. S24H). Finally, AMOVA analyses comparing two- and three-grouping scenarios (with PAN or FLO representing the third group) reflect a consistent pattern where a significant proportion of genetic variation (25.27–34.27%; Table S8) is found between groups. In the two-grouping scenario, PRI-VIR are significantly different and represent the highest portion of genetic variation (34.37%), aligning with a strong effect by the ECB. Overall, the largest proportion of genetic variation is found within individuals across all scenarios (61.68–71.33%), indicating a high degree of genetic diversity at the individual level.

***M. triangulatus*.** The phylogenomic tree recovers three main clades. Clade A encompasses three subclades: the first consists of Florida (FLO) and Veracruz (VER); the second is made up of Belize (BLZ) and Yucatán (YUC); and the third includes Panama (PAN), Limón (LIM), and Tayrona (TAY). Clade B also comprises three subclades: the first is represented by San Andrés (AND); the second includes individuals from Mona (MON), the Bahamas (BAH), and Cayman Islands (CAY); and the third subclade includes Virgin Islands (VIR), Puerto Rico (PRI), Curaçao (CUR), and Dominica (DOM). Lastly, clade C showcases three primary subclades: the first represented by Trinidad and Tobago (TRT), the second by Fernando da Noronha (NOR), and the third also by individuals from TRT (Fig. S25A). FastStructure analyses identified three main clusters of populations as the optimal *k* scenario. The first cluster is comprised of populations from FLO, BLZ, VER, YUC, LIM, PAN, TAY, and AND; the second group consists of BAH, CAY, MON, PRI, VIR, CUR, and DOM; the third group is made up of TRT and NOR. These analyses identify genetic flow from populations in the second group to AND, as well as two individuals from TRT that exhibit a shared genetic makeup with the other two genetic groups (Fig. S25B). ADMIXTURE analyses recovered a higher genetic partitioning, suggesting a best *k* scenario of 5 populations. Here, the first group identified by fastStructure is further subdivided, with the populations of LIM, PAN, TAY, and AND emerging as a separate genetic cluster. The second fastStructure group splits into two clusters: CUR and DOM form an independent group, whereas PRI and VIR display approximately 50% genetic information shared with DOM and CUR. However, the cross-validation results from ADMIXTURE were marginal, pointing to $k = 6$ as another potential scenario. Under this framework, AND constitutes its own distinct cluster (Fig. S25C). Both PCA and DAPC analyses, conducted with and without *a priori* groupings, revealed similar patterns in which PC1 distinguishes the populations of TRT and NOR from the others, accounting for 18.1 to 33.9% of the variance (Fig. S25D–F). The PC2 axis in the PCA, accounting for 11.4%, segregates four major groups: the first includes populations from FLO, VER, BLZ, and YUC; the second contains individuals from LIM, PAN, and TAY; the third is composed of AND, ant the fourth encompasses the remaining populations (Fig. S25D). For DAPC's PC2 (21.4%), estimated with *a priori* groupings, a slightly different partitioning is evident, with only three groups: the first comprising FLO and VEN; the second consisting of YUC, TAY, AND, PAN, and LIM; and the third integrated by VIR, PRI, CUR, DOM, and CAY (Fig. S25E). In contrast, DAPC's PC2 (11.3%), without *a priori* groupings, differentiates the TRT and NOR group from the individuals from TRT who exhibited admixture patters (Fig. S25F). The EEMS plot highlights several areas impeding genetic flow. These areas correspond to previously recognized marine breaks, including the Bahamas Break (BB), Yucatán Break (YUC), Western Caribbean Break (WCB), Mona Passage Break (MPB), the Eastern Caribbean Break (ECB). These analyses were also in agreement with the identification of two potentially new breaks: one between the coast of Honduras-El Salvador and Cayman Islands (PNB6), and another between DOM and TRT (PNB10) (Fig. S25G). In this scenario, the sPCA analyses similarly identify TRT and NOR as a distinct genetic group supporting PNB10 as a potential marine break for this species (Fig. S25H).

***M. versicolor*.** The phylogenomic tree distinguishes each of the three geographic locations as a distinct lineage (Fig. S26A). The fastStructure analyses align with the phylogenomic tree, revealing three main genetic clusters (k=3), each corresponding to a specific geographical location (Fig. S26B). Conversely, ADMIXTURE's cross-validation procedure suggests an optimal *k*

scenario of two populations. The first group comprises individuals from the Bahamas (BAH), while the second encompasses individuals from Yucatán (YUC). Interestingly, the sole? Puerto Rico (PRI) individual sequenced displays genetic admixture from both groups (Fig. S26C). Both PCA and DAPC, with and without *a priori* groupings, also recover three genetic clusters, consistent with the geographic locations and the phylogenomic tree (Fig. S26D–F). The PC1 axis clearly segregates all clusters, accounting for 43.5 to 64.5% of the overall variation, while the PC2 axis primarily differentiates the PRI population, representing 14.1 to 36.6% of the variance. The EEMS plot highlights an area hindering population connectivity that coincides with the Bahamas Break (BB) (Fig. S26G), while the sPCA plot signals YUC population as a unique genetic cluster (Fig. S26H).

***M. costaricanus*, *M. gigas*, *M. margaritae*, and *M. zonogaster*.** For these species, we collected several individuals but only in one or two localities across the TEP. We estimated phylogenomic trees to evaluate the degree of genetic disparity (see Supplementary Materials and Methods). As expected, the phylogenomic trees failed to recover any significant structure among individuals from either a single population (*M. costaricanus* and *M. margaritae*) or two geographically proximate populations (*M. gigas* and *M. zonogaster*) (Fig. S27).

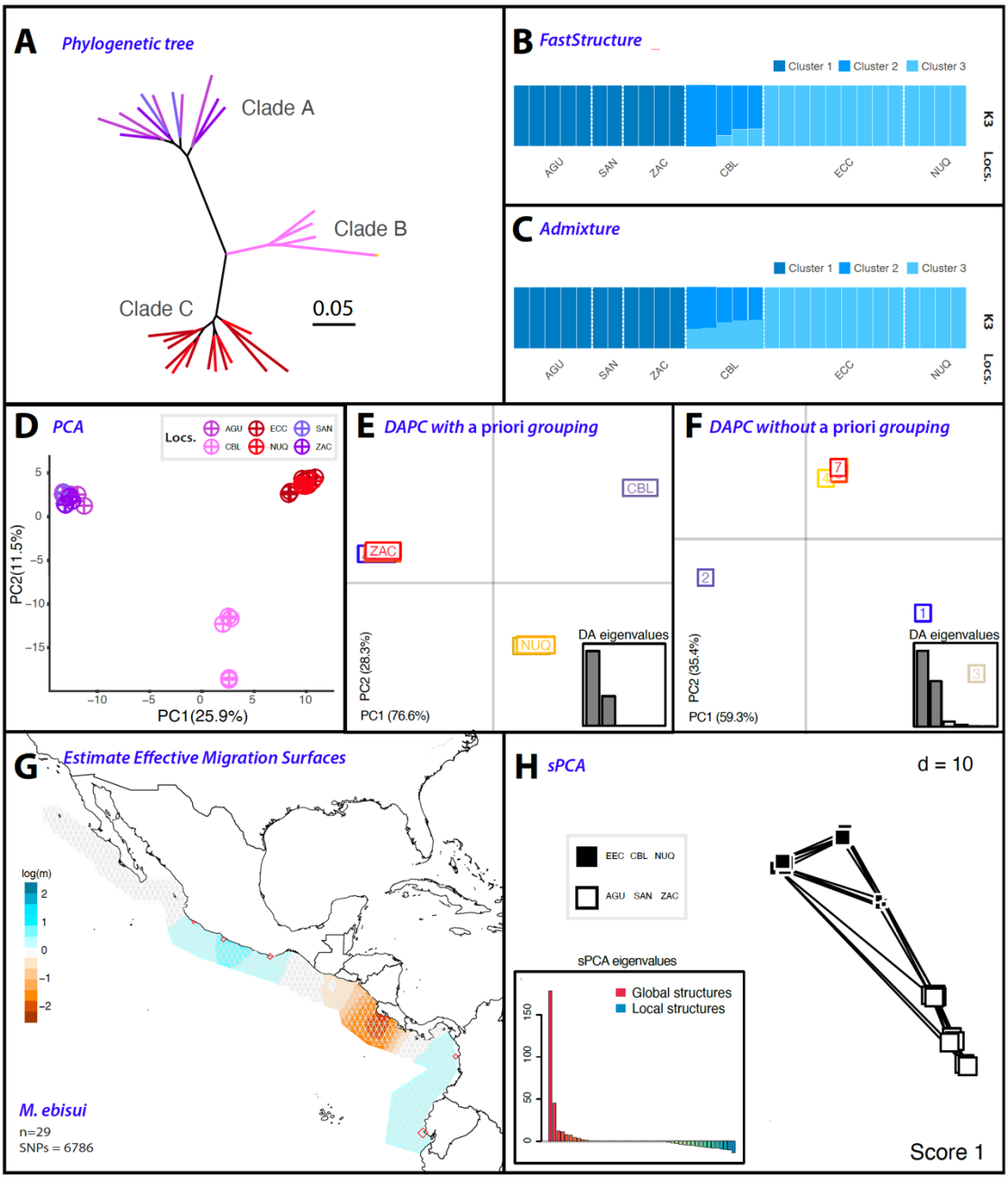


Figure S13. Population differentiation analyses on *M. ebisui* based on 6,786 SNPs. A) Phylogenetic tree, with branches are color-coded to match sampling sites in Fig. 3.1. B & C) FastStructure and ADMIXTURE analyses, respectively, highlight the best *k* scenario. Each bar denotes an individual; colors indicate the inferred membership of each genomic group. D) Principal component analysis (PCA) where each point represents an individual, color-coded by their population. E & F) Discriminant analysis of principal components (DAPC), both with and without *a priori* grouping, respectively. G) Model of the estimated effective migration surfaces (EEMS) illustrating the spatial structure of the populations. White areas correlate with estimates from an isolation by distance (IBD) model, while migration rates (*m*) above average are in blue, and those below average are in brown. H) Spatial principal component analysis (sPCA) illustrate global and local genotype patterns. Each square symbolizes a population, positioned according to its spatial coordinates and color-coded based on its genetic cluster.

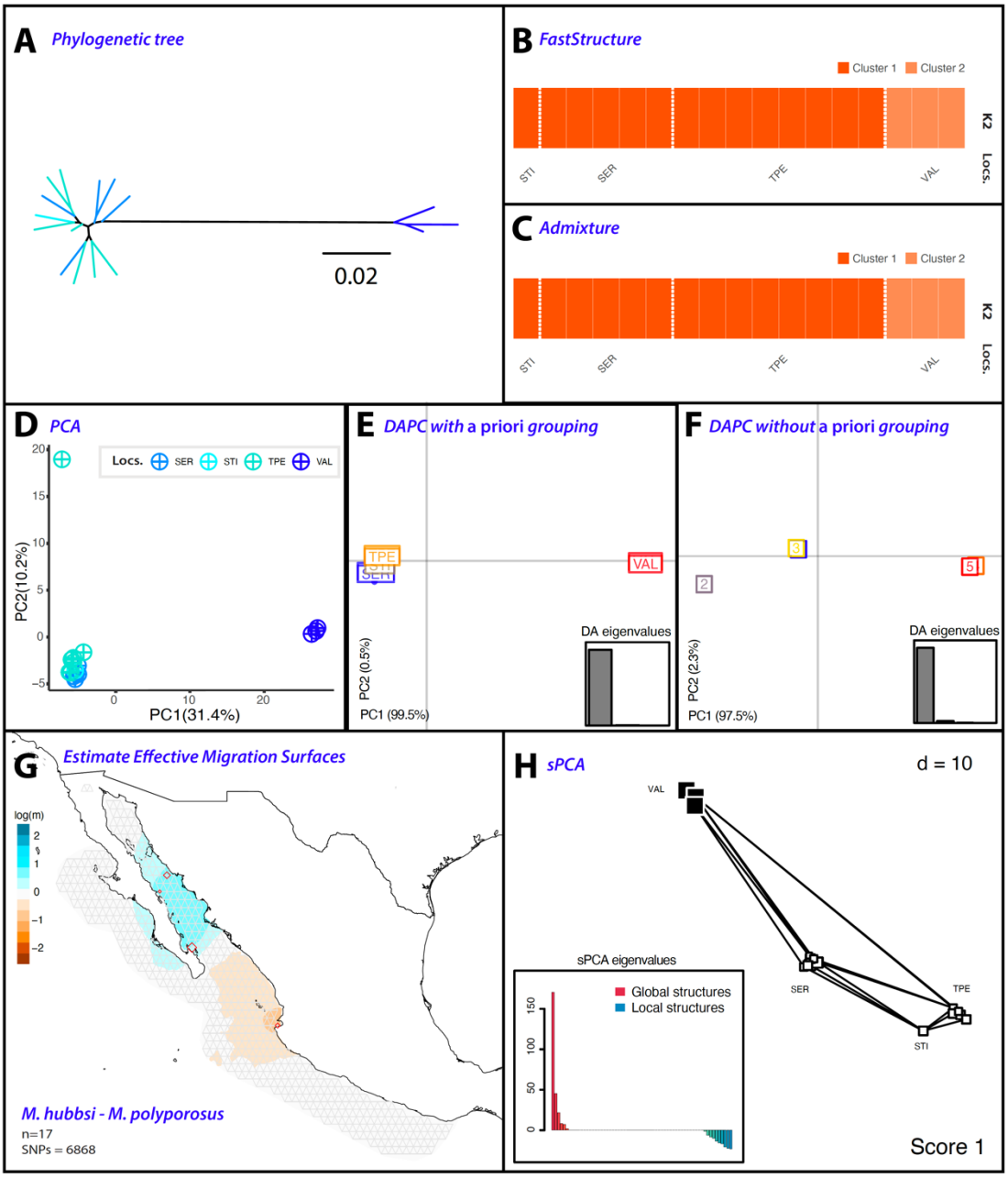


Figure S14. Population differentiation analyses on *M. hubbsi-M. polyporosus* based on 6,868 SNPs. A) Phylogenetic tree, with branches are color-coded to match sampling sites in Fig. 3.1. B & C) FastStructure and ADMIXTURE analyses, respectively, highlight the best *k* scenario. Each bar denotes an individual; colors indicate the inferred membership of each genomic group. D) Principal component analysis (PCA) where each point represents an individual, color-coded by their population. E & F) Discriminant analysis of principal components (DAPC), both with and without *a priori* grouping, respectively. G) Model of the estimated effective migration surfaces (EEMS) illustrating the spatial structure of the populations. White areas correlate with estimates from an isolation by distance (IBD) model, while migration rates (*m*) above average are in blue, and those below average are in brown. H) Spatial principal component analysis (sPCA) illustrate global and local genotype patterns. Each square symbolizes a population, positioned according to its spatial coordinates and color-coded based on its genetic cluster.

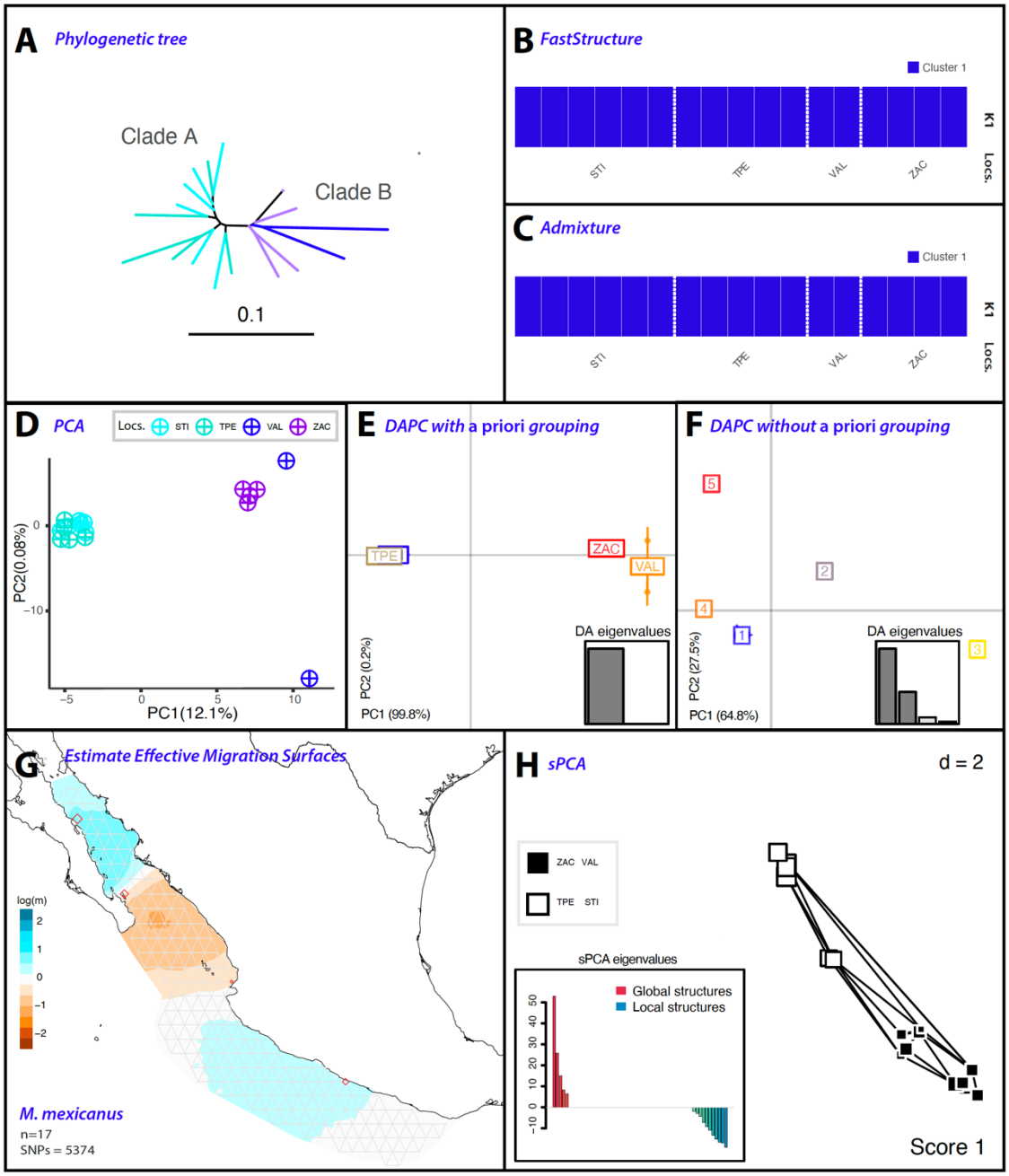


Figure S15. Population differentiation analyses on *M. mexicanus* based on 5,374 SNPs. A) Phylogenetic tree, with branches are color-coded to match sampling sites in Fig. 3.1. B & C) FastStructure and ADMIXTURE analyses, respectively, highlight the best *k* scenario. Each bar denotes an individual; colors indicate the inferred membership of each genomic group. D) Principal component analysis (PCA) where each point represents an individual, color-coded by their population. E & F) Discriminant analysis of principal components (DAPC), both with and without *a priori* grouping, respectively. G) Model of the estimated effective migration surfaces (EEMS) illustrating the spatial structure of the populations. White areas correlate with estimates from an isolation by distance (IBD) model, while migration rates (*m*) above average are in blue, and those below average are in brown. H) Spatial principal component analysis (sPCA) illustrate global and local genotype patterns. Each square symbolizes a population, positioned according to its spatial coordinates and color-coded based on its genetic cluster.

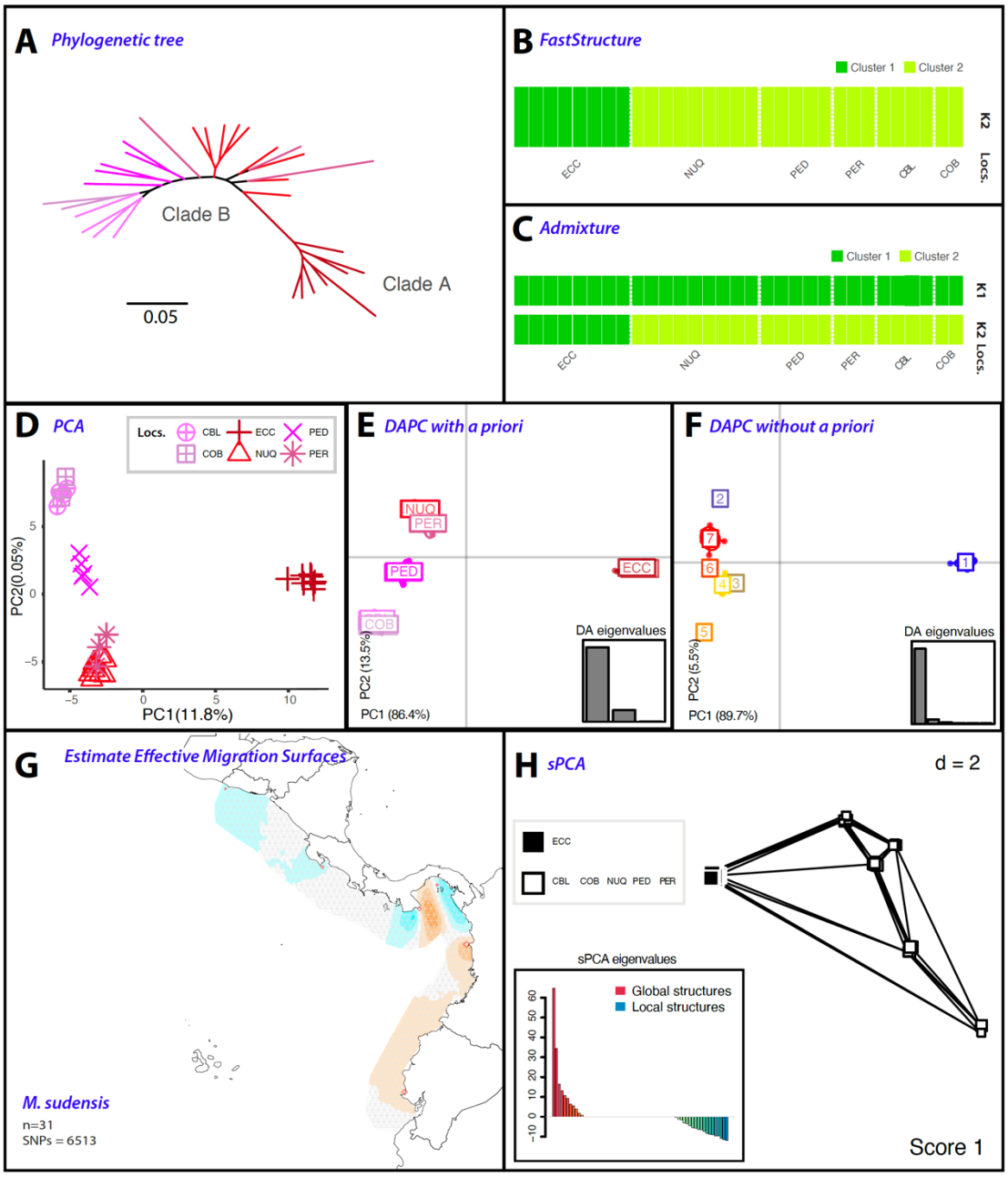


Figure S16. Population differentiation analyses on *M. sudensis* based on 6,513 SNPs. A) Phylogenetic tree, with branches are color-coded to match sampling sites in Fig. 3.1. B & C) FastStructure and ADMIXTURE analyses, respectively, highlight the best *k* scenario. Each bar denotes an individual; colors indicate the inferred membership of each genomic group. D) Principal component analysis (PCA) where each point represents an individual, color-coded by their population. E & F) Discriminant analysis of principal components (DAPC), both with and without *a priori* grouping, respectively. G) Model of the estimated effective migration surfaces (EEMS) illustrating the spatial structure of the populations. White areas correlate with estimates from an isolation by distance (IBD) model, while migration rates (*m*) above average are in blue, and those below average are in brown. H) Spatial principal component analysis (sPCA) illustrate global and local genotype patterns. Each square symbolizes a population, positioned according to its spatial coordinates and color-coded based on its genetic cluster.

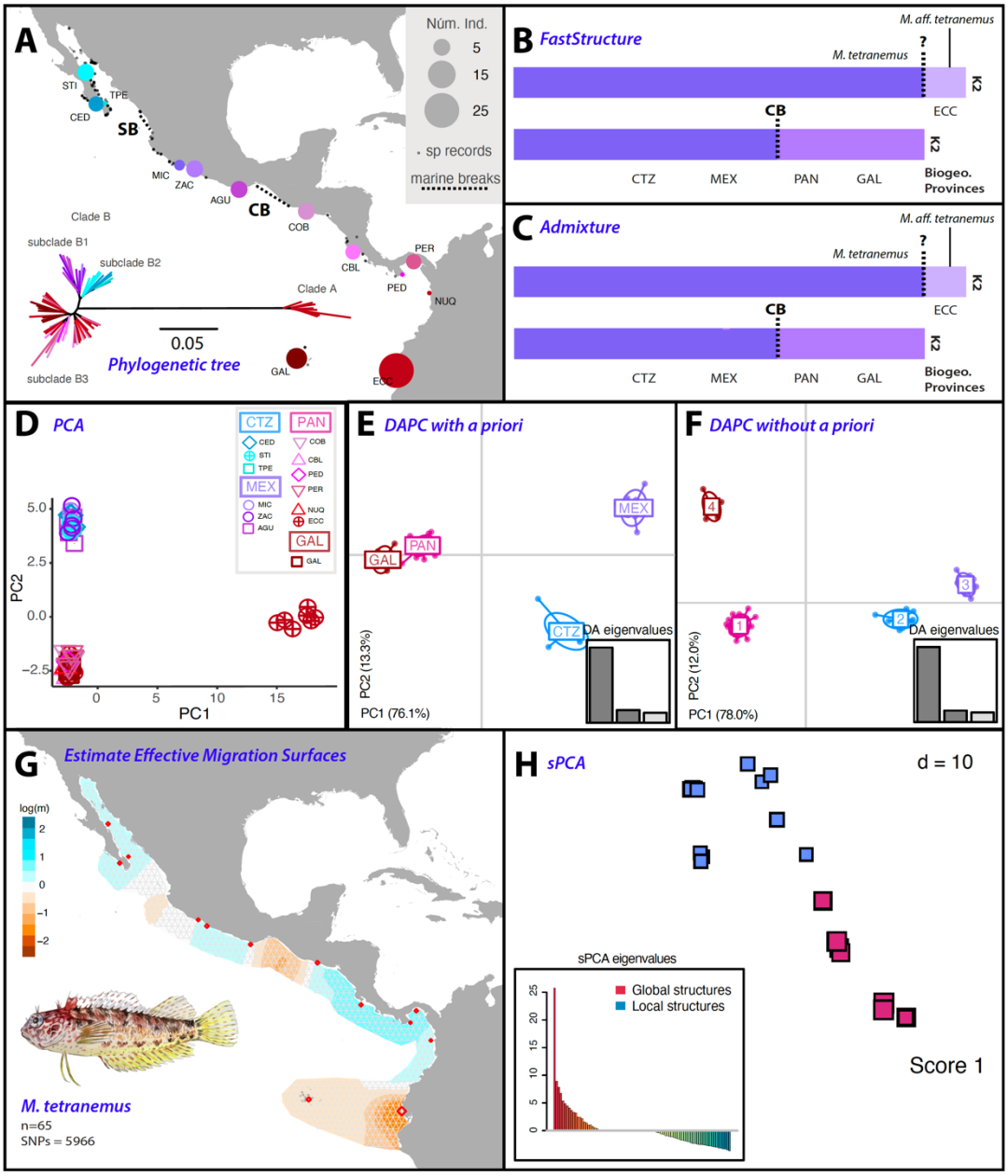


Figure S17. Population differentiation analyses on *M. tetranemus* based on 5,996 SNPs. A) Map showing sampling sites and the proportion of individuals collected from each. Adjacent to the map is a phylogenetic tree, with branches are color-coded to match sampling sites. Dashed lines on the map indicate marine barriers to dispersal evaluated for this species. B & C) FastStructure and ADMIXTURE analyses, respectively, highlight the best *k* scenario. Each bar denotes an individual; colors indicate the inferred membership of each genomic group. D) Principal component analysis (PCA) where each point represents an individual, color-coded by their population. E & F) Discriminant analysis of principal components (DAPC), both with and without *a priori* grouping, respectively. G) Model of the estimated effective migration surfaces (EEMS) illustrating the spatial structure of the populations. White areas correlate with estimates from an isolation by distance (IBD) model, while migration rates (*m*) above average are in blue, and those below average are in brown. H) Spatial principal component analysis (sPCA) illustrate global and local genotype patterns. Each square symbolizes a population, positioned according to its spatial coordinates and color-coded based on its genetic cluster.

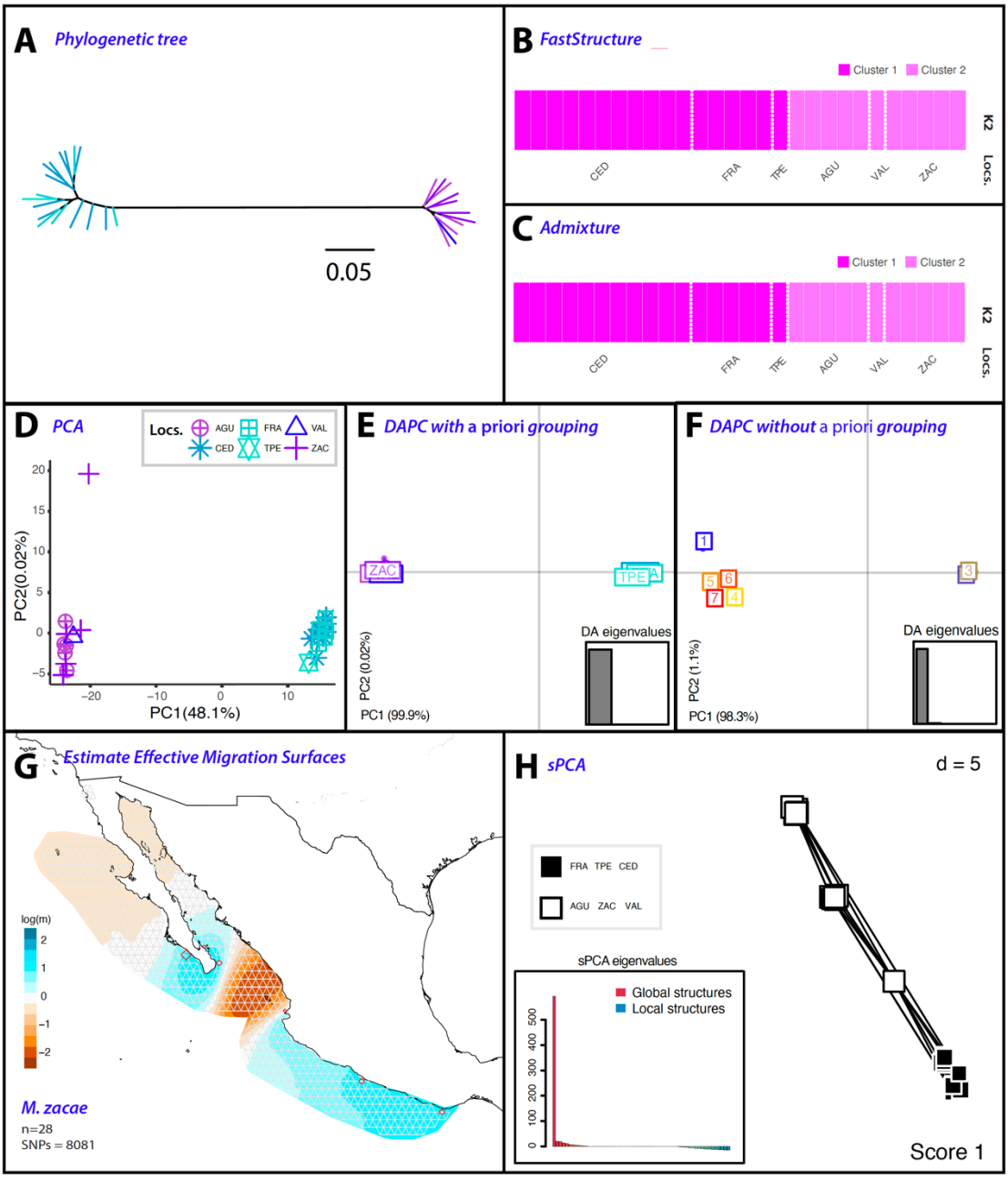


Figure S18. Population differentiation analyses on *M. zacae* based on 8,081 SNPs. A) Phylogenetic tree, with branches are color-coded to match sampling sites in Fig. 3.1. B & C) FastStructure and ADMIXTURE analyses, respectively, highlight the best *k* scenario. Each bar denotes an individual; colors indicate the inferred membership of each genomic group. D) Principal component analysis (PCA) where each point represents an individual, color-coded by their population. E & F) Discriminant analysis of principal components (DAPC), both with and without *a priori* grouping, respectively. G) Model of the estimated effective migration surfaces (EEMS) illustrating the spatial structure of the populations. White areas correlate with estimates from an isolation by distance (IBD) model, while migration rates (*m*) above average are in blue, and those below average are in brown. H) Spatial principal component analysis (sPCA) illustrate global and local genotype patterns. Each square symbolizes a population, positioned according to its spatial coordinates and color-coded based on its genetic cluster.

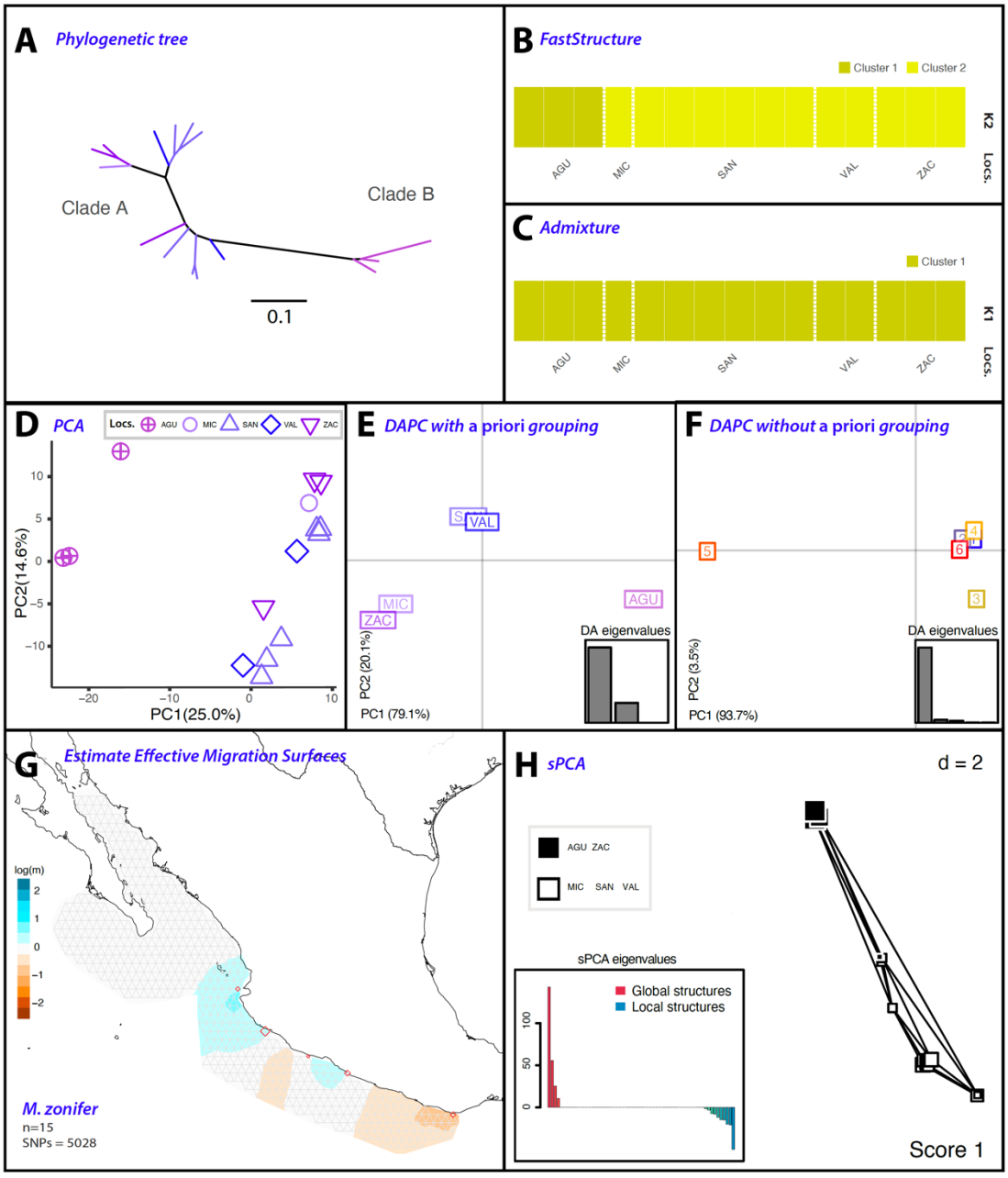


Figure S19. Population differentiation analyses on *M. zonifer* based on 5,028 SNPs. A) Phylogenetic tree, with branches are color-coded to match sampling sites in Fig. 3.1. B & C) FastStructure and ADMIXTURE analyses, respectively, highlight the best *k* scenario. Each bar denotes an individual; colors indicate the inferred membership of each genomic group. D) Principal component analysis (PCA) where each point represents an individual, color-coded by their population. E & F) Discriminant analysis of principal components (DAPC), both with and without *a priori* grouping, respectively. G) Model of the estimated effective migration surfaces (EEMS) illustrating the spatial structure of the populations. White areas correlate with estimates from an isolation by distance (IBD) model, while migration rates (*m*) above average are in blue, and those below average are in brown. H) Spatial principal component analysis (sPCA) illustrate global and local genotype patterns. Each square symbolizes a population, positioned according to its spatial coordinates and color-coded based on its genetic cluster.

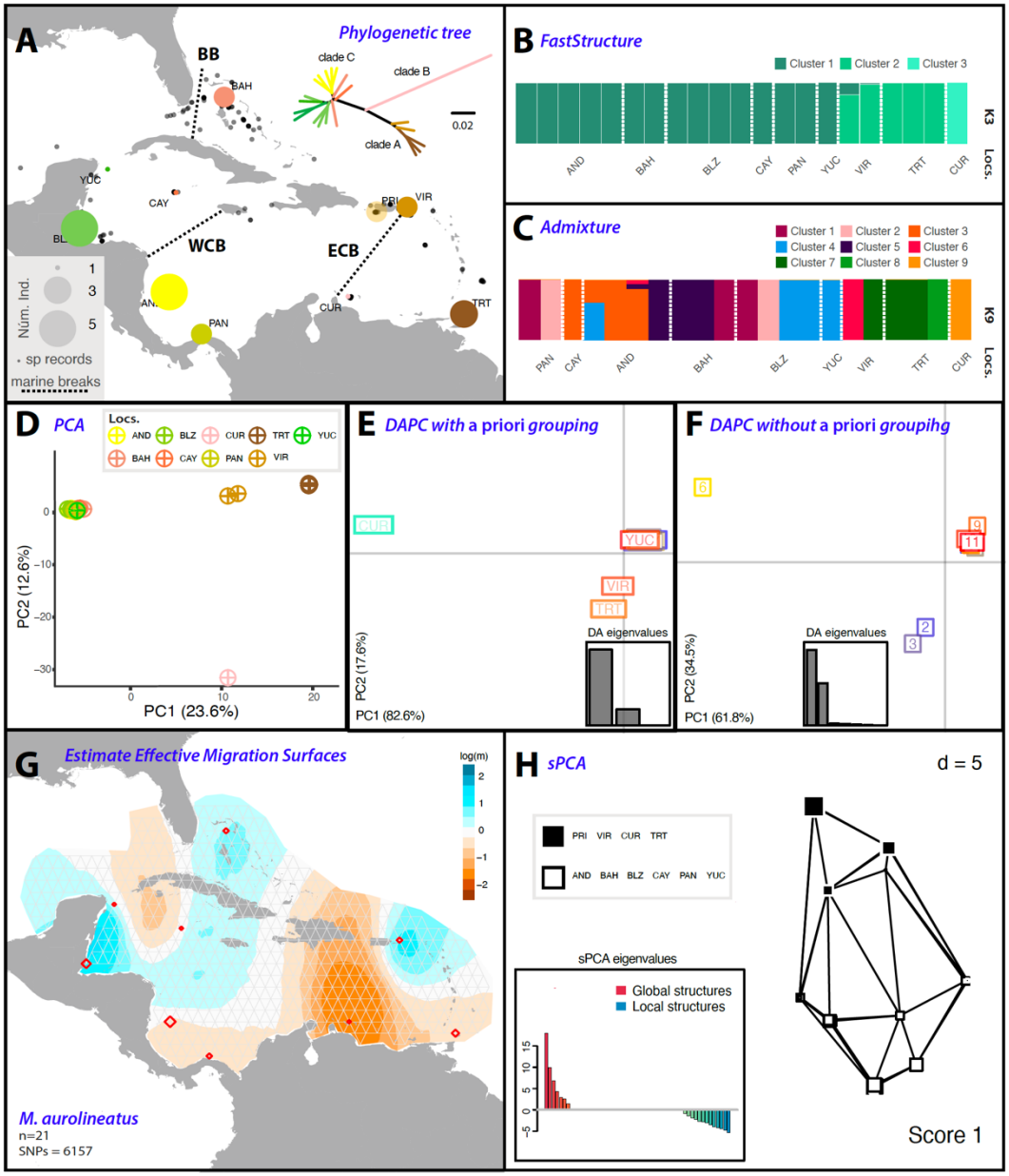


Figure S20. Population differentiation analyses on *M. aurolineatus* based on 6,157 SNPs. A) Map showing sampling sites and the proportion of individuals collected from each. Adjacent to the map is a phylogenetic tree, with branches are color-coded to match sampling sites. Dashed lines on the map indicate marine barriers to dispersal evaluated for this species. B & C) FastStructure and ADMIXTURE analyses, respectively, highlight the best *k* scenario. Each bar denotes an individual; colors indicate the inferred membership of each genomic group. D) Principal component analysis (PCA) where each point represents an individual, color-coded by their population. E & F) Discriminant analysis of principal components (DAPC), both with and without *a priori* grouping, respectively. G) Model of the estimated effective migration surfaces (EEMS) illustrating the spatial structure of the populations. White areas correlate with estimates from an isolation by distance (IBD) model, while migration rates (*m*) above average are in blue, and those below average are in brown. H) Spatial principal component analysis (sPCA) illustrate global and local genotype patterns. Each square symbolizes a population, positioned according to its spatial coordinates and color-coded based on its genetic cluster.

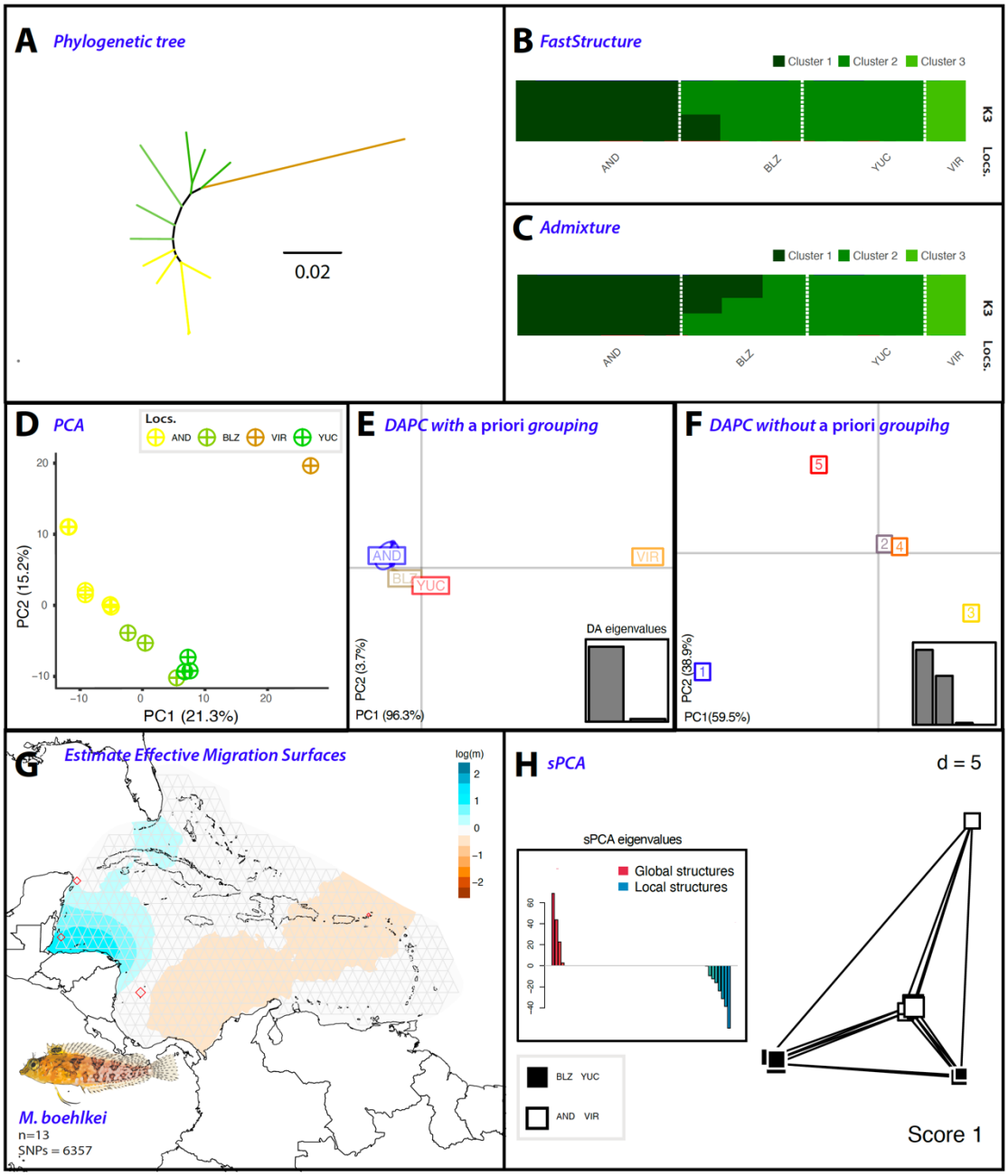


Figure S21. Population differentiation analyses on *M. boehlkei* based on 6,357 SNPs. A) Phylogenetic tree, with branches are color-coded to match sampling sites in Fig. 3.1. B & C) FastStructure and ADMIXTURE analyses, respectively, highlight the best *k* scenario. Each bar denotes an individual; colors indicate the inferred membership of each genomic group. D) Principal component analysis (PCA) where each point represents an individual, color-coded by their population. E & F) Discriminant analysis of principal components (DAPC), both with and without *a priori* grouping, respectively. G) Model of the estimated effective migration surfaces (EEMS) illustrating the spatial structure of the populations. White areas correlate with estimates from an isolation by distance (IBD) model, while migration rates (*m*) above average are in blue, and those below average are in brown. H) Spatial principal component analysis (sPCA) illustrate global and local genotype patterns. Each square symbolizes a population, positioned according to its spatial coordinates and color-coded based on its genetic cluster.

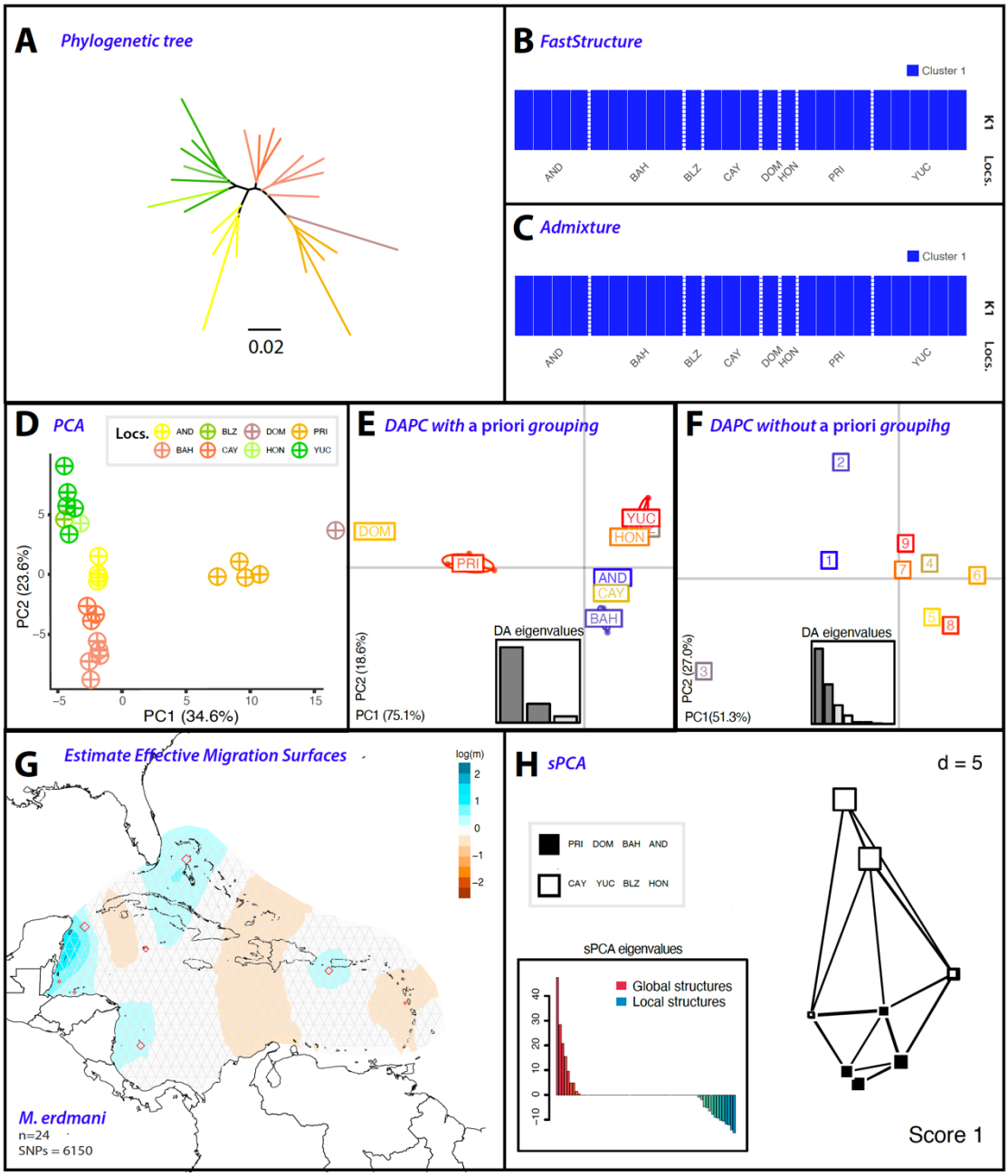


Figure S22. Population differentiation analyses on *M. erdmani* based on 6,150 SNPs. A) Phylogenetic tree, with branches are color-coded to match sampling sites in Fig. 3.1. B & C) FastStructure and ADMIXTURE analyses, respectively, highlight the best *k* scenario. Each bar denotes an individual; colors indicate the inferred membership of each genomic group. D) Principal component analysis (PCA) where each point represents an individual, color-coded by their population. E & F) Discriminant analysis of principal components (DAPC), both with and without *a priori* grouping, respectively. G) Model of the estimated effective migration surfaces (EEMS) illustrating the spatial structure of the populations. White areas correlate with estimates from an isolation by distance (IBD) model, while migration rates (*m*) above average are in blue, and those below average are in brown. H) Spatial principal component analysis (sPCA) illustrate global and local genotype patterns. Each square symbolizes a population, positioned according to its spatial coordinates and color-coded based on its genetic cluster.

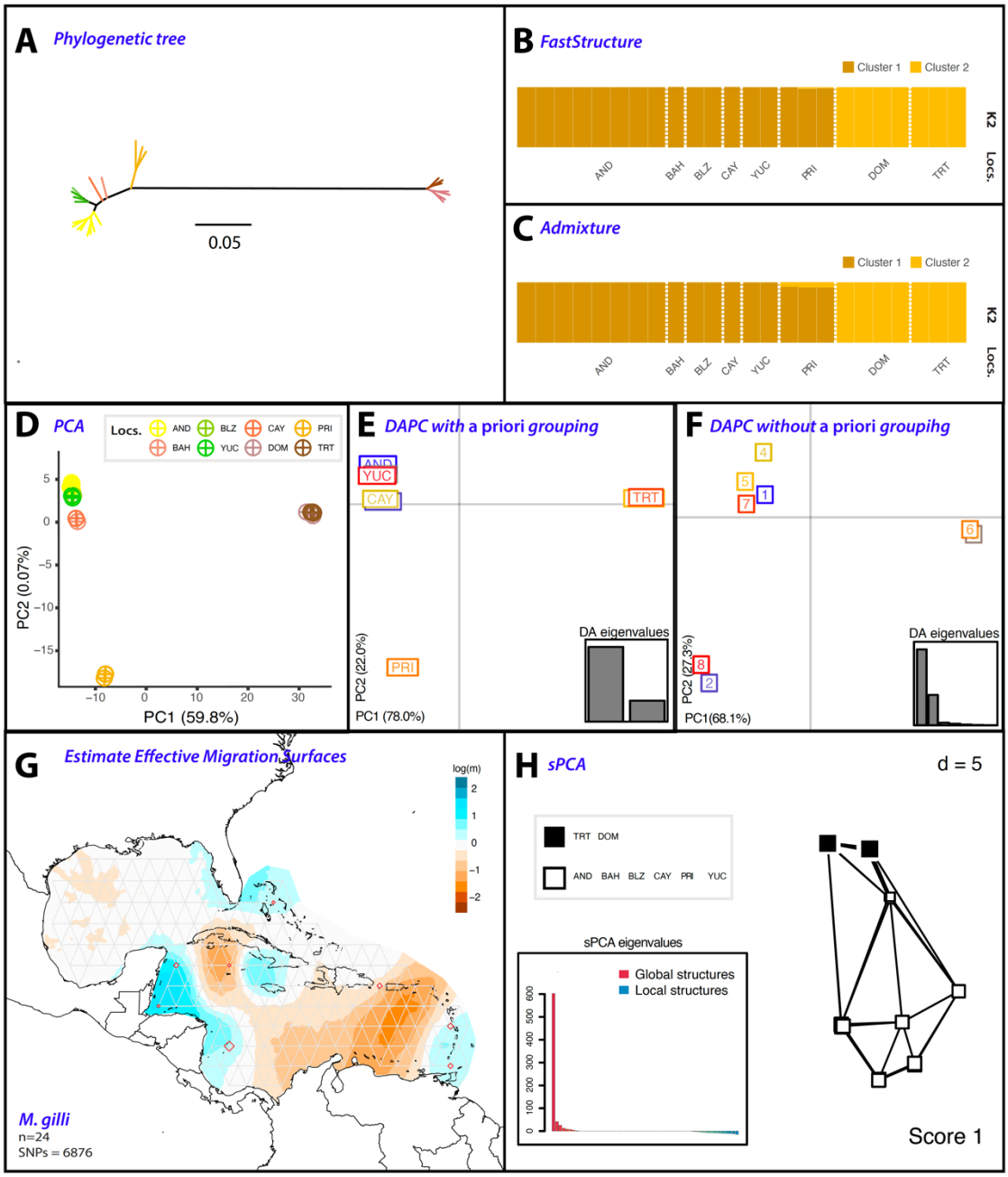


Figure S23. Population differentiation analyses on *M. gilli* based on 6,876 SNPs. A) Phylogenetic tree, with branches are color-coded to match sampling sites in Fig. 3.1. B & C) FastStructure and ADMIXTURE analyses, respectively, highlight the best *k* scenario. Each bar denotes an individual; colors indicate the inferred membership of each genomic group. D) Principal component analysis (PCA) where each point represents an individual, color-coded by their population. E & F) Discriminant analysis of principal components (DAPC), both with and without *a priori* grouping, respectively. G) Model of the estimated effective migration surfaces (EEMS) illustrating the spatial structure of the populations. White areas correlate with estimates from an isolation by distance (IBD) model, while migration rates (*m*) above average are in blue, and those below average are in brown. H) Spatial principal component analysis (sPCA) illustrate global and local genotype patterns. Each square symbolizes a population, positioned according to its spatial coordinates and color-coded based on its genetic cluster.

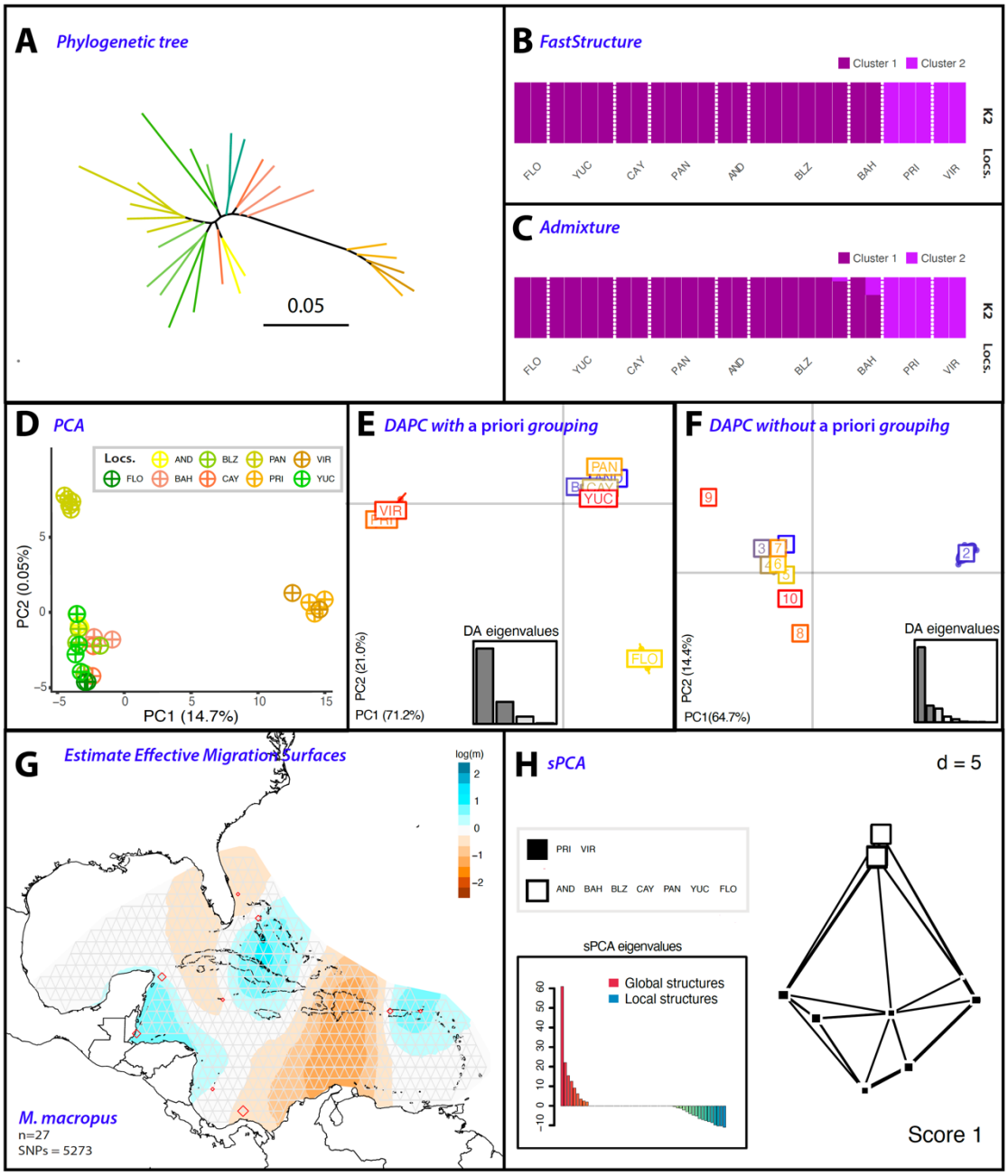


Figure S24. Population differentiation analyses on *M. macropus* based on 5,273 SNPs. A) Phylogenetic tree, with branches are color-coded to match sampling sites in Fig. 3.1. B & C) FastStructure and ADMIXTURE analyses, respectively, highlight the best *k* scenario. Each bar denotes an individual; colors indicate the inferred membership of each genomic group. D) Principal component analysis (PCA) where each point represents an individual, color-coded by their population. E & F) Discriminant analysis of principal components (DAPC), both with and without *a priori* grouping, respectively. G) Model of the estimated effective migration surfaces (EEMS) illustrating the spatial structure of the populations. White areas correlate with estimates from an isolation by distance (IBD) model, while migration rates (*m*) above average are in blue, and those below average are in brown. H) Spatial principal component analysis (sPCA) illustrate global and local genotype patterns. Each square symbolizes a population, positioned according to its spatial coordinates and color-coded based on its genetic cluster.

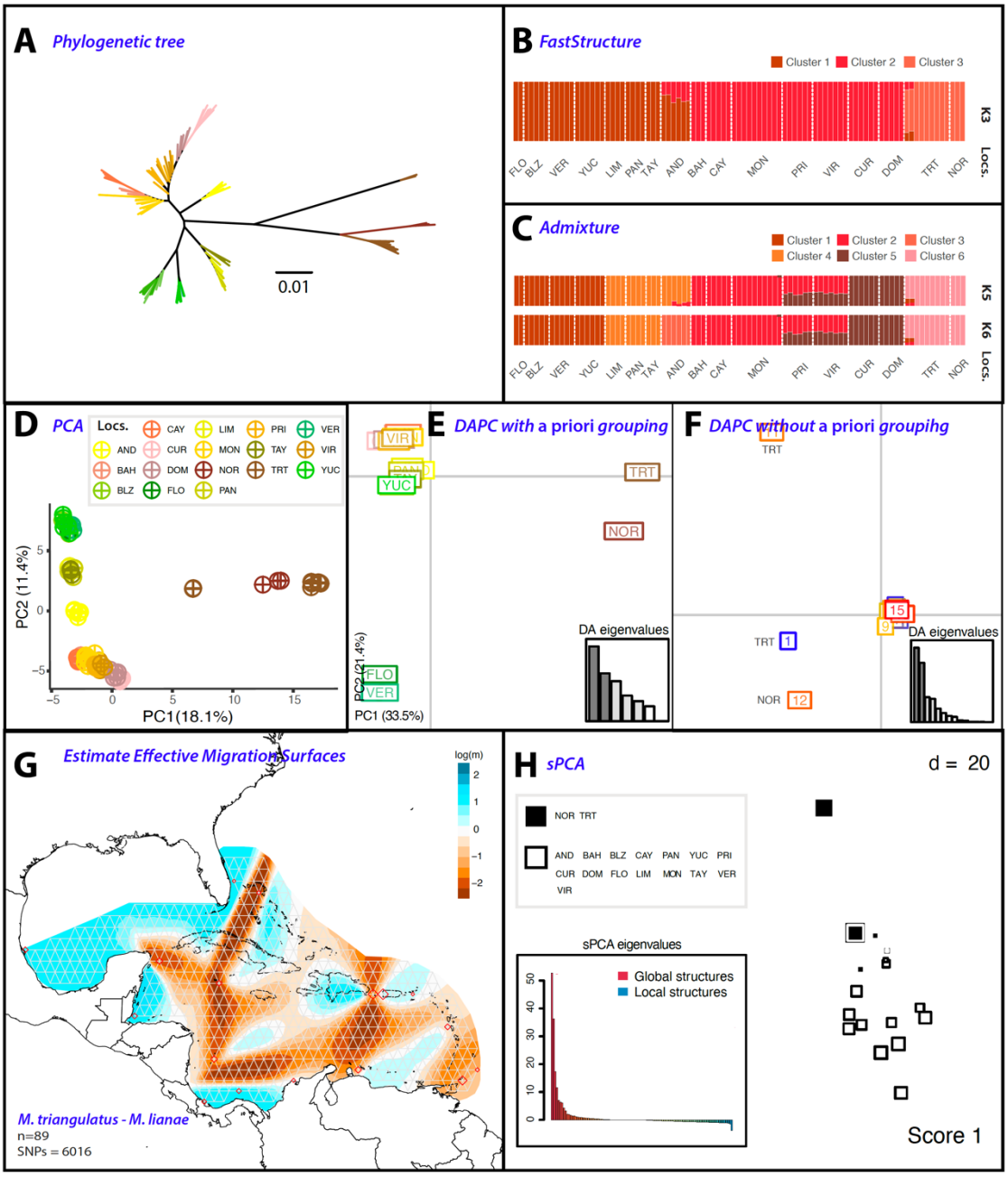


Figure S25. Population differentiation analyses on *M. triangulatus-M. lianae* based on 6,016 SNPs. A) Phylogenetic tree, with branches are color-coded to match sampling sites in Fig. 3.1. B & C) FastStructure and ADMIXTURE analyses, respectively, highlight the best *k* scenario. Each bar denotes an individual; colors indicate the inferred membership of each genomic group. D) Principal component analysis (PCA) where each point represents an individual, color-coded by their population. E & F) Discriminant analysis of principal components (DAPC), both with and without *a priori* grouping, respectively. G) Model of the estimated effective migration surfaces (EEMS) illustrating the spatial structure of the populations. White areas correlate with estimates from an isolation by distance (IBD) model, while migration rates (*m*) above average are in blue, and those below average are in brown. H) Spatial principal component analysis (sPCA) illustrate global and local genotype patterns. Each square symbolizes a population, positioned according to its spatial coordinates and color-coded based on its genetic cluster.

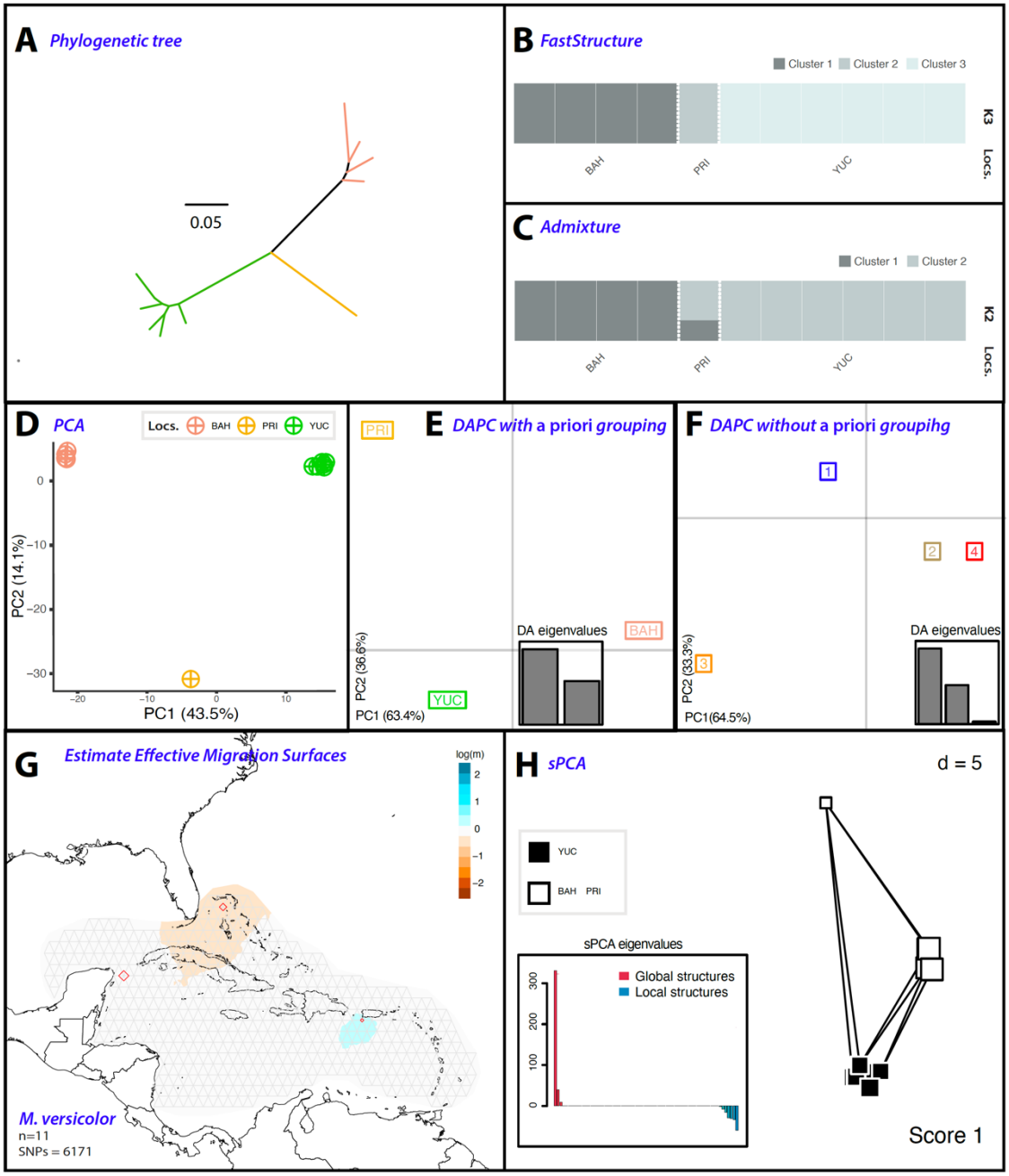


Figure S26. Population differentiation analyses on *versicolor* based on 6,171 SNPs. A) Phylogenetic tree, with branches are color-coded to match sampling sites in Fig. 3.1. B & C) FastStructure and ADMIXTURE analyses, respectively, highlight the best *k* scenario. Each bar denotes an individual; colors indicate the inferred membership of each genomic group. D) Principal component analysis (PCA) where each point represents an individual, color-coded by their population. E & F) Discriminant analysis of principal components (DAPC), both with and without *a priori* grouping, respectively. G) Model of the estimated effective migration surfaces (EEMS) illustrating the spatial structure of the populations. White areas correlate with estimates from an isolation by distance (IBD) model, while migration rates (*m*) above average are in blue, and those below average are in brown. H) Spatial principal component analysis (sPCA) illustrate global and local genotype patterns. Each square symbolizes a population, positioned according to its spatial coordinates and color-coded based on its genetic cluster.

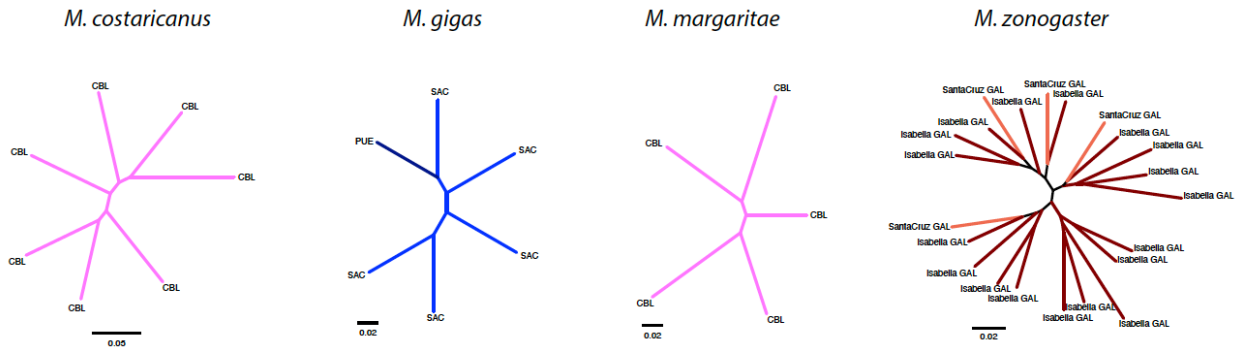


Figure S27. Phylogenetic trees of *M. costaricanus, M. gigas, M. margaritae* and *M. zonogaster.*

Table S8. Hierarchical analysis of molecular variance (AMOVA) partitioning genetic variation according to each population clustering hypothesis. Total variance of the observations (Sum sq); proportion of total variance explained at each hierarchy (%var); and levels of genetic differentiation sigma (ϕ_{ST}). Analyses for the species in the TEP and TA .

Source of Variation	Source of Variation	Sum sq	%var	ϕ_{ST}
***M. tetranemus* – 2 groups**	Between groups	1276.90	11.27	0.14*
Group A: Gulf of California and Mexico	Between individuals within groups	3995.99	3.17	0.04*
Group B: Central and South America	Within individuals	11860.74	85.55	0.11*
***M. tetranemus* – 3 groups**				
Group A: Gulf of California	Between groups	1738.20	11.83	0.13*
Group B: Mexico	Between individuals within groups	3534.69	2.12	0.02*
Group C: South America and Galapagos	Within individuals	11860.75	86.05	0.11*
***M. tetranemus* – 4 groups**				
Group A: Gulf of California	Between groups	2209.72	11.60	0.12*
Group B: Mexico	Between individuals within groups	3063.17	0.82	0.01
Group C: South America	Within individuals	11860.75	87.57	0.11*
Group D: Galapagos				
***M. sudensis* – 2 groups**	Between groups	1549.45	17.02	0.24*
Group A: Continental Ecuador	Between individuals within groups	2007.22	7.80	0.09*
Group B: South America	Within individuals	8686.30	75.18	0.17*
***M. sudensis* – 4 groups**				
Group A: Costa Rica and El Salvador	Between groups	2710.92	13.67	0.19
Group B: Continental Panama	Between individuals within groups	845.75	4.93	0.05*
Group C: Continental Ecuador	Within individuals	8686.30	81.40	0.13*
Group D: Panama Islands and Colombia				
***M. erdmani* – 2 groups**	Between groups	720.74	13.94	0.25*
Group A: Eastern Caribbean	Between individuals within groups	2033.37	11.16	0.13*
Group B: Western Caribbean	Within individuals	3843.64	74.90	0.14
***M. erdmani* – 3 groups**	Between groups	1110.76	17.69	0.26*
Group A: Eastern Caribbean	Between individuals within groups	1643.35	9.02	0.11*
Group B: Puerto Rico	Within individuals	3843.65	73.29	0.17
Group C: Dominica				
***M. erdmani* – 4 groups**				
Group A: Belize, Honduras, Yucatán	Between groups	1541.57	9.14	0.19*
Group B: San Andrés Island	Between individuals within groups	1212.54	9.98	0.11
Group C: Bahamas and Cayman Islands	Within individuals	3843.65	80.88	0.09

Group D: Puerto Rico and Dominica				
***M. gilli* – 2 groups** **Group A:** Eastern Caribbean **Group B:** Dominica and Trinidad and Tobago	Between groups Between individuals within groups Within individuals	13696.62 3674.00 4214.46	76.57 7.93 15.50	0.85* 0.34* 0.77
***M. gilli* – 3 groups** **Group A:** Eastern Caribbean **Group B:** Puerto Rico **Group C:** Dominica and Trinidad and Tobago	Between groups Between individuals within groups Within individuals	15329.11 2041.51 4214.46	76.82 4.28 18.90	0.81* 0.18* 0.77
***M. macropus* – 2 groups** **Group A:** Eastern Caribbean **Group B:** Western Caribbean	Between groups Between individuals within groups Within individuals	741.99 1085.89 2337.99	34.27 4.06 61.68	0.38* 0.06* 0.34
***M. macropus* – 3 groups-PAN** **Group A:** Eastern Caribbean **Group B:** Western Caribbean **Group C:** Panama	Between groups Between individuals within groups Within individuals	925.71 1020.89 2219.27	25.27 3.41 71.33	0.29* 0.05 0.25
***M. macropus* – 3 groups-FLO** **Group A:** Eastern Caribbean **Group B:** Western Caribbean **Group C:** Florida	Between groups Between individuals within groups Within individuals	923.11 904.77 2337.99	29.31 3.51 67.19	0.33* 0.05 0.29
***M. triangulatus-M. lianae* – 2 groups** **Group A:** *M. triangulatus* **Group B:** *M. triangulatus* (TRT) and *M. lianae*	Between groups Between individuals within groups Within individuals	107.55 46.44 60.34	50.18 21.69 28.15	0.72* 0.44* 0.50
***M. triangulatus-M. lianae* – 3 groups** **Group A:** *M. triangulatus* **Group B:** *M. lianae* **Group C:** *M. triangulatus* (TRT)	Between groups Between individuals within groups Within individuals	3033.82 4018.15 4344.66	51.58 20.57 27.85	0.72* 0.42* 0.52
***M. triangulatus-M. lianae* – 6 groups** **Group A:** *M. lianae* and *M. triangulatus* (TRT) **Group B:** CUR, DOM **Group C:** BAH, CAY, MON, PRI, VIR **Group D:** San Andrés **Group E:** Western Caribbean **Group F:** Gulf of Mexico	Between groups Between individuals within groups Within individuals	5330.07 1721.90 4344.66	44.47 13.58 41.95	0.58* 0.24* 0.44*
***M. triangulatus-M. lianae* – 9 groups** **Group A:** *M. lianae* **Group B:** *M. triangulatus* (TRT) **Group C:** TRT hybrids **Group D:** San Andrés **Group E:** BAH, CAY, MON **Group F:** FLO, VER **Group G:** PRI, VIR, DOM, CUR **Group H:** PAN, LIM, TAY **Group I:** BLZ, YUC	Between groups Between individuals within groups Within individuals	6942.46 906.21 3547.96	57.71 7.26 35.02	* * *

$p < 0.001$*

PHENOTIPIC ANALYSES OF BODY SHAPE

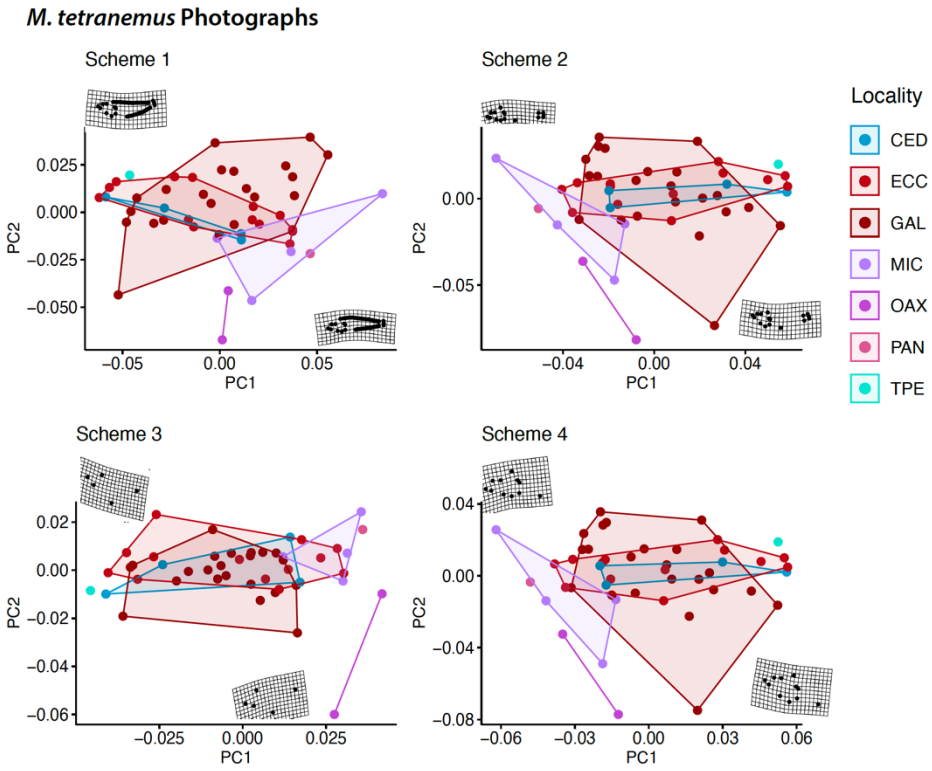


Figure S28. Geometric morphometric analyses of *M. tetranemus*. A Scheme 1, landmarks and semi-landmarks, B) Landmarks only, C) head only, and D) anterior body shape.

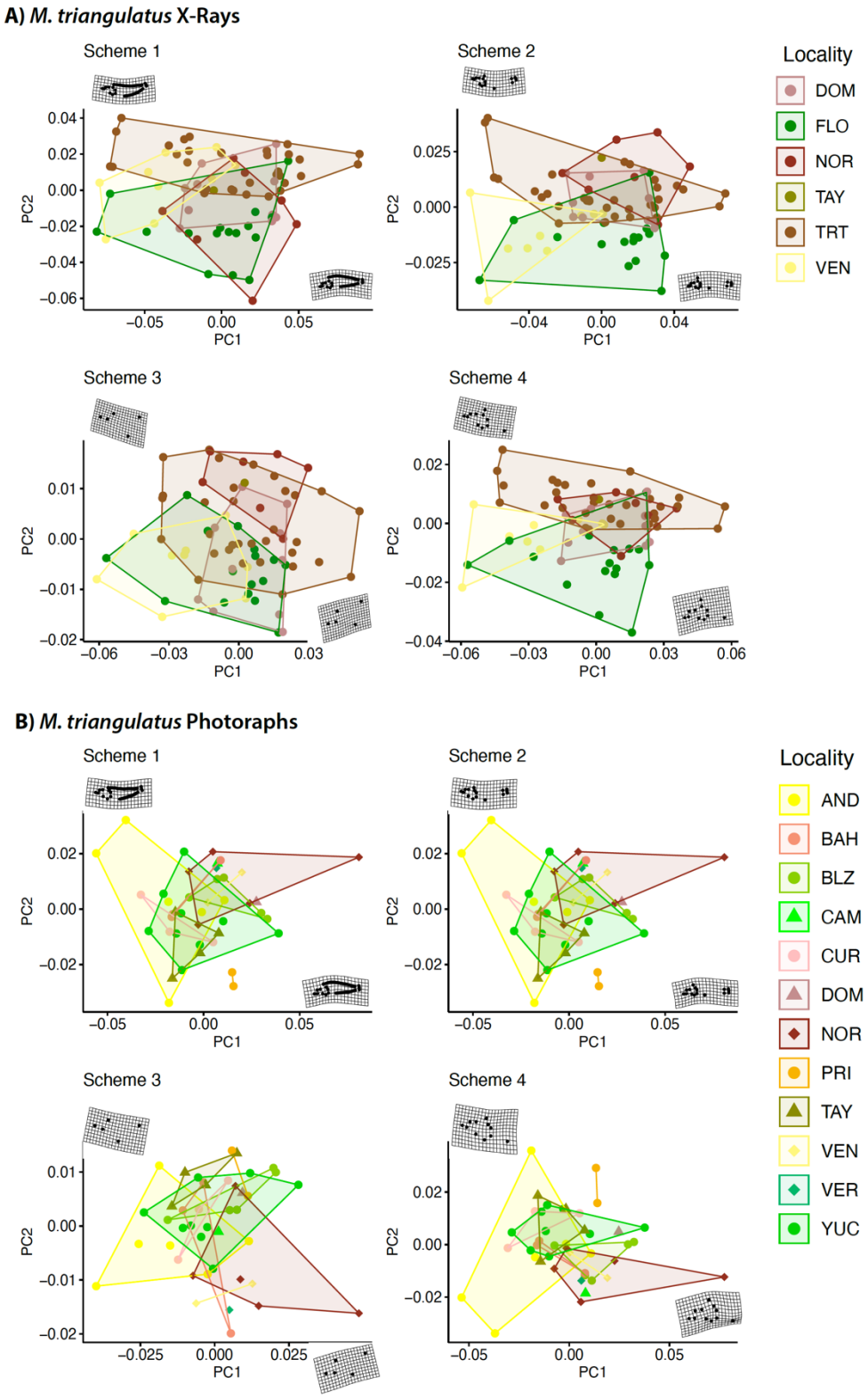


Figure S29. Geometric morphometric analyses of *M. triangulatus*. A Scheme 1, landmarks and semi-landmarks, B) Landmarks only, C) head only, and D) anterior body shape.

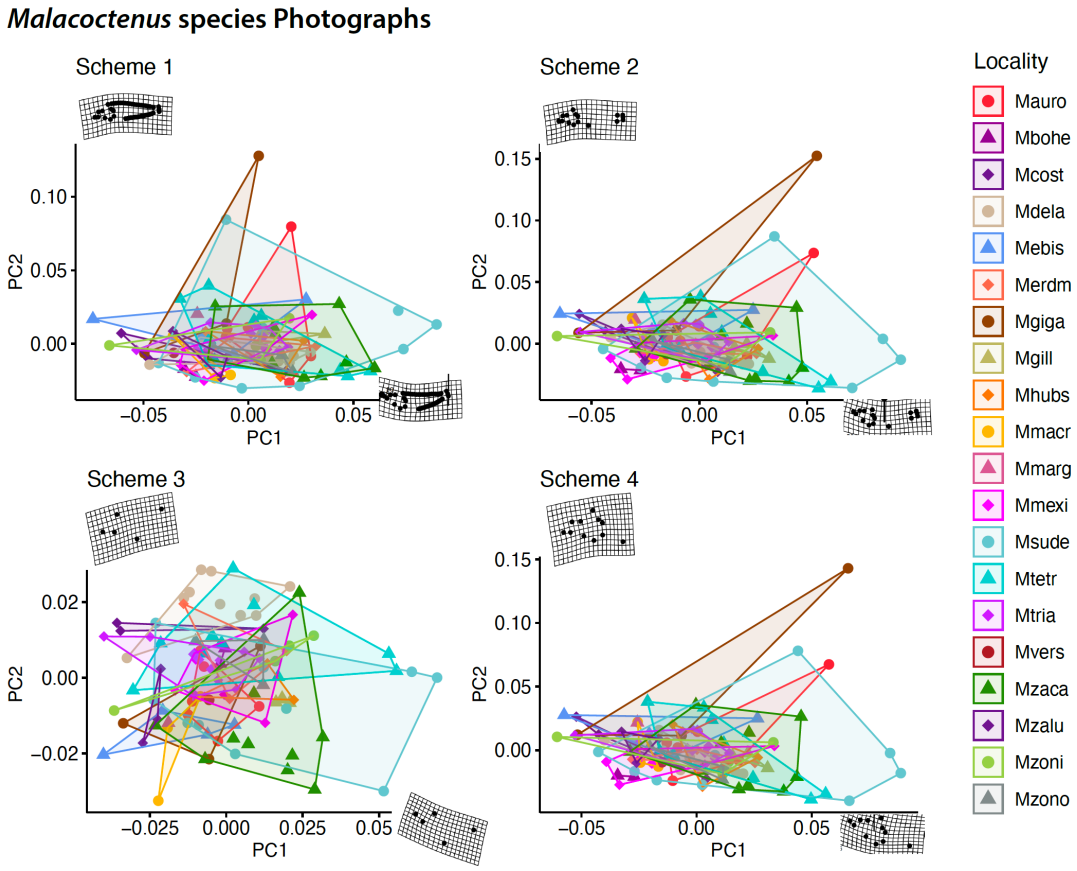


Figure S30. Geometric morphometric analyses of 20 *Malacoctenus* species. A Scheme 1, landmarks and semi-landmarks, B) Landmarks only, C) head only, and D) anterior body shape.

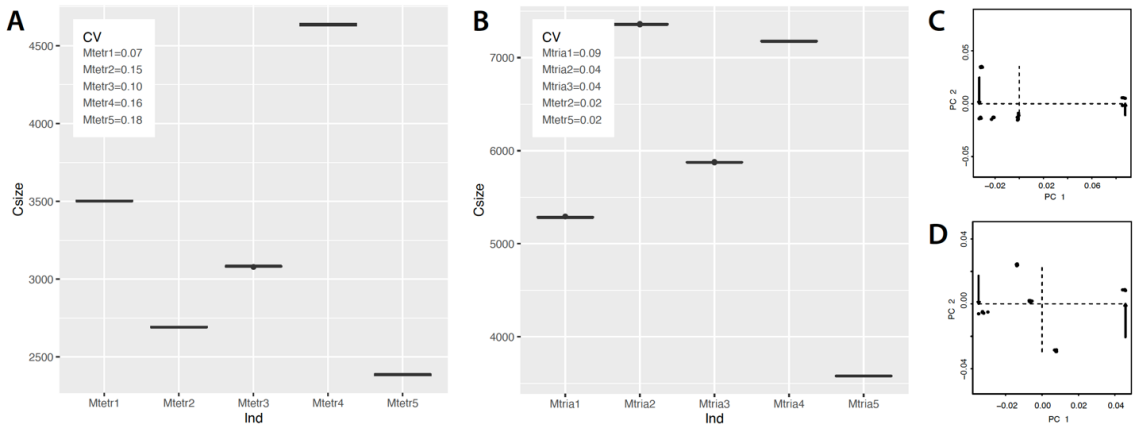


Figure S31. Error rate of landmark digitization, measured using the centroid size from five replicates for each of five individuals of: A) *M. tetranemus* and B) *M. triangulatus*. Each plot displays centroid size values and the coefficient of variation (CV) for each individual. C and D) depict the morphospaces for *M. tetranemus* and *M. triangulatus*, respectively.

SEASCAPE GENOMICS

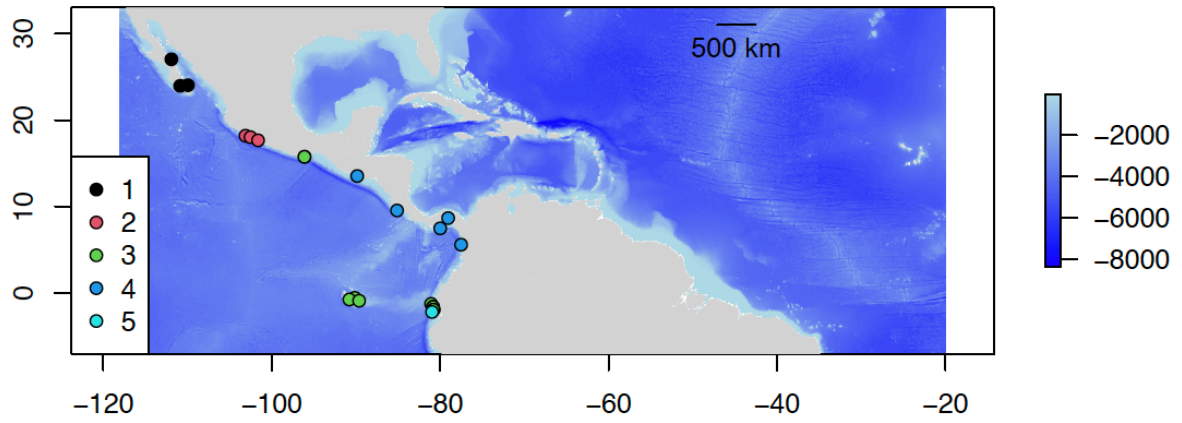


Figure S32. Environmental clusters of the sampling locations in the Tropical Eastern Pacific.

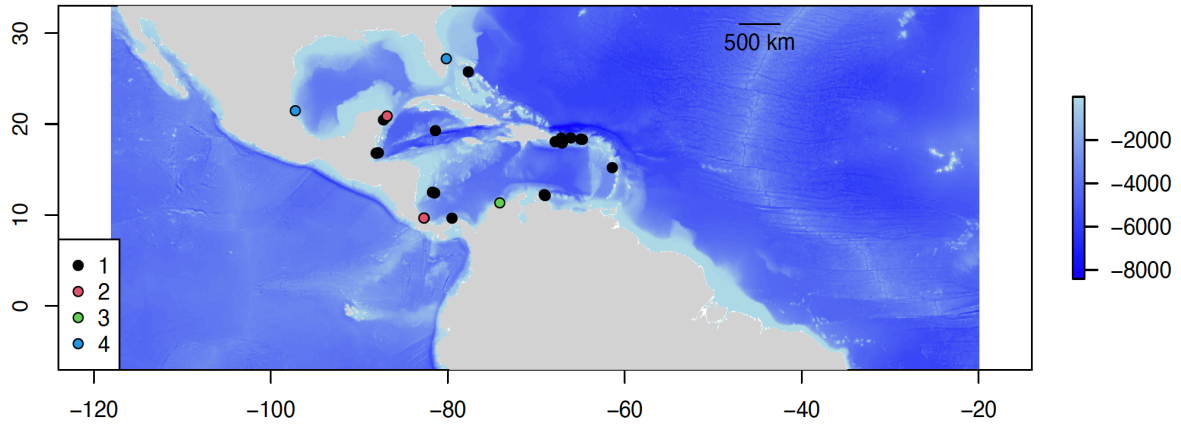


Figure S33. Environmental clusters of the sampling locations in the Tropical Atlantic.

MACROEVOLUTIONARY ANALYSES, PHYLOGENOMIC INFERENCE AND BIOGEOGRAPHIC HISTORY

Phylogenomic inference and mito-nuclear discordance. Phylogenomic trees estimated using matrices assembled from *de novo*, *quasi de novo*, and reference-based approaches displayed similar phylogenetic relationships, with the exception of *M. versicolor*. However, the difference in the phylogenetic placement of this species—either as the basal lineage of the genus (in *de novo*, and *quasi de novo*) or as the basal lineage of clade B (reference-based)—is weakly supported in both scenarios (Fig. S34). Due to the large amount of missing data and the computationally

intensive nature of *de novo* and *quasi de novo* methods, phylogenetic inferences were carried out using the reference-based assembled matrix at different levels of missing data (see Methods).

In these phylogenomic trees, *M. versicolor* is placed as the basal lineage within the *Malacoctenus* genus, except in the 28K SNPs tree (Fig. S35). Unlike the comparison of assembly approaches in the previous case, every phylogenetic tree strongly supports this placement. All trees identify two main clades. Clade A consists of the species *M. aurolineatus*, *M. gilli*, *M. erdmani*, and *M. macropus*. Clade B1 comprises *M. costaricanus*, *M. polyporosus*, *M. hubbsi*, *M. gigas*, *M. zacae*, *M. zonogaster*, *M. zonifer*, and *M. sudensis*, while clade B2 includes *M. tetranemus*, *M. delalandii*, *M. margaritae*, *M. mexicanus*, *M. triangulatus*, *M. brunoi*, and *M. lianae*, and *M. boehlkei* as the basal lineage of clade B, except in the 28K SNPs tree. In general, all trees showed consistent phylogenetic relationships for clades A and B2. However, within clade B1, a significant discrepancy in clade B1 was noted considering the positions of *M. gigas*, *M. zacae*, and the *M. hubbsi* species complex. In most instances, *M. zacae* represents the sister lineage to the *M. hubbsi* species complex. Yet, in the 28K SNPs tree, *M. gigas* appears as the sister lineage of the *M. hubbsi* species complex.

All species trees (Fig. S36), estimated using reference-based assembled matrices, identify *M. versicolor* as the basal lineage within the *Malacoctenus* genus and *M. boehlkei* as the oldest lineage in clade B. Contrary to the findings in phylogenetic trees, the geminate species pair (*M. tetranemus* and *M. delalandii*) is presented as the basal lineage in clade B1. Nevertheless, we consistently observed a pattern where the sister species to the *M. hubbsi* species complex is either *M. gigas* or *M. zacae*. Notably, the species tree with the fewest SNPs (1891) showed the most discrepancies compared to the other trees. One such discrepancy is where *M. aurolineatus*, initially the basal lineage of *M. gilli*, *M. erdmani*, and *M. macropus*, appears as the sister species to *M. gilli*.

The mitochondrial tree depicts *M. boehlkei* as the basal lineage of the *Malacoctenus* genus, mirroring the 28K SNPs phylogenomic tree (Fig. S37). Mitochondrial inferences suggest *M. versicolor* is the sister species of *M. aurolineatus*, though this relationship has low support. Minimal genetic variation was observed between *M. macropus* and *M. gilli*, which appear as sister species. Notably, *M. erdmani*, the sister species of *M. macropus* based on phylogenomic inferences, was not assessed using mitochondrial data. Despite the low support for relationships among the basal lineages, the mitochondrial tree still distinguishes clades B1 and B2. It's worth highlighting that *M. gigas* represents the sister species of *M. zacae*, a finding also supported by some phylogenomic trees (Fig. S35–S37).

Test for shared events of divergence. Results from the test of synchronous cladogenetic events triggered by the rise of the Isthmus of Panama, conducted using ecoevolity, supported a model with no shared divergent events across trans-isthmic clades (Fig. S38). This finding was consistent in analyses that included both *ca.* 200K and more than 1000K bp. These results align with the divergence times among trans-isthmic clades estimated in the time-calibrated species tree (Fig. 4).

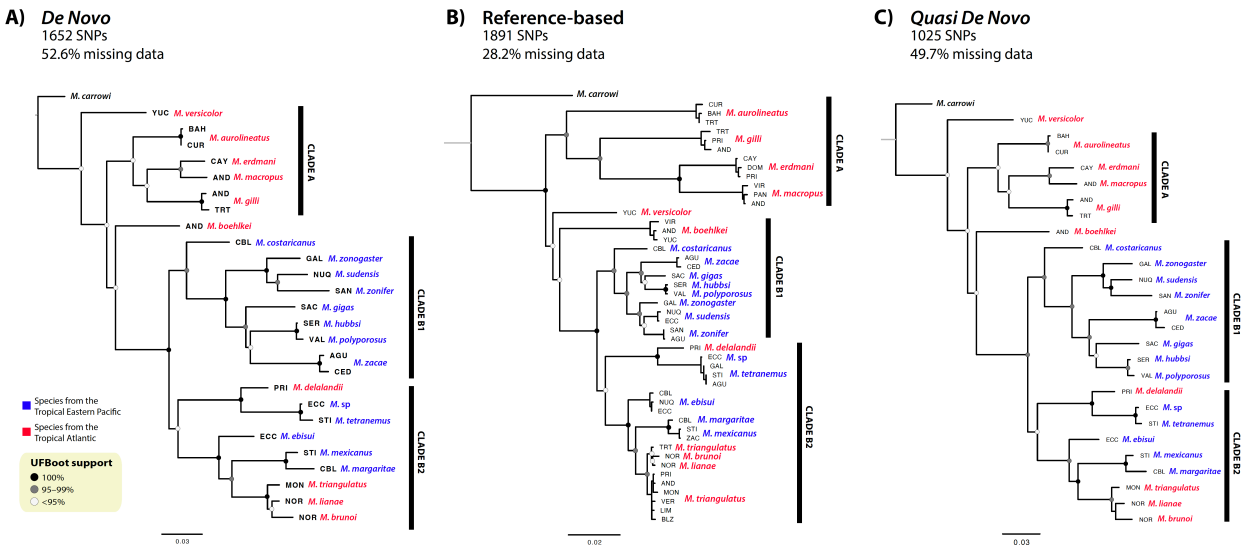


Figure S34. Maximum-likelihood trees estimated with matrices generated by different approaches. A) *De novo*, B) Reference-based, and C) *Quasi de novo* assemblies.

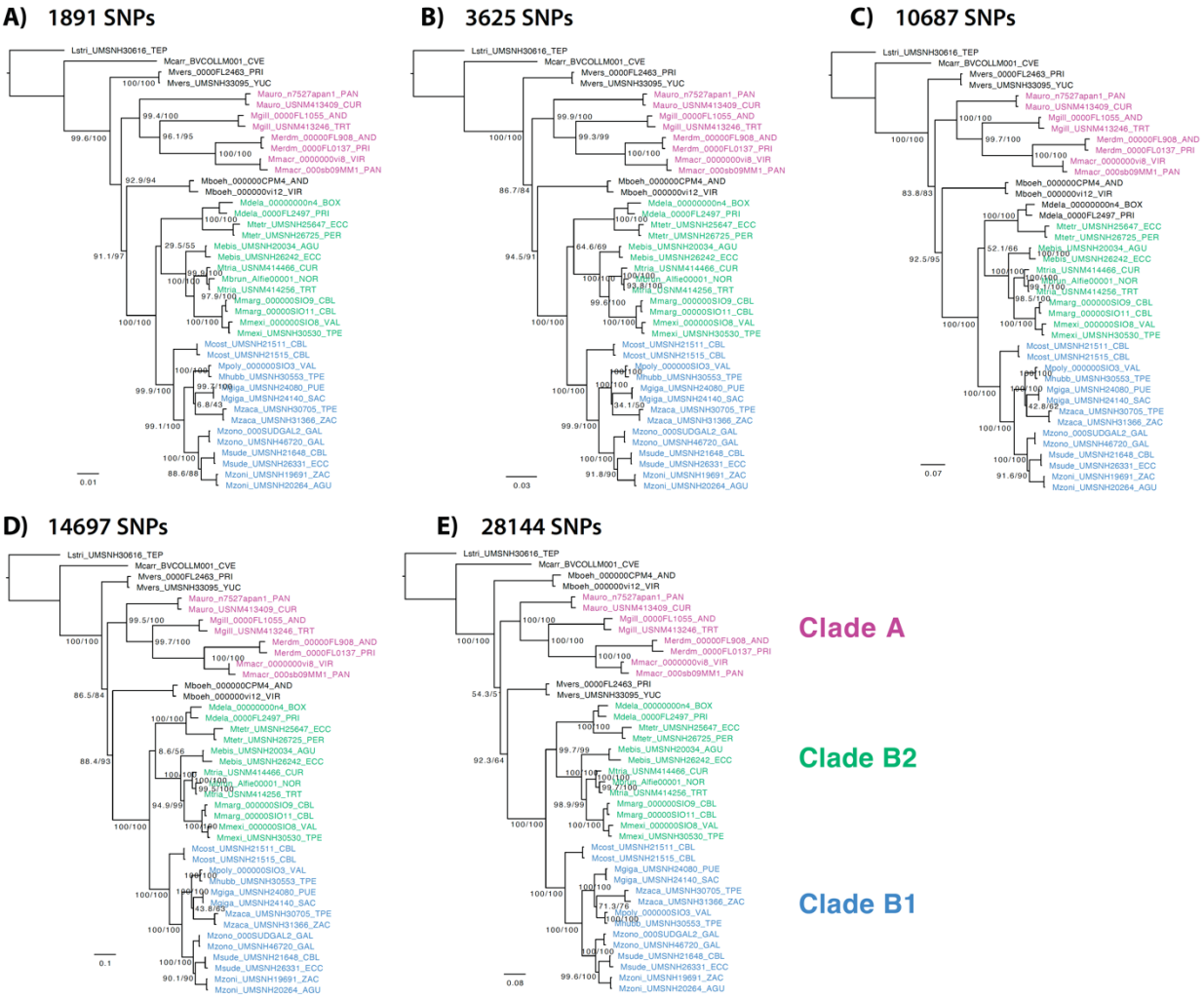


Figure S35. Maximum-likelihood trees estimated with matrices with varying levels of missing data.

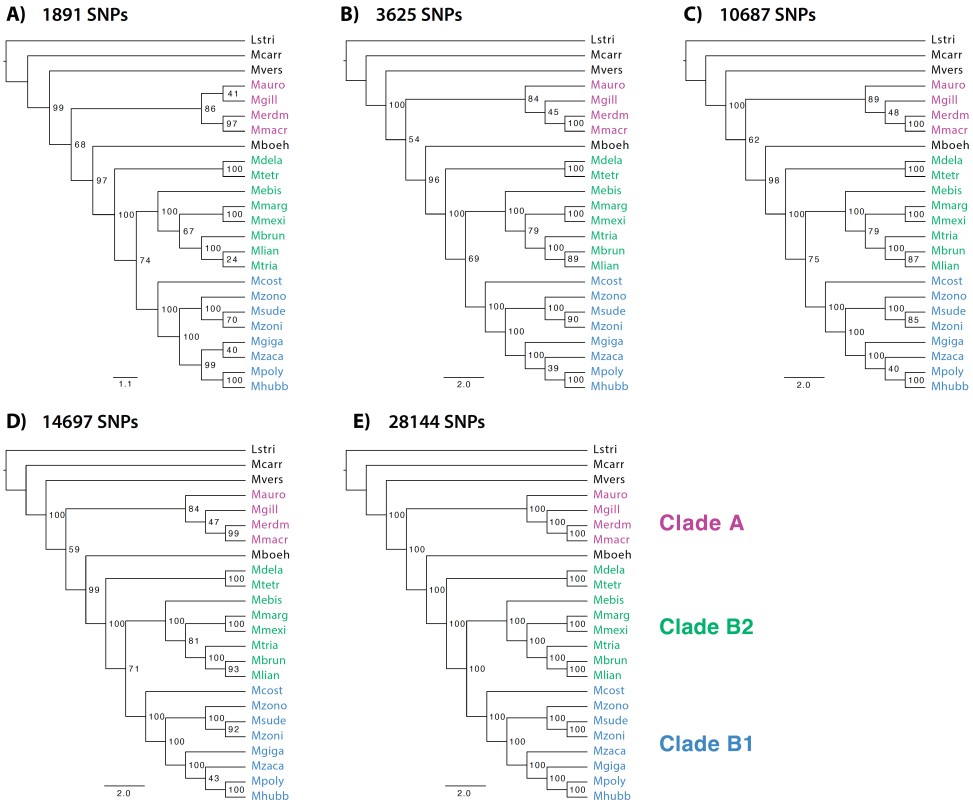


Figure S36. Species trees based on the multispecies coalescent model, estimated with matrices with varying levels of missing data.

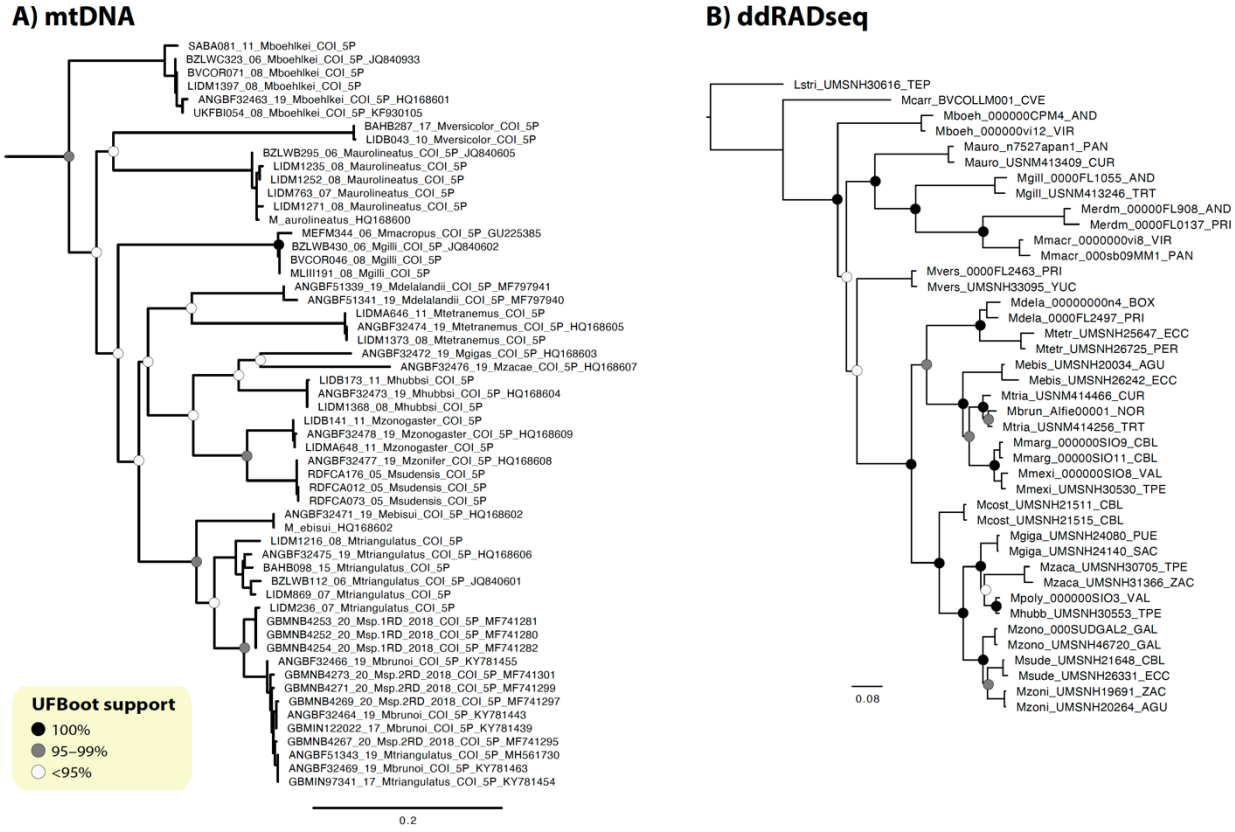


Figure S37. Maximum-likelihood A) mitochondrial tree based on COI sequences; and B) phylogenomic tree based on 28K SNPs estimated using IQtree software.

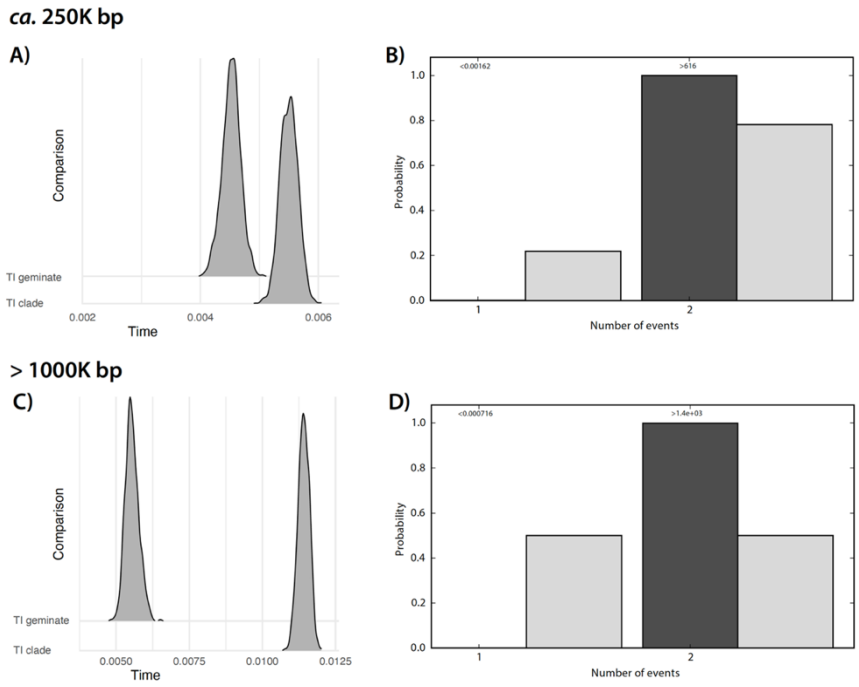


Figure S38. Results from the test of synchronous cladogenetic events across trans-isthmic pairs using ecoevolity. Approximate marginal posterior densities of divergence times for the trans-isthmic geminate pair (TI geminate) and the trans-isthmic clade (TI clade) using A) *ca.* 200K and C) over 1000K bp. Approximate prior (light grey bars) and posterior (dark grey bars) probabilities of the number of divergence events are shown in B) *ca.* 200K and D) more than 1000K bp.

Table 10. Weir and Cockerham's 1984 *Fst* estimates among highly independent evolutionary lines identified by marine barriers

Genetic groups	***Fst* value**
Species complexes	
Malacoctenus hubbsi vs *Malacoctenus polyporosus* (separated by SB; TEP)	0.11
Malacoctenus triangulatus vs Malacoctenus lianae (separated by PNB10; TA)	0.07
Independent evolutionary lineages	
Malacoctenus tetranemus vs M. aff. tetranemus (ECC populations)	0.06
Malacoctenus zacae northern populations *vs* southern populations (SB)	0.14
Malacoctenus zonifer northern populations *vs* southern populations (PNB2)	0.10
Malacoctenus aurolineatus vs M. aurolineatus from Curacao (PNB9)	0.09
Malacoctenus gilli west PNB11 *vs* east PNB11 (TRT-DOM populations)	0.19

References

1. H.-C. Lin, P. A. Hastings, Phylogeny and biogeography of a shallow water fish clade (Teleostei: Blenniiformes). *BMC Evol. Biol.* **13**, 210 (2013).

2. V. G. Springer, "Systematics and Zoogeography of the Clinid Fishes of the Subtribe Labrisomini Hubbs," University of Texas. (1959).

3. M. D. López, M. U. Alcocer, P. D. Jaimes, Phylogeography and historical demography of the Pacific Sierra mackerel (Scomberomorus sierra) in the Eastern Pacific. 1–12 (2010).

4. E. R. Sandoval-Huerta, *et al.*, The evolutionary history of the goby Elacatinus puncticulatus in the tropical eastern pacific: Effects of habitat discontinuities and local environmental variability. *Mol. Phylogenet. Evol.* **130**, 269–285 (2019).

5. M. Biol, L. A. H. M. Frey, P. G. E. Pfeiler, T. A. Markow, Geographical subdivision , demographic history and gene X ow in two sympatric species of intertidal snails , Nerita scabricosta and Nerita funiculata , from the tropical eastern Paci W c (2007) https:/doi.org/10.1007/s00227-007-0620-5.

6. M. S. Taylor, M. E. Hellberg, Genetic evidence for local retention of pelagic larvae in a Caribbean reef fish. *Science (80-.).* **299**, 107–109 (2003).

7. M. S. & H. M. E. Taylor, Comparative phylogeography in a genus of coral reef fishes : biogeographic and genetic concordance in the Caribbean. *Mol. Ecol.* **15**, 695–707 (2006).

8. R. Betancur-r, *et al.*, Reconstructing the lionfish invasion : insights into Greater Caribbean biogeography. *J. Biogeogr.* **38**, 1281–1293 (2011).

9. J. S. S. Butterfield, *et al.*, Wide-ranging phylogeographic structure of invasive red lionfish in the Western Atlantic and Greater Caribbean. *Mar. Biol.* **162**, 773–781 (2015).

10. J. P. Rippe, *et al.*, Population structure and connectivity of the mountainous star coral, Orbicella faveolata, throughout the wider Caribbean region. *Ecol. Evol.* **7**, 9234–9246 (2017).

11. D. C. Ballesteros-Contreras, L. M. Barrios, R. Preziosi, Population structure of the shallow coral Madracis auretenra in the Caribbean Sea. *Front. Mar. Sci.* **9**, 1–15 (2022).

12. R. Betancur-R, P. Arturo Acero, H. Duque-Caro, S. R. Santos, Phylogenetic and morphologic analyses of a coastal fish reveals a marine biogeographic break of terrestrial origin in the Southern Caribbean. *PLoS One* **5**, 1–10 (2010).

13. M. S. Taylor, M. E. Hellberg, Comparative phylogeography in a genus of coral reef fishes: Biogeographic and genetic concordance in the Caribbean. *Mol. Ecol.* **15**, 695–707 (2006).

14. I. B. Baums, M. W. Miller, M. E. Hellberg, Regionally isolated populations of an imperiled Caribbean coral, Acropora palmata. *Mol. Ecol.* **14**, 1377–1390 (2005).

15. R. J. Elshire, *et al.*, A robust, simple genotyping-by-sequencing (GBS) approach for high

diversity species. *PLoS One* **6**, 1–10 (2011).

16. K. R. Andrews, J. M. Good, M. R. Miller, G. Luikart, P. A. Hohenlohe, Harnessing the power of RADseq for ecological and evolutionary genomics. *Nat Rev Genet* **advance on**, 81–92 (2016).

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

18. H. Li, R. Durbin, Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).

19. H. Li, *et al.*, The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).

20. K. R. Andrews, J. M. Good, M. R. Miller, G. Luikart, P. A. Hohenlohe, Harnessing the power of RADseq for ecological and evolutionary genomics. *Nat. Publ. Gr.* (2016) https:/doi.org/10.1038/nrg.2015.28.

21. D. B. Lowry, *et al.*, Breaking RAD: an evaluation of the utility of restriction site-associated DNA sequencing for genome scans of adaptation. *Mol. Ecol. Resour.* **17**, 142–152 (2017).

22. N. C. Rochette, J. M. Catchen, Deriving genotypes from RAD-seq short-read data using Stacks. *Nat. Publ. Gr.* **12**, 2640–2659 (2017).

23. J. L. Davey, M. W. Blaxter, RADseq: Next-generation population genetics. *Brief. Funct. Genomics* **9**, 416–423 (2010).

24. P. Kirschner, *et al.*, Performance comparison of two reduced-representation based genome-wide marker-discovery strategies in a multi-taxon phylogeographic framework. *Sci. Rep.* **11**, 1–12 (2021).

25. A. Mastretta-Yanes, *et al.*, Restriction site-associated DNA sequencing, genotyping error estimation and de novo assembly optimization for population genetic inference. *Mol. Ecol.*, 28–41 (2015).

26. P. A. Hohenlohe, S. J. Amish, J. M. Catchen, F. W. Allendorf, G. Luikart, Next-generation RAD sequencing identifies thousands of SNPs for assessing hybridization between rainbow and westslope cutthroat trout. *Mol. Ecol. Resour.* **11**, 117–122 (2011).

27. S. J. O'Leary, J. B. Puritz, S. C. Willis, C. M. Hollenbeck, D. S. Portnoy, These aren't the loci you'e looking for: Principles of effective SNP filtering for molecular ecologists. *Mol. Ecol.* **27**, 3193–3206 (2018).

28. J. R. Paris, J. R. Stevens, J. M. Catchen, Lost in parameter space: a road map for stacks. *Methods Ecol. Evol.* **8**, 1360–1373 (2017).

29. J. Ferrer Obiol, *et al.*, Integrating Sequence Capture and Restriction Site-Associated DNA Sequencing to Resolve Recent Radiations of Pelagic Seabirds. *Syst. Biol.* **70**, 976–996

(2021).

30. J. R. Paris, J. R. Stevens, J. M. Catchen, Lost in parameter space: A road map for Stacks. *Methods Ecol. Evol.* **8**, 1360–1373 (2017).

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

32. A. Mastretta-Yanes, *et al.*, Restriction site-associated DNA sequencing, genotyping error estimation and de novo assembly optimization for population genetic inference. *Mol. Ecol. Resour.* **15**, 28–41 (2015).

33. J. M. Catchen, A. Amores, P. Hohenlohe, W. Cresko, J. H. Postlethwait, Stacks : Building and Genotyping Loci De Novo From Short-Read Sequences. *Genes, Genomes* **1**, 171–182 (2011).

34. C. del R. Pedraza-Marrón, *et al.*, Genomics overrules mitochondrial DNA, siding with morphology on a controversial case of species delimitation. *Proc. R. Soc. B Biol. Sci.* **286** (2019).

35. J. W. Davey, *et al.*, Special features of RAD Sequencing data: Implications for genotyping. *Mol. Ecol.* **22**, 3151–3164 (2013).

36. B. J. Knaus, N. J. Grünwald, vcfr: a package to manipulate and visualize variant call format data in R. *Mol. Ecol. Resour.* **17**, 44–53 (2017).

37. B. Ulaszewski, J. Meger, J. Burczyk, Comparative analysis of SNP discovery and genotyping in Fagus sylvatica L. and Quercus robur L. using RADseq, GBS, and ddRAD methods. *Forests* **12**, 1–17 (2021).

38. E. Linck, C. J. Battey, Minor allele frequency thresholds strongly affect population structure inference with genomic data sets. *Mol. Ecol. Resour.* **19**, 639–647 (2019).

39. D. A. DeRaad, snpfiltr: An R package for interactive and reproducible SNP filtering. *Mol. Ecol. Resour.* **00**, 1–11 (2022).

40. P. Danecek, *et al.*, The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158 (2011).

41. J. Cerca, *et al.*, Removing the bad apples: A simple bioinformatic method to improve loci-recovery in de novo RADseq data for non-model organisms. *Methods Ecol. Evol.* **12**, 805–817 (2021).

42. J. Goudet, HIERFSTAT , a package for R to compute and test hierarchical F -statistics. *Mol. Ecol. Notes* **2**, 184–186 (2005).

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

44. M. C. Whitlock, K. E. Lotterhos, Reliable detection of loci responsible for local adaptation:

Inference of a null model through trimming the distribution of FST. *Am. Nat.* **186**, S24–S36 (2015).

45. K. Luu, E. Bazin, M. G. B. Blum, pcadapt: an R package to perform genome scans for selection based on principal component analysis. *Mol. Ecol. Resour.* **17**, 67–77 (2017).

46. R. R. Betancur, *et al.*, Phylogenetic classification of bony fishes. *BMC Evol. Biol.* **17**, 1–40 (2017).

47. T. Jombart, Adegenet: A R package for the multivariate analysis of genetic markers. *Bioinformatics* **24**, 1403–1405 (2008).

48. J. M. Miller, C. I. Cullingham, R. M. Peery, The influence of a priori grouping on inference of genetic clusters: simulation study and literature review of the DAPC method. *Heredity (Edinb).* **125**, 269–280 (2020).

49. A. Raj, M. Stephens, J. K. Pritchard, FastSTRUCTURE: Variational inference of population structure in large SNP data sets. *Genetics* **197**, 573–589 (2014).

50. D. H. Alexander, J. Novembre, K. Lange, Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* **19**, 1655–1664 (2009).

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

52. L. T. Nguyen, H. A. Schmidt, A. Von Haeseler, B. Q. Minh, IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* **32**, 268–274 (2015).

53. D. Petkova, J. Novembre, M. Stephens, Visualizing spatial population structure with estimated effective migration surfaces. *Nat. Genet.* **48**, 94–100 (2015).

54. S. J. Oyler-McCance, *et al.*, New strategies for characterizing genetic structure in wide-ranging, continuously distributed species: A Greater Sage-grouse case study. *PLoS One* **17**, 1–22 (2022).

55. QGIS Development Team, QGIS Geographic Information System (2009).

56. T. Jombart, S. Devillard, A. B. Dufour, D. Pontier, Revealing cryptic spatial patterns in genetic variability by a new multivariate method. *Heredity (Edinb).* **101**, 92–103 (2008).

57. Z. N. Kamvar, J. F. Tabima, N. J. Grünwald, Poppr: an R package for genetic analysis of populations with clonal, partially clonal, and/or sexual reproduction. *PeerJ* **2**, e281 (2014).

58. E. Paradis, pegas: an R package for population genetics with an integrated--modular approach. *Bioinformatics* **26**, 419–420 (2010).

59. N. V. Gordeeva, O. G. Nanova, Application of geometric morphometrics for intraspecific variability analysis in mesopelagic fishes of Sternoptychidae and Myctophidae families. *J. Ichthyol.* **57**, 29–36 (2017).

60. P. J. Park, W. E. Aguirre, D. A. Spikes, J. M. Miyazaki, Landmark-Based Geometric Morphometrics: What Fish Shapes Can Tell Us about Fish Evolution. *Test. Stud. Lab. Teach. Proc. Assoc. Biol. Lab. Educ.* **34**, 361–371 (2013).

61. W. M. Berbel-Filho, U. P. Jacobina, P. A. Martinez, Preservation effects in geometric morphometric approaches: Freezing and alcohol in a freshwater fish. *Ichthyol. Res.* **60**, 268–271 (2013).

62. C. D. McMahan, *et al.*, Objectively measuring subjectively described traits: geographic variation in body shape and caudal coloration pattern within Vieja melanura (Teleostei: Cichlidae). *Rev. Biol. Trop.* **65**, 623–631 (2017).

63. A. M. Olsen, M. W. Westneat, StereoMorph: An R package for the collection of 3D landmarks and curves using a stereo camera set-up. *Methods Ecol. Evol.* **6**, 351–356 (2015).

64. D. C. Adams, E. Otárola-Castillo, Geomorph: An r package for the collection and analysis of geometric morphometric shape data. *Methods Ecol. Evol.* **4**, 393–399 (2013).

65. C. Fruciano, Measurement error in geometric morphometrics. *Dev. Genes Evol.* **226**, 139–158 (2016).

66. Q. Rougemont, *et al.*, Long-distance migration is a major factor driving local adaptation at continental scale in Coho salmon. *Mol. Ecol.* **32**, 542–559 (2023).

67. L. Benestan, *et al.*, Seascape genomics provides evidence for thermal adaptation and current-mediated population structure in American lobster (Homarus americanus). *Mol. Ecol.* **25**, 5073–5092 (2016).

68. L. Benestan, *et al.*, Restricted dispersal in a sea of gene flow. *Proc. R. Soc. B Biol. Sci.* **288** (2021).

69. R. J. Hijmans, raster: Geographic Data Analysis and Modeling (2023).

70. S. Dray, A. B. Dufour, The ade4 package: Implementing the duality diagram for ecologists. *J. Stat. Softw.* **22**, 1–20 (2007).

71. A. Xuereb, *et al.*, Asymmetric oceanographic processes mediate connectivity and population genetic structure, as revealed by RADseq, in a highly dispersive marine invertebrate (Parastichopus californicus). *Mol. Ecol.* **27**, 2347–2364 (2018).

72. P. Legendre, D. Borcard, P. R. Peres-Neto, Analyzing Beta Diversity: Partitioning the Spatial Variation of Community Composition Data. *Ecol. Monogr.* **75**, 435–450 (2005).

73. S. Dray, P. Legendre, P. R. Peres-Neto, Spatial modelling: a comprehensive framework for principal coordinate analysis of neighbour matrices (PCNM). *Ecol. Modell.* **196**, 483–493 (2006).

74. P. R. Peres-Neto, P. Legendre, Estimating and controlling for spatial structure in the study of ecological communities. *Glob. Ecol. Biogeogr.* **19**, 174–184 (2010).

75. E. Pante, B. Simon-Bouhet, marmap: A Package for Importing, Plotting and Analyzing Bathymetric and Topographic Data in R. *PLoS One* **8**, 6–9 (2013).

76. S. Dray, *et al.*, adespatial: Multivariate Multiscale Spatial Analysis (2023).

77. P. Legendre, J. Oksanen, C. J. F. ter Braak, Testing the significance of canonical axes in redundancy analysis. *Methods Ecol. Evol.* **2**, 269–277 (2011).

78. P. R. Peres-Neto, P. Legendre, S. Dray, D. Borcard, Variation partitioning of species data matrices: Estimation and comparison of fractions. *Ecology* **87**, 2614–2625 (2006).

79. D. Kahle, H. Wickham, ggmap: Spatial Visualization with ggplot2. *R J.* **5**, 144–161 (2013).

80. P. Legendre, M. J. Anderson, Distance-based redundancy analysis: Testing multispecies responses in multifactorial ecological experiments. *Ecol. Monogr.* **69**, 1–24 (1999).

81. L. M. Benestan, *et al.*, Population genomics and history of speciation reveal fishery management gaps in two related redfish species (Sebastes mentella and Sebastes fasciatus). *Evol. Appl.* **14**, 588–606 (2021).

82. D. Borcard, P. Legendre, All-scale spatial analysis of ecological data by means of principal coordinates of neighbour matrices. *Ecol. Modell.* **153**, 51–68 (2002).

83. G. Guénard, P. Legendre, B. Pages, codep: Multiscale Codependence Analysis. R package version 0.5-1 (2015).

84. M. Kearse, *et al.*, Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* **28**, 1647–1649 (2012).

85. S. V. Edwards, *et al.*, Implementing and testing the multispecies coalescent model: A valuable paradigm for phylogenomics. *Mol. Phylogenet. Evol.* **94**, 447–462 (2016).

86. L. Liu, L. Yu, S. V. Edwards, A maximum pseudo-likelihood approach for estimating species trees under the coalescent model. *BMC Evol. Biol.* **10**, 25–27 (2010).

87. X. Jiang, S. V. Edwards, L. Liu, B. Faircloth, The Multispecies Coalescent Model Outperforms Concatenation across Diverse Phylogenomic Data Sets. *Syst. Biol.* **69**, 795–812 (2020).

88. J. D. M. Bryan C. Carstens, Phylogenetic Model Choice: Justifying a Species Tree or Concatenation Analysis. *J. Phylogenetics Evol. Biol.* **01**, 1–8 (2013).

89. J. Chifman, L. Kubatko, Quartet inference from SNP data under the coalescent model. *Bioinformatics* **30**, 3317–3324 (2014).

90. D. L. Swofford, D. L. Swofford, D. L. Swofford, PAUP*: Phylogenetic analysis using parsimony (*and other methods), Version 4.0b10 in (2002).

91. R. Bouckaert, *et al.*, BEAST 2: A Software Platform for Bayesian Evolutionary Analysis. *PLoS Comput. Biol.* **10**, 1–6 (2014).

92. M. Stange, M. R. Sánchez-Villagra, W. Salzburger, M. Matschiner, Bayesian divergence-time estimation with genome-wide single-nucleotide polymorphism data of sea catfishes (Ariidae) supports miocene closure of the Panamanian Isthmus. *Syst. Biol.* **67**, 681–699 (2018).

93. R. McGirr, M. Seton, S. Williams, Kinematic and geodynamic evolution of the Isthmus of Panamaregion: Implications for Central American Seaway closure. *GSA Bull.*, 1–18 (2020).

94. G. Carnevale, The first fossil ribbonfish (Teleostei, Lampridiformes, Trachipteridae). *Geol. Mag.* **141**, 573–582 (2004).

95. A. Rambaut, M. A. Suchard, D. Xie, A. J. Drummond, Tracer v1.6. (2014).

96. D. L. Rabosky, *et al.*, An inverse latitudinal gradient in speciation rate for marine fishes. *Nature* **559**, 392–395 (2018).

97. K. Tamura, *et al.*, Estimating divergence times in large molecular phylogenies. *Proc. Natl. Acad. Sci. U. S. A.* **109**, 19333–19338 (2012).

98. S. Kumar, G. Stecher, M. Li, C. Knyaz, K. Tamura, MEGA X: Molecular evolutionary genetics analysis across computing platforms. *Mol. Biol. Evol.* **35**, 1547–1549 (2018).

99. E. Paradis, K. Schliep, Ape 5.0: An environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics* **35**, 526–528 (2019).

100. J. R. Oaks, Full Bayesian comparative phylogeography from genomic data. *Syst. Biol.* **68**, 371–395 (2019).

101. J. R. Oaks, C. D. Siler, R. M. Brown, The comparative biogeography of Philippine geckos challenges predictions from a paradigm of climate-driven vicariant diversification across an island archipelago. *Evolution (N. Y).* **73**, 1151–1167 (2019).

102. A. Papadopoulou, L. L. Knowles, Toward a paradigm shift in comparative phylogeography driven by trait-based hypotheses. *Proc. Natl. Acad. Sci. U. S. A.* **113**, 8018–24 (2016).

103. N. J. Matzke, nmatzke/BioGeoBEARS: BioGeoBEARS: BioGeography with Bayesian (and likelihood) Evolutionary Analysis with R Scripts (2018) https:/doi.org/10.5281/zenodo.1463216.

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

105. M. D. Spalding, *et al.*, Marine Ecoregions of the World : A Bioregionalization of Coastal and Shelf Areas. **57**, 573–583 (2007).

106. D. R. Robertson, K. L. Cramer, Defining and dividing the Greater Caribbean: Insights from the biogeography of shorefishes. *PLoS One* **9** (2014).

107. P. April, O. Pen, D. R. Robertson, K. L. Cramer, Shore fishes and biogeographic subdivisions of the Tropical Eastern Pacific. **380**, 1–17 (2009).

108. A. C. Siqueira, D. R. Bellwood, P. F. Cowman, Historical biogeography of herbivorous coral reef fishes: The formation of an Atlantic fauna. *J. Biogeogr.* **46**, 1611–1624 (2019).

109. A. Santaquiteria, *et al.*, Phylogenomics and Historical Biogeography of Seahorses, Dragonets, Goatfishes, and Allies (Teleostei: Syngnatharia): Assessing Factors Driving Uncertainty in Biogeographic Inferences. *Syst. Biol.* **70**, 1145–1162 (2021).

110. M. Rincon-Sandoval, *et al.*, Evolutionary determinism and convergence associated with water-column transitions in marine fishes. *Proc. Natl. Acad. Sci. U. S. A.* **117**, 33396–33403 (2021).

111. N. J. Matzke, Model selection in historical biogeography reveals that founder-event speciation is a crucial process in island clades. *Syst. Biol.* **63**, 951–970 (2014).

112. J. Dupin, *et al.*, Bayesian estimation of the global biogeographical history of the Solanaceae. *J. Biogeogr.* **44**, 887–899 (2017).

113. W. M. White, R. a Duncan, Petrology and Geochemistry of the Galapagos Islands: Portrait of a Pathological Mantle Plume. *J. Geophys. Res.* **98**, 533–563 (1993).