Received Date : 17-Jun-2014

Revised Date : 28-Aug-2014

Accepted Date : 12-Sep-2014

Article type : Original Article

Genomic atolls of differentiation in coral reef fishes (*Hypoplectrus spp, Serranidae*)

O. Puebla[*†], E. Bermingham[‡], W.O. McMillan[†],

[*]*GEOMAR Helmholtz Centre for Ocean Research Kiel, Evolutionary Ecology of Marine Fishes, Düsternbrooker Weg 20, D-24105 Kiel, Germany.* [†]*Smithsonian Tropical Research Institute, Apartado Postal 0843-03092, Panamá, República de Panamá.* [‡]*Patricia and Phillip Frost Museum of Science, 3280 South Miami Avenue, Miami, FL*

Keywords: genomic architecture, speciation, RAD sequencing, *Hox* genes, marine

Correspondence: Oscar Puebla, GEOMAR Helmholtz Centre for Ocean Research Kiel, Düsternbrooker Weg 20, 24105 Kiel, Germany, Tel +49 431 600 4559, Fax +49 431 600 4553, email oscar.puebla@mail.mcgill.ca

Running title: Genomic atolls in *Hypoplectrus*

Abstract

Because the vast majority of species are well-diverged, relatively little is known about the genomic architecture of speciation during the early stages of divergence. Species within recent evolutionary radiations are often minimally diverged from a genomic perspective, and therefore provide rare

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1111/mec.12926

opportunities to address this question. Here, we leverage the hamlet radiation (*Hypoplectrus* spp, brightly colored reef fishes from the tropical western Atlantic) to characterize genomic divergence during the early stages of speciation. Transect surveys and spawning observations in Belize, Honduras, and Panama confirm that sympatric barred (*H. puella*), black (*H. nigricans*) and butter (*H. unicolor*) hamlets are phenotypically distinct and reproductively isolated, although hybrid spawnings and individuals with intermediate phenotypes are seen on rare occasions. A survey of approximately 100,000 restriction-site associated SNPs in 126 samples from the three species across the three replicate populations reveals extremely slight genome-wide divergence among species (F_{st}=0.0038), indicating that ecomorphological differences and functional reproductive isolation are maintained in sympatry in a backdrop of extraordinary genomic similarity. Nonetheless, a very small proportion of SNPs (0.05% on average) are identified as F_{st} outliers among sympatric species. Remarkably, a single SNP is identified as an outlier in repeated populations for the same species pair. A mini-contig assembled *de novo* around this SNP falls into the genomic region containing the *HoxCa10 and HoxCa11* genes in 10 teleost species, suggesting an important role for *Hox* gene evolution in this radiation. This finding, if confirmed, would provide a better understanding of the links between micro- and macroevolutionary processes.

Introduction

Recent evolutionary radiations such as Darwin's finches or East African cichlids have served as model systems to understand how new species arise (Schluter 2000), and have arguably transformed our understanding of the origins of biodiversity (Grant & Grant 2014; Wagner *et al*. 2012). With the advent of next-generation sequencing, it is now possible to extrapolate this discovery process to the genomic level. Recent radiations are particularly interesting in this regard, because they provide rare windows into the early stages of genomic divergence (Seehausen *et al*. 2014). Recently diverged genomes also represent good opportunities to detect genomic elements that may be under divergent

selection, because such elements are expected to clearly stand out with high F_{st} estimates against a backdrop of low genetic divergence (Pérez-Figueroa *et al*. 2010; Vilas *et al*. 2012). An important limitation, though, is that loci with high F_{st} estimates are also expected by chance, or due to demographic processes that have little to do with divergent selection (Bierne *et al*. 2011). One strategy to filter out such 'false positives' consists in repeating comparisons in multiple populations; some loci may show high F_{st} estimates in a given population by chance or due to specific demographic processes, but loci linked to genomic elements under divergent selection are expected to present consistently high F_{st} estimates across populations.

Here, we leverage the recent radiation in the hamlets (*Hypoplectrus* spp, Serranidae), brightly colored coral reef fishes from the tropical western Atlantic, to explore the genomic architecture during the early stages of divergence. The hamlets provide an important marine equivalent to the classic terrestrial and freshwater radiations that promises to promote our understanding of the origin of variation in the sea. As in other recent radiations, mitochondrial phylogenies of Caribbean hamlets show very little divergence among species (McCartney *et al*. 2003), a result confirmed by microsatellite (Puebla *et al*. 2007, 2012) and AFLP (Barretto & McCartney 2007; Holt *et al*. 2011) data. Notwithstanding low levels of genetic differentiation and a history of debate among ichthyologists (reviewed in Domeier 1994 and Lobel 2011), there are 18 hamlet species recognized today, 6 of which have been described recently (Lobel 2011; Del Moral Flores *et al*. 2011; Victor 2012; Tavera & Acero 2013). The hamlets are highly sympatric, with up to 9 species found on a single reef (Puebla *et al*. 2012). Color pattern and mate choice are the only traits that have been found to consistently differentiate hamlet species, although ecological differences have been noted on a few occasions (Whiteman *et al*. 2007). A combination of natural (Puebla *et al*. 2007) and sexual (Puebla *et al*. 2012) selection on color pattern has been proposed to explain the origin and maintenance of species within the radiation.

The goal of our study is to identify regions of the genome that are more divergent among incipient species than expected by chance. These regions likely contain the loci that underlie species differences in morphology and behavior. To this end, we take a genotyping-by-sequencing (GBS) approach to scan the genome of three sympatric hamlet species with very low levels of genetic divergence, and repeat these comparisons in three populations. GBS provides the opportunity to finely characterize genetic variation at tens of thousands of loci across the genome (Hohenlohe *et al*. 2010; Keller *et al*. 2013). Among the various GBS methods developed (reviewed in Davey *et al*. 2011), the *Restriction-site Associated D*NA (RAD) protocol by Etter *et al*. (2011) is particularly appealing because random shearing of the paired-ends provides an opportunity to filter out PCR clones and assemble mini-contigs *de novo* around SNPs of interest, two procedures that cannot be easily applied with double-digest approaches. Of the approximately 100,000 SNPs surveyed with this method, only one showed consistent and significant differences among the morphologically and behaviorally distinct species we sampled. Remarkably, this SNP falls into a region of the genome containing a *Hox* gene cluster—a class of genes known to be important in animal diversification.

Materials and methods

Sampling, transects and spawning observations

The sampling design targeted three sympatric species (the barred hamlet *Hypoplectrus puella*, the black hamlet *Hypoplectrus nigricans*, and the butter hamlet *Hypoplectrus unicolor*) from three locations (Belize, Honduras, Panama), providing the opportunity to compare multiple pairs of sympatric species and repeat each comparison multiple times. These species were chosen because microsatellite data indicate that they are very close genetically at these locations (Puebla *et al*. 2007, 2012). A sample size of 14 individuals per species per location was targeted, totaling 42 individuals per species and 126 individuals overall. Samples were collected between 2004 and 2006 in the vicinity of Carrie Bow Cay (Belize), the Becerro Keys (Honduras) and Guna Yala (Panama) as detailed in Puebla *et al*. (2007).

In order to link the genomic patterns to phenotypic variation, behavioral reproductive isolation and variation at microsatellite loci, data from field survey transects, spawning observations, and 10 microsatellite loci from Puebla *et al*. (2007, 2012) were recompiled and reanalyzed for the populations and species considered here specifically.

Library preparation and sequencing

DNA was extracted with DNeasy Blood & Tissue Kit columns (QIAGEN) from gill tissue preserved in salt-saturated DMSO and treated with RNase A (QIAGEN). Libraries were prepared following the RAD protocol by Etter *et al*. (2011) using 500-1000 ng of DNA per sample (mean=936, data from a pilot library indicated that coverage is not affected by variation within this range) and the *Sbf*I restriction enzyme (NEB). Samples were identified with 63 5-bp indices on the P1 adapter (Table S1, Supporting information), divided into two libraries of 63 samples each. DNA was sheared with a Covaris sonicator using a duty cycle of 10%, an intensity of 4 and 200 cycles per burst for a total of 48 seconds. The sheared libraries were run on separate agarose gels and the 300-500 bp range was manually excised, purified, and enriched with 16 amplification cycles in 10 individual PCR reactions containing 5 µl of Phusion High-Fidelity PCR Master Mix with HF Buffer (NEB), 0.2 µl of each amplification primer (10 µ*M*), 2 µl of library (9 ng/µl) and 2.6 µl of water (total 10 µl). Each library was sequenced on a lane of a HiSeq 2000 Illumina sequencer (paired-end, 2x101 bp).

Raw sequences filtering

Raw sequences were filtered using the *process_radtags* pipeline in Stacks version 1.08 (Catchen *et al*. 2011, 2013). This included the removal of low quality reads (with an average raw phred score <10 within a 15-bp sliding-window), of reads with an ambiguous index at position 1-5 of read 1 (where one of the 63 indices is expected), of reads with an ambiguous *Sbf*I restriction site at position 6-12 of read 1 (where TGCAGG is expected), and of reads including adapter sequences. Since all pairs of indices differed by at least two bp, sequences that differed by a single bp from any expected index (or

the restriction site sequence) were corrected and retained. The pipeline *clone_filter* in Stacks was used to filter out pair of paired-end reads that match exactly, since these are expected to represent PCR clones.

Assembly

In the absence of a reference genome for *Hypoplectrus*, reads were assembled *de novo* using the *denovo_map.pl* pipeline in Stacks. For the main analyses presented below, the number of raw reads required to form a stack (stack depth parameter, *m*) was set to 3 and the number of allowed nucleotides mismatch between two stacks (mismatch parameter, *M*) to 2. In order to test the robustness of the results to these assembly parameters, the main analyses were rerun with *m*=3 *M*=3, *m*=4 *M*=2, *m*=5 *M*=4 and *m*=10 *M*=4.

Population genetic statistics

Because an important objective of the study was to screen as many loci as reasonably possible to identify outliers, moderate filtering was applied to the dataset. For consistency, similar levels of filtering were applied for all analyses. As indicated in the Results, similar genomic patterns are observed with more stringent filtering.

Samples were initially pooled by species (*n*=42 samples per species) and stacks with coverage ≥10x in ≥15 individuals per species in ≥2 species were retained. From this pool, stacks that included a SNP with observed heterozygosity >0.5 in >1 species were removed. The dataset was reformatted with PGDSpider version 2.0.5.2 (Lischer & Excoffier 2012) and F_{st} were estimated following a standard ANOVA approach (Weir & Cockerham 1984) using Genepop version 4.2.1 (Rousset 2008). We note that F_{st} estimates may take negative values under this framework.

Samples were then grouped by species and location (*n*=14 samples per group) and stacks with coverage ≥10x in ≥5 individuals per group in ≥7 groups were retained. Here again, stacks that included a SNP with observed heterozygosity >0.5 in >1 species were removed. This reduced dataset

was used to estimate F_{st} between sympatric species in each population, following the same approach as above.

Clustering analyses

Clustering analyses were performed to further explore genetic structure. All stacks with coverage ≥10x in ≥15 individuals per species in ≥2 species were considered for these analyses, but this time considering a single SNP per stack (the first one). Structure version 2.3.4 (Pritchard *et al*. 2000) was used for these analyses, considering the admixture model with correlated frequencies (Falush *et al*. 2003). The λ parameter was estimated from 10 initial runs with the number of presumed clusters (K) set to 1, and the mean estimated value (0.373) was used for all subsequent runs. Species/location information was not used to pre-assign individuals to clusters or to improve clustering. A burnin period of 100,000 MCMC iterations was used, followed by 100,000 iterations for each run. K was set from 1 to 10 and 10 replicate analyses were run for each value of K (100 runs per analysis).

A combination of two approaches was adopted to infer the number of clusters present in the dataset. First, the number of clusters was considered to correspond to the K value with highest ln Pr($X|K$), or the one after which the trend plateaus and that also provided consistent groupings across repeated runs (Pritchard *et al*. 2000). The *ad hoc* statistic ΔK (Evanno *et al*. 2005) was also considered, keeping in mind that this approach does not apply when K=1.

In order to get a sense of what proportion of the genome might be differentiated between species, genetic structure was analyzed with different SNP subsets. These were established according to global F_{st} estimates among species (Figure 1a), considering the interval above the 90th percentile (F_{st}≥0.0243), between the 80th and 90th percentiles (0.0094≤F_{st}<0.0243), between the 70th and 80th percentiles (0.0027≤F_{st}<0.0094), between the 60th and 70th percentiles (0.0006≤F_{st}<0.0027), and below the 60th percentile (F_{st}<0.0006).

SNP trees

In order to also adopt a phylogenetic perspective, maximum likelihood trees were generated from SNP data (i.e. sites variant among individuals, concatenated and exported using Stacks). The fact that all sites are variable in this situation generates an ascertainment bias, which is problematic for branch length and topology inference (Lewis 2001). One strategy to address this difficulty, which we adopted here, consists in applying an ascertainment bias correction to the likelihood calculations (Lewis 2001). Such trees should nonetheless be interpreted with caution, since questions remain regarding phylogenetic inference from large genomic datasets (Wagner *et al*. 2013).

As for the clustering analyses, all stacks with coverage ≥10x in ≥15 individuals per species in ≥2 species were considered, keeping a single SNP per stack. The dataset was analyzed with jModeltest version 2.1 (Darriba *et al*. 2012) and the Akaike Information Criterion, Bayesian Information Criterion, and Decision Theory all indicated that a GTR+G model of nucleotide substitution was appropriate. RAxML version 8.0.5 (Stamatakis 2014) was used for these analyses, implementing the GTR+G model with ascertainment bias correction as recommended by the author for SNP data requiring a model of rate heterogeneity (RAxML v8.0.x manual, Feb 2014). The rapid bootstrap procedure was implemented (Stamatakis *et al*. 2008), with 100 replicates per run. Analyses were run with the entire SNP dataset, and repeated with the same SNP subsets considered for the clustering analyses.

F_{st} outlier analyses

In order to identify SNPs that may be under divergent selection, outlier scans were run between all species pairs in all populations. All loci with coverage ≥10x in ≥5 individuals in both populations were considered for these analyses, selecting a single SNP per stack. (In order to make sure that no potential outlier SNP was missed, the analyses were rerun with different SNP subsets from the same stacks). Bayescan version 2.1 (Foll & Gaggiotti 2008) was used for these analyses, with default

parameters for run length. The prior odds for the neutral model were initially set to 10 (default value), implying a prior belief that the neutral model is 10 times more likely than the model with selection at any locus. In order to evaluate the effect of these prior odds on the results, all analyses were rerun with this parameter set to 100 and 1. A locus was considered an outlier if it had a *q*-value <0.2, corresponding to an expected false discovery rate of 20%. Additional population differentiation methods were not applied because they have been shown to be less efficient than Bayescan (Pérez-Figueroa *et al*. 2010; Vilas *et al*. 2012). We note that outlier detection approaches that incorporate geographic or environmental variables do not strictly apply in this case, since we are comparing sympatric species.

Mini-contigs

In an attempt to characterize the single repeated outlier SNP, a mini-contig was assembled *de novo* around it using the paired-end reads for this locus. The pipeline *sort_read_pairs.pl* in Stacks was used to export these reads, and *de novo* assembly was carried out using Velvet version 1.2.03 (Zerbino & Birney 2008). Sequences matching the consensus sequence were searched using megablast on the NCBI server (http://www.ncbi.nlm.nih.gov/blast) and Blastn searches to a variety of teleost genomes on the Ensembl genome browser (Flicek *et al*. 2014, http://www.ensembl.org/index.html). This procedure was also applied to the most differentiated SNPs (F_{st}≥0.296, the 99.99th percentile).

Negative control

In order to contrast our results to what may be expected by chance, we randomly grouped the 126 samples into 3 ‘species’ and 9 ‘populations’, and rerun the analyses on this dataset.

Results

Sampling, transects and spawning observations

A list of the 126 samples with identification number, collection date, collection site and coordinates is presented in Table S1 (Supporting information). Voucher specimen, tissue samples and photographs of all samples are available upon request.

Location, depth, date and raw data from 144 transects made at the time of sampling in the same area are presented in Table S2 (Supporting information). A total of 984 barred, black and butter hamlets were observed, all of which could be unambiguously identified to species except for two individuals that were phenotypic intermediates (one *H. nigricans*/tan intermediate and one *H. puella*/*H. indigo* intermediate).

Location, area, depth, date and raw data of 123 spawning observations involving barred, black and butter hamlets at the time of sampling in the same area are presented in Table S3 (Supporting information). All spawnings were assortative, i.e. between members of the same species, except for three hybrid spawnings *(H. puella*/*H. aberrans*, tan/*H. nigricans*, and *H. nigricans*/*H. aberrans*), and two cases where one of the partners could not be unambiguously identified (one *H. puella*/*H. unicolor* intermediate and one *H. puella*/*H. aberrans* intermediate).

Raw sequences filtering

A total of 673,488,582 reads of 101 bp each were obtained. Of these 8.6% were discarded due to low quality, 5.9% because pairs of paired-end reads matched exactly (PCR clones), 1.0% due to the presence of adapter sequence in the read, 0.6% due to ambiguous restriction site and 0.1% due to ambiguous indices. Overall 565,253,125 sequences (83.9% of the raw reads) were retained.

Assembly

The assembly with stack depth parameter m=3 and mismatch parameter M=2 provided an average of 53,811 stacks per sample, with a mean coverage per stack per sample of 31x. Coverage was relatively balanced with respect to species (40x for *H. puella*, 31x for *H. nigricans* and 21x for *H.*

unicolor), but samples from Panama had lower mean coverage (14x) than samples from Belize (32x) and Honduras (46x). The weak correlation between initial DNA concentration and coverage (R^2=0.007) suggests that disparities in coverage may be due to DNA (or adapter) quality rather than quantity. Preliminary analyses indicated that results were broadly consistent across species and population notwithstanding disparities in coverage, so all species/populations were retained for analyses. As expected (Catchen *et al*. 2013), the number of stacks decreased with increasing *m* and *M* parameter values (Table S4, Supporting information). Yet the five assemblies with different combinations of *m* and *M* parameters provided similar global F_{st} estimates (0.0036-0.0041) and proportion of outliers (0.04%-0.05%), and the same repeated outlier was consistently identified.

Population genetic statistics

With samples pooled by species, a total of 52,459 stacks were retained after data filtering. A total of 96,418 SNPs were identified, i.e. 1.8 SNP per stack on average. Number of sites, number and proportion of polymorphic sites, mean number of individuals sampled per site, nucleotide diversity (π) and expected heterozygosity for each species are presented in Table 1. The three species presented similar parameters with a slightly lower diversity (π) for *H. unicolor*, which is consistent with the smaller population size of this species in the sampled areas.

Global F_{st} among the three species (considering all SNPs) was estimated to 0.0038. Close estimates of 0.0041 and 0.0034 were obtained when considering only one SNP per locus or when applying more stringent filtering (loci present in ≥32 individuals per species instead of 15), respectively. Pairwise F_{st} were estimated to 0.0025 (*H. puella/H. unicolor*), 0.0040 (*H. puella/H. nigricans*) and 0.0057 (*H. nigricans*/*H. unicolor*). The distribution of SNP F_{st} estimates was characterized by a sharp mode close to 0, and a long tail extending to a value of 0.512 (Figure 1a).

With samples grouped by species and population, a total of 31,059 stacks were retained after filtering, providing 55,195 SNPs. F_{st} estimates between sympatric species in each population ranged

between 0.0028 (*H. puella*/*H. nigricans* from Honduras) and 0.0198 (*H. nigricans*/*H. unicolor* from Panama, Table 2). F_{st} estimates based on microsatellite data from the same area and collected at the same time (Puebla *et al*. 2007, 2012) are also presented in Table 2.

Clustering analyses

Results of the clustering analyses are illustrated in Figure 1b-f and detailed in Table S5 (Supporting information). Considering a single SNP per stack, a total of 41,690 SNPs were retained. Using the entire dataset, no evidence of clustering was found. For the 10 replicate runs, ln Pr(*X*|*K*) was systematically higher for *K*=1 than for any other value of *K* (Table S5, Supporting information).

In sharp contrast, the SNP subsets from the 90^{th}-100^{th}, 80^{th}-90^{th}, and 70^{th}-80^{th} percentiles provided strong evidence of clustering. The highest mean ln Pr(*X*|*K*) corresponded to *K*=4 (90^{th}-100^{th} percentile), *K*=3 (80^{th}-90^{th} percentile) and *K*=2 (70^{th}-80^{th} percentile), the ΔK statistic presented a sharp peak at these same values of *K* (Table S5, Supporting information), and the 10 replicated runs provided identical groupings. For the 90^{th}-100^{th} percentile, the four clusters matched the three species, and also differentiated between populations of *H. nigricans* (Figure 1c). For the 80^{th}-90^{th} percentile, the three clusters matched the three species except for one *H. puella* individual from Panama, which clustered with *H. nigricans* in the ten replicate runs (Figure 1d). For the 70^{th}-80^{th} percentile, the two clusters matched broadly *H. puella* and *H. unicolor* (Figure 1e). Clusters were also identified with SNPs from the 60^{th}-70^{th} percentile, but the evidence was less clear in this case and the structure very subtle. No evidence of clustering was found with SNPs from the 0-60^{th} percentile (Table S5, Supporting information). Similar patterns (no clustering with all the data, clustering by species with the most divergent SNPs) were observed with more stringent filtering (loci present in ≥32 individuals per species instead of 15, data not shown).

SNP trees

No phylogenetic signal was detected when considering the entire dataset, with a bootstrap support value of 0 for the central node and samples from different species and populations mixed over the

tree (Figure 1g). The SNP subset from the 90^{th}-100^{th} percentile presented a much different picture, with samples grouped by species (but not by populations) and supported by a bootstrap value of 77 (Figure 1h). The SNP subsets from the 80^{th}-90^{th} percentile, 70^{th}-80^{th} percentile, 60^{th}-70^{th} percentile and below the 60^{th} percentile did not reveal any phylogenetic signal, with trees similar to Figure 1g (data not shown).

F_{st} outlier analyses

Results of the F_{st} outlier analyses are detailed in Table 3. With the prior odds for the neutral model set to 10, a total of 84 outliers were identified, representing 0.05% of the SNPs analyzed. A single one of these was identified as an outlier in more than one population for the same species pair, in this case between *H. nigricans* and *H. puella* in Belize (F_{st} estimate=0.751) and Honduras (F_{st} estimate=0.713). It was not considered in Panama because it was below the minimum coverage threshold for this species pair, but this SNP was strongly differentiated in Panama as well (F_{st} estimate=0.378). This illustrates the fact that in order to be identified as repeated outliers, candidate SNPs need to be sequenced with good coverage in repeated species and populations. In our dataset 80% of the outlier SNPs identified in one population were also genotyped in at least one other population for the same species pair, indicating that the small number of repeated outliers identified is not mainly due to coverage issues. The absence of outliers in Panama is probably due to the lower coverage and therefore lower number of SNPs surveyed in this population.

With the prior odds for the neutral model set to 100, a total of 16 outliers were identified, representing 0.01% of the SNPs analyzed. Here again a single repeated outlier was identified, the same as above.

With the prior odds for the neutral model set to 1, a total of 1005 outliers were identified, representing 0.56% of the SNPs analyzed. Of these, 24 were identified as repeated outliers. Given the relaxed prior odds for the neutral model in this case, a large proportion of these outliers are expected to be false positives. Nevertheless, the single outlier SNP identified above with more stringent

parameters became a 'quadruple outlier' in this case: between *H. nigricans* and *H. puella* in Belize (F_{st}=0.751) and Honduras (F_{st}=0.713), and also between *H. nigricans* and *H. unicolor* in Belize (F_{st}=0.394) and Honduras (F_{st}=0.584).

Mini-contigs

A total of 3,786 paired-end reads were extracted from the locus containing the repeated outlier SNP and assembled. The mini-contig aligned with the consensus sequence of the P1 read over 58 bp, so this sequence was included in the final assembly, providing a consensus sequence of 460 bp. Coverage was bell-shaped, with a mean of 735x and a maximum of 2,313x at position 245. The blast searches returned 11 hits, summarized in Table 4. All were between the *HoxC10a* and *HoxC11a* genes of other teleosts (E-values between 2E-6 and 1E-95). A single hit per species was found in nine cases, suggesting that the stack containing this SNP did not merge paralogs from the three genome duplication events thought to have happened throughout the evolutionary history of teleosts (Jaillon *et al*. 2004). The only exception was the Atlantic salmon, with two hits (one between *HoxC10a*α and *HoxC11a*α, and another between *HoxC10a*β and *HoxC11a*β). This is consistent with the hypothesis that salmonids have gone through a fourth, more recent round of genome duplication (Allendorf & Thorgaard 1984). No hits were found to more distantly related taxa, e.g. the spotted gar (*Lepisosteus oculatus*, Holostei).

Five hits were found for the mini-contigs assembled around the other nine most differentiated SNPs (F_{st}≥0.296, the 99.99th percentile). These included coding regions of cysteine conjugate-beta lyase 2 (*Ccbl2*, E-values between 9E-12 and 2E-119), spermine oxidase (*Smox*, 3E-30 - 1E-67), THO complex subunit 6 homolog (*Thoc6*, 1E-10 - 2E-14), and a non-coding region between 1.6 and 3 kb from UDP glycosyltransferase 8 (*Ugt8*, 7E-09 - 6E-29). Another hit was near the stickleback *HoxC11a* (3E-16). Close examination of this alignment indicated that it is directly flanking the stack containing our repeated outlier SNP, i.e. on the other side of the same *Sbf*I restriction site. This locus

is a non-repeated outlier between black and barred hamlets in Honduras, also strongly differentiated between these two species in Belize (F_{st}=0.254) and Honduras (F_{st}=0.387).

Negative control

The randomized dataset provided a global F_{st} estimate of 0.0001 and a distribution of SNP F_{st} estimates similar to Figure 1a, but with a shorter tail (all estimates <0.394). No evidence of clustering and no phylogenetic signal were detected, even when considering only the SNPs above the 90th percentile (Table S5 and Figure S1, Supporting information). A total of 7 outliers (0.003% of the SNPs analyzed) were identified with the prior odds for the neutral model set to 10, and no repeated outlier was found. These results suggest that the genomic patterns reported here are not expected by chance.

Discussion

The data presented here indicate that ecomorphological differences and functional reproductive isolation can persist in sympatry in a backdrop of extraordinary genomic similarity. Global F_{st} between sympatric *H. puella*, *H. nigricans* and *H. unicolor* across 96,418 SNPs was estimated to 0.0038, slightly higher than the estimate of 0.0022 provided by 10 microsatellite loci for the same populations (the difference is consistent with the higher diversity and larger sample size of the microsatellite dataset). The distribution of SNP F_{st} estimates presented a sharp mode close to 0, with 99% of the estimates <0.1 (Figure 1a). Considering a single SNP per locus, both clustering and phylogenetic analyses failed to differentiate between the three species, a remarkable outcome given the number of loci involved (41,690 SNPs). It is important to keep in mind that the species/populations considered in this study were targeted precisely because microsatellite data provided F_{st} estimates in the lower range of observed values across the Caribbean (Puebla *et al*. 2007, 2012). Nonetheless, levels of genomic differentiation are lower than that observed between sympatric Lake Victoria cichlids (Keller *et al*. 2013), adjunct populations of marine and freshwater

sticklebacks (Hohenlohe *et al*. 2010), and even freely hybridizing parapatric races of *Heliconius* butterflies (F_{st}=0.009, Martin *et al*. 2013).

The striking genomic similarity contrasts with the data from transect surveys and spawning observations, made at the time of sampling in the same populations. A total of 144 transects covering 57,600 m^2 of reef and 123 spawning observations confirm that black, barred and butter hamlets from these populations are clearly distinct in terms of color pattern, and reproductively isolated from a behavioral perspective. Nonetheless, the species barrier is porous and reproductive isolation not complete, which could contribute to explain the genomic patterns. About 2.5% of the spawnings were between different species, and cross-fertilization experiments have shown that eggs and larvae from hybrid spawnings grow, develop and survive normally (Whiteman & Gage 2007). Of the 1107 barred, black and butter hamlets seen during the transect surveys and spawning observations, 4 individuals had intermediate phenotypes. Laboratory raised F_1 hybrids between *H. unicolor* and *H. gemma* have been shown to have color patterns that are intermediate between the parental species (Domeier 1994), but little is known about the phenotype of other interspecific crosses. Thus, whether the oddly patterned individuals we observed were F_1 or later generation hybrids remains to be confirmed. In any event, the occurrence of hybrid spawnings in the wild, the apparent lack of intrinsic post-zygotic barriers between species, and the observation of individuals with intermediate color patterns all point to ongoing gene flow among species.

Under a scenario of ongoing gene flow, genetic homogenization is expected throughout the genome, except at regions under strong divergent selection (Wu 2001). In our case, such genomic regions may be linked to color pattern and/or mate choice, the two traits that have been shown to be consistently differentiated between hamlet species so far. Most SNP F_{st} estimates between species were close to zero, but the distribution of F_{st} estimates was long-tailed, with ~1% of the estimates >0.1. About 20% of the SNPs presented a consistent signal of genetic structure between species, and about 10% a phylogenetic signal, two patterns that were not observed in the randomized dataset that

we used as a negative control. Thus, a fraction of the SNPs appear to be differentiated among species. Yet levels of genetic structure were generally low even at these SNPs, which explains why the vast majority of them were not identified as F_{st} outliers. Nonetheless, their large number provided enough resolution to distinguish species. The fact that these SNPs differentiate all species pairs in all populations is notable, since high global F_{st} estimates could in principle be driven by divergence in only one species or population. It is also interesting to note that phylogenetic resolution of species was obtained with the most diverged SNPs, not with the highest number of SNPs as in Lake Victoria cichlids (Wagner *et al*. 2013). The hamlets considered in this study are so closely related that the subtle phylogenetic signal from the most differentiated SNPs appears to be covered by background genetic variation at the other SNPs.

A single repeated outlier maps to HoxCa

A total of 84 F_{st} outliers were identified, representing 0.05% of the SNPs analyzed. This is one order of magnitude less than the 0.71% outliers detected in Lake Victoria cichlid fishes with the same methodology and parameters (Keller *et al*. 2013). Of these, only one was identified as an outlier in more than one population for the same species pair. This SNP was consistently identified as a repeated outlier with more stringent parameters in the F_{st} outlier analysis, across all combinations of parameters tested for the initial assembly of stacks, and with more stringent filtering of the data. Indeed, it is the most differentiated SNP of the entire dataset (F_{st} estimate=0.512) and stands well out in the F_{st} distribution (Figure 1a, highlighted with a red arrow). The next most-diverged SNP has an F_{st} estimate of 0.359, lower than the most differentiated SNP in our randomized dataset (F_{st}=0.394).

Remarkably, the contig containing this SNP maps uniquely to a sequence between the *HoxC10a* and *HoxC11a* genes in ten different teleost species. In addition, another strongly differentiated, non-repeated outlier SNP (F_{st}=0.340) also maps to this region in the stickleback genome. It is tempting to speculate that the *HoxCa* gene cluster plays a role in the color pattern differences that define species. *Hox* genes are well known for their organization in tight genomic

clusters, highly conserved among vertebrates. They code for homeodomain-containing transcription factors and are involved the anterior-posterior patterning of tissues along the body axis during development. They have been shown to play a role in large-scale divergence of homologous structures in animals (reviewed in Carroll *et al*. 2005), but can also be redeployed later in development to play a role in terminal color pattern phenotype. Two examples include the regulation of body pigmentation in *Drosophila* (Jeong *et al*. 2006) and eyespot formation on the wings of Satyrinae butterflies (Saenko *et al*. 2011).

Color pattern is a highly complex trait in teleosts. Structurally, one observes an array of markings—bars, stripes, lines, spots, dots—and extraordinary variation in color, many of which are represented in the *Hypoplectrus* radiation (Puebla 2009). Developmentally, there are no less than five types of pigment cells involved in the composition of color pattern (Braasch *et al*. 2008). Genetically, over 20 pigmentation genes have been identified from mutant screens of the zebrafish and rice fish (Braasch *et al*. 2008), with certainly more to be found. Fewer genes have been shown to be involved in the color patterns observed in East African cichlids (Maan & Sefc 2013), yet these comprise a diverse array of genomic elements, including transcription factors (Roberts *et al*. 2009).

Longer sequences around our outlier SNP, characterized in many individuals, will be needed to confirm the association between this region of the genome and species differences between black and barred hamlets. Longer sequences will also provide the opportunity to determine to what extent high levels of genetic differentiation are due to reduced gene flow (Wu 2001) or low diversity (Cruickshank & Hahn 2014) in this region of the genome. In both cases, fine mapping of the association will help clarify its origin and functional bases.

It is possible that we missed other loci of interest. Assuming a 1Gb genome size typical of many serranids (Gregory 2014), we would expect an average of 2 stacks (one on each side of *Sbf*I restriction sites) every 40 kb or so across the genome. Physical linkage between SNPs and genomic elements of interest could easily erode within this distance. This is especially true if linkage

disequilibrium decays to background levels rapidly, as expected in populations with large effective population sizes (Counterman *et al*. 2010). Moreover, non-repeated outliers may be not all be 'false positives'. They could be implicated in adaptation to the specific conditions of each population, and the genomic bases of reproductive isolation may even differ between populations. The latter hypothesis would be consistent with the genetic structure between populations of *H. nigricans* revealed with the most diverged SNP subset (Figure 1c). It would also be consistent with the stronger population genetic structure observed within this species at microsatellite markers and our hypothesis that *H. nigricans* may have evolved repeatedly from *H. puella* in different populations (Puebla *et al*. 2008).

Concluding remarks

With a single repeated outlier SNP and no clustering or phylogenetic signal when considering all the RAD loci, the hamlets sampled in this study stand at the lower end of the 'speciation continuum' (Seehausen *et al*. 2014), even more closely related than freely hybridizing parapatric races of *Heliconius* butterflies. Such extraordinary low levels of genetic differentiation may be due to a combination of ongoing gene flow and/or vast effective population sizes. The genomic architecture in this situation appears to be characterized by one or a few genomic 'islands' of differentiation against a background 'sea level' of almost no differentiation. Nonetheless, even in this extreme case there are 16,176 SNPs with $F_{st} \geq 0.0094$ (80^{th} percentile) that altogether identify the three species in the clustering analyses, a pattern that we did not observe in the randomized dataset that we used as a control. These are what we refer to as 'genomic atolls', and may represent the early stages of genome hitchhiking (*sensu* Feder *et al*. 2012). More data will be needed to establish to what extent these SNPs are clustered or spread out across the genome.

The possibility that *Hox* genes may be involved in a rapid radiation is intriguing and permits a consistent set of predictions that can be tested with genomic approaches. What is more, the *Hypoplectrus* radiation—including 18 species and as many color patterns—provides the opportunity

to put this genomic hypothesis into a broader evolutionary context, in which specific sources of natural selection on color pattern have been identified (Puebla *et al.* 2007; Puebla 2009), and where color pattern appears to play a strong role in mate choice (Puebla *et al*. 2012). Considering that a significant proportion of global biodiversity dwells in the oceans, astonishingly little is known about the origin of species in the sea. Tractable systems like the hamlets are sorely needed to fill this gap.

Acknowledgements

We thank the Belizean, Honduran, Panamanian and Guna Yala authorities for support with collecting, export, and import permits. This study was funded by a Smithsonian Institution Scholarly Studies grant to O. Puebla, E. Bermingham and W.O. McMillan. We are grateful to Carlos Arias, Paul Etter, Andy Jones, Claudia Rosales, Chris Smith and Megan Supple for their invaluable help, and to Christophe Eizaguirre for comments on an early version of the manuscript.

References

Allendorf FW, Thorgaard GH (1984). Tetraploidy and the evolution of salmonid fishes. In: *Evolutionary genetics of fishes* (ed Turner BJ), pp. 1–53. Plenum Press, New York, NY.

Barreto FS, McCartney MA (2007) Extraordinary AFLP fingerprint similarity despite strong assortative mating between reef fish color morphospecies. *Evolution*, **62**, 226–233.

Bierne N, Welch J, Loire E, Bonhomme F, David P (2011) The coupling hypothesis: why genome scans may fail to map local adaptation genes. *Molecular Ecology*, **20**, 2044–2072.

Braasch I, Volff J-N, Schartl M (2008). The evolution of teleost pigmentation and the fish-specific genome duplication. *Journal of Fish Biology*, **73**, 1891–1918.

Carroll SB, Grenier JK, Weatherbee SD (2005) From DNA to diversity: molecular genetics and the evolution of animal design, 3rd edn. Blackwell Science Press, Malden, MA.

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

Catchen J, Hohenlohe PA, Bassham S, Amores A, Cresko WA (2013) Stacks: an analysis tool set for population genomics. *Molecular Ecology*, **22**, 3124–3140.

Counterman BA, Araujo-Perez F, Hines HM *et al*. (2010) Genomic hotspots for adaptation: the population genetics of Müllerian mimicry in *Heliconius erato*. *PLoS Genetics*, **6**, e1000796.

Cruickshank TE, Hahn MW (2014) Reanalysis suggests that genomic islands of speciation are due to reduced diversity, not reduced gene flow. *Molecular Ecology*, **23**, 3133–3157.

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

Davey JW, Hohenlohe PA, Etter PD, Boone JQ, Catchen JM, Blaxter ML (2011) Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nature Reviews Genetics*, **12**, 499–510.

Del Moral Flores LF, Tello-Musi JL, Martínez-Pérez JA (2011) Descripción de una nueva especie del género *Hypoplectrus* (Actinopterigy: Serranidae) del Sistema Arrecifal Veracruzano, suroeste del Golfo de México. *Revista de Zoología*, **22**, 1–10.

Domeier ML (1994) Speciation in the serranid fish *Hypoplectrus*. *Bulletin of Marine Science*, **54**, 103–141.

Etter PD, Bassham S, Hohenlohe PA, Johnson EA, Cresko W (2011) SNP discovery and genotyping for evolutionary genetics using RAD sequencing. In: *Molecular Methods for Evolutionary Genetics* (eds Orgogozo V, Rockman MV), pp. 157–178. Humana Press, New York, NY.

Evanno G, Regnaut S, Goudet J (2005) Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Molecular Ecology*, 14, 2611–2620.

Falush D, Stephens M, Pritchard JK (2003). Inference of population structure: Extensions to linked loci and correlated allele frequencies. *Genetics*, **164**, 1567–1587.

Feder JL, Egan SP, Nosil P (2012) The genomics of speciation-with-gene-flow. *Trends in Genetics*, **28**, 342–350.

Flicek P, Amode MR, Barrell D *et al*. (2014) Ensembl 2014. *Nucleic Acids Research*, **42**, D749–D755.

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

Grant PR, Grant BR (2014) *40 Years of Evolution: Darwin's Finches on Daphne Major Island*. Princeton University Press, Princeton, NJ.

Gregory TR (2014) Animal Genome Size Database. http://www.genomesize.com.

Hoegg S, Boore JL, Kuehl JV, Meyer A (2007) Comparative phylogenomic analyses of teleost fish Hox gene clusters: lessons from the cichlid fish *Astatotilapia burtoni*. *BMC Genomics*, **8**, 317.

Hohenlohe PA, Bassham S, Etter PD, Stiffler N, Johnson EA, Cresko WA (2010) Population genomics of parallel adaptation in threespine stickleback using sequenced RAD tags. *PLoS Genetics*, **6**, e1000862.

Holt BG, Côté IM, Emerson BC (2011) Searching for speciation genes: molecular evidence for selection associated with colour morphotypes in the Caribbean reef fish genus *Hypoplectrus*. *Plos One*, **6**, e20394.

Jaillon O, Aury JM, Brunet F *et al*. (2004) Genome duplication in the teleost fish *Tetraodon nigroviridis* reveals the early vertebrate proto-karyotype. *Nature*, **431**, 946–957.

Jeong S, Rokas A, Carroll SB (2006) Regulation of body pigmentation by the Abdominal-B Hox protein and its gain and loss in *Drosophila* evolution. *Cell*, **125**, 1387–1399.

Keller I, Wagner CE, Greuter L *et al*. (2013) Population genomic signatures of divergent adaptation, gene flow and hybrid speciation in the rapid radiation of Lake Victoria cichlid fishes. *Molecular Ecology*, **22**, 2848–2863.

Kurosawa G, Takamatsu N, Takahashi M *et al*. (2006) Organization and structure of hox gene loci in medaka genome and comparison with those of pufferfish and zebrafish genomes. *Gene*, **370**, 75–82.

Lee AP, Koh EGL, Tay A, Brenner S & Venkatesh B (2006) Highly conserved syntenic blocks at the vertebrate Hox loci and conserved regulatory elements within and outside Hox gene clusters. *Proceedings of the National Academy of Sciences of the United States of America*, **103**, 6994–6999.

Lewis PO (2001) A likelihood approach to estimating phylogeny from discrete morphological character data. *Systematic Biology*, **50**, 913–925.

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

Lobel PS (2011). A review of the Caribbean hamlets (Serranidae, *Hypoplectrus*) with description of two new species. *Zootaxa*, **3096**, 1–17.

Maan ME, Sefc KM (2013) Colour variation in cichlid fish: developmental mechanisms, selective pressures and evolutionary consequences. *Seminars in Cell & Developmental Biology*, **24**, 516–528.

Martin SH, Dasmahapatra KK, Nadeau NJ *et al*. (2013) Genome-wide evidence for speciation with gene flow in *Heliconius* butterflies. *Genome Research*, **23**, 1817–1828.

McCartney MA, Acevedo J, Heredia C *et al*. (2003) Genetic mosaic in a marine species flock. *Molecular Ecology*, **12**, 2963–2973.

Mungpakdee S, Seo HC, Angotzi AR, Dong X, Akalin A, Chourrout D (2008) Differential evolution of the 13 Atlantic salmon *Hox* clusters. *Molecular Biology and Evolution*, **25**, 1333–1343.

Pérez-Figueroa A, García-Pereira MJ, Saura M, Rolán-Alvarez E, Caballero A (2010) Comparing three different methods to detect selective loci using dominant markers. *Journal of Evolutionary Biology*, **23**, 2267–2276.

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

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

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

Puebla O, Bermingham E, Guichard F (2008) Population genetic analyses of *Hypoplectrus* coral reef fishes provide evidence that local processes are operating during the early stages of marine adaptive radiations. *Molecular Ecology*, **17**, 1405–1415.

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

Roberts RB, Ser JR, Kocher TD (2009) Sexual conflict resolved by invasion of a novel sex determiner in Lake Malawi cichlid fishes. *Science*, **326**, 998–1001.

Rousset F (2008) GENEPOP'007: a complete re-implementation of the GENEPOP software for Windows and Linux. *Molecular Ecology Resources*, **8**, 103–106.

Saenko SV, Marialva MSP, Beldade P (2011) Involvement of the conserved *Hox* gene *Antennapedia* in the development and evolution of a novel trait. *EvoDevo*, **2**:9.

Schluter D (2000) The ecology of adaptive radiation. Oxford University Press, Oxford, UK.

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

Stamatakis A, Hoover P, Rougemont J (2008) A rapid bootstrap algorithm for the RAxML web servers. *Systematic Biology*, **57**, 758–771.

Stamatakis A (2014) RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, **30**, 1312–1313.

Tavera J, Acero AP (2013) Description of a new species of *Hypoplectrus* (Perciformes: Serranidae) from the Southern Gulf of Mexico. *aqua, International Journal of Ichthyology*, **19**, 29–38.

Victor BC (2012) *Hypoplectrus floridae* n. sp. and *Hypoplectrus ecosur* n. sp., two new Barred Hamlets from the Gulf of Mexico (Pisces: Serranidae): more than 3% different in COI mtDNA

sequence from the Caribbean *Hypoplectrus* species flock. *Journal of the Ocean Science Foundation*, **5**, 1–19.

Vilas A, Pérez-Figueroa A, Caballero A (2012) A simulation study on the performance of differentiation based methods to detect selected loci using linked neutral markers. *Journal of Evolutionary Biology*, **25**, 1364–1376.

Wagner CE, Harmon LJ, Seehausen O (2012) Ecological opportunity and sexual selection together predict adaptive radiation. *Nature*, **487**, 366–369.

Wagner CE, Keller I, Wittwer S *et al*. (2013) Genome-wide RAD sequence data provide unprecedented resolution of species boundaries and relationships in the Lake Victoria cichlid adaptive radiation. *Molecular Ecology*, **22**, 787–798.

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

Whiteman EA, Côté IM, Reynolds JD (2007) Ecological differences between hamlet (*Hypoplectrus*: Serranidae) colour morphs: between-morph variation in diet. *Journal of Fish Biology*, **71**, 235–244.

Whiteman EA, Gage MJG (2007). No barriers to fertilization between sympatric colour morphs in the marine species flock *Hypoplectrus* (Serranidae). *Journal of Zoology*, **272**, 305–310.

Wu CI (2001) The genic view of the process of speciation. *Journal of Evolutionary Biology*, **14**, 851–865.

Zerbino DR & Birney E (2008). Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Research*, **18**, 821–829.

Data accessibility

Raw data for all samples, transect surveys and spawning observations: provided as Supporting information. Raw microsatellite data, sequence data, stacks consensus sequences, SNP genotype

calls, genepop input file, structure input files (all data and SNP subsets), tree files (all data and above 90th percentile) and mini-contig sequences: Dryad doi:10.5061/dryad.nv1f0.

Author contributions

O. Puebla, W.O. McMillan and E. Bermingham contributed to the sampling, transect surveys and spawning observations. O. Puebla wrote the Scholarly Studies grant for this study with input from W.O. McMillan. O. Puebla designed the study, performed the lab work, did the data analysis, and wrote the manuscript with input from W.O. McMillan and E. Bermingham.

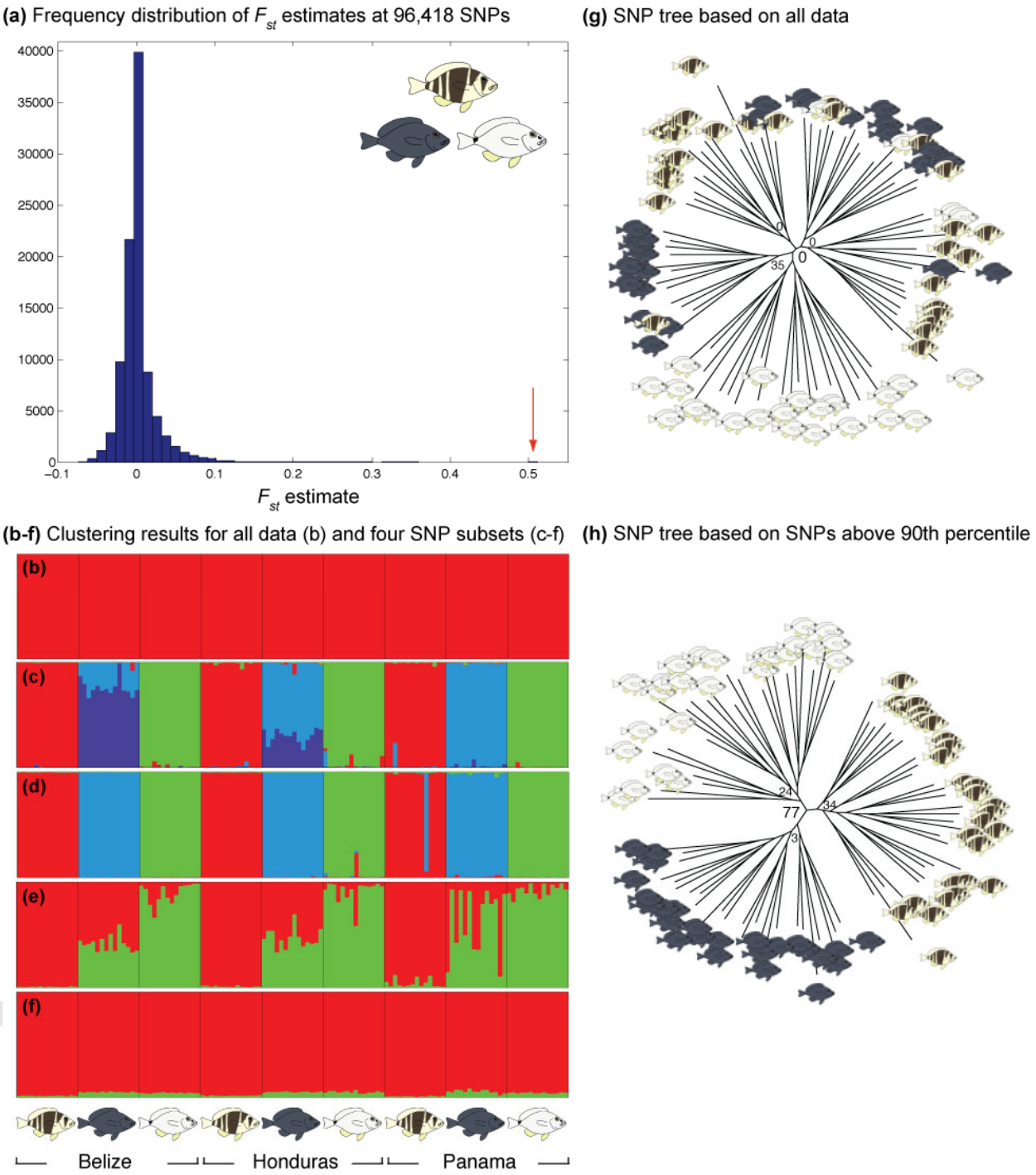

Figure 1. (a) Distribution of global F_{st} estimates between *H. puella*, *H. nigricans* and *H. unicolor* from Belize, Honduras and Panama at 96,418 SNPs from 52,459 loci. Red arrow: the single repeated outlier SNP between sympatric species identified in this study. (b) Clustering results for the entire dataset (1 SNP per locus, 41,690 SNPs). (c) For the SNP subset above the 90th percentile (F_{st}≥0.0243,

1 SNP per locus, 7,841 SNPs). (d) For the SNP subset between the 80th and 90th percentile ($0.0094 \leq F_{st} < 0.0243$, 1 SNP per locus, 8,335 SNPs). (e) For the SNP subset between the 70th and 80th percentile ($0.0027 \leq F_{st} < 0.0094$, 1 SNP per locus, 8,184 SNPs). (f) For the SNP subset between the 60th and 70th percentile ($0.0006 \leq F_{st} < 0.0027$, 1 SNP per locus, 8,036 SNPs). (g) Maximum-likelihood SNP tree based on all the data. (h) Maximum-likelihood SNP tree for the SNP subset above the 90th percentile. Bootstrap values within groups not shown.

	H. puella (*n*=42)	*H. nigricans* (*n*=42)	*H. unicolor* (*n*=42)
Number of sites	5,025,161	5,007,529	4,354,085
Number of polymorphic sites	66,779	63,575	54,219
Proportion of polymorphic sites (%)	1.329	1.270	1.245
Mean *n* per site	28.5	26.0	27.8
Nucleotide diversity (π)	0.00242	0.00241	0.00233
Expected heterozygosity	0.125	0.124	0.118

Table 1. Number of sites, number and proportion of polymorphic sites, mean number of individuals sampled per site per locus, nucleotide diversity (π) and expected heterozygosity of the three species considered in this study.

Species 1	Species 2	Location	F_{st} estimate (sample size)		
			10 μsatellite loci	*96,418 SNPs*	*1 outlier SNP*
All species		All locations	0.0022 (*n*=418)	0.0038 (mean *n*=78.5)	0.5119 (*n*=107)
H. nigricans	*H. puella*	All locations	0.0029 (*n*=310)	0.0040 (mean *n*=53.4)	0.7222 (*n*=70)
H. nigricans	*H. unicolor*	All locations	0.0035 (*n*=264)	0.0057 (mean *n*=50.5)	0.1858 (*n*=69)
H. puella	*H. unicolor*	All locations	-0.0002 (*n*=262)	0.0025 (mean *n*=53.0)	0.3843 (*n*=75)
H. nigricans	*H. puella*	Belize	0.0133 (*n*=101)	0.0152 (mean *n*=24.8)	0.7506 (*n*=26)
H. nigricans	*H. unicolor*	Belize	-	0.0166 (mean *n*=22.0)	0.3940 (*n*=24)
H. puella	*H. unicolor*	Belize	-	0.0064 (mean *n*=22.7)	0.2709 (*n*=26)
H. nigricans	*H. puella*	Honduras	0.0005 (*n*=101)	0.0028 (mean *n*=26.6)	0.7131 (*n*=28)
H. nigricans	*H. unicolor*	Honduras	0.0040 (*n*=105)	0.0135 (mean *n*=19.4)	0.5843 (*n*=25)
H. puella	*H. unicolor*	Honduras	-0.0016 (*n*=104)	0.0118 (mean *n*=19.5)	-0.0120 (*n*=25)
H. nigricans	*H. puella*	Panama	0.0039 (*n*=108)	0.0048 (mean *n*=11.7)	0.3783 (*n*=16)*
H. nigricans	*H. unicolor*	Panama	0.0011 (*n*=108)	0.0196 (mean *n*=17.5)	-0.0354 (*n*=20)*
H. puella	*H. unicolor*	Panama	0.0008 (*n*=108)	0.0112 (mean *n*=17.9)	0.4116 (*n*=24)*

Table 2. F_{st} estimates between sympatric *H. puella*, *H. nigricans*, and *H. unicolor* from Belize, Honduras and Panama at 10 microsatellite loci, 96,418 SNPs, and at the repeated outlier identified in this study. *Low coverage (mean=15x for these three pairs).

Species 1	Species 2	Location	N. loci	N. outliers / ratio (%)					
				Prior odds=10		*Prior odds=100*		*Prior odds=1*	
H. nigricans	*H. puella*	Belize	35584	25	0.07	5	0.01	341	0.96
H. nigricans	*H. unicolor*	Belize	28205	24	0.09	4	0.01	271	0.96
H. puella	*H. unicolor*	Belize	28857	13	0.05	4	0.01	135	0.47
H. nigricans	*H. puella*	Honduras	38886	9	0.02	2	0.01	137	0.35
H. nigricans	*H. unicolor*	Honduras	16932	6	0.04	0	<0.01	69	0.41
H. puella	*H. unicolor*	Honduras	17025	7	0.04	1	0.01	52	0.31
H. nigricans	*H. puella*	Panama	1315	0	<0.08	0	<0.08	0	<0.08
H. nigricans	*H. unicolor*	Panama	2089	0	<0.05	0	<0.05	0	<0.05
H. puella	*H. unicolor*	Panama	10137	0	<0.01	0	<0.01	0	<0.01
		Total	178830	84	0.05	16	0.01	1005	0.56
		Repeated outliers		1		1		18	

Table 3. F_{st} outlier analyses. The same repeated outlier SNP was identified with the prior odds for the neutral model set to 10 and 100. Most outliers detected with the prior odds set to 1 are expected to be false positives given the relaxed prior odds for the neutral model in this case, which imply a prior belief that the model with selection is as likely as the neutral model at any locus.

Species	Region	Distance from *HoxC11a* (bp)	Identity (%)	Alignment length (bp)	E-value	Reference
Astatotilapia burtoni (East African cichlid)	*HoxC10a* - *HoxC11a*	2220	85	323	1E-85	Hoegg *et al.* (2007)
Oreochromis niloticus (Nile tilapia)	*HoxC10a* - *HoxC11a*	2335	80	423	1E-83	-
Oryzias latipes (Japanese rice fish)	*HoxC10a* - *HoxC11a*	1864	73	377	6E-54	Kurosawa *et al.* (2006)
Gasterosteus aculeatus (three-spined stickleback)	*HoxC10a* - *HoxC11a*	2594	89	409	1E-45	-
Xiphophorus maculatus (Southern platyfish)	*HoxC10a* - *HoxC11a*	1665	77	235	2E-40	-
Takifugu rubripes (Japanese pufferfish)	*HoxC11a*	-	77	219	5E-38	Lee *et al.* (2008)
Tetraodon nigroviridis (spotted green pufferfish)	*HoxC10a* - *HoxC11a*	2333	75	213	3E-31	-
Salmo salar (Atlantic salmon)	*HoxC10aβ* - *HoxC11aβ*	2466	74	182	6E-18	Mungpakdee *et al.* (2008)
Salmo salar (Atlantic salmon)	*HoxC10aα* - *HoxC11aα*	2602	74	189	2E-18	Mungpakdee *et al.* (2008)
Astyanax mexicanus (Mexican tetra)	*HoxC11a*	-	78	83	5E-07	-
Danio rerio (zebrafish)	*HoxC10a* - *HoxC11a*	1691	79	75	2E-06	-

Table 4. Result of the blast searches for the consensus sequence of the mini-contig containing the repeated outlier SNP identified in this study.