Rapid radiation in a highly diverse marine environment

Kosmas Hench\textsuperscript{a,1}, Martin Helmkampf\textsuperscript{b}, W. Owen McMillan\textsuperscript{b}, and Oscar Puebla\textsuperscript{a,b,c,d,1}

\textsuperscript{a}Ecology Department, Leibniz Centre for Tropical Marine Research, 28359 Bremen, Germany; \textsuperscript{b}Smithsonian Tropical Research Institute, Apartado Postal 0843-03092, Republic of Panama; \textsuperscript{c}Institute for Chemistry and Biology of the Marine Environment, 26111 Oldenburg, Germany; and \textsuperscript{d}Marine Evolutionary Ecology, GEOMAR Helmholtz Centre for Ocean Research Kiel, 24105 Kiel, Germany

Edited by Nancy Knowlton, National Museum of Natural History, Smithsonian Institution, Washington, DC; received September 30, 2020; accepted November 18, 2021

Rapid diversification is often observed when founding species invade isolated or newly formed habitats that provide ecological opportunity for adaptive radiation. However, most of the Earth’s diversity arose in diverse environments where ecological opportunities appear to be more constrained. Here, we present a striking example of a rapid radiation in a highly diverse marine habitat. The hamlets, a group of reef fishes from the wider Caribbean, have radiated into a stunning diversity of color patterns but show low divergence across other ecological axes. Although the hamlet lineage is \textasciitilde 26 My old, the radiation appears to have occurred within the last 10,000 generations in a burst of diversification that ranks among the fastest in fishes. As such, the hamlets provide a compelling backdrop to uncover the genomic elements associated with phenotypic diversification and an excellent opportunity to build a broader comparative framework for understanding the drivers of adaptive radiation. The analysis of 170 genomes suggests that color pattern diversity is generated by different combinations of alleles at a few large-effect loci. Such a modular genomic architecture of diversification has been documented before in Heliconius butterflies, capucino finches, and munia finches, three other tropical radiations that took place in highly diverse and complex environments. The hamlet radiation also occurred in a context of high effective population size, which is typical of marine populations. This allows for the accumulation of new variants through mutation and the retention of ancestral genetic variation, both of which appear to be important in this radiation.

Adaptive radiation is driven by ecological opportunity, whereby newly accessible niches provide potential for diversification (1). This process often takes place in geographically isolated and/or newly formed habitats such as lakes or oceanic islands, where founding species are exposed to depauperate environments in which competition is relaxed. This is, for example, the case in exemplary adaptive radiations, like Darwin’s finches (2), East African cichlids (3, 4), and Caribbean Anolis lizards (5). Nevertheless, these radiations represent extremes of an evolutionary process that may also operate in diverse and complex habitats, where the relationship between biotic and abiotic drivers and rapid speciation is often more subtle. Unfortunately, with the notable exception of Heliconius butterflies (6), these radiations remain poorly explored despite the fact that they occur in environments that contain most of the diversity on Earth. As a result, we lack a broader comparative framework for identifying and generally understanding the main drivers of adaptive radiation.

The hamlets (Hypoplectrus spp., Serranidae), a group of reef fishes from the wider Caribbean, present an excellent opportunity to investigate the genomic basis of radiation in diverse and complex environments where most niches are already occupied. Typical of marine species (7), most hamlets are characterized by extensive geographic ranges, large population sizes, high fecundity, and high potential for dispersal through a 3-wk pelagic larval stage (8–10). These characteristics are expected to shape their evolutionary potential in complex ways, possibly reducing the opportunity for speciation (11, 12). Against this backdrop, the genus diversified into at least 18 species that are highly sympatric (9, 13) and very similar in terms of habitat and diet (14, 15). They differ essentially in color pattern, a trait that is thought to be ecologically relevant through crypsis and mimicry (16–20). According to the aggressive mimicry hypothesis, some hamlet species (the mimics) achieve higher preying success by resembling other fishes from different families (the models) that are harmless to the hamlets’ prey. In agreement with this hypothesis, behavioral differences in resource acquisition have been documented between hamlet species (19, 20). Yet, resemblance between models and mimics is imperfect, and some species appear to be neither cryptic nor mimetic. Color pattern is an important cue for mate choice, and hamlets show strong assortative mating both in the field and in the laboratory (10, 13, 18, 19, 21). In addition, hamlets are simultaneously hermaphroditic, and mate choice is mutual (8, 22). This particular mating system results in complex pairing dynamics among individuals that can contribute to diversification by generating strong sexual selection (13, 23). Despite strong assortative mating, interspecific spawnings are occasionally observed in natural populations (10, 13, 18, 19, 21). Fertilization is external, and eggs are planktonic. The available evidence suggests that there are no barriers to fertilization among species (24) and that gene flow is ongoing (25). Genetically, the Caribbean hamlets are very closely related. Similar to East

Significance

Adaptive radiation, the evolutionary process whereby a lineage diversifies over a short period of time, often occurs in geographically isolated or newly formed habitats where colonizing species encounter unoccupied niches and reduced selective pressures. Rapid radiations may also occur in diverse and complex environments, but these cases are less well documented. Here, we show that the hamlets, a group of Caribbean reef fishes, radiated within the last 10,000 generations in a burst of diversification that ranks among the fastest in fishes. Genomic analysis suggests that color pattern diversity is generated by different combinations of alleles at a few genes with large effect. Such a modular genomic architecture of diversification is emerging as a common denominator to a variety of radiations.

Author contributions: K.H., M.H., W.O.M., and O.P. designed research; K.H., M.H., and O.P. performed research; K.H., M.H., and O.P. analyzed data; and K.H., M.H., W.O.M., and O.P. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

This article is distributed under Creative Commons Attribution-NonCommercial-NoDerivatives License 4.0 (CC BY-NC-ND).

1To whom correspondence may be addressed. Email: k.hench@posteo.de or oscar.puebla@leibniz-zmt.de.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2020457119/DCSupplemental.

Published January 18, 2022.

PNAS 2022 Vol. 119 No. 4 e2020457119

https://doi.org/10.1073/pnas.2020457119 | 1 of 11
African cichlids from Lake Victoria (4), they show generally low levels of genetic divergence (13, 21, 25–27) and do not sort into distinct mitochondrial haplogroups (28–30). A chromosome-resolution reference genome is now available for the group (25), which provides the foundation to more fully explore the genomic basis of this radiation. Here, we examine the genomes of 170 individuals from 28 pairs of sympatric species to 1) uncover the genomic architecture of the radiation, 2) identify genomic regions associated with major phenotypic differences, and 3) ask how the hamlet radiation compares with other rapid radiations.

**Results**

**Rapid Radiation, Old Lineage.** We started by considering the hamlet radiation in the context of the phylogeny of the subfamily Serraninae and the Fish Tree of Life (FToL) estimates of evolutionary rates (31). This broad phylogenetic perspective indicates that the hamlets exhibit speciation rates that are among the highest in fishes (mean clade-specific speciation rate: 2.44) (SI Appendix, Fig. S1). Within the Serraninae, a single burst of diversification was identified at the base of the hamlet radiation (Fig. L4 and SI Appendix, Fig. S2). No other serranine lineage radiated rapidly (subfamily background speciation rate excluding hamlets: 0.11). Notably, this includes the chalk bass (*Serranus tortugarum*) and the tobaccofish (*Serran tabacarius*), two species that are closely related to the hamlets and have a similar mating system (33). The two clades that are simultaneously hermaphroditic (marked by a star in Fig. L4 and SI Appendix, Fig. S24) do not present an increase in diversification rate either. Finally, the time-calibrated FToL phylogeny illustrates that although the extant hamlet species appear to be very young, the *Hypoplectrus* lineage last shared a common ancestor with other species in the subfamily ~26 Mya (SI Appendix, Fig. S24).

**Population Genetic Patterns.** Given the low levels of genetic divergence among hamlets, we first used our genomic data from Belize, Honduras, and Panama to examine to what extent sympatric hamlets represent genetic clusters. Principal component analysis indicates that this is generally the case, albeit with varying degrees of overlap between species (Fig. 1 C–E). We next grouped samples from each species and location together to examine patterns of genome-wide differentiation (\(F_{ST}\)) between pairs of sympatric species. This analysis reveals a continuous range of differentiation, from <0.003 to 0.1 (Fig. 1B and SI Appendix, Table S1). It is important to stress that this differentiation continuum does not constitute a temporal sequence of speciation, as it includes different species from different locations. Moreover, the barred, black, and butter hamlets are over-represented in this dataset since they were sampled at the three

---

**Fig. 1.** Phylogenetic context and population genetic patterns. (A) Maximum likelihood phylogeny of the Serraninae subfamily based on 23 nuclear and mitochondrial genes. The hamlet radiation is highlighted in gray, and species considered in this study are marked with an asterisk. Gene sequences for the other species were obtained from the FToL project (31). Stars denote clades that are simultaneously hermaphroditic following ref. 32 (note that the mating system of *Serranus accraenis* is not known). Ch., Chelidoperca; Cp., Centropristis; D., Diplectrum; E., Epinephelus; H., Hypoplectrus; Pa., Paralabrax; Pl., Plectranthias; Sc., Schultzzea; Se., Serranus. (B) Genetic differentiation (\(F_{ST}\)) between pairs of sympatric hamlets. The 28 pairs are numbered and presented in order of increasing genome-wide differentiation, and they are color coded with respect to location (red, Belize; blue, Honduras; green, Panama). The pairs that are significantly differentiated at the \(\alpha = 0.05\) level are highlighted with filled bars. (C–E) Principal component analysis (PCA) for each location. The percentage of variation explained by each principal component (PC) is shown. The PCAs were repeated with the highly diverged genomic regions excluded and produced very similar results (SI Appendix, Fig. S3), indicating that the clustering patterns are not driven by these regions.

---

Hench et al.  
Rapid radiation in a highly diverse marine environment  
https://doi.org/10.1073/pnas.2020457119
locations. Nevertheless, levels of genetic differentiation among populations within species are within the range of differentiation between sympatric species (SI Appendix, Table S1). From a population genetic perspective, populations of the same species can, therefore, be as distinct as different species (34, 35).

Genomic Architecture of Radiation. We then divided the genome into 50-kb windows to reveal how the genomic architecture of the hamlet radiation unfolded across the differentiation continuum. The radiation is characterized by a small number of sharp peaks (so-called “islands” (36)) of differentiation that do not expand with increasing genome-wide differentiation (Fig. 2A and SI Appendix, Figs. S4 and S5). Their number does not increase substantially either, with ≤6 peaks with $F_{ST} > 0.7$ per species pair throughout the entire continuum. It is also noteworthy that at the lower end of the differentiation continuum, differentiation is largely independent from recombination rate (SI Appendix, Fig. S6A, row 1). In this respect, our results contrast with previous studies conducted across higher levels of differentiation that report a marked effect of recombination on differentiation (37–41). We nonetheless capture the onset of the effect of recombination, with differentiation accumulating disproportionately in regions of low recombination as genome-wide differentiation increases (SI Appendix, Fig. S6). This effect is particularly strong in a large section of linkage group (LG) (chromosome) 8, which we had previously identified as a low-recombining region (25, 42) (Fig. 2A). In contrast to genetic differentiation ($F_{ST}$, which captures allele frequency differences), the genomic architecture of divergence ($d_{XY}$, which captures sequence divergence) is similar across all species pairs (SI Appendix, Fig. S7). Divergence is generally elevated in the chromosome peripheries where recombination rate tends to be higher (SI Appendix, Fig. S8 B and D). This is likely an effect of the recombination landscape that shaped ancestral variation, resulting in higher diversity in regions of high recombination.

Recent Radiation. In order to estimate the age of the hamlet radiation, we inferred the demographic histories of all species in all locations using the multiple sequentially Markovian coalescent (MSMC2) (44). The cross-coalescence patterns suggest that hamlets diverged within the last 10,000 generations (Fig. 3B). These results are consistent with previous analyses using a different approach that does not rely on phasing (45). Alternatively, gene flow among sympatric species may have obliterated the signature of older demographic events, which could be complex and involve several cycles of “fission–fusion–fission” (46). Nevertheless, reconstructing such events is challenging at the low levels of divergence that characterize the Caribbean hamlets. The MSMC2 analyses also indicate that ancestral effective population sizes remained high, in the order of $10^5 – 10^6$ (Fig. 3A).

Ongoing Gene Flow and Introgression. The overall landscape of genomic differentiation suggests that it is shaped in part by gene flow and introgression. In order to test for ongoing gene flow, we searched for genetic hybrids and backcrosses in our dataset. A total of 12 high-probability hybrids and backcrosses were identified (SI Appendix, Fig. S11), representing 7% of the individuals analyzed. This proportion is likely an underestimate since our sampling design explicitly excluded individuals with intermediate color patterns. It is also substantially higher than

![Fig. 2.](https://doi.org/10.1073/pnas.2020457119)

Hench et al.
Rapid radiation in a highly diverse marine environment

Fig. 2. Whole-genome patterns. The alternating white and gray blocks represent the 24 LGs (chromosomes). All statistics are calculated over 50-kb sliding windows with 5-kb increments unless stated otherwise. (A) Joint differentiation ($F_{ST}$) among the 14 groups of samples (species within locations). The red vertical lines highlight regions above the 99.8th $F_{ST}$ percentile (SI Appendix, Table S2). Letters A, B, and C correspond to the three genomic regions that are highlighted in Fig. 5. (B and C) Topology weighting for Belize and Honduras, respectively, along nonoverlapping 200 SNP windows. The different colors correspond to different topologies, and the white horizontal lines indicate the null weighting (i.e., all topologies equally likely). (D–F) G × P association for bars, dark saddle on the caudal peduncle, and spot on the snout, respectively. An expanded version of this figure is presented in SI Appendix, Fig. S8.
the 2% interspecific spawnings that were observed at the three locations (13, 19). However, behavioral observations indicate that hybridization rate can vary dramatically in the hamlets depending on the social context of individuals at the time of spawning (13).

In order to test for past introgression events, we calculated D statistics for all possible trios of species/populations using Hypoplectrus floridae—which belongs to the Gulf of Mexico hamlet clade (47)—as the out-group. In the most treelike configuration (BBAA), a total of 69 trios show evidence of introgression, representing 19% of the 364 trios that we tested. This proportion is slightly higher than the 14% reported in East African cichlids from Lake Victoria (4). Importantly, all species/populations and 44% of all pairs show evidence of introgression (SI Appendix, Fig. S12), indicating that it is pervasive across the radiation.

**Weak Background Phylogenetic Signal.** The low levels of divergence, ongoing gene flow, and history of extensive introgression observed in the hamlets are expected to obscure—possibly even rewrite—the phylogenetic relationships among species and populations. As a result, the genome-wide phylogeny of the radiation is not well resolved, particularly with respect to the deeper nodes (Fig. 4A). The reconstruction only recovers the most phylogenetically well-resolved clades, whereas unsupported groups appear diffuse and weakly connected. Several subnetworks across species boundaries also emerge from the network, consistent with a history of rapid divergence and introgression.

**Genetic Basis of Phenotypic Diversification.** The weak phylogenetic signal in the hamlets provides the context to localize regions of the genome that are likely responsible for phenotypic differences among species. The underlying logic is that functionally important regions should group individuals by phenotype. We used topology weighting by iterative sampling of sub-trees [Twiss (48)] to dissect the phylogenetic signal along the genome. Briefly, this method considers sliding-windows phylogenies along the genome and weights the contribution of each possible taxon topology to the full tree. As expected given the weak phylogenetic signal in the hamlets and the small size of the windows considered (200 single-nucleotide polymorphism [SNPs]), this approach failed to identify a leading taxon topology at most windows throughout the genome. Nevertheless, it revealed a number of sharp topology weighting peaks (Fig. 2 B and C). For example, in Belize, the topologies in which H. indigo and H. maya—the two blue hamlet species—are sister species dominate a narrow region on LG 4 (Fig. 5A), while two other regions on LG 12 are dominated by topologies grouping the two hamlet species that display vertical bars (H. indigo and H. puella) (Fig. 5 B and C). These patterns are also visible in phylogenies based on the entire contiguous sequence of these regions (SI Appendix, Figs. S14 and S15).

In order to further explore the association between genetic variation and specific components of color pattern, we scored all fishes for the presence or absence of vertical bars, saddle mark on the caudal peduncle, and spot on the snout (SI Appendix, Fig. S16). These traits were chosen because they are polymorphic and can be scored unambiguously. Genotype \times phenotype (G \times P) association analysis revealed a strong association between the presence or absence of vertical bars and genetic variation in a narrow genomic interval on LG 12 (Fig. 2D). Associations with the other two traits were more diffuse, but here again, association peaks emerged, notably on LG 12 for the saddle mark and on LG 4 for the snout spot (Fig. 2 E and F). Clustering patterns at these genomic regions (SI Appendix, Fig. S16) are consistent with the topology weighting, phylogenetic, and G \times P analyses.

The genomic regions identified by topology weighting and G \times P analyses also match regions of high differentiation (FST) among species (Figs. 24 and 5). Altogether, these analyses allow us to start dissecting the genetic variation associated with phenotypic diversity in the hamlets. The strongest signal was observed in a narrow region of LG 12 that shows a strong association with the presence or absence of vertical bars (letter B in Fig. 2). The 13 pairs of sympatric species that include one species with vertical bars and one without present high differentiation at this locus, while the 15 species pairs that include two species with bars or two species without bars do not (Fig. 5B). In line with this pattern, the region is dominated by topologies in which the two hamlets with vertical bars are sister species. This locus, which we had previously identified (25), is centered on cas1 (Fig. 5B). This gene encodes a castor zinc finger transcription factor that is involved in a number of processes through development.

Our data also provide insights into the hoxca gene cluster on LG 12. In line with our preliminary analyses (25), we observe a strong association between variation at the hoxca3a locus and the presence or absence of a saddle mark on the caudal peduncle (Fig. 5C), which is characteristic of the butter hamlet (H. unicolor). The pairs of sympatric species that are most differentiated at this locus include H. unicolor, and the region is dominated by topologies that single out the butter hamlet. In addition, we observe a secondary association with vertical bars between hoxca8a

![Figure 3](image-url). Demographic inference. (A) Inferred history of effective population size. Each line is based on three to four genomes per group (species within locations). (B) Cross-coalescence rates for the 28 pairs of sympatric species color coded by genome-wide locations. Each line represents an independent run based on two genomes from two sympatric species (total of four). A cross-coalescence rate of one indicates completely shared ancestry, and a rate of zero indicates no shared ancestry. All estimates are scaled with a per site mutation rate $\mu = 3.7 \times 10^{-3}$. The most ancient and the two most recent time segments are omitted due to unreliable inference at these extremes.
and hoxc11a. Hox genes play an important role in patterning tissues along the body axis and have been shown to be involved in color pattern development in insects (49, 50) and vertebrates (51). They are arranged and expressed in a sequence that follows the body axis, with 3′ genes expressed anteriorly and 5′ genes posteriorly (52). This pattern is consistent with our results as hoxc13a is the most 5′ gene of the hoxca cluster, the saddle on the caudal peduncle is the most posterior mark in the hamlets, and hoxc13a is known to be expressed in the caudal peduncle and at the pigment appearance stage in fishes (53, 54). Vertical bars, on the other hand, are anterior to the saddle mark just as the hoxc8-11a gene is on the 3′ side of hoxc13a. A list of the genes found in the 18 genomic regions above the 99.8th £ST percentile is presented in SI Appendix, Table S2 and includes a number of genes that are involved in pigmentation/skin development (e.g., tmem79, mafb, kitlg) and vision/photoreceptor development (e.g., grk7a, rab8a, slc12a5). Functional analyses are needed to test the role played by these candidate loci in vision and pigmentation in the hamlets.

**Origin of Genomic Variants.** The old age (Fig. 1A) and high effective population size of the Hypoplectrus lineage (Fig. 3A) suggest that ancestral variation may have played an important role in the hamlet radiation. In order to test this hypothesis, we estimated the age of the genomic variants within our three regions of interest. We used a nonparametric approach that combines probability distributions of the time to the most recent common ancestor of a large number of haplotypes, allowing us to estimate the age of genomic variants over a continuous timescale and without relying on a priori assumptions about demography or selection (55). We reasoned that the SNPs that are most strongly associated with the three phenotypic traits that we scored (spot on the snout, vertical bars, and saddle on the caudal peduncle) are the ones that are most likely to have a functional role and compared the strength of G × P association with the estimated age of the derived allele for all the SNPs within our three candidate genomic regions. The results indicate that the majority of genomic variants that are strongly associated with each trait predate the radiation (i.e., are older than 10,000 generations), although a smaller number of younger variants also show strong associations (Fig. 6).

**Discussion**

Hamlets live on coral reefs, a highly diverse and complex environment. They feed on small invertebrates and fishes (14, 56) but are not particularly specialized, and their 3-wk pelagic larval phase (10) provides potential for long-distance dispersal. Within this ecological backdrop, the hamlets show an extraordinary burst of speciation that is on par with radiations occurring in depauperate and geographically isolated environments. Moreover,
our genomic data suggest that diversification happened very recently and in a backdrop of substantial gene flow and introgression, resulting in a genetic architecture that is characterized by sharp peaks of differentiation. These peaks contain genes that are known to be involved in vision and pigmentation in other groups and suggest a modular architecture of color pattern variation in the group. Here, the term modular [or combinatorial (46)] refers to the observation that major patterning elements (e.g., vertical bars) and at least one color [black (25)] are associated with one or a few genomic regions and that different species present different combinations of alleles at these loci. There is much more to be explored about the relationship between genomic and phenotypic variation in the hamlets, but such a modular architecture is similar to that observed in Heliconius butterflies (57), capuchino finches (58), and munia finches (59). These radiations also took place in diverse and complex environments and together with hamlets, provide the context for a more general understanding of the drivers of rapid speciation.
In this light, our data also suggest that the hamlet radiation may have been further catalyzed by a tight coupling of loci involved in patterning and vision. The strongest and clearest signal in our data is the association between the presence/absence of vertical bars and genetic variation at the *casz1* locus. This transcription factor has been shown to be involved in the development of photoreceptors in mice (60, 61), and a role in vision is also likely in the hamlets since *casz1* is strongly and consistently expressed in the retinal tissue (25). The strong association with vertical bars that we report here suggests that *casz1*, or a locus in close proximity, might also be involved in patterning. This is similar to what has been observed in *Heliconius* butterflies, a system where strong natural and sexual selection acts on wing color patterns (62). The genetic basis of pattern variation is well characterized in this group, and recent work is demonstrating a tight physical linkage between loci responsible for color pattern variation and preferences for this variation (63). The possibility of physical linkage between patterning and sensory loci in hamlets is significant as it would both facilitate the evolution of reproductive isolation through visually-based assortative mating and help maintain it in the face of ongoing gene flow.

In any event, the occurrence of genes that are known to be involved in vision and pigmentation in peaks of differentiation is broadly in line with the genic view of speciation (64, 65), whereby differentiation between species is initially restricted to genes that are involved in reproductive isolation. Nevertheless, these highly differentiated regions do not expand with increasing genome-wide differentiation as predicted by this framework. This parallels what has been observed in *Heliconius* butterflies (66) and *Ficedula* flycatchers (38) and indicates that divergence hitchhiking does not play the prominent role in the buildup of genomic differences that is implied by the genic view of speciation in these groups. In the hamlets, this finding is consistent with the rapid decay of linkage disequilibrium along chromosomes (25, 45).

While marine systems are generally more open than terrestrial and freshwater habitats due to the scarcity of geographic barriers and the high mobility of many species (7), the hamlets remain nevertheless geographically constrained to the wider Caribbean. Within this region, they are essentially restricted to coral reefs, a highly discrete and patchy habitat. Their pelagic larval duration of 3 wk (10) is relatively short for a reef fish, and larval dispersal has been shown to be more restricted than previously thought in reef fishes generally (67) and in the hamlets in particular (68, 69). This explains the occurrence of local evolutionary processes (e.g., different levels of genetic clustering, differentiation, and hybridization) at our three study sites that are separated by just a few hundred kilometers. The hamlet lineage is also characterized by a history of high effective population size, which is typical of marine populations (11, 12, 70). This allows for both the accumulation of new genetic variants through mutation and the retention of ancestral genetic variants. While both ancestral and new variants are strongly associated with color pattern components, the majority of these variants predate the variation, pointing to an important role of ancestral variation.

We posit that the rapid hamlet radiation is driven by strong selection on color pattern, a modular genetic architecture for this trait, and a mating system that is conducive to the evolution of reproductive isolation among color morphs. Here, the historically high effective population size of hamlets provides a rich genomic substrate from which hybridization can rapidly assemble new phenotypic variation. This attribute is emerging as the common denominator to a variety of radiations on land and in the sea.

**Materials and Methods**

**Software Versions, Parameter Settings, and Scripts.** Software versions and parameter settings were omitted from the text for readability; software versions are listed in *SI Appendix*. Data analysis was managed using nextflow (71). The workflows used to produce our results from raw data to figures are...
Provided in the accompanying repository [accessible from the accompanying repository (72) and the documentation therein; hereafter git].

Sequencing. This study is based on a total of 170 genomes obtained from 167 hamlets and three out-group samples (2 × S. tortugarum and 1 × S. tabacarius). Fifty genomes are new to this study, 110 are from ref. 25, and 10 are from ref. 45. All new tissue samples were available from private individuals via their informed consent (except for two individuals from ref. 23, which were collected in 2017 in Bocas del Toro (Panama) under the Smithsonian Tropical Research Institute Institutional Animal Care and Use Committee protocol 2017-0101-2020-2, the Panamanian Ministry of Environment permits SCA-53-16 and SEXA-35-17, and the Access and Benefit-Sharing Clearing-House identifier ABSCH-IRCC-PA-241203-1). Genomic DNA was extracted from gill tissue using Qiagen’s DNeasy Blood & Tissue kit. Clade-specific rates were performed and sequenced by Novogene and the Institute of Clinical Molecular Biology (Kiel University) on an Illumina HiSeq 4000 (PE; 2 × 151) to a mean postfiltering sequencing depth of 17 ×.

Variant Calling. All the samples considered in this study were genotyped jointly and anew. The variant calling procedure was adapted from the best practice recommendations for the Genome Analysis Toolkit (GATK) workflow (73) provided by the Broad Institute (74, 75). The general workflow is presented below, and the exact parameters used for each step are provided in SI Appendix, Table S1 (git 1.5 to 1.9). GATK was used to transform the raw reads from _fq_ to _uBAM_ format, assign read groups, and mark adapters (git 1.2 to 1.4). The sequences were then back transformed to _fq_ format using GATK, mapped to the hamlet reference genome using BWA (76), and merged with the _uBAM_ files containing the read group information with GATK (git 1.5). Duplicated reads were removed (git 1.6), and genotype likelihoods were computed for a minor allele count to quality and missing data (git 1.14 and 2.6 to 2.7). The SNPs only dataset was phased with SHAPEIT (78) (git 1.16 to 1.17). Bioinformatic phasing is notoriously difficult when it relies on population genetic data only (without, e.g., parent–offspring trios or linked reads). In order to mitigate this issue, we used the read-aware phasing approach implemented in SHAPEIT, which takes the phase information contained within the raw sequencing data into account. Demographic inference (see below) is the only analysis that relies on phase information.

Serraninae Phylogeny and Specification Rates. To reconstruct the phylogenetic position of the hamlets within the Serraninae subfamily, we searched the _H. puella_ reference genome for the 27 genes considered in the FToL project (31) with Basic Local Alignment Search Tool (79) (23 of which were found; git 19.1 to 19.2). Ambiguously aligned positions were automatically removed with GBLOCKS (81), and minor adjustments were made by hand to finalize the mitochondrial gene alignments (we also dropped two poorly aligned genes in the two _Serranus_ species altogether instead of manually editing them). Maximum likelihood reconstruction was performed with IQ-TREE (99) based on a concatenated approach with edge-linked partition model and 1,000 ultrafast bootstrap replicates (git 19.7). For divergence time estimates and evolutionary rate analysis, we used the time-calibrated FToL phylogeny and Bayesian Analysis of Macroevolutionary Mixtures-estimated rates from the project’s data repository (83). Tip-specific speciation rates (λ_BAMM) were plotted for all ray-finned fish species included in the original data. These were then subset to the Serraninae subfamily, and mean speciation rates along the phylogeny were extracted using the BAMM tools. The mean speciation rates for hamlets and Serraninae excluding hamlets, the 95% credible set of rate shift configurations, and macroevolutionary rate cohorts were estimated following the BAMM documentation (git 20.8).

Population Genetic Statistics. Throughout the study, all windowed statistics were computed over 50-kb sliding windows with 5-kb increments unless stated otherwise. _F_ST_. Genetic differentiation was computed from the SNPs only dataset with VCFTools following Weir and Cockerham (84) and using the weighted mean. It was calculated within 50-kb windows for each species pair within each (80) to their FToL homologs (git 19.3 to 19.6). Regarding the latter, only species in the Serraninae were retained, with species considered in the _F_{ST}_{max}_. Genetic divergence (86) was computed from the all BP data within 50-kb windows. The data were reformatted to a custom genotype format, and divergence was computed using popgenWindows.py (87) (git 4.3 to 4.9). The _Δ_ _d_{XY}_. Nucleotide diversity was calculated for each hamlet group (species/population) within 50-kb windows (git 4.19) with popgenWindows.py using the all BP dataset. _G × P_. _G × P_ associations were based on the SNPs only dataset and estimated using a linear model with GEMMA (88). This approach takes population structure into account by considering a matrix of relatedness among individuals. The dataset was transformed to the _f_ vector format using VCFTools and _f_ (89) (git 3.11 to 3.10). _G × P_ association was calculated on an _f_ SNP basis for the presence/absence of three phenotypic traits: vertical bars, saddle mark on the caudal peduncle, and spot on the snout (git 3.17). Phenotyping was based on photographs of all but five samples for which photographs were not available. A Wald test was conducted using GEMMA to determine the association between the phenotype and the genotype on population. The results were averaged over 50-kb windows (10 and 50 kb; git 3.20). Note that Wald test _P_ values were _log_{10} transformed_ before averaging, so _−log_{10}(P_ is reported for every window. GEMMA was also run under the linear mixed model, which provided similar results (SI Appendix, Fig. S20).

Population recombination rate (ρ) was estimated using the _R_ package _FastEPRR_ (90). Briefly, this approach uses boosting (a machine learning approach) to select the best regression model between recombination rate and a set of summary statistics. The analysis was based on the SNPs only dataset considering all samples (except out-groups) and calculated within nonoverlapping windows of 50 kb using 250 parallel jobs (git 6.4 to 6.10). These results were also used to explore the correlation between _ρ_ , _F_{ST}_, and _τ_ with linear regression (SI Appendix, Figs. S6 and S9).

Population genetic divergence (86) was computed from the all BP data within 50-kb windows. The data were reformatted to a custom genotype format, and divergence was computed using popgenWindows.py (87) (git 4.3 to 4.9). The _Δ_ _d_{XY}_. Nucleotide diversity was calculated for each hamlet group (species/population) within 50-kb windows (git 4.19) with popgenWindows.py using the all BP dataset. _G × P_. _G × P_ associations were based on the SNPs only dataset and estimated using a linear model with GEMMA (88). This approach takes population structure into account by considering a matrix of relatedness among individuals. The dataset was transformed to the _f_ vector format using VCFTools and _f_ (89) (git 3.11 to 3.10). _G × P_ association was calculated on an _f_ SNP basis for the presence/absence of three phenotypic traits: vertical bars, saddle mark on the caudal peduncle, and spot on the snout (git 3.17). Phenotyping was based on photographs of all but five samples for which photographs were not available. A Wald test was conducted using GEMMA to determine the association between the phenotype and the genotype on population. The results were averaged over 50-kb windows (10 and 50 kb; git 3.20). Note that Wald test _P_ values were _log_{10} transformed_ before averaging, so _−log_{10}(P_ is reported for every window. GEMMA was also run under the linear mixed model, which provided similar results (SI Appendix, Fig. S20).

Population recombination rate (ρ) was estimated using the _R_ package _FastEPRR_ (90). Briefly, this approach uses boosting (a machine learning approach) to select the best regression model between recombination rate and a set of summary statistics. The analysis was based on the SNPs only dataset considering all samples (except out-groups) and calculated within nonoverlapping windows of 50 kb using 250 parallel jobs (git 6.4 to 6.10). These results were also used to explore the correlation between _ρ_ , _F_{ST}_, and _τ_ with linear regression (SI Appendix, Figs. S6 and S9).

N. and Coalescence Rate. Demographic history was inferred using the multiple sequentially Markovian coalescent method implemented in _msM2_ (44). This analysis was based on the phased SNPs only dataset, which was prepared following the _msM2_ authors recommendations (91) as detailed in ref. 45. This included masking the data on the basis of mappability to the reference genome and the occurrence of indels. The data were also filtered with respect to coverage for each individual (between 10× and twice the indivs). Cladean coverage; git 8.10). Individuals from each species and location were randomly grouped into sets of three or four, with each individual included in only one set (git 8.12), and individual masks were combined to create the _msM2_ input files (git 8.16 to 8.21). Individuals
were also grouped for the cross-coalescence rate analysis, with each group containing two individuals of each species for all pairs of sympatric species. Here, each individual was assigned to only one group for each species pair but reused across species pairs. All MSMC2 analyses were run with a time segment of 1.0 million years and a mutation rate of \(10^{-8}\) substitutions per site per generation. The mutation rate was set to \(3.7 \times 10^{-8}\) based on the closest relative for which we could find a reliable mutation rate estimate (91, 92) (git 8.18 and 8.23).

**Identification of Putative Hybrids and Backcrosses.** We used the approach implemented in NewHybrids (93) to evaluate ongoing gene flow among sympatric species. This method is based on patterns of Mendelian inheritance at highly differentiated loci. This analysis was based on small subsets of the SNPs only dataset. First, the 800 most differentiated SNPs were selected for each pair of sympatric species (git 5.2 to 6.6). These were then filtered for a minimum physical distance of 5 kb to reduce linkage among them using VCFtools (git 9.7). From this SNP set, 80 SNPs were randomly chosen using bash scripting and converted to the NewHybrids input format using PGDSpider (94). The assignment to hybrid classes with NewHybrids was performed using the Bonferroni procedure (git 17.5 to 17.8).

**Genome-Wide Hamlet Phylogeny.** To reconstruct the phylogenetic relationships among hamlets, we divided the genome into nonoverlapping windows of 5 kb. A total of 5,000 were selected randomly from those windows that contained less than 20% missing sites and at least 50 SNPs within hamlets in the all BP genotype dataset (69% of all windows met these criteria; git 13.1 to 13.11). The selected windows were extracted using VCFtools, and Serranus samples were removed in one copy of the dataset. After conversion to continuous sequence (Fasta format) using a custom Perl script, windows were individually realigned with MAFFT (80) (git 13.12 to 13.17). Maximum likelihood phylogenies were calculated for each window in IQ-TREE (82) based on the automatically selected best-fit model and default parameters (git 13.18). A coalescent-based species tree was then estimated from all local trees with ASTRAL-III (99), with internal branch length measured in coalescent units and support values given as local posterior probabilities (100) (git 13.19). Two species trees were produced that way, one rooted with H. floridae from the Gulf of Mexico clade and another one rooted with the three Serranus samples.

**Region-Specific Phylogenies.** The three genomic regions of interest were extracted from the all BP data (i.e., including invariant sites) set using VCFtools. All Serranus and hybrid samples identified by NewHybrids were removed alongside indel positions (git 14.1 to 14.3) before conversion to Fasta format and per site allele frequencies for each group using the ciftlib library. Region-specific phylogenies were then inferred at the level of individual samples with RaxML-NG (101) based on the GTR + G model, 10 each of random and parsimony starting trees, and 100 bootstrap replicates (git 14.5 to 14.6). Group-level maximum likelihood trees for each region were also estimated with IQ-TREE (82) using a polymorphism-aware model (102) and 100 nonparametric bootstrap replicates (git 14.4).

**Topology Weighting.** Topology weighting was conducted for Belize and Honduran indigenously (the number of taxa in Panama is too small to conduct this analysis). The SNPs only dataset was subset to include only hamlets from the respective location and filtered to include only SNPs with a minor allele count greater than or equal to three (git 5.12). The data were then split by LG and converted to a custom genotype format (git 5.14). Using phylm sliding_windows.py (87), we applied PhyML (103) to build phylogenies from nonoverlapping sliding windows of 200 SNPs each along all LGS (git 5.17). Topology weighting was conducted on the resulting phylogenies using Twiss (48) (git 5.18).

**IBD Fragments.** Fragments of IBD were identified using trufife (104). Outgroup samples were removed from the SNPs only dataset with VCFtools (git 16.2), and trufife was run with three different minimum sequence length thresholds for IBDS1 and IBDS2 (25/20, 15/7.5, and 10/5 × 10^{-3} consecutive SNPs; git 16.3 and 16.5). IBD fragment lengths were then converted from base pairs to centimorgans based on two available linkage maps for Hypoplectrus (42) (git 16.6). IBD networks were then constructed from the pairwise genome-wide linkage intervals. The edges in the networks were defined as edge weights in a force-directed graph (105). In order to limit the effect of low recombination and stabilizing selection on the detection of IBD fragments, the IBD fragments that overlapped with regions of the genome above the 95th IBD score percentile (git 16.4 to 16.5) and that were smaller than 0.2 cM were also filtered out.

**Admixture Analysis.** We used admixture (106) to analyze the population genetic patterns at three genomic regions of interest. For this, the out-group species were removed from the dataset, and the genomes were subset to the respective candidate regions and converted to plink format (git 10.4). Admixture was run for all k values between 2 and 15 for each region (git 10.6).

**Allele Age Estimation.** This was done using Genealogical Estimation of Variance Age, a nonparametric method that combines probability distributions of the time to the most recent common ancestor of a large number of haplotypes based on empirically constructed hidden Markov models to estimate the age of genomic variants over a continuous timescale and without relying on a priori assumptions about demography or selection (55). The ancestral state of all SNPs was determined using the Serranus out-groups and allele frequency (git 11.3). For SNPs that were invariant in both S. fabaricus and S. tortugatus, the Serranus allelic state was set as ancestral. For the SNPs that were variant within Serranus, the major allele was set as ancestral. The information about the ancestral state was added as an annotation to the vcf file (git 11.3) using vcf-annotate, and the vcf file was recoded using a custom java script in combination with Javakit (107) (git 11.4). GEVA was looped over the entire dataset in batches of 250 SNPs, with a mutation rate of 3.7 × 10^{-8} and an estimated effective population size of 3 × 10^{5}. Recombination rate was fixed at the average of each LG and based on the FastEPR estimation, ranging from 1.95 × 10^{-7} to 5.13 × 10^{-10} (git 11.6). The results were then compiled for each LG (git 11.7).

**Visualization.** All results were plotted using R (git 20). The details of the visualization are provided in the R scripts and their documentation (git docs/index.html; file within the git repository). Other than the scripts within the Github repository, the visualization relied on the three custom R packages (hypogen, hypoimg, and GenomicOriginsScripts), which can also be accessed via GitHub (108). The R packages used are listed and referenced in SI Appendix. Package versions were managed using the R package renv; thus, the precise R configuration used for this study can be restored based on the provided lock file (git renv.lock; file within the git repository).

**Data Availability.** All raw sequencing data are deposited in the European Nucleotide Archive (project accession no. PRED354599). Whole-genome sequencing data, genotype data, population genetic summary statistics, and code used for data analysis have been deposited in Dryad (https://doi.org/10.5061/dryad.280b16wy6; (109) and Zenodo (https://doi.org/10.5281/zenodo.4709890) (72) and https://doi.org/10.5281/zenodo.4709767) (108). Individual sample accession numbers are provided in SI Appendix, Table S3.

**This study was funded by a grant from the Smithso-
Rapid radiation in a highly diverse marine environment

56. 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).

85. Y. Benjamini, D. Yekutieli, The control of the false discovery rate in multiple testing. Ann. Stat. 29, 1165–1188 (2001).

99. C. Zhang, M. Rabiee, E. Sayyari, S. Mirarab, ASTRAL-III: Polynomial time species tree reconstruction from partially resolved gene trees. BMC Bioinformatics 19 (suppl. 6), 152 (2018).

100. E. Sayyari, S. Mirarab, Fast coalescent-based computation of local branch support under dependency. Bioinformatics 36, 1294–1301 (2019).

101. A. Dimitromanolakis, A. D. Paterson, L. Sun, Fast and accurate shared segment inference and ambiguously aligned blocks from protein sequence alignments. Mol. Biol. Evol. 33, 268–274 (2016).

103. S. Guindon et al., New algorithms and methods to estimate maximum-likelihood phylogenies: Assessing the performance of PhyML 3.0. Syst. Biol. 59, 307–321 (2010).

105. T. M. J. Fruchterman, E. M. Reingold, Graph drawing by force-directed placement. Softw. Pract. Exper. 21, 1129–1164 (1991).

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

107. P. Lindenbaum, R. Redon, Bioalddna, samjs and vcffilterjs: Object-oriented formatters and filters for bioinformatics files. Bioinformatics 34, 1224–1225 (2018).

108. K. Hench, k-hench/GenomicOriginsScripts: PNAS revisions. Zenodo. https://doi.org/10.5281/zenodo.4709767. Deposited 22 April 2021.

109. K. Hench, W. O. McMillan, O. Puebla, M. Helmkampf, Data from: Rapid radiation in a highly diverse marine environment. Dryad. https://doi.org/10.5061/dryad.280gb5mmt. Deposited 8 November 2021.