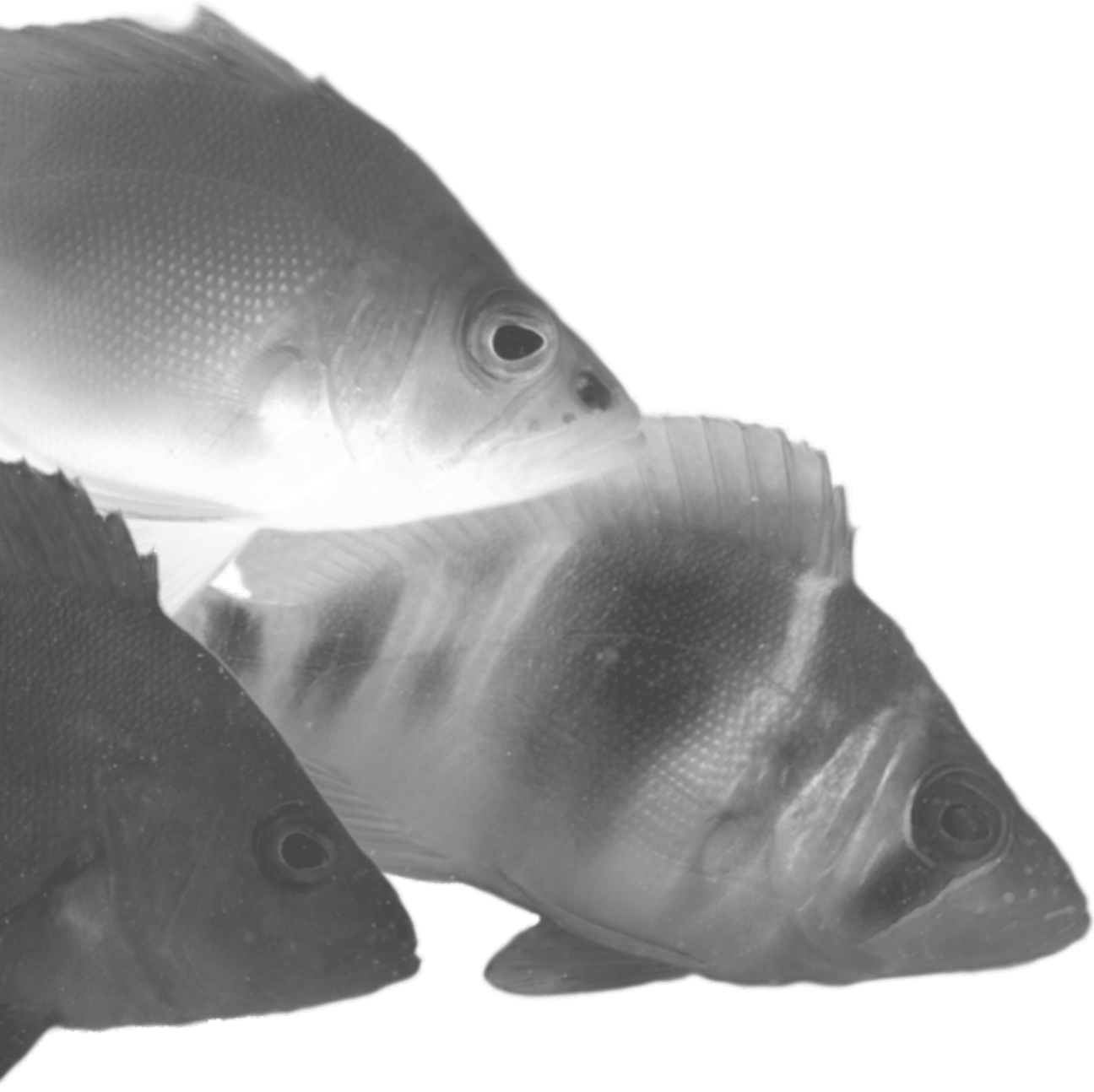
Genomics and the Origin of Marine Species

Kosmas Hench



Genomics and the Origin of Marine Species

Dissertation zur Erlangung des akademischen Grades

Doctor rerum naturalium

der Mathematisch-Naturwissenschaftlichen Fakultät
der Christian-Albrechts-Universität zu Kiel

vorgelegt von
Kosmas Hench

Kiel, 26.07.2020

Erster Gutachter: Prof. Oscar Puabla
Zweite Gutachterin: Prof. Tal Dagan
Tag der Disputation: 23.07.2020
Zum Druck genehmigt: 23.07.2020

Contents

Summary 3
Zusammenfassung 5

1 Introduction **7**
1.1 Speciation, Gene Flow and Evolutionary Radiations 7
1.2 Genomics and Speciation 10
1.3 Evolution within the Ocean 12
1.4 Hamlets 14
1.5 Thesis Outline and Objective 17
Intro References 18

2 Temporal Changes in Hamlet Communities (*Hypoplectrus* spp., Serranidae) over 17 Years **23**
Kosmas Hench, W. Owen McMillan, Ricardo Betancur-R., Oscar Puebla
Original publication 23
2.1 Introduction 24
2.2 Materials and Methods 25
2.3 Results 27
2.4 Discussion 30
Chapter 2 References 35

3 Inter-chromosomal Coupling between Vision and Pigmentation Genes during Genomic Divergence **39**
Kosmas Hench, Marta Vargas, Marc P. Höppner, W. Owen McMillan, Oscar Puebla
Original publication 39
3.1 Introduction 40
3.2 Results and Discussion 42
3.3 Methods 62
Chapter 3 References 77

4 The Evolution of Microendemism in a Reef Fish *(Hypoplectrus maya)* **87**
Benjamin M. Moran, **Kosmas Hench**, Robin S. Waples, Marc P. Höppner, Carole C. Baldwin, W. Owen McMillan, Oscar Puebla
Original publication 87
4.1 Introduction 88
4.2 Methods 91

4.3 Results 96
4.4 Discussion 104
Chapter 4 References 115

5 The Genomic Origins of a Marine Radiation **123**
Kosmas Hench, W. Owen McMillan, Oscar Puebla
5.1 Results and Discussion 124
5.2 Methods 140
Chapter 5 References 146

6 Synthesis and Perspective **151**
6.1 Synthesis 151
6.2 Perspective 154
6.3 Concluding Remarks 156
Synthesis References 156

Appendix **159**
List of Figures 159
List of Supplementary Figures 159
List of Tables 161
List of Supplementary Tables 161
Declaration of Author Contributions 163
Curriculum Vitae 164
Acknowledgments 166
Eidesstattliche Erklärung 168

Summary

The great biological *mystery of mysteries* as coined by John Herschel and adopted by Charles Darwin, is the origin of new species that fuels the stunning biodiversity witnessed around the globe today. The last decades have seen a rapid development in the field of DNA sequencing that today provides an exciting new tool-set to probe this mystery. The investigation of the genetic signatures underlying speciation processes has given rise to the biological field of speciation genomics. Yet, while the approaches unlocked by our ability to sequence whole genomes of large numbers of samples have vastly improved our understanding of the speciation process, most of these discoveries have come from studies on terrestrial or limnic organisms. For historical reasons, contributions from marine systems are scarce regardless of their long evolutionary history and importance.

The work within this doctoral thesis introduces the Caribbean reef fish genus *Hypoplectrus* (hamlets) into the field of speciation genomics. This marine species flock of eighteen distinct coral reef fish species presents an exciting model system to study the speciation process *in action*. There is a rich scientific background for the hamlets, from which we know that the different hamlet species mate assortatively with respect to their species specific bright color patterns, yet they are only very shallowly differentiated genetically. Also, there is no known ecological divergence within this evolutionary radiation, posing the question if speciation in hamlets includes adaptive aspects. The overarching theme within this thesis is the investigation of the underlying evolutionary drivers that are acting at the origin of this marine radiation and facilitate rapid speciation within the ocean. Distributed over four separate manuscripts, this work addresses several aspects impacting the dynamics of the *Hypoplectrus* radiation. Within the first manuscript, the temporal stability of the hamlet community in a patch of reefs in Puerto Rico is investigated. The findings indicate that the hamlet community composition is dynamic and potentially impacted by ecological factors such as turbidity or the presence of specific coral species. This implicates that in fact hamlet species might differ ecologically, which possibly provides an angle for natural selection to work within the radiation. Within the second manuscript the hamlet reference genome is introduced and whole genome resequencing is applied to investigate the signals of speciation within three of the most common hamlet species. The results show that, against a genome wide background of very low differentiation, a small number of color pattern and vision genes are highly differentiated between species and apparently co-selected for. The third manuscript explores the demographic history of a rare endemic hamlet species. It uses a coalescent approach to show the decline in population size of this particular species since the recent evolutionary split from the remaining genus. In the last manuscript, nine different hamlet species are sequenced to provide a cross section through the hamlet radiation. The results of population and phylogenomics indicate ongoing inter-species gene flow throughout the majority of the genome with only a small set of putative barrier genes. Phylogenetic relationships through most of the genome are diffuse, yet the signal within the few differentiated genomic intervals is discordant, pointing to introgession

events or differential lineage sorting at those major effect loci.

In sum, the emerging picture is that of a very young radiation, which is driven by a modular system of co-selected vision and color pattern genes that is maintained by a highly assortative mating system. Yet, imperfection in assortativenes allows for occasional hybridization, which facilitates gene flow and can promote diversity by shuffling the modules characterizing the hamlet species. The rearrangement of the genetic basis underlying co-selected discrete phenotypic traits through recombination can quickly generate a large variety in hamlet phenotypes. The hamlet radiation thus seems to employ evolutionary mechanisms that are also known from other systems like the *Heliconius* butterflies, despite the fundamentally different preconditions prevailing in marine versus terrestrial habitats.

Zusammenfassung

Das große biologische *Geheimnis der Geheimnisse*, wie es von John Herschel geprägt und von Charles Darwin übernommen wurde, ist die Entstehung neuer Arten, welche die faszinierende Artenvielfalt rund um den Globus antreibt. In den letzten Jahrzehnten hat sich auf dem Gebiet der DNS-Sequenzierung eine rasante Entwicklung vollzogen, welche uns heute spannende Werkzeuge an die Hand gibt, um dieses Geheimnis zu erkunden. Die Untersuchung der genetischen Siganturen des Artbildungs-Prozesses hat die biologische Disziplin der Speziations-Genomik hervorgebracht. Diese neuen Verfahren erlauben uns die gesamten Genome von einer goßen Anzahl an Proben zu sequenzieren, was zu beachtlichen Fortschritten in der Evolutionsbiologie geführt hat. Allerdings sind diese Entwicklungen größten Teils auf terrestrische und limnische Systeme beschränkt. Aufgrund historischer Entwicklung des Feldes sind Beiträge zu marinen Systemen immernoch selten — trotz ihrer langen evolutionären Geschichte und Relevanz.

Die in dieser Doktorarbeit vorgestellte Arbeit beinhaltet die Einführung der Karibischen Riff-Fisch Gattung *Hypoplectrus* (der Hamlets) in das Feld der Speziations-Genomik. Diese aus achtzehn Unterarten bestehende marine Sammelart von Korallen-Riff-Fischen stellt ein spannendes Model-System für die Erforschung des Artbildungsprozesses *in Aktion* dar. Dank des umfangreichen wissenschaftlichen Hintergrundes über Hamlets wissen wir, dass diese sich bevorzugt innerhalb ihreres arttypischen Farbmusters verpaaren, wobei nur eine sehr geringe genetische Differenzierung zwischen Arten vorliegt. Des Weiteren sind keine nennenswerten ökologischen Unterschiede innerhalb dieser evolutionären Radiation bekannt, was die Frage aufwirft inwiefern die Artbildung innerhalb der Hamlets adaptive Züge aufweist. Die übergeordnete Thematik der vorliegenden Arbeit umfasst die zugrunde liegenden evolutionären Mechanismen, welche dieser marinen Radiation zugrunde liegen und eine schnelle Artbildung im Ozean ermöglichen. Verteilt über vier separate Manuskripte werden verschiedene Aspekte der evolutionären Dynamik der *Hypoplectrus* Radiation beleuchtet. Das erste Manuskript behandelt die zeitliche Stabilität der Hamlet-Gemeinschaft innerhalb einer Gruppe von Korallenriffe Puerto Ricos. Die Ergebnisse deuten darauf hin, dass die Artzusammensetzung dynamisch ist und potentiell auf Veränderungen von ökologischen Faktoren wie Wassertrübung und die Anwesenheit bestimmeter Korallenarten reagiert. Dies impliziert ökologische Unterschiede zwischen den einzelnen Hamlet Arten, was einen Angriffspunkt für natürlich Selektion innerhalb der Radiation bieten könnte. Im zweiten Manuskript wird das neu assemblierte Hamelt Referenz-Genom vorgestellt. Mit Hilfe von *whole genome resequencing* werden die Signale der Artbildung innerhalb von drei weitverbreiteten Hamlet Arten untersucht. Die Ergebnisse zeigen einen genom-weiten niedrigen genetischen Differenzierungs-Hintergrund gegen den sich einige wenige stark ausgeprägte Differenzierungs-Spitzen abzeichnen. Diese scheienen sich auf Farbgebungs- und Seh-Vermögens-Gene zu knozentrieren, welche scheinbar co-selektiert werden. Das dritte Manuskript befasst sich mit der demographischen Geschichte einer seltenen endemischen Hamlet Art und nutzt *coalecsent*-basierte Verfahren, um den

Rückgang der effektiven Populationsgröße seit kürzlichen evolutionären der Trennung vom restlichen Genus zu beschreiben. Im letzten Manuskript wird anhand von neun unterschiedlichen Hamlet Arten ein Querschnitt der Radiation sequenziert. Die populations- und phylo-genetischen Ergebnisse deuten auf einen genomweit ausgeprägten Inter-Art-Genfluss hin, der nur an wenigen Barriere-Genen unterbrochen scheint. Die phylogenetischen Beziehungen sind über weite Teile des Genoms diffus, an den wenigen genomischen Regionen mit erhöhter genetischer Differenzierung wird das Signal zwar ausgeprägter, allerdings sind die Beziehungen innerhalb dieser Regionen wiedersprüchlich. Dies deutet auf Introgression oder *incomplete lineage sorting* an diesen einflussreichen Loci hin.

Insgesamt entsteht das Bild einer sehr jungen evolutionären Radiation, welche von einem modularen System von co-selektierten Farbgebungs- und Seh-Vermögens-Genen angetrieben wird. Dieses wird aufrecht erhalten von farb-speziefischer Partnerwahl, welche Hybridisierung zwischen unterschiedlichen Arten aber nicht vollständig ausschließt. Dies erlaubt Gefluss zwischen den Arten, was durch Rekombination zum Vermischen der diskreten art-spezifischen Musterungselemente führen kann. Das Umorganisieren der genetischen Basis hinter diskreten co-selektierten Eigenschaften kann auf diese Weise schnell zu einer Steigerung der phenotypischen Diveristät innerhalb der Hamlets führen. Innerhalb der Hamlet-Radiation scheinen also vergleichbare evolutionäre Mechanismen zu greifen wie beispielsweise bei *Heliconius*-Schmetterligen — ungeachtet der fundamental unterschiedlichen Vorraussetzungen durch die verschiedenen Lebensräume.

– 1 –
Introduction

"In considering the Origin of Species, it is quite conceivable that a naturalist, reflecting on the mutual affinities of organic beings, on their [...] geographical distribution, geological succession, and other such facts, might come to the conclusion that each species had not been independently created, but had descended, like varieties, from other species. Nevertheless, such a conclusion, even if well founded, would be unsatisfactory, until it could be shown how the innumerable species inhabiting this world have been modified, so as to acquire that perfection of structure and coadaptation which most justly excites our admiration."

(Darwin, 1859)

1.1. Speciation, Gene Flow and Evolutionary Radiations

The existence of the huge variety of lifeforms has fascinated biologists for centuries. How this diversity originates and which forces act to maintain the plurality of life is central to evolutionary biology and tightly linked to the processes of extinction and speciation. While the concept of evolution by natural selection was established by Darwin (1859), the term *speciation* describing the formation of separate independent evolutionary units from an ancestral lineage was introduced only later by Cook (1906). Yet, how exactly to define these evolutionary units representing species proved more complicated than one might have anticipated. The biological species concept by Mayr (1942) and Dobzhansky (1937) is a long held model that defines species as interbreeding populations which produce fertile offspring and are reproductively isolated from any organism outside this set of populations. Even though the dogmatic view of complete reproductive isolation between species has now been put into perspective (Mallet,

2001), discontinuities in *gene flow* between populations are still regarded as key to species delineation. A timely definition of speciation is given by Seehausen *et al.* (2014) who describe speciation as the "origin of reproductive barriers among populations that permit the maintenance of genetic and phenotypic distinctiveness of these populations in geographical proximity".

Under this concept, gene flow and speciation are antagonistic processes where a reduction of gene flow for a prolonged period of time allows for the accumulation of enough divergence between two populations to form individual species (Figure 1.1 a). At later stages, the relationship between gene flow and speciation might flip though: Hybridization between two closely related species not only jeopardizes the biological species concept, but it also poses potential for the creation of a new species of hybrid origin (Kronforst *et al.*, 2013). A prerequisite for hybrid speciation are at least two populations that are diverged enough to be recognized as separate multi-locus 'genotypic clusters', but that are not yet reproductively isolated (Mallet, 2007). Hybridization of these genotypic clusters shuffles elements of the co-adapted genotypes, thus boosting genetic variation. This enables the instant creation of new genotypes which may have access to novel ecological niches and thus thrive as independent new species (Figure 1.1 b).

Another aspect of hybridization also contradicting the biological species concept is that it allows for *introgression*. Here, hybrids of two diverged populations are not forming an independent population, but backcross into one of their parental lineages. This allows for gene flow between distinct populations, where alleles from one population move into the other without eroding the defining differences between the two populations (Harrison and Larson, 2014). In this case, the hybrids merely present a vector of alleles from one parental lineage into the other (Figure 1.1 c).

The changing role of hybridization over time highlights the fact that speciation is not an in-

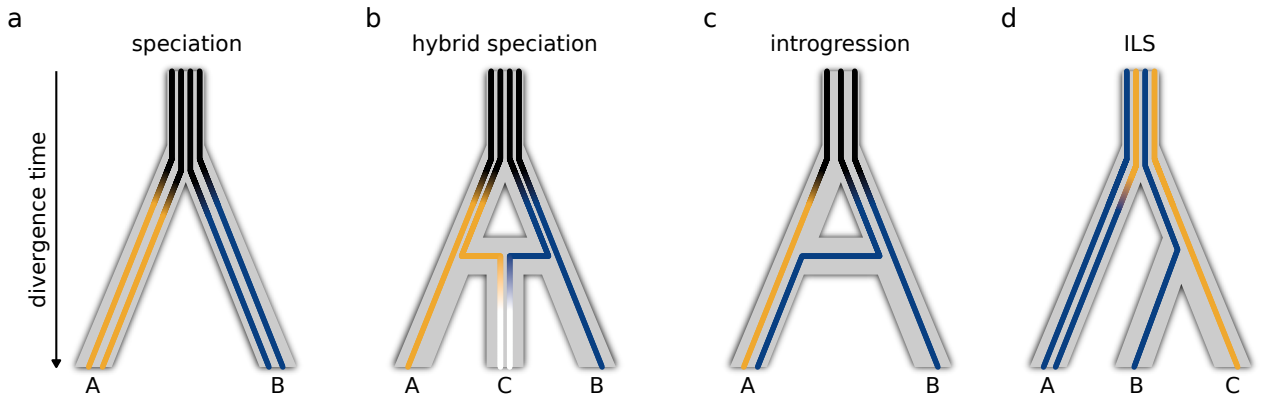


Figure 1.1: Models of speciation and gene flow. **a**, the null model of speciation, the two lineages A and B start to diverge soon after the interruption of gene flow. **b**, hybridization of two recently diverged lineages A and B gives rise to a third independent lineage C. **c**, hybridization between the recently diverged lineages A and B and subsequent backcrossing into lineage A introduces alleles originating from lineage B into lineage A. **d**, in the case of incomplete lineage sorting (ILS), an ancient polymorphism is inherited in both sister lineages of the first speciation process. The polymorphism is maintained within the common ancestor of the lineages B and C and fixes differentially after the second speciation process.

stant event, but rather a continuous process. The extent of gene flow is reduced gradually and during this time window of separation, different processes might interact. The gradient of differentiation between sister lineages is referred to as the *speciation continuum* (Shaw and Mullen, 2014), indicating that it is hard to pinpoint the exact point in time when two lineages are separating.

The expected divergence during the process of speciation describes the accumulation of different alleles between the splitting lineages. This might be caused by chance, due to new mutations that only effect one of the diverging populations given the restricted gene flow. Alternatively, shared standing variation that is inherited from the common ancestor might evolve differently in the sister species — either neutrally (through drift), or directed (through differential selection). While this sorting of alleles can be swift in the case of differential selection, the duration of neutral processes (mutation and drift) require more time depending on the effective population sizes and mutation rates of the young species. As long as species share a polymorphism that predates their species split, this might resemble an introgression scenario despite the actual absence of gene flow (Guerrero and Hahn, 2017; Cruickshank and Hahn, 2014). If further speciation events occur while the polymorphism is maintained, this can lead to *incomplete lineage sorting* (ILS, Figure 1.1 d). This implies that later, after fixation of the originally polymorphic loci, the genomic regions affected by ILS might not coalesce according to the species tree. As ILS is dependent on standing variation, it is most likely to occur if the duration between successive speciation events is short, since this decreases the chances of intermediate fixation of the variant loci.

In the case of an *evolutionary radiation*, a lineage speciates multiple times in quick succession, which leads to a rapid diversification of the clade (Simões *et al.*, 2016). It is believed that radiations are a main contributor to biodiversity with such prominent examples as the Cambrian explosion (Marshall, 2006), the diversification of vertebrates (Alfaro *et al.*, 2009) or the adaptive radiation of the east African cichlids (Malinsky *et al.*, 2018). They are characterized by the great number of descendant lineages and the short waiting time between the independent speciations. It is therefore likely that radiations are facilitated in situations where high diversity in an ancestral population can be utilized to circumvent the dependency on gradual divergence through new mutations for successive speciations (Marques *et al.*, 2019). Thus, hybridization can aid evolutionary radiation by providing an excess in diversity. But, rapid speed also implies that there might be ample opportunities for interspecies gene flo in the early stage of the radiations, which also increases the chance of introducing ILS into the radiating lineage (Mallet *et al.*, 2016).

The best studied cases of evolutionary radiations are *adaptive radiations* (Simões *et al.*, 2016). These are radiations where the rapid speciations are triggered by ecological opportunity such as sudden access to open niche spaces through the evolution of novel key traits (Stroud and Losos, 2016). The diversification of the radiating lineages is classically realized through adaptation of the emerging new species to divergent ecological niches and niche partitioning with an increasing degree of specialization of the individual species. The

evolutionary driver behind this process is natural selection, which implies that the emerging sister species are ecologically distinct and thus offer different points of attack for natural selection (Schluter, 2000). In contrast to this, in sexual radiations, sexual selection is the driving evolutionary force. Sexual selection is not constrained by available ecological niches, as is does not imply any ecological divergence between the emerging species (Martin and Richards, 2019). Finally, natural and sexual selection can interact during a radiation (Wagner *et al.*, 2012). If both act on the same trait, this is considered a *magic trait* — this is for example believed to be the case for the *optix* gene in the *Heliconius* radiation (Merrill *et al.*, 2019). In this case, the genes effect on wing pattern has both ecological and reproductive implications through its effect on mimicry and assortative mating.

Summary

- *Speciation* is the origin of reproductive barriers between populations.
- *Gene flow* is opposing speciation in early stages but can lead to the formation of *hybrid species* or *introgression* once populations are diverged.
- Speciation usually represents a process rather than an instant event, thus leading to a *speciation continuum*.
- If an ancestral polymorphism persists throughout several speciation processes, it can lead to *incomplete lineage sorting* (ILS).
- Evolutionary *radiation* describes a quick succession of speciations, leading to a rapid diversification within one lineage.
- Adaptive radiations are driven by natural selection, sexual radiations by sexual selection. If both act together, this can lead to the evolution of a *magic trait*.

1.2. Genomics and Speciation

Focusing more on the evolutionary mechanisms that provide variation as the basis for selection to work upon, the field of *genetics* revolves around the question how traits are inherited from one generation to the next. Since its early origins, posed by Mendel's pea experiments (Mendel, 1866), the field has served as backbone for the development of evolutionary theory. Especially in the early twentieth century, Fisher, Wright and Haldane formulated an extensive statistical background for population genetics (Thompson, 1990). The discovery of deoxyribonucleic acid (DNA) as carrier of genetic information (Watson and Crick, 1953) and the subsequent ability to sequence genes opened up the field to more direct acquisition of genetic data (Reid and Ross, 2011; Thompson, 1990). A massive drop in sequencing costs, roughly since the turn of the millennium, caused by the advent of *next generation sequencing* (NGS) led to an enormous global up-scaling in sequencing capacity (Metzker, 2010). Coupled with the increasing availability of computation power, this facilitated a shift in the focus from individual genes to the sequencing of whole genomes of large sample numbers per population, thus promoting a transition from genetics to *genomics*. Spearheaded by the human genome project led by Craig Venter (Lander *et al.*, 2001) and medical model organ-

isms like *Drosophila* (Adams *et al.*, 2000) and mice (Chinwalla *et al.*, 2002), genome assembly and resequencing soon also became feasible for classical systems in speciation research like Darwin's finches (Lamichhaney *et al.*, 2015) or the African cichlids (Hulsey and Renn, 2009).

The implications of the shift from classical genetics to genomics for speciation research are severalfold: Maybe the most immediate consequence is a massive expanse in scope (da Fonseca *et al.*, 2016). While genetic methods usually consider low tens (microsatellites) or hundreds Single-nucleotide polymorphism of markers (SNPs), reduced representation sequencing (eg. RAD) can produce tens of thousands and whole genome sequencing millions of SNPs. Reduced representation sequencing presents a kind of bridging technology, but is considered *genomics* here — if applied in concert with a reference genome.

In contrast to microsatellites, which are very diverse markers and thus effective for statistical inference, the sheer quantity of markers renders genomic approaches statistically very powerful. It also provides the option to split the data set for example to compare neutral markers with those under selection.

A second benefit of genomic approaches is their high resolution due to the high marker density across the genome. This is also the area where, naturally, whole genome resequencing clearly outperforms reduced representation sequencing. Again, the resolution of genomic studies is not unprecedented — classical Sanger sequencing can provide the same detail, but in genomic studies this density of markers is provided across the entire genome.

Besides those quantitative benefits of genomics, there is also a qualitative advantage when a reference genome of the study organism is available. When mapping sequences to a genome, the spatial context of population genetic statistics is revealed (Ekblom and Wolf, 2014), which allows to study the entire *genomic architecture* of evolutionary processes. In speciation research, this opens up new possibilities, like screening for adaptive loci or investigating local barriers to gene flow (Steiner *et al.*, 2013). Especially when it is possible to phase genotypes to acquire haplotype information, genomics can open access to more complex methods to investigate for example the effect of linkage and recombination on introgression and hybridization or the demographic history of populations.

The power of genomic approaches lies in the combination of these three aspects (resolution/ SNP density, genome wide extent, spatial context) and in particular whole genome resequencing provides flexible data basis enabeling the investigatigation of diverse evolutionary questions (da Fonseca *et al.*, 2016).

Besides the aforementioned Darwin's finches and African chichlids, genomics studies have been applied in various terrestrial or fresh water systems to understand evolutionary processes underlying speciation. Early systems with available reference genomes and thus access to genomic methods include for example *Heliconius* butterflies (Dasmahapatra *et al.*, 2012), the three-spined stickleback (*Gasterosteus aculeatus*, Roesti *et al.* 2013) or sunflowers (*Helianthus*, Kane *et al.* 2011).

Summary

- *Genetics* is the scientific field studying the creation and inheritance of genetic variation.
- When switching from focal genes to the entire genome, genetics transforms into *genomics*.
- The strengths of genomics lie in combining high spatial resolution with maximizing the extent to genome wide coverage.
- Inference about the influence of the genomic architecture on evolutionary processes is often only possible using genomic methods.

1.3. Evolution within the Ocean

When looking at the models in evolutionary biology discussed so far, one particular bias becomes obvious: the lack of marine organisms. This is striking considering both the enormous spatial extent as well as the long evolutionary history within the marine habitat (Labandeira, 2005). Given that evolution is believed to have started and (except for microbes) long remained exclusively within the ocean and that the ocean covers the majority of the globe, the evolutionary dynamics within this habitat are surely of relevance when studying the underlying dynamics of speciation. Since particularly the conditions for gene flow differ between terrestrial and marine systems, marine study systems are needed for a comprehensive understanding of evolution.

One major structural difference between marine and terrestrial (as well as limnic) habitats is the apparent homogeneity of the ocean compared to the more fragmented terrestrial habitats (Palumbi, 1994). Although ocean currents and differences in thermohaline conditions can act as soft barriers, the world oceans in principle present one huge single environment. Coupled with the often complex life histories of marine organisms, which

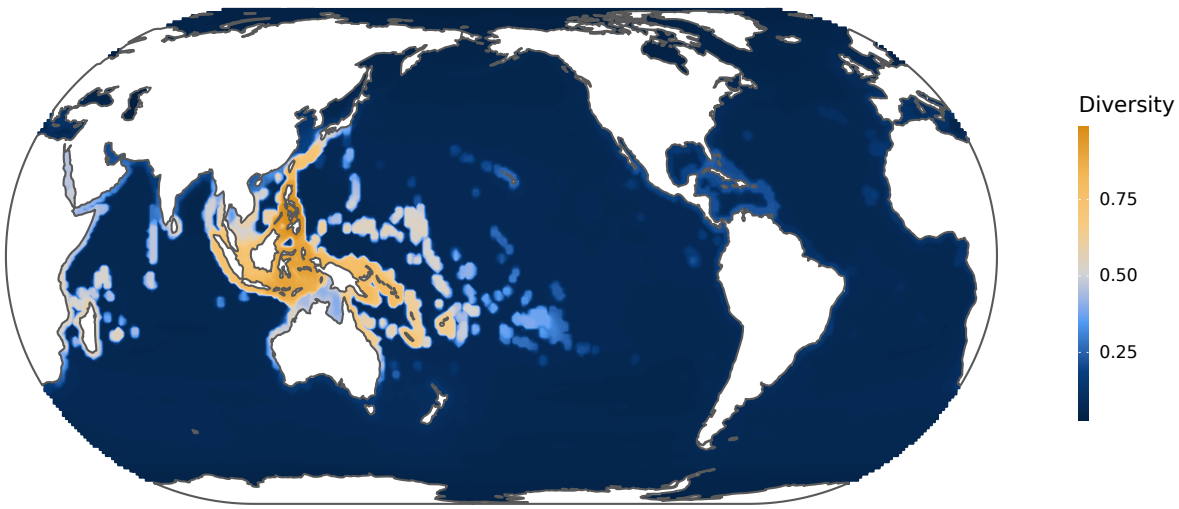


Figure 1.2: Global distribution of marine biodiversity. Diversity patterns are modeled based on distributions of 4352 species from nine broad taxa (plants, fish, echinoderms, crustaceans, cnidarians, molluscs, mammals, reptiles and birds). Data from Jenkins and Van Houtan 2016.

often include a pelagic larval phase, this allows for huge species ranges and generally high connectivity within the ocean (Kinlan and Gaines, 2003). Still, this does not mean that speciation is impeded in the ocean. Many of the events that modulate large scale evolutionary dynamics on land have just as much impact in the ocean: The glaciation cycles that fragmented terrestrial habitats also led to oscillations in sea levels which repeatedly separated and merged marine habitats. And while the closing of the Isthmus of Panama allowed for the *Great American Interchange*, it also separated the East Pacific from the Caribbean allowing for independent evolutionary trajectories of previously connected populations (Palumbi, 1994).

This shows that the ocean is a dynamic stage for evolution, which in fact has produced a great amount of biodiversity. Globally however, this is not homogeneous: the most diverse marine habitats are found in the tropics where large coral reef systems present a marine analog to rainforests in terms of biodiversity (Figure 1.2). The most extreme global hotspot in terms of marine biodiversity is the *Coral Triangle* in the Indo-West Pacific which is considered a *marine center of speciation* (Bowen *et al.*, 2013). Secondary centers of speciation have been identified in Caribbean, the Antarctic and the Northern Pacific based on the levels of endemism within these ares (Briggs, 2003). This is not to say that speciation runs faster in those areas, but merely that lineages that arise within those areas appear surpassingly successful evolutionary, which leads to a large influence of these centers on the global biodiversity (Briggs, 2003).

A further peculiarity of the marine habitat is that many species are characterized by huge effective population sizes compared to terrestrial organisms. As population size has direct impact on the efficiency of selection, this further underscores the point that evolutionary models based on terrestrial systems might be rather inept for marine organisms (Palumbi, 1994). An illustration of this might be the comparison of marine and terrestrial/ limnic *species flocks*. A species flock is a cluster of closely related species that are the result of an evolutionary radiation and are characterized by shared endemism to a specific area (Bowen *et al.*, 2020). While these are found frequently on marine islands and freshwater lakes, marine species flocks are quite rare. Bowen *et al.* (2020) propose that this is likely due to the dispersal potential of marine organisms which result in different scales on which adaptation and speciation work in the sea compared to on land. As a result, the sea is dominated by ancient, monotypic lineages and provides fewer potential for radiations. From an evolutionary biologist's perspective this is somewhat unfortunate, since the most promising setting to study the emergence of reproductive isolation are precisely species pairs that speciated recently and thus have not accumulated many evolutionary changes that could blur the patterns of speciation itself (Palumbi, 1994). While the ocean clearly offers many examples of speciation and ongoing evolution (as discussed above), species flocks allow for replication when studying a single divergence history.

Summary

- Marine model systems in speciation genomics are lacking.
- Central aspects in the establishment of reproductive isolation differ from

terrestrial systems due to intrinsic (life history stages) and extrinsic factors (lack of barriers).

- Adaptation and speciation might work on different scales, thus resulting in fewer young radiations and species flocks.

1.4. Hamlets

One exception to the paucity in species flocks in the marine realm is presented by the genus *Hypoplectrus* from the wider Caribbean — a supposed hotspot of speciation in the west Atlantic. Their distribution ranges from Trinidad and Tobago in the southeast, throughout the whole Caribbean and large parts of the Gulf of Mexico, covers the Bahamas and has a northeasterly outpost on Bermuda (Figure 1.3, Robertson and Tassell, 2019). This genus is comprised by eighteen currently described species about half of which have only been recognized in the last ten years (Del Moral Flores *et al.*, 2011; Lobel, 2011; Victor, 2012; Tavera and Acero, 2013). Some species are endemic to specific regions such as the maya hamlet (*H. maya*) that occurs only on the Belizean part of the mesoamerican barrier reef (Lobel, 2011) or the striped hamlet (*H. liberte*) that is restricted to northeastern Haiti. Nevertheless, most hamlets have pan-caribbean distributions leading to high levels of sympatry through most of the genus range with up to nine species co-occuring on the same reef (Thresher, 1978; Puebla *et al.*, 2012). Historically, their status as separate species or single species has long been debated, even though the existence of different color variants has been acknowledged since the days of the hamlets first scientific description:

> After an examination of the large series of typical forms sent by Professor Poey to the Museum at Cambridge, we find ourselves driven to the conclusion that all the common forms of *Hypoplectrus* probably constitute but a single species, subject to almost endless variations in color. This view we here adopt, leaving for convenience sake the various nominal species to stand as color varieties or subspecies, produced by the action of some agencies as yet unknown.
>
> (Jordan and Evermann, 1896)

While the discussion about the status as species versus color morphs carried through

Figure 1.3: Distribution of the genus *Hypoplectrus*. In the majority of their range several hamlet species are present in sympatry with up to nine different species on a single coral reef. Data from (Robertson and Tassell, 2019).

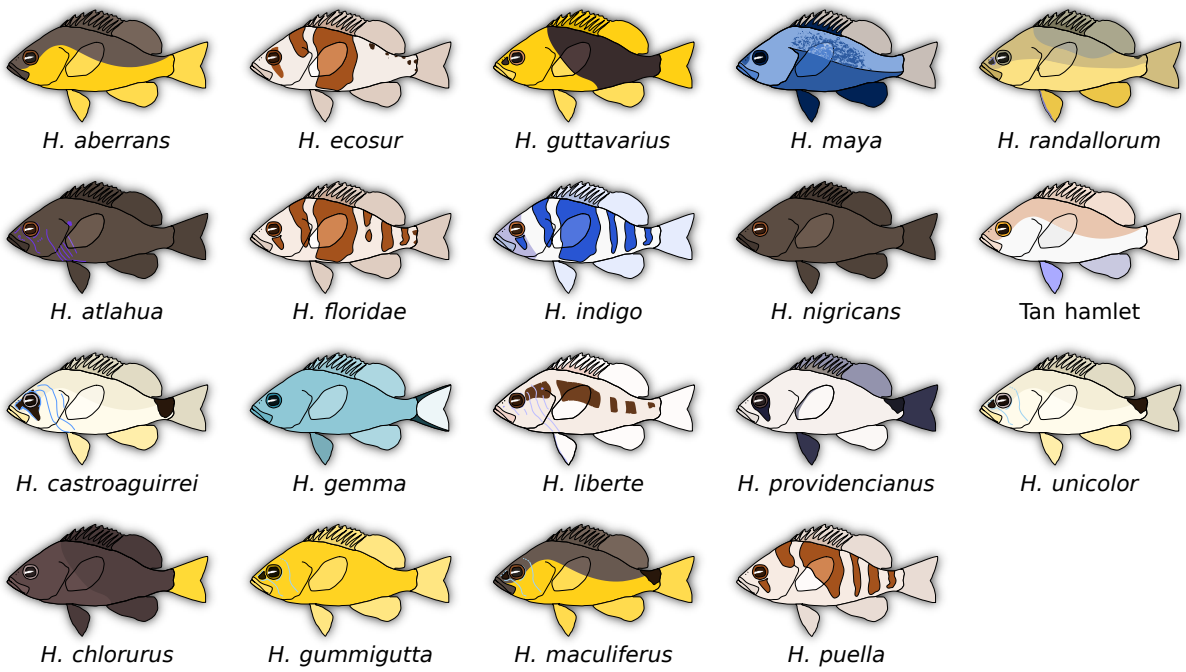


Figure 1.4: Overview of all currently described hamlet species. The different hamlet species display a wide variety of color patterns but are morphologically uniform. Note, that while the schematics resembles the typical phenotype of the respective species, color pattern are also variable within species. Note, that there are two different phenotypes of "tan hamlets" — *H. randallorum* and the "tan hamlet". These are assumed to be independent species, but currently, only *H. randallorum* is scientifically described.

the largest part of the twentieth century (Randall and Randall, 1960; Thresher, 1978; Graves and Rosenblatt, 1980), the classification as separate species has been widely accepted by now. This discussion stems from the broad sympatry and the fact that hamlets show no decisive differences in ecology, morphology or (early) genetic markers and are thus only distinguishable based on coloration. Yet, based on their striking color patterns the different species are easily identified (Figure 1.4) and mate choice in hamlets is highly assortative with respect to color pattern (Puebla *et al.*, 2007). Hamlets are simultaneous hermaphrodites that evolved a tit for tat egg trading, so mate choice is mutual. However, hybridization between the different hamlet species seems possible as hybrid spawnings can be observed in the wild, hybrids have been raised to adulthood in captivity (Domeier, 1994) and occasional intermediate color variants can be seen in the field. Moreover Puebla *et al.* (2012) showed that mate choice and the tendency for assortative mating appears to be context specific in the sense that hamlets seem more inclined to hybridize if they lack available conspecifics. Thus, while assortative mating largely restricts gene flow between species, they are likely still not completely reproductively isolated. The whole ambiguity surrounding the question of hamlets as *good species* according to the biological species concept leads back to the conceptual issues regarding the process of speciation itself. For the following work, the current opinion on the hamlets as separate species will be followed, while recognizing that they are spanning a rather early section of the speciation continuum (Puebla *et al.*, 2012). This pragmatic approach is not uncommon or

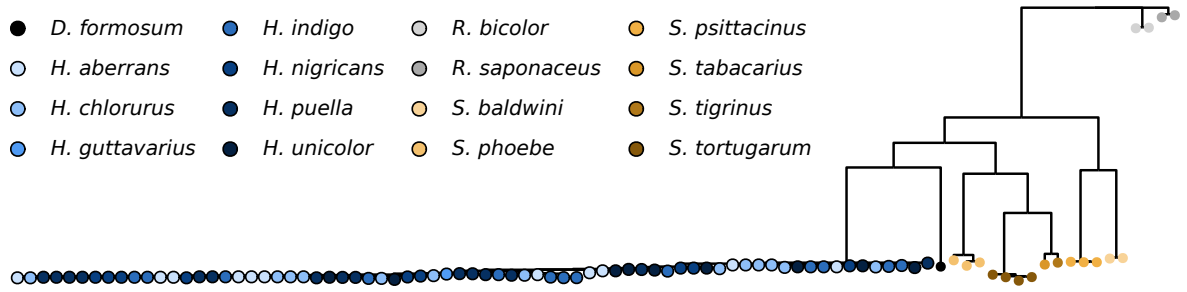


Figure 1.5: Phylogeny based on cytochrome b. Approximately-maximum-likelihood phylogeny of *Hypoplectrus* spp. and closely related species (*Diplectrum formosum*, *Serranus* spp. and *Rypticus* spp.), based on cytochrome *b* sequences. Sequence data from Mccartney *et al.* 2003.

novel considering that it resembles the species concept as described by Darwin:

> I look at the term species, as one arbitrarily given for the sake of convenience to a set of individuals closely resembling each other, and that it does not essentially differ from the term variety, which is given to less distinct and more fluctuating forms. The term variety, again, in comparison with mere individual differences, is also applied arbitrarily, and for mere convenience sake. (Darwin, 1859)

The genetic differentiation between the hamlet species is very low based on mtDNA (García-Machado *et al.*, 2004), AFLPs (Barreto and McCartney, 2008), microsatellites (Puebla *et al.*, 2008) and RAD sequencing (Puebla *et al.*, 2014), indicating a quite recent origin of the radiation. In contrast to this, the genus *Hypoplectrus* appears to be quite substantially diverged from other closely related Serranids based on mtDNA (Figure 1.5, Mccartney *et al.* 2003).

The shallow level of divergence, the large degree of sympatry and the endemism with respect to the greater Caribbean makes the hamlets a rare picture-book example of a marine species flock. It is note worthy though that, while the hamlets are considered an evolutionary radiation, the extent to which this is an *adaptive* one is less clear, since all hamlet species are considdered ecologically fairly similar (Thresher, 1978).

All hamlets are predatory coral reef fishes that prey mostly on benthic invertebrates and inhabit reefs from very shallow to intermediate depths. While slight variation in diet (Whiteman *et al.*, 2007) and depth distribution (Aguilar-Perera, 2003) have been reported, those are not substantial (Holt *et al.*, 2008). The color pattern itself could be under natural selection in some species (in addition to sexual selection by assortative mate choice), given that some hamlet species are considered aggressive mimics. For several hamlet species (Thresher, 1978) — especially for the blue (*H. gemma*) and the butter hamlet (*H. unicolor*) putative sympatric models have been proposed (Randall and Randall, 1960; Puebla *et al.*, 2007) and for *H. unicolor* active tracking behavior with increased predation success has been shown (Puebla *et al.*, 2007, 2018). The hypothesis behind aggressive mimicry is that the predatory but comparably rare hamlets imitate a far more abundant and non-predatory model fish species

both in appearance and in behavior. They could then utilize this cover to approach their prey, which is unalert because the model species that is mimicked is not perceived as threat (Thresher, 1978). Yet, while aggressive mimicry could render the coloration of specific hamlet species adaptive, it is unlikely that this is the case for all species. A second type of natural selection on the hamlet color pattern could be the promotion of disruptive color patterns like bars, which are believed to improve camouflage in the complex reef habitat (Phillips *et al.*, 2017). Still, that leaves a considerable share of brightly colored, non-barred hamlets that lack a model for aggressive mimicry without an obvious angle for natural selection. As a result the question, which forces specifically act to reduce gene flow between the different hamlet species, is still open — while sexual selection appears to be of importance, natural selection may or may not.

Summary

- Hamlets (*Hypoplectrus* spp., Serranidae) are a marine species flock of coral reef fishes.
- The 18 different species are largely sympatric through most of the Caribbean and the Gulf of Mexico.
- The species are exclusively distinguishable based on their bright color patterns and mate assortatively.
- Genetically, they are not strongly differentiated, thus covering the lower range of the speciation continuum.

1.5. Thesis Outline and Objective

The work presented within this doctoral thesis marks the transition from genetic to genomic research in hamlets. It thus introduces the hamlets as a model system for marine speciation and aims at contributing to the understanding of the early stages of speciation and evolutionary radiations in the marine environment. The central focus lies on the question which evolutionary mechanisms are acting on gene flow between the different species. In this context, potential targets of differential selection, influences of the genomic architecture and signs of past and present gene flow are explored. The work is distributed over four separate manuscripts which form the following chapters:

Manuscript 1: Hench, K. *et al*. (2017). Temporal changes in hamlet communities (*Hypoplectrus* spp., Serranidae) over 17 years. *Journal of Fish Biology*, 91(5): 1475–1490.

The first manuscript investigates the temporal stability of the hamlet community composition within a group of Puerto Rican reefs. Comparing relative species abundances in the context of differing ecological conditions over time and space, we examine the effect of potential selective factors on the local hamlet community.

Manuscript 2: Hench, K. *et al*. (2019). Inter-chromosomal coupling between vision and pigmentation genes during genomic divergence. *NatureEcology & Evolution*, 3(4): 657–667.

The second manuscript introduces the newly assembled hamlet reference genome, thus opening up the system for population ge-

nomic research. Based on whole genome re-sequencing data from three hamlet species, we investigate the interaction of pigmentation genes and genes underlying the visual system — two decisive factors regarding the assortative mating system in hamlets.

Manuscript 3: Moran, B.M. *et al*. (2019). The evolution of microendemism in a reef fish (*Hypoplectrus maya*). *Molecular Ecology*, 28(11): 2872–2885.

In the third manuscript we explore the demographic history of a rare hamlet species endemic to the Belizean part of the Mesoamerican Barrier Reef. We investigate the historical dynamics of this species' effective population size and put it into the context of three widespread and sympatric sister species as well as the phenotypically most similar hamlet species.

Manuscript 4: Hench, K. *et al*. (in revision). The genomic origins of a marine radiation. *Current Biology*

The last manuscript provides a cross-section through the hamlet radiation comparing eight different hamlet species. Using the low background levels of divergence, we look for signs of incipient divergent selection and screen the phylogenic history of those genomic ares for discordances whith each other and the genome wide background signal.

Intro References

Adams, M.D. *et al*. (2000). The Genome Sequence of Drosophila melanogaster. *Science*, 287(5461):2185–2195.

Aguilar-Perera, A. (2003). Abundance and distribution of hamlets (Teleostei : Hypoplectrus) in coral reefs off southwestern Puerto Rico: Support for the multiple-species hypothesis. *CARIBBEAN JOURNAL OF SCIENCE*, 39(1):147–151.

Alfaro, M.E. *et al*. (2009). Nine exceptional radiations plus high turnover explain species diversity in jawed vertebrates. *Proceedings of the National Academy of Sciences*, 106(32):13410–13414.

Barreto, F.S. and McCartney, M.A. (2008). Extraordinary AFLP fingerprint similarity despite strong assortative mating between reef fish color morphospecies. *Evolution*, 62(1):226–233.

Bowen, B.W. *et al*. (2020). Species Radiations in the Sea: What the Flock? *Journal of Heredity*, 111(1):70–83.

Bowen, B.W. *et al*. (2013). The origins of tropical marine biodiversity. *Trends in Ecology & Evolution*, 28(6):359–366.

Briggs, J.C. (2003). Marine centres of origin as evolutionary engines. *Journal of Biogeography*, 30(1):1–18.

Chinwalla, A.T. *et al*. (2002). Initial sequencing and comparative analysis of the mouse genome. *Nature*, 420(6915):520–562.

Cook, O.F. (1906). Factors of Species-Formation. *Science*, 23(587):506–507.

Cruickshank, T.E. and Hahn, M.W. (2014). Reanalysis suggests that genomic islands of speciation are due to reduced diversity, not reduced gene flow. *Molecular Ecology*, 23(13):3133–3157.

da Fonseca, R.R. *et al*. (2016). Next-generation biology: Sequencing and data

analysis approaches for non-model organisms. *Marine Genomics*, 30:3 – 13.

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

Dasmahapatra, K.K. *et al*. (2012). Butterfly genome reveals promiscuous exchange of mimicry adaptations among species. *Nature*, 487(7405):94–98.

Del Moral Flores, L.F., Tello-Musi, J.L. and Martínez-Pérez, J.A. (2011). Descripción de una nueva especie del género Hypoplectrus (Actinopterigy: Serranidae) del Sistema Arrecifal Veracruzano, suroeste del Golfo de México. *Revista de Zoología*, pages 1–10.

Dobzhansky, T. (1937). *Genetics and the Origin of Species*. Columbia University biological series. Columbia University Press.

Domeier, M.L. (1994). Speciation in the serranid fish Hypoplectrus. *Bulletin of Marine Science*, 54(1):103–141.

Ekblom, R. and Wolf, J.B.W. (2014). A field guide to whole-genome sequencing, assembly and annotation. *Evolutionary applications*, 7(9):1026–1042.

García-Machado, E., Chevalier Monteagudo, P.P. and Solignac, M. (2004). Lack of mtDNA differentiation among hamlets (Hypoplectrus, Serranidae). *Marine Biology*, 144(1):147–152.

Graves, J.E. and Rosenblatt, R.H. (1980). Genetic Relationships of the Color Morphs of the Serranid Fish Hypoplectrus unicolor. *Evolution*, 34(2):240.

Guerrero, R.F. and Hahn, M.W. (2017). Speciation as a sieve for ancestral polymorphism. *Molecular Ecology*, 26(20):5362–5368.

Harrison, R.G. and Larson, E.L. (2014). Hybridization, introgression, and the nature of species boundaries. *The Journal of heredity*, 105 Suppl:795–809.

Holt, B.G. *et al*. (2008). Stable isotope analysis of the Hypoplectrus species complex reveals no evidence for dietary niche divergence. *Marine Ecology Progress Series*, 357:283–289.

Hulsey, C.D. and Renn, S.C.P. (2009). Genomics and vertebrate adaptive radiation: A celebration of the first cichlid genome. *Integrative and Comparative Biology*, 49(6):613–617.

Jenkins, C.N. and Van Houtan, K.S. (2016). Global and regional priorities for marine biodiversity protection. *Biological Conservation*, 204:333–339.

Jordan, D.S. and Evermann, B.W. (1896). *The fishes of North and Middle America: a descriptive catalogue of the species of fish-like vertebrates found in the waters of North America north of the Isthmus of Panama. Part I*. Government Printing Office, Washington.

Kane, N.C. *et al*. (2011). Progress towards a reference genome for sunflower. *Botany*, 89(7):429–437.

Kinlan, B.P. and Gaines, S.D. (2003). Propagule dispersal in marine and terrestrial environments: a community perspective. *Ecology*, 84(8):2007–2020.

Kronforst, M.R. *et al*. (2013). Hybridization Reveals the Evolving Genomic Architecture of Speciation. *Cell Reports*, *5*(3):666–677.

Labandeira, C.C. (2005). Invasion of the continents: cyanobacterial crusts to tree-inhabiting arthropods. *Trends in Ecology & Evolution*, 20(5):253–262.

Lamichhaney, S. *et al*. (2015). Evolution of Darwin's finches and their beaks revealed by genome sequencing. *Nature*, 518(7539):371–375.

Lander, E.S. *et al*. (2001). Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921.

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

Malinsky, M. *et al*. (2018). Whole-genome sequences of Malawi cichlids reveal multiple radiations interconnected by gene flow. *Nature Ecology & Evolution*, 2(12):1940–1955.

Mallet, J. (2001). The speciation revolution. *Journal of Evolutionary Biology*, 14(6):887–888.

Mallet, J. (2007). Hybrid speciation. *Nature*, 446(7133):279–283.

Mallet, J., Besansky, N. and Hahn, M.W. (2016). How reticulated are species? *BioEssays*, 38(2):140–149.

Marques, D.A., Meier, J.I. and Seehausen, O. (2019). A Combinatorial View on Speciation and Adaptive Radiation. *Trends in Ecology & Evolution*, 34(6):531–544.

Marshall, C.R. (2006). Explaining the Cambrian "Explosion" of Animals". *Annual Review of Earth and Planetary Sciences*, 34(1):355–384.

Martin, C.H. and Richards, E.J. (2019). The Paradox Behind the Pattern of Rapid Adaptive Radiation: How Can the Speciation Process Sustain Itself Through an Early Burst? *Annual Review of Ecology, Evolution, and Systematics*, 50(1):569–593.

Mayr, E. (1942). *Systematics and the origin of species, from the viewpoint of a zoologist*. New York: Columbia University Press.

Mccartney, M.A. *et al*. (2003). Genetic mosaic in a marine species flock. *Molecular Ecology*, 12(11):2963–2973.

Mendel, G. (1866). *Versuche über Pflanzen-Hybriden*, volume IV 13 für das Jahr 1865, Abhandlungen. Verhandlungen des naturforschenden Vereines in Brünn.

Merrill, R.M. *et al*. (2019). Genetic dissection of assortative mating behavior. *PLoS biology*, 17(2).

Metzker, M.L. (2010). Sequencing technologies — the next generation. *Nature Reviews Genetics*, 11(1):31–46.

Palumbi, S.R. (1994). Genetic-divergence, reproductive isolation, and marine speciation. *Annual Review Of Ecology And Systematics*, 25:547–572.

Phillips, G.A.C. *et al*. (2017). Disruptive colouration in reef fish: does matching the background reduce predation risk? *The Journal of Experimental Biology*, 220(11):1962 LP – 1974.

Puebla, O., Bermingham, E. and 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(6):1405–1415.

Puebla, O., Bermingham, E. and Guichard, F. (2012). Pairing dynamics and the origin of species. *Proceedings of the Royal Society B: Biological Sciences*, 279(1731):1085–1092.

Puebla, O. *et al*. (2007). Colour pattern as a single trait driving speciation in Hypoplectrus coral reef fishes? *Proceedings of the Royal Society B: Biological Sciences*, 274(1615):1265–1271.

Puebla, O., Bermingham, E. and McMillan, W.O. (2014). Genomic atolls of differentiation in coral reef fishes (Hypoplectrus spp., Serranidae). *Molecular Ecology*, 23(21):5291–5303.

Puebla, O. *et al*. (2018). Social-trap or mimicry? An empirical evaluation of the Hypoplectrus unicolor–Chaetodon capistratus association in Bocas del Toro, Panama. *Coral Reefs*, 37(4):1127–1137.

Randall, J.E. and Randall, H.A. (1960). Examples of mimicry and protective resemblance in tropical marine fishes. *Bulletin of Marine Science*, 10(4):444–480.

Reid, J.B. and Ross, J.J. (2011). Mendel's genes: Toward a full molecular characterization. *Genetics*, 189(1):3–10.

Robertson, D.R. and Tassell, J.V. (2019). Shorefishes of the Greater Caribbean: online information system. Version 2.0. Smithsonian Tropical Research Institute, Balboa, Panamá. http://biogeodb.stri.si.edu/caribbean/en/research/index/range.

Roesti, M., Moser, D. and Berner, D. (2013). Recombination in the threespine stickleback genome—patterns and consequences. *Molecular Ecology*, 22(11):3014–3027.

Schluter, D. (2000). *The Ecology of Adaptive Radiation (Oxford Series in Ecology and Evolution)*. Oxford University Press.

Seehausen, O. *et al*. (2014). Genomics and the origin of species. *Nature Reviews Genetics*, 15(3):176–192.

Shaw, K.L. and Mullen, S.P. (2014). Speciation Continuum. *Journal of Heredity*, 105(S1):741–742.

Simões, M. *et al*. (2016). The Evolving Theory of Evolutionary Radiations. *Trends in Ecology & Evolution*, 31(1):27–34.

Steiner, C.C. *et al*. (2013). Conservation genomics of threatened animal species. *Annual Review of Animal Biosciences*, 1(1):261–281. PMID: 25387020.

Stroud, J.T. and Losos, J.B. (2016). Ecological Opportunity and Adaptive Radiation. *Annual Review of Ecology, Evolution, and Systematics*, 47(1):507–532.

Tavera, J. and Acero, A. (2013). Description of a new species of Hypoplectrus (Perciformes:Serrnidae) from the southern Gulf of Mexico. *Aqua, Journal of Ichthyology*, 19(1):29–38.

Thompson, E.A. (1990). R. a. fisher's contributions to genetical statistics. *Biometrics*, 46(4):905–914.

Thresher, R.E. (1978). Polymorphism, mimicry, and the evolution of the hamlets (Hypoplectrus, Serranidae). *Bulletin of Marine Science*, 28(2):345–353.

Victor, B.C. (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*, 5:2–19.

Wagner, C.E., Harmon, L.J. and Seehausen, O. (2012). Ecological opportunity and sexual selection together predict adaptive radiation. *Nature*, 487(7407):366–369.

Watson, J.D. and Crick, F.H.C. (1953). Genetical Implications of the Structure of Deoxyribonucleic Acid. *Nature*, 171(4361):964–967.

Whiteman, E.A., Côté, I.M. and Reynolds, J.D. (2007). Ecological differences between hamlet (Hypoplectrus: Serranidae) colour morphs: Between-morph variation in diet. *Journal of Fish Biology*, 71(1):235–244.

– 2 –

Temporal Changes in Hamlet Communities (*Hypoplectrus* spp., Serranidae) over 17 Years

Kosmas Hench [1], W. Owen McMillan [2], Ricardo Betancur-R. [3], Oscar Puebla [1]

1 GEOMAR Helmholtz Centre for Ocean Research Kiel, Evolutionary Ecology of Marine Fishes, Düsternbrooker Weg 20, 24105 Kiel, Germany

2 Smithsonian Tropical Research Institute, Apartado Postal 0843-03092, Panamá, República de Panamá

3 Department of Biology, University of Puerto Rico – Río Piedras, PO Box 23360, San Juan, Puerto Rico

This study was originally published in *Journal of Fish Biology*. Figures were re-drawn for this thesis but no additional changes were made. The re-print within this thesis is in agreement with *John Wiley and Sons* under license number 4826450091578.

Original publication

Hench, K. *et al*. (2017). Temporal changes in hamlet communities (*Hypoplectrus* spp., Serranidae) over 17 years. *Journal of Fish Biology*, 91(5):1475–1490. doi: 10.1111/jfb.13481.

Abstract

Transect surveys of hamlet communities (*Hypoplectrus* spp., Serranidae) covering $14\ 000\ m^2$ across 16 reefs off La Parguera, Puerto Rico, are presented and compared with a previous survey conducted in the year 2000. The hamlet community has noticeably changed over 17 years, with a $> 30\%$ increase in relative abundance of the yellow tail hamlet *Hypoplectrus chlorurus* on the inner reefs at the expense of the other hamlet species. The data also suggest that the density of *H. chlorurus* has declined and that its distribution has shifted towards shallower depths. Considering that *H. chlorurus* has been previously identified as one of the few fish showing a positive association with seawater turbidity on the innerreefs of La Parguera and that sedimentation of terrestrial origin has increased over recent decades on these reefs, it is proposed that turbidity may constitute an important but so far overlooked ecological driver of hamlet communities.

Keywords: community stability, Parguera, Puerto Rico, sedimentation, turbidity.

2.1. Introduction

The hamlets, simultaneously hermaphroditic sea basses from the tropical western Atlantic Ocean (*Hypoplectrus* spp., Perciformes: Serranidae), have intrigued ichthyologists for decades (Barlow, 1975; Fischer and Petersen, 1987; Domeier, 1994; McCartney et al., 2003; Theodosiou et al., 2016). Seventeen species are now recognized, a third of which have been described in the past few years (Del Moral Flores et al., 2011; Lobel, 2011; Victor, 2012; Tavera and Acero, 2013). Hamlets from the Gulf of Mexico appear to be well-diverged from the Caribbean hamlets at mitochondrial DNA markers (Victor (2012); Tavera and Acero (2013)), yet hamlets tend to be very closely related genetically within these two regions (McCartney et al., 2003; Barreto and McCartney, 2008; Tavera and Acero, 2013; Puebla et al., 2014). Hamlets are also very similar from an ecomorphological perspective and, to date, colour pattern is the only trait that has been found to consistently differentiate species (Randall, 1968; Lobel, 2011; Tavera and Acero, 2013). Yet colour pattern also varies within species, both within and between locations (Thresher, 1978; Aguilar-Perera, 2004), complicating their taxonomy and identification.

Hamlets vary in their distribution (Aguilar-Perera and Gonzalez-Salas, 2010; Holt et al., 2010), but tend to be highly sympatric, with up to nine species found on a single reef (Puebla et al., 2012a). Hamlets are reef-associated predators that feed on small invertebrates and fishes (Randall 1967; Holt et al. 2008, G. M. Serviss, unpubl. data). Sympatric species tend to live in the same habitat and have similar diets, except for the indigo hamlet *Hypoplectrus indigo* (Poey 1851) that appears to feed mostly on fishes (Whiteman et al., 2007). Spawning occurs before sunset on a daily basis throughout the year. Sympatric species spawn at the same time and in the same area, often within sight of each other. Yet mating is strongly assortative with respect to colour pattern, with $> 98\%$ of spawnings occurring

among members of the same species (Fischer, 1980a; Barreto and McCartney, 2008; Puebla et al., 2007, 2012a). Apparently, there are no strong intrinsic post-fertilization barriers in hamlets (Whiteman and Gage, 2007) and in the only case where hybrids were bred in aquaria, they appeared intermediate between parental species in terms of colour pattern (Domeier, 1994).

Hamlets have served as a distinctive model system for the study of a variety of ecological and evolutionary processes including the evolution and maintenance of simultaneous hermaphroditism (Fischer, 1980b), sex allocation (Fischer, 1981), egg trading (Fischer and Petersen, 1987), sexual selection (Puebla et al., 2011), dispersal (Puebla et al., 2009), local adaptation (Picq et al., 2016), speciation (Holt et al., 2011), evolutionary radiation (Puebla et al., 2008) and recombination (Theodosiou et al., 2016). Temporal changes in hamlet communities can potentially affect or be affected by such processes. For example, changes in population densities and relative abundances are expected to affect effective population sizes as well as the potential for hybridization among the different species (Puebla et al., 2012a). Temporal changes in hamlet densities and relative abundances could also provide hints about the ecological factors that shape hamlet communities, which are still eluding ecologists. Yet detailed hamlet surveys are scarce and very little is known about the dynamics of hamlet communities. Data from general fish surveys are to be treated with caution due to extensive colour pattern variation in the group that complicates species identification and typically do not provide a detailed picture of local communities. Hamlet population densities can be relatively low (of the order 10 fish $1000m^{-2}$ of reef) and several species are rare (of the order 1 fish $1000m^{-}2$ or less (Puebla et al., 2012a)), requiring extensive surveys.

The hamlets from La Parguera, Puerto Rico, constitute a notable exception with a thorough survey available for the year 2000 (Aguilar-Perera, 2003). This survey stands out because it targets the hamlet community specifically, is spread across 16 reefs identified with reef names and GPS coordinates, provides raw fish counts with depth and reef type (inner v. outer reefs) and is complemented by a note on colour pattern variation at this location (Aguilar-Perera, 2004). Here, the opportunity is taken to evaluate the temporal dynamics of hamlet communities. Seventeen years later the same reefs were revisited and transect surveys covering $14000m^2$ across 16 reefs were conducted. The transect data indicate that hamlet relative abundances have noticeably changed over 17 years, with most notably a 30% increase of the yellowtail hamlet *Hypoplectrus chlorurus* (Cuvier 1828) on the inner reefs at the expense of the other hamlet species. The potential drivers of this change are discussed in light of the literature and it is proposed that recent changes in water turbidity regimes might play an important, but previously overlooked role in this context.

2.2. Materials and Methods

Fieldwork was conducted under the Institutional Animal Care and Use Committee (IACUC) protocol 2017-0101–2020-2 and the Puerto Rico Departamento de Recursos Naturales y Ambientales research permit $\#2016 - IC - 127(E)$ between 13 and 24

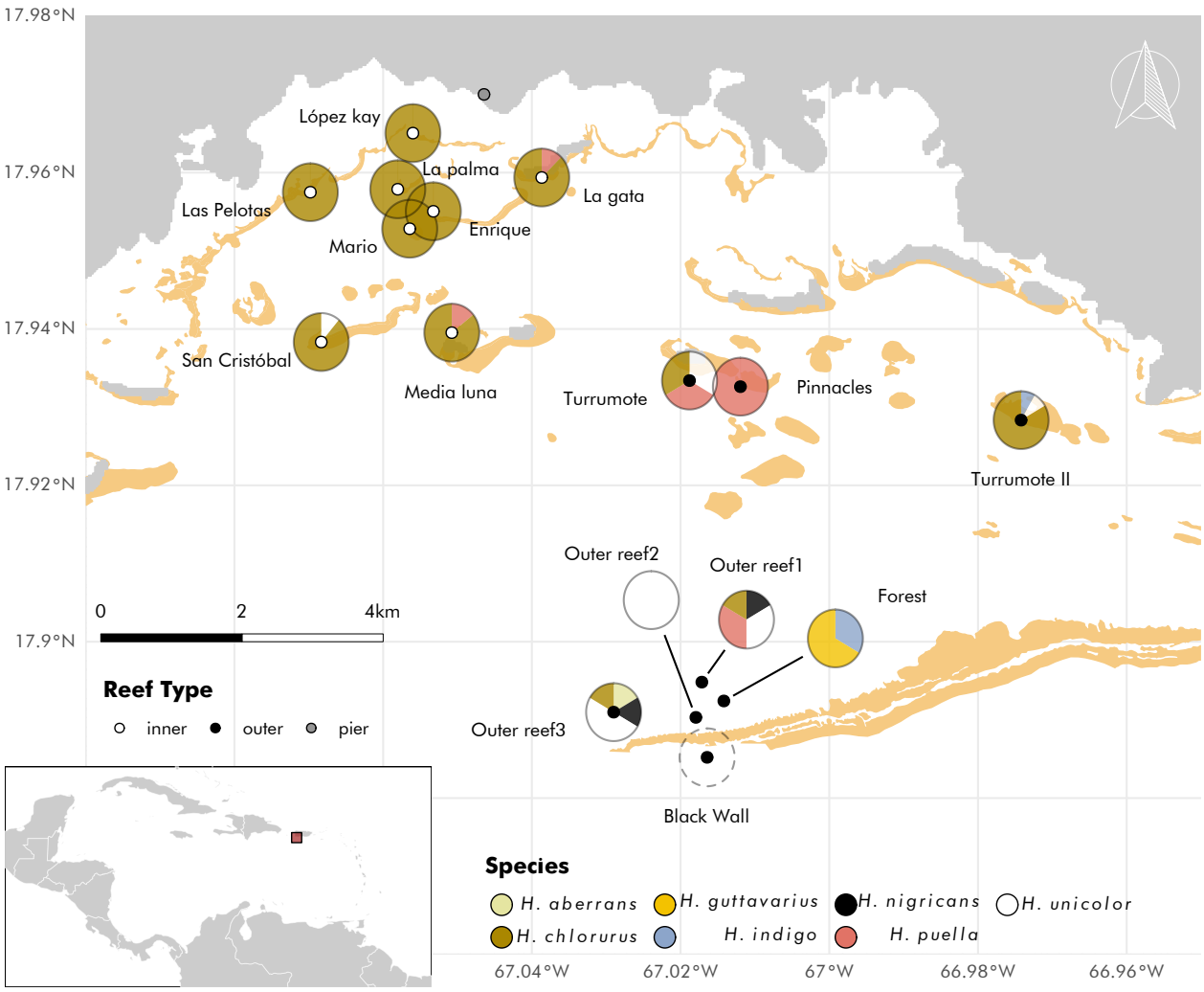


Figure 2.1: Sampling location and hamlet relative abundances. Location (dots) and hamlet relative abundances (pie charts) from 35 survey transects covering a total of 14,000 m2 of reef on 16 reefs off La Parguera (Puerto Rico). Reefs indicated in grey (spatial data on reef extent taken from UNEP-WCMC, WorldFish Centre, WRI, TNC (2010), http://data.unep-wcmc.org/datasets/1).

March 2017 on 16 reefs in the vicinity of La Parguera, Puerto Rico (Figure 2.1), targeting the same reefs surveyed by Aguilar-Perera (2003).

Transect Surveys: Hamlet population densities and relative abundances were estimated using scuba visual censuses. Belt transects were preferred over the roaming surveys adopted by Aguilar-Perera (2003) since they provide standardized population density estimates (i.e. number of individuals per unit of reef area) that can be compared across time and space and also used to estimate population sizes (Puebla et al., 2009, 2012b). Yet hamlet densities are quite low in La Parguera, resulting in a large proportion of empty $25 \times 2m$ transects (Aguilar-Perera, 2003). In order to address this limitation, $100 \times 4m$ transects were adopted, following the approach used in previous hamlet surveys (Puebla et al., 2007, 2008, 2009, 2012b), resulting in an eightfold increase of the area surveyed per transect ($400m^2$ v. $50m^2$ for $25 \times 2m$ transects). Briefly, two divers swam in parallel a few feet above the reef with each diver counting all

the Hypoplectrus spp. observed within $2m$ on each side of a $100m$ transect tape. Fishes swimming across the transect tape were signalled to avoid counting the same fish twice. Coral-reef habitat was specifically targeted for the transects (i.e. avoiding sandy or seagrass areas) and an effort was made to cover a variety of depths and reef zones (e.g. reef slope v. reef flat), but the exact location of each transect was randomly chosen. An effort was made to broadly match the sampling effort of Aguilar-Perera (2003) and avoid comparatively under or over-sampling, with also a broadly similar proportion of surveys conducted in the inner v. outer reefs.

Data Analysis: Densities (number of fish $1000\ m^{-}2$ of reef) and hamlet relative abundances (%) were estimated for each species, reef site, reef type (inner v. outer reefs) and overall. Similar data were recompiled from Aguilar-Perera (2003)'s raw species counts (it is noted that some minor numerical errors were found in the total counts in Aguilar-Perera (2003); (Table I), hence the relative abundances reported here differ slightly from those in the original paper).

Differences in relative abundances were tested with permutational analysis of variance (PERMANOVA; Anderson 2001) with 1999 permutations per test and visualized using non-metric multidimensional scaling as implemented in the vegan package in R (Oksanen et al. 2017; www.r-project.org), using the Bray-Curtis measure of ecological distance in both cases (Bray and Curtis, 1957). This analysis was done at the reef level, i.e. considering relative abundances per reef (not per transect). Differences between the inner and outer reefs were first tested and given the significant outcome (see Results) changes in hamlet relative abundances between 2000 (Aguilar-Perera, 2003) and 2017 (this study) were tested for the inner and outer reefs separately. Hamlet diversity in the inner and outer reefs for the years 2000 (Aguilar-Perera, 2003) and 2017 were estimated using the effective number of species of the first order, corresponding to the exponential of the Shannon entropy and referred from here on as the effective number of species (Hill, 1973; Jost, 2006). Differences in diversity between 2000 (Aguilar-Perera, 2003) and 2017 were tested for the inner and outer reefs separately with a Mann–Whitney U-test and finally the depth distribution of *H. chlorurus* was tested for significant differences between shallow ($< 5.5\ m$) and deep ($\geq 5.5\ m$) sections of the inner reefs using a χ^2 -test following Aguilar-Perera (2003). This test was not repeated for the other species due to the relatively low number of sightings.

2.3. Results

Transect Surveys: A total of 35 non-overlapping transects were conducted, covering an area of $14\ 000\ m^2$ across 16 reefs at depths ranging between 2 and $18m$ (Table 2.1). One hundred and seventeen hamlets from seven species were sighted within the transects, providing an overall hamlet density estimate of 8.4 ± 1.5 fish $1000\ m^{-2}$ of reef (mean $\pm$ S.E.). The most abundant species by far was H. chlorurus, representing 80.3% of all hamlets seen, followed by the butter hamlet *Hypoplectrus unicolor* (Walbaum 1792), 7.7% and the barred hamlet *Hypoplectrus puella* (Cuvier 1828), 6.0%. The shy *Hypoplectrus guttavarius* (Poey 1851), *H. indigo*, black *Hypoplectrus nigricans* (Poey 1851) and yellow-belly *Hypoplectrus aberrans* (Poey 1868) hamlets were rare, with only one or two individuals

Table 2.1: Hamlet counts, population densities and relative abundances from 35 survey transects of $400\ m^2$ each covering a total of $14\ 000\ m^2$ of reef across 16 reefs off La Parguera, Puerto Rico, in March 2017. Data from 2000 recompiled from Aguilar-Perera (2003). The hamlet species is indicated by the first three letters: *H. chlorurus* (chl), *H. unicolor* (uni), *H. puella* (pue), *H. guttavarius* (gut), *H. indigo* (ind), *H. nigricans* (nig) and *H. aberrans* (abe).

Inner reefs	Position N (decimal degrees)	Position W (decimal degrees)	Depth (feet)	Total count	*chl*	*uni*	*pue*	*gut*	*ind*	*nig*	*abe*	Density (ind 1000 m-2)	Rel. ab. (%)
López kay	17.965	-67.056	11	10	0	0	0	0	0	0	10	25	8.5
López kay	17.965	-67.056	11	6	0	0	0	0	0	0	6	15	5.1
La gata	17.959	-67.039	26	3	0	0	0	0	0	0	3	8	2.6
La gata	17.959	-67.039	13	4	0	1	0	0	0	0	5	13	4.3
Mario	17.953	-67.056	18	0	0	0	0	0	0	0	0	0	0
Mario	17.953	-67.056	13	2	0	0	0	0	0	0	2	5	1.7
Mario	17.953	-67.056	6	0	0	0	0	0	0	0	0	0	0
Mario	17.953	-67.056	12	15	0	0	0	0	0	0	15	38	12.8
San Cristóbal	17.938	-67.068	22	0	1	0	0	0	0	0	1	3	0.9
San Cristóbal	17.938	-67.068	7	3	0	0	0	0	0	0	3	8	2.6
San Cristóbal	17.938	-67.068	9	5	0	0	0	0	0	0	5	13	4.3
La palma	17.958	-67.058	15	10	0	0	0	0	0	0	10	25	8.5
La palma	17.958	-67.058	15	7	0	0	0	0	0	0	7	18	6
Enrique	17.955	-67.053	14	8	0	0	0	0	0	0	8	20	6.8
Media luna	17.94	-67.051	15	3	0	1	0	0	0	0	4	10	3.4
Media luna	17.94	-67.051	13	3	0	0	0	0	0	0	3	8	2.6
Las Pelotas	17.957	-67.07	20	1	0	0	0	0	0	0	1	3	0.9
Las Pelotas	17.957	-67.07	10	1	0	0	0	0	0	0	1	3	0.9
Turrumote	17.934	-67.019	30	0	1	1	0	0	0	0	2	5	1.7
Turrumote	17.934	-67.019	15	0	0	0	0	0	0	0	0	0	0
Turrumote	17.934	-67.019	42-7	1	0	0	0	0	0	0	1	3	0.9
Turrumote II	17.928	-66.974	42	5	0	0	0	0	0	0	5	13	4.3
Turrumote II	17.928	-66.974	19	5	1	0	0	1	0	0	7	18	6
Pinnacles	17.97	-67.046	45	0	0	2	0	0	0	0	2	5	1.7
Total count				**92**	**3**	**5**	**0**	**1**	**0**[a]	**0**	**101**		
Density (inds 1000 m^{-2} of reef, mean ± SE)				9.6 ± 0.8	0.3 ± 0.1	0.5 ± 0.1	0.0 ± 0.0	0.1 ± 0.0	< 0.1[a]	0.0 ± 0.0		10.5 ± 1.9	
Total relative abundance (%)				91.1	3	5	0	1	< 1a	0			86.3
Total relative abundance Aguilar-Perera (2003) (%)				58.2	5.9	10.5	0	7.2	10.5	7.8			86.9
Outer reefs													
Black Wall	17.885	-67.016	59	0	0	0	0	0	0	0	0	0	0
Black Wall	17.885	-67.016	59	0	0	0	0	0	0	0	0	0	0
Forest	17.892	-67.014	44	0	0	0	0	1	0	0	1	3	0.9
Forest	17.892	-67.014	44	0	0	0	2	0	0	0	2	5	1.7
Outer reef 1	17.895	-67.017	52	0	0	1	0	0	1	0	2	5	1.7
Outer reef 1	17.895	-67.017	52	1	2	1	0	0	0	0	4	10	3.4
Outer reef 2	17.89	-67.018	51	0	1	0	0	0	0	0	1	3	0.9
Outer reef 2	17.89	-67.018	51	0	0	0	0	0	0	0	0	0	0
Outer reef 3	17.891	-67.029	50	0	2	0	0	0	1	0	3	8	2.6
Outer reef 3	17.891	-67.029	50	1	0	0	0	0	0	0	1	3	0.9
Outer reef 3	17.891	-67.029	50	0	1	0	0	0	0	1	2	5	1.7
Total count				**2**	**6**	**2**	**2**	**1**	**2**	**1**	**16**		
Density (inds 1000m^{-2} of reef, mean ± SE)				0.5 ± 0.1	1.4 ± 0.2	0.5 ± 0.1	0.5 ± 0.2	0.2 ± 0.1	0.5 ± 0.1	0.2 ± 0.1		3.6 ± 1.0	
Total relative abundance (%)				12.5	37.5	12.5	12.5	6.3	12.5	6.3			13.7
Total relative abundance Aguilar-Perera (2003) (%)				0	21.7	60.9	0	0	0	17.4			13.1
Overall													
Total count				**94**	**9**	**7**	**2**	**2**	**2**	**1**	**117**		
Density (inds 1000m^{-2} of reef, mean ± SE)				6.7 ± 0.6	0.6 ± 0.1	0.5 ± 0.1	0.1 ± 0.1	0.1 ± 0.0	0.1 ± 0.0	0.1 ± 0.0		8.4 ± 1.5	
Relative abundance (%)				80.3	7.7	6	1.7	1.7	0.7	0.9			100
Total relative abundance Aguilar-Perera (2003) (%)				50.6	8	17	0	6.3	9.1	9.1			100

[a] sighted outside of the transects on the inner reefs

seen over all transects. No other hamlets were sighted outside of the transects.

There was a significant difference in relative abundances between the inner and outer reefs (Figure 2.2 a; PERMANOVA $P < 0.01$). The inner reefs were dominated by *H. chlorurus* (91.1% of all hamlets seen) followed by *H. puella* (5.0%) and *H. unicolor* (3.0%) and *H. indigo* was rare (1.0%). No other species was sighted within the transects, but it is noted that *H. nigricans* was observed outside of the transects on the inner reefs, implying presence on these reefs at a density < 0.1 fish $1000m^{-2}$. The outer reefs were dominated

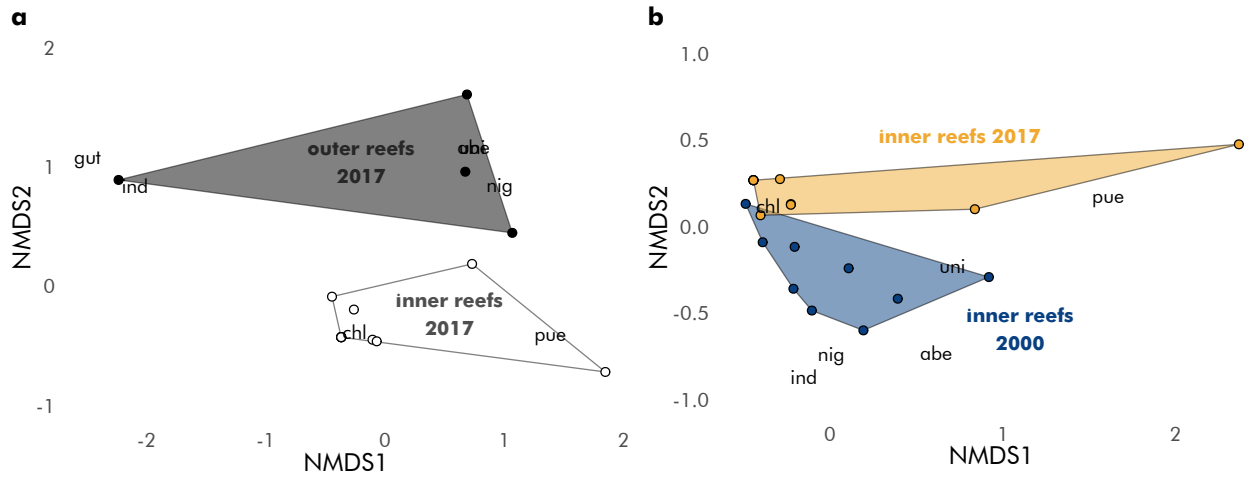
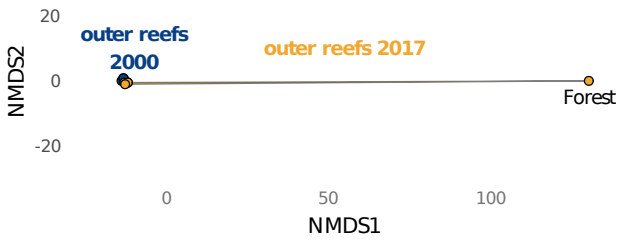


Figure 2.2: Non-metric multidimensional scaling (NMDS) plots. **a** for the comparison between inner and outer reefs (this study), **b** between the years 2000 (Aguilar-Perera, 2003) and 2017 (this study) for the inner reefs.

by *H. unicolor* (37.5% of all hamlets seen) followed by *H. chlorurus*, *H. puella*, *H. nigricans* and *H. guttavarius* (12.5% each) and finally *H. indigo* and *H. aberrans* (6.3% each). No other species was sighted outside of the transects on these reefs. The difference in community composition between the inner and outer reefs was accompanied by a significant difference in density, with an average of 10.5 ± 1.9 v. 3.6 ± 1.0 fish $1000m^{-2}$ of reef in the inner and outer reefs, respectively (mean $\pm$ S.E., Mann–Whitney U-test $P < 0.05$).

There was a significant change in relative abundances between the years 2000 (Aguilar-Perera, 2003) and 2017 on the inner reefs (Figure 2.2 b; PERMANOVA $P < 0.05$). The most notable difference was the relative increase of *H. chlorurus*, from 58.2% in 2000 to 91.1% in 2017 at the expense of *H. nigricans* (10.5% to $<$ 1%), *H. aberrans* (7.8 to 0%) and *H. indigo* (7.2 to 1.0%). There was a marginally significant change in relative abundances between 2000 and 2017 on the outer reefs (Suppl. Fig. 2.1; PERMANOVA $P = 0.05$), but it should be noted that this result is to be interpreted with caution given the low hamlet densities and counts on the outer reefs and the over-dispersion of the 2017 data compared with 2000 (due to the occurrence of both *H. guttavarius* and *H. indigo* in one outer reef in 2017, none of which were observed on the outer reefs in 2000). The most notable change on the outer reefs was a decrease of *H. puella* from 60.9% in 2000 to 12.5% in 2017 and an increase of *H. chlorurus* and *H. nigri-*



Suppl. Figure 2.1: NMDS outer reefs. Non-metric multidimensional scaling (NMDS) of *Hypoplectrus* spp. communities on the outer reefs off La Parguera, Puerto Rico, for the years 2000 (Aguilar-Perera, 2003) and 2017. The over-dispersion of the 2017 data compared with 2000 on the first NMDS axis is due to the occurrence of both *H. guttavarius* and *H. indigo* on one outer reef in 2017, none of which were observed in 2000.

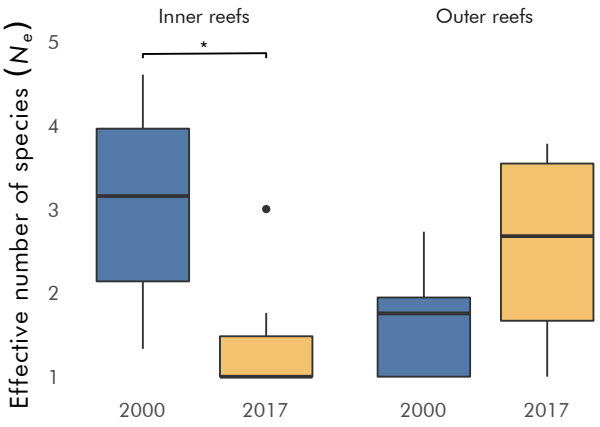


Figure 2.3: Comparison of the temporal change in the diversity of the first order for inner and outer reefs. Significance was tested independently according to Man Whitney U test.

cans from 0 to 12.5%. The presence of *H. guttavarius* on the outer reefs is also to be noted since this species was not reported by Aguilar-Perera (2003) in either the inner or outer reefs.

Effective number of species on the inner and outer reefs from Aguilar-Perera's data and this study are presented in Figure 2.3. The increased dominance of *H. chlorurus* on the inner reefs in 2017 compared with 2000 is reflected by a significant decrease in effective number of species, from 3.1 ± 1.1 in 2000 to 1.4 ± 0.6 in 2017 (mean $\pm$ S.D., Mann–Whitney U-test $P < 0.01$). The opposite trend was observed on the outer reefs (1.62 ± 0.25 in 2000 v. 2.53 ± 0.66 in 2017), but this difference was not significant (Mann–Whitney U-test $P > 0.05$) and here again caution is warranted due to the low species densities and counts on the outer reefs. Overall depth distributions of all hamlets sighted are presented in Figure 2.4 *Hypoplectrus chlorurus* was significantly more abundant in the shallow ($< 5.5m$) sections of the inner reefs than in the deeper areas ($\geq 5.5m$, χ^2-square test $P < 0.001$).

2.4. Discussion

Temporal Changes In The Hamlet Community: Hamlet relative abundances have noticeably changed between the years 2000 and 2017 in La Parguera, with most notably a $>$ 30% increase of *H. chlorurus* on the inner reefs, from 58.2% in 2000 (Aguilar-Perera,

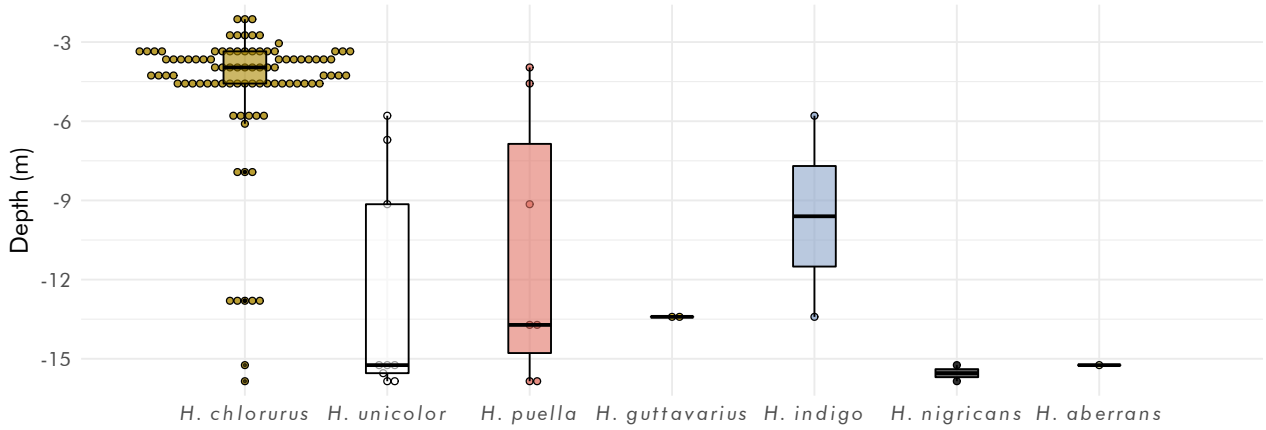


Figure 2.4: Hamlet depth distribution from 35 transect surveys covering $14\,000\ m^2$ of reef on 16 reefs off La Parguera (Puerto Rico).

2003) to 91.1% in 2017. Both surveys targeted the hamlets specifically, were extensive and highly replicated at the inner-reef level, conducted on the same reefs and in the same depth range. It is also noteworthy that the hamlets are quite conspicuous. Thus, differences in relative abundance between Aguilar-Perera (2003) and this study are unlikely to be due to methodological biases. The relative increase of *H. chlorurus* on the inner reefs occurred at the expense of the other, less abundant species, resulting in a significant decrease in effective number of species from 2000 to 2017.

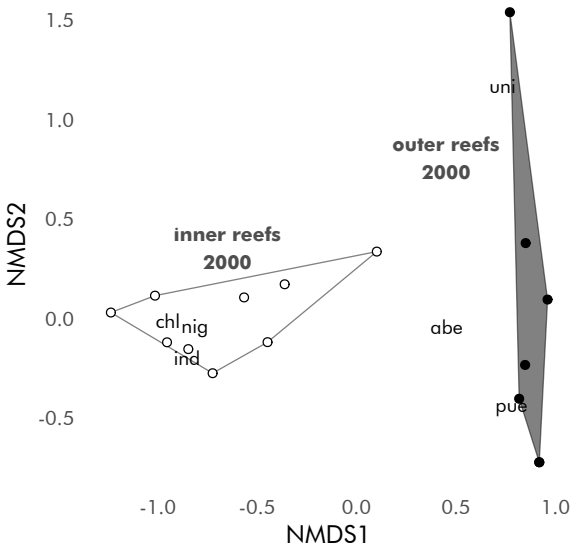
While the relative abundance of *H. chlorurus* had increased since the year 2000, it also appears that its density declined, implying that overall hamlet densities have declined. As mentioned above, Aguilar-Perera (2003) does not provide density estimates, only relative abundances, but densities of *H. chlorurus* in La Parguera have been estimated at 35 fish $1000m^{-2}$ of reef in 1988–1989 (McGehee, 1994), 18 ± 3.8 in 2005 (Bejarano and Appeldoorn, 2013) and 6.7 ± 0.6 in this study (mean $\pm$ S.E.). These estimates are to be interpreted with caution due to differences in methodologies and reefs surveyed between the three studies, but the magnitude and consistency of the decline suggests that they reflect a real trend.

Hypoplectrus chlorurus was significantly more abundant at shallow depths on the inner reefs in 2017, with a density of 4.5 ± 0.9 fish $1000m^{-2}$ of reef at $< 5.5m$ v. $2.1 \pm 0.6 at \geq 5.5m$ (mean $\pm$ S.E.). Interestingly, Aguilar-Perera (2003) did not find a significant difference in abundance of *H. chlorurus* between shallow ($< 5.5m$) and deep ($> 5.5m$) sections of the inner reefs using the same methodology. Further back in time, McGehee (1994) reports the opposite pattern in 1988–1989, with a significant increase in abundance of *H. chlorurus* with depth in La Parguera. Here again, the sequence from higher abundances of *H. chlorurus* in deeper waters in 1988–1989 (McGehee, 1994), no difference with depth in 2000 (Aguilar-Perera, 2003) and higher abundances at shallow depths in 2017 suggest that the relative increase of this species on the inner reefs has been accompanied by not only a decrease in density, but also a shift in depth distribution towards shallower depths.

Implications: One implication from these results is that it cannot be assumed that hamlet relative abundance and densities are stable over a temporal scale of < 20 years. Retrospectively, this does not come as a surprise given the dynamic nature of the coral reefs with which these fishes are tightly associated and even more so in the face of the large-scale anthropogenic influence on these ecosystems. Thus, data on hamlet relative abundances and densities need to be updated when possible and relevant. Another implication is that given the rarity and uncertain taxonomic status of some species, global databases on hamlet distributions and abundance are to be interpreted with caution (not to mention the pervasive occurrence of errors in such databases (Robertson, 2008)). Clearly, the distribution of e.g. the tan hamlet (*Hypoplectrus randallorum* Lobel 2011) depends on what is considered a tan hamlet in the first place (see below). Extensive surveys are required to capture rare species, which are commonplace in the hamlets as illustrated in the present study; for four species only one or two individuals were sighted over 35 transects

covering a total of $14000m^2$ across 16 reefs. The hamlets appear to be able to persist at low densities, which might be due to the fact that they are simultaneously hermaphroditic (implying that two individuals can reproduce regardless of their sex). Hamlets are also able to find conspecifics on the reef when present (Puebla et al., 2012a), which is confirmed here by the observation that three of the four rare species were sighted in pairs when found. Finally, given the data presented here, it is also possible that hamlet distributions might be dynamic at the regional scale over a few decades, which could have implications for the understanding of speciation in this group (population-centre hypothesis (Domeier, 1994)).

Ecology: What factors could have driven the observed temporal changes in the hamlet community? The simplest explanation that comes to mind is neutral, stochastic variation. Assuming ecological equivalence among species and individuals, relative species abundances are expected to fluctuate following a process analogous to genetic drift referred to as ecological drift, with the expectation that less abundant species have a higher probability of going extinct (Hubbell, 2001). Since *H. chlorurus* was already the most abundant species on the inner reefs in 2000, it is in principle plausible that stochastic fluctuations in the other, less abundant species led to their relative decline. Nevertheless, specific patterns in the distribution and abundance of hamlets from La Parguera suggest that the hamlet community is not behaving neutrally, but is at least in part shaped by ecological factors. The most notable of these patterns is the difference in densities and relative abundances between the inner and outer reefs, despite the fact that these are in close geographic proximity ($< 12km$). The inner reefs were clearly dominated by *H. chlorurus*, which represented 91.1% of all hamlets sighted. The outer reefs had in contrast a much more even species distribution (resulting in a higher effective number of species), with the most abundant species, *H. unicolor*, representing 37.5% of the hamlets sighted and *H. chlorurus* only 12.5%. Densities were also three times lower on the outer reefs than on the inner reefs, with an average of 3.6 ± 1.0 (mean $\pm$ S.E.) hamlets $1000m^{-2}$ of outer reef versus 10.5 ± 1.9 on the inner reefs all ham-



Suppl. Figure 2.2: NMDS outer reefs. Non-metric multidimensional scaling (NMDS) of *Hypoplectrus* spp. communities on 16 inner and outer reefs off La Parguera, Puerto Rico, for the year 2000 (Aguilar-Perera, 2003). Each circle represents one reef; species abbreviations indicate the direction (from the center of the plot) in which each species drives the community.

lets confounded. This difference between the inner and outer reefs appears to be temporally stable since Aguilar-Perera (2003) also reports differences in relative abundance and densities between the inner and outer reefs (Suppl. Fig. 2.2), with lower densities and fewer *H. chlorurus* on the outer reefs. In this context it is noted that the inner reefs are clearly structurally distinct from the outer reefs, justifying the decision to contrast these two types of reefs in both Aguilar-Perera (2003) and this study. The inner reefs are characterized by a shallow reef flat and a reef slope that goes down to ca. $14\ m$ while the outer reefs are exclusively deep (13–$18\ m$) and characterized by a spur-and-groove formation. The outer reefs are also clear blue reefs while the inner reefs are more turbid. In addition to differences in hamlet communities between the inner and outer reefs, the non-random depth distribution of *H. chlorurus* also suggests a role for ecology in shaping its distribution. If ecology drove the observed changes in the hamlet community from La Parguera, what ecological factors in particular might be involved? So far ecologists have failed to identify clear ecological differences among hamlets. Sympatric species are commonly found in the same habitat, often within sight of each other. Hamlets also have broadly similar diets as revealed by stomach-content and stable-isotope analysis in a variety of locations across the wider Caribbean, including La Parguera, except for *H. indigo* that appears to feed mostly on fishes (Randall 1967; Whiteman et al. 2007; Holt et al. 2008; G. M. Serviss, unpubl. data). This being said, broad differences in distribution and abundances between reef sections and types have been noted (Thresher, 1978; Fischer, 1980a; McGehee, 1994). Regarding *H. chlorurus* in particular, significant differences have been found in its distribution in Deep Water Cay (Grand Bahama) with higher abundances on the shallow *Acropora cerviconis* zone (G. M. Serviss, unpubl. data). This resonates with the observation that *H. chlorurus* was often (although not exclusively) found in association with *A. cervicornis* in La Parguera.

Turbiditiy: One intriguing point to be noted about *H. chlorurus* is that an extensive study on seawater turbidity and fish communities conducted between February and October 2005 on 21 reefs off La Parguera indicated that it is the only fish together with the sharknose goby *Elacatinus evelynae* (Böhlke & Robins 1968) that shows a positive association with turbidity, with higher abundances on more turbid reefs (Bejarano and Appeldoorn, 2013). In addition, sediment cores indicate that sedimentation of terrestrial origin has significantly increased over recent decades on the inner reefs in La Parguera (Ryan et al., 2008), providing a parallel between changes in turbidity regimes and the relative abundance of *H. chlorurus*. Ryan et al. (2008) also report that sedimentation rates are higher on the inner reefs ($0.47 \pm 0.02 cm year^{-1}$) than on the outer reefs ($0.19 \pm 0.01 cm year^{-1}$, mean $\pm$ S.D.), once more providing here a parallel between turbidity and relative abundances of *H. chlorurus* on the inner v. outer reefs. This pattern is consistent with the observation that the outer reefs were in clear, open blue waters while the inner reefs were noticeably murkier. It has been noted before that different hamlets tend to associate with different turbidity regimes (Thresher, 1978), but the hypothesis that water turbidity per se could constitute an important ecological factor in shaping hamlet communities has not been evaluated in

depth. If *H. chlorurus* responds positively to turbidity, an increase in turbidity on the inner reefs over recent decades would be expected to result in higher relative abundances of this species on these reefs, as observed here. Water turbidity is also negatively correlated with coral cover ($r^2 = 0.50$; (Bejarano and Appeldoorn, 2013). A decrease in coral cover associated with the increase in sedimentation rates on the inner reefs would, therefore, be expected to result in a decrease in hamlet densities, as observed here. One point to consider is seasonality since this survey was conducted during the dry season while Aguilar-Perera's survey as well as Bejarano & Appeldoorn's study on turbidity, were conducted over both the dry and rainy seasons. Relative abundances are not expected to vary seasonally at the reef level since the hamlets are reef-associated organisms and as such do not move between reefs after settlement (which would imply swimming over extensive non-reef areas). In addition, transect and tagging data over several seasons at other sites in Panama and Belize (Puebla et al., 2007, 2012a) indicate that hamlet communities do not change substantially at this time scale and that individuals are long-lived (several years) and quite sedentary. This being said, it is possible that the distribution of individuals within reefs might vary seasonally and the data on depth distribution are therefore to be considered with caution.

Aggressive Mimicry: Several hamlets including *H. chlorurus*, have been proposed as aggressive mimics, whereby the predatory hamlets (the putative mimics) gain an advantage in the approach and attack of prey by resembling and sometimes actively associating with other non-predatory fishes (the putative models; (Randall and Randall, 1960; Thresher, 1978; Puebla et al., 2007), a hypothesis that is still debated (Robertson, 2013). One prediction generated by this hypothesis is that the distribution of *H. chlorurus* would be expected to match the distribution of its putative model, the yellowtail damselfish *Microspathodon chrysurus* (Cuvier 1830) (Thresher, 1978). This does not appear to be the case in La Parguera since *M. chrysurus* shows a negative association with water turbidity (i.e. more abundant on less turbid reefs), the exact opposite of the pattern found for *H. chlorurus* (Bejarano and Appeldoorn, 2013). This lack of association between putative-model and mimic distributions does not necessarily invalidate the aggressive mimicry hypothesis, but it is noted that specific aggressive mimicry behaviours were not observed in hamlets from La Parguera during this survey.

Colour Pattern Variation: Hamlets from La Parguera form discrete phenotypic clusters that correspond to described species, but the taxonomic status of *H. nigricans* is still to be clarified. *Hypoplectrus nigricans* from La Parguera matches the description by Aguilar-Perera (2004) and resembles *H. nigricans* from Barbados (Puebla et al., 2008), but differs from *H. nigricans* from Panama, Belize and Mexico that is smaller, darker, with short and round pelvic fins (Aguilar-Perera, 2004; Puebla et al., 2008). This suggests that this nominal species might in fact constitute a species complex as proposed by Aguilar-Perera (2004); Puebla et al. (2008) and Lobel (2011). It is also noted that some individuals were tan coloured, but differed from the recently described tan hamlet *H. randallorum* in lacking the spots on the nose, at

the base of the pectoral fin and on the upper part of the caudal peduncle that are diagnostic of this species (Lobel, 2011). Additional data from other locations are needed to clarify the taxonomic status of *H. nigricans* and establish whether or not there is another tan hamlet species. In addition, a few individuals appeared intermediates between species, notably between *H. chlorurus* and *H. nigricans* and *H. chlorurus* and *H. puella*, respectively. Yet it is important to underscore that such individuals were rare, representing $< 2\%$ of all hamlets sighted within and outside of transects. Considering that hybrid pairings and spawnings have been observed in natural populations (Fischer, 1980a; Barreto and McCartney, 2008; Puebla et al., 2007, 2012a), that there do not appear to be intrinsic post-fertilization barriers between species in hamlets (Whiteman and Gage, 2007) and that in the only case where hybrids were bred in aquaria, they appeared intermediate between parental species (Domeier, 1994), it is plausible that such intermediate individuals might be hybrids.

Perspectives: The hypothesis that hamlet species might respond differentially to turbidity is intriguing and deserves further evaluation since turbidity correlates with a variety of ecological factors whose effect on hamlet communities need to be disentangled. One possibility is that hamlets might differ in their visual sensitivities, which is currently being tested using whole-genome analysis with a particular focus on opsin genes. Whole genomes will also allow testing whether individuals of intermediate appearance are actually hybrids, if so what type of hybrids (F 1 , F 2 , backcross, ...) and also clarify the taxonomic status of *H. nigricans*.

Acknowledgements

The authors are grateful to V. and C. McMillan, R. Pappa, C. Milton, O. Espinosa and the Magueyes Island Marine Laboratories staff for their assistance. This project was funded by grants from the Global Genome Initiative and Smithsonian Institute for Biodiversity Genomics, DFG and The Future Ocean Cluster of Excellence.

Chapter 2 References

Aguilar-Perera, A. (2003). Abundance and distribution of hamlets (Teleostei: *Hypoplectrus*) in coral reefs off southwestern Puerto Rico: support for the multiple-species hypothesis. *Caribbean Journal Of Science*, 39(1):147–151.

Aguilar-Perera, A. (2004). Variations in morphology and coloration in the black hamlet, *Hypoplectrus nigricans* (Teleostei: Serranidae). *Caribbean Journal Of Science*, 40(1):150–154.

Aguilar-Perera, A. and Gonzalez-Salas, C. (2010). Distribution of the genus *Hypoplectrus* (Teleostei: Serranidae) in the greater caribbean region: support for a color-based speciation. *Marine Ecology-An Evolutionary Perspective*, 31(2):375–387.

Anderson, M. (2001). A new method for non-parametric multivariate analysis of variance. *Austral Ecology*, 26(1):32–46.

Barlow, G. (1975). Sociobiology of some hermaphroditic serranid fishes, hamlets, in Puerto Rico. *Marine Biology*, 33(4):295–300.

Barreto, F. S. and McCartney, M. A. (2008). Extraordinary AFLP fingerprint similarity despite strong assortative mating between reef fish color morphospecies. *Evolution*, 62(1):226–233.

Bejarano, I. and Appeldoorn, R. S. (2013). Seawater turbidity and fish communities on coral reefs of Puerto Rico. *Marine Ecology Progress Series*, 474:217–226.

Bray, J. R. and Curtis, J. T. (1957). An ordination of the upland forest communities of southern Wisconsin. *Ecological Monographs*, 27(4):326–349.

Del Moral Flores, L. F., Tello-Musi, J. L., and MartÍnez-Pérez, J. A. (2011). Descripción de una nueva especie del género *Hypoplectrus* (Actinopterigy: Serranidae) del Sistema Arrecifal Veracruzano, suroeste del Golfo de México. *Revista De ZoologÍA*, pages 1–10.

Domeier, M. (1994). Speciation in the serranid fish *Hypoplectrus*. *Bulletin Of Marine Science*, 54(1):103–141.

Fischer, E. A. (1980a). Speciation in the hamlets (*Hypoplectrus*, Serranidae) - a continuing enigma. *Copeia*, (4):649–659.

Fischer, E. A. (1980b). The relationship between mating system and simultaneous hermaphroditism in the coral-reef fish, *Hypoplectrus nigricans* (Serranidae). *Animal Behaviour*, 28(MAY):620–633.

Fischer, E. A. (1981). Sexual allocation in a simultaneously hermaphroditic coral-reef fish. *American Naturalist*, 117(1):64–82.

Fischer, E. A. and Petersen, C. W. (1987). The evolution of sexual patterns in the seabasses. *Bioscience*, 37(7):482–489.

Hill, M. (1973). Diversity and evenness: a unifying notation and its consequences. *Ecology*, 54(2):427–432.

Holt, B. G., Cote, I. M., and Emerson, B. C. (2010). Signatures of speciation? Distribution and diversity of *Hypoplectrus* (Teleostei: Serranidae) colour morphotypes. *Global Ecology And Biogeography*, 19(4):432–441.

Holt, B. G., Cote, I. M., and Emerson, B. C. (2011). Searching for speciation genes: molecular evidence for selection associated with colour morphotypes in the caribbean reef fish genus *Hypoplectrus*. *Plos One*, 6(6).

Holt, B. G., Emerson, B. C., Newton, J., Gage, M. J. G., and Cote, I. M. (2008). Stable isotope analysis of the *Hypoplectrus* species complex reveals no evidence for dietary niche divergence. *Marine Ecology Progress Series*, 357:283–289.

Hubbell, S. (2001). *Monographs in population biology. The unified neutral theory of biodiversity and biogeography*, volume 32. Princeton University Press, Princeton.

Jost, L. (2006). Entropy and diversity. *Oikos*, 113(2):363–375.

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

McCartney, M. A., Acevedo, J., Heredia, C., Rico, C., Quenoville, B., Bermingham, E., and McMillan, W. O. (2003). Genetic mosaic in a marine species flock. *Molecular Ecology*, 12(11):2963–2973.

McGehee, M. (1994). Correspondence between assemblages of coral-reef fishes and gradients of water motion, depth, and substrate size off Puerto Rico. *Marine Ecology Progress Series*, 105(3):243–255.

Oksanen, J., Blanchet, F. G., Friendly, M., Kindt, R., Legendre, P., McGlinn, D., Minchin, P. R., O'hara, R. B., Simpson, G. L., Solymos, P., Stevens, M. H. H., Szoecs, E., and Wagner, H. (2017). *Vegan: community ecology package*.

Picq, S., McMillan, W. O., and Puebla, O. (2016). Population genomics of local adaptation versus speciation in coral reef fishes (*Hypoplectrus* spp, Serranidae). *Ecology And Evolution*, 6(7):2109–2124.

Puebla, O., Bermingham, E., and 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(6):1405–1415.

Puebla, O., Bermingham, E., and Guichard, F. (2009). Estimating dispersal from genetic isolation by distance in a coral reef fish (*Hypoplectrus puella*). *Ecology*, 90(11):3087–3098.

Puebla, O., Bermingham, E., and Guichard, F. (2011). Perspective: matching, mate choice, and speciation. *Integrative And Comparative Biology*, 51(3):485–491.

Puebla, O., Bermingham, E., and Guichard, F. (2012a). Pairing dynamics and the origin of species. *Proceedings Of The Royal Society B-Biological Sciences*, 279(1731):1085–1092.

Puebla, O., Bermingham, E., Guichard, F., and Whiteman, E. A. (2007). Colour pattern as a single trait driving speciation in *Hypoplectrus* coral reef fishes? *Proceedings Of The Royal Society B-Biological Sciences*, 274(1615):1265–1271.

Puebla, O., Bermingham, E., and McMillan, W. O. (2012b). On the spatial scale of dispersal in coral reef fishes. *Molecular Ecology*, 21(23):5675–5688.

Puebla, O., Bermingham, E., and McMillan, W. O. (2014). Genomic atolls of differentiation in coral reef fishes (*Hypoplectrus* spp., Serranidae). *Molecular Ecology*, 23(21):5291–5303.

Randall, J. E. (1967). Food habits of reef fishes of the West Indies. *Studies In Tropical Oceanography*, (5):655–847.

Randall, J. E. (1968). *Caribbean reef fishes*, volume 1. T.f.h. Publications, Neptune City, Nj.

Randall, J. E. and Randall, H. A. (1960). Examples of mimicry and protective resemblance in tropical marine fishes. *Bulletin Of Marine Science*, 10(4):444–480.

Robertson, D. (2008). Global biogeographical data bases on marine fishes: caveat emptor. *Diversity And Distributions*, 14(6):891–892.

Robertson, D. (2013). Who resembles whom? Mimetic and coincidental lookalikes among tropical reef fishes. *Plos One*, 8(1).

Ryan, K. E., Walsh, J. P., Corbett, D. R., and Winter, A. (2008). A record of recent change in terrestrial sedimentation in a coral-reef environment, La Parguera,

Puerto Rico: a response to coastal development? *Marine Pollution Bulletin*, 56(6):1177–1183.

Tavera, J. J. and Acero, A. P. (2013). Description of a new species of *Hypoplectrus* (Perciformes: Serrnidae) from the southern Gulf of Mexico. *Aqua, Journal Of Ichthyology*, 19(1):29–38.

Theodosiou, L., McMillan, W. O., and Puebla, O. (2016). Recombination in the eggs and sperm in a simultaneously hermaphroditic vertebrate. *Proceedings Of The Royal Society B-Biological Sciences*, 283(1844).

Thresher, R. E. (1978). Polymorphism, mimicry, and evolution of hamlets (*Hypoplectrus*, Serranidae). *Bulletin Of Marine Science*, 28(2):345–353.

Victor, B. (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*, 5:2–19.

Whiteman, E. A., Cote, I. M., and Reynolds, J. D. (2007). Ecological differences between hamlet (*Hypoplectrus*: Serranidae) colour morphs: between-morph variation in diet. *Journal Of Fish Biology*, 71(1):235–244.

Whiteman, E. A. and Gage, M. J. G. (2007). No barriers to fertilization between sympatric colour morphs in the marine species flock *Hypoplectrus* (Serranidae). *Journal Of Zoology*, 272(3):305–310.

– 3 –

Inter-chromosomal Coupling between Vision and Pigmentation Genes during Genomic Divergence

Kosmas Hench [1], Marta Vargas [2], Marc P. Höppner [3], W. Owen McMillan [2], Oscar Puebla [1,2,4]

1 GEOMAR Helmholtz Centre for Ocean Research Kiel, Evolutionary Ecology of Marine Fishes, Düsternbrooker Weg 20, 24105 Kiel, Germany

2 Smithsonian Tropical Research Institute, Apartado Postal 0843-03092, Panamá, República de Panamá

3 Institute of Clinical Molecular Biology, Christian Albrechts University of Kiel, Kiel, Germany

4 University of Kiel, Faculty of Mathematics and Natural Sciences, Christian-Albrechts-Platz 4, 24118 Kiel, Germany

This study was originally published in *Nature Ecology & Evolution*. Figures were re-drawn for this thesis but no additional changes were made. The re-print within this thesis is in agreement with *Springer Nature* as the original publication is under Creative Commons CC BY license.

Original publication

Hench, K. *et al*. (2019). Inter-chromosomal coupling between vision and pigmentation genes during genomic divergence. *Nature Ecology & Evolution*, 3(4):657–667. doi: 10.1038/s41559-019-0814-5.

Abstract

The evolution of linkage disequilibrium between genes underlying mate choice and ecological traits is thought to be a fundamental step in the course of speciation with gene flow. Here, we capture this process in the hamlets, a group of closely related reef fishes from the wider Caribbean that differ essentially in colour pattern and are reproductively isolated through strong visually-based assortative mating. Using full-genome analysis, we identify four narrow genomic intervals that are consistently differentiated among sympatric species in a backdrop of extremely low genomic divergence. These four intervals include genes involved in pigmentation *(sox10)*, axial patterning *(hoxc13a)*, photoreceptor development *(casz1)* and visual sensitivity (*SWS* and *LWS* opsins), respectively, that develop islands of long-distance and inter-chromosomal linkage disequilibrium as species diverge. The relatively simple genomic architecture of species differences allows linkage disequilibrium to be maintained in the presence of gene flow.

Keywords: speciation, linkage disequilibrium, vision, colour pattern, hamlets, *Hypoplectrus*.

3.1. Introduction

How new species may arise and persist in the presence of gene flow is a fundamental and unresolved question in our understanding of the origins of biological diversity. This issue is particularly relevant in the ocean, where physical barriers are often poorly defined and pelagic larvae provide potential for extensive gene flow, but which nonetheless harbours some of the most diverse communities on earth (Palumbi, 1994). Indeed, diversity on coral reefs rivals the diversity seen in tropical forests (Reaka-Kudla, 1997) and coral reef fish communities are among the most species-rich assemblages of vertebrates. The origin of coral reef fish families and functional groups dates back to the Paleocene (66 Mya); however, the vast majority of species arose within the last 5.3 My, with closely related species often differing primarily with respect to colour and patterning (Bellwood et al., 2017).

The hamlets (*Hypoplectrus* spp., Serranidae), a complex of 18 closely related reef fishes from the wider Caribbean (Figure 3.1 a), provide an excellent context to explore speciation in the sea. Hamlets differ most notably in colour pattern, a trait that has been suggested to have direct ecological implications in terms of crypsis (Thresher, 1978; Fischer, 1980) and mimicry (Randall and Randall, 1960; Thresher, 1978; Puebla et al., 2007, 2008). Additionally, colour pattern plays a central role for reproductive isolation in this complex. Individuals mate assortatively with respect to colour pattern (Fischer, 1980; Barreto and McCartney, 2008; Puebla et al., 2007, 2012) and it has been experimentally established that mate choice is driven by visual cues (Domeier, 1994). Nonetheless, spawnings between different species are observed at a low frequency ($< 2\%$) in natural populations (Fischer, 1980; Puebla et al., 2007; Barreto and McCartney, 2008; Puebla et al., 2012). Larvae from inter-specific crosses grow and develop normally (Whiteman and Gage, 2007); the ones that have been raised

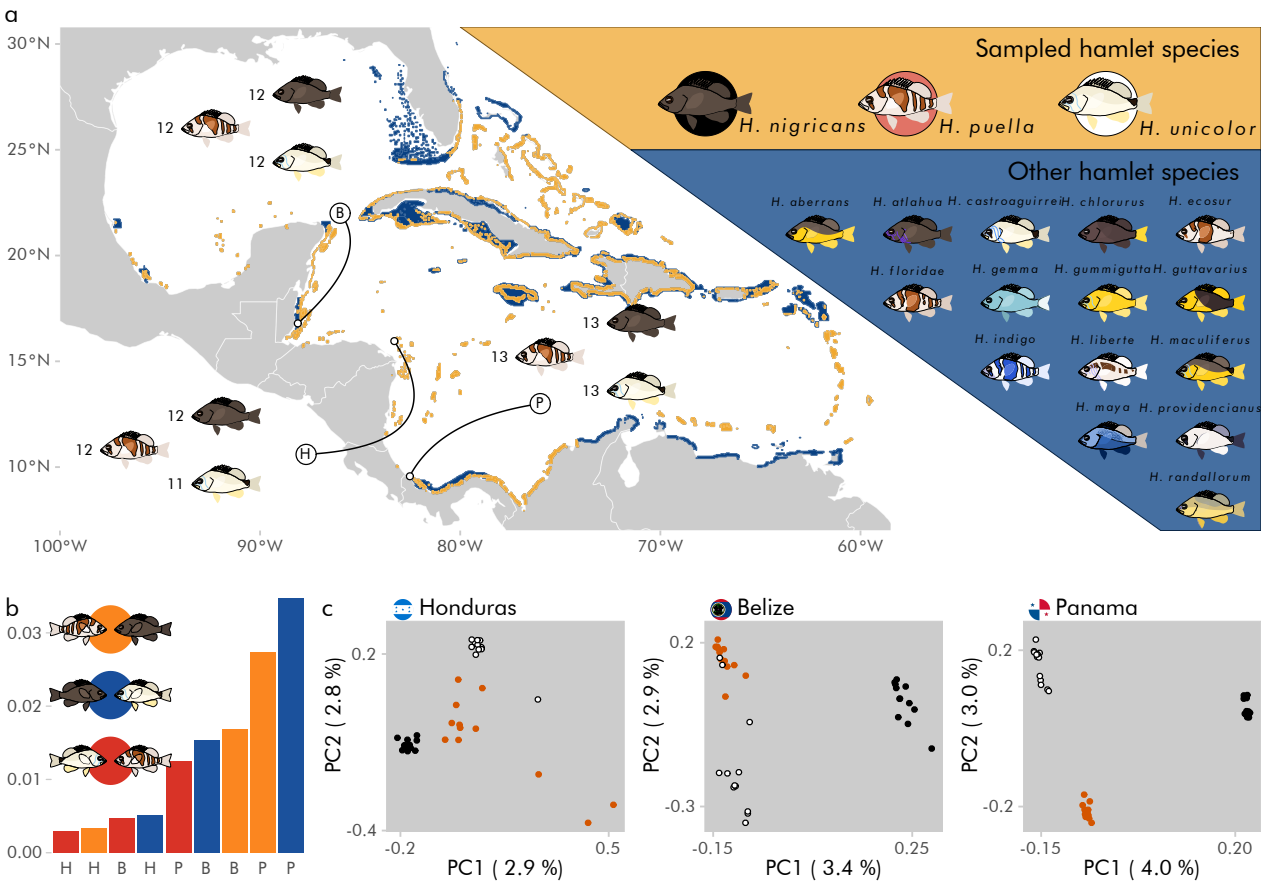


Figure 3.1: Sampling design and whole-genome population genetic patterns. **a**, three sympatric species from three locations (B: Belize, H: Honduras, P: Panama) were targeted for resequencing. The area of sympatry among the three sampled species is highlighted in orange, and the distribution of the whole genus in blue (Robertson and Tassell, 2015). The three sampled species are the most common and widely distributed, but they represent just a fraction of the full hamlet diversity. Numbers indicate the sample size. **b**, F_{ST} estimates among pairs of sympatric species, in order of increasing F_{ST}. Colors indicate the species pair and labels on the x axis the location (B: Belize, H: Honduras, P: Panama). **c**, Principal Component Analysis (PCA) within each location. Genomic data from a total of 8,247,395 SNPs.

past the juvenile stage showed intermediate colour pattern phenotypes (Domeier, 1994), and individuals with intermediate phenotypes are also observed in natural populations at a low frequency (Puebla et al., 2008). Patterns of genetic divergence among species indicate that the radiation encompasses the entire range of genomic divergence (referred to as the *speciation continuum* (Seehausen et al., 2014)), from species that are nearly genetically indistinguishable (Puebla et al., 2007; Barreto and McCartney, 2008; Puebla et al., 2012; McCartney et al., 2003; Puebla et al., 2014) to those that are well-diverged (Victor, 2012; Tavera and Acero, 2013). There is extensive sympatry among hamlet species, with up to nine species co-occurring on Caribbean reefs (Puebla et al., 2012) with a high degree of overlap in feeding ecology and habitat (Whiteman et al., 2007; Holt et al., 2008).

Here we focus on the lower end of the speciation continuum and examine patterns of genomic divergence among the three most abundant, widespread, and genetically similar hamlets - the black hamlet (*H. nigricans*), the barred hamlet (*H. puella*)

and the butter hamlet (*H. unicolor*) (Figure 3.1a). We take advantage of their extensive and overlapping distributions to sample the three species in three reef systems in Panama, Honduras and Belize. This sampling design provides the opportunity to identify the genomic regions that are consistently differentiated among sympatric species across locations. Furthermore, microsatellite and RAD-seq data from the same species and locations indicate that levels of genetic differentiation among sympatric species are similar to the levels of differentiation among populations within species (Puebla et al., 2008, 2014; Picq et al., 2016), providing the opportunity to contrast between-species and between-population genetic architectures.

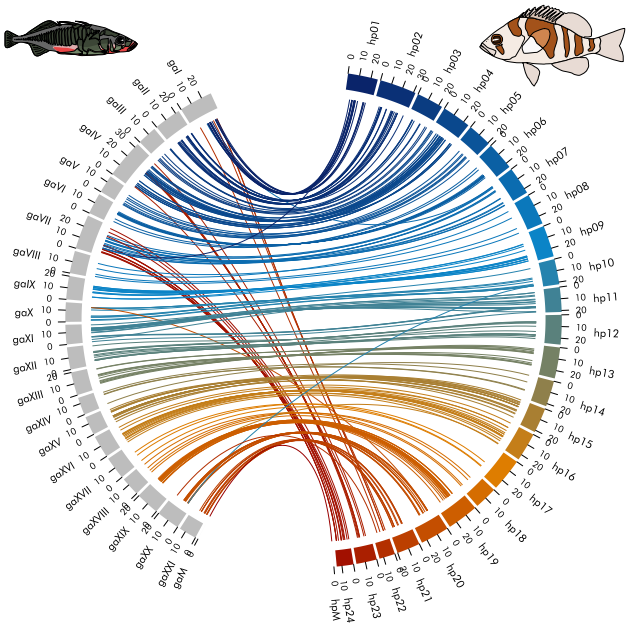
Given the slight genetic differences among species and the link between colour pattern, natural selection and mate choice, we made two predictions regarding genome-wide patterns of differentiation and divergence among the three species. First, we predicted that regions showing elevated and consistent differentiation between species would contain loci with strong functional links to either the development or the perception of colour pattern. Second, we reasoned that linkage disequilibrium (LD, the non-random association of alleles at different loci) among these regions would develop as species diverge. Our second prediction derives from an influential theoretical paper by Felsenstein (Felsenstein, 1981) who identified recombination between loci underlying mate choice and ecological traits as a major evolutionary force acting against speciation with gene flow, with the corollary that the evolution of linkage disequilibrium (LD) between such loci is a fundamental step in the origin of species (Felsenstein, 1981). Empirical studies have shown that pleiotropy or physical linkage provide a direct way to generate associations between mate choice and ecology (Hawthorne and Via, 2001; Kronforst et al., 2006; Bay et al., 2017), but it remains unclear whether and how long-distance or inter-chromosomal linkage disequilibrium (ILD) between loci underlying mate choice and ecological traits may develop in the presence of gene flow (Seehausen et al., 2014).

3.2. Results and Discussion

Genome Assembly and Resequencing

To test these hypotheses, we assembled a reference genome for the hamlets. We used a combination of *Illumina* ($245\times$) and *PacBio* ($10\times$) data to assemble scaffolds, which were then anchored to a high-density linkage map that includes 24 linkage groups (LGs) (Theodosiou et al., 2016), matching the 24 chromosomes expected in serranids (Arai, 2011). The resulting assembly was 612 Mb long with ninety-two percent of scaffolds anchored to the linkage map, resulting in a super-scaffold n50 of 24 Mb. We annotated 27,469 genes using a combination of *ab initio* gene predictions and RNAseq data from a variety of tissue types. Overall, there was broad synteny between the hamlet genome and the genome of the most closely related species with a similar high-quality genome, the three-spined stickleback (*Gasterosteus aculeatus*, Suppl. Fig. 3.1).

Whole-genome analysis of 110 individuals (Figure 3.1a) confirmed the striking genetic



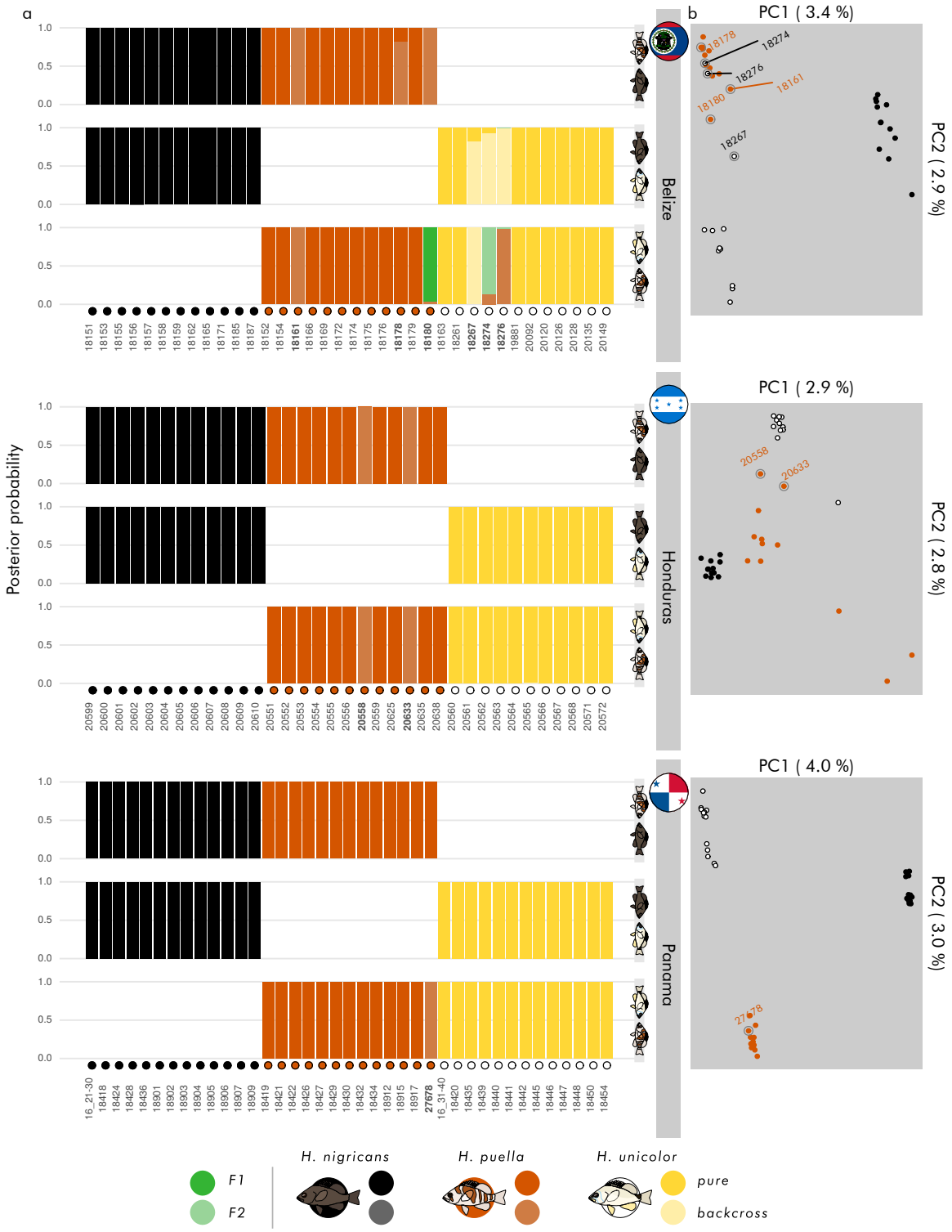
Suppl. Figure 3.1: Broad-scale synteny between the hamlet and stickleback genomes. The comparison is based on a whole-genome alignment using last. Only alignments > 5000 bp are shown. Left: stickleback (*Gasterosteus aculeatus*, 21 LGs & mitochondria). Right: hamlet (*Hypoplectrus puella*, 24 LGs & mitochondria).

similarity among species and revealed differences in patterns of genetic differentiation among species in the three locations. Pairwise F_{ST} among sympatric species ranged from 0.003 between *H. puella* and *H. unicolor* in Honduras to 0.035 between *H. unicolor* and *H. nigricans* in Panama (Figure 3.1b). In all three locations, Principal Component Analysis (PCA) clustered individuals by species; however, overall genetic differentiation among species showed differences among locations and was lowest in Honduras (F_{ST} among the three species $= 0.004$), intermediate in Belize ($F_{ST} = 0.012$) and highest in Panama ($F_{ST} = 0.025$, Figure 3.1c). PCA also suggested that some individuals might be of hybrid origin (e.g. the two butter hamlets from Belize that clustered with barred hamlets). This hypothesis was corroborated by additional analyses based on Mendelian inheritance patterns of a small subset of highly differentiated SNPs. A total of eight high-probability hybrids or backcrosses were identified out of the 110 samples (five in Belize, two in Honduras and one in Panama, Suppl. Fig. 3.2), establishing that gene flow is ongoing among species.

Similar to previous studies in other taxa (Malinsky et al., 2015; Vijay et al., 2016; Van Belleghem et al., 2017), differentiation was highly heterogeneous across the genome in local comparisons (Figure 3.2), and a similar pattern was also evident when considering Genotype x Phenotype ($G \times P$) association (Suppl. Fig. 3.3). Notably, a large section of LG08 exhibited generally elevated levels of differentiation in all local comparisons. This may be explained by low levels of recombination along this LG (Suppl. Fig. 3.4), which might harbour a large chromosomal inversion. Nevertheless, patterns of differentiation in LG08 were not entirely consistent across locations or species, resulting in relatively weak differentiation in our global comparisons where samples were pooled across locations (Figure 3.2, top panel).

Vision and Pigmentation Genes

In contrast to the pattern in LG08, four small (50-100 kb) genomic intervals were strongly and consistently differentiated among species,



Suppl. Figure 3.2: Identification of putative backcrosses and hybrids. a, bars indicate the posterior probability of assignment to the different hybrid classes for each pairwise comparison. A total of nine individuals, highlighted in bold, were identified as putative hybrids or backcrosses, eight of which with high (>0.99) posterior probabilities (five in Belize, two in Honduras and one in Panama). Five of these involved butter hamlets (*H. unicolor*) from Belize, which is consistent with the fact that this species is rare in this location. The same individuals were often identified as putative hybrids or backcrosses in two pairwise comparisons, suggesting either multi-species exchanges or false positives in one species pair resulting from exchanges with a third species. **b**, putative hybrids and backcrosses (highlighted) were also intermediate (e.g. 18267) or clustered with other species (e.g. 18274) in the whole-genome PCA. Note that the sampling design explicitly excluded individuals with intermediate colour patterns, thereby reducing the probability of recovering F1 hybrids in our data set.

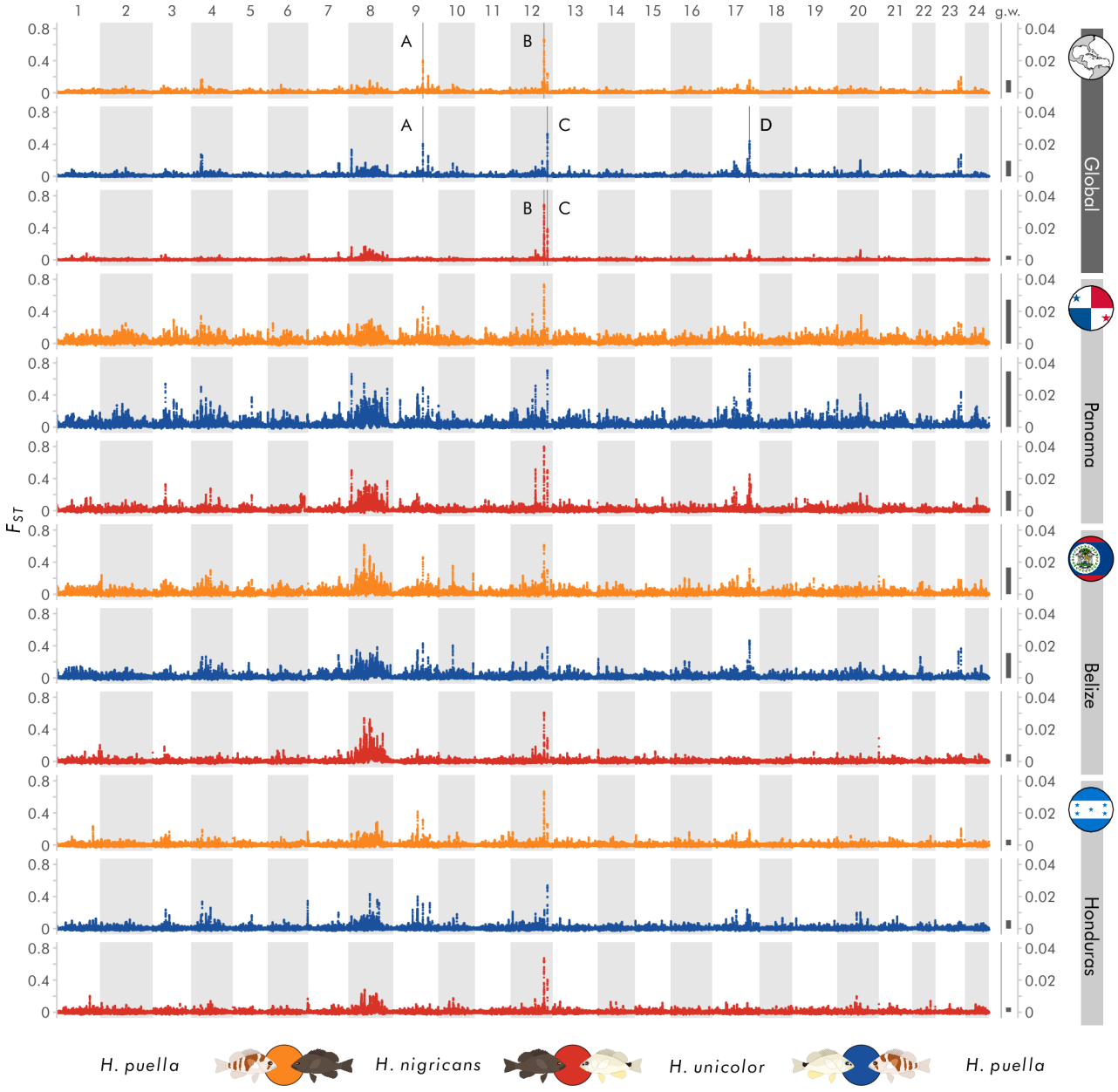
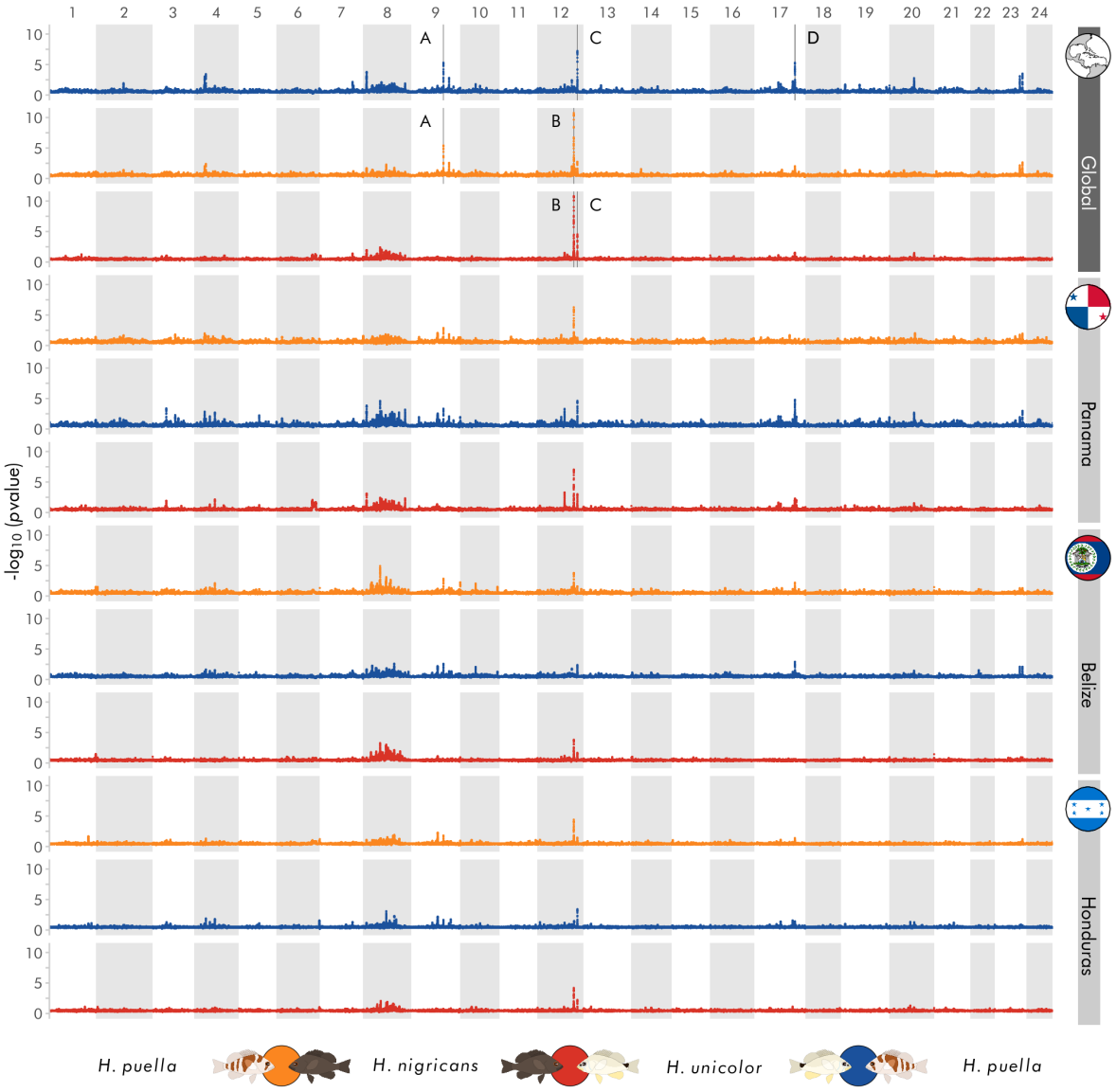


Figure 3.2: Patterns of genomic differentiation among black (*H. nigricans*), barred (*H. puella*) and butter (*H. unicolor*) hamlets. The alternating white and grey blocks represent the 24 linkage groups (LGs). Each species pair is represented by one colour, pooled across locations (Global) and within each location (Belize, Honduras & Panama). F_{ST} values correspond to the weighted mean per 50 kb window with 5 kb increments. Vertical bars on the right indicate the genome-wide weighted mean F_{ST} (note the different scale). The four genomic intervals above the 99.98^{th} F_{ST} percentile in the global comparison are highlighted with a vertical line.

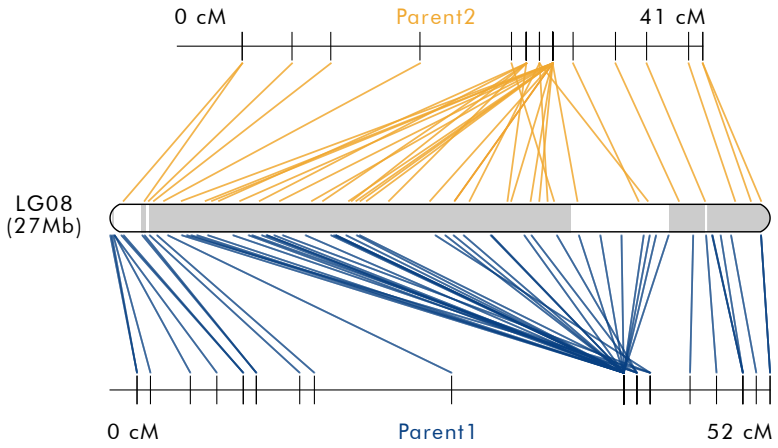
forming sharp "genomic islands" (Turner et al., 2005) that stood out above the 99.98^{th} percentile in the global comparisons considering either F_{ST} (Figure 3.2) or $G \times P$ association (Suppl. Fig. 3.3). In agreement with our first hypothesis, each contained at least one candidate gene with a strong functional connection to either the development of colour pattern or sensory processes involved in pattern perception.

The sharp peak on LG09 (A in Figure 3.2) contained *sox10* (Figure 3.3a). This gene encodes a transcription factor that has been



Suppl. Figure 3.3: Genotype by phenotype (G × P) association among black (*H. nigricans*), barred (*H. puella*) and butter (*H. unicolor*) hamlets. Each species pair is represented by one colour, pooled across locations (Global) as well as within each location (Belize, Honduras & Panama). The *p* values are from the linear model with Wald test, transformed using the negative of the common logarithm and averaged across 50 kb window with 5 kb increments ($\overline{(-log_{10}(p))}$). The four genomic regions highlighted with a vertical line, included as reference, correspond to the four intervals identified in Figure 3.2

shown to be involved in the development of melanophores in zebrafish (Dutton et al., 2001; Elworthy et al., 2003). The role of this gene in melanisation is consistent with the finding of strong differentiation at this locus between the melanic species (*H. nigricans*) and the other two non-melanic species (*H. puella* and *H. unicolor*). Similarly, a strongly differentiated interval on LG12 (C in Figure 3.2) was centred on the *hoxca* gene cluster (Figure 3.3c). This region was identified in a previous genome scan using RAD sequencing (Puebla et al., 2014), but our new reference genome allowed us to localise the interval far more precisely. *Hox* genes code for homeodomain-containing transcription factors that play a central role in the patterning of tissues along the body axis, with 3′



Suppl. Figure 3.4: Large low-recombining region on linkage group 08 (LG08). Top and bottom: linkage maps of the two parents used for the F_1 cross, from(Theodosiou et al., 2016). Middle: assembled linkage group 08. Lines connect individual RAD markers that are identified in both the linkage maps and the assembly. Gray and white blocks represent individual scaffolds. A large number of RAD markers that are in close proximity on the linkage maps are distributed over a wide region on the assembled linkage group, providing direct evidence of low recombination in this region

genes expressed anteriorly and 5′ genes posteriorly (Carroll et al., 2005). *Hox* genes can also be involved in the development of colour pattern phenotype. They have for example been shown to play a role in the regulation of body pigmentation in birds (Poelstra et al., 2015) and *Drosophila* (Jeong et al., 2006), as well as in eyespot formation on butterfly wings (Saenko et al., 2011). The strongest F_{ST} signal was positioned on *hoxc13a* specifically, the most 5′ gene of the *hoxca* cluster. This gene is known to be expressed in the caudal peduncle and at the pigment appearance stage in fishes (Thummel et al., 2004; Jakovlić and Wang, 2016). Again, the specific role of this locus in patterning is consistent with pattern differences among hamlet species. This interval strongly differentiated *H. unicolor*, which has a prominent dark saddle on the caudal peduncle, from the other two species that lack this pattern. The possibility that *hox* genes may be involved in the development of colour pattern differences at a very shallow phylogenetic level in the hamlets is intriguing and may provide an opportunity to better understand the links between micro- and macro-evolutionary processes.

The remaining two highly differentiated genomic intervals contained candidate loci with strong functional connections to vision. One of these two intervals was on LG12 (B in Figure 3.2) and fell in an apparently non-genic region upstream of *casz1* that strongly differentiated *H. puella* from the other two uniformly coloured species (Figure 3.3b). *casz1* is a castor zinc finger transcription factor involved in a number of processes through development, including the development of photoreceptors (Mattar et al., 2015). Given that the visual system grows continuously in teleost fishes, we examined RNA expression in the retinas of 24 adult black, barred and butter hamlets from Panama. We confirmed that *casz1* is consistently and strongly expressed in the retina, and also identified two splice variants of *casz1* that extend the coding region across a large part of this peak (Figure 3.3b). The other interval, on LG17 (D in Figure 3.2), contained a cluster of short- and long-wave sensitive opsin genes

(*sws2a*β, *sws2a*α, *sws2b* & *lws*, Figure 3.3d) that play a key role in the fine-tuning of visual sensitivity (Yokoyama, 2008). Unlike the previous intervals, which each differentiated a particular species from the other two, differentiation at this interval was not clearly species-specific. It was strongest in the comparison between the melanic (*H. nigricans*) and white (*H. unicolor*) species, where it presented a peak-valley-peak pattern that may reflect parallel adaptation from standing genetic variation (Bierne, 2010; Roesti et al., 2014).

These four highly differentiated genomic intervals were narrow and our highlighted candidate genes were not selectively picked from a large set of loci: the first peak on LG12 (B) contained only *casz1*, the second one (C) contained only *hox* genes except for the *calcoco1* locus and was centred on *hoxc13a* specifically, and the peak on LG09 (A) contained only two genes, *sox10* and *rnaseh2a* (Figure 3.3). The last highly differentiated interval on LG17 (D) contained more genes, but the peak-valley peak pattern was centred on the opsin genes specifically, with *sws2b* in the valleyand *sws2a*β, *sws2a*α & *lws* in the two flanking peaks (Figure 3.3). In line with (Feder and Nosil, 2010), simulations indicate that a combination of large effective

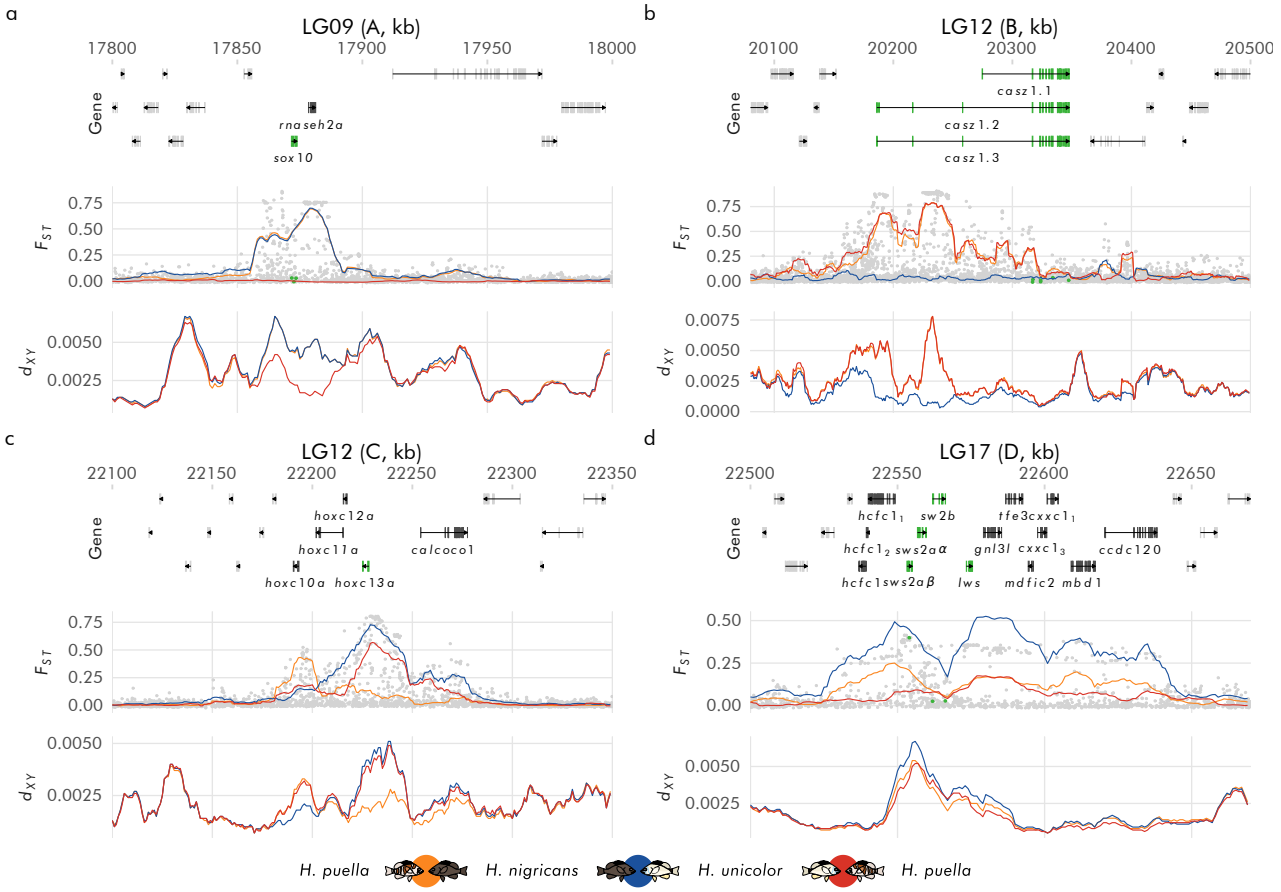
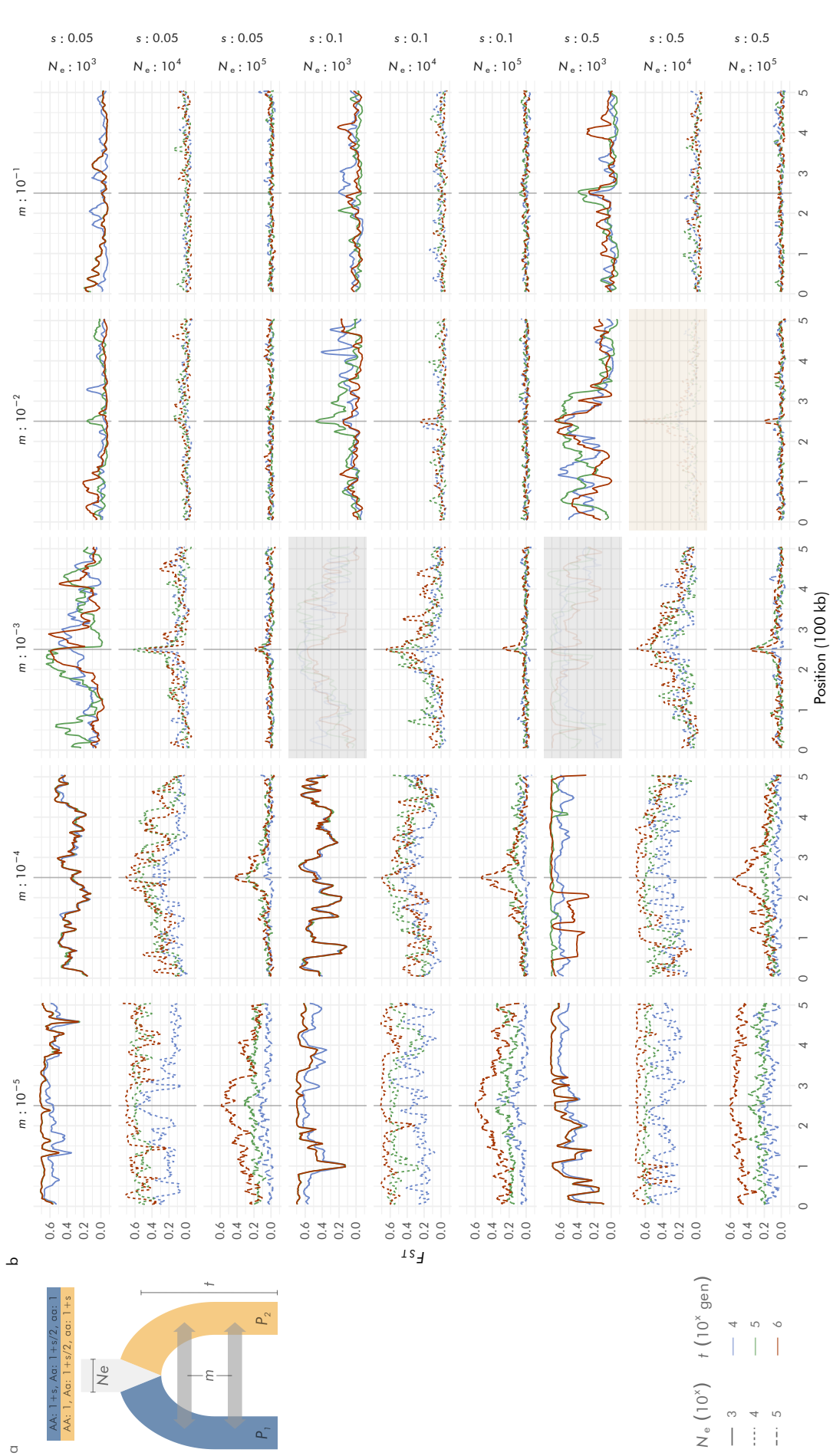
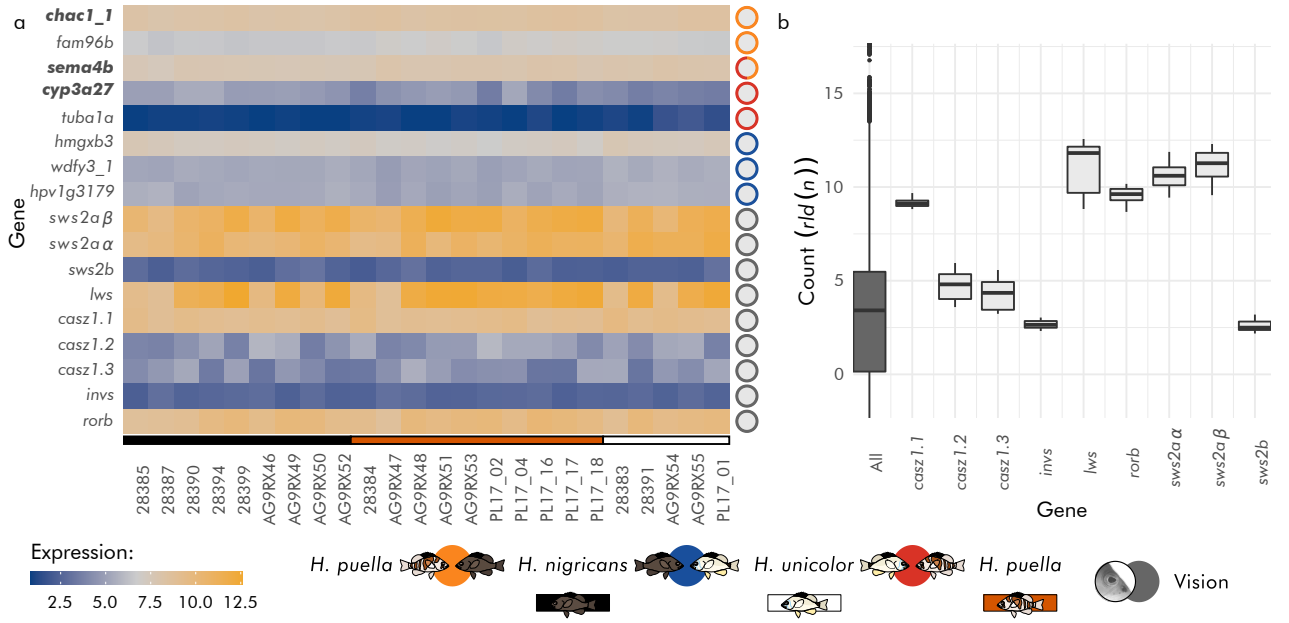


Figure 3.3: Close-up on the candidate intervals. The four panels (**a-d**) correspond to the four intervals above the 99.98^{th} F_{ST} percentile (A-D). Each panel shows the LG, genomic position, annotation, F_{ST} and d_{XY}. For clarity only the genes in high F_{ST} regions are labelled, with candidate genes and non-synonymous SNPs within these genes highlighted in green. Coloured lines correspond to pairwise species comparisons (weighted mean, 10 kb window, 1 kb increments) and dots to global F_{ST} among the three species on a SNP basis. All comparisons with species samples pooled across the three locations.



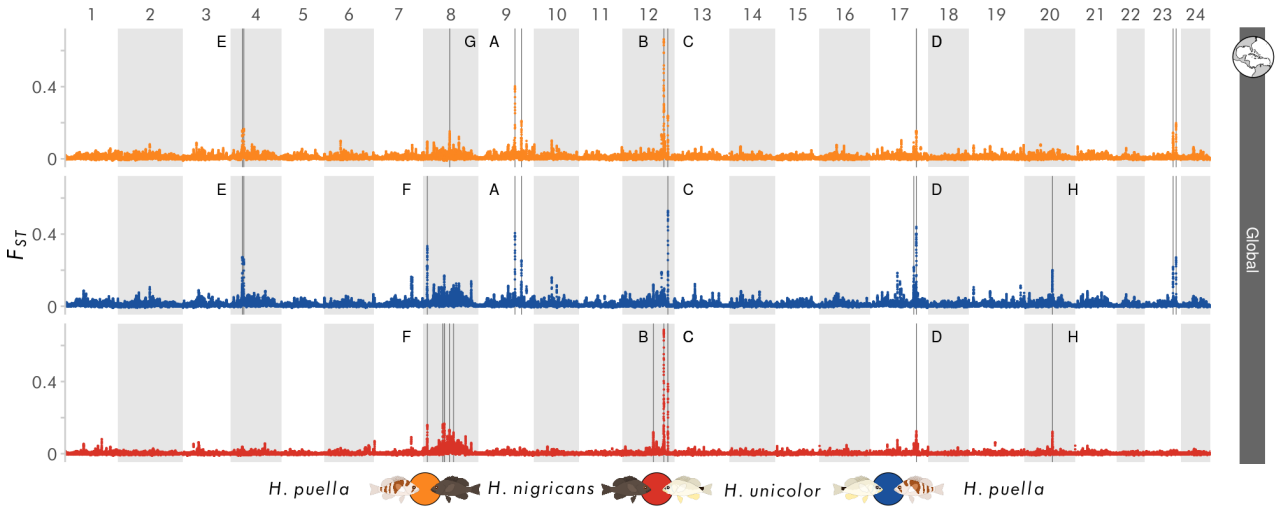
Suppl. Figure 3.5: Extent of differentiation under various simulated scenarios. The demographic history underlying the simulations consisted of two populations (1 & 2) of constant size N_e that split t generations ago and experienced constant and symmetrical migration (m) since then. A selected site was considered in the middle of a 500-kb chromosome, consisting of a codominant locus with two alleles A and a that are advantageous in population 1 and 2, respectively, with a fitness of $1+s$ for homozygotes and $1+s/2$ for heterozygotes where s is the selection coefficient. Scenarios highlighted in grey are similar to those explored by Charlesworth et al., 1997. Patterns similar to the ones observed in this study (highlighted in orange) were obtained with larger N_e (10,000) and higher m (0.01), which we suggest may be more representative of the situation in the hamlets.



Suppl. Figure 3.6: Gene expression in the retinal tissue. **a**, only three genes, highlighted in bold, were significantly differentially expressed among species. Genes included in the figure correspond, from top to bottom, to the three most differentially expressed genes for each species pair (labelled by the colour-coded rings on the right) followed by the candidate genes related to vision identified in this study (grey rings, not differentially expressed). Data from 10 adult barred, 9 black and 5 butter hamlets from Panama (labelled by the colour bar on the bottom). **b**, albeit not differentially expressed, many of the candidate genes showed consistent above-average expression levels. Expression data transformed with regularised logarithm in both panels.

population size (N_e = 10,000), intermediate migration rate (m = 0.01) and strong selection (s = 0.1-0.5) may generate sharp peaks of differentiation as observed in the hamlets (Suppl. Fig. 3.5). It is noteworthy that all but one of the diverged SNPs in the four regions are either in non-coding regions or synonymous, suggesting that species differences are mainly driven by regulatory mechanisms. The only exception was one diverged non-synonymous SNP on *sws2a*β that corresponds to the bovine rhodopsin amino acid 200. Although not a known spectral tuning site, the location of this amino acid suggests that it might possibly be involved in spectral tuning (Yokoyama, 2008). We also note that only three genes (*chac1_1*, *sema4b* and *cyp3a27*) showed significant differences in expression among species in the retinal tissue (Suppl. Fig. 3.6), yet our methodology does not allow to capture differences in expression that may occur during development, in specific light environments (e.g. at dusk at the time of spawning) or in specific cell types.

Additional vision and pigmentation genes were identified by extending our analyses to the genomic regions above the 99.90^{th} F_{ST} percentile that presented weaker or less consistent differentiation among species. This less stringent selection identified 14 additional intervals across seven LGs (Suppl. Fig. 3.7, Suppl. Fig. 3.8, Suppl. Tab. 3.1), four of which contained further vision or pigmentation genes (Suppl. Fig. 3.9). *ednrb* on LG04 (E in Suppl. Fig. 3.7) is involved in zebrafish melanophore and iridophore development (Parichy et al., 2000) and again differentiated *H. nigricans* from the other non-melanic species. One interval on LG08 (F in Suppl. Fig. 3.7) presented



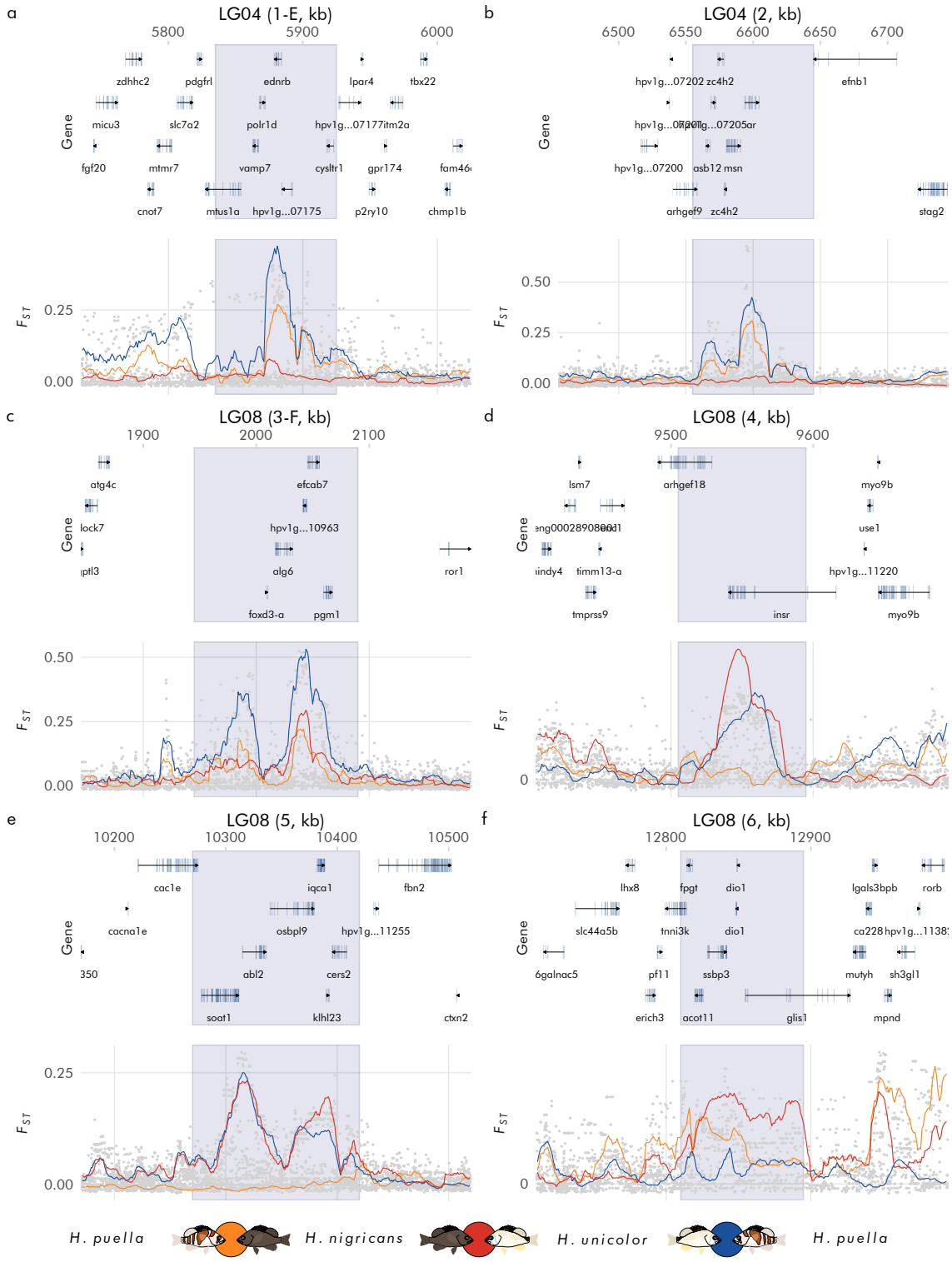
Suppl. Figure 3.7: Genomic intervals above the 99.90^{th} F_{ST} percentile. F_{ST} values were estimated as the weighted mean per 50 kb window with 5 kb increments, considering all locations for each species. A total of 19 genomic intervals above the 99.90^{th} F_{ST} percentile, highlighted with a vertical bar, were identified (Suppl. Tab. 3.1). Intervals containing candidate genes are labelled with capital letters. Peaks A-D are also above the 99.98^{th} F_{ST} percentile, peaks E-H are not.

a non species-specific peak-valley-peak pattern centred on *foxd3*, a transcription factor also involved in melanophore differentiation in zebrafish (Curran et al., 2009). A further interval on LG08 (G in Suppl. Fig. 3.7) included *rorb*, which plays a critical role during photoreceptor differentiation in mice (Jia et al., 2009). Similar to *casz1*, the other gene involved in photoreceptor development, *rorb* singled out *H. puella* and was consistently and strongly expressed in the retina (Suppl. Fig. 3.6). Finally, *invs* on LG20 (H in Suppl. Fig. 3.7) is involved in the transport of opsins into the outer segment of photoreceptors (Zhao and Malicki, 2011).

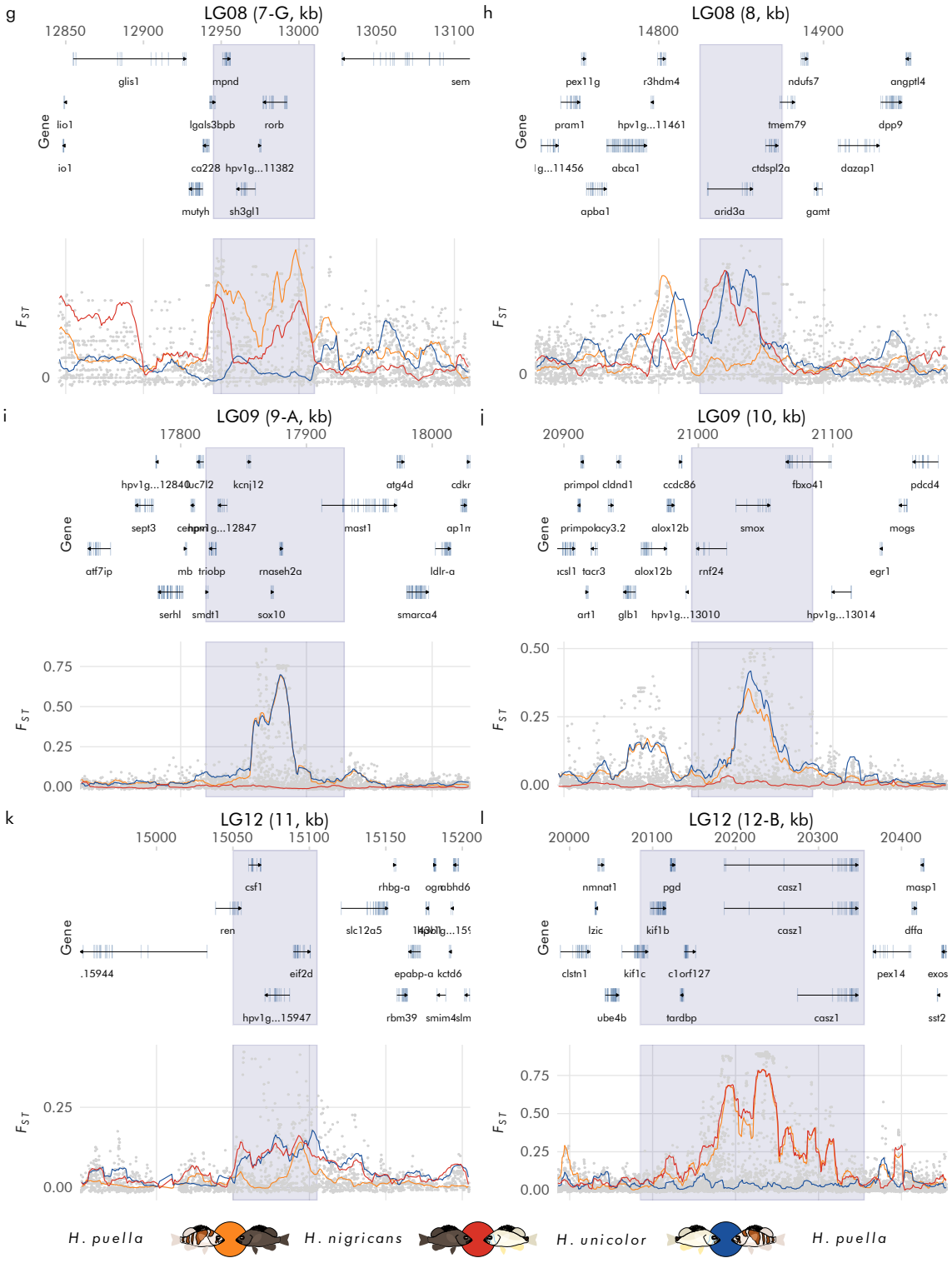
Long-Distance LD and Barrier Genes

The four intervals that showed marked differentiation in our genome-wide comparison were either on different LGs or 2 Mb apart on the same LG (B and C), which is well beyond physical linkage in the hamlets (Suppl. Fig. 3.10d). The four candidate intervals are therefore not physically linked. Nevertheless, in line with our second hypothesis, these intervals showed islands of elevated long-distance and inter-chromosomal LD in a backdrop of nearly zero genome-wide ILD (Suppl. Fig. 3.4a,b). In addition, there was a buildup of ILD with increasing genome-wide differentiation, with weakest ILD in Honduras, intermediate in Belize and most pronounced in Panama (Suppl. Fig. 3.4c,d). As expected, there was no ILD among these intervals within species (Suppl. Fig. 3.4e). The same patterns were observed when considering the four additional vision and pigmentation genes above the 99.90^{th} F_{ST} percentile (Suppl. Fig. 3.11).

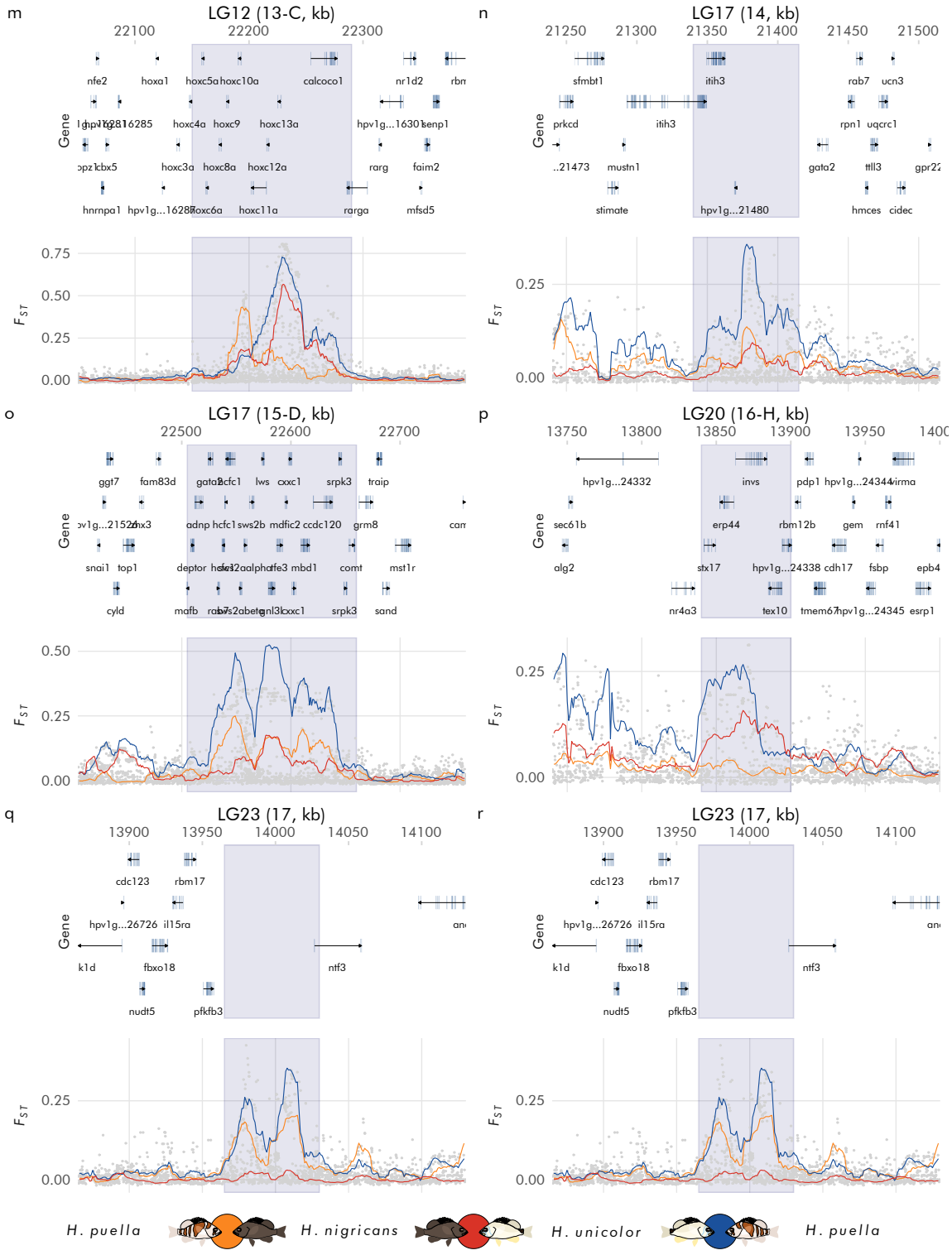
Local regions of strong differentiation can arise for a number of reasons (Ravinet et al., 2017; Wolf and Ellegren, 2017), including processes unrelated to speciation such as background selection (Charlesworth et al., 1993) or the sorting of ancestral polymorphisms (Guerrero and Hahn, 2017). These



Suppl. Figure 3.8: Close-up on all the intervals above the 99.90^{th} F_{ST} percentile The panels (**a-r**, figure continued below) correspond to the 18 intervals above the 99.90^{th} F_{ST} percentile. From top to bottom, each panel includes the respective linkage group (LG) and interval ID, the position on the LG, the gene model annotation and F_{ST} values. Gene models include the extent and direction of genes as well as exon boundaries. The F_{ST} plots show the pairwise comparisons among species (lines, weighted mean per 10 kb window with 1 kb increments). Additionally, the global F_{ST} values among the three species are shown as dots on a SNP basis. All comparisons with species samples pooled across the three locations. The highlighted area corresponds to the whole intervals as defined in Suppl. Tab. 3.1.



Suppl. Figure 3.8: (continued I) Close-up on all the intervals above the 99.90^{th} F_{ST} percentile The panels (**a-r**, figure continued below) correspond to the 18 intervals above the 99.90^{th} F_{ST} percentile. From top to bottom, each panel includes the respective linkage group (LG) and interval ID, the position on the LG, the gene model annotation and F_{ST} values. Gene models include the extent and direction of genes as well as exon boundaries. The F_{ST} plots show the pairwise comparisons among species (lines, weighted mean per 10 kb window with 1 kb increments). Additionally, the global F_{ST} values among the three species are shown as dots on a SNP basis. All comparisons with species samples pooled across the three locations. The highlighted area corresponds to the whole intervals as defined in Suppl. Tab. 3.1.



Suppl. Figure 3.8: (continued II) Close-up on all the intervals above the 99.90^{th} F_{ST} percentile The panels (**a-r**) correspond to the 18 intervals above the 99.90^{th} F_{ST} percentile. From top to bottom, each panel includes the respective linkage group (LG) and interval ID, the position on the LG, the gene model annotation and F_{ST} values. Gene models include the extent and direction of genes as well as exon boundaries. The F_{ST} plots show the pairwise comparisons among species (lines, weighted mean per 10 kb window with 1 kb increments). Additionally, the global F_{ST} values among the three species are shown as dots on a SNP basis. All comparisons with species samples pooled across the three locations. The highlighted area corresponds to the whole intervals as defined in Suppl. Tab. 3.1.

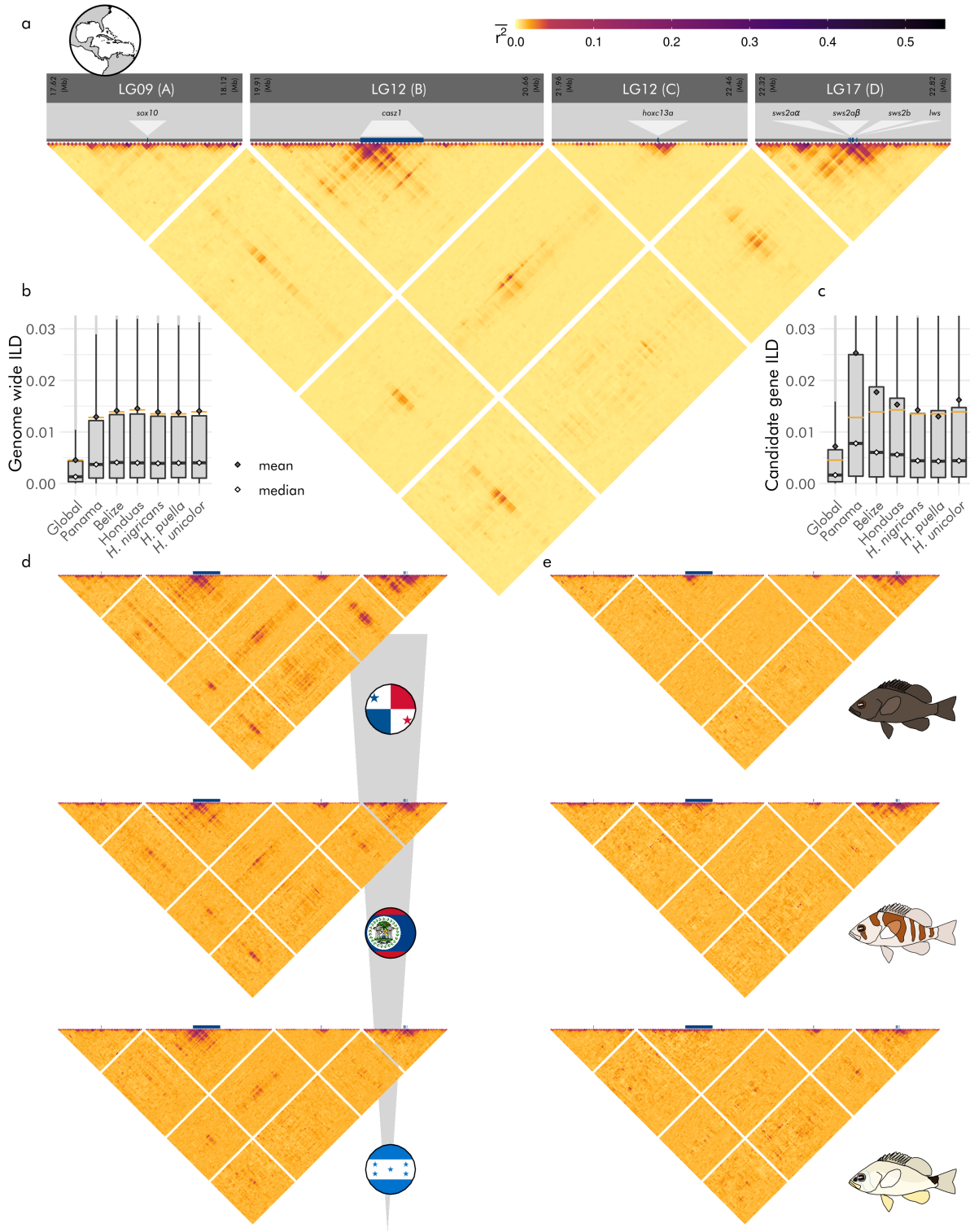
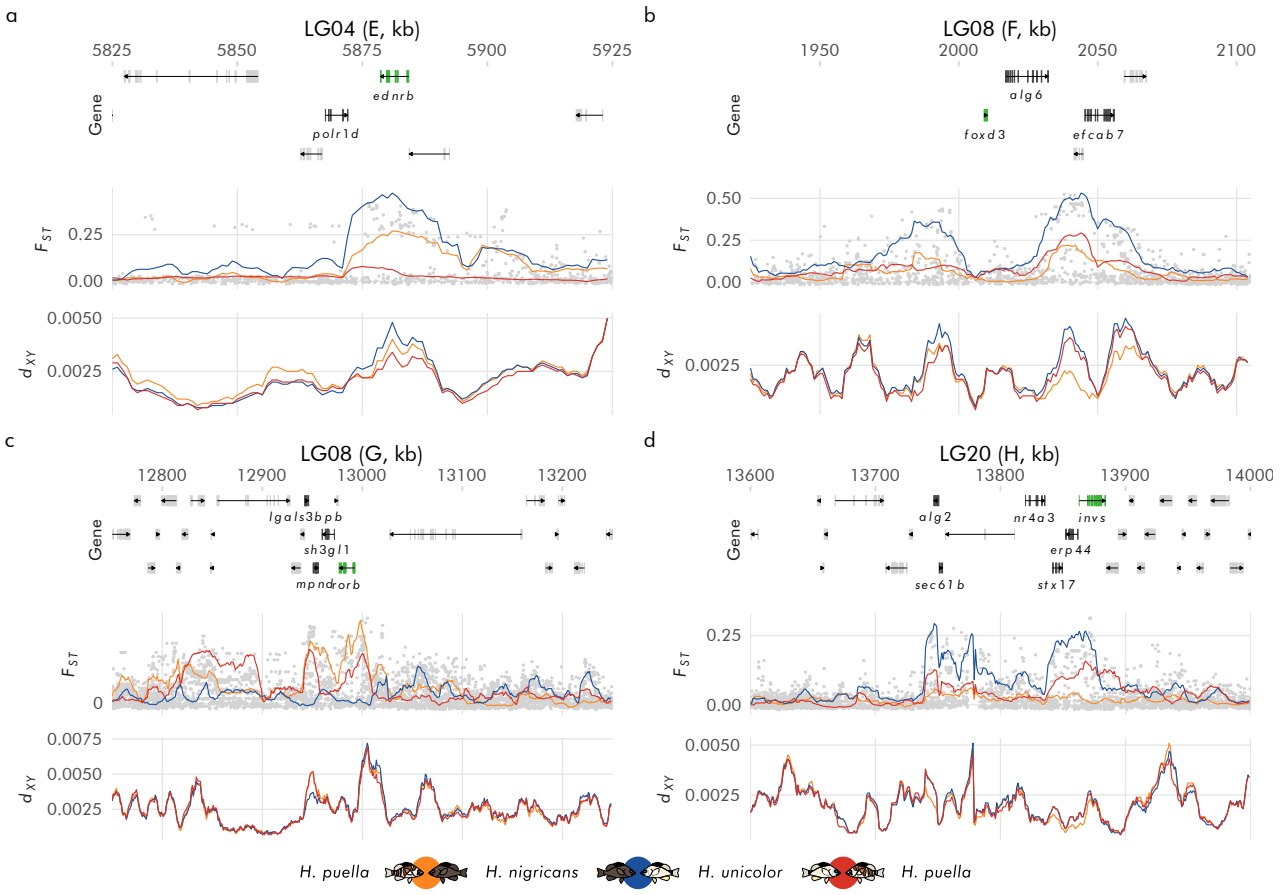


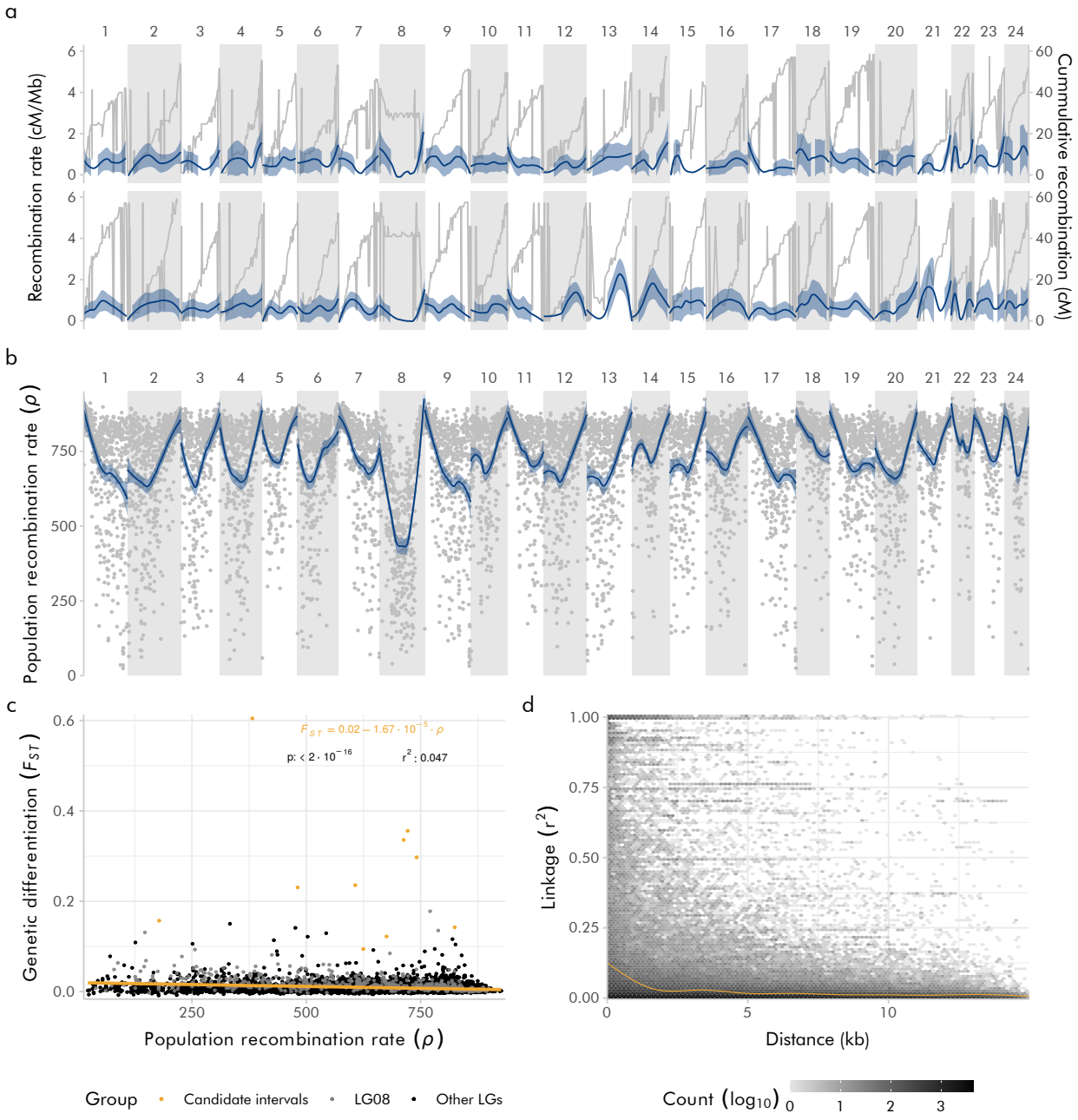
Figure 3.4: Long-distance and inter-chromosomal linkage disequilibrium (r^2) among the four candidate intervals. **a**, the four intervals displayed islands of increased LD. **b**, genome-wide inter-chromosomal LD (ILD). Boxes: $25^{th} - 75^{th}$ percentile interval, whiskers: $1.5 \times$ interquartile range, dots: outliers, red lines: r^2 expectation for the mean ($= 1/2n$ where *n* is sample size). Genome-wide ILD was lower for the global data set due to the larger sample size (n=110) compared to the location- and species-specific data sets (n=35-39). **c**, ILD among the four candidate intervals. **d**, LD among the four intervals increased with increasing differentiation among species (grey gradient). **e**, in contrast, LD among the four intervals was low or absent within species. r^2 values are shown on a SNP basis in **b**, **c** and averaged over two-dimensional bins of $10 \times 10\ kb$ in **a**, **d**, **e**.



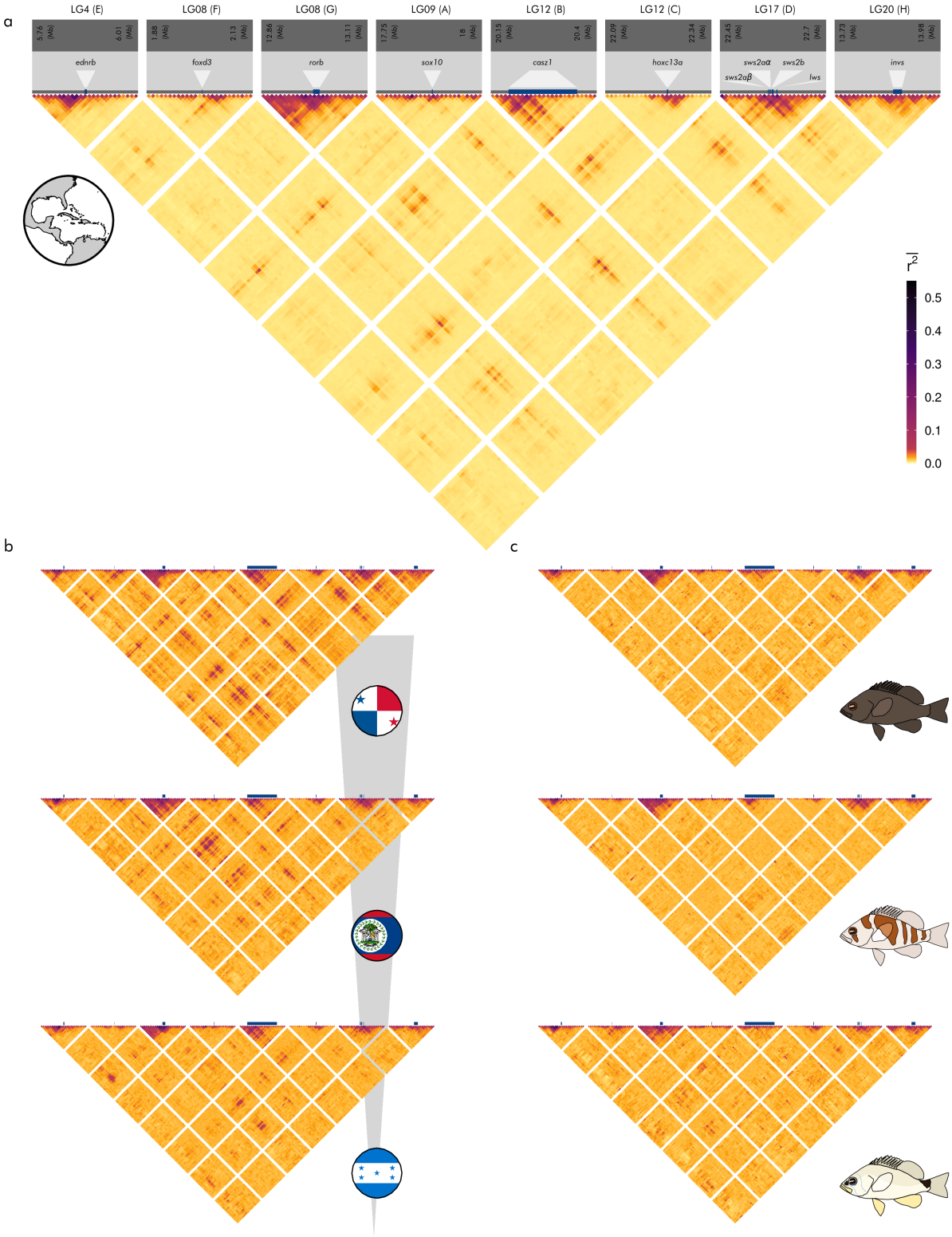
Suppl. Figure 3.9: Close-up on the four additional intervals containing candidate vision and pigmentation genes. The four panels (**a-d**) correspond to the four additional intervals above the 99.90^{th} F_{ST} percentile (but not above the 99.98^{th}) that include candidate vision and pigmentation genes. From top to bottom, each panel includes the respective linkage group (LG), the position on the LG, the gene model annotation, F_{ST} and d_{XY}. Gene models include the extent and direction of genes as well as exon boundaries. For clarity only the genes in high F_{ST} intervals are labelled, and candidate genes are highlighted in green. The F_{ST} plots show the pairwise comparisons among species (lines, weighted mean per 10 kb window with 1 kb increments). Additionally, the global F_{ST} values among the three species are shown as dots on a SNP basis. The d_{XY} values are also averaged over 10 kb windows with 1 kb increments. All comparisons with species samples pooled across the three locations.

processes are almost certainly operating within hamlet species. Indeed, we see an expected buildup of genetic differentiation across a large region on LG08 with very low recombination. This region may be a large chromosomal inversion and is exceptional as it contains six of the 14 intervals that showed moderate levels of differentiation among species. Nonetheless, the sharp erosion in overall levels of differentiation among species in our global comparisons coupled with the elevated differentiation among populations of the same species (Suppl. Fig. 3.12) suggest that this region does not contain loci that are essential for the maintenance of species differences and that, if it does contain an inversion, it is polymorphic both within and between species.

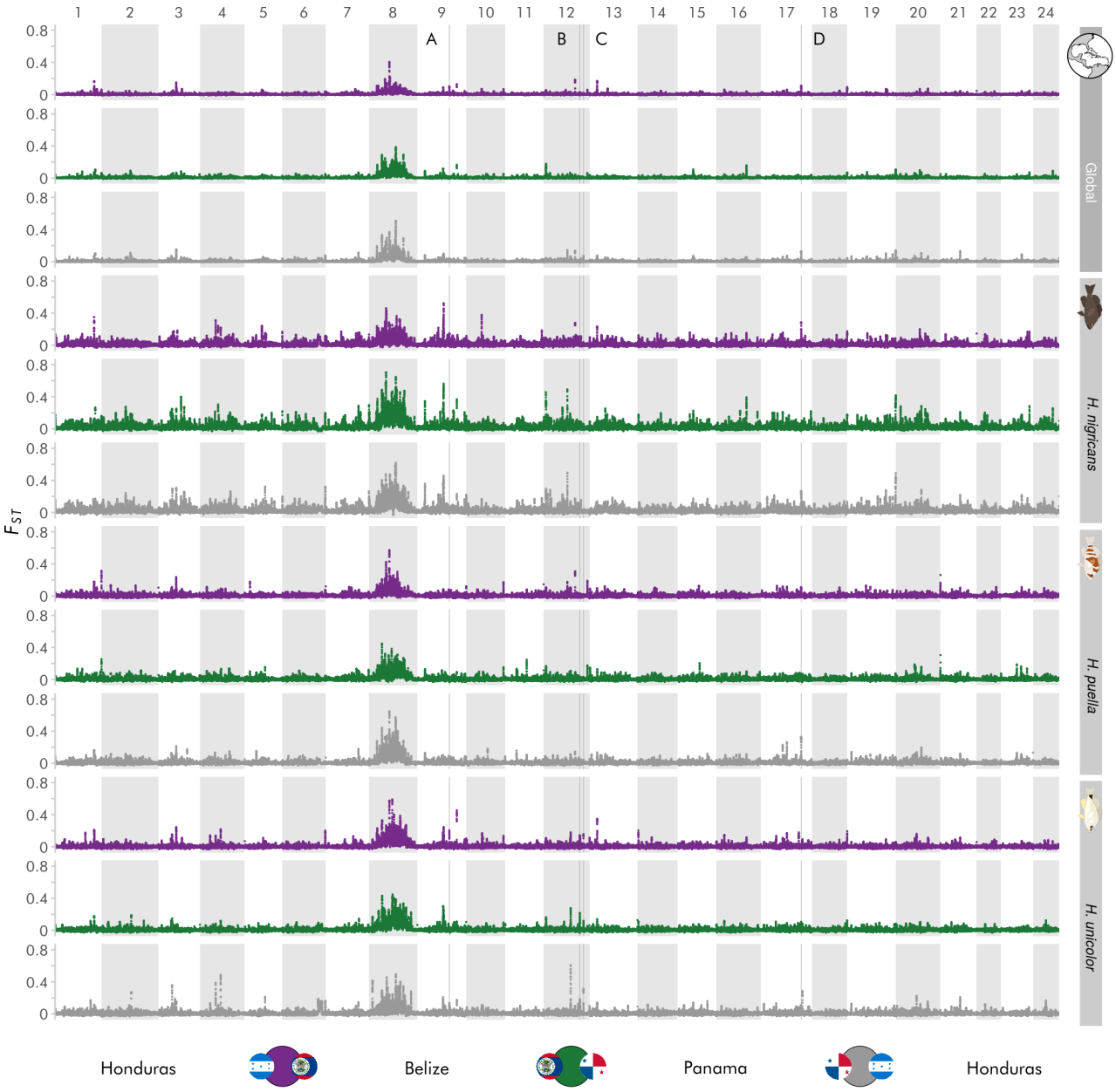
In contrast, there are a number of compelling reasons to argue that the four intervals that showed much stronger and consistent differentiation among species do contain the loci responsible for reproductive isolation. Foremost, all contained genes involved in vision, pigmentation or patterning



Suppl. Figure 3.10: Genome-wide recombination patterns. **a**, recombination landscape inferred from the combination of a within-species cross and the genome assembly. Recombination rate was inferred by mapping the linkage map(Theodosiou et al., 2016) markers onto the genome assembly and dividing linkage (cM) by physical distance (Mb). **b**, recombination landscape inferred from population genomic data considering all species and locations. As expected due to the different data sets considered, the two types of recombination maps differ substantially. Yet both identify a large low-recombining region in LG08. **c**, correlation between population recombination rate and genetic differentiation among the three species considering non-overlapping 50-kb windows. As expected, a negative relationship is observed. The correlation and regression slope are nevertheless weak, indicating that recombination does not have a strong impact on differentiation at this stage of genomic divergence. Red dots correspond to the 50 kb windows that are within our four candidate regions; these windows do not show particularly low recombination rates. **d**, decay in linkage disequilibrium with physical distance, estimated over 20 randomly placed 15 kb windows. The shading of the hexagonal bins indicates the log_{10} count for each combination of distance and r^2 values. The red lines indicates a smoothing spline (gam, cubic regression spline) of the original data. Physical linkage decays rapidly within 2 kb.

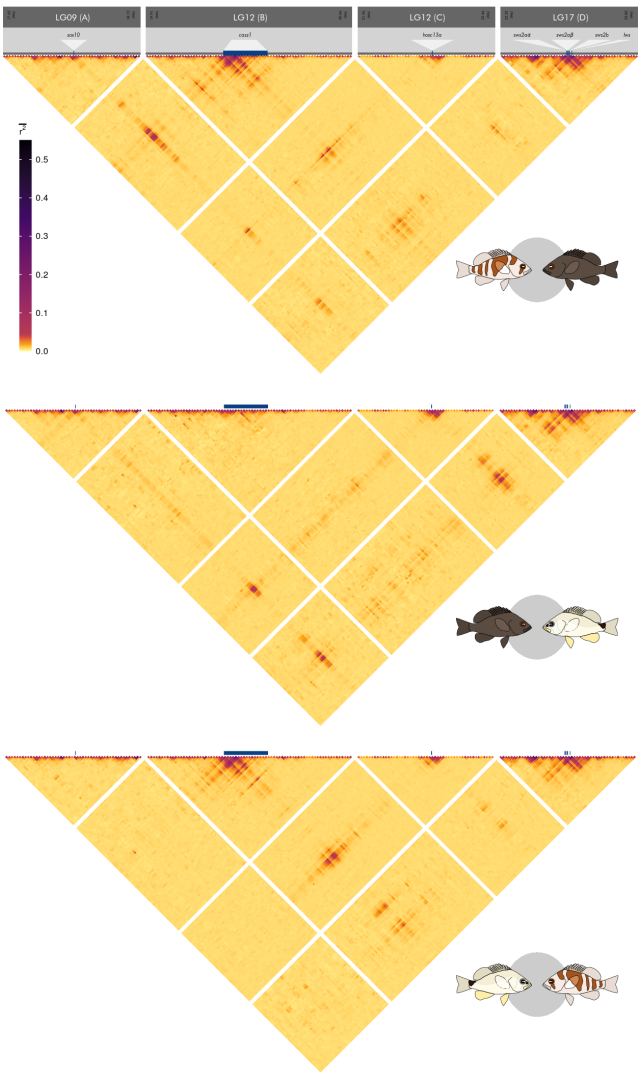


Suppl. Figure 3.11: Long-distance and inter-chromosomal linkage disequilibrium (LD) among the eight intervals containing candidate vision and pigmentation genes. **a**, the intervals identified in Figure 3.2 & Suppl. Fig. 3.7 displayed increased long-distance and inter-chromosomal LD. LD was calculated between individual SNP pairs and averaged over $10\ kb \times 10\ kb$ areas. **b**, LD among the eight intervals ordered by increasing differentiation among the three species (indicated by gray gradient, Figure 3.1b,c, Suppl. Tab. 3.6). **c**, in contrast, LD was very low or absent within each of the three species.



Suppl. Figure 3.12: Patterns of genomic differentiation among hamlets from Belize, Honduras and Panama. The alternating white and grey blocks represent the 24 linkage groups (LGs). Each population comparison is represented by one colour, pooled across species (Global) as well as within each species (*H. nigricans*, *H. puella* & *H. unicolor*). F_{ST} values were estimated as the weighted mean per 50 kb window with 5 kb increments. The four genomic regions highlighted with a vertical line, included as reference, correspond to the four intervals identified in Figure 3.2.

in vertebrates, fitting our initial expectations about the types of loci involved in speciation based on the ecology and reproductive biology of these species. This pattern is even more compelling when considering that variation at the candidate loci for pigmentation (*sox10*) and patterning (*hoxc13a*) parallels the specific colour pattern differences that characterise hamlets (melanisation and marking on the caudal peduncle). Moreover, our sampling design permits us to isolate genomic intervals that are consistently differentiated among species across locations, effectively filtering out processes acting within populations,

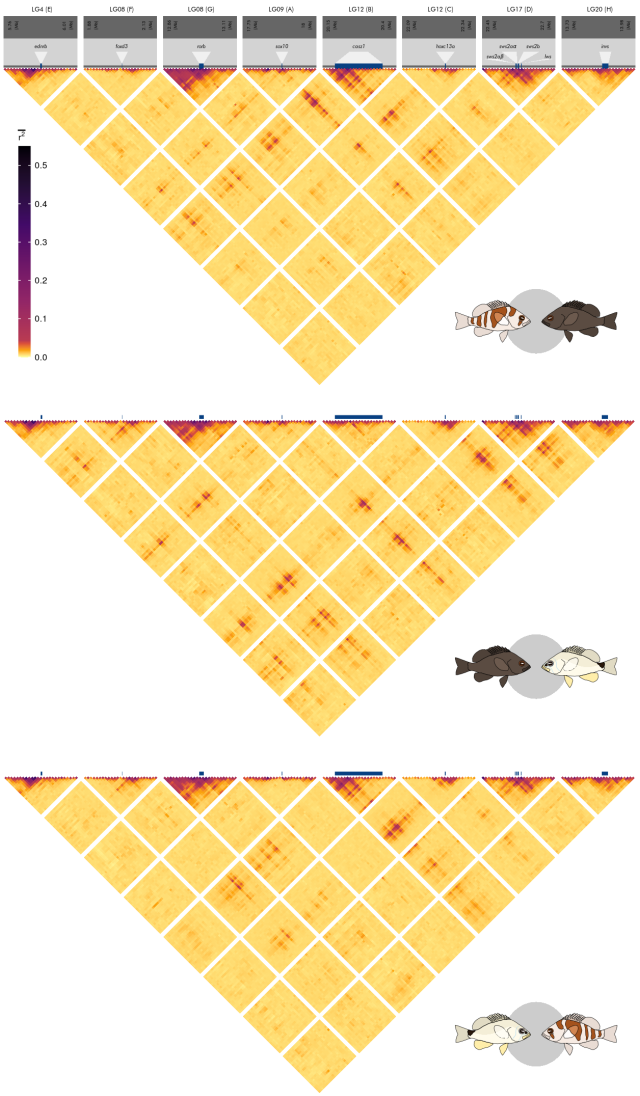


Suppl. Figure 3.13: Long-distance and inter-chromosomal linkage disequilibrium (LD) among the four candidate intervals for each species pair. The intervals identified in Figure 3.2 displayed increased long-distance and inter-chromosomal LD, yet different pairs of intervals were in LD in each species pair. LD calculated between individual SNP pairs and averaged over $10\ kb \times 10\ kb$ areas.

and to establish that differentiation is specific to between-species comparisons. In contrast with the low-recombining region on LG08, differentiation in the four intervals was weaker or absent when comparing populations within species (Suppl. Fig. 3.12). Furthermore, the effects of background selection are unlikely to be important in the earliest stages of differentiation studied here (Burri, 2017). This is confirmed by the weak genome-wide correlation between recombination rate and differentiation, and by the fact that our four candidate intervals do not show particularly low recombination rates (Suppl. Fig. 3.10). Finally, patterns of differentiation (F_{ST}) across these intervals were paralleled by genetic divergence (d_{XY}, Figure 3.3). This is the expected genomic signature of so-called barrier genes (Noor and Feder, 2006) that maintain species differences in the face of gene flow (Cruickshank and Hahn, 2014).

LD, Gene Flow and Speciation in the Sea

In the presence of gene flow, the extent of selection that is required to maintain long-distance or inter-chromosomal LD increases with the number of loci involved (Flaxman et al., 2014). This is because the number of possible genotypes in backcrosses increases exponentially with the number of loci (Slatkin,



Suppl. Figure 3.14: Long-distance and inter-chromosomal linkage disequilibrium (LD) among the eight intervals containing vision and pigmentation candidate genes. The intervals identified in & Suppl. Fig. 3.7 displayed increased long-distance and inter-chromosomal LD, yet different pairs of genes were in LD in each species pair. LD was calculated between individual SNP pairs and averaged over $10\ kb \times 10\ kb$ areas.

2008). Thus, disproportionally stronger selection is required to filter species-specific genotypes as the number of loci increases (Cruickshank and Hahn, 2014). In the hamlets, a small number of genomic intervals are strongly and consistently differentiated among species. This simple genetic architecture is expected to facilitate the build-up of ILD. Furthermore, differentiation is species-specific at three of these genomic intervals. Accordingly, long-distance and inter-chromosomal LD is not systematically observed among all pairs of intervals in all species pairs (Suppl. Fig. 3.13).

Once gene flow is sufficiently reduced through strong assortative mating, divergence and LD can accumulate rapidly by a combination of extrinsic and intrinsic forces (Flaxman et al., 2014). This is exactly the pattern we capture within the three hamlet species, which show a gradient of increasing differentiation and LD among populations (Figure 3.1b, c, Suppl. Fig. 3.4c, d, Suppl. Fig. 3.11b). The build-up of more pervasive ILD might be aided by epistatic interactions among loci. For example, *foxd3* on LG08 and *sox10* on LG09 both regulate the expression of *mitf*, a transcription factor involved in the development of melanophores in zebrafish (Elworthy et al., 2003; Curran et al., 2009). *ednrb* on LG04 (Parichy et al., 2000) is also involved in the development of melanophores in zebrafish,

and these three intervals show strong ILD in the comparison between the melanic (*H. nigricans*) and white (*H. unicolor*) species (Suppl. Fig. 3.14).

Our data provide a compelling scenario where speciation is driven by a combination of assortative mating and natural selection acting on a small number of large-effect loci, among which long-distance and interchromosomal LD is maintained in the presence of gene flow. The relatively simple genomic architecture underlying species differences in the hamlets parallels that observed in parapatric bird subspecies (Vijay et al., 2016), parapatric butterflies races (Van Belleghem et al., 2017) or depth-segregated cichlid ecomorphs (Malinsky et al., 2015) and we suggest that such a simple genomic architecture may be an important initial condition for the origin of many new species. The hamlets stand out from these other case studies by being fully sympatric at both the macro (overlapping distributions) and micro (overlapping habitats) geographic scales. In this respect they provide a counter-example to the idea that divergence tends to be genomically widespread among species that are fully sympatric (Seehausen et al., 2014). The relatively simple genomic architecture observed in the hamlets also contrasts with other systems in which differentiation is more widespread across the genome (Tine et al., 2014; Feulner et al., 2015; Meier et al., 2018). Factors contributing to this difference may include recent divergence, relatively high levels gene flow associated with extensive sympatry, and a simple genetic basis of the traits involved in reproductive isolation in the hamlets. In addition, the two-week planktonic larval stage of the hamlets provides potential for long-distance dispersal (Domeier, 1994). Nonetheless, our genomic data show that local evolutionary processes are operating in three communities separated by only a few hundreds of kilometres despite this dispersal phase. For example, *H. puella* and *H. unicolor* present two marked peaks of differentiation on LG17 in Panama that are not observed in Belize and Honduras for the same species pair. Marine speciation can therefore be characterised by local, heterogeneous and complex processes as observed in terrestrial and freshwater systems notwithstanding the apparent homogeneity of the marine environment.

3.3. Methods

Sampling. The majority of samples considered in this study were already available from previous studies(Puebla et al., 2007, 2012). New samples were only collected in Bocas del Toro (Panama) for RNA expression analysis following (Puebla et al., 2011) and relevant ethical regulations under the STRI IACUC protocol 2017-0101-2020-2 and the Panamanian Ministry of Environment permits SC/A-53-16 and SEX/A-35-17. Samples for expression analysis were collected in the early afternoon, kept in tanks overnight under natural light conditions and processed at noon on the following day. Only samples that could be unambiguously assigned to species on the basis of their colour pattern were considered.

Software versions, parameter settings and scripts. Software versions and parameter settings were omitted from the text for readability. Software versions are instead listed in Suppl. Tab. 3.2. All software pa-

rameter settings and scripts needed to reproduce our results from raw data to figures are provided in the accompanying repository (doi:10.3289/SW_2_2018, hereafter git).

De novo Genome Assembly

Library preparation and sequencing. The Genome assembly was based on a single barred hamlet (*H. puella*) from Panama (id 27678, Suppl. Tab. 3a). Genomic DNA was extracted from gill and muscle tissue using *Qiagen* MagAttract kits. Four paired-end (PE) 2×151 bp libraries with insert sizes ranging from roughly 250 to 320 bp were prepared, as well as one PE 2×251 bp PCR-free library with 580 bp insert size (Suppl. Tab. 3.4). Furthermore, two mate-pair (MP) libraries with insert sizes of about 2.5 and 4.3 kb were prepared. All PE and MP libraries were sequenced on *Illumina* HiSeq 2000/2500 platforms. Finally, *Illumina* data were complemented with longer *PacBio* (PB) reads from 20 SMRT cells. All sequencing for genome assembly was done at the *Duke Center for Genomic and Computational Biology*.

For annotation, RNA was extracted from gill, liver and muscle tissue from a single individual (id 16_21-30, Suppl. Tab. 3a) with an *Invitrogen* PureLink mRNA Mini Kit and sequenced on an *Illumina* MiSeq at the *Smithsonian Tropical Research Institute in Panama*. Additionally, RNA was extracted from the retinal tissue of 24 hamlets from Bocas del Toro, Panama (Suppl. Tab. 3.5), and sequenced on an *Illumina* NovaSeq platform by *Novogene*.

Data preparation – *Illumina* sequences. Prior to assembly, the sequencing data were preprocessed to remove low quality reads and possible contamination. As a first step, *Illumina* adapters and low quality reads were trimmed or filtered using Trimmomatic (Bolger et al., 2014) (PE & MP libraries, git 1.1.1.1) and NextClip (Leggett et al., 2014) (MP libraries, git 1.1.1.2 & 1.1.1.3).

To check for contamination, the filtered data were screened for bacterial and viral content using Kraken (Wood and Salzberg, 2014) (default database & settings, git 1.1.1.4 & 1.1.1.5) and classified reads were discarded using seqtk (`https://github.com/lh3/seqtk`, git 1.1.1.6). To remove possible human contamination, a two-step approach was applied. First, the reads were mapped against the human genome (GRCh38.p5) using Bowtie2 (Langmead and Salzberg, 2012)and hits were removed from the sample (git 1.1.1.7). The discarded reads were then mapped against the genome of the Asian Seabass (*Lates calcarifer*) (Vij et al., 2016) (git 1.1.1.8). The aim was to identify reads from conserved regions shared between the hamlet and the human genome. These reads were then merged back into the original samples (git 1.1.1.9).

Data preparation – *PacBio* sequences. The preparation of the *PacBio* data was done with proovread (Hackl et al., 2014) and Trimmomatic (git 1.1.2.0 – 1.1.2.5). The first 25 bp of all reads were trimmed to remove *PacBio* adapters. Then, a subset ($\sim 40\times$) of the filtered 2×251 bp PCR-free *Illumina* library (nr. 5 in Suppl. Tab. 3.4) was mapped

against the *PacBio* data. The mapping results were used to correct the *PacBio* reads and break apart chimeric reads. The whole process was parallelised using the SeqChunker script distributed with proovread. The results of every step were monitored with FastQC (Andrews, 2012) and MultiQC (Ewels et al., 2016) throughout the preparation phase.

***De novo* genome assembly.** After exploring a number of assemblers, Platanus (Kajitani et al., 2014) was chosen due to its good performance with the relatively heterozygous hamlet genome, using only the *Illumina* data in a first step. The contiging was based on the PE libraries and both PE & MP libraries were used for scaffolding and gap-closing (git 1.2.1 – 1.2.3). The resulting scaffolds were additionally gap-closed with the *Illumina*-corrected *PacBio* data using PBjelly (English et al., 2012) (git 1.2.4.1 – 1.2.4.6). Finally, the twofold gap-closed scaffolds were anchored and oriented to two RAD-based hamlet linkage maps (Theodosiou et al., 2016) using Allmaps (Tang et al., 2015). Briefly, the linkage map RAD tags were mapped onto the assembly scaffolds using Bowtie2 and the physical positions of the markers on the scaffolds were retrieved. Using custom R (R Core Team, 2017) scripts, the physical positions (bp) from the mapping were combined with the linkage map positions (cM, git 1.2.5). The resulting maps were merged into a single file and used for anchoring with Allmaps (git 1.2.6).

Manual curation. The anchored assembly was unmasked to capitalise lowercase sections resulting from PBjelly. The mitochondrial scaffold was identified by mapping the mitochondrial genome of the blue hamlet (*Hypoplectrus gemma*, GenBank accession nr: FJ848375) to the assembly. Finally, scaffolds smaller than 500 bp were removed from the assembly using SAMtools (Li et al., 2009) and bedtools (Quinlan and Hall, 2010) (git 1.2.7). At this point the assembly was considered complete and used as reference throughout the study (hereafter *hamlet reference genome*).

Quality assessment. The final assembly was aligned with the stickleback genome (Roesti et al., 2013) (*Gasterosteus aculeatus*, doi:10.5061/dryad.846nj), the most closely related high-quality genome, using LAST (Kiełbasa et al., 2011) (git 1.2.8). The alignments were visualised using Circos (Krzywinski et al., 2009) based on matches larger than 5 kb (git 2.2.0.1). The large scale synteny among the two genomes (Suppl. Fig. 3.1) was interpreted as a validation of the general structure of the hamlet genome assembly, and the hamlet linkage groups (LGs) were numbered following the numbering of the homologous stickleback LGs. Furthermore, the presence of genes highly conserved in vertebrates was assessed using BUSCO (Simão et al., 2015) and summary statistics for the assembly were generated with the summarizeAssembly.py script provided in the PBjelly suite.

Recombination landscape. To assess the large-scale recombination landscape, the RAD tags from the linkage map were mapped with Bowtie2 to the final assembly (git 1.2.9). A rough estimate of recombination density was provided by dividing linkage (cM) by physical distances (Mb) using R.

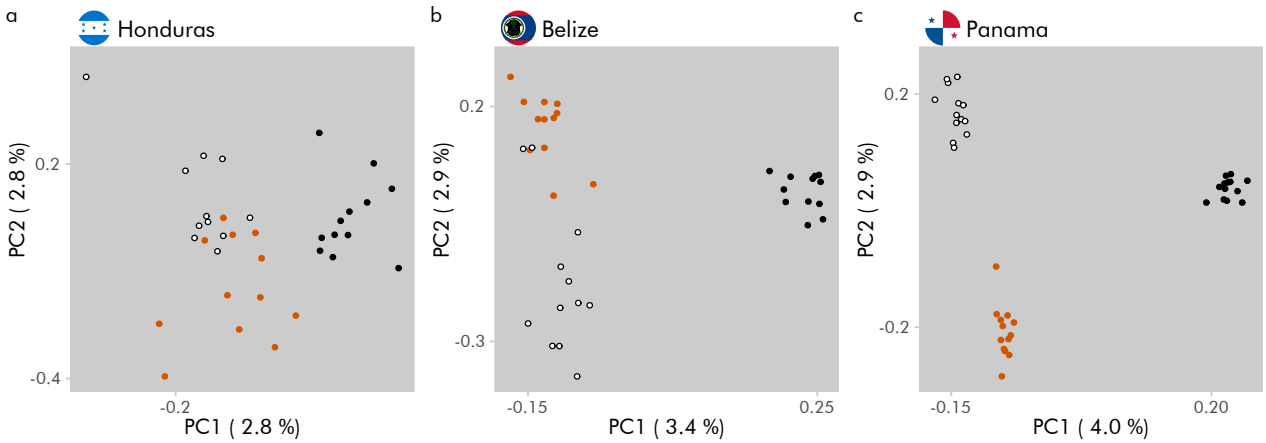
Annotation. The RNA libraries were assembled into a combined transcriptome as a basis for genome annotation. To prepare the reference genome, it was screened for specific repeat families using RepeatModler (Smit et al., 2015) (git 1.3.1) and repeats were masked for mapping using RepeatMasker (Smit and Hubley, 2015) (git 1.3.2). Scaffolds that contained only masked sequence were removed from the assembly. The RNA sequences were quality checked using FastQC, quality filtered using Trimmomatic (git 1.3.3) and mapped onto the masked version of the reference genome using HISAT2 (Kim et al., 2015) (git 1.3.4). The transcriptome was assembled from the mapped sequences using Trinity (Grabherr et al., 2011) in *genome_guided* mode (git 1.3.5). Preliminary gene models were constructed using the Maker package (Campbell et al., 2012), combining the information from *de novo* assembled transcripts with evidence-based, full-length protein sequences from zebrafish and stickleback Uniprot (The Uniprot Consortium, 2015). Functional inferences were generated using similarity searches of the annotated gene models against Uniprot/Swissprot and Interproscan (Finn et al., 2017), followed by manual curation of selected genes of interest with the Webapollo platform (Lee et al., 2013).

Resequencing and Variant Calling

Resequencing. The black (*H. nigricans*), barred (*H. puella*) and butter (*H. unicolor*) hamlets from Panama, Honduras and Belize were considered for resequencing. Eleven to thirteen individuals were sequenced per species and location, adding up to a total of 110 samples (Figure 3.1**a**, Suppl. Tab. 3a). An additional golden hamlet (*H. gumigutta*) was genotyped but excluded prior to analysis. Genomic libraries were prepared at the *Smithsonian Tropical Research Institute* in Panama (*STRI*, Belize & Honduras samples) and at the *Institute of Clinical Molecular Biology*in Kiel, Germany (*IKMB*, Panama samples) and sequenced on an *Illumina* HiSeq 4000 (PE, 2x151) by *Novogene* (Belize & Honduras Samples) and *IKMB* (Panama samples) at a mean sequencing depth of 24× (Suppl. Tab. 3a).

Variant calling. The genotyping procedure followed the best practice recommendations for the GATK (McKenna et al., 2010) work flow provided by the *Broad Institute* (Depristo et al., 2011; Van Der Auwera et al., 2013). We describe here the general work flow while the exact parameters used for each step are specified in git 2.1.1 – 2.1.12. Note that the samples from Panama were sequenced first and prepared independently from the Belize and Honduras samples, but processed together from the variant calling stage on (git 2.1.7).

Picard Tools (Broad Institute, 2015) was used to transform the sequences from *fq* to *uBAM* format, assign read groups, mark adapters and back-transform into *fq* format (git 2.1.1 – 2.1.3). They were then mapped to the hamlet reference genome using BWA (Li and Durbin, 2009) and merged with the *uBAM* files containing the read group information with Picard Tools (git 2.1.3). Afterwards, duplicated reads were removed (git 2.1.4). Then, us-



Suppl. Figure 3.15: PCA based on filtered data set. Principal Component Analysis (PCA) within each location. Genomic data was filtered for a minimum distance of 25 kb between SNPs to rule out physical linkage. After filtration 22,266 SNPs (0.3%) of the original data set remained.

ing GATK, genotype likelihoods were called (git 2.1.6) and all 110 samples were genotyped jointly (git 2.1.7.1). This step was duplicated, generating one data set with variant sites only (git 2.1.7.1.call_variants.sh) to be used for phasing and another data set including every callable site - even invariant ones (git 2.1.7.all_Variants_temp.sh) to calculate π and d_{XY}. SNPs were extracted from the raw genotypes and hard filtered with respect to quality (git 2.1.8). Furthermore, SNPs with missing data in more than 11 genotypes as well as multiallelic SNPs were removed using VCFtools (Danecek et al., 2011) (git 2.1.8).

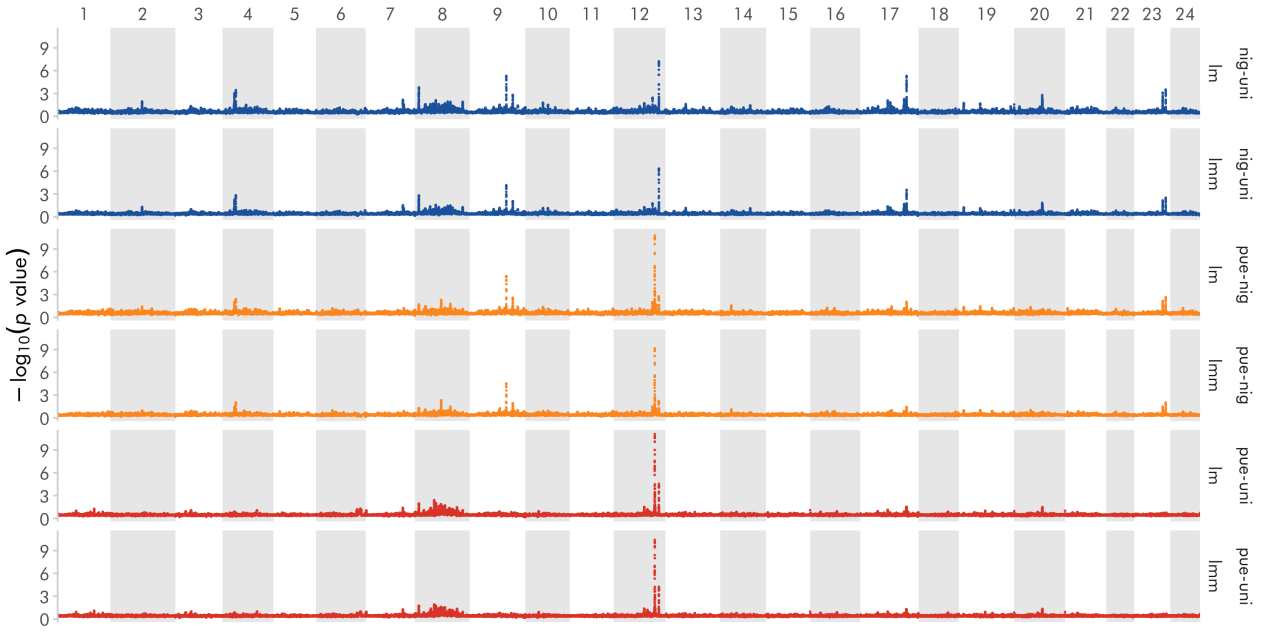
For the *all callable sites* data set, the genotypes (*vcf*) were subset by LG (git 2.1.9) and transformed to a custom genotype format required for popgenWindows.py (Martin, 2016) (git 2.1.10.vcf_2_geno_temp.sh).

The *SNPs only* data set was subset by LG as well (git 2.1.9.subset_LGs.sh). *Phase informative reads* were extracted (git 2.1.10.loop_extractPIRs.sh) and the genotypes were phased (git 2.1.11) using SHAPEIT (Delaneau et al., 2012). As a final step, SNPs with a minor allele count of one (minor allele frequency $< 0.9\%$) were removed from the data set (git 2.1.12).

Population Genomics

Most population genomic statistics were calculated within sliding windows along the genome. This was done at two resolutions: for a genome-wide overview a window size of 50 kb with 5 kb increments was chosen, and for more fine-scale analysis a window size of 10 kb with 1 kb increments was applied. In the following, these two resolutions are referred to as *broad* and *fine scale*.

PCA. For the principal component analysis, the *SNPs only* data set was subset by location using VCFtools and the subsets were reformatted. The three PCAs were then run independently using the R package pcadapt (Luu and Blum, 2017) (git 2.2.4). Similar re-



Suppl. Figure 3.16: Comparison of linear model and linear mixed model results of the genotype by phenotype (G × P) association among black (*H. nigricans*), barred (*H. puella*) and butter (*H. unicolor*) hamlets. Each species pair is represented by one colour, pooled across locations. Species comparisons are indicated on the right as well as model type (lm: linar model; lmm: linear mixed model). The *p* values are from the linear model with Wald test, transformed using the negative of the common logarithm and averaged across 50 kb window with 5 kb increments ($\overline{-log_{10}(p)}$).

sults were obtained when considering physically unlinked SNPs only (Suppl. Fig. 3.15).

π. Nucleotide diversity was based on the *all callable sites* data set and computed with VCFtools. It was calculated for each species as well as each species pair within each population at *fine scale* resolution (git 2.2.2 & 2.2.3).

d_{XY}. Genetic divergence (Nei, 1987) was based on the reformatted *all callable sites* data set. It was calculated for each population pair within each location in both resolutions using popgenWindows.py (Martin, 2016) (git 2.2.1).

F_{ST}. Genetic differentiation (Weir and Cockerham, 1984) was based on the *SNPs only* data set and computed with VCFtools, using the weighted mean following Weir & Cockerham (Weir and Cockerham, 1984). It was calculated at both resolutions for each species pair, both within each location and globally, as well as on a SNP basis among the three species pooled over locations. Additionally, it was calculated in *broad* resolution for every pair of locations within each species and globally (git 2.2.5).

G × P. Genotype × Phenotype associations were based on the *SNPs only* data set and estimated using a linear model (-lm) with GEMMA (Zhou and Stephens, 2012). For this, the data set was transformed to the plink format using VCFtools and plink (Purcell et al., 2007). G × P association was calculated on

a SNP basis for every species pair both within each location and globally, as well as for every pair of locations both within each species and globally (git 2.2.6.2 & 2.2.6.3). The results were then averaged over windows at both resolutions for the species comparisons, and at *broad* resolution for the location comparisons using custom shell scripts (git 2.2.6.4 – 2.2.6.6). Note that Wald-test *p* values were $-log_{10}$ transformed before averaging, so that $\overline{-log_{10}(p)}$ was reported for every window. GEMMA was additionally run under the linear mixed model, which provided similar results (Suppl. Fig. 3.16). Note also that $G \times P$ association, when applied to discrete phenotypes as done here, introduces some degree of redundancy with respect to F_{ST}.

***r*².** Linkage disequilibrium was calculated using VCFtools (git 2.2.8.1) at four different levels: first, to estimate the decay of linkage disequilibrium with physical distance, pairwise r^2 for all SNPs within 20 randomly selected windows of 15 kb was calculated. Second, to establish a baseline, genome-wide levels of LD were estimated from a random subset of 570 SNPs separated by at least 1 Mb (to rule out physical linkage) and considering interchromosomal pairs of SNPs only (ILD). Third, r^2 was calculated for all SNPs in and between broad regions around the candidate intervals, and fourth considering only SNPs within the candidate intervals (exact regions: git LD.bed & extendedLD.bed). The SNPs within the regions considered were then collated to allow a continuous visualisation (git 2.2.8.2 & 2.2.9.2). The pairwise r^2 values were sorted into 2-dimensional bins of $10 \times 10\ kb$ each and the average r^2 value for every bin was then calculated using R (git 2.2.8.3 & 2.2.9.3). Note that $r^2 = 0$ when there is no linkage, as opposed to the recombination rate *r* (not considered here), which equals 0.5 in the absence of linkage.

Hybrids and backcrosses. The approach implemented in NewHybrids (Anderson and Thompson, 2002) was used to test our hypothesis that some individuals might be of hybrid origin. This method does not require the *a priori* identification of pure individuals and relies on an explicit genetic model based on Mendelian inheritance. These analyses were run on small subsets of the *SNPs only* data set. First, for every pairwise species comparison within each location, the 800 SNPs with highest F_{ST} were selected. These SNP subsets were further filtered to include only SNPs that are separated by at least 5 kb to limit the effect of physical linkage among SNPs. From the resulting SNPs, 80 were randomly chosen to ensure that all analyses are based on the same number of markers. Note that the results were robust to alternative SNP selection strategies. Sub-setting was done using a combination of R, VCFtools and unix commands (git 2.2.10.1 & 2.2.10.2). The SNP subsets were transformed to the NewHybrids input format using PGDSpider (Lischer and Excoffier, 2012) (git 2.2.10.2). NewHybrids (Anderson and Thompson, 2002) was run in parallel using the R package parallelnewhybrid (Wringe et al., 2016) with a burnin of 10^6 n and $10 \cdot 10^6$ sweeps (git 2.2.10.3).

ρ. Population recombination rate was estimated using the machine learning R package FastEPRR (Gao et al., 2016). It was based on the *SNPs only* data set and calculated within

non-overlapping windows of 50 kb using 250 parallel jobs (git 2.2.11.1 - 2.2.11.3). For visualisation, the results where reformatted using a custom bash script (git 2.2.11.4).

RNA Expression

A *fasta* version of the transcriptome was created from the genome annotation file (*gff* in combination with the hamlet reference genome using gffread (Pertea, 2015) (git 2.3.1). The reference transcriptome was then indexed (git 2.3.2) and transcript abundances of the filtered retina RNA samples (also used for annotation) were estimated using kallisto (Bray et al., 2016) (git 2.3.3). Expression was analysed using the R package DSeq2 (Love et al., 2014) (git 3_figures & docs/index.html).

Simulations

Simulations were conducted to explore what combination of parameters may generate patterns of differentiation as sharp as the ones observed in the four candidate regions. Several demographic scenarios were simulated using the coalescent simulator msms (Ewing and Hermisson, 2010), considering a selected site located in the middle of a 500-kb chromosome. The simulations consisted of two populations of constant size N_e that split t generations ago and experienced constant and symmetrical migration (m) since then. The selected site was a single codominant locus with two alleles A and a that are advantageous in population 1 and 2, respectively, with a fitness of $1+s$ for homozygotes and $1+s/2$ for heterozygotes where s is the selection coefficient. We explored the parameter space spanned by the combinations of $N_e \in \{1000, 10000, 100000\}$, $t \in \{10000, 100000, 1000000\}$ generations, $m \in \{0.00001, 0.0001, 0.001, 0.01, 0.10\}$ & $s \in \{0.05, 0.1, 0.5\}$. The simulations were conducted with a recombination rate r of 0.02, providing a population recombination rate $4N_er$ that is similar to the one estimated from the empirical data with FastEPRR.

Sequences were simulated on the basis of the simulated genealogies using seq-gen (Rambaut and Grass, 1997) and variable sites were exported to the vcf format using msa2vcf (Lindenbaum, 2015). The vcf files were then used to calculate F_{ST} with VCFtools over 10 kb windows with 1 kb increments. NextFlow (Di Tommaso et al., 2017) was used to manage the simulations and analysis across the entire parameter space. Visualisation of the results was done within R (git 3_figures & docs/index.html).

Visualisation

All results were plotted using R with the exception of the synteny plot (Suppl. Fig. 3.1, Circos) and the LG08 low-recombination plot (Suppl. Fig. 3.4, Allmaps & Inkscape). The details of the visualisation are provided in the R scripts and their documentation (git 3_figures & docs/index.html). Within those R scripts, the following packages were used:

bookdown (0.7) (Xie, 2016), colorspace (1.3-2) (Ihaka et al., 2016), cowplot (0.9.2)

(Wilke, 2017), DESeq2 (1.16.1) (Love et al., 2014), dplyr (0.7.4) (Wickham et al., 2017), FastEPRR (1.0) (Gao et al., 2016), gdata (2.18.0) (Warnes et al., 2017), ggforce (0.1.2) (Pedersen, 2018), ggmap (2.6.1) (Kahle and Wickham, 2013), ggplot2 (3.0.0) (Wickham, 2016a), ggrepel (0.7.0) (Slowikowski, 2017), gplots (3.0.1) (Warnes et al., 2016), grConvert (0.1-0) (Potter, 2018a), grid (3.4.3) (R Core Team, 2017), gridExtra (2.3) (Auguie, 2017), gridSVG (1.6-0) (Murrell and Potter, 2017), grImport2 (0.1-4) (Potter, 2018b), gtable (0.2.0) (Wickham, 2016b), hrbrthemes (0.1.0) (Rudis, 2017), knitr (1.20) (Xie, 2014, 2015, 2018), maptools (0.9-2) (Bivand and Lewin-Koh, 2017), marmap (0.9.6) (Pante and Simon-Bouhet, 2013), parallelnewhybrid (0.0.0.9002) (Wringe et al., 2016), PBSmapping (2.70.4) (Schnute et al., 2017), pcadapt (3.0.4) (Luu and Blum, 2017), RColorBrewer (1.1-2) (Neuwirth, 2014), rtracklayer (1.36.6) (Lawrence et al., 2009), scales (0.5.0.9000) (Wickham, 2018), scatterpie (0.0.9) (Yu, 2018), sp (1.2-7) (Pebesma and Bivand, 2005; Bivand et al., 2013), tidyverse (1.2.1) (Wickham, 2017), tximport (1.4.0) (Soneson et al., 2015), vsn (3.44.0) (Huber et al., 2002)

Acknowledgements

This study was funded by an individual German Research Foundation (DFG) grant to OP (PU571/1-1) and a grant from the Smithsonian Institute for Biodiversity Genomics and the Global Genome Initiative Grants Program to WOM, OP and Carole Baldwin. The authors thank Till Bayer, Biff Bermingham, Katharina Fietz, Paul Frandsen, Bernhard Haubold, Chris Jiggins, Richard Merrill, Thorsten Reusch, Montserrat Torres Oliva and Steven Van Belleghem as well as the Belizean, Honduran and Panamanian authorities for support.

Data Accessibility

The raw sequencing data and genome assembly (fasta format) are deposited in the European Nucleotide Archive (ENA, project accession number PRJEB27858). Accession numbers for all sample sequences are provided in Suppl. Tab. 3a and Suppl. Tab. 3.5. Additional data underlying this study, including the genome annotation (gff format), the genotypes (vcf format) and the per locus F_{ST} and $G \times P$ results are deposited in Dryad (doi:10.5061/dryad.pg8q56g).

Supplementary Tables

Suppl. Table 3.1: Genomic regions above the 99.90^{th} F_{ST} percentile, based on sliding window analysis with 50 kb windows and 5 kb increments. The Comparison column refers to the specific pairs in which the region is above the 99.90^{th} F_{ST} percentile: *H. nigricans vs. H. puella* (NP), *H. nigricans vs. H. unicolor* (NU) & *H. puella vs. H. unicolor* (NP). Regions containing candidate genes are labelled with capital letters. Regions $A - D$ are also above the 99.90^{th} F_{ST} percentile. The **Other genes** column includes all genes overlapping with the 50-kb windows above the 99.90^{th} F_{ST} percentile. Note that this approach conservatively includes genes situated before and after peaks of differentiation at a 10-kb window resolution (Suppl. Fig. 3.8).

Nr	ID	LG	Start (kb)	End (kb)	Candidate genes	Other genes	Comparison
1	E	04	5835	5925	*ednrb*	*polr1d, hpv1g...7175, cysltr1, mtus1a, vamp7*	NP,NU
2		04	6555	6645		*asb12, zc4h2,hpv1g...7205, msn,ar, arhgef9, efnb1*	NP,NU
3	F	08	1945	2090	*foxd3*	*hpv1g...10963, alg6, efcab7, pgm1*	NU,PU
4		08	9505	9595		*arhgef18, insr*	PU
5		08	10270	10420		*klhl23, soat1, abl2, iqca1, cers2, cac1e, osbpl9*	PU
6		08	12810	12895		*fpgt, acot11, dio1, tnni3k, ssbp3, glis1*	PU
7	G	08	12945	13010	*rorb*	*mpnd, sh3gl1, hpv1g...11382, lgals3bpb*	NP
8		08	14825	14875		*ctdspl2a, arid3a, tmem79*	PU
9	A	09	17821	17930	*sox10*	*smdt1, hpv1g...12847, kcnj12, rnaseh2a, mast1, triobp*	NP,NU
10		09	20995	21085		*rnf24, smox, fbxo41*	NP,NU
11		12	15050	15105		*csf1, ren, eif2d, hpv1g...15947*	PU
12	B	12	20085	20355	*casz1*	*pgd, c1orf127, kif1c, kif1b, tardbp*	NP,PU
13	C	12	22150	22290	*hoxc13a*	*hoxc5a, hoxc6a, hoxc8a, hoxc9, hoxc10a, hoxc11a, hoxc12a, calcoco1, rarga*	NP,NU,PU
14		17	21340	21415		*itih3, hpv1g...21480*	NU
15	D	17	22505	22660	*lws, sws2aβ, sws2aα, sws2b*	*mafb, deptor, adnp, rab7,hcfc1, gnl3l, tfe3, mdfic2, cxxc1, srpk3, comt, gata2, mbd1, ccdc120*	NP,NU,PU
16	H	20	13840	13900	*invs*	*hpv1g...24338, stx17, erp44, tex10*	NU,PU
17		23	13965	14030		*ntf3*	NP,NU
18		23	15445	15530		*crys, glipr1l1, ache, nxpe3, krr1, st3gal1, gp2*	NP,NU

Suppl. Table 3.2: Software versions used in this study

Software	**version**	**Software**	**version**
Allmaps	Version 1	NextFlow	0.31.1
bedtools	v2.27.1	NewHybrids	2.0+ Developmental
Bowtie2	version 2.3.4.1	PBjelly	v14.1
BUSCO	2PGDSpider	2.1.1.5	
BWA	0.7.12-r1044	Picard Tools	2.9.2-SNAPSHOT
Circos	v 0.69	Platanus	1.2.4
FastQC	v0.11.3	plink	v1.90b4 64-bit
GATK	v3.7-0-gcfedb67	proovread	2.13.13
GEMMA	0.97.2	R	3.4.1 (calculations)
gffread	v0.9.12		3.4.3 (visualisations)
HISAT2	2.0.4	RepeatMasker	Open-4.0.6
Inkscape	0.91 r13725	RepeatModler	open-1-0-8
kallisto	0.43.1	SAMtools	1.7
Kraken	0.10.6-unreleased	selscan	v1.2.0a
LAST	737	SeqChunker	v0.22.2
Maker	v 3.0	seq-gen	1.3.4
msa2vcf		seqtk	1.2-r94
98d97d07d6101fab1b0bef757b4ceee279e171d9		SHAPEIT	v2.r837
msms	3.2rc	Trimmomatic	0.33
MultiQC	Version 0.8	Trinity	v2.2.0
NextClip	v1.3.1	VCFtools	0.1.15

Suppl. Table 3a: Samples used for resequencing (Samples 1-50).

ID	Species	Location	Date	Latitude	Longitude	Cov.	Acces. Nr.
18151	*H. nigricans*	Belize	2004-07-25	16.7653	-088.1442	22.6	ERS2619600
18153	*H. nigricans*	Belize	2004-07-25	16.7653	-088.1442	23.7	ERS2619601
18155	*H. nigricans*	Belize	2004-07-25	16.8008	-088.0789	21.2	ERS2619602
18156	*H. nigricans*	Belize	2004-07-25	16.8008	-088.0789	26.5	ERS2619603
18157	*H. nigricans*	Belize	2004-07-25	16.8008	-088.0789	23.2	ERS2619604
18158	*H. nigricans*	Belize	2004-07-25	16.8008	-088.0789	21.6	ERS2619605
18159	*H. nigricans*	Belize	2004-07-25	16.8008	-088.0789	22.5	ERS2619606
18162	*H. nigricans*	Belize	2004-07-25	16.7653	-088.1442	19.1	ERS2619607
18165	*H. nigricans*	Belize	2004-07-25	16.7653	-088.1442	22.9	ERS2619608
18171	*H. nigricans*	Belize	2004-07-25	16.7653	-088.1442	26.4	ERS2619609
18185	*H. nigricans*	Belize	2004-07-26	16.8058	-088.0792	25.6	ERS2619610
18187	*H. nigricans*	Belize	2004-07-26	16.8058	-088.0792	26.6	ERS2619611
20599	*H. nigricans*	Honduras	2006-06-04	15.9558	-083.2931	24.2	ERS2619612
20600	*H. nigricans*	Honduras	2006-06-04	15.9558	-083.2931	25.6	ERS2619613
20601	*H. nigricans*	Honduras	2006-06-04	15.9558	-083.2931	22.8	ERS2619614
20602	*H. nigricans*	Honduras	2006-06-04	15.9558	-083.2931	23	ERS2619615
20603	*H. nigricans*	Honduras	2006-06-04	15.9558	-083.2931	26	ERS2619616
20604	*H. nigricans*	Honduras	2006-06-04	15.9558	-083.2931	20.3	ERS2619617
20605	*H. nigricans*	Honduras	2006-06-04	15.9558	-083.2931	22.9	ERS2619618
20606	*H. nigricans*	Honduras	2006-06-04	15.9558	-083.2931	22.6	ERS2619619
20607	*H. nigricans*	Honduras	2006-06-04	15.9558	-083.2931	25.9	ERS2619620
20608	*H. nigricans*	Honduras	2006-06-04	15.9558	-083.2931	24.6	ERS2619621
20609	*H. nigricans*	Honduras	2006-06-04	15.9558	-083.2931	26.2	ERS2619622
20610	*H. nigricans*	Honduras	2006-06-04	15.9558	-083.2931	22.9	ERS2619623
16_21-30	*H. nigricans*	Panama	2016	-	-	89.1	ERS2619624
18418	*H. nigricans*	Panama	2004-05-12	09.3775	-082.3039	18.7	ERS2619625
18424	*H. nigricans*	Panama	2004-05-12	10.2392	-083.1731	19.3	ERS2619626
18428	*H. nigricans*	Panama	2004-05-12	10.2392	-083.1731	19.1	ERS2619627
18436	*H. nigricans*	Panama	2004-05-12	09.3775	-082.3039	9.8	ERS2619628
18901	*H. nigricans*	Panama	2005-03-25	09.2983	-082.2894	25.6	ERS2619629
18902	*H. nigricans*	Panama	2005-03-25	09.2983	-082.2894	22.4	ERS2619630
18903	*H. nigricans*	Panama	2005-03-25	09.2983	-082.2894	17.1	ERS2619631
18904	*H. nigricans*	Panama	2005-03-25	09.2983	-082.2894	20.1	ERS2619632
18905	*H. nigricans*	Panama	2005-03-25	09.2983	-082.2894	22.2	ERS2619633
18906	*H. nigricans*	Panama	2005-03-25	09.2983	-082.2894	26.7	ERS2619634
18907	*H. nigricans*	Panama	2005-03-25	09.2983	-082.2894	18.1	ERS2619635
18909	*H. nigricans*	Panama	2005-03-25	09.2983	-082.2894	21.9	ERS2619636
18152	*H. puella*	Belize	2004-07-25	16.7653	-088.1442	23.4	ERS2619637
18154	*H. puella*	Belize	2004-07-25	16.7653	-088.1442	24.6	ERS2619638
18161	*H. puella*	Belize	2004-07-26	16.8058	-088.0792	18.4	ERS2619639
18166	*H. puella*	Belize	2004-07-25	16.7653	-088.1442	24.2	ERS2619640
18169	*H. puella*	Belize	2004-07-25	16.7653	-088.1442	19.6	ERS2619641
18172	*H. puella*	Belize	2004-07-25	16.7653	-088.1442	23.4	ERS2619642
18174	*H. puella*	Belize	2004-07-25	16.7653	-088.1442	24.8	ERS2619643
18175	*H. puella*	Belize	2004-07-25	16.7653	-088.1442	22.2	ERS2619644
18176	*H. puella*	Belize	2004-07-25	16.7653	-088.1442	29	ERS2619645
18178	*H. puella*	Belize	2004-07-25	16.7653	-088.1442	22.9	ERS2619646
18179	*H. puella*	Belize	2004-07-25	16.7653	-088.1442	22.1	ERS2619647
18180	*H. puella*	Belize	2004-07-25	16.7653	-088.1442	21.7	ERS2619648
20551	*H. puella*	Honduras	2006-06-04	15.9558	-083.2931	23	ERS2619649

Suppl. Table 3b: Samples used for resequencing (continued, Samples 51-100).

ID	Species	Location	Date	Latitude	Longitude	Cov.	Acces. Nr.
20552	*H. puella*	Honduras	2006-06-04	15.9558	-83.2931	24.5	ERS2619650
20553	*H. puella*	Honduras	2006-06-04	15.9558	-83.2931	26.2	ERS2619651
20554	*H. puella*	Honduras	2006-06-04	15.9558	-83.2931	22	ERS2619652
20555	*H. puella*	Honduras	2006-06-04	15.9558	-83.2931	23.2	ERS2619653
20556	*H. puella*	Honduras	2006-06-04	15.9558	-83.2931	29.1	ERS2619654
20558	*H. puella*	Honduras	2006-06-04	15.9558	-83.2931	25.5	ERS2619655
20559	*H. puella*	Honduras	2006-06-04	15.9558	-83.2931	26	ERS2619656
20625	*H. puella*	Honduras	2006-06-05	15.9558	-83.2931	21.8	ERS2619657
20633	*H. puella*	Honduras	2006-06-05	15.9558	-83.2931	27.6	ERS2619658
20635	*H. puella*	Honduras	2006-06-05	15.9558	-83.2931	26.3	ERS2619659
20638	*H. puella*	Honduras	2006-06-05	15.9558	-83.2931	21.7	ERS2619660
18419	*H. puella*	Panama	2004-05-12	10.2392	-83.1731	19.8	ERS2619661
18421	*H. puella*	Panama	2004-05-12	10.2392	-83.1731	22.7	ERS2619662
18422	*H. puella*	Panama	2004-05-12	9.3775	-82.3039	22.9	ERS2619663
18426	*H. puella*	Panama	2004-05-12	10.2392	-83.1731	20.9	ERS2619664
18427	*H. puella*	Panama	2004-05-12	10.2392	-83.1731	21.3	ERS2619665
18429	*H. puella*	Panama	2004-05-12	10.2392	-83.1731	23.5	ERS2619666
18430	*H. puella*	Panama	2004-05-12	10.2392	-83.1731	19.5	ERS2619667
18432	*H. puella*	Panama	2004-05-12	10.2392	-83.1731	21.6	ERS2619668
18434	*H. puella*	Panama	2004-05-12	9.3775	-82.3039	25.7	ERS2619669
18912	*H. puella*	Panama	2005-03-25	9.2983	-82.2894	23	ERS2619670
18915	*H. puella*	Panama	2005-03-25	9.2983	-82.2894	19.4	ERS2619671
18917	*H. puella*	Panama	2005-03-25	9.2983	-82.2894	18.8	ERS2619672
27678	*H. puella*	Panama	2013-04-12	9.3681	-82.2928	89.8	ERS2619673
18163	*H. unicolor*	Belize	2004-07-25	16.7653	-88.1442	24.5	ERS2619674
18261	*H. unicolor*	Belize	2004-07-24	-	-	21.8	ERS2619675
18267	*H. unicolor*	Belize	2004-07-24	-	-	28.4	ERS2619676
18274	*H. unicolor*	Belize	2004-07-24	-	-	24.4	ERS2619677
18276	*H. unicolor*	Belize	2004-07-25	16.7653	-88.1442	23.8	ERS2619678
19881	*H. unicolor*	Belize	2005-08-16	16.7078	-87.8598	23.7	ERS2619679
20092	*H. unicolor*	Belize	2005-08-15	16.8936	-88.1226	23.5	ERS2619680
20120	*H. unicolor*	Belize	2005-08-11	16.8008	-88.0789	30.3	ERS2619681
20126	*H. unicolor*	Belize	2005-08-12	16.8936	-88.1226	22.4	ERS2619682
20128	*H. unicolor*	Belize	2005-08-12	16.8936	-88.1226	22.6	ERS2619683
20135	*H. unicolor*	Belize	2005-08-12	16.8936	-88.1226	21.2	ERS2619684
20149	*H. unicolor*	Belize	2005-08-12	16.8936	-88.1226	23.5	ERS2619685
20560	*H. unicolor*	Honduras	2006-06-04	15.9558	-83.2931	21.3	ERS2619686
20561	*H. unicolor*	Honduras	2006-06-04	15.9558	-83.2931	24.2	ERS2619687
20562	*H. unicolor*	Honduras	2006-06-04	15.9558	-83.2931	25.4	ERS2619688
20563	*H. unicolor*	Honduras	2006-06-04	15.9558	-83.2931	23.1	ERS2619689
20564	*H. unicolor*	Honduras	2006-06-04	15.9558	-83.2931	28.9	ERS2619690
20565	*H. unicolor*	Honduras	2006-06-04	15.9558	-83.2931	25	ERS2619691
20566	*H. unicolor*	Honduras	2006-06-04	15.9558	-83.2931	27.4	ERS2619692
20567	*H. unicolor*	Honduras	2006-06-04	15.9558	-83.2931	27.3	ERS2619693
20568	*H. unicolor*	Honduras	2006-06-04	15.9558	-83.2931	30.1	ERS2619694
20571	*H. unicolor*	Honduras	2006-06-04	15.9558	-83.2931	24.6	ERS2619695
20572	*H. unicolor*	Honduras	2006-06-04	15.9558	-83.2931	23.8	ERS2619696
16_31-40	*H. unicolor*	Panama	2016	-	-	53.4	ERS2619697
18420	*H. unicolor*	Panama	2004-05-12	10.2392	-83.1731	20.3	ERS2619698
18435	*H. unicolor*	Panama	2004-06-23	9.3328	-82.2547	20.2	ERS2619699

Suppl. Table 3c: Samples used for resequencing (continued, Samples 101-110).

ID	Species	Location	Date	Latitude	Longitude	Cov.	Acces. Nr.
18439	*H. unicolor*	Panama	2004-06-25	9.3328	-82.2547	23.8	ERS2619700
18440	*H. unicolor*	Panama	2004-06-25	9.3328	-82.2547	19.7	ERS2619701
18441	*H. unicolor*	Panama	2004-06-25	9.3328	-82.2547	25.1	ERS2619702
18442	*H. unicolor*	Panama	2004-07-08	9.2983	-82.2894	20.6	ERS2619703
18445	*H. unicolor*	Panama	2004-06-28	9.3328	-82.2547	19	ERS2619704
18446	*H. unicolor*	Panama	2004-06-29	9.3328	-82.2547	26	ERS2619705
18447	*H. unicolor*	Panama	2004-07-09	9.2894	-82.2589	16.8	ERS2619706
18448	*H. unicolor*	Panama	2004-06-28	9.3328	-82.2547	26.1	ERS2619707
18450	*H. unicolor*	Panama	2004-06-25	9.3328	-82.2547	17.7	ERS2619708
18454	*H. unicolor*	Panama	2004-06-30	9.3481	-82.2633	22.7	ERS2619709

Suppl. Table 3.4: Overview of the sequencing data generated for the assembly of the *Hypoplectrus* genome (PE: paired end, MP: mate pair, PB: PacBio)

Nr	Tissue	Type	Type	Targeted insert size (bp)	Mapped insert size (bp, mean $\pm$ SD)	Read length (bp)	Coverage ($\times$)
1	gill	PE	DNA	300	264 ± 94	151	24
2	gill	PE	DNA	300	255 ± 80	151	144
3	gill	PE	DNA	500	299 ± 106	151	46
4	gill	PE	DNA	800	321 ± 123	151	67
5	muscle	PE (PCR-free)	DNA	550	579 ± 155	251	100
6	muscle	MP	DNA	3000	2457 ± 639	101	34
7	muscle	MP	DNA	6000	4329 ± 1110	101	31
8	muscle	PB	DNA	-	-	50–33680	16
9	gill	PE	RNA	-	-	251	-
10	muscle	PE	RNA	-	-	251	-
11	liver	PE	RNA	-	-	251	-

Suppl. Table 3.5: Samples used for RNA sequencing

ID	Date	Species	Latitude	Longitude	Raw reads (n, 10^6)	Filtered reads (n, 10^6)	Alignment rate (%)	Acession Number
28385	2017-02-06	*H. nigricans*	09.318	-082.222	2x7.4	2x4.7	60.84	ERS2619746
28387	2017-02-06	*H. nigricans*	09.318	-082.222	2x6.3	2x4.7	59.34	ERS2619747
28390	2017-02-06	*H. nigricans*	09.318	-082.222	2x5.9	2x4.4	61.72	ERS2619748
28394	2017-02-07	*H. nigricans*	09.301	-082.294	2x4.8	2x3.0	63.55	ERS2619749
28399	2017-02-07	*H. nigricans*	09.301	-082.294	2x7.3	2x5.2	62.11	ERS2619750
AG9RX46	2017-02-06	*H. nigricans*	09.318	-082.222	2x7.1	2x5.3	60.73	ERS2619751
AG9RX49	2017-02-07	*H. nigricans*	09.301	-082.294	2x7.1	2x5.3	63.38	ERS2619752
AG9RX50	2017-02-07	*H. nigricans*	09.301	-082.294	2x6.8	2x4.8	60.68	ERS2619753
AG9RX52	2017-02-07	*H. nigricans*	09.301	-082.294	2x6.7	2x4.9	59.98	ERS2619754
28384	2017-02-06	*H. puella*	9.318	-082.222	2x6.1	2x4.2	63.13	ERS2619755
AG9RX47	2017-02-06	*H. puella*	09.318	-082.222	2x5.5	2x4.0	68.03	ERS2619756
AG9RX48	2017-02-07	*H. puella*	09.301	-082.294	2x6.2	2x4.5	66.079	ERS2619757
AG9RX51	2017-02-07	*H. puella*	09.301	-082.294	2x6.7	2x5.0	61.41	ERS2619758
AG9RX53	2017-02-07	*H. puella*	09.301	-082.294	2x6.3	2x4.5	62	ERS2619759
PL17_02	2017-02-07	*H. puella*	09.301	-082.294	2x6.1	2x4.6	62.61	ERS2619760
PL17_04	2017-02-07	*H. puella*	09.301	-082.294	2x6.5	2x4.7	63.17	ERS2619761
PL17_16	2017-02-09	*H. puella*	09.367	-082.291	2x7.5	2x5.3	62.1	ERS2619762
PL17_17	2017-02-09	*H. puella*	09.367	-082.291	2x7.3	2x5.3	61.72	ERS2619763
PL17_18	2017-02-09	*H. puella*	09.367	-082.291	2x6.3	2x4.6	63.68	ERS2619764
28383	2017-02-06	*H. unicolor*	09.318	-082.222	2x6.8	2x4.7	61.04	ERS2619765
28391	2017-02-07	*H. unicolor*	09.301	-082.294	2x7.1	2x5.2	61.75	ERS2619766
AG9RX54	2017-02-06	*H. unicolor*	09.318	-082.222	2x5.5	2x3.9	59.2	ERS2619767
AG9RX55	2017-02-07	*H. unicolor*	09.301	-082.294	2x6.3	2x4.7	61.54	ERS2619768
PL17_01	2017-02-07	*H. unicolor*	09.301	-082.294	2x8.1	2x5.9	62.11	ERS2619769

Suppl. Table 3.6: Whole-genome weighted mean F_{ST} estimates among *H. nigricans*, *H. puella* & *H. unicolor*.

Location	***H. nigricans* *H. puella***	***H. nigricans* *H. unicolor***	***H. puella* *H. unicolor***	All
Global	0.0079	0.0098	0.0027	0.0068
Belize	0.0168	0.0153	0.0047	0.0123
Honduras	0.0033	0.0051	0.0030	0.0038
Panama	0.0274	0.0348	0.0125	0.0249

Chapter 3 References

Anderson, E. C. and Thompson, E. A. (2002). A model-based method for identifying species hybrids using multilocus genetic data. *Genetics*, 160(3):1217–1229.

Andrews, S. (2012). Fastqc: a quality control tool for high throughput sequence data.

Arai, R. (2011). *Fish karyotypes. A check list*. Springer, Tokyo.

Auguie, B. (2017). *Gridextra: miscellaneous functions for "grid" graphics*. R package version 2.3.

Barreto, F. S. and McCartney, M. A. (2008). Extraordinary AFLP fingerprint similarity despite strong assortative mating between reef fish color morphospecies. *Evolution*, 62(1):226–233.

Bay, R. A., Arnegard, M. E., Conte, G. L., Best, J., Bedford, N. L., McCann, S. R., Dubin, M. E., Chan, Y. F., Jones, F. C., Kingsley, D. M., Schluter, D., and Peichel, C. L. (2017). Genetic coupling of female mate choice with polygenic ecological divergence facilitates stickleback speciation. *Current Biology*, 27(21):3344–3349.

Bellwood, D. R., Goatley, C. H., and Bellwood, O. (2017). The evolution of fishes and corals on reefs: form, function and interdependence. *Biological Reviews*, 92(2):878–901.

Bierne, N. (2010). The distinctive footprints of local hitchhiking in a varied environment and global hitchhiking in a subdivided population. *Evolution*, 64(11):3254–3272.

Bivand, R. and Lewin-Koh, N. (2017). *Maptools: tools for reading and handling spatial objects*. R package version 0.9-2.

Bivand, R. S., Pebesma, E., and Gomez-Rubio, V. (2013). *Applied spatial data analysis with r, second edition*. Springer, Ny.

Bolger, A. M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for illumina sequence data. *Bioinformatics*, 30(15):2114–2120.

Bray, N. L., Pimentel, H., Melsted, P., and Pachter, L. (2016). Near-optimal probabilistic rna-seq quantification. *Nature Biotechnology*, 34(5):525–527.

Broad Institute (2015). Picard tools.

Burri, R. (2017). Interpreting differentiation landscapes in the light of long-term linked selection. *Evolution Letters*, 1(3):118–131.

Campbell, M. S., Holt, C., Moore, B., and Yandell, M. (2012). Genome annotation and curation using maker and maker-p. *Current Protocols in Bioinformatics*, 48(1):4.11.1–4.11.39.

Carroll, S. B., Grenier, J. K., and Weatherbee, S. D. (2005). *From DNA to diversity, molecular genetics and the evolution of animal design*. Blackwell Publishing Ltd, Oxford, 2 edition.

Charlesworth, B., Morgan, M., and Charlesworth, D. (1993). The effect of deleterious mutations on neutral molecular variation. *Genetics*, 134(4):1289–1303.

Charlesworth, B., Nordborg, M., and Charlesworth, D. (1997). The effects of local selection, balanced polymorphism and background selection on equilibrium

patterns of genetic diversity in subdivided populations. *Genetical Research*, 70(2):155–174.

Cruickshank, T. E. and Hahn, M. W. (2014). Reanalysis suggests that genomic islands of speciation are due to reduced diversity, not reduced gene flow. *Molecular Ecology*, 23(13):3133–3157.

Curran, K., Raible, D. W., and Lister, J. A. (2009). Foxd3 controls melanophore specification in the zebrafish neural crest by regulation of mitf. *Developmental Biology*, 332(2):408–417.

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., Depristo, M. A., Handsaker, R. E., Lunter, G., Marth, G. T., Sherry, S. T., McVean, G., Durbin, R., and 1000 Genomes Project Analysis Group (2011). The variant call format and vcftools. *Bioinformatics*, 27(15):2156–2158.

Delaneau, O., Marchini, J., and Zagury, J.-F. (2012). A linear complexity phasing method for thousands of genomes. *Nature Methods*, 9:179.

Depristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., Philippakis, A. A., Del Angel, G., Rivas, M. A., Hanna, M., McKenna, A., Fennell, T. J., Kernytsky, A. M., Sivachenko, A. Y., Cibulskis, K., Gabriel, S. B., Altshuler, D., and Daly, M. J. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics*, 43(5):491–498.

Di Tommaso, P., Chatzou, M., Floden, E. W., Barja, P. P., Palumbo, E., and Notredame, C. (2017). Nextflow enables reproducible computational workflows. *Nat. Biotechnol.*, 35:316.

Domeier, M. (1994). Speciation in the serranid fish *Hypoplectrus*. *Bulletin of Marine Science*, 54(1):103–141.

Dutton, K. A., Pauliny, A., Lopes, S. S., Elworthy, S., Carney, T. J., Rauch, J., Geisler, R., Haffter, P., and Kelsh, R. N. (2001). Zebrafish colourless encodes sox10 and specifies non-ectomesenchymal neural crest fates. *Development*, 128(21):4113–4125.

Elworthy, S., Lister, J. A., Carney, T. J., Raible, D. W., and Kelsh, R. N. (2003). Transcriptional regulation of mitfa accounts for the sox10 requirement in zebrafish melanophore development. *Development*, 130(12):2809–2818.

English, A. C., Richards, S., Han, Y., Wang, M., Vee, V., Qu, J., Qin, X., Muzny, D. M., Reid, J. G., Worley, K. C., and Gibbs, R. A. (2012). Mind the gap: upgrading genomes with pacific biosciences rs long-read sequencing technology. *PLoS One*, 7.

Ewels, P., Magnusson, M., Lundin, S., and Käller, M. (2016). Multiqc: summarize analysis results for multiple tools and samples in a single report.

Ewing, G. and Hermisson, J. (2010). MSMS: a coalescent simulation program including recombination, demographic structure and selection at a single locus. *Bioinformatics*, 26(16):2064–2065.

Feder, J. L. and Nosil, P. (2010). The efficacy of divergence hitchhiking in generating genomic islands during ecological speciation. *Evolution*, 64(6):1729–1747.

Felsenstein, J. (1981). Skepticism towards santa rosalia, or why are there so few kinds of animals. *Evolution*, 35(1):124–138. 10.2307/2407946.

Feulner, P. G., Chain, F. J., Panchal, M., Huang, Y., Eizaguirre, C., Kalbe, M., Lenz, T. L., Samonte, I. E., Stoll, M., Bornberg-Bauer, E., et al. (2015). Genomics of divergence along a continuum of parapatric population differentiation. *PLoS genetics*, 11(2):e1004966.

Finn, R. D., Attwood, T. K., Babbitt, P. C., Bateman, A., Bork, P., Bridge, A. J., Chang, H.-Y., Dosztányi, Z., El-Gebali, S., Fraser, M., Gough, J., Haft, D., Holliday, G. L., Huang, H., Huang, X., Letunic, I., Lopez, R., Lu, S., Marchler-Bauer, A., Mi, H., Mistry, J., Natale, D. A., Necci, M., Nuka, G., Orengo, C. A., Park, Y., Pesseat, S., Piovesan, D., Potter, S. C., Rawlings, N. D., Redaschi, N., Richardson, L., Rivoire, C., Sangrador-Vegas, A., Sigrist, C., Sillitoe, I., Smithers, B., Squizzato, S., Sutton, G., Thanki, N., Thomas, P. D., Tosatto, S. C. E., Wu, C. H., Xenarios, I., Yeh, L.-S., Young, S.-Y., and Mitchell, A. L. (2017). Interpro in 2017—beyond protein family and domain annotations. *Nucleic Acids Research*, 45(D1):D190–D199.

Fischer, E. (1980). The relationship between mating system and simultaneous hermaphroditism in the coral-reef fish, *Hypoplectrus nigricans* (Serranidae). *Animal Behaviour*, 28(May):620–633.

Flaxman, S. M., Wacholder, A. C., Feder, J. L., and Nosil, P. (2014). Theoretical models of the influence of genomic architecture on the dynamics of speciation. *Molecular Ecology*, 23(16):4074–4088.

Gao, F., Ming, C., Hu, W., and Li, H. (2016). New software for the fast estimation of population recombination rates (fasteprr) in the genomic era. *G3: Genes, Genomes, Genetics*, 6(6):1563–1571.

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., Chen, Z., Mauceli, E., Hacohen, N., Gnirke, A., Rhind, N., Di Palma, F., Birren, B. W., Nusbaum, C., Lindblad-Toh, K., Friedman, N., and Regev, A. (2011). Full-length transcriptome assembly from RNA-seq data without a reference genome. *Nature Biotechnology*, 29(7):644–U130.

Guerrero, R. F. and Hahn, M. W. (2017). Speciation as a sieve for ancestral polymorphism. *Molecular Ecology*, 26(20):5362–5368.

Hackl, T., Hedrich, R., Schultz, J., and Förster, F. (2014). Proovread: large-scale high-accuracy pacbio correction through iterative short read consensus. *Bioinformatics*, 30(21):3004–3011.

Hawthorne, D. J. and Via, S. (2001). Genetic linkage of ecological specialization and reproductive isolation in pea aphids. *Nature*, 412(6850):904–907.

Holt, B. G., Emerson, B. C., Newton, J., Gage, M. J. G., and Cote, I. M. (2008). Stable isotope analysis of the *Hypoplectrus* species complex reveals no evidence for dietary niche divergence. *Marine Ecology Progress Series*, 357:283–289.

Huber, W., Von Heydebreck, A., Sueltmann, H., Poustka, A., and Vingron, M. (2002). Variance stabilization applied to microarray data calibration and to the quantification of

differential expression. *Bioinformatics*, 18 Suppl. 1:S96–S104.

Ihaka, R., Murrell, P., Hornik, K., Fisher, J. C., and Zeileis, A. (2016). *Colorspace: color space manipulation*. R package version 1.3-2.

Jakovlić, I. and Wang, W.-M. (2016). Expression of hox paralog group 13 genes in adult and developing *Megalobrama amblycephala*. *Gene Expression Patterns*, 21(2):63–68.

Jeong, S., Rokas, A., and Carroll, S. B. (2006). Regulation of body pigmentation by the abdominal-b hox protein and its gain and loss in drosophila evolution. *Cell*, 125(7):1387–1399.

Jia, L., Oh, E. C. T., Ng, L., Srinivas, M., Brooks, M., Swaroop, A., and Forrest, D. (2009). Retinoid-related orphan nuclear receptor ror beta is an early-acting factor in rod photoreceptor development. *Proceedings of the National Academy of Sciences of the United States Of America*, 106(41):17534–17539.

Kahle, D. and Wickham, H. (2013). Ggmap: spatial visualization with ggplot2. *The R Journal*, *5*(1):144–161.

Kajitani, R., Toshimoto, K., Noguchi, H., Toyoda, A., Ogura, Y., Okuno, M., Yabana, M., Harada, M., Nagayasu, E., Maruyama, H., Kohara, Y., Fujiyama, A., Hayashi, T., and Itoh, T. (2014). Efficient de novo assembly of highly heterozygous genomes from whole-genome shotgun short reads. *Genome Research*, 24(8):1384–1395.

Kiełbasa, S. M., Wan, R., Sato, K., Horton, P., and Frith, M. C. (2011). Adaptive seeds tame genomic sequence comparison. *Genome Research*, 21(3):487–493.

Kim, D., Langmead, B., and Salzberg, S. L. (2015). Hisat: a fast spliced aligner with low memory requirements. *Nature Methods*, 12(4):357–U121.

Kronforst, M. R., Young, L. G., Kapan, D. D., McNeely, C., O'neill, R. J., and Gilbert, L. E. (2006). Linkage of butterfly mate preference and wing color preference cue at the genomic location of wingless. *Proceedings of the National Academy of Sciences of the United States Of America*, 103(17):6575–6580.

Krzywinski, M. I., Schein, J. E., Birol, I., Connors, J., Gascoyne, R., Horsman, D., Jones, S. J., and Marra, M. A. (2009). Circos: an information aesthetic for comparative genomics. *Genome Research*.

Langmead, B. and Salzberg, S. L. (2012). Fast gapped-read alignment with bowtie 2. *Nature Methods*, 9(4):357–U54.

Lawrence, M., Gentleman, R., and Carey, V. (2009). Rtracklayer: an r package for interfacing with genome browsers. *Bioinformatics*, 25:1841–1842.

Lee, E., Helt, G. A., Reese, J. T., Munoz-Torres, M. C., Childers, C. P., Buels, R. M., Stein, L., Holmes, I. H., Elsik, C. G., and Lewis, S. E. (2013). Web apollo: a web-based genomic annotation editing platform. *Genome Biology*, 14(8):R93.

Leggett, R. M., Clavijo, B. J., Clissold, L., Clark, M. D., and Caccamo, M. (2014). Nextclip: an analysis and read preparation tool for nextera long mate pair libraries. *Bioinformatics*, 30(4):566–568.

Li, H. and Durbin, R. (2009). Fast and accurate short read alignment with burrows–wheeler transform. *Bioinformatics*, 25(14):1754.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and Subgroup, . G. P. D. P. (2009). The sequence alignment/map format and samtools. *Bioinformatics*, 25(16):2078–2079.

Lindenbaum, P. (2015). JVarkit: java-based utilities for Bioinformatics.

Lischer, H. E. L. and Excoffier, L. (2012). PGDspider: an automated data conversion tool for connecting population genetics and genomics programs. *Bioinformatics*, 28(2):298–299.

Love, M. I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with deseq2. *Genome Biology*, 15(12).

Luu, K. and Blum, M. (2017). *Pcadapt: fast principal component analysis for outlier detection*. R package version 3.0.4.

Malinsky, M., Challis, R. J., Tyers, A. M., Schiffels, S., Terai, Y., Ngatunga, B. P., Miska, E. A., Durbin, R., Genner, M. J., and Turner, G. F. (2015). Genomic islands of speciation separate cichlid ecomorphs in an east african crater lake. *Science*, 350(6267):1493–1498.

Martin, S. (2016). Genomics_general: general tools for genomic analyses; a github repository.

Mattar, P., Ericson, J., Blackshaw, S., and Cayouette, M. (2015). A conserved regulatory logic controls temporal identity in mouse neural progenitors. *Neuron*, 85(3):497–504.

McCartney, M. A., Acevedo, J., Heredia, C., Rico, C., Quenoville, B., Bermingham, E., and McMillan, W. O. (2003). Genetic mosaic in a marine species flock. *Molecular Ecology*, 12(11):2963–2973.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., and Depristo, M. A. (2010). The genome analysis toolkit: a mapreduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, 20(9):1297–1303.

Meier, J. I., Marques, D. A., Wagner, C. E., Excoffier, L., and Seehausen, O. (2018). Genomics of parallel ecological speciation in lake victoria cichlids. *Molecular Biology and Evolution*, 35(6):1489–1506.

Murrell, P. and Potter, S. (2017). *Gridsvg: export 'grid' graphics as svg*. R package version 1.6-0.

Nei, M. (1987). *Molecular evolutionary genetics*. Columbia University Press, New York.

Neuwirth, E. (2014). *Rcolorbrewer: colorbrewer palettes*. R package version 1.1-2.

Noor, M. A. F. and Feder, J. L. (2006). Speciation genetics: evolving approaches. *Nature Reviews Genetics*, 7(11):851–861.

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

Pante, E. and Simon-Bouhet, B. (2013). Marmap: a package for importing, plot-

ting and analyzing bathymetric and topographic data in r. *PLoS One*, 8(9):e73051. doi:10.1371/journal.pone.0073051.

Parichy, D. M., Mellgren, E. M., Rawls, J. F., Lopes, S. S., Kelsh, R. N., and Johnson, S. L. (2000). Mutational analysis of endothelin receptor b1 (rose) during neural crest and pigment pattern development in the zebrafish *Danio rerio*. *Developmental Biology*, 227(2):294–306.

Pebesma, E. J. and Bivand, R. S. (2005). Classes and methods for spatial data in R. *R News*, 5(2):9–13.

Pedersen, T. (2018). *Ggforce: accelerating 'ggplot2'*. R package version 0.1.2.

Pertea, G. (2015). Gffread: gff/gtf utility providing format conversions, region filtering, fasta sequence extraction and more; a github repository.

Picq, S., McMillan, W. O., and Puebla, O. (2016). Population genomics of local adaptation versus speciation in coral reef fishes (*Hypoplectrus* spp, Serranidae). *Ecology and Evolution*, 6(7):2109–2124.

Poelstra, J. W., Vijay, N., Hoeppner, M. P., and Wolf, J. B. W. (2015). Transcriptomics of colour patterning and coloration shifts in crows. *Molecular Ecology*, 24(18):4617–4628.

Potter, S. (2018a). *Grconvert: converting vector graphics*. R package version 0.1-0.

Potter, S. (2018b). *Grimport2: importing 'svg' graphics*. R package version 0.1-4.

Puebla, O., Bermingham, E., and 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(6):1405–1415.

Puebla, O., Bermingham, E., and Guichard, F. (2011). Perspective: matching, mate choice, and speciation. *Integrative and Comparative Biology*, 51(3):485–491.

Puebla, O., Bermingham, E., and Guichard, F. (2012). Pairing dynamics and the origin of species. *Proceedings of the Royal Society of London B: Biological Sciences*, 279(1731):1085–1092.

Puebla, O., Bermingham, E., Guichard, F., and Whiteman, E. (2007). Colour pattern as a single trait driving speciation in *Hypoplectrus* coral reef fishes? *Proceedings of the Royal Society of London B: Biological Sciences*, 274(1615):1265–1271.

Puebla, O., Bermingham, E., and McMillan, W. O. (2014). Genomic atolls of differentiation in coral reef fishes (*Hypoplectrus* spp., Serranidae). *Molecular Ecology*, 23(21):5291–5303.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., Maller, J., Sklar, P., De Bakker, P. I. W., Daly, M. J., and Sham, P. C. (2007). Plink: a tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics*, 81(3):559–575.

Quinlan, A. R. and Hall, I. M. (2010). Bedtools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6):841–842.

R Core Team (2017). *R: a language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rambaut, A. and Grass, N. C. (1997). Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. *Bioinformatics*, 13(3):235–238.

Randall, J. E. and Randall, H. A. (1960). Examples of mimicry and protective resemblance in tropical marine fishes. *Bulletin of Marine Science*, 10(4):444–480.

Ravinet, M., Faria, R., Butlin, R., Galindo, J., Bierne, N., Rafajlović, M., Noor, M., Mehlig, B., and Westram, A. (2017). Interpreting the genomic landscape of speciation: a road map for finding barriers to gene flow. *Journal of evolutionary biology*, 30(8):1450–1477.

Reaka-Kudla, M. (1997). The global biodiversity of coral reefs: a comparison with rain forests. *Biodiversity II: Understanding and protecting our biological resources*, 2:551.

Robertson, D. R. and Tassell, J. V. (2015). Shorefishes of the greater caribbean: online information system. Version 1.0. Smithsonian Tropical Research Institute, Balboa, Panamá. http://biogeodb.stri.si.edu/caribbean/en/research/index/range.

Roesti, M., Gavrilets, S., Hendry, A. P., Salzburger, W., and Berner, D. (2014). The genomic signature of parallel adaptation from shared genetic variation. *Molecular Ecology*, 23(16):3944–3956.

Roesti, M., Moser, D., and Berner, D. (2013). Recombination in the threespine stickleback genome—patterns and consequences. *Molecular Ecology*, 22(11):3014–3027.

Rudis, B. (2017). *Hrbrthemes: additional themes, theme components and utilities for 'ggplot2'*. R package version 0.1.0.

Saenko, S. V., Marialva, M. S., and Beldade, P. (2011). Involvement of the conserved hox gene antennapedia in the development and evolution of a novel trait. *EvoDevo*, 2(1):9.

Schnute, J. T., Boers, N., and Haigh, R. (2017). *Pbsmapping: mapping fisheries data and spatial analysis tools*. R package version 2.70.4.

Seehausen, O., Butlin, R. K., Keller, I., Wagner, C. E., Boughman, J. W., Hohenlohe, P. A., Peichel, C. L., Saetre, G.-P., Bank, C., Brannstrom, A., Brelsford, A., Clarkson, C. S., Eroukhmanoff, F., Feder, J. L., Fischer, M. C., Foote, A. D., Franchini, P., Jiggins, C. D., Jones, F. C., Lindholm, A. K., Lucek, K., Maan, M. E., Marques, D. A., Martin, S. H., Matthews, B., Meier, J. I., Most, M., Nachman, M. W., Nonaka, E., Rennison, D. J., Schwarzer, J., Watson, E. T., Westram, A. M., and Widmer, A. (2014). Genomics and the origin of species. *Nature Reviews Genetics*, 15(3):176–192.

Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., Zdobnov, E. M., Simao, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., and Zdobnov, E. M. (2015). Busco: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, 31(19):3210–3212.

Slatkin, M. (2008). Linkage disequilibrium — understanding the evolutionary past and

mapping the medical future. *Nature Reviews Genetics*, 9:477.

Slowikowski, K. (2017). *Ggrepel: repulsive text and label geoms for 'ggplot2'*. R package version 0.7.0.

Smit, A. F. A. and Hubley, R. (2015). Repeatmodeler open-1.0.

Smit, A. F. A., Hubley, R., and Green, P. (2015). Repeatmasker open-4.0.

Soneson, C., Love, M. I., and Robinson, M. D. (2015). Differential analyses for rna-seq: transcript-level estimates improve gene-level inferences. *F1000Research*, 4.

Tang, H., Zhang, X., Miao, C., Zhang, J., Ming, R., Schnable, J. C., Schnable, P. S., Lyons, E., and Lu, J. (2015). Allmaps: robust scaffold ordering based on multiple maps. *Genome Biology*, 16(1):1–15.

Tavera, J. and Acero, A. P. (2013). Description of a new species of *Hypoplectrus* (perciformes: Serranidae) from the southern gulf of mexico. *aqua: International Journal of Ichthyology*, 19(1):29–38.

The Uniprot Consortium (2015). Uniprot: a hub for protein information. *Nucleic Acids Research*, 43(D1):D204–D212.

Theodosiou, L., McMillan, W. O., and Puebla, O. (2016). Recombination in the eggs and sperm in a simultaneously hermaphroditic vertebrate. *Proceedings of the Royal Society of London B: Biological Sciences*, 283(1844).

Thresher, R. (1978). Polymorphism, mimicry, and the evolution of the hamlets (*Hypoplectrus*, Serranidae). *Bulletin of Marine Science*, 28(2):345–353.

Thummel, R., Li, L., Tanase, C., Sarras, M. P., and Godwin, A. R. (2004). Differences in expression pattern and function between zebrafish hoxc13 orthologs: recruitment of hoxc13b into an early embryonic role. *Developmental Biology*, 274(2):318–333.

Tine, M., Kuhl, H., Gagnaire, P.-A., Louro, B., Desmarais, E., Martins, R. S., Hecht, J., Knaust, F., Belkhir, K., Klages, S., et al. (2014). European sea bass genome and its variation provide insights into adaptation to euryhalinity and speciation. *Nature communications*, 5:5770.

Turner, T. L., Hahn, M. W., and Nuzhdin, S. V. (2005). Genomic islands of speciation in *Anopheles gambiae*. *PLoS Biology*, 3(9):e285.

Van Belleghem, S. M., Rastas, P., Papanicolaou, A., Martin, S. H., Arias, C. F., Supple, M. A., Hanly, J. J., Mallet, J., Lewis, J. J., Hines, H. M., Ruiz, M., Salazar, C., Linares, M., Moreira, G. R. P., Jiggins, C. D., Counterman, B. A., McMillan, W. O., and Papa, R. (2017). Complex modular architecture around a simple toolkit of wing pattern genes. *Nature Ecology & Evolution*, 1:52.

Van Der Auwera, G. A., Carneiro, M. O., Hartl, C., Poplin, R., Del Angel, G., Levy-Moonshine, A., Jordan, T., Shakir, K., Roazen, D., Thibault, J., Banks, E., Garimella, K. V., Altshuler, D., Gabriel, S., and Depristo, M. A. (2013). From fastq data to high-confidence variant calls: the genome analysis toolkit best practices pipeline. *Current Protocols in Bioinformatics*, (SUPL.43).

Victor, B. (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*, 5:2–19.

Vij, S., Kuhl, H., Kuznetsova, I. S., Komissarov, A., Yurchenko, A. A., Van Heusden, P., Singh, S., Thevasagayam, N. M., Prakki, S. R. S., Purushothaman, K., Saju, J. M., Jiang, J., Mbandi, S. K., Jonas, M., Hin Yan Tong, A., Mwangi, S., Lau, D., Ngoh, S. Y., Liew, W. C., Shen, X., Hon, L. S., Drake, J. P., Boitano, M., Hall, R., Chin, C.-S., Lachumanan, R., Korlach, J., Trifonov, V., Kabilov, M., Tupikin, A., Green, D., Moxon, S., Garvin, T., Sedlazeck, F. J., Vurture, G. W., Gopalapillai, G., Kumar Katneni, V., Noble, T. H., Scaria, V., Sivasubbu, S., Jerry, D. R., O'brien, S. J., Schatz, M. C., Dalmay, T., Turner, S. W., Lok, S., Christoffels, A., and Orbán, L. (2016). Chromosomal-level assembly of the asian seabass genome using long sequence reads and multi-layered scaffolding. *PLoS Genetics*, 12.

Vijay, N., Bossu, C. M., Poelstra, J. W., Weissensteiner, M. H., Suh, A., Kryukov, A. P., and Wolf, J. B. W. (2016). Evolution of heterogeneous genome differentiation across multiple contact zones in a crow species complex. *Nature Communications*, 7.

Warnes, G. R., Bolker, B., Bonebakker, L., Gentleman, R., Liaw, W. H. A., Lumley, T., Maechler, M., Magnusson, A., Moeller, S., Schwartz, M., and Venables, B. (2016). *Gplots: various R programming tools for plotting data*. R package version 3.0.1.

Warnes, G. R., Bolker, B., Gorjanc, G., Grothendieck, G., Korosec, A., Lumley, T., Macqueen, D., Magnusson, A., Rogers, J., and Others (2017). *Gdata: various R programming tools for data manipulation*. R package version 2.18.0.

Weir, B. S. and Cockerham, C. C. (1984). Estimating f-statistics for the analysis of population-structure. *Evolution*, 38(6):1358–1370.

Whiteman, E. A., Côté, I. M., and Reynolds, J. D. (2007). Ecological differences between hamlet (*Hypoplectrus*: Serranidae) colour morphs: between-morph variation in diet. *Journal of Fish Biology*, 71(1):235–244.

Whiteman, E. A. and Gage, M. J. G. (2007). No barriers to fertilization between sympatric colour morphs in the marine species flock *Hypoplectrus* (Serranidae). *Journal of Zoology*, 272(3):305–310.

Wickham, H. (2016a). *Ggplot2: elegant graphics for data analysis*. Springer-Verlag New York.

Wickham, H. (2016b). *Gtable: arrange 'grobs' in tables*. R package version 0.2.0.

Wickham, H. (2017). *Tidyverse: easily install and load the 'tidyverse'*. R package version 1.2.1.

Wickham, H. (2018). *Scales: scale functions for visualization*. R package version 0.5.0.9000.

Wickham, H., Francois, R., Henry, L., and Müller, K. (2017). *Dplyr: a grammar of data manipulation*. R package version 0.7.4.

Wilke, C. (2017). *Cowplot: streamlined plot theme and plot annotations for 'ggplot2'*. R package version 0.9.2.

Wolf, J. B. and Ellegren, H. (2017). Making sense of genomic islands of differentiation in light of speciation. *Nature Reviews Genetics*, 18(2):87.

Wood, D. E. and Salzberg, S. L. (2014). Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biology*, 15(3):1–12.

Wringe, B., Stanley, R., Jeffery, N., Anderson, E., and Bradbury, I. (2016). *Parallelnewhybrid: an R package for the parallelization of hybrid detection using newhybrids*.

Xie, Y. (2014). Knitr: a comprehensive tool for reproducible research in R. In Stodden, V., Leisch, F., and Peng, R. D., editors, *Implementing Reproducible Computational Research*. Chapman And Hall/Crc. ISBN 978-1466561595.

Xie, Y. (2015). *Dynamic documents with R and knitr*. Chapman And Hall/Crc, Boca Raton, Florida, 2nd edition. ISBN 978-1498716963.

Xie, Y. (2016). *Bookdown: authoring books and technical documents with R markdown*. Chapman And Hall/Crc, Boca Raton, Florida. ISBN 978-1138700109.

Xie, Y. (2018). *Knitr: a general-purpose package for dynamic report generation in r*. R package version 1.20.

Yokoyama, S. (2008). Evolution of dim-light and color vision pigments. *Annual Review of Genomics and Human Genetics*, 9:259–282.

Yu, G. (2018). *Scatterpie: scatter pie plot*. R package version 0.0.9.

Zhao, C. and Malicki, J. (2011). Nephrocystins and mks proteins interact with ift particle and facilitate transport of selected ciliary cargos. *Embo Journal*, 30(13):2532–2544.

Zhou, X. and Stephens, M. (2012). Genome-wide efficient mixed-model analysis for association studies. *Nature Genetics*, 44:821.

– 4 –

The Evolution of Microendemism in a Reef Fish (*Hypoplectrus maya*)

Benjamin M. Moran [1,2], **Kosmas Hench** [1], Robin S. Waples [3], Marc P. Höppner [4], Carole C. Baldwin [5], W. Owen McMillan [6], Oscar Puebla [1,6,7]

1 GEOMAR Helmholtz Centre for Ocean Research Kiel, Marine Evolutionary Ecology, Düsternbrooker Weg 20, 24105 Kiel, Germany

2 Department of Marine and Environmental Sciences, Northeastern University, 360 Huntington Ave., Boston, MA 02115, USA

3 Northwest Fisheries Science Center, National Marine Fisheries Service, NOAA, Seattle, Washington, USA

4 Institute of Clinical Molecular Biology, Kiel University, Rosalind-Franklin-Strasse 12, 24105 Kiel, Germany

5 Department of Vertebrate Zoology, National Museum of Natural History, Smithsonian Institution, 10th and Constitution Ave, NW, Washington, DC 20560, USA

6 Smithsonian Tropical Research Institute, Apartado Postal 0843-03092, Panamá, República de Panamá

7 University of Kiel, Faculty of Mathematics and Natural Sciences, Christian-Albrechts-Platz 4, 24118 Kiel, Germany

This study was originally published in *Molecular Ecology*. While the figures were re-drawn for this thesis, no additional changes were made except re-labeling colors/ shapes on updated figures. The re-print within this thesis is in agreement with *John Wiley and Sons* under license number 4826441243245.

Original publication

Moran, B.M. *et al*. (2019). The evolution of microendemism in a reef fish (*Hypoplectrus maya*). *Molecular Ecology*, 28(11):2872–2885. doi: 10.1111/mec.15110.

Abstract

Marine species tend to have extensive distributions, which are commonly attributed to the dispersal potential provided by planktonic larvae and the rarity of absolute barriers to dispersal in the ocean. Under this paradigm, the occurrence of marine microendemism without geographic isolation in species with planktonic larvae poses a dilemma. The recently described Maya hamlet (*Hypoplectrus maya*, Serranidae) is exactly such a case, being endemic to a 50-km segment of the Mesoamerican Barrier Reef System (MBRS). We use whole-genome analysis to infer the demographic history of the Maya hamlet and contrast it with the sympatric and pan-Caribbean black *(H. nigricans)*, barred *(H. puella)* and butter *(H. unicolor)* hamlets, as well as the allopatric but phenotypically similar blue hamlet *(H. gemma)*. We show that *H. maya* is indeed a distinct evolutionary lineage, with genomic signatures of inbreeding and a unique demographic history of continuous decrease in effective population size since it diverged from congeners just ∼3000 generations ago. We suggest that this case of microendemism may be driven by the combination of a narrow ecological niche and restrictive oceanographic conditions in the southern MBRS, which is consistent with the occurrence of an unusually high number of marine microendemics in this region. The restricted distribution of the Maya hamlet, its decline in both census and effective population sizes, and the degradation of its habitat place it at risk of extinction. We conclude that the evolution of marine microendemism can be a fast and dynamic process, with extinction possibly occurring before speciation is complete.

Keywords: hamlets, *Hypoplectrus*, endemism, speciation, demographic inference.

4.1. Introduction

Islands, in the sense of isolation, are much more rare in the ocean than on land. Many marine organisms are planktonic for a portion of their life cycle, allowing them to cross the pelagic expanses that limit terrestrial organisms (Palumbi, 1992). Planktonic dispersal leads to a greater rate of cosmopolitanism than in terrestrial communities and a corresponding rarity of microendemic species (that is, species that are endemic to unusally small areas), as colonists are less likely to exist in isolation long enough for reproductive isolation to evolve (Kay and Palumbi, 1987; Randall, 1998; Rocha and Bowen, 2008). Cases of marine microendemism that have been identified are generally in taxa with short or non-existent planktonic phases (Paulay and Meyer, 2002; Meyer et al., 2005). This paradigm suggests planktonic larval duration (PLD) as a potential driver of range size in reef fishes; however, syntheses have shown that these factors are poorly correlated (Lester and Ruttenberg, 2005; Mora et al., 2012; Luiz et al., 2013). Instead, growing knowledge of marine dispersal suggests that reef fishes show lower average dispersal distances than expected based on PLD (Jones et al., 2009). This deviation is driven in part by local oceanographic processes, and by natal homing and habitat selectivity among larvae (Leis, 2006). As such, bio-physical coupling of oceanic currents and larval behavior is currently regarded as the primary determinant of dispersal patterns (Jones et al., 2009).

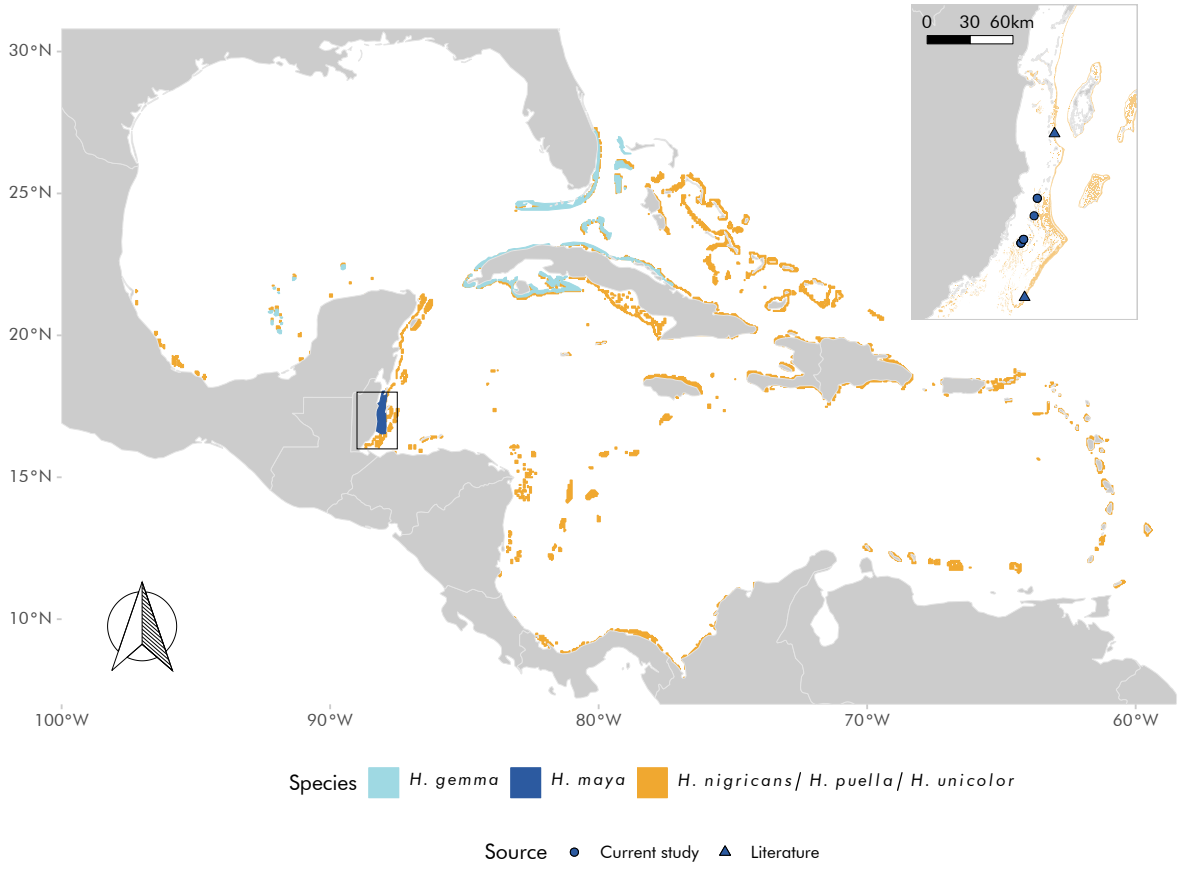


Figure 4.1: Ranges of the *Hypoplectrus* species considered in this study. *H. puella*, *H. nigricans*, and *H. unicolor* occur throughout the Greater Caribbean (orange). *H. gemma* (light blue) is restricted to the Northern Caribbean, and *H. maya* (dark blue) to a section of the Mesoamerican Barrier Reef System (MBRS) in Belize. Inset: reports of *H. maya* from the literature (triangles) and the current study (circles, note that some of these locations had been reported before). Distribution of pan-Caribbean hamlets extrapolated from the WCMC-008 Global Distribution of Coral Reefs (UNEP-WCMC et al., 2010).

While mean dispersal in reef fishes is often limited to the scale of tens of kilometers, the potential for rare long-distance dispersal events remains (Simpson et al., 2014). Consequently, rates of microendemism are generally lower in marine fishes than in terrestrial animals (Kay and Palumbi, 1987; Paulay and Meyer, 2002). Marine fishes with exceptionally small ranges are generally found in the most isolated islands of habitat, or in zones where circulation patterns lead to a unidirectional transport of larvae towards inhospitable habitat (Roberts et al., 2002). Nonetheless, given the importance of local-scale oceanography and behavior in determining reef-fish dispersal, marine microendemism might emerge without severe geographic isolation.

The Maya hamlet, *Hypoplectrus maya* (Serranidae), is a clear example of such microendemism without geographic isolation. This reef fish was identified by P.S. Lobel in the coastal lagoon of the Mesoamerican Barrier Reef System (MBRS) in Belize in 1993, and thereafter variously referred to as the "Belize" (Heemstra et al., 2002), "Belize Blue" (Ramon et al., 2003) or "Mayan" (Smith et al., 2003) hamlet. Domeier (1994) identified the Maya hamlet as a population of the similarly-colored blue hamlet, *H. gemma*, which oc-

curs in Florida, Cuba, and the northern Yucatan (Aguila-Perera and Tuz-Sulub (2010); Figure 4.1); however, based on the lack of melanized upper and lower margins on the caudal fin which are diagnostic of *H. gemma*, Lobel (2011) described the Maya hamlet (*H. maya*) as a distinct species. The Maya hamlet has been reported on the lagoon side of the MBRS between Wee Wee Cay and the Sapodilla Cays, corresponding to a range of approximately 50 linear kilometers of reef (though one vagrant individual was collected northward in 2010 on the seaward reef wall off Alligator Cay, (Lobel 2011; Figure 4.1). This is an exceptionally small range considering that reef-fish distributions typically range between 2000 and 13,000 km in the Atlantic (Ruttenberg and Lester, 2015). As of 2003, the species was described as "common and abundant" in the Pelican Cays and the surrounding Rhomboidal Cays (Smith et al., 2003). In particular, Lobel (2011) noted the frequent occurrence of *H. maya* among mangrove roots, suggesting an ecological specificity to the complex array of shallow coral ridges and mangroves found within these cays.

The restricted range of the Maya hamlet contrasts sharply with the pan-Caribbean barred (*H. puella*), black (*H. nigricans*), and butter (*H. unicolor*) hamlets, which are also found in the coastal lagoon of the MBRS in Belize (Holt et al. 2010; Figure 4.1). These species are sympatric throughout most of their range, with ongoing gene flow maintaining low levels of genetic differentiation despite strong assortative mating (Puebla et al., 2007, 2014; Hench et al., 2019). More broadly, the genus includes a total of 19 species that vary widely in range size and abundance (Holt et al., 2010). It encompasses the entire continuum of genomic divergence, from species that are almost genetically identical (Barreto and McCartney, 2008; Puebla et al., 2012a, 2014) to well-diverged species (Victor, 2012; Tavera and Acero, 2013). Hamlets are nonetheless reproductively isolated from a behavioral perspective through strong assortative mating (Fischer, 1980; Puebla et al., 2007; Barreto and McCartney, 2008), and described as valid species by ichthyologists (Lobel, 2011; Victor, 2012; Tavera and Acero, 2013; Victor and Marks, 2018). We herein follow this current, accepted nomenclature and consider them species, even though reproductive isolation is not always complete. In this regard we note that the biological species concept does not necessarily imply absolute isolation; many species that are considered good species do hybridize in nature, and hybridization also occurs above the species level (Mallet, 2005).

Hamlets are very similar from an ecological perspective (Whiteman and Gage, 2007; Holt et al., 2008), yet the color patterns that characterize the different species appear to be ecologically relevant through crypsis (Thresher, 1978; Fischer, 1980) and mimicry (Randall and Randall, 1960; Thresher, 1978; Puebla et al., 2007, 2018). It has been suggested that speciation in the hamlets may be driven by a combination of natural (Thresher, 1978; Puebla et al., 2007) and sexual (Puebla et al., 2012a) selection, but it remains unclear whether the hamlets diverged in full sympatry or in allopatry followed by secondary contact as suggested by Domeier (1994). Regardless, the two-week planktonic larval stage of the hamlets (Domeier, 1994), the occurrence of hybrid spawnings in natural populations (Fischer, 1980; Puebla et al., 2007; Barreto and

McCartney, 2008; Puebla et al., 2012a), the apparent lack of post-zygotic barriers between species (Whiteman and Gage, 2007), the identification of hybrid and backcrossed individuals in the field (Hench et al., 2019), and the low levels of genetic differentiation among sympatric species (McCartney et al., 2003; Ramon et al., 2003; Puebla et al., 2012a) as well as allopatric populations within species (Puebla et al., 2008, 2009; Picq et al., 2016) indicate that gene flow is pervasive among Caribbean hamlets, in contrast to the genetic isolation usually implied by microendemism. Given the biogeographic disparity observed in the hamlets, *H. maya* presents an ideal opportunity to understand the processes by which marine microendemism might arise or persist in the absence of geographic isolation.

The recent publication of a chromosome-resolution reference genome for the hamlets offers a new opportunity to understand the evolution of marine microendemism from a genomic perspective (Hench et al., 2019). Here, we test whether *H. maya* represents an evolutionarily distinct lineage from its three sympatric pan-Caribbean congeners (*H. puella*, *H. nigricans*, and *H. unicolor*) and from the allopatric but phenotypically similar species *H. gemma*. Considering the restricted range of *H. maya*, we also test the hypothesis that it may present genomic signatures of inbreeding (in terms of nucleotide diversity, heterozygosity, coefficient of inbreeding, relatedness and runs of homozygosity) relative to its more widely distributed congeners. Following the same line of thought, we then infer the demographic histories of the five species using Markovian Coalescent analyses of past effective population size (N_e). Finally, we estimate the recent effective population size of *H. maya* and discuss the potential causes and consequences of microendemism in marine species.

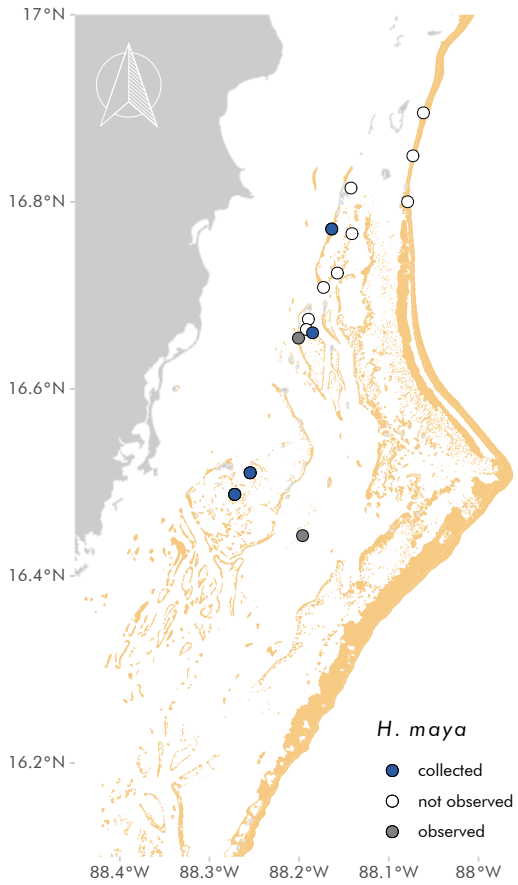
4.2. Methods

Sampling

We considered whole genomes of 12 individuals each of *H. puella*, *H. nigricans*, and *H. unicolor* from the Belize portion of the MBRS, available from a previous study (Hench et al., 2019). To this we added 10 *H. maya* samples collected in Belize in May 2017 under STRI IACUC protocol 2017-0101-2020-2, Northeastern University IACUC protocol 17-0206R, and Belize Fisheries Department permit 000026-17, as well as 5 *H. gemma* samples collected in the Florida Keys in July 2017 under the prior IACUC protocols, NOAA ONMS permit 2017-042, and Florida FWCC permit SAL-17-1890A-SR. Gill tissue for sequencing was preserved in salt-saturated DMSO buffer, and entire fish were preserved in 10% formalin until accessioned and stored as voucher specimens in 70-75% ethanol at the Smithsonian National Museum of Natural History (Suppl. Tab. 4.2).

Field Surveys

In Belize, *H. maya* surveys were carried out opportunistically in May 2017 in the center of the species' known distribution (Figure 4.1; Suppl. Fig. 4.1). Surveys targeted reef and mangrove habitat, including the specific cays where *H. maya* had been previously reported (Domeier, 1994; Smith et al., 2003; Lobel, 2011). In the latter case, snorkelers surveyed all mangrove habitat encircling the cay and



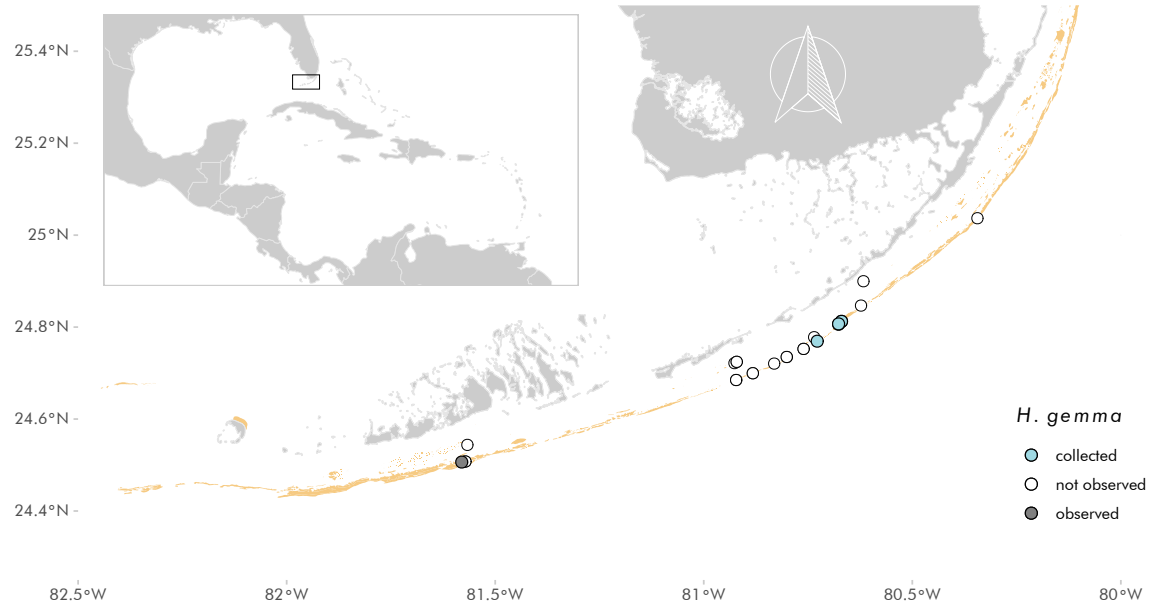
Suppl. Figure 4.1: Map of *H. maya* survey and sampling locations within the Belize section of the Mesoamerican Barrier Reef System. Locations surveyed are marked by circles. Circle color denotes sites at which *H. maya* was not observed (white), observed (gray) or collected (blue). Available reef habitat (orange) gathered from the WCMC-008 Global Distribution of Coral Reefs (UNEP-WCMC et al., 2010).

those interior ponds which were accessible by boat. A combination of snorkeling (0-5 m) and SCUBA diving (5-15 m) was moreover used to haphazardly survey reef habitat on the MBRS exterior wall, fringing reefs around cays, and patch reefs within the lagoon.

Field surveys were also conducted in the Florida Keys in July 2017 using 4 x 100 m linear SCUBA transects to assess the densities of *H. gemma* and all other hamlets, and to evaluate whether densities and relative abundances changed over the last 15 years (Suppl. Fig. 4.2). Average densities over all transects were compared to the yearly averages from stationary surveys (15 m^2 diameter) conducted throughout the Keys by the Florida Keys Reef Visual Census, which took place in all years between 2002 and 2016, except for 2013 and 2015 (Smith et al., 2011). To test for changes in community composition over time, years were divided into two periods with equal sampling effort, 2002–2008 and 2009–2017. The dissimilarity of community compositioning was tested using PERMANOVA (Anderson, 2001) with 999 permutations and the Bray-Curtis measure of ecological distance, as implemented in the vegan package in R (Oksanen et al., 2018).

Genotyping

Hypoplectrus gemma and *H. maya* genomic DNA was extracted from gill tissue using a Qiagen MagAttract High Molecular Weight Kit and sequenced to a mean genome-wide coverage of $\sim$ 22X on an Illumina HiSeq 4000 (PE, 2x151) at the Institute of Clinical Molecular Biology (IKMB) in Kiel, Germany (Suppl. Tab. 4.2), following the same sequencing approach that was taken for the



Suppl. Figure 4.2: Map of *H. gemma* survey and sampling locations within the Florida Keys. Locations of transects surveys are marked by circles. Circle color denotes sites at which *H. gemma* was not observed (white), observed (gray) or collected (blue). Available reef habitat (orange) gathered from the WCMC-008 Global Distribution of Coral Reefs (UNEP-WCMC et al., 2010). Inset shows location of sampling within the greater Caribbean.

H. puella, *H. nigricans*, and *H. unicolor* samples (Hench et al., 2019). Raw reads were mapped to the *H. puella* reference genome (Hench et al., 2019) using BWA v0.7.12 (Li and Durbin, 2009), with an average mapping efficiency of 97.24% for *H. maya*, 98.62% for *H. gemma*, 99.16% for *H. unicolor*, 99.20% for *H. nigricans*, and 99.21% for *H. puella*. All samples considered in this study were genotyped together with a workflow adapted from GATK Best Practices, with hard filters for quality control (Van der Auwera et al., 2013). Specifically, reads were filtered to remove outliers in the ratio of Phred-scaled probability of the genotype to sequencing depth (QD < 2.5), the Phred-scaled *p*-value from a Fisher's Exact Test for strand bias (FS > 25.0), the Strand Odds Ratio (SOR > 3.0), the root-mean-squared mapping quality across samples (MQ < 58.0 or > 62.0) and the *U*-values from rank-sum tests for differences in mapping quality (|MQRankSum| > 2.5) and variant position within read (|ReadPosRankSum| > 2.5) in reference vs. alternate alleles. Additional filtering with respect to minor allele frequency and coverage was specific to each analysis and mentioned explicitly when applied. Two VCF data sets were generated: one including all (variant and invariant) callable sites (555 379 974 sites, 2.5% missing data), and another including only biallelic SNPs (11 419 868 sites, 0.4% missing data). A phased data set was generated from the biallelic VCF by using phase-informative reads and SHAPEIT2 (Delaneau et al., 2013).

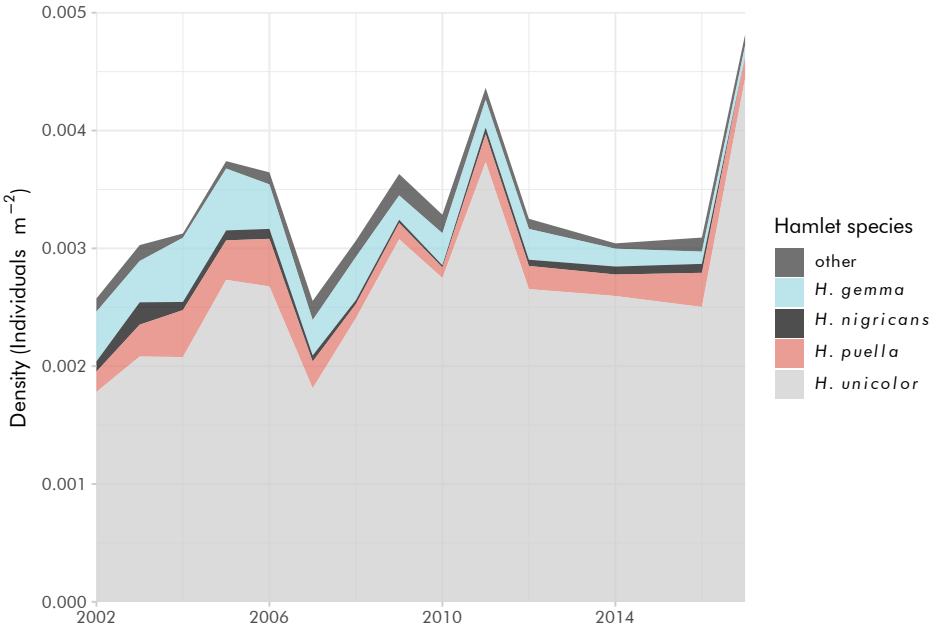
Population Genomics

To estimate the extent of physical linkage, r^2 was calculated using VCFtools between all pairs of SNPs with a minor allele frequency greater than 10% in 200 randomly placed windows of 30 kb each. Principal Component Analysis (PCA) was performed using the R package SNPRelate (Zheng et al., 2012).

This analysis was conducted on all samples, repeated considering the Belize samples only (i.e. excluding *H. gemma*), and repeated again in Belize considering a minimum distance of 15 kb between SNPs to minimize the effect of linkage. Genome-wide differentiation (F_{ST}) was calculated for all pairs of species using the weighted mean approach implemented in VCFtools (Weir and Cockerham, 1984). F_{ST} was also calculated in sliding windows (50 kb window with 5 kb increments) between *H. maya* and each other species in order to explore the distribution of differentiation across the genome. Heterozygosity and inbreeding coefficient (F) were calculated for each sample using VCFtools, and the data set including all callable sites was used to calculate nucleotide diversity (π) in non-overlapping 10 kb windows for each species using VCFtools (Danecek et al., 2011). The data set including all callable sites was also used to calculate absolute divergence (d_{XY}; Nei 1987) for each species pair in non-overlapping 50 kb windows using popgenWindows.py (Martin, 2016). Genome-wide absolute divergence was then calculated by averaging over the windows using the number of SNPs as weights. Relatedness was calculated between all pairs of individuals using the Maximum Likelihood Estimation (MLE) method implemented in SNPRelate as well as the unadjusted A_{jk} statistic in VCFtools (Zheng et al., 2012; Yang et al., 2010). Runs of homozygosity (ROH) greater than 150 kb in length were identified using PLINK and located in the genome after filtering SNPs for a minor allele frequency > 5% across all individuals (Purcell et al., 2007). iHH12 (Torres et al., 2018) was also calculated over the entire genome (50 kb windows, 5 kb increments) using selscan (Szpiech and Hernandez, 2014) to look for signs of recent positive selection.

Demographic Inference

Demographic history was inferred for each species using the Multiple Sequentially Markovian Coalescent (MSMC) v2.0.0 (Schiffels and Durbin, 2014). Data preparation was performed following `https://github.com/stschiff/msmc-tools`. Based on the recommendations of Nadachowska-Brzyska et al. (2016), variant sites were filtered for a minimum depth of 10X, and a maximum depth of twice the individual's mean depth. MSMC inference may be affected by deviations from neutrality caused by selection (Schrider et al., 2016); as such, runs were repeated after excluding the regions above the 99.90th F_{ST} percentile identified in Hench et al. (2019). Each MSMC run included 4 individuals, with the exception of 2 runs of 3 individuals in *H. maya* (Suppl. Tab. 4.4). Each individual was included in only one MSMC analysis; replicate runs are therefore independent samplewise. To explore the history of divergence among species, we also considered the cross-population coalescence rate, scaled relative to the within-population coalescence rates (relative cross-coalescence rates, Schiffels and Durbin 2014). Cross-coalescence rates were inferred for the maximum possible number of independent runs for each species pair, considering two individuals per species (four individuals per run). All MSMC runs were performed with a time segmentation pattern of $1 \cdot 2 + 25 \cdot 1 + 1 \cdot 2 + 1 \cdot 3$, and the average of Watterson's estimator across input data sets, $\theta = 2.55 * 10^{-3}$. To explore whether the recent demographic trends observed in MSMC were an artifact of phasing switch errors, we also

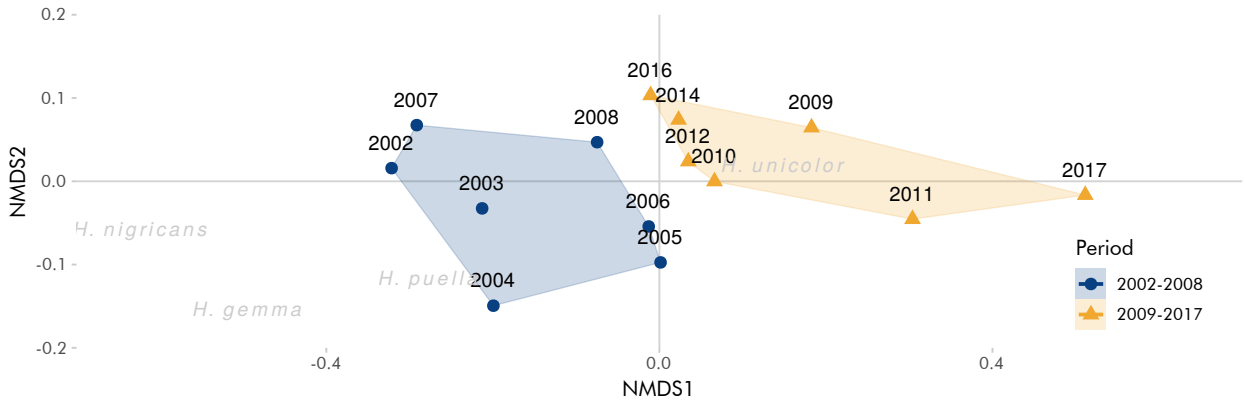


Suppl. Figure 4.3: Density of *Hypoplectrus* spp. on Florida Keys coral reefs, 2000–2017. 2017 data were collected through linear transects by the authors. All other data (2002–2016) gathered from stationary counts in the Florida Keys Reef Visual Census. Note the decline in relative abundance of *H. gemma*.

applied SMC++ v1.14.0, an extension of the Sequentially Markovian Coalescent that does not rely on phasing (Terhorst et al., 2017). A single SMC++ composite likelihood estimate for each species was created from the product of estimates across chromosomes, and across each possible "distinguished individual" in a species (see Terhorst et al. 2017). In both MSMC and SMC++, the mutation rate was set at $\mu = 3.7 * 10^{-8}$, based on the closest relative for which the value was known (Liu et al., 2016). Generation times (that is, the mean age of successfully reproducing individuals) for hamlets are uncertain, but likely fall between 1 and 3 years based on size, taxonomy, and habitat. Due to the resultant uncertainty, time is presented in terms of generations, with potential years on a secondary axis.

To complement estimates of past effective population size (N_e), we used a novel whole-genome implementation of recent N_e estimation based on linkage disequilibrium (Hill, 1981; Waples, 2006). We first used GATK to subset the biallelic SNP set by species, then selected sites with no missing genotype calls and minor allele counts > 2 (i.e. minor allele frequency > 0.1 in *H. maya*). Each SNP set was then randomly subset into 100 non-overlapping sets, to which N_e estimations were applied independently. We utilized a new feature of the LD method in NeEstimator v2.1 (released December 2017), calculating N_e based on only interchromosomal comparisons (Do et al., 2014). Confidence intervals were obtained from the per-individual jackknife of Jones et al. (2016), as well as the distribution of N_e across the 100 SNP subsets. This analysis was applied to all species except for *H. gemma* due to the low sample size for this species.

All scripts to reproduce our results from the raw data are available at `https://github.com/benmoran11/hamlets_endemism.git`.



Suppl. Figure 4.4: Non-metric multidimensional scaling (NMDS) of *Hypoplectrus* spp. densities in the Florida Keys, 2002-2017. Comparisons were drawn between surveys from 2002-2008 (blue circles) and 2009-2017 (orange triangles). Species abbreviations (gray text) indicate directions in which each species drives the ordination.

4.3. Results

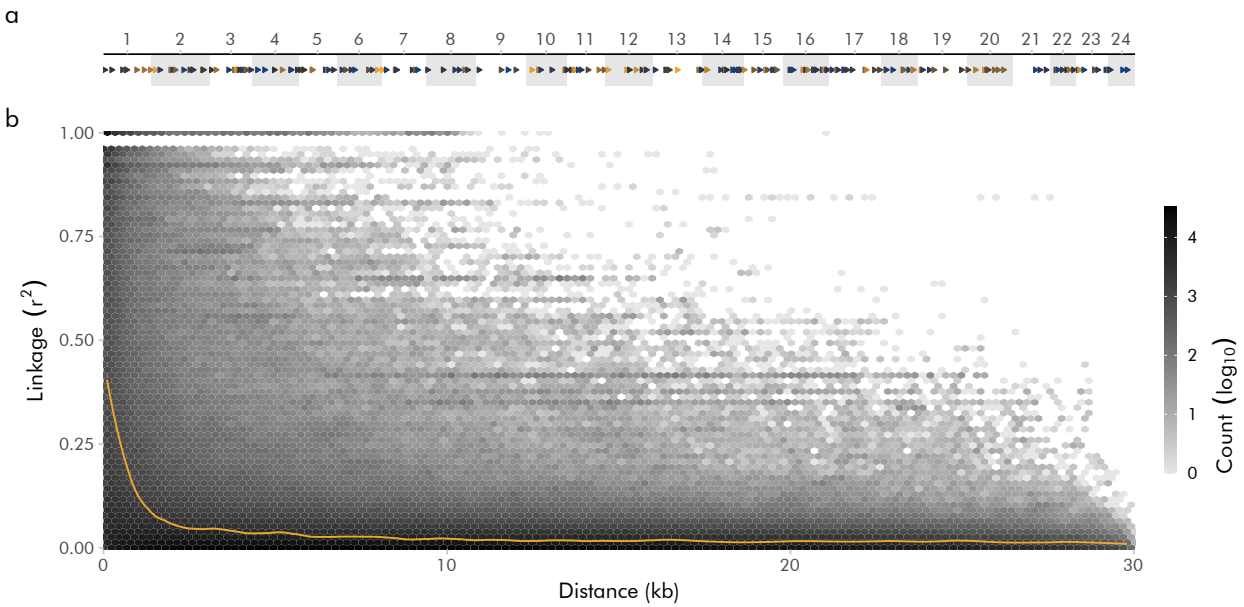
Field Surveys

Only two Maya hamlets were sighted in the Pelican Cays and surrounding Rhomboidal Cays where *H. maya* was described as "common and abundant" by Smith et al. (2003). One individual was a juvenile found on the reef flat adjacent to Little Cat Cay at a depth of 1.5 m next to an *Orbicella* coral head. The other individual was found in proximity to a barrel sponge surrounded by *Acropora cervicornis* rubble at a depth of 3 m in "Tunicate Cove", a honeycomb of coral ridges adjacent to Cat Cay where Lobel (2011) collected the holotype and eight paratypes over seven years. In contrast, *H. maya* was the most abundant hamlet species in the shallow (1-5 m) *A. cervicornis* patch reefs near Laughing Bird Cay National Park, where the majority of samples were collected for this study. Two individuals were also sighted at a depth of 2 m on an *Orbicella*-dominated fringing reef in Bread and Butter Cay (Suppl. Fig. 4.1).

In Florida, 27 non-overlapping transects were conducted between Geiger Key and French Reef, encompassing the majority of the range surveyed by the FL Reef Visual Census (Key West to Key Largo; Suppl. Fig. 4.2). Total mean density of hamlets in 2017 was 4.8 $\pm$ 1.0 (SE) fish 1000 m^{-2}, while RVC estimates in 2002-2016 fell between 2.5 and 4.5 fish 1000 m^{-2} (Suppl. Fig. 4.3). Hamlet community composition changed significantly between the first and second half of the temporal data set (Suppl. Fig. 4.4; PERMANOVA $P = 0.002$), with *H. unicolor* increasing in relative abundance at the expense of *H. gemma*. Between the two periods, mean *H. gemma* densities declined by more than 50%, from 0.41 $\pm$ 0.03 (SE) to 0.18 $\pm$ 0.03 (SE) fish 1000 m^{-2}.

Population Genomics

Our linkage analysis indicates that physical linkage decays rapidly within 5 kb (Suppl. Fig. 2). Genome-wide PCA showed clear clustering of *H. maya*, *H. gemma*, and *H. nigricans*, with partial overlap between *H. puella* and *H. unicolor* (Figure 4.2a). Similar pat-



Suppl. Figure 4.5: Decay in linkage disequilibrium with physical distance, estimated from 200 randomly placed windows of 30 kb each. a) Location of the windows along the genome (note that for visualization purposes the triangles exceed the actual extent of the windows). **b)** The shading of the hexagonal bins indicates the log_{10} count for each combination of distance and r^2 values. Orange line: LOWESS (Locally Weighted Scatterplot Smoothing) regression of a subset of the original data (the $5.9 \cdot 10^6$ pairwise comparisons were divided into 1.5 kb bins and 1000 pairwise LD values were sampled within each bin, providing a subset of $20 \cdot 10^3$ pairwise comparisons that were used for smoothing). Physical linkage decays rapidly within 5 kb.

terns were obtained when considering only SNPs > 15 kb apart to minimize physical linkage (Suppl. Fig. 4.6). Genome-wide differentiation was greatest between *H. maya* and *H. gemma* ($F_{ST} = 0.060$), lowest between *H. puella* and *H. unicolor*, ($F_{ST} = 0.004$), and intermediate for the other species pairs ($F_{ST} = 0.014 - 0.040$; Table 4.1). Sliding-window analysis revealed heterogeneous patterns of differentiation between *H. maya* and the four other species, with an accumulation of differentiation on linkage groups (LGs) 8 and 9 (likely due to large inversions, Hench et al. 2019) and a number of sharp peaks, some of which were repeated across different species comparisons (Suppl. Fig. 4.7a). The genomic regions above the 99.99^{th} F_{ST} percentile in comparisons involving *H. maya* are highlighted in Suppl. Fig. 4.7 and the genes found within these regions are listed in Suppl. Tab. 4.3.

Table 4.1: Estimates of genome-wide differentiation and divergence among the *Hypoplectrus* species considered in this study; F_{ST} above the diagonal, and d_{XY} below.

Species	*H. gemma*	*H. maya*	*H. nigricans*	*H. puella*	*H. unicolor*
H. gemma	—	0.060	0.040	0.037	0.033
H. maya	0.00379	—	0.039	0.026	0.028
H. nigricans	0.00378	0.00360	—	0.016	0.014
H. puella	0.00379	0.00357	0.00360	—	0.004
H. unicolor	0.00379	0.00360	0.00361	0.00359	—

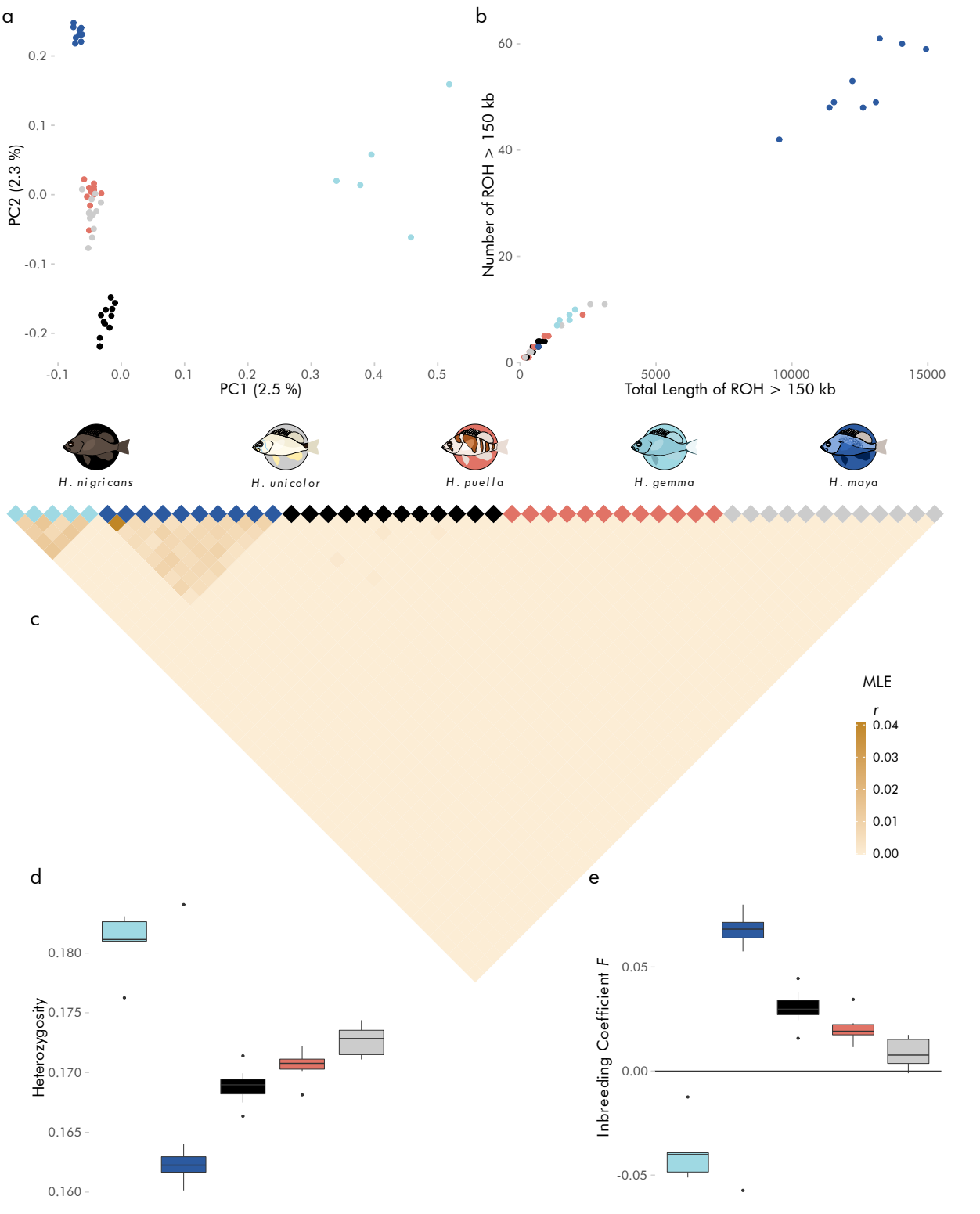
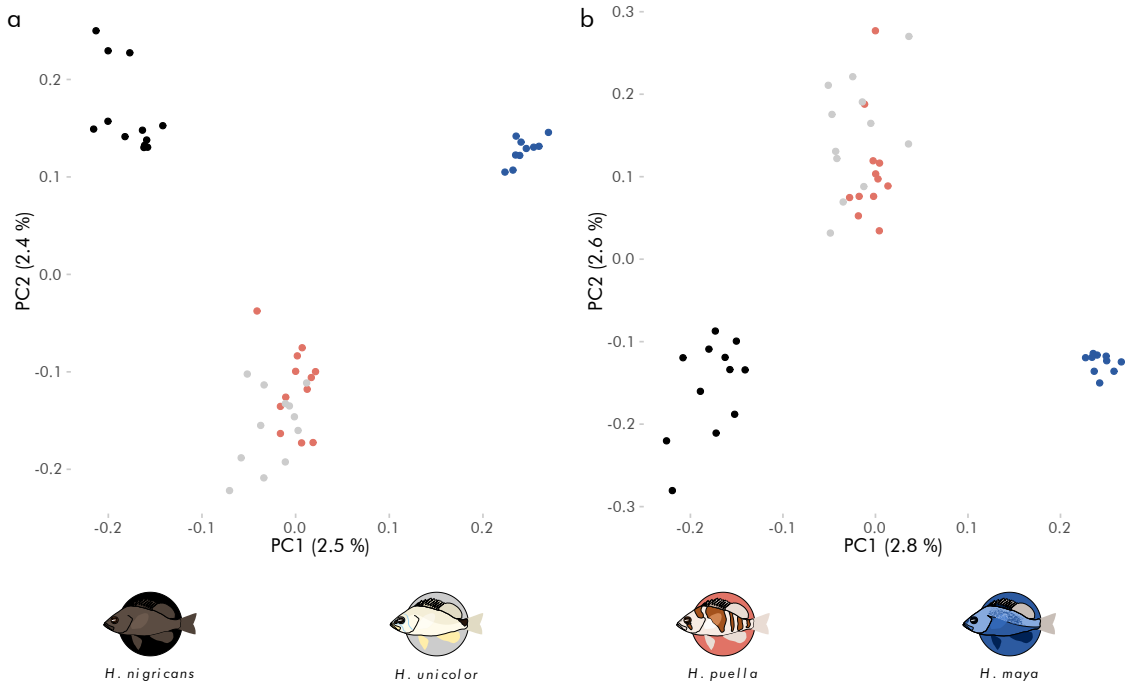


Figure 4.2: Population genomics of five *Hypoplectrus* species. a) Principal Component Analysis (PCA) based on whole genome data from all individuals in this study. Proportion of explained variance for the first two PCs listed on axes. **b)** Runs of Homozygosity (ROH) > 150 kb in each individual. The total number of ROH with length > 150 kb is plotted against the summed length of those ROH. **c)** Genome-wide Maximum Likelihood Estimation (MLE) of relatedness between all pairs of samples. **d)** Heterozygosity, calculated genome-wide for each individual. **e)** Inbreeding coefficient F, calculated genome-wide for each individual. Central bars represent median values, and boxes 25th – 75th percentile intervals. Whiskers show data within 1.5 × interquartile range, and dots are outliers beyond this range.



Suppl. Figure 4.6: Principal Component Analysis (PCA) based on whole genome data from Belize samples only (i.e. excluding *H. gemma*). Proportion of explained variance for the first two PCs listed on axes. **a)** PCA based on the entire data set. **b)** PCA based on filtered SNP set with a miniumim distance of 15 kb between individual SNPs to limit the effect of linkage.

Heterozygosity was depressed in *H. maya* (median = 0.162) relative to other Belizean hamlets (median = 0.169 – 0.173) and to *H. gemma* (median = 0.181; Figure 4.2d). Nucleotide diversity was also lowest in *H. maya* (median π = 0.0047 versus 0.0049 – 0.0053 in other species), but the difference was small relative to the variation among 10 kb windows (Suppl. Fig. 4.8). For both maximum likelihood estimates and A_{jk}, mean relatedness was highest in *H. gemma* (mean MLE $r = 0.012$, $A_{jk} = 0.054$), followed by *H. maya* (mean MLE r = 0.008, A_{jk} = 0.032) and the other Belizean hamlet species (mean MLE $r = 0$, A_{jk} = -0.011 – 0.003 Figure 4.2c, Suppl. Fig. 4.9). A positive outlier was observed between two *H. maya* individuals, suggesting inbreeding beyond background relatedness (MLE r = 0.041, A_{jk} = 0.093; Figure 4.2c, Suppl. Fig. 4.9). Inbreeding in *H. maya* was also suggested by the higher inbreeding coefficients observed in this species (median F = 0.068) relative to the other Belizean species (median F = 0.008 – 0.030, Figure 4.2e, note that this includes *H. unicolor* which is rare in Belize) as well as the markedly higher number of runs of homozygosity > 150 kb in *H. maya* relative to other species (Figure 4.2b). The ROH were located all over the genome, indicating that the higher prevalence of ROH in *H. maya* is a genome-wide phenomenon. Nevertheless, ROH were disproportionately represented on LG2, LG9 and LG12, matching the F_{ST} patterns (Suppl. Fig. 4.7b). This result was confirmed by the integrated haplotype homozygosity pooled (iHH12, Torres et al. 2018, Suppl. Fig. 4.7c), which is often used to detect signs of recent positive selection and thereby suggests that selection is also playing a role in these regions. The blue hamlet showed negative inbreeding coefficients (median = -

0.040), yet this result should be interpreted with caution due to the low sample size for this species ($n = 5$).

Demographic Inference

We used MSMC to identify demographic trends leading to current biogeographic patterns. The most ancient and two most recent time segments provided highly inconsistent N_e estimates within species (Suppl. Fig. 4.10) and were therefore not considered, since this suggests unreliable inference (S. Schiffels, personal communication). All species presented very similar trends earlier than 3000 generations before present (gbp), suggesting that they diverged only recently (Figure 4.3). Following an expansion until 2000 gbp, *H. maya* N_e decreased continuously to a minimum of 12000 at 290 gbp (Figure 4.3). The *H. gemma* and *H. nigricans* populations also decreased beginning 2000 gbp, but rebounded to a final N_e of 50000 and 100000 $\pm$ 15000 (mean $\pm$ SE across *H. nigricans* runs), respectively. In contrast, *H. puella* and *H. unicolor* N_e increased to final values of 120000 $\pm$ 18000 and 110000 $\pm$ 13000, respectively (Figure 4.3). SMC++ analysis, which does not rely on phasing, confirmed that these general trends were not due to phasing switch errors (Suppl. Fig. 4.12). Though the heuristic calculation of time points limited SMC++ inference to 10^3–10^5 gbp, we nonetheless observed a population expansion beginning 10^4 gbp in all species, a sharp decline in *H. maya*, and a limited decline in *H. gemma* (Suppl. Fig. 4.12). The most notable differences in the SMC++ results were large N_e fluctuations between 10^5 and 10^4 gbp and a shift towards older times for the beginning of the declines in *H. maya* and *H. gemma* (Suppl. Fig. 4.12).

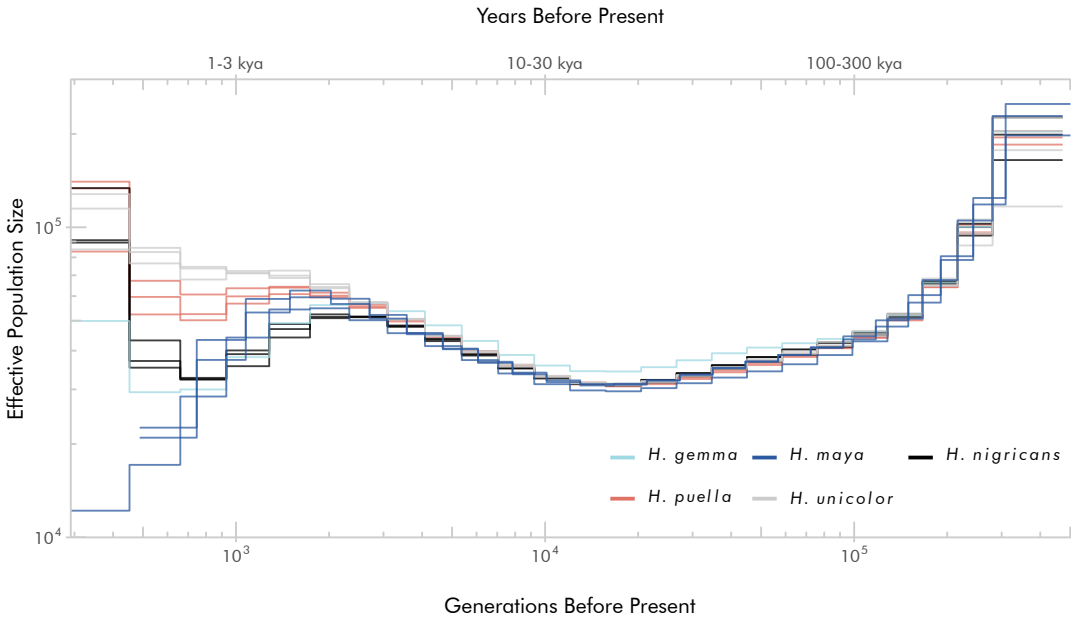
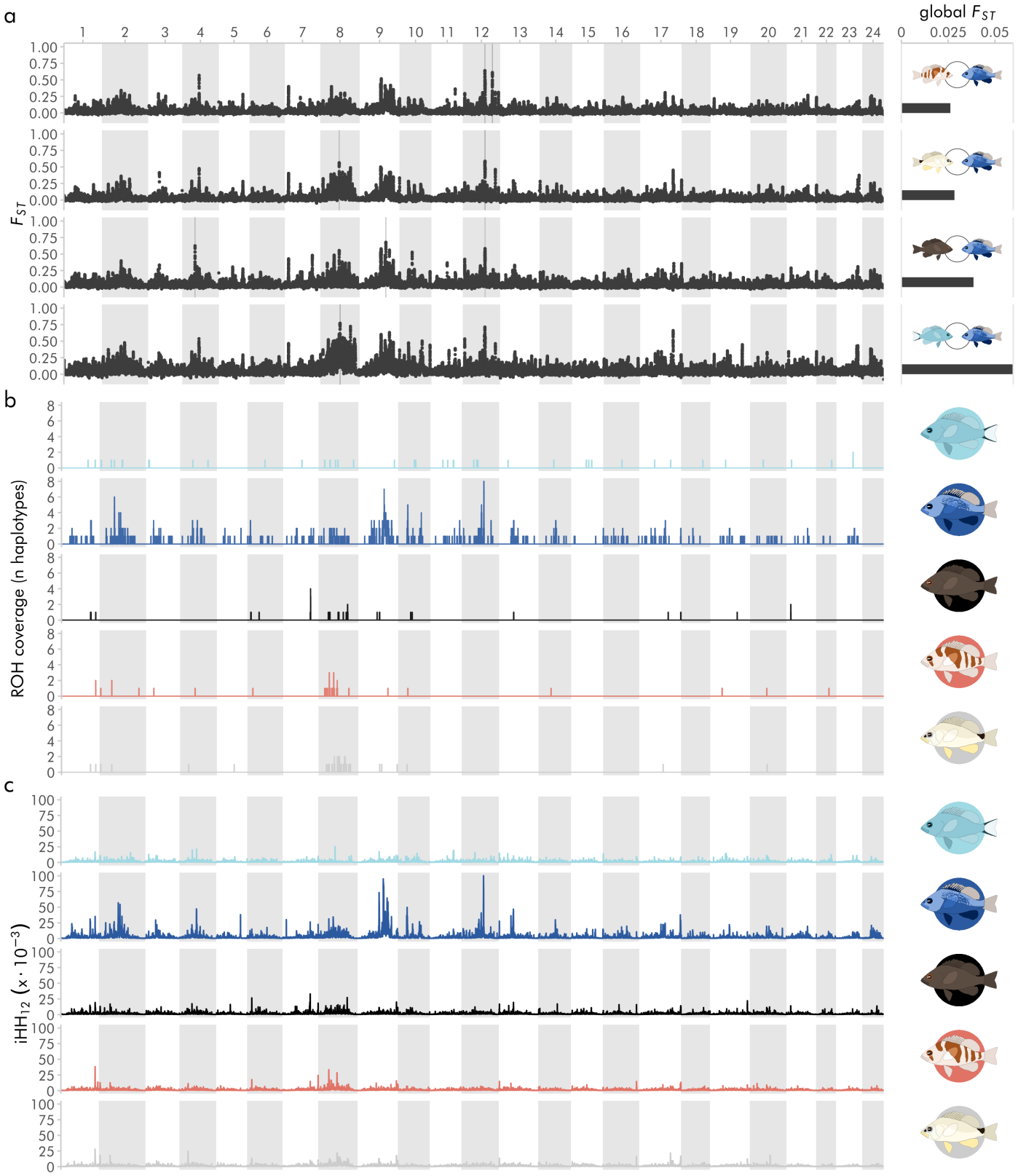
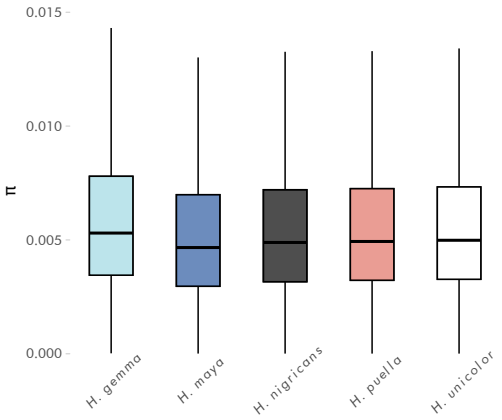


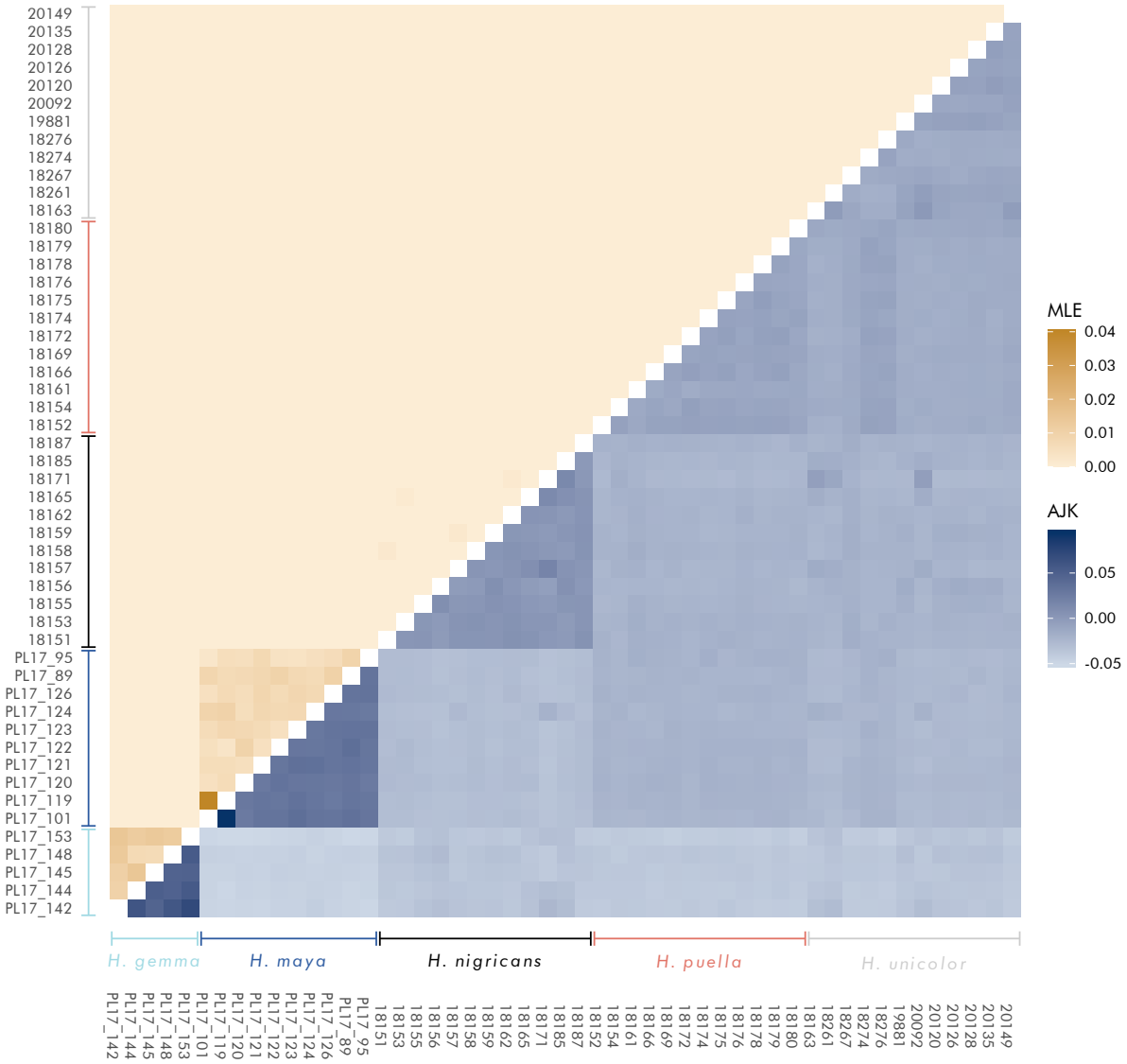
Figure 4.3: MSMC inference of effective population size over time in the five species. Each analysis is based on 3-4 genomes and each genome is used in only one analysis. All estimates are scaled with a per-site mutation rate $\mu = 3.7 \cdot 10^{-8}$. The most ancient and two most recent time segments are omitted due to unreliable inference (see text).



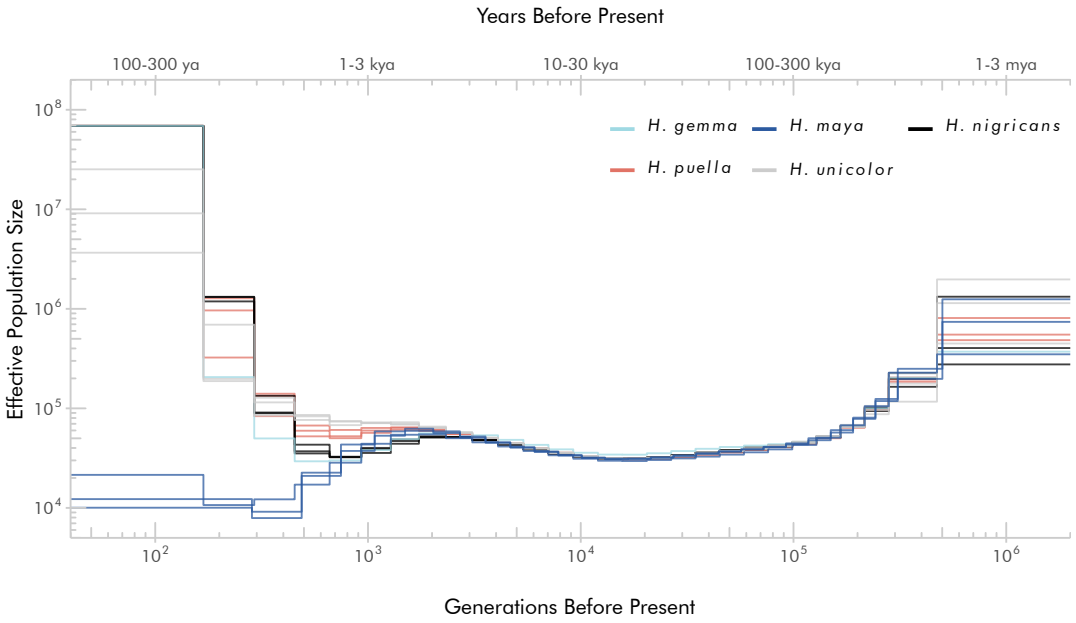
Suppl. Figure 4.7: Genome-wide patterns. Alternating white and grey blocks represent the 24 linkage groups (LGs). a) Patterns of genomic differentiation between *H. maya*, three sympatric species (*H. puella*, *H. nigricans* and *H. unicolor*), and the allopatric but phenotypically similar *H. gemma*. F_{ST} values correspond to the weighted mean per 50 kb window with 5 kb increments. Horizontal bars on the right represent genome-wide weighted mean F_{ST}. Genomic intervals above the 99.99^{th} F_{ST} percentile in each comparison are highlighted with a vertical line. **b)** Location and coverage of Runs of Homozygosity (ROH, **Fig. 2b**) for each species. **c)** Integrated haplotype homozygosity pooled (iHH12) for each species (50 kb windows, 5 kb increments).



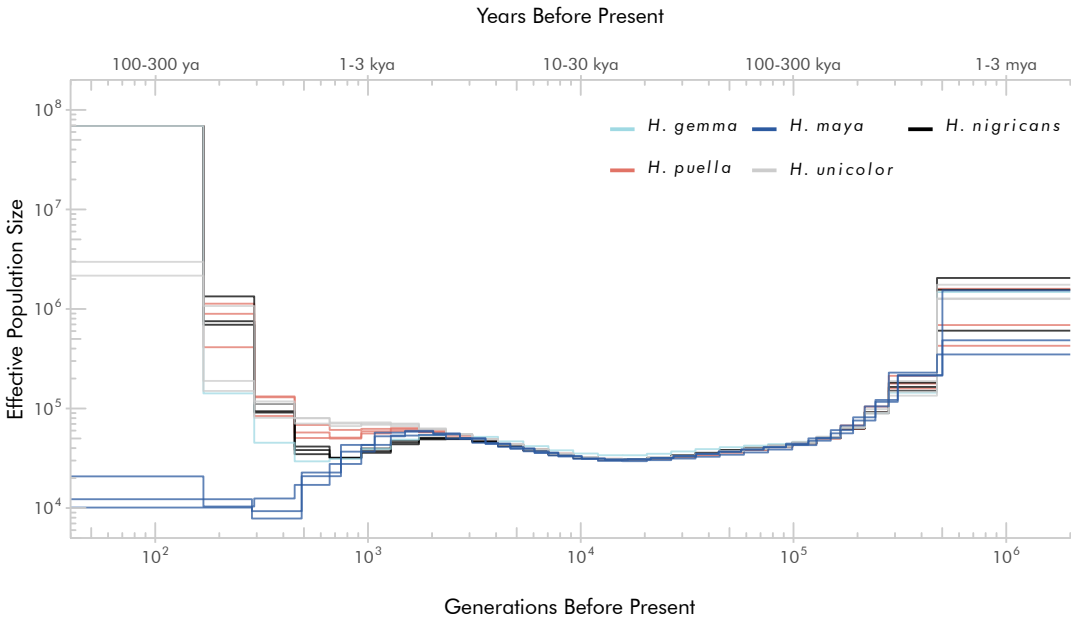
Suppl. Figure 4.8: Distribution of nucleotide diversity (π) in the five species considered in this study, calculated in non-overlapping 10 kb windows across all individuals in a population. Outlier windows are not shown. Central bars represent median values, and boxes $25^{th} - 75^{th}$ percentile intervals. Whiskers show data within 1.5 $\times$ interquartile range.



Suppl. Figure 4.9: Comparison between estimates of the coefficient of relationship among all samples by Maximum Likelihood Estimation (MLE) and A_{jk} estimator. Individual IDs are shown in the axis labels, with species indicated by colored lines. gem = *H. gemma*, may = *H. maya*, nig = *H. ngricans*, pue = *H. puella* and uni = *H. unicolor*.



Suppl. Figure 4.10: Raw MSMC estimates of effective population size over time in the five species considered in this study.



Suppl. Figure 4.11: MSMC estimates without masking of 99.90^{th} percentile F_{ST} peaks, as identified in Hench et al. (2019).

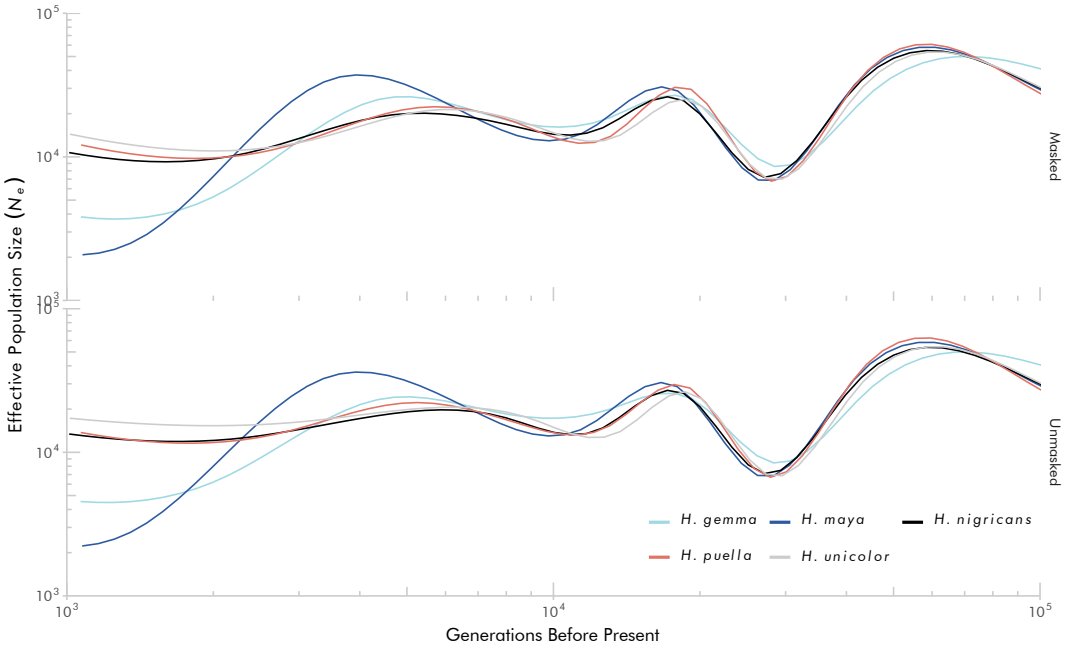
For both analyses, results were qualitatively identical with and without the most diverged genomic regions that are likely under selection (Suppl. Fig. 4.10; Suppl. Fig. 4.11; Suppl. Fig. 4.12).

The cross-coalescence results indicate that *H. gemma* diverged from the other species within $\sim$ 6000 gbp, followed by *H. nigricans* ($\sim$ 5000 gbp) and *H. maya* ($\sim$ 3000 gbp, Figure 4.4). The barred and butter hamlets appear to have diverged even more recently ($\sim$ 2000 gbp), yet these results should be interpreted with caution due to ongoing gene flow between these two species in Belize (Hench et al. 2019, which may explain the observed cross-coalescence rates $>$ 1.0). Relative cross-coalescence was $>$ 0.01 in all comparisons until $<$ 500 gbp, and remained $>$ 0.05 throughout inference in two *H. puella*–*H. unicolor* runs (Figure 4.4). As such, MSMC relative cross-coalescence supports other evidence of ongoing gene flow within the genus, especially between *H. puella* and *H. unicolor*.

For the estimation of recent N_e, quality filters left 3,296,967 suitable variant sites in *H. maya*, which were split into 100 non-overlapping data sets. Median estimated N_e was 1584 individuals, with a minimum of 1002, and a maximum of 9478 (Figure 4.5). Based on the Jones et al. (2016) jackknife variance method, NeEstimator estimated that the effective degrees of freedom associated with the 100 subsets ranged from 229647 to 532857; jackknife 95% confidence intervals had lower bounds between 277 and 528, and a consistent upper bound of infinity (Figure 4.5). In contrast, the 100 replicates provided an empirical 95% CI of 1073 – 4426 effective individuals (Figure 4.5). All 100 analyses for *H. puella*, *H. unicolor* and *H. nigricans* produced N_e point estimates, as well as lower and upper confidence bounds, of infinity.

4.4. Discussion

Our data confirm that *H. maya* represents a rare case of microendemism in reef fishes. From the moment of its scientific documentation, this species was confused with the phenotypically similar *H. gemma* of the northern Caribbean (Domeier, 1994). The diagnostic color pattern used to describe the new species and distinguish it from *H. gemma* (absence of black margins on the caudal fin, Lobel 2011) is only found within the MBRS; however, such characteristics are strained as taxonomic identifiers in the hamlets, where intermediate phenotypes, polymorphism, and regional variants of described species are frequently observed. In particular, black margins on the caudal fin are polymorphic within other hamlet species and populations (O. Puebla, personal observation). As such, we sought first to establish the status of *H. maya* as a distinct evolutionary unit. Our analyses demonstrate that *H. maya* and *H. gemma* are distinct evolutionary lineages, despite their phenotypic similarity. In fact, whole-genome differentiation between these two species is markedly higher than any other allopatric or sympatric comparison within this study, and *H. maya* is also differentiated from the other three sympatric pan-Caribbean hamlets (Table 4.1; Figure 4.2). The Maya hamlet can therefore be considered a separate species, so far as the biological species concept applies to the low differentiation and ongoing gene flow regime within *Hypoplectrus*.



Suppl. Figure 4.12: SMC++ estimates of effective population sizes over time in the five *Hypoplectrus* species considered in this study. Inference was repeated with (top) and without (bottom) masking of 99.90^{th} percentile F_{ST} peaks, as identified in Hench et al. (2019).

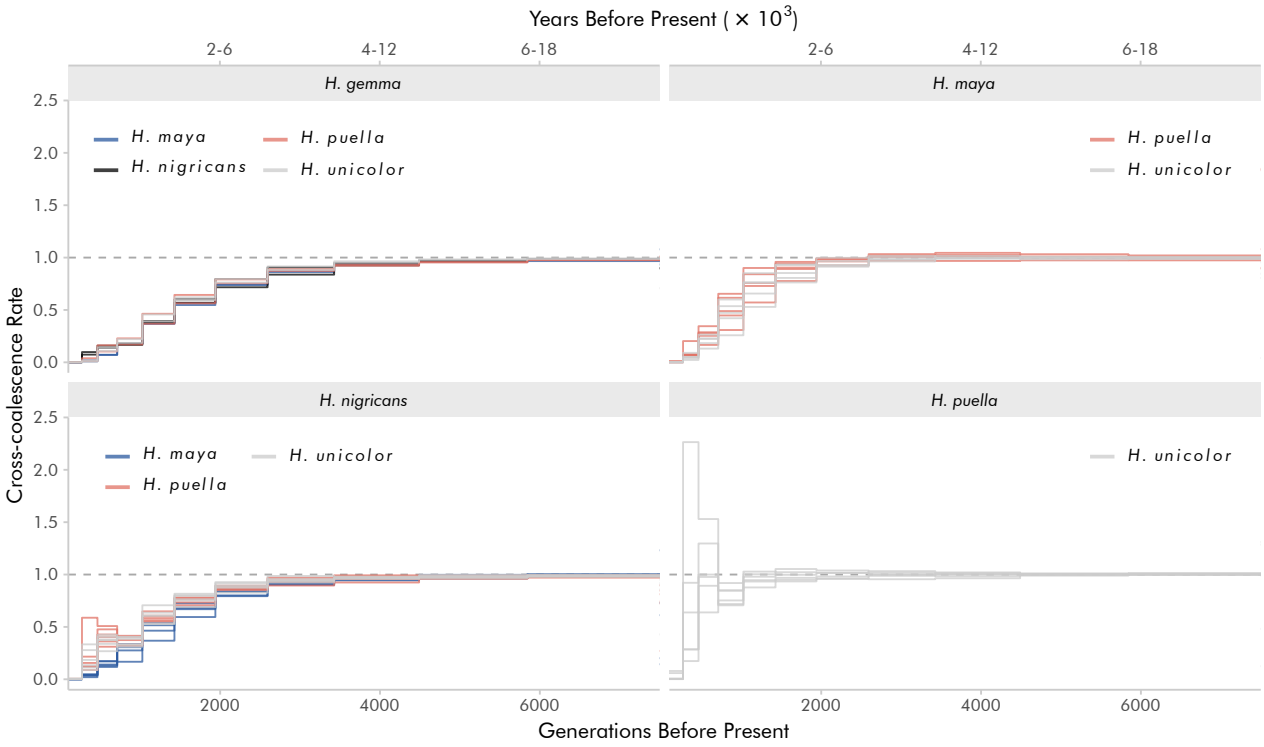


Figure 4.4: MSMC cross-coalescence inference of divergence times between all pairs of species. Each line represents an independent run including 2 individuals from each species. Panel headers identify the first species in the comparison, and colors the second. All estimates are scaled with a per-site mutation rate $\mu = 3.7 * 10^{-8}$. In a given time interval, a relative cross-coalescence rate of 1 (dashed line) indicates totally shared ancestry, and a rate of 0 indicates no shared ancestry.

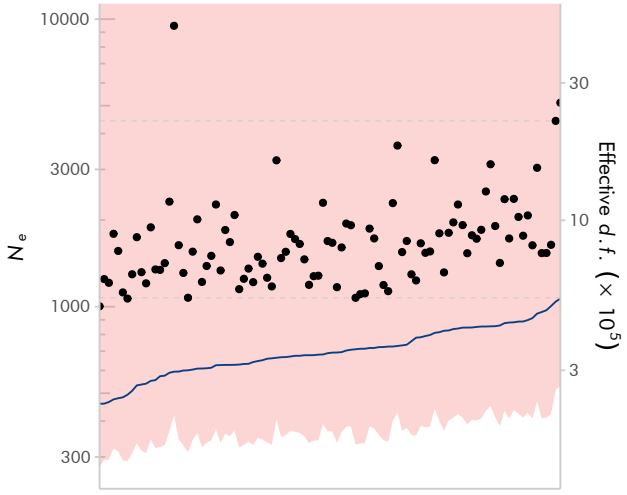


Figure 4.5: Estimates of *H. maya* recent effective population size from inter-chromosomal LD among 100 non-overlapping SNP subsets. N_e point estimates (black points) are ordered by effective degrees of freedom (blue line) inferred from the individual-wise jackknife procedure of Jones et al. (2016). Corresponding N_e 95% CIs (orange shading) extend to positive infinity in all estimates. Empirical 95% CI is denoted by dashed horizontal lines. Both vertical axes are log-scaled.

The Evolution of Microendemism

Considering the restricted distribution of *H. maya* and its recent divergence, it provides a rare window into the evolution of marine microendemism. The heterogeneous landscape of genomic differentiation between *H. maya* and other *Hypoplectrus* species suggests that *H. maya* evolved under the effect of selection and may be locally adapted (Suppl. Fig. 4.7). Some of the highly differentiated regions evidenced here have been previously identified, and include genes involved in vision (*rorb*) and pigmentation (*sox10*) that may play a role in reproductive isolation through visually-based assortative mating (Hench et al., 2019). We also note the presence of a sharp peak of differentiation on LG07 centered on the androgen receptor (*AR*) gene, which, although not above the 99.99th F_{ST} percentile, is consistently and exclusively observed in comparisons involving *H. maya* (Suppl. Fig. 4.7a). A iHH12 signal was also observed at this locus (Suppl. Fig. 4.7c), suggesting that it is under positive selection. Androgens are involved in the development of sex-specific traits, including vision (Shao et al., 2014) and pigmentation (Lindsay et al., 2011). It remains to be shown whether this is the case in the hamlets, which have a very specific simultaneously hermaphroditic mating system whereby individuals reciprocally trade eggs for fertilization (Fischer, 1980).

All measures point to reduced genomic diversity and increased inbreeding in *H. maya* relative to pan-Caribbean congeners (Figure 4.2). The Maya hamlet shows decreased heterozygosity, higher inbreeding coefficients, and more runs of homozygosity than sympatric congeners, as expected following a bottleneck or ongoing population decline (Nei et al., 1975; Frankham, 1998). In contrast to the three pan-Caribbean species, background levels of relatedness are also > 0 in *H. maya*. Furthermore, we identified one pair of Maya hamlets that are much more related than background levels ($r = 0.041$, which corresponds to the level of relatedness that is expected between second cousins with a most recent common ancestor 3 generations ago; Wright 1922; Figure 4.2). These individuals were collected 34 km apart, at opposite ends of the sampling area, which is within the estimated dispersal potential of Belizean hamlets across three generations (Puebla et al., 2012b). Median nucleotide diversity was also 4-12% lower in *H. maya* than

congeners (Suppl. Fig. 4.8). This difference may appear small, particularly in comparison to observed π in other taxa: *H. maya* nucleotide diversity is $\sim$ 2 times higher than that observed in *Ficedula* flycatchers, and $\sim$ 6 times higher than that in humans (Primmer et al., 2002; International SNP Map Working Group, 2001). This high diversity is expected within the framework of high marine effective population sizes, and is concordant with our inferred demographies: the hamlets experienced a pre-divergence bottleneck of $N_e \approx 30 \times 10^4$ (Figure 4.3), as opposed to 20×10^4 and 1×10^4 in flycatchers and humans, respectively (Nadachowska-Brzyska et al., 2016; Li and Durbin, 2011).

We note that *H. gemma* presents striking population genomic patterns, with higher levels of heterozygosity and background relatedness, and lower (negative) inbreeding coefficients relative to the four other species (Figure 4.2). We suggest that the high heterozygosity and apparent outbreeding observed in this species may be associated with the mixing of two lineages, from the Gulf of Mexico and Caribbean, in the Florida Keys (Ramon et al., 2003). As for the high levels of relatedness, they may be due to the ongoing decline of *H. gemma* populations in the Florida Keys documented by the transect data (Suppl. Fig. 4.3; Suppl. Fig. 4.4). We nevertheless reiterate caution with these hypotheses since they rely on only five *H. gemma*.

The analysis of present-day diversity and divergence is complemented by an understanding of the historical population dynamics in which they arose. Our approach allowed us to infer *Hypoplectrus* demographic histories up to $<$ 300 generations before present, with a likely historical range of $\sim$ 300 – 900 years ago (Figure 4.3). Regardless of uncertainty in *Hypoplectrus* generation times, inference provided clear support for widely divergent demographic trends in *H. maya*, beginning near the last glacial maximum. While pan-Caribbean species began a growth trajectory ending with effective population sizes around 100000, *H. maya* began a monotonic decrease to $N_e \approx$ 12000. In contrast, *H. gemma* N_e declined to $\sim$ 30000, and rebounded to $\sim$ 50000. The divergent trajectories of these taxa provide further support for their evolutionary distinction. Cross-coalescence rates, too, support the developing picture of *Hypoplectrus* as an ongoing speciation event. Our analyses suggest four independent divergence windows, all falling during or after the last glacial maximum (Figure 4.4). Extended gene flow is also suggested by this coalescent approach, with gene flow continuing into the current millennium in all lineages, and ongoing between *H. puella* and *H. unicolor*, the species pair between which high-probability hybrid and back-crossed individuals have been previously identified (Hench et al., 2019). An explicit analysis of the history of gene flow—which may be complex—is beyond the scope of this study, and we note that the decrease in N_e inferred in *H. maya* may also be interpreted in terms of a decrease in gene flow from other hamlet species and populations. Regardless, given the recent divergence of *H. maya*, it is likely that it arose within the MBRS and is thereby neoendemic to this area.

Recent Effective Population Size

The estimation of recent effective population size from linkage disequilibrium using whole-

genome data has been limited by the computational scale of the necessary number of pairwise comparisons, as well as physical linkage, which decreases the effective degrees of freedom presented by each pair of loci (Waples et al., 2016). To eliminate bias due to physical linkage, we considered only interchromosomal comparisons. The remaining effects of non-independence among pairwise comparisons of loci were accounted for by the per-individual jackknife procedure of Jones et al. (2016), which calculates "effective degrees of freedom" and corresponding confidence intervals. In addition, we leveraged the scale of our data set to calculate N_e estimates from 100 non-overlapping sets of markers, allowing an empirical evaluation of uncertainty in our estimate. These replicates display much less uncertainty than the jackknife confidence intervals would suggest; though no finite upper bounds could be placed on the jackknife CIs, 95% of our estimates fell between $\sim$ 1000 to 4500 (Figure 4.5). Simulation-based analysis of pseudo-replication in genomic-scale LD data sets suggests that the Jones et al. (2016) jackknife confidence interval generally underestimates precision in LDNe, and that subsetting loci provides a more realistic assessment. On the other hand, genetic indices (like r^2 for unlinked loci) that reflect very recent demography are sensitive to the pedigree structure of the individuals in the sample. Replicating across many subsets of loci, all generated by the same pedigree, will not capture uncertainty associated with differences between the pedigree structure of the sample and the pedigree structure of the population as a whole (King et al., 2018). This argues for some caution in interpreting CIs for estimates of Ne for the hamlets, all of which are based on small samples of individuals. Nonetheless, given the order of magnitude of the N_e estimates, this does not change our interpretation of the *H. maya* population as orders of magnitude smaller than its size at the beginning of speciation, including a tenfold reduction within the last few hundred generations.

Our recent N_e estimate of $\sim$ 1600 contrasts with the rarity of *H. maya* in the field and its restricted distribution. Considering the dramatic decline of *H. maya* in the Pelican and surrounding Rhomboidal Cays within the last two decades documented here, this number may nevertheless be inflated by much higher effective population sizes just a few generations ago. It is also possible that *H. maya* N_e is still affected by gene flow from pan-Caribbean hamlets, or that the population center of *H. maya* may not be in the Pelican and surrounding Rhomboidal Cays but around Laughing Bird Cay and further south, beyond the area surveyed here. Though effective population sizes as low as 500 were originally theorized as stable from a mutation-drift equilibrium perspective, the body of empirical evidence suggests that sizes of 1000-5000 are likely necessary to maintain fitness in perpetuity (Lande, 1995; Frankham et al., 2014). As such, we suggest that the past and present effective population size of *H. maya* is by itself sufficient cause for concern regarding its long-term survival.

Our data also support the disparity in effective population size between *H. maya* and its congeners. N_e estimation for *H. gemma* was not possible due to the low sample size for this species (n=5, which is below the validated range for LD-based estimation). Such a limitation is unfortunate given the recent decline in *H. gemma* census population reported here, and a renewed effort to estimate this species'

N_e is advised. In other species, though, infinite N_e estimates were obtained with a larger sample size (n = 12) than in *H. maya* (n = 10), which indicates that the pan-Caribbean species' N_e can be reliably inferred as 'much larger' than that of *H. maya*. This is compatible with a previous Approximate Bayesian Computation estimate of N_e of $\sim$ 15000 for *H. nigricans* on the BBR, an order of magnitude higher than our *H. maya* estimate (Puebla et al., 2012b).

Microendemism in the MBRS

While the case of the Maya hamlet is remarkable, it is not unique. Twelve fish species are known to be endemic to the Belize section of the MBRS and the adjacent Honduran Bay Islands, representing over 20% of those endemic to the continental Caribbean (Floeter et al. 2008; Robertson and Van Tassell 2015; Suppl. Tab. 4.1). Similar levels of microendemism are found among invertebrates (Rützler et al., 2000; Miloslavich et al., 2010). In the MBRS, the endemic fishes are distributed variably between the landward lagoon, the seaward barrier wall, and the associated atolls (Lobel et al., 2009). This high level of microendemism may be due in part to the intense sampling and exploration of the southern MBRS (Miloslavich et al., 2010). Yet analogous cases of microendemism have been documented in the less intensively sampled Indo-Pacific (Allen et al., 2018b,a), suggesting that such patterns may be more prevalent among reef-fish communities than previously recognized. Should broader sampling reveal similar concentrations of microendemics elsewhere in the Caribbean, in particular among small cryptobenthic fishes or in the mesophotic zone, the question of the underlying evolutionary processes will become even more pressing.

In accordance with the recognition of ocean currents as a limiting factor in marine dispersal (Jones et al., 2009), we suggest that local oceanography may be a primary cause of high microendemism in the MBRS. Drifters and numerical models have identified a system of temporally variable eddies that occur along the Belizean MBRS. Areas south of Glover's Reef ($\sim$ 16.75° N) experience slow, invariant transport to the south, while those found at or north of this point experience variable transport dependent on the season: transport may be rapidly southward, or weakly northwestward (Ezer et al., 2005; Tang et al., 2006). Particles (e.g. planktonic larvae) which are transported southward either encounter the interior MBRS lagoon and the Honduran Bay Islands, or are carried into a gyre within the Gulf of Honduras (Richardson, 2005; Paris et al., 2007). Of the 12 species endemic to the MBRS and southward Honduran islands, ten have northward boundaries at Carrie Bow Cay and Glover's Reef (Floeter et al., 2008; Robertson and Van Tassell, 2015). D'Aloia et al. (2015) estimated the dispersal kernel of one of these endemics (*Elacatinus lori*) at the proposed oceanographic divide, and recovered an isotropic kernel of extremely small dispersal range. Such a pattern is consistent with an oceanographic limitation to range expansion, so long as these species originated in the southern MBRS under the current oceanographic regime. Multiple independent estimates of these species kernels across their entire range, extending the work of D'Aloia et al. (2015), could shed further light on this hypothesis.

The case of the Maya hamlet is remarkable in that it is currently sympatric with congeners in terms of both distribution and microhabitat. This contrasts with other cases of microendemism in reef fishes, which show either allopatry or habitat divergence (Allen et al., 2018b,a). Though *H. maya* overlaps in habitat with sympatric congeners, it may differ in its habitat specificity. Our qualitative observations indicate that *H. maya* is strongly associated with shallow (1–3 m) reef habitat. The Maya hamlet was nearly extirpated from the Pelican Cays as of 2017, coinciding with the degradation of shallow coral communities on the Cays' characteristics polygonal ridges (O. Puebla and B. Moran, personal observation). In contrast, *H. maya* was the dominant hamlet species on the shallow reefs west of Laughing Bird Cay, which harbored high coverage of *A. cervicornis* (O. Puebla and B. Moran, personal observation). Specialist adaptation to shallow *A. cervicornis* reefs would provide another explanation for the long-term N_e decline of *H. maya* inferred by our MSMC analyses, given the geological history of their range. The cays of the southern MBRS lagoon began as Pleistocene limestone surfaces, which were submerged by sea-level rise after the last glacial maximum (Macintyre et al., 2000). *Acropora cervicornis* colonized this substrate, growing towards the surface at a rate of up to 8 m/1000 years (Westphall, 1986; Macintyre et al., 2000; Aronson et al., 2002). Where reef accretion outpaced sea level rise, the reef crest was colonized by the shallow-specialist coral *Porites divaricata*, and later red mangrove (*Rhizophora mangle*) trees (Neumann, 1985; Macintyre et al., 2000). The MBRS lagoon thus represents a non-equilibrium habitat in relative isolation, presenting the exceptional opportunity for reduced gene flow with outside populations, unfilled niches, and founder effects. If *H. maya* is indeed an *A. cervicornis* specialist that appeared in the mid-Holocene as inferred by our MSMC analyses, ecological succession after the last glacial maximum would have created a long-term natural decline in habitat availability throughout its existence. This, combined with a relatively short PLD of 14-22 days (Domeier, 1994) and the aforementioned oceanographic characteristics of the southern MBRS, may explain this case of micro-endemism and a long-term decline of *H. maya*. The generality of such forces could be tested in other cases of microendemism, both in geographically distinct cases within *Hypoplectrus* (Victor and Marks, 2018) and in phylogenetically distinct cases within the MBRS.

Microendemism and Extinction

Species with small ranges are particularly vulnerable to extinction, due to a combination of low total population size and increased threat presented by local extirpations (Gaston, 1998). This risk is further elevated in the case of ecological specialists, which exhibit a synergistic combination of lower population densities and lower tolerance to change (Munday, 2004). While *H. maya* population declines predated human influence, the reduction in habitat available to *H. maya* was likely accelerated in the last century by the drastic decline in Caribbean corals, and acroporids in particular. This trend of reef degradation is largely attributable to coral disease outbreaks (Aronson and Precht, 2001), coastal development (Murray, 2007), decline of herbivorous fishes and invertebrates (Hughes, 1994), and ocean

warming (Aronson et al., 2000). The MBRS lagoon, in particular, is currently threatened by clear-cutting of mangroves and dredging of shallow patch reefs to increase land values for real estate and touristic development (McKee and Vervaeke, 2009). Furthermore, the invasive lionfish constitutes a direct threat to the Maya hamlet and other Caribbean microendemic fishes (Rocha et al., 2015). Such a combination of stressors provides a plausible explanation for the recent reduction of the *H. maya* population evidenced here by both genetic data and field surveys. Likewise, the recent decline in *H. gemma* evidenced by transect surveys (Suppl. Fig. 4.3) coincides with the loss of Florida Keys reef communities to disease and warming (Precht et al., 2016). Collection by the aquarium trade may also play a role in the case of *H. gemma*, given the popularity of this species among public and private aquarists (O. Puebla and B. Moran, personal observation). Given the exceptionally small range of *H. maya*, its rarity, its long-term and recent decline in population size, its strong association with *A. cervicornis* and the ongoing degradation of its habitat, the persistence of this recently-diverged species is in jeopardy. The case of the Maya hamlet shows that the evolution of marine microendemism can be a fast and dynamic process, with extinction possibly occurring before speciation is complete.

Acknowledgements

We thank Ryan Waples for feedback during the implementation of whole-genome LDNe, Stephan Schiffels for his guidance regarding MSMC, Derya Akkaynak for assistance in the field, as well as Katie Lotterhos and Phil Lobel for their input regarding population genomics and Belizean endemics. We are indebted to the staff of Carrie Bow Cay Field Station and Keys Marine Laboratory for their assistance during sampling, to Allison and Carlos Estapé, and to Lisa Carne of the Belizean coral restoration non-profit Fragments of Hope. This work was conducted under National Geographic Society Young Explorers Grant #WW-037ER-17, a Summer Independent Research Fellowship from Northeastern University, and two scholarships (Undergraduate and Short-Term Research) from the German Academic Exchange Service to BM, a German Research Foundation and a Future Ocean Cluster of Excellence grant to OP, and a Global Genome Initiative and Smithsonian Institute for Biodiversity Genomics grant to CB, WOM, and OP.

Data Accessibility

All new raw sequences (*H. maya* and *H. gemma*) have been deposited at the European Nucleotide Archive (ENA) under project accession number PRJEB29705; the individual accession numbers for these samples are provided in Suppl. Tab. 4.2. Previously sequenced Belizean samples (*H. nigricans*, *H. puella*, and *H. unicolor*) are available under ENA project PRJEB27858. The biallelic SNP genotypes in VCF format for all samples are also available in Dryad (doi:10.5061/dryad.hp388dm).

Author Contributions

BM conceived of the study, conducted field work and data analyses, and wrote the

manuscript. OP conceived of the study, conducted field work, and contributed to the manuscript. RWS contributed to data analyses and the manuscript. KH contributed to the data analyses. WOM and MH contributed to genome sequencing. CB contributed to the curation of specimens. All co-authors provided feedback on the manuscript.

Supplementary Tables

Suppl. Table 4.1: List of species considered endemic to the Belize Barrier Reef (BBR) System and Honduran Bay Islands (HBI). All data retrieved from IUCN Red List website on 14 November 2018. CBC = Carrie Bow Cay, AOO = estimated Area of Occupancy (km^2), DD = Data Deficient, LC = Least Concern, VU = Vulnerable, EN = Endangered.

Species	Family	Known Range	AOO	Status
Cathorops belizensis	Ariidae	Mangroves near Belize City	2.5	DD
Sanopus astrifer	Batrachoididae	Turneffe Atoll and Glover's Reef	700	VU
Sanopus greenfieldorum	Batrachoididae	CBC to South Water Cay	705	VU
Vladicthys gloverensis	Batrachoididae	Turneffe Atoll to HBI	813	VU
Hypoplectrus maya	Serranidae	South Water Cay to Sapodilla Cays	1483	VU
Halichoeres socialis	Labridae	BBR Lagoon, esp. Pelican Cays	24	EN
Emblemariopsis dianae	Chaenopsidae	Pelican Cays	?	DD
Emblemariopsis pricei	Chaenopsidae	BBR to HBI	?	VU
Tomicodon clarkei	Gobiesocidae	Carrie Bow Cay	?	DD
Tomicodon lavettsmithi	Gobiesocidae	Pelican Cays and Twin Cays	2234	DD
Elacatinus lori	Gobiidae	CBC to HBI	3341	LC
Psilotris amblyrhynchus	Gobiidae	CBC to Puerto Cortes, HN	2037	DD

Suppl. Table 4.2: Metadata and accession numbers of the new samples considered in this study. Mean genome-wide coverage assessed from BAM files prior to GATK genotyping. Accession numbers presented for raw reads in European Nucleotide Archive (ENA), for tissue samples in the Smithsonian National Museum of Natural History Biorepository (Tissue), and for voucher specimens deposited in the NMNH vertebrate collection (Voucher).

ID	Species	Site	Date	Lat.	Lon.	Cov.	ENA	Tissue	Voucher
PL17_89	*H. maya*	Belize	2017-04-07	16.660	-88.185	19.845	ERS2899590	AH0SC11	446507
PL17_95	*H. maya*	Belize	2017-04-08	16.771	-88.164	24.183	ERS2899591	AH0SC16	446512
PL17_101	*H. maya*	Belize	2017-04-06	16.771	-88.164	17.290	ERS2899592	AH0SC09	446518
PL17_119	*H. maya*	Belize	2017-04-13	16.487	-88.272	20.625	ERS2899593	AH0SC38	446535
PL17_120	*H. maya*	Belize	2017-04-14	16.487	-88.272	19.317	ERS2899594	AH0SC39	446536
PL17_121	*H. maya*	Belize	2017-04-15	16.487	-88.272	18.465	ERS2899595	AH0SC40	446537
PL17_122	*H. maya*	Belize	2017-04-16	16.487	-88.272	26.143	ERS2899596	AH0SC41	446538
PL17_123	*H. maya*	Belize	2017-04-17	16.510	-88.255	20.546	ERS2899597	AH0SC42	446539
PL17_124	*H. maya*	Belize	2017-04-18	16.510	-88.255	21.769	ERS2899598	AH0SC43	446540
PL17_126	*H. maya*	Belize	2017-04-19	16.510	-88.255	21.336	ERS2899599	AH0SC45	446542
PL17_142	*H. gemma*	Florida	2017-07-08	24.806	-80.677	22.668	ERS2899137	AH0SC97	446558
PL17_144	*H. gemma*	Florida	2017-07-09	24.812	-80.670	26.630	ERS2899138	AH0SC99	446560
PL17_145	*H. gemma*	Florida	2017-07-09	24.812	-80.670	21.917	ERS2899139	AH0SD00	446561
PL17_148	*H. gemma*	Florida	2017-07-11	24.769	-80.728	21.157	ERS2899140	AH0SD03	446564
PL17_153	*H. gemma*	Florida	2017-07-13	24.807	-80.677	24.3975	ERS2899141	AH0SD07	446568

Suppl. Table 4.3: F_{ST} outlier regions (above 99.99th F_{ST} percentile) across all pairwise species comparisons that include *H. maya*. Start and End correspond to the positions along linkage groups (chromsomomes) in bp, Comparisons to the number of pairwise comparisons in which this interval is identified as an outlier (above 99.99th F_{ST} percentile), and Genes to the number of annotated genes found within these regions (listed in the rightmost column).

LG	Nr	Start	End	Comparisons (n)	Genes (n)	Genes (names)
LG04	1	8780001	8845000	1	7	*hbegf*, *rufy1*, *hnrnph1*, *ankhd1*, *higd2a*, *clta*, *pdlim7*
LG08	1	12980001	13040000	1	2	*rorb*, *sema6b*
LG08	2	13545001	13660000	1	9	*ssbp3*, *fkbp8*, *ell2*, *dot1l*, *amh*, *hpv1g...11406*, *oaz1a*, *prag1*, *hpv1g...11409*
LG09	1	17835001	17915000	1	5	*hpv1g...12847*, *kcnj12*, *sox10*, *rnaseh2a*, *mast1*
LG12	1	15045001	15130000	2	5	*ren*, *csf1*, *hpv1g...15947*, *eif2d*, *slc12a5*
LG12	2	15140001	15215000	2	12	*slc12a5*, *rhbg-a*, *rbm39*, *epabp-a*, *143b1*, *ogn*, *smim4*, *kctd6*, *hpv1g...15957*, *abhd6*, *slmap*, *cers5*
LG12	3	20190001	20255000	2	1	*casz1*

Suppl. Table 4.4: Sample ID and coverage for all effective population size (MSMC) and cross-coalescence (CC) analyses.

Species	MSMC Run	CC Run	ID	Coverage
H. maya	1	5	PL17_119	18.2060
	1	3	PL17_120	17.0390
	1	1	PL17_122	23.1169
	1	5	PL17_123	18.0826
	2	2	PL17_121	16.0652
	2	3	PL17_124	19.2426
	2	4	PL17_126	18.8595
	3	4	PL17_89	17.6444
	3	2	PL17_95	21.5546
	3	1	PL17_101	15.2245
H. gemma	1	2	PL17_142	21.3610
	1	1	PL17_144	25.3755
	1	1	PL17_145	20.5592
	1	2	PL17_153	22.8527
H. nigricans	1	5	18159	20.0515
	1	1	18162	16.5132
	1	6	18165	20.955
	1	4	18187	22.5938
	2	4	18151	19.7292
	2	5	18157	21.0023
	2	3	18158	19.5513
	2	3	18185	23.2919
	3	6	18153	20.9889
	3	2	18155	19.3689
	3	2	18156	23.5517
	3	1	18171	23.8413
H. puella	1	1	18161	16.5969
	1	4	18166	21.6082
	1	2	18174	22.5343
	1	5	18179	20.1783
	2	5	18152	21.1712
	2	1	18176	25.2119
	2	4	18178	19.7819
	2	3	18180	19.5115
	3	3	18154	22.2604
	3	2	18169	17.7687
	3	6	18172	20.8912
	3	6	18175	20.2601
H. unicolor	1	6	19881	20.7515
	1	1	20120	27.3212
	1	3	20128	19.9119
	1	1	20135	19.0213
	2	4	18163	21.4329
	2	2	18267	25.1234
	2	6	18276	21.1962
	2	4	20126	20.0994
	3	2	18261	19.7475
	3	3	18274	22.0874
	3	5	20092	21.4006
	3	5	20149	20.5865

Suppl. Table 4.5: Software versions used in this study.

Software	Version
bcftools	1.9
bedtools	2.26.0-148-gd1953b6
BWA	0.7.12-r1039
extractPIRs	1.r68.x86_64
FastQC	0.11.7
GATK	3.8.0-ge9d806836
htslib	1.9
MSMC	2.0.0
msmc-tools	12758d9283f47fee173eab840c3ce364c6eb3495
NeEstimator	2.1
NextFlow	0.31.1
PGDSpider	2.1.1.3
Picard Tools	2.11.0-SNAPSHOT
PLINK	v1.90b4 64-bit
R	3.5.1
samtools	1.9
SHAPEIT	2.r837
SMC++	1.14.0
vcftools	0.1.15

Chapter 4 References

Aguila-Perera, A. and Tuz-Sulub, A. N. (2010). Scientific note Hypoplectrus gemma (Teleostei, Serranidae) is not endemic to southern Florida waters. *Pan-American Journal of Aquatic Sciences*, 5(1):143–146.

Allen, G., Erdmann, M., and Cahyani, N. D. (2018a). Chrysiptera uswanasi, a new microendemic species of damselfish (Teleostei: Pomacentridae) from West Papua Province, Indonesia. *Journal of the Ocean Science Foundation*, 31:74–86.

Allen, G., Erdmann, M., and Hidayat, N. (2018b). Pomacentrus bellipictus, a new microendemic species of damselfish (Pisces: Pomacentridae) from the Fakfak Peninsula, West Papua, Indonesia. *Journal of the Ocean Science Foundation*, 30:1–10.

Anderson, M. J. (2001). A new method for non-parametric multivariate analysis of variance. *Austral Ecology*, 26(1):32–46.

Aronson, R. B., Macintyre, I. G., Precht, W. F., Murdoch, T. J., and Wapnick, C. M. (2002). The expanding scale of species turnover events on coral reefs in Belize. *Ecological Monographs*, 72(2):233–249.

Aronson, R. B. and Precht, W. F. (2001). White-band disease and the changing face of Caribbean coral reefs. *Hydrobiologia*, 460(1):25–38.

Aronson, R. B., Precht, W. F., Macintyre, I. G., and Murdoch, T. J. (2000). Ecosystems: coral bleach-out in Belize. *Nature*, 405(6782):36.

Barreto, F. S. and McCartney, M. A. (2008). Extraordinary AFLP fingerprint similarity despite strong assortative mating between reef fish color morphospecies. *Evolution: International Journal of Organic Evolution*, 62(1):226–233.

D'Aloia, C. C., Bogdanowicz, S. M., Francis, R. K., Majoris, J. E., Harrison, R. G., and Buston, P. M. (2015). Patterns, causes, and consequences of marine larval dispersal. *Proceedings of the National Academy of Sciences*, 112(45):13940–13945.

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., Handsaker, R. E., Lunter, G., Marth, G. T., Sherry, S. T., McVean, G., Durbin, R., and 1000 Genomes Project Analysis Group (2011). The variant call format and vcftools. *Bioinformatics*, 27(15):2156–2158.

Delaneau, O., Howie, B., Cox, A. J., Zagury, J.-F., and Marchini, J. (2013). Haplotype estimation using sequencing reads. *The American Journal of Human Genetics*, 93(4):687–696.

Do, C., Waples, R. S., Peel, D., Macbeth, G., Tillett, B. J., and Ovenden, J. R. (2014). Neestimator v2: re-implementation of software for the estimation of contemporary effective population size (Ne) from genetic data. *Molecular Ecology Resources*, 14(1):209–214.

Domeier, M. L. (1994). Speciation in the serranid fish Hypoplectrus. *Bulletin of Marine Science*, 54(1):103–141.

Ezer, T., Thattai, D. V., Kjerfve, B., and Heyman, W. D. (2005). On the variability of the flow along the Meso-American Barrier Reef system: a numerical model study of the influence of the Caribbean current and eddies. *Ocean Dynamics*, 55(5-6):458–475.

Fischer, E. A. (1980). The relationship between mating system and simultaneous hermaphroditism in the coral reef fish, Hypoplectrus nigricans (Serranidae). *Animal Behaviour*, 28(2):620–633.

Floeter, S. R., Rocha, L. A., Robertson, D. R., Joyeux, J., Smith-Vaniz, W. F., Wirtz, P., Edwards, A., Barreiros, J. P., Ferreira, C., Gasparini, J. L., Brito, A., Falcón, J. M., Bowen, B. M., and Bernardi, G. (2008). Atlantic reef fish biogeography and evolution. *Journal of Biogeography*, 35(1):22–47.

Frankham, R. (1998). Inbreeding and extinction: island populations. *Conservation Biology*, 12(3):665–675.

Frankham, R., Bradshaw, C. J., and Brook, B. W. (2014). Genetics in conservation management: revised recommendations for the 50/500 rules, Red List criteria and population viability analyses. *Biological Conservation*, 170:56–63.

Gaston, K. J. (1998). Ecology: rarity as double jeopardy. *Nature*, 394(6690):229.

Heemstra, P. C., Anderson Jr., W. D., and Lobel, P. S. (2002). Serranidae: Groupers (seabasses, creolefish, coney, hamlets, anthiines, and soapfishes). In Carpenter, K. E., editor, *The living marine resources of the Western Central Atlantic, Volume 2: Bony fishes, part 1 (Acipenseridae to Grammatidae)*, pages 1308–1369. FAO Species Identification Guide for Fishery Purposes and American Society of Ichthyologists and Herpetologists Special Publication No. 5, FAO, Rome, Italy, 601–1374.

Hench, K., Vargas, M., Höppner, M. P., McMillan, W. O., and Puebla, O. (2019). Inter-chromosomal coupling between vision and pigmentation genes during genomic divergence. *Nature Ecology & Evolution*.

Hill, W. G. (1981). Estimation of effective population size from data on linkage disequilibrium. *Genetics Research*, 38(3):209–216.

Holt, B., Côté, I., and Emerson, B. (2010). Signatures of speciation? distribution and diversity of Hypoplectrus (Teleostei: Serranidae) colour morphotypes. *Global Ecology and Biogeography*, 19(4):432–441.

Holt, B. G., Emerson, B. C., Newton, J., Gage, M. J., and Côté, I. M. (2008). Stable isotope analysis of the Hypoplectrus species complex reveals no evidence for dietary niche divergence. *Marine Ecology Progress Series*, 357:283–289.

Hughes, T. P. (1994). Catastrophes, phase shifts, and large-scale degradation of a Caribbean coral reef. *Science*, 265(5178):1547–1551.

International SNP Map Working Group (2001). A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature*, 409(6822):928.

Jones, A., Ovenden, J., and Wang, Y. (2016). Improved confidence intervals for the linkage disequilibrium method for estimating effective population size. *Heredity*, 117(4):217.

Jones, G., Almany, G., Russ, G., Sale, P., Steneck, R., Van Oppen, M., and Willis, B. (2009). Larval retention and connectivity among populations of corals and reef fishes: history, advances and challenges. *Coral Reefs*, 28(2):307–325.

Kay, E. A. and Palumbi, S. R. (1987). Endemism and evolution in Hawaiian marine invertebrates. *Trends in Ecology & Evolution*, 2(7):183–186.

King, L., Wakeley, J., and Carmi, S. (2018). A non-zero variance of tajima's estimator for two sequences even for infinitely many unlinked loci. *Theoretical Population Biology*, 122:22–29.

Lande, R. (1995). Mutation and conservation. *Conservation Biology*, 9(4):782–791.

Leis, J. M. (2006). Are larvae of demersal fishes plankton or nekton? *Advances in Marine Biology*, 51:57–141.

Lester, S. E. and Ruttenberg, B. I. (2005). The relationship between pelagic larval duration and range size in tropical reef fishes: a synthetic analysis. *Proceedings of the Royal Society of London B: Biological Sciences*, 272(1563):585–591.

Li, H. and Durbin, R. (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics*, 25(14):1754.

Li, H. and Durbin, R. (2011). Inference of human population history from individual whole-genome sequences. *Nature*, 475(7357):493.

Lindsay, W. R., Webster, M. S., and Schwabl, H. (2011). Sexually selected male plumage color is testosterone dependent in a tropical passerine bird, the red-backed fairy-wren (malurus melanocephalus). *PLoS One*, 6(10):e26067.

Liu, S., Hansen, M. M., and Jacobsen, M. W. (2016). Region-wide and ecotype-specific differences in demographic histories of threespine stickleback populations, estimated from whole genome sequences. *Molecular Ecology*, 25(20):5187–5202.

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

Lobel, P. S., Rocha, L. A., and Randall, J. E. (2009). Color phases and distribution of the western Atlantic labrid fish, Halichoeres socialis. *Copeia*, (1):171–174.

Luiz, O. J., Allen, A. P., Robertson, D. R., Floeter, S. R., Kulbicki, M., Vigliola, L., Becheler, R., and Madin, J. S. (2013). Adult and larval traits as determinants of geographic range size among tropical reef fishes. *Proceedings of the National Academy of Sciences*, page 201304074.

Macintyre, I. G., Precht, W. F., and Aronson, R. B. (2000). Origin of the Pelican Cays ponds, Belize. *Atoll Research Bulletin*, (466).

Mallet, J. (2005). Hybridization as an invasion of the genome. *Trends in Ecology & Evolution*, 20(5):229–237.

Martin, S. H. (2016). genomics_general: General tools for genomic analyses; a github repository.

McCartney, M. A., Acevedo, J., Heredia, C., Rico, C., Quenoville, B., Bermingham, E., and McMillan, W. O. (2003). Genetic mosaic in a marine species flock. *Molecular Ecology*, 12(11):2963–2973.

McKee, K. L. and Vervaeke, W. C. (2009). Impacts of human disturbance on soil erosion potential and habitat stability of mangrove-dominated islands in the Pelican Cays and Twin Cays Ranges, Belize. *Smithsonian Contributions to the Marine Sciences*, (38).

Meyer, C. P., Geller, J. B., and Paulay, G. (2005). Fine scale endemism on coral reefs: archipelagic differentiation in turbinid gastropods. *Evolution*, 59(1):113–125.

Miloslavich, P., Díaz, J. M., Klein, E., Alvarado, J. J., Díaz, C., Gobin, J., Escobar-Briones, E., Cruz-Motta, J. J., Weil, E., Cortes, J., Bastidas, A. C., Robertson, R., Zapata, F., Martín, A., Castillo, J., Kazandjian, A., and Ortiz, M. (2010). Marine biodiversity in the Caribbean: regional estimates and distribution patterns. *PloS One*, 5(8):e11916.

Mora, C., Treml, E. A., Roberts, J., Crosby, K., Roy, D., and Tittensor, D. P. (2012). High connectivity among habitats precludes the relationship between dispersal and range size in tropical reef fishes. *Ecography*, 35(1):89–96.

Munday, P. L. (2004). Habitat loss, resource specialization, and extinction on coral reefs. *Global Change Biology*, 10(10):1642–1647.

Murray, G. (2007). Constructing paradise: the impacts of big tourism in the Mexican coastal zone. *Coastal Management*, 35(2-3):339–355.

Nadachowska-Brzyska, K., Burri, R., Smeds, L., and Ellegren, H. (2016). PSMC analysis of effective population sizes in molecular ecology and its application to black-and-white Ficedula flycatchers. *Molecular Ecology*, 25(5):1058–1072.

Nei, M. (1987). *Molecular evolutionary genetics*. Columbia University Press.

Nei, M., Maruyama, T., and Chakraborty, R. (1975). The bottleneck effect and genetic variability in populations. *Evolution*, 29(1):1–10.

Neumann, A. (1985). Reef response to sea-level rise: keep-up, catch-up, or give-up. In *Proceedings of the Fifth International Coral Reef Congress, Tahiti*, volume 3, pages 105–110.

Oksanen, J., Blanchet, F. G., Friendly, M., Kindt, R., Legendre, P., McGlinn, D., Minchin, P. R., O'Hara, R. B., Simpson, G. L., Solymos, P., Stevens, M. H. H., Szoecs, E., and Wagner, H. (2018). *vegan: Community Ecology Package*. R package version 2.5-2.

Palumbi, S. R. (1992). Marine speciation on a small planet. *Trends in Ecology & Evolution*, 7(4):114–118.

Paris, C. B., Chérubin, L. M., and Cowen, R. K. (2007). Surfing, spinning, or diving from reef to reef: effects on population connectivity. *Marine Ecology Progress Series*, 347:285–300.

Paulay, G. and Meyer, C. (2002). Diversification in the tropical Pacific: comparisons between marine and terrestrial systems and the importance of founder speciation. *Integrative and Comparative Biology*, 42(5):922–934.

Picq, S., McMillan, W. O., and Puebla, O. (2016). Population genomics of local adaptation versus speciation in coral reef fishes (Hypoplectrus spp, Serranidae). *Ecology and Evolution*, 6(7):2109–2124.

Precht, W. F., Gintert, B. E., Robbart, M. L., Fura, R., and Van Woesik, R. (2016). Unprecedented disease-related coral mortality in Southeastern Florida. *Scientific Reports*, 6:31374.

Primmer, C., Borge, T., Lindell, J., and Sætre, G.-P. (2002). Single-nucleotide polymorphism characterization in species with limited available sequence information: high nucleotide diversity revealed in the avian genome. *Molecular Ecology*, 11(3):603–612.

Puebla, O., Bermingham, E., and 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(6):1405–1415.

Puebla, O., Bermingham, E., and Guichard, F. (2009). Estimating dispersal from genetic isolation by distance in a coral reef fish (Hypoplectrus puella). *Ecology*, 90(11):3087–3098.

Puebla, O., Bermingham, E., and Guichard, F. (2012a). Pairing dynamics and the origin of species. *Proceedings of the Royal Society of London B: Biological Sciences*, 279(1731):1085–1092.

Puebla, O., Bermingham, E., Guichard, F., and Whiteman, E. (2007). Colour pattern as a single trait driving speciation in Hypoplectrus coral reef fishes? *Proceedings of the Royal Society of London B: Biological Sciences*, 274(1615):1265–1271.

Puebla, O., Bermingham, E., and McMillan, W. O. (2012b). On the spatial scale of dis-

persal in coral reef fishes. *Molecular Ecology*, 21(23):5675–5688.

Puebla, O., Bermingham, E., and McMillan, W. O. (2014). Genomic atolls of differentiation in coral reef fishes (Hypoplectrus spp., Serranidae). *Molecular Ecology*, 23(21):5291–5303.

Puebla, O., Picq, S., Lesser, J. S., and Moran, B. (2018). Social-trap or mimicry? an empirical evaluation of the Hypoplectrus unicolor–Chaetodon capistratus association in Bocas del Toro, Panama. *Coral Reefs*, 37(4):1127–1137.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., Maller, J., Sklar, P., De Bakker, P. I., Daly, M. J., and Sham, P. (2007). Plink: a tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics*, 81(3):559–575.

Ramon, M. L., Lobel, P. S., and Sorenson, M. D. (2003). Lack of mitochondrial genetic structure in hamlets (Hypoplectrus spp.): recent speciation or ongoing hybridization? *Molecular Ecology*, 12(11):2975–2980.

Randall, J. E. (1998). Zoogeography of shore fishes of the Indo-Pacific region. *Zoological Studies*, 37(4):227–268.

Randall, J. E. and Randall, H. A. (1960). Examples of mimicry and protective resemblance in tropical marine fishes. *Bulletin of Marine Science*, 10(4):444–480.

Richardson, P. L. (2005). Caribbean current and eddies as observed by surface drifters. *Deep Sea Research Part II: Topical Studies in Oceanography*, 52(3-4):429–463.

Roberts, C. M., McClean, C. J., Veron, J. E., Hawkins, J. P., Allen, G. R., McAllister, D. E., Mittermeier, C. G., Schueler, F. W., Spalding, M., Wells, F. a., Vynne, C., and Werner, T. B. (2002). Marine biodiversity hotspots and conservation priorities for tropical reefs. *Science*, 295(5558):1280–1284.

Robertson, D. R. and Van Tassell, J. (2015). Shorefishes of the Greater Caribbean: online information system.

Rocha, L. and Bowen, B. (2008). Speciation in coral-reef fishes. *Journal of Fish Biology*, 72(5):1101–1121.

Rocha, L. A., Rocha, C. R., Baldwin, C. C., Weigt, L. A., and McField, M. (2015). Invasive lionfish preying on critically endangered reef fish. *Coral Reefs*, 34(3):803–806.

Ruttenberg, B. I. and Lester, S. E. (2015). Patterns and processes in geographic range size in coral reef fishes. In Mora, C., editor, *Ecology of Fishes on Coral Reefs*, page 97. Cambridge University Press.

Rützler, K., Díaz, M. C., van Soest, R. W. M., Zea, S., Smith, K. P., Alvarez, B., and Wulff, J. (2000). Diversity of sponge fauna in mangrove ponds, Pelican Cays, Belize. *Atoll Research Bulletin*, (476).

Schiffels, S. and Durbin, R. (2014). Inferring human population size and separation history from multiple genome sequences. *Nature Genetics*, 46(8):919.

Schrider, D. R., Shanku, A. G., and Kern, A. D. (2016). Effects of linked selective sweeps on demographic inference and model selection. *Genetics*, pages genetics–116.

Shao, Y. T., Wang, F.-Y., Fu, W.-C., Yan, H. Y., Anraku, K., Chen, I.-S., and Borg, B. (2014). Androgens increase lws opsin expression and red sensitivity in male three-spined sticklebacks. *PLoS One*, 9(6):e100330.

Simpson, S. D., Harrison, H. B., Claereboudt, M. R., and Planes, S. (2014). Long-distance dispersal via ocean currents connects Omani clownfish populations throughout entire species range. *PLoS One*, 9(9):e107610.

Smith, C. L., Tyler, J. C., Davis, W. P., Jones, R. S., Smith, D. G., and Baldwin, C. C. (2003). Fishes of the Pelican Cays, Belize. *Atoll Research Bulletin*, 497.

Smith, S. G., Ault, J. S., Bohnsack, J. A., Harper, D. E., Luo, J., and McClellan, D. B. (2011). Multispecies survey design for assessing reef-fish stocks, spatially explicit management performance, and ecosystem condition. *Fisheries Research*, 109(1):25–41.

Szpiech, Z. A. and Hernandez, R. D. (2014). selscan: an efficient multithreaded program to perform EHH-based scans for positive selection. *Molecular biology and evolution*, 31(10):2824–2827.

Tang, L., Sheng, J., Hatcher, B. G., and Sale, P. F. (2006). Numerical study of circulation, dispersion, and hydrodynamic connectivity of surface waters on the Belize shelf. *Journal of Geophysical Research: Oceans*, 111(C1).

Tavera, J. and Acero, A. P. (2013). Description of a new species of Hypoplectrus (Perciformes: Serranidae) from the Southern Gulf of Mexico. *Aqua: International Journal of Ichthyology*, 19(1):29–39.

Terhorst, J., Kamm, J. A., and Song, Y. S. (2017). Robust and scalable inference of population history from hundreds of unphased whole genomes. *Nature Genetics*, 49(2):303.

Thresher, R. E. (1978). Polymorphism, mimicry, and the evolution of the hamlets (Hypoplectrus, Serranidae). *Bulletin of Marine Science*, 28(2):345–353.

Torres, R., Szpiech, Z. A., and Hernandez, R. D. (2018). Human demographic history has amplified the effects of background selection across the genome. *PLoS Genetics*, 14(6):1–27.

UNEP-WCMC, WorldFish Centre, WRI, and TNC (2010). *Global distribution of coral reefs, compiled from multiple sources including the Millennium Coral Reef Mapping Project. Version 2.0, updated by UNEP-WCMC. Includes contributions from IMaRS-USF and IRD (2005), IMaRS-USF (2005) and Spalding et al. (2001). Cambridge (UK): UNEP World Conservation Monitoring Centre*. URL: http://data.unepwcmc.org/datasets/1.

Van der Auwera, G. A., Carneiro, M. O., Hartl, C., Poplin, R., del Angel, G., Levy-Moonshine, A., Jordan, T., Shakir, K., Roazen, D., Thibault, J., Banks, E., Garimella, K. V., Altshuler, D., Gabriel, S., and DePristo, M. A. (2013). From fastQ data to high-confidence variant calls: The genome analysis toolkit best practices pipeline. *Current Protocols in Bioinformatics*, (SUPL.43).

Victor, B. C. (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.

Victor, B. C. and Marks, K. (2018). Hypoplectrus liberte, a new and endangered microendemic hamlet from Haiti (Teleostei: Serranidae). *Journal of the Ocean Science Foundation*, 31:8–17.

Waples, R. K., Larson, W. A., and Waples, R. S. (2016). Estimating contemporary effective population size in non-model species using linkage disequilibrium across thousands of loci. *Heredity*, 117(4):233–240.

Waples, R. S. (2006). A bias correction for estimates of effective population size based on linkage disequilibrium at unlinked gene loci. *Conservation Genetics*, 7(2):167–184.

Weir, B. S. and Cockerham, C. C. (1984). Estimating f-statistics for the analysis of population structure. *Evolution*, 38(6):1358–1370.

Westphall, M. (1986). Anatomy and history of a ringed-reef complex, Belize, Central America. Master's thesis, University of Miami, Coral Gables, Florida.

Whiteman, E. and Gage, M. (2007). No barriers to fertilization between sympatric colour morphs in the marine species flock Hypoplectrus (Serranidae). *Journal of Zoology*, 272(3):305–310.

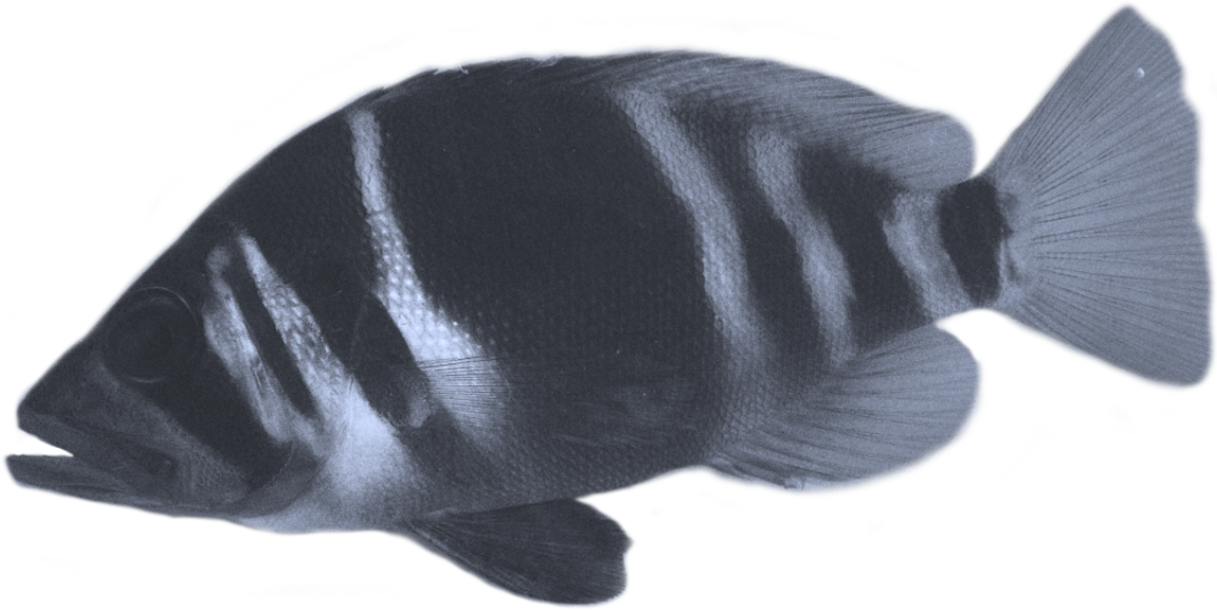
Wright, S. (1922). Coefficients of inbreeding and relationship. *The American Naturalist*, 56(645):330–338.

Yang, J., Benyamin, B., McEvoy, B. P., Gordon, S., Henders, A. K., Nyholt, D. R., Madden, P. A., Heath, A. C., Martin, N. G., Montgomery, G. W., Goddard, M. E., and Visscher, P. M. (2010). Common SNPs explain a large proportion of the heritability for human height. *Nature Genetics*, 42(7):565–569.

Zheng, X., Levine, D., Shen, J., Gogarten, S. M., Laurie, C., and Weir, B. S. (2012). A high-performance computing toolset for relatedness and principal component analysis of SNP data. *Bioinformatics*, 28(24):3326–3328.

– 5 –

The Genomic Origins of a Marine Radiation



Kosmas Hench [1], W. Owen McMillan [2], Oscar Puebla [1,2,3,4]

1 Leibniz Centre for Tropical Marine Research (ZMT), Ecology Department, Fahrenheitstraße 6, 28359 Bremen, Germany
2 Smithsonian Tropical Research Institute, Apartado Postal 0843-03092, Panamá, República de Panamá
3 Institute for Chemistry and Biology of the Marine Environment (ICBM), Carl-von-Ossietzky-Straße 9-11, 26111 Oldenburg, Germany
4 GEOMAR Helmholtz Centre for Ocean Research Kiel, Evolutionary Ecology of Marine Fishes, Düsternbrooker Weg 20, 24105 Kiel, Germany

The included version of the manuscript represents the initial submission of this study to the *Current Biology* prior to any changes that were included during review process.

Original manuscript

Hench, K. *et al*. (in revision). The genomic origins of a marine radiation. *Current Biology*

Abstract

Adaptive radiation, the evolutionary process whereby a lineage diversifies over a short period of time, is an important source of biological diversity (Schluter, 2000). While substantial progress has been made in our understanding of the ecological contexts that provide opportunity for radiation, how this potential is realized from a genetic perspective remains largely unknown (Berner and Salzburger, 2015). The hamlets, a group of reef fishes from the wider Caribbean that radiated into a stunning diversity of color patterns, provide a compelling backdrop to investigate how genomes diverge during the earliest stages of adaptive radiation. Cross-coalescence analyses based on a dataset of 170 genomes representing 28 species pairs suggest that the radiation is very recent. At the lowest levels of genome-wide differentiation, genetic differentiation (F_{ST}) is restricted to a few narrow genomic intervals. Contrary to a central prediction of the genic view of speciation (Wu, 2001; Wu and Ting, 2004), these intervals do not expand as species differentiate. Background levels of differentiation rise instead. Genetic divergence (d_{XY}) remains largely dominated by ancestral variation, except again at a few narrow genomic regions. These genomic intervals are associated with specific components of color pattern variation (bars, marks, color) that form the basis of phenotypic variation in the group. Together our data reveal a modular genetic basis of radiation whereby phenotypic diversity is generated by different combinations of ancestral alleles at these loci. We suggest that such a modular genetic basis of diversification may underlie a variety of adaptive radiations on land, in freshwater and in the sea.

Keywords: adaptive radiation, genomic bases, genetic differentiation, genetic divergence, genomic architecture, *Hypoplectrus*.

5.1. Results and Discussion

The Genomic Architecture of Differentiation and Divergence

Caribbean hamlets encompass extremely low levels of genetic differentiation, with genome-wide F_{ST} between pairs of sympatric species ranging from < 0.003 to 0.1 (Figure 5.1, Suppl. Tab. 5.1). This low and continuous range of differentiation provides a rare window into the earliest genomic stages of adaptive radiation. It notably allows us to describe how the genomic architecture of species differences evolves as genome-wide differentiation develops, and to do so at much shallower levels of differentiation than previous studies (in e.g. sunflowers (Renaut *et al.*, 2013), *Heliconius* butterflies (Kronforst *et al.*, 2013), *Ficedula* flycatchers (Burri *et al.*, 2015), Darwin's finches (Han *et al.*, 2017) or monkeyflowers (Stankowski *et al.*, 2019), Suppl. Fig. 5.1). In line with previous analyses of three hamlet species (Hench *et al.*, 2019), sharp peaks (so-called 'islands' (Turner *et al.*, 2005)) of differentiation are observed at the lowest levels of genome-wide differentiation (Suppl. Fig. 5.2). Yet contrary to a central prediction of the genic view of speciation (Wu, 2001; Wu and Ting, 2004), these peaks do not expand as species differentiate. Their

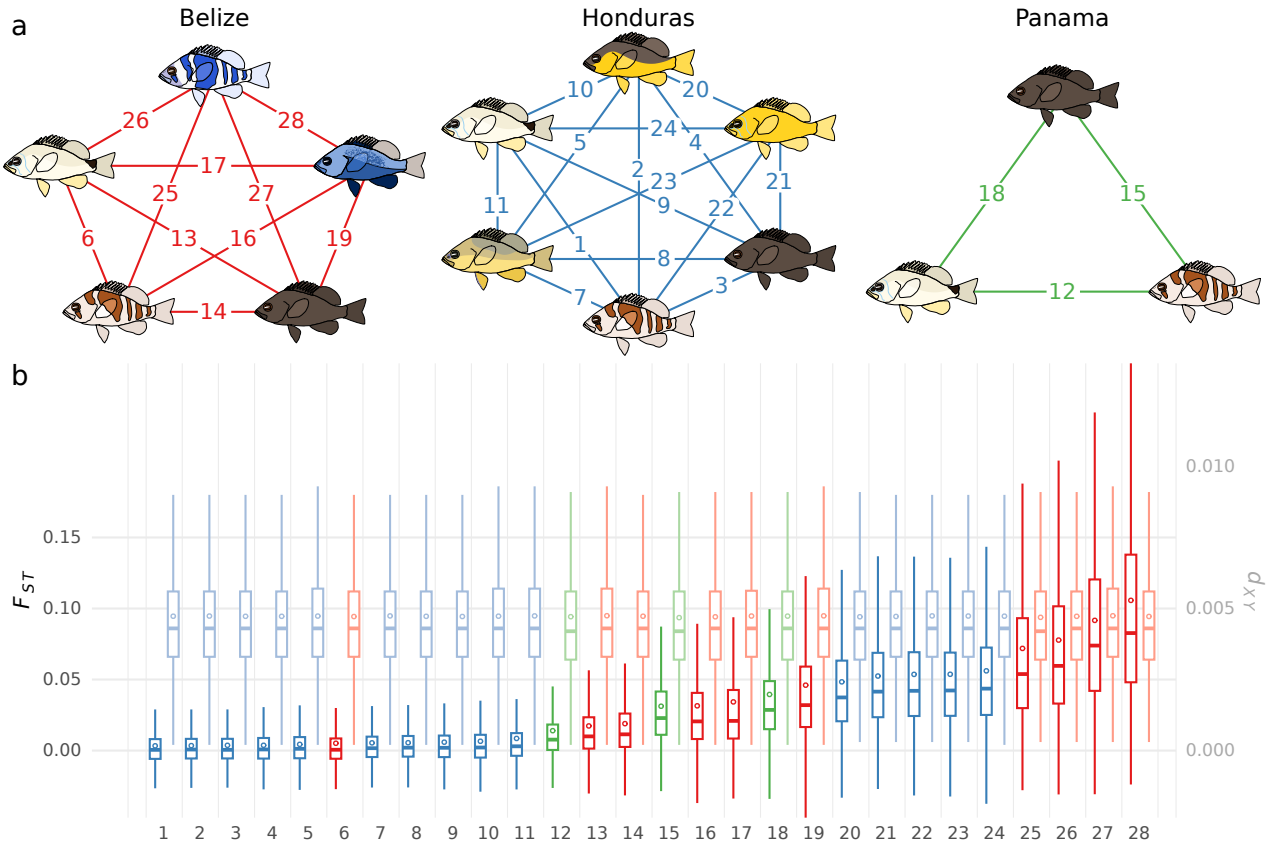
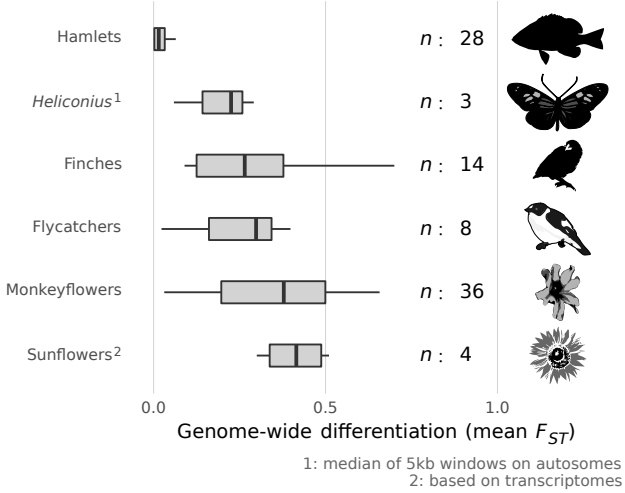
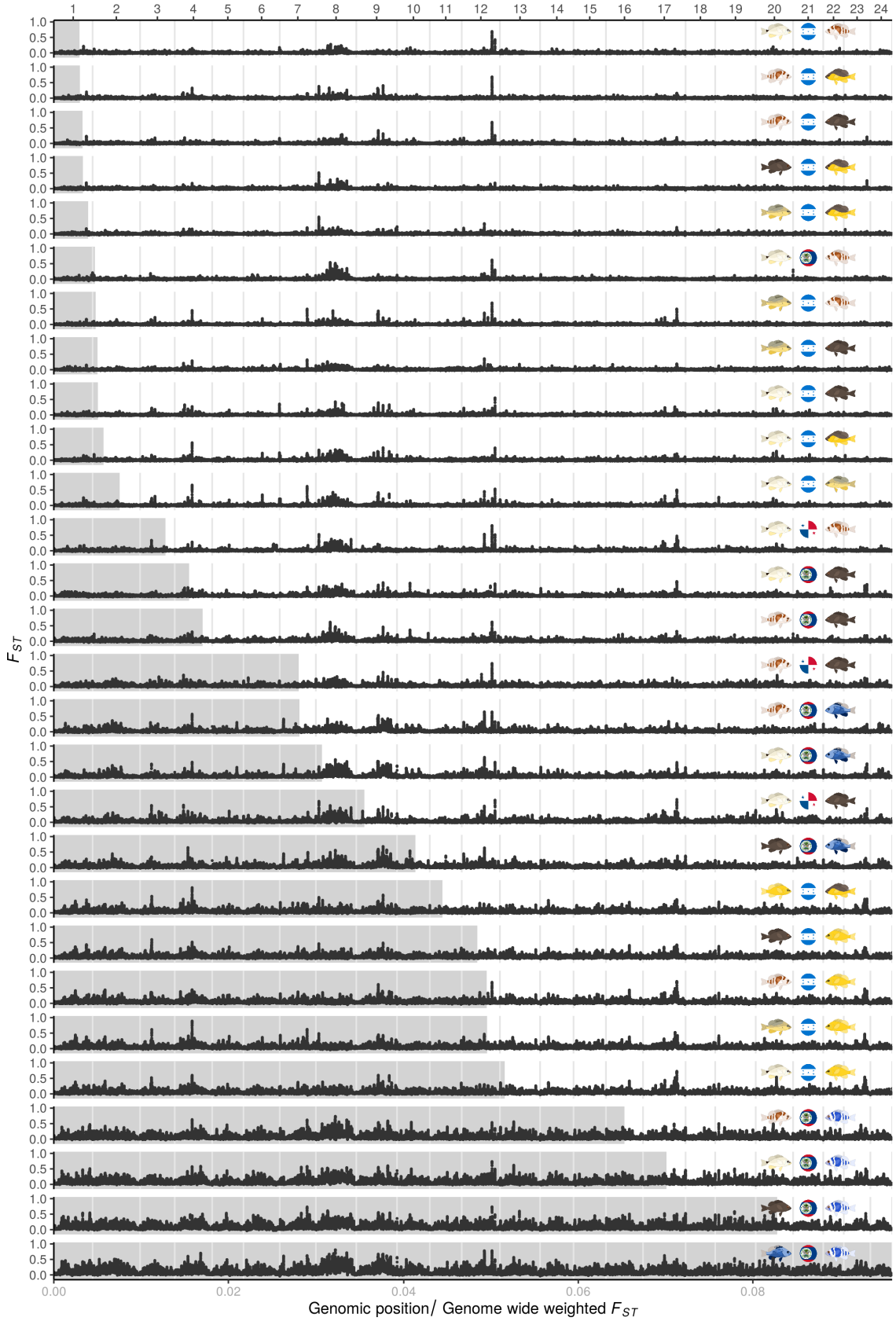


Figure 5.1: Genome-wide differentiation (F_{ST}) and divergence (d_{XY}) among pairs of sympatric species. **a**, the 28 pairs of sympatric species considered in this study, numbered in order of increasing genome-wide differentiation. *N* = 10-13 genomes per species and location. **b**, genetic differentiation (weighted mean F_{ST}, dark colors) and divergence (mean d_{XY}, light colors) between species pairs, averaged over non-overlapping 50 kb windows along the genome. dot: mean, bar: median, box: quartiles, vertical line: outermost data points within $1.5 \times$ the interquartile range from the quartiles.

number does not increase substantially either, with ≤6 peaks with F_{ST}>0.7 per species pair throughout the entire continuum. What happens instead is that background levels of differentiation rise. This process can be dissected by considering the genomic regions that exceed arbitrarily chosen F_{ST} thresholds in the 28 species pairs (Figure 5.2). The number and



Suppl. Figure 5.1: F_{ST} range covered by this versus previous studies on the evolution of genomic architecture. From top to bottom: this study (*Hypoplectrus*), *Heliconius* butterflies (Kronforst *et al.*, 2013), Darwin's finches (*Camarhynchus, Geospiza, Pinaroloxias* and *Platyspiza*)(Han *et al.*, 2017), *Ficedula* flycatchers (Burri *et al.*, 2015), monkeyflowers (*Mimulus*)(Stankowski *et al.*, 2019) and wild sunflowers (*Helianthus*) (Renaut *et al.*, 2013). Bar: median, box: quartiles, horizontal line: outermost data points within $1.5 \times$ the interquartile range from the quartiles. *n*: number of species pairs considered.



Suppl. Figure 5.2: Genetic differentiation (F_{ST}) between pairs of sympatric species. The 28 pairs of sympatric species are shown, in order of increasing genome-wide differentiation. Vertical lines delineate the 24 linkage groups (chromosomes). The grey bar in the background indicates the genome-wide (weighted mean) F_{ST} of the respective pairwise comparison (scale at the bottom).

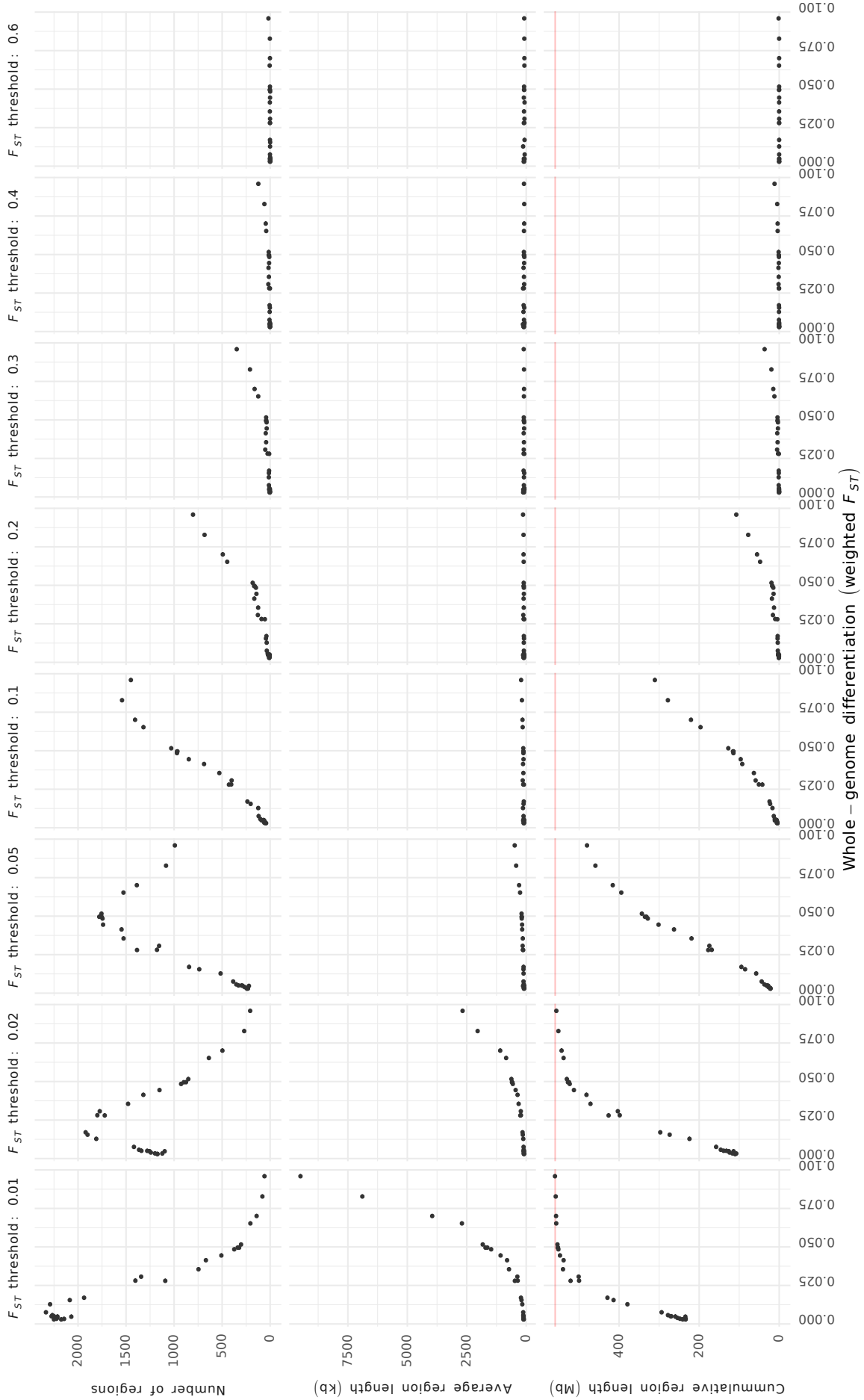
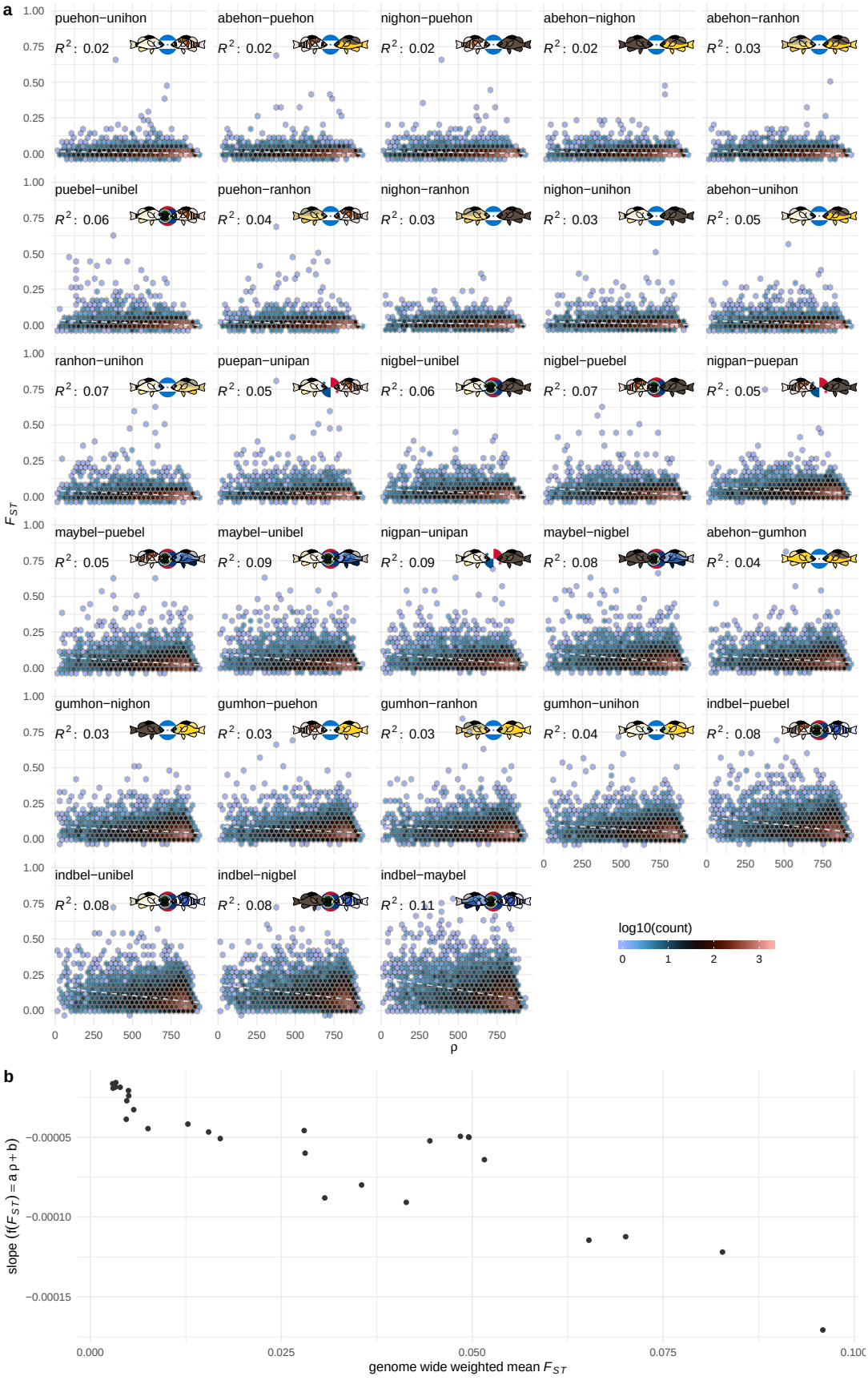
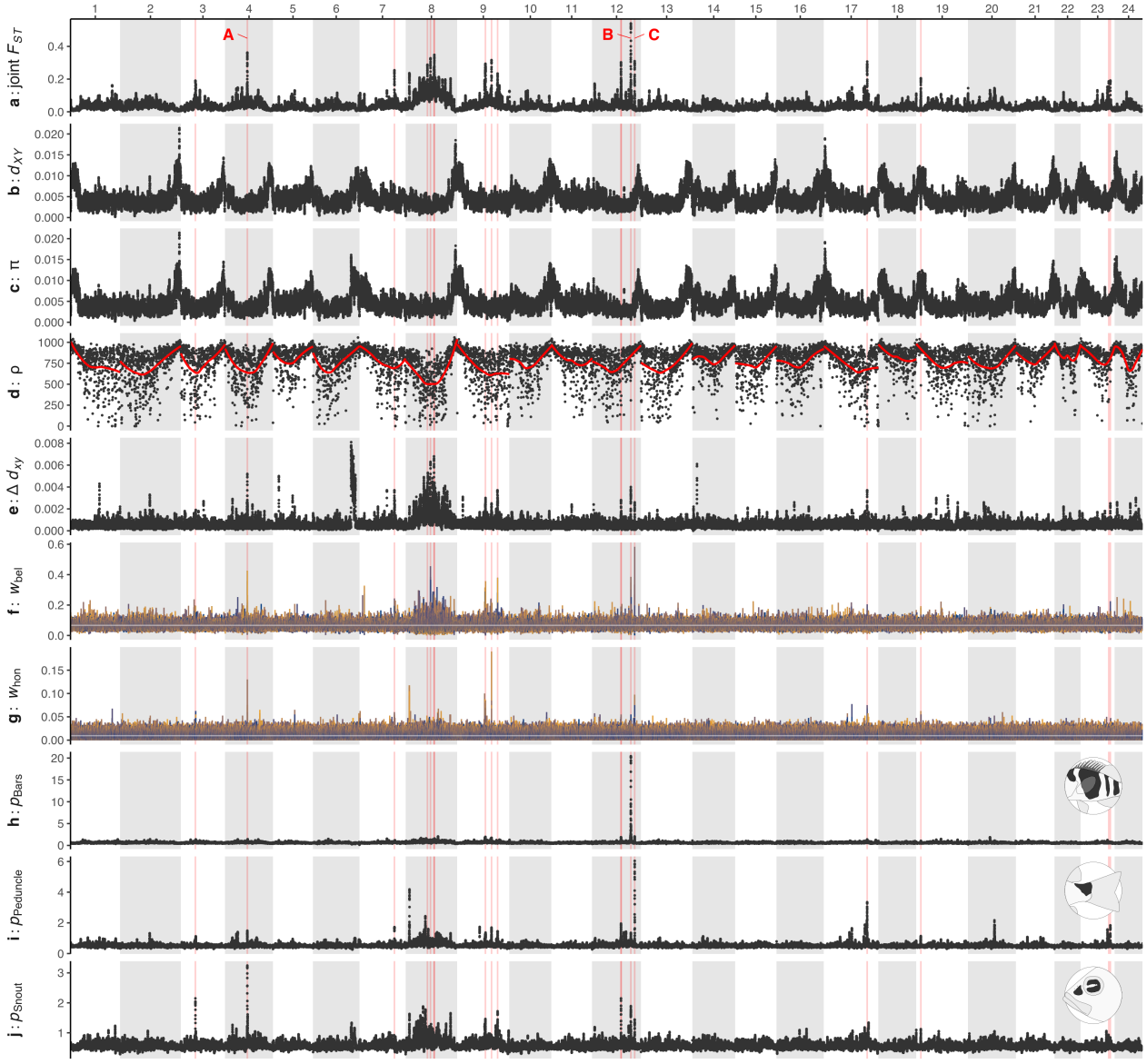


Figure 5.2: Progression of genetic differentiation (F_{ST}) across the 28 species pairs. Each panel shows the 28 species pairs, in order of increasing whole-genome differentiation on the x axis. The columns represent eight arbitrarily chosen F_{ST} threshold values ($> 0.01, ..., 0.6$) and the rows show the number, average length and cumulative length of genomic regions above these threshold values. Analysis based on 50-kb windows. Red line: size of the reference genome.



Suppl. Figure 5.3: Genetic differentiation (F_{ST}) vs. population recombination rate (ρ). Relationship between differentiation and population recombination rate in the 28 pairs of sympatric species, in order of increasing whole-genome differentiation. The slope and correlation coefficient (R^2) between the two variables tend to increase (in absolute value) as species get more differentiated, with higher differentiation in regions of low recombination. Analysis based on 50-kb windows.



Suppl. Figure 5.4: Whole-genome patterns. The alternating white and grey blocks represent the 24 linkage groups (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 samples (species/location). The red vertical lines highlight regions above the 99.8^{th} F_{ST} percentile. **b**, divergence (d_{XY}) between one pair of sympatric species (*H. nigricans-H. puella* from Panama). A similar pattern is observed in all species pairs (Suppl. Fig. 5.5). **c**, diversity (π) of one sample (*H. unicolor* from Panama). A similar pattern is observed in all species (Suppl. Fig. 5.6). **d**, population recombination rate (ρ), calculated over non-overlapping 50 kb windows (loess smoothing in red). **e**, Δ d_{XY} $(max(d_{XY}) - min(d_{XY})$ among the 28 pairs). **f, g**, topology weighting for Belize and Honduras, respectively, along non-overlapping 200 SNP windows. The different colors correspond to different topologies and the white horizontal line indicates the null weighting (i.e. all topologies equally likely). **h, i, j**, Genotype $\times$ Phenotype association for bars, dark saddle on the caudal peduncle and spot on the snout, respectively.

average length of these regions initially increase as species differentiate. As they keep expanding they start fusing with each other, resulting in a reduction in their number, until they span entire chromosomes. Importantly this process is initiated from low F_{ST} values (i.e. from the left in Figure 5.2), indicating that it is really background levels of differentiation that increase. The fact that differentiation peaks do not expand as species differenti-

ate indicates that divergence hitchhiking (Via, 2009; Feder and Nosil, 2010) does not play the prominent role implied by the genic view of speciation in the hamlets. This finding is consistent with the rapid decay of genetic linkage along chromosomes that is observed in this group (Hench *et al.*, 2019; Moran *et al.*, 2019).

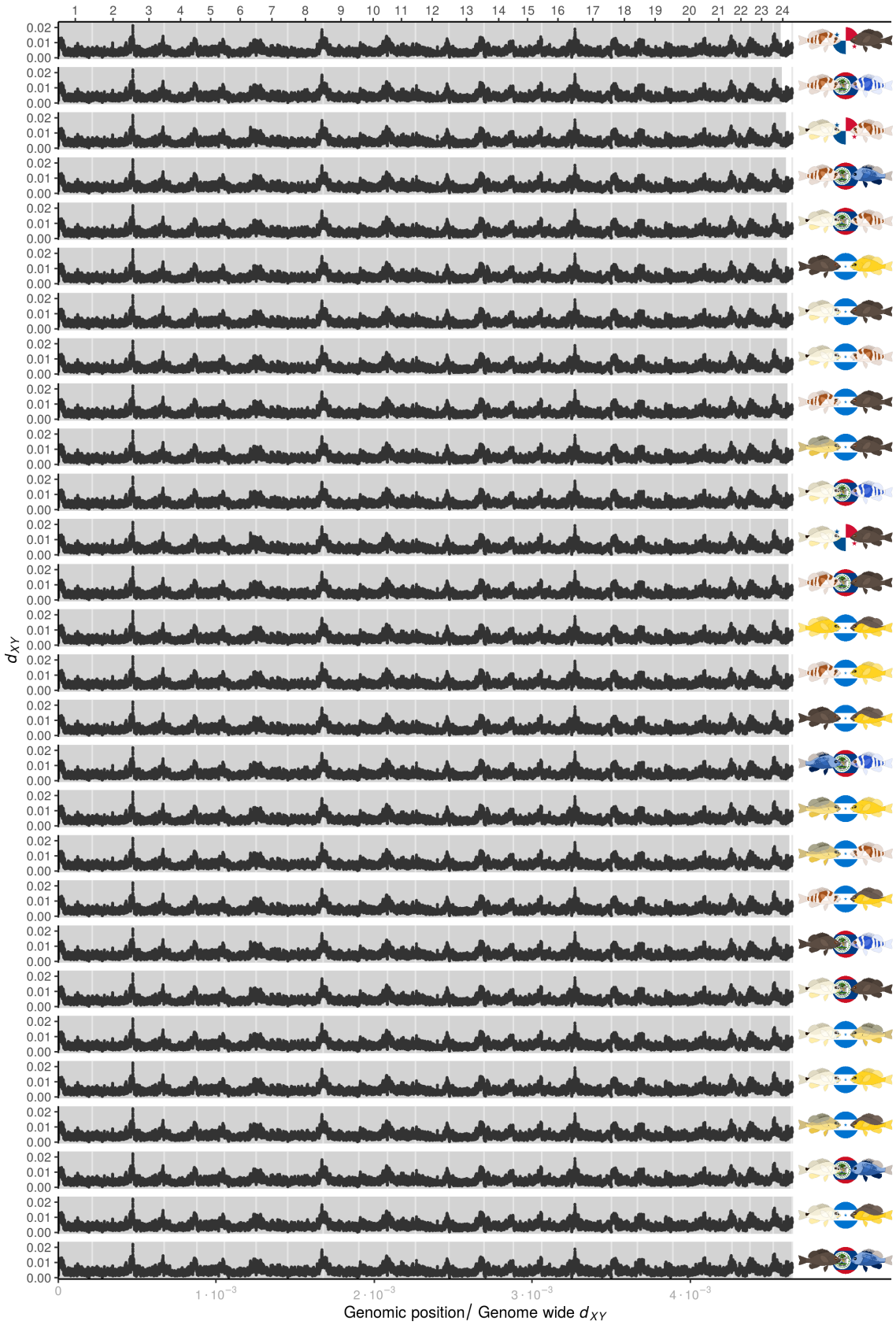
It is also noteworthy that at the lower end of the differentiation continuum, differentiation is largely independent from recombination rate (Suppl. Fig. 5.3 a, top row). In particular, highly differentiated regions are not located in low recombining regions. We nonetheless capture the onset of the effect of recombination, with differentiation accumulating disproportionately in regions of low recombination as species differentiate (Suppl. Fig. 5.3). This effect is particularly strong in a large section of linkage group (chromosome) 8, which we had previously identified as a low-recombining region on the basis of high-density linkage mapping and *de novo* genome assembly (Theodosiou *et al.*, 2016; Hench *et al.*, 2019) (Suppl. Fig. 5.4 a, d).

Genetic differentiation captures changes in allele frequency but not sequence divergence specifically. Caribbean hamlets also present low levels of genetic divergence, with genome-wide d_{XY} between pairs of sympatric species <0.005 (Suppl. Tab. 5.1). Yet genome-wide divergence does not increase along the differentiation continuum (Figure 5.1, Suppl. Fig. 5.5), indicating that divergence is largely driven by ancestral variation. This interpretation is supported by the observation that divergence between species pairs is almost exactly equal to diversity within species (R^2 of the linear regression between the two = 0.955 - 0.996 across all species pairs, Suppl. Fig. 5.4 b, c). Considering the theoretical expectation for divergence (Gillespie and Langley, 1979)

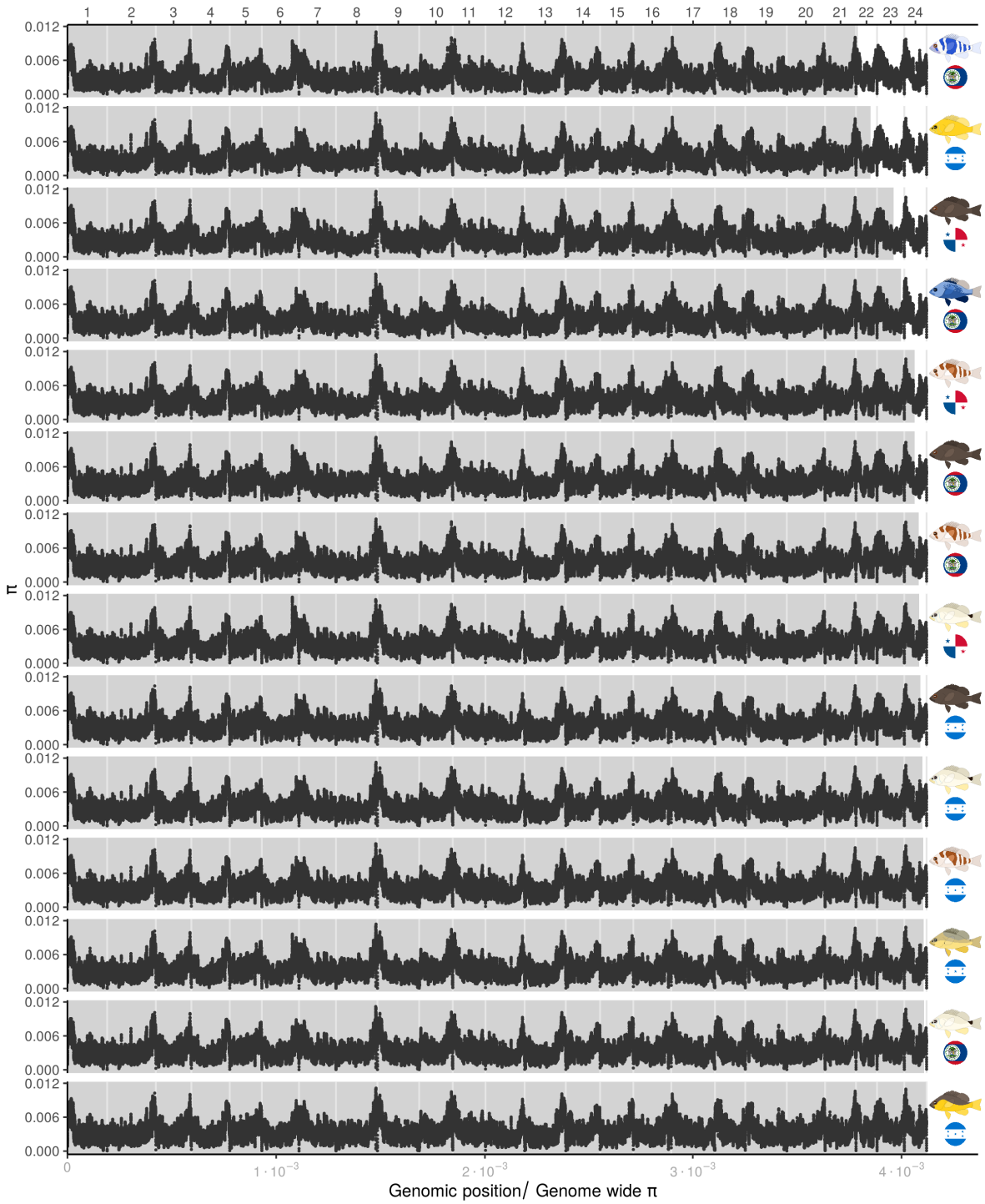
$$E(d_{XY}) = \pi_{Anc} + 2\mu T \qquad (5.1)$$

where π_{Anc} is ancestral diversity, μ mutation rate and T divergence time, d_{XY} reduces to π when T approaches zero. Divergence is also generally elevated in the chromosome peripheries where recombination rate tends to be higher (Suppl. Fig. 5.4 b, d). This is likely an effect of the recombination landscape that shaped ancestral variation, resulting in higher diversity in regions of high recombination (Suppl. Fig. 5.4 d, Suppl. Fig. 5.7). In order to filter out the effect of ancestral diversity from the divergence signal, we considered the range of variation in divergence among species pairs, which we defined as $\Delta\ d_{XY} = max(d_{XY}) - min(d_{XY})$ among the 28 pairs. This statistic varied markedly along the genome (Suppl. Fig. 5.4 e), paralleling in large parts patterns of genetic differentiation (Suppl. Fig. 5.4 a) and indicating that the selection regime varies among species at these loci.

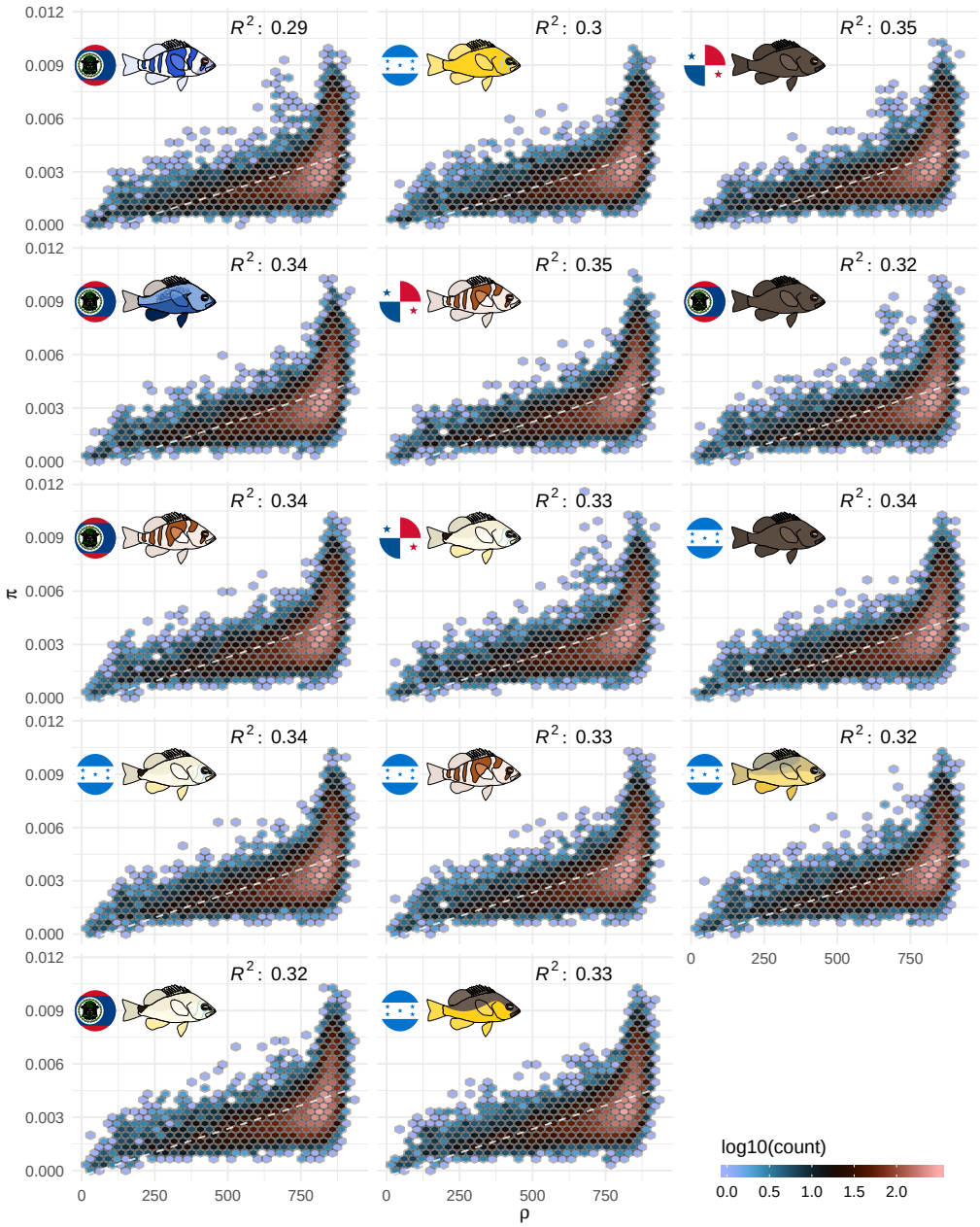
In sum, the genomic origins of adaptive radiation in the hamlets are characterized by low levels of differentiation and divergence except at a few narrow genomic intervals, an increase in background levels of differentiation but not divergence, and an effect of recombination that is clear on diversity (and thereby divergence) but still weak on differentiation. In this latter respect our results contrast with previous studies conducted across higher levels of differentiation (Suppl. Fig. 5.1) that report a marked effect of recombination on



Suppl. Figure 5.5: Genetic divergence (d_{XY}) between pairs of sympatric species. The 28 pairs of sympatric species (Figure 5.1 a) are shown. The grey bar in the background indicates the genome-wide mean d_{XY} of the respective pairwise comparison (scale at the bottom). The genome-wide patterns are similar in all pairs of sympatric species. Analysis based on 50-kb windows.



Suppl. Figure 5.6: Genetic diversity (π). Genome-wide patterns of diversity in the 14 samples (species/populations). The grey bar in the background indicates the genome-wide (mean) π of the respective sample (scale at the bottom). The genome-wide patterns are similar in all species/populations.



Suppl. Figure 5.7: Genetic diversity (π) vs. population recombination rate (ρ). Relationship between diversity and population recombination rate in the 14 samples (species/population). The patterns are similar in all species/populations, with an increase in diversity in regions of high recombination. Analysis based on 50-kb windows.

differentiation (Renaut *et al.*, 2013; Kronforst *et al.*, 2013; Burri *et al.*, 2015; Han *et al.*, 2017; Stankowski *et al.*, 2019). Nevertheless, the accumulation of differentiation in low-recombining regions builds up quickly. Differentiation peaks that stand out clearly at the lowest levels of differentiation become swamped when genome-wide differentiation is still below 0.1 and before any genome-wide divergence develops.

Recent Radiation, old Variation

In order to get a broader perspective on the process of adaptive radiation in the hamlets, we inferred the demographic histories

of all samples (species/populations) using the multiple sequentially Markovian coalescent (MSMC) (Schiffels and Durbin, 2014). The results suggest that hamlets started to diverge less than 10,000 generations ago (Figure 5.3). The MSMC analyses also indicate that historical effective population sizes are high in the hamlets (in the order of $10^4 - 10^5$, Figure 5.3a), which provides the opportunity for the accumulation of standing genetic variation though mutation. Furthermore, the genus *Hypoplectrus* appears much older than the radiation (McCartney *et al.*, 2003). An old lineage/recent radiation scenario is consistent with the stronger effect of recombination rate on diversity and divergence than on differentiation that we report here. Alternatively, gene flow among sympatric species may have been high enough to erase the signature of older demographic events throughout the genome, except at a few narrow genomic intervals that would be resistant to gene flow. Such demographic events could be complex and involve several cycles of 'fission-fusion-fission' (Marques *et al.*, 2019). It would nonetheless take extensive gene flow to completely obliterate the mark of such older demographic events considering that our MSMC

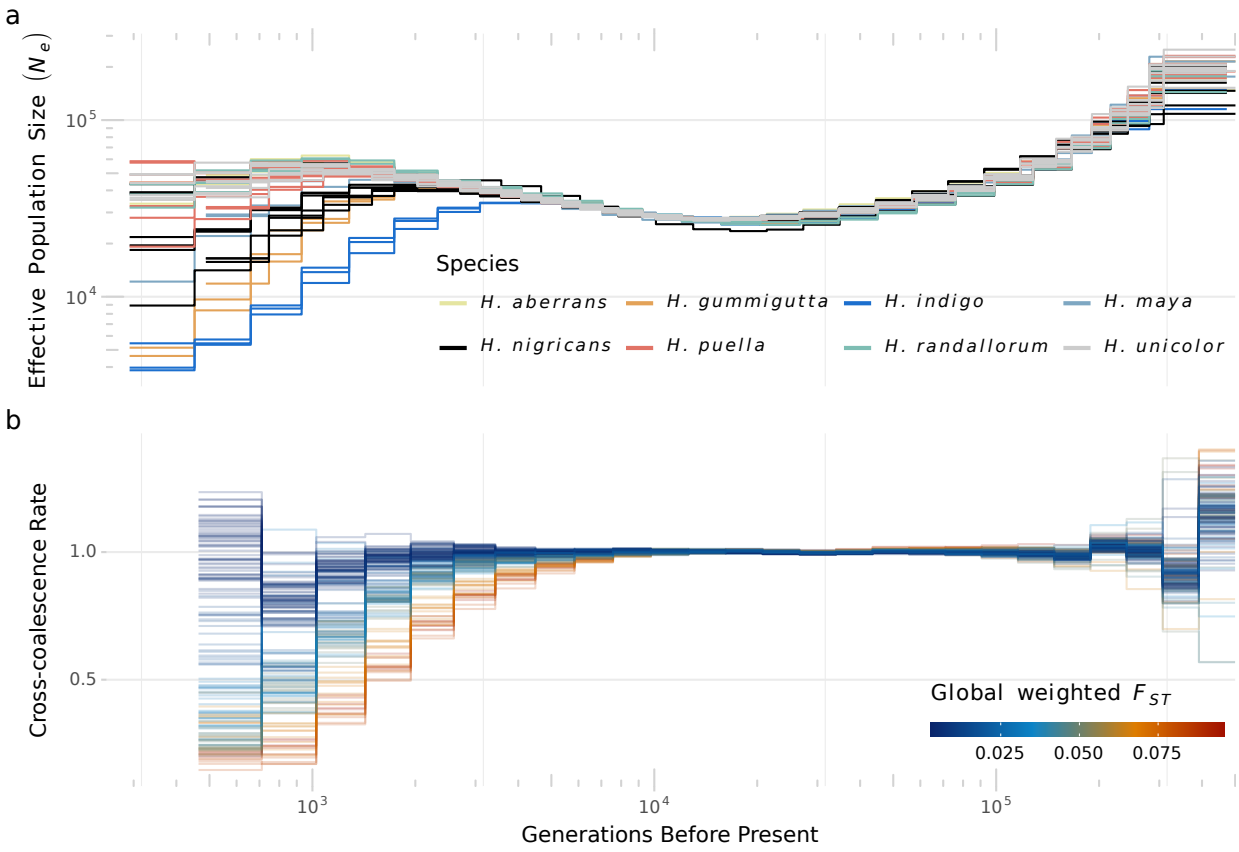
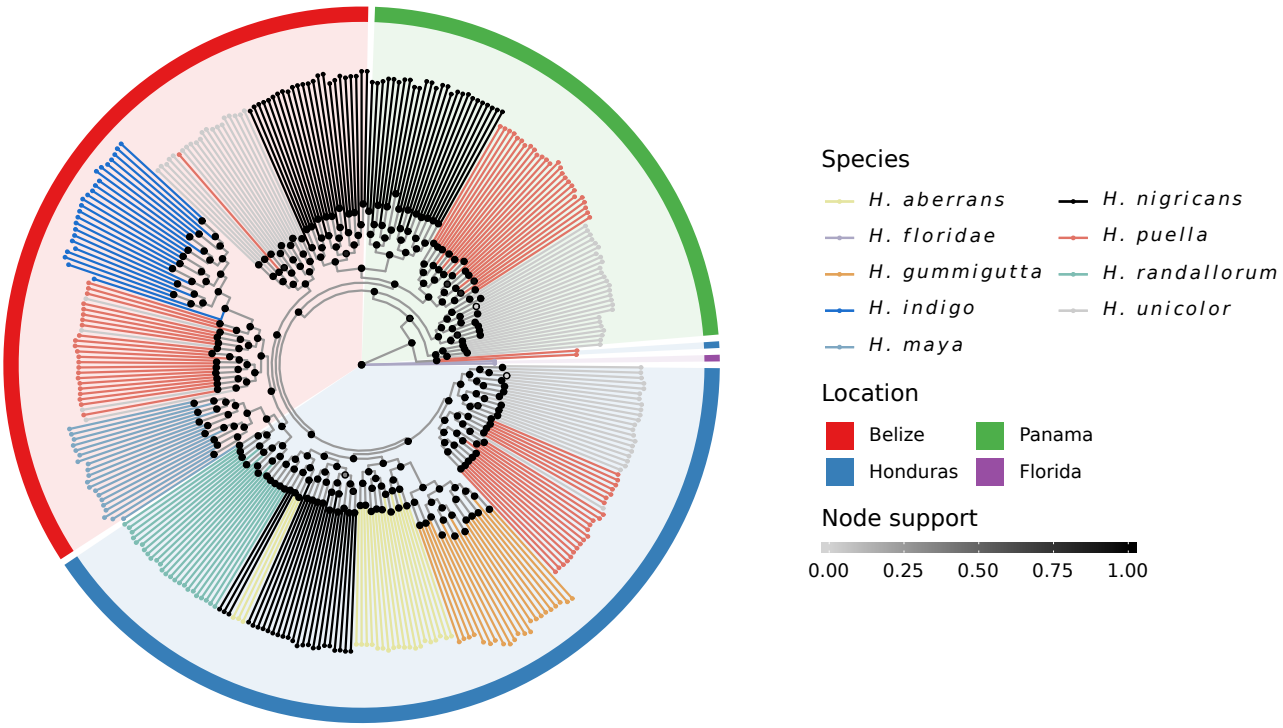


Figure 5.3: Demographic inference. a, inferred history of effective population size. Each line is based on 3-4 genomes from one species/population, with each genome used in only one analysis. **b**, cross-coalescence rates for the 28 pairs of sympatric species, colour-coded by genome-wide F_{ST}. Each line represents an independent run based on two genomes from two sympatric species (total four), with each genome used in only one analysis. A cross-coalescence rate of 1 indicates completely shared ancestry, and a rate of 0 indicates no shared ancestry. All estimates are scaled with a per-site mutation rate $\mu = 3.7 \times 10^{-8}$. The most ancient and two most recent time segments are omitted due to unreliable inference at these extremes.

analyses are based on blocks of common ancestry along the entire genome (Schiffels and Durbin, 2014).

Genetic Bases of Phenotypic Diversification

The young age of the hamlet radiation provides a backdrop to identify the genomic intervals associated with phenotypic diversification before the signal gets swamped by the accumulation of differentiation in low-recombining regions. Given the low levels of divergence and ongoing gene flow observed in the Caribbean hamlets (Hench *et al.*, 2019), the phylogeny of the group is not expected to be well resolved and this is exactly what we observe (Suppl. Fig. 5.8). Nevertheless, a whole-genome bifurcating tree is not an appropriate representation of the data at these early genomic stages of adaptive radiation that are highly reticulated (Mallet *et al.*, 2016). Yet if the phylogenies of specific genomic intervals do sort individuals by phenotype, these are likely to be functionally significant. We used topology weighting by iterative sampling of subtrees (Martin and Van Belleghem, 2017) to dissect the phylogenetic signal along the genome. As expected, this approach failed to identify a leading topology at most 200 SNP (Single Nucleotide Polymorphism) windows throughout the genome. Nevertheless, it revealed a number of topology weighting peaks that match peaks of differentiation and $\Delta\ d_{XY}$ (Suppl. Fig. 5.4 f, g). 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 linkage group 4 (Figure 5.4 a),



Suppl. Figure 5.8: Whole-genome phylogeny. The phylogeny is rooted with *H. floridae* from Florida, which is part of the Gulf of Mexico clade. The tips are color-coded by species and the internal nodes are color-coded with respect to bootstrap support. Note that although the phylogeny is sorted with respect to geography, the three locations Belize, Honduras and Panama do not represent monophyletic groups.

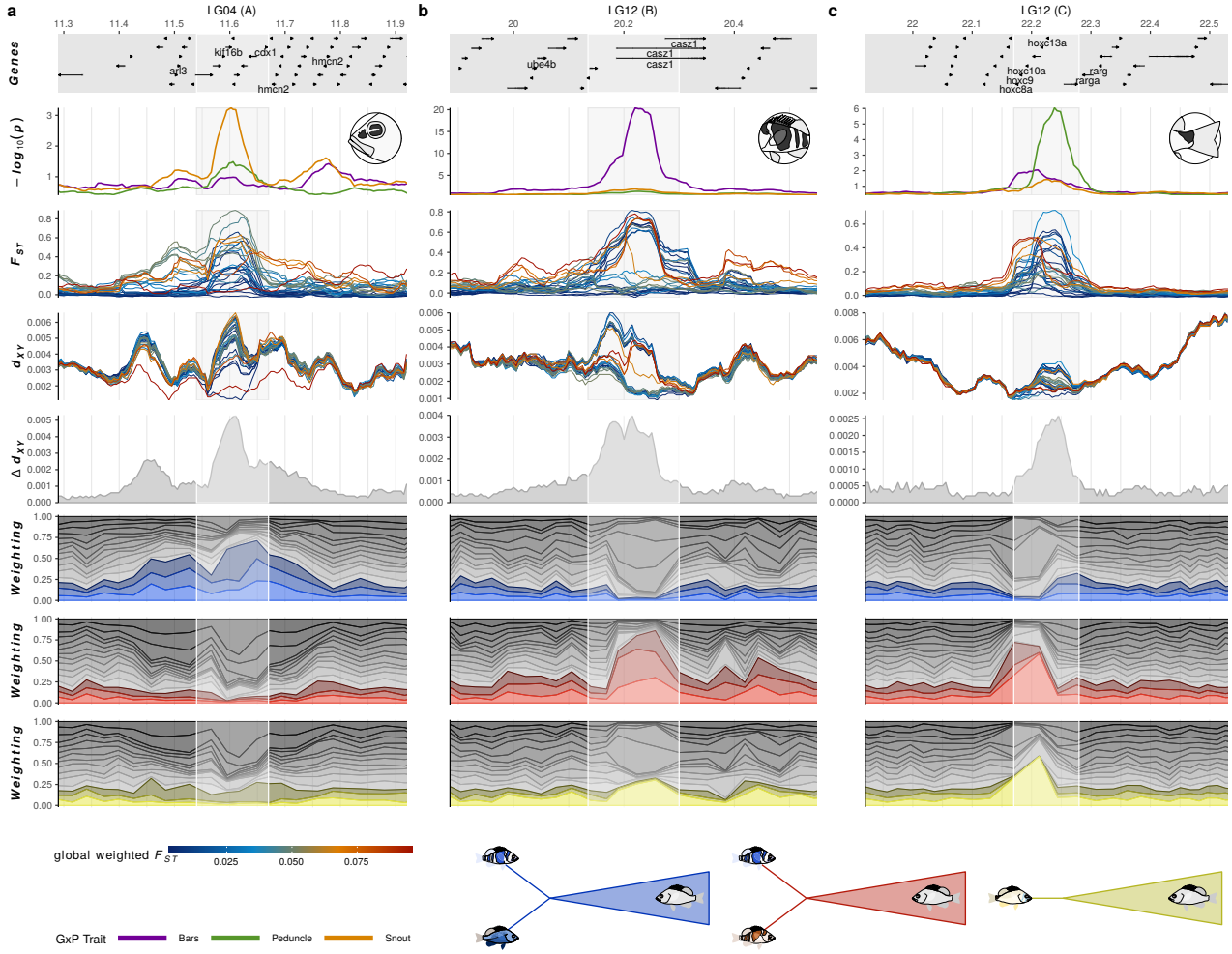
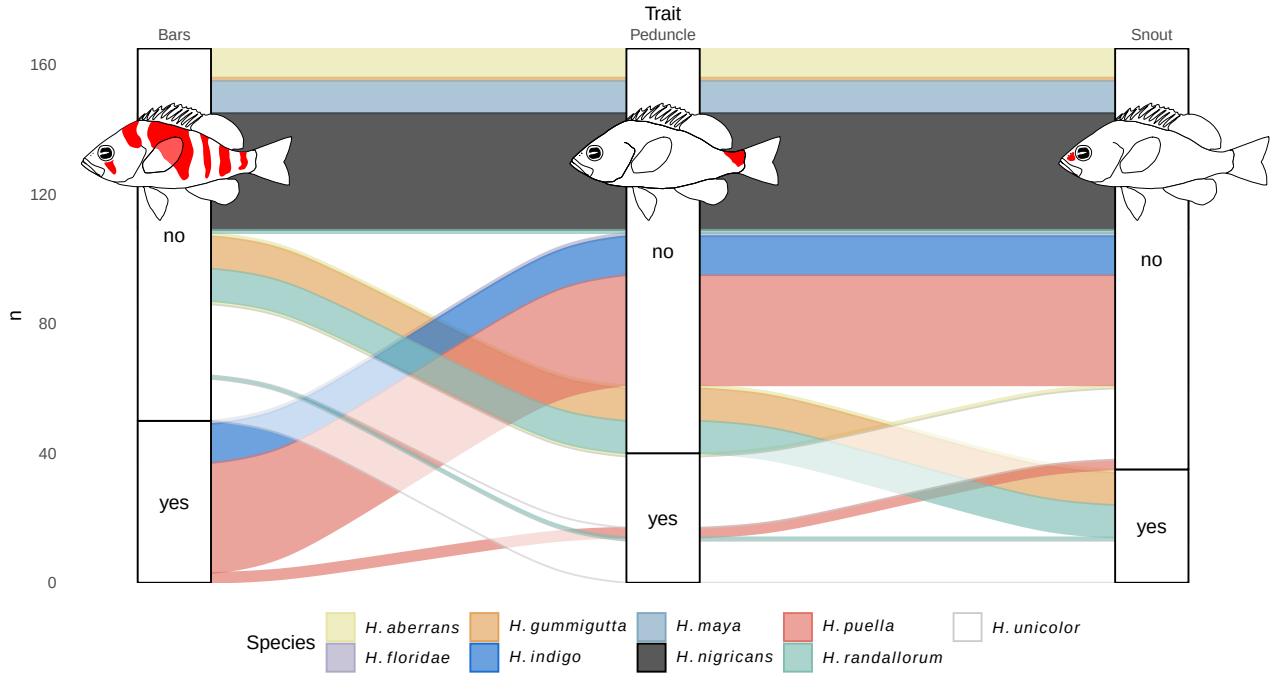


Figure 5.4: Close-up on three genomic regions of interest. The **a, b, c** panels correspond to the A, B, C regions highlighted in Suppl. Fig. 5.4. The x axis shows the position on the respective linkage group (in Mb), with regions above the 99.8^{th} F_{ST} percentile highlighted in light grey. The sub-panels correspond (from top to bottom) to the gene annotation, the log-transformed p value of the Genotype × Phenotype association, genetic differentiation (F_{ST}), genetic divergence (d_{XY}) and $\Delta\, d_{XY}$ ($max(d_{XY}) - min(d_{XY})$ among the 28 pairs of sympatric species), color-coded by genome-wide F_{ST}. All these statistics are calculated in 50 kb sliding-windows with 5 kb increments. The last three sub-panels show the topology weighting for Belize, with particular sets of topologies highlighted. These correspond to the topologies in which the two blue species (*H. indigo* and *H. maya*) are sister species (blue), in which the two species with vertical bars (*H. indigo* and *H. puella*) are sister species (red), and in which *H. unicolor* is singled out without a sister species (yellow). Note that the y scale varies between panels for Genotype × Phenotype association, F_{ST}, d_{XY} and $\Delta\, d_{XY}$.

while two other regions on linkage group 12 are dominated by topologies grouping the two hamlet species that display vertical bars (*H. indigo* and *H. puella*, Figure 5.4 b, c).

In order to further explore the association between genetic variation and specific components of color pattern, we scored all individual samples for the presence or absence of vertical bars, saddle mark on the caudal peduncle and spot on the snout (Suppl. Fig. 5.9). These traits were chosen because they are polymorphic and can be scored unambiguously. Genotype × Phenotype (G × P) association analysis revealed a strong association between the presence or absence of vertical bars and genetic variation in a narrow genomic in-



Suppl. Figure 5.9: Phenotype overview. Distribution of the three binary phenotypes: vertical bars, saddle on caudal peduncle and spot on the snout. Note that these traits can also be polymorphic within species.

terval on linkage group 12 that is also a major differentiation, $\Delta\, d_{XY}$ and topology weighting peak (Suppl. Fig. 5.4 h). Associations with the other two traits were more diffuse, but here again association peaks emerged, notably on linkage group 12 for the saddle mark and on linkage group 4 for the snout spot (Suppl. Fig. 5.4 i,j).

The combination of differentiation, $\Delta\, d_{XY}$, topology weighting and G $\times$ P analysis allows us to start dissecting the genetic bases of phenotypic diversification in the hamlets. The clearest signal was observed in a narrow region of linkage group 12 that shows a strong association with the presence or absence of vertical bars (B in Suppl. Fig. 5.4). The 13 pairs of sympatric species that include one species with vertical bars and one without present high differentiation and divergence at this locus, while the 15 species pairs that include two species with bars or two species without bars do not (Figure 5.4 b). 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 (Hench *et al.*, 2019), is centered on *casz1* (Figure 5.4 b). This gene encodes a castor zinc finger transcription factor that is involved in a number of processes through development, including the development of photoreceptors in mice (Mattar *et al.*, 2015, 2018). A role in vision is also likely in the hamlets since *casz1* is strongly and consistently expressed in the retinal tissue (Hench *et al.*, 2019). 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 possibility is significant from an evolutionary perspective as it would result in a genetic coupling between vision and pigmentation, which would facilitate the evolution of reproductive isolation though visually-based assortative mating that we observe in the hamlets. This is remi-

niscent of the *optix* locus in *Heliconius* butterflies that controls red wing pattern variation (Reed *et al.*, 2011), is expressed in the optic lobe of pupal *Heliconius* (Martin *et al.*, 2014), and falls within a major quantitative trait locus responsible for visually-based assortative mating (Merrill *et al.*, 2019). Alternatively, *casz1* may be involved in vision only and the association with vertical bars may be a by-product of differences among species in their perception.

Our data also provide new insights into the *hoxca* gene cluster on linkage group 12. In line with our preliminary analyses (Hench *et al.*, 2019), we observe a strong association between variation at the *hoxc13a* locus and the presence or absence of a saddle mark on the caudal peduncle (Figure 5.4 c), 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 *hoxc8a* 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 (Jeong *et al.*, 2006; Saenko *et al.*, 2011) and vertebrates (Poelstra *et al.*, 2015). They are arranged and expressed in a sequence that follows the body axis, with 3′ genes expressed anteriorly and 5′ genes posteriorly (Carroll *et al.*, 2005). 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 (Thummel *et al.*, 2004; Jakovlić and Wang, 2016). Vertical bars, on the other hand, are anterior to the saddle mark and the *hoxc8-11a* genes are on the 3′ side of *hoxc13a*.

Other loci are associated with color variation. This is for example the case for a narrow region of linkage group 4 that is centered on the *cdx1* gene and dominated by topologies grouping together the two species that are blue (Figure 5.4 c). Similarly, we have previously shown that black hamlets are differentiated at the *sox10* gene on linkage group 9 that is involved in melanism in zebrafish (Dutton *et al.*, 2001; Elworthy *et al.*, 2003; Hench *et al.*, 2019). A list of all the genes found in the 18 genomic regions above the 99.8^{th} F_{ST} percentile is presented in Suppl. Tab. 5.2 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*).

The genomic origins of adaptive radiation appear largely independent from genomic architecture in the hamlets. Different loci throughout the genome are associated with different components of color pattern variation (bars, marks, color) that form the basis of phenotypic variation in the group. This suggests a modular genetic basis of radiation whereby phenotypic diversity is generated by different combinations of alleles at these loci. Such a modular genetic basis of diversification has been documented in *Heliconius* butterflies (Van Belleghem *et al.*, 2017) and may underlie a variety of adaptive radiations generally.

The Origins of Adaptive Radiation

The buildup of standing genetic variation through a history of high effective population size provides the genomic substrate for phenotypic diversification. Hamlets live on coral reefs, a highly visual environment. They are predators that feed on small invertebrates and fishes (Randall, 1967; Whiteman *et al.*, 2007) but are not particularly specialized and compete with a number of other specialized and generalist fishes on the reef (Thresher, 1978). Furthermore, they are themselves prey to larger visual predators such as groupers. In this context, any variation in color pattern that contributes to improve their hunting efficiency or survival is expected to be strongly selected for. Several hamlets have been proposed to be aggressive mimics, whereby their resemblance in color pattern with other reef fishes that are more abundant and non-predatory improve their ability to approach and attack their prey (Randall, 1968; Thresher, 1978; Puebla *et al.*, 2007, 2018). The vertical bars of the barred and indigo hamlets have also been suggested to be cryptic in specific reef environments (Thresher, 1978; Fischer, 1980a). The other important aspect is the evolution of reproductive isolation between color types. In this respect the simultaneously hermaphroditic mating system of the hamlets (Fischer, 1980b) provides a strong source of sexual selection through mutual mate choice that is expected to facilitate the evolution of assortative mating (Puebla *et al.*, 2012). We suggest that adaptive radiation in the hamlets is driven by the combination of high standing genetic variation, strong selection on color pattern, a modular genetic basis of this trait and a mating system that is conducive to the evolution of reproductive isolation between color morphs, and that these conditions may be common denominators to a variety of adaptive radiations on land, in freshwater and in the sea.

Author Contributions

Conceptualization, K.H., O.P. and W.O.M.; Methodology, K.H. and O.P.; Investigation, K.H. and O.P; Software, K.H; Data Curation, K.H.; Visualization, K.H.; Writing – Original Draft, K.H. and O.P.; Writing – Review & Editing, K.H. and O.P.; Funding Acquisition, O.P. and W.O.M.; Resources, O.P.; Supervision, O.P.

Declaration of Interests

The authors declare they have no competing interests.

Data Accessibility and Supplemental Information

All raw sequencing data are deposited at the European Nucleotide Archive (ENA, project accession number PRJEB35459), individual sample accession numbers are provided in Suppl. Tab. 5.3. Genotypes, phenotypes and population genetic results are deposited on dryad (doi: 10.5061/dryad.280gb5mmt).

5.2. Methods

Software Versions, Parameter Settings and Scripts

Software versions and parameter settings were omitted from the text for readability; software versions are instead listed at the end of the Methods section. Data analysis was managed using nextflow (Di Tommaso *et al.*, 2017). The workflows used to produce our results from raw data to figures are provided in the accompanying repository (accessible from the links highlighted in grey, repository, documentation, hereafter git).

Sequencing

This study is based on a total of 170 genomes that include 167 hamlets and three outgroups (2 x *Serranus tortugarum* and 1 x *Serranus tabacarius*). Fifty genomes are new to this study, 110 are from (Hench *et al.*, 2019) and 10 from (Moran *et al.*, 2019). All new tissue samples were available from previous studies (Puebla *et al.*, 2007, 2012), except for sample #28393 which was collected in 2017 in Bocas del Toro (Panama) under the Smithsonian Tropcial Research Institute IACUC protocol 2017-0101-2020-2, the Panamanian Ministry of Environment permits SC/A-53-16 and SEX/A-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 MagAttract High Molecular Weight kits. Libraries were prepared and sequenced by Novogene and the Institute of Clinical Molecular Biology (Kiel University) on an Illumina HiSeq 4000 (PE, 2x151) to a mean 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 GATK workflow (McKenna *et al.*, 2010) provided by the Broad Institute (DePristo *et al.*, 2011; Van der Auwera *et al.*, 2013). The general workflow is presented below and the exact parameters used for each step are provided in git 1.1 - 1.17 & 2.1 - 2.7. GATK was used to transform the sequences from *fq* to *uBAM* format, assign read groups and mark adapters (git 1.2 – 1.4). The sequences were then back-transformed to *fq* format using GATK, mapped to the hamlet reference genome using BWA (Li and Durbin, 2009) and merged with the *uBAM* files containing the read group information again with GATK (git 1.5). Duplicated reads were removed (git 1.6) and genotype likelihoods were called for each individual (git 1.9) and then merged for all samples (git 1.10). Then, based on the genotype likelihoods from all samples, all individuals were genotyped jointly. This step was duplicated to create two variants of the data set (git 1.11/2.4): a lightweight version with variant sites only (*SNPs only*, git 1.10) and a full version including every callable site - even invariant ones - to calculate π and d_{XY} (*all BP*, git 2.4). SNPs were extracted from the raw genotypes and hard-filtered with respect to quality and missing data (git 1.14/2.6 & 2.7). The *SNPs only* data set was also filtered for a minor allele count $\geqslant 2$ and reduced to bialleic SNPs only using VCFtools (Danecek

et al., 2011) (git 1.14). In preparation of the phasing, the *SNPs only* data set was subset by LG and phase-informative reads were extracted based on the original alignments and the SNPs (git 1.15). Finally, genotypes were phased with SHAPEIT (Delaneau *et al.*, 2012) (git 1.16 & 1.17).

Population Genetic Statistics

Unless stated otherwise, statistics were computed over 50 kb sliding-windows with 5 kb increments for genome-wide displays and over 10 kb windows with 1 kb increments for close-up plots. These two resolutions are referred to as *broad scale* and *fine scale*, respectively.

***F_{ST}*.** Genetic differentiation was computed from the *SNPs only* data set with VCFtools, following Weir & Cockerham (Weir and Cockerham, 1984) and using the weighted mean. It was calculated at both resolutions for each species pair within each location (git 3.9), as well as among all samples (species/ location) jointly (git 3.7).

***d_{XY}*.** Genetic divergence (Nei, 1987) was computed from the *all BP* data set at both resolutions. The data were reformatted to a custom genotype format and divergence was computed using popgenWindows.py (Martin, 2016) (git 4.3 - 4.9).

Δ *d_{XY}*. This statistic, computed at both resolutions, was defined as the range of divergence among pairs of sympatric species: $\Delta\ d_{XY} = max(d_{XY}) - min(d_{XY})$ among the 28 pairs.

π. Nucleotide diversity was based on the *all BP* data set and computed with VCFtools. It was calculated for each population at both resolutions (git 4.19).

***G* $\times$ *P*.** Genotype $\times$ Phenotype associations were based on the *SNPs only* data set and estimated using a linear model with GEMMA (Zhou and Stephens, 2012). The data set was transformed to the plink format using VCFtools and plink (Purcell *et al.*, 2007) (git 3.11 & 3.12). G $\times$ P association was calculated on a SNP basis for the presence/ absence of three phenotypic traits: vertical bars, saddle mark on the caudal peduncle and sport on the snout (git 3.17). The results were averaged over windows at both resolutions (git 3.20). Note that Wald-test *p* values were $-log_{10}$ transformed before averaging, so $\overline{-log_{10}(p)}$ is reported for every window. GEMMA was also run under the linear mixed model, which provided similar results.

ρ. Population recombination rate was estimated using the machine learning R package FastEPRR (Gao *et al.*, 2016). It was based on the *SNPs only* data set considering all samples (except outgroups) and calculated within non-overlapping windows of 50 kb using 250 parallel jobs (git 6.4 - 6.11).

H_o. Observed heterozygosity was estimated from the *all BP* data set on an individual basis. For each genome, allele counts were recorded using VCFtools and re-coded as ($0 =$ homozygous, $1 =$ heterozygous) using bash (git 7.2). Heterozygosity was then averaged over non-overlapping 50 kb windows using R (git 7.3).

N_e and cross-coalescence rate. Demographic history was inferred using the multiple sequentially Markovian coalescent method implemented in MSMC2 (Schiffels and Durbin, 2014). This analysis was based on the phased *SNPs only* data set, which was prepared following the MSMC2 authors recommendations (Schiffels, 2014) as detailed in

(Moran *et al.*, 2019). 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 $10\times$ and twice the individual mean coverage, git 8.10). Individuals from each species and location were randomly grouped into sets of 4 or 3, with each individual included in only one set (git 8.12), and individual masks were combined to create the MSMC2 input files (git 8.16 & 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 segmentation pattern of $1 \cdot 2 + 25 \cdot 1 + 1 \cdot 2 + 1 \cdot 3$, and the average of Watterson's estimator across input data sets ($\theta = 2.55 \cdot 10^{-3}$, git 8.18 & 8.23).

Whole-genome phylogeny. The whole-genome phylogeny was based on the *SNPs only* data set, which was filtered to include only SNPs with a minor allele count ≥ 3 (git 5.6). The data were transformed from vcf to a custom genotype format using parseVCF.py (Martin, 2016) and concatenated to generate two whole-genome pseudo-haplotypes per sample in fasta format (git 5.7). The phylogeny was constructed from the pseudo-haplotypes using fasttree with default settings (git 5.8).

Topology weighting. Topology weighting was conducted for the Belize and Honduras samples independently. The *SNPs only* data set was subsetted to include only hamlets from the respective location, and filtered to include only SNPs with a minor allele count ≥ 3 (git 5.12). The data were then split by LG and converted to a custom genotype format (git 5.14). Using phyml_sliding_windows.py (Martin, 2016), we applied PhyML (Guindon *et al.*, 2010) to build phylogenies within 200 SNP, non-overlapping sliding-windows along all LGs (git 5.17). Topology weighting was conducted on the resulting phylogenies using twisst (Martin and Van Belleghem, 2017) (git 5.18).

Visualisation

All results were plotted using R (git 9). The details of the visualisation are provided in the R scripts and their documentation (git docs/index.html). Besides the scripts within the github repository, the visualisation relied on the three custom R packages (hypogen, hypoimg and GenomicOriginsScripts) which can also be accessed via github (link). The following R packages (including their dependencies) were used: ape (Paradis and Schliep, 2018) (5.3), FastEPRR (Hao and Gao, 2019) (1.0), furrr (Vaughan and Dancho, 2019) (0.1.0.9002), GenomicOriginsScripts (Hench, 2019a) (0.0.0.9000), geomfactory (Hench, 2019b) (0.0.3.1), ggalluvial (Brunson, 2019) (0.10.0.1), ggrepel (Slowikowski, 2019) (0.8.1), ggtree (Yu *et al.*, 2018) (1.16.5), hypogen (Hench, 2019c) (1.0.0.0), hypoimg (Hench, 2019d) (1.0.0.5), igraph (Csardi and Nepusz, 2006) (1.2.4.1), logisticPCA (Landgraf and Lee, 2015) (0.2.9000), patchwork (Pedersen, 2019) (1.0.0.9000), phytools (Revell, 2012) (0.6.99), tidyverse (Wickham *et al.*, 2019) (1.2.1.9000) and vroom (Hester and Wickham, 2019) (1.0.2.9000).

Software Versions

The following software versions were used: bwa (0.7.17-r1188), Fasttree (2.1.10), gatk (4.0.8.1), gemma (0.98), msmsc2 (2.0.0), nextflow (0.31.1.4886), plink (v1.90b4), Python (2.7.15), R (3.5.2, for analysis), R (3.6.1, for visualisation), samtools (1.9), shapeit (v2.r837) and vcftools (0.1.15).

Supplementary Tables

Suppl. Table 5.1: Genome-wide F_{ST} and d_{XY} between pairs of sympatric species. F_{ST} values are displayed below the diagonal and d_{XY} above. The first three letters of each population indicates the species (*H. aberrans* (abe), *H. gummigutta* (gum), *H. indigo* (ind), *H. maya* (may), *H. nigricans* (nig), *H. puella* (pue), *H. randallorum* (ran), *H. unicolor* (uni) and the last three letters the location (Belize (bel), Honduras (hon) and Panama (pan)). Note that the F_{ST} values differ slightly from Figure 5.1 since the table shows the genome-wide average over all SNPs while Figure 5.1 is based on 50 kb sliding-windows.

a) Belize

Population	ind|bel	may|bel	nig|bel	pue|bel	uni|bel
ind|bel	-	0.00472	0.00472	0.00469	0.00473
may|bel	0.09590	-	0.00474	0.00471	0.00473
nig|bel	0.08276	0.04136	-	0.00473	0.00475
pue|bel	0.06527	0.02812	0.01701	-	0.00468
uni|bel	0.07009	0.03069	0.01548	0.00471	-

b) Honduras

Population	abe|hon	gum|hon	nig|hon	pue|hon	ran|hon	uni|hon
abe|hon	-	0.00471	0.00472	0.00470	0.00470	0.00473
gum|hon	0.04446	-	0.00472	0.00473	0.00474	0.00470
nig|hon	0.00334	0.04845	-	0.00472	0.00469	0.00468
pue|hon	0.00297	0.04952	0.00327	-	0.00471	0.00469
ran|hon	0.00391	0.04957	0.00499	0.00477	-	0.00471
uni|hon	0.00569	0.05159	0.00502	0.00293	0.00755	-

c) Panama

Population	nig|pan	pue|pan	uni|pan
nig|pan	-	0.00465	0.00473
pue|pan	0.02799	-	0.00468
uni|pan	0.03552	0.01277	-

Suppl. Table 5.2: Genes located in the 50 kb windows that are above the 99.8^{th} F_{ST} percentile.

LG	Id	Start	End	n	Genes
LG03	LG03_1	7575001	7645000	6	*grk7a*, *atp1b3*, *gk5*, *hp...06295* *hp...06296*, *hp...06297*
LG04	LG04_1	11540001	11670000	9	*hp...07427*, *ndst1*, *cd74*, *kif16b* *slc35a4*, *hmgxb3*, *csf1r1*, *pdgfrb* *cdx1*
LG07	LG07_1	18215001	18315000	4	*dpysl3*, *jakmip2*, *slc6a13*, *p4ha2*
LG08	LG08_1	11355001	11440000	8	*ccl20*, *hp...11291*, *slc1a3*, *tax1bp3* *hsh2d*, *rab8a*, *rps27l*, *tpm1*
LG08	LG08_2	12955001	13060000	5	*mpnd*, *sh3gl1*, *hp...11382*, *rorb* *sema6b*
LG08	LG08_3	14805001	14890000	4	*arid3a*, *ctdspl2a*, *tmem79*, *ndufs7*
LG08	LG08_4	14955001	15070000	5	*kank3*, *hp...11472*, *pik3r2*, *ifi30* *pde4d*
LG09	LG09_1	14535001	14815000	2	*ctnna2*, *lrrtm1*
LG09	LG09_2	17820001	17925000	7	*smdt1*, *triobp*, *hp...12847*, *kcnj12* *sox10*, *rnaseh2a*, *mast1*
LG09	LG09_3	21005001	21080000	3	*rnf24*, *smox*, *fbxo41*
LG12	LG12_1	15030001	15140000	6	*hp...15944*, *ren*, *csf1*, *hp...15947* *eif2d*, *slc12a5*
LG12	LG12_2	15160001	15210000	9	*rbm39*, *epabp-a*, *143b1*, *ogn*, *smim4* *kctd6*, *hp...15957*, *abhd6*, *slmap*
LG12	LG12_3	20135001	20300000	3	*tardbp*, *c1orf127*, *casz1*
LG12	LG12_4	22170001	22280000	7	*hoxc8a*, *hoxc9*, *hoxc10a*, *hoxc11a* *hoxc12a*, *hoxc13a*, *calcoco1*
LG17	LG17_1	22505001	22645000	17	*mafb*, *deptor*, *adnp*, *gata2* *rab7*, *hcfc1*, *sws2a*β, *sws2a*α *sws2b*, *lws*, *gnl3l*, *tfe3* *mdfic2*, *cxxc1*, *mbd1*, *ccdc120* *srpk3*
LG19	LG19_1	2360001	2445000	2	*kitlg*, *slc23a2*
LG23	LG23_1	14635001	14685000	2	*hp...26758*, *nxpe3*
LG23	LG23_2	15455001	15530000	5	*crys*, *gp2*, *krr1*, *glipr1l1* *st3gal1*

Suppl. Table 5.3: New samples sequenced for this study. Locations correspond to Belize (bel), Honduras (hon) and Panama (pan) and species to *H. aberrans* (abe), *H. gummigutta* (gum), *H. indigo* (ind), *H. randallorum* (ran), *S. tabacarius* (tab) and *S. tortugarum* (tor).

Nr	ID	Species	Location	Date	Latitude	Longitude	Accession Number
1	17996	ind	bel	2004-07-27	16.8008	-88.0789	ERS4141229
2	17997	ind	bel	2004-07-27	16.8008	-88.0789	ERS4141230
3	17998	ind	bel	2004-07-27	16.8008	-88.0789	ERS4141231
4	17999	ind	bel	2004-07-27	16.8008	-88.0789	ERS4141232
5	18000	ind	bel	2004-07-27	16.8008	-88.0789	ERS4141233
6	18195	ind	bel	2004-07-26	16.8058	-88.0792	ERS4141234
7	18222	ind	bel	2004-07-27	16.8008	-88.0789	ERS4141235
8	18225	ind	bel	2004-07-27	16.8008	-88.0789	ERS4141236
9	18226	ind	bel	2004-07-27	16.7839	-88.0767	ERS4141237
10	18227	ind	bel	2004-07-27	16.7839	-88.0767	ERS4141238
11	18237	ind	bel	2004-07-27	16.7839	-88.0767	ERS4141239
12	18238	ind	bel	2004-07-27	16.7839	-88.0767	ERS4141240
13	20418	gum	hon	2006-06-03	16.03	-83.3286	ERS4141241
14	20419	gum	hon	2006-06-03	16.03	-83.3286	ERS4141242
15	20420	gum	hon	2006-06-03	16.03	-83.3286	ERS4141243
16	20421	ran	hon	2006-06-03	16.03	-83.3286	ERS4141244
17	20425	abe	hon	2006-06-03	16.03	-83.3286	ERS4141245
18	20426	gum	hon	2006-06-04	15.9558	-83.2931	ERS4141246
19	20427	gum	hon	2006-06-04	15.9558	-83.2931	ERS4141247
20	20428	gum	hon	2006-06-04	15.9558	-83.2931	ERS4141248
21	20429	ran	hon	2006-06-04	15.9558	-83.2931	ERS4141249
22	20430	ran	hon	2006-06-04	15.9558	-83.2931	ERS4141250
23	20433	abe	hon	2006-06-04	15.9558	-83.2931	ERS4141251
24	20478	tab	hon	2006-06-06	-	-	ERS4141252
25	20613	ran	hon	2006-06-05	15.9558	-83.2931	ERS4141253
26	20615	gum	hon	2006-06-05	15.9558	-83.2931	ERS4141254
27	20617	gum	hon	2006-06-05	15.9558	-83.2931	ERS4141255
28	20641	gum	hon	2006-06-05	15.9558	-83.2931	ERS4141256
29	20642	gum	hon	2006-06-05	15.9558	-83.2931	ERS4141257
30	20643	gum	hon	2006-06-05	15.9558	-83.2931	ERS4141258
31	20644	abe	hon	2006-06-05	15.9558	-83.2931	ERS4141259
32	20696	ran	hon	2006-06-12	16.1103	-86.9539	ERS4141260
33	20759	abe	hon	2006-06-06	15.25	-82.6167	ERS4141261
34	20761	abe	hon	2006-06-06	15.25	-82.6167	ERS4141262
35	20762	abe	hon	2006-06-06	15.25	-82.6167	ERS4141263
36	20861	abe	hon	2006-06-07	15.25	-82.6167	ERS4141264
37	20862	abe	hon	2006-06-07	15.25	-82.6167	ERS4141265
38	20864	abe	hon	2006-06-07	15.25	-82.6167	ERS4141266
39	20866	abe	hon	2006-06-07	15.25	-82.6167	ERS4141267
40	20867	abe	hon	2006-06-07	15.25	-82.6167	ERS4141268
41	20892	ran	hon	2006-06-08	16.445	-85.875	ERS4141269
42	20893	ran	hon	2006-06-08	16.445	-85.875	ERS4141270
43	20894	ran	hon	2006-06-08	16.445	-85.875	ERS4141271
44	20896	ran	hon	2006-06-08	16.445	-85.875	ERS4141272
45	20922	ran	hon	2006-06-09	16.4736	-85.9239	ERS4141273
46	20923	ran	hon	2006-06-09	16.4736	-85.9239	ERS4141274
47	20980	ran	hon	2006-06-10	16.4975	-85.9028	ERS4141275
48	28393	tor	pan	2017-02-07	9.3014	-82.2941	ERS4141276
49	PL17_160	flo	flo	2017-07-07	24.5077	-81.5714	ERS4141277
50	s_tort_3	tor	pan	2016	-	-	ERS4141278

Chapter 5 References

Berner, D. and Salzburger, W. (2015). The genomics of organismal diversification illuminated by adaptive radiations. *Trends in Genetics*, 31(9):491–499.

Brunson, J.C. (2019). *ggalluvial: Alluvial Diagrams in 'ggplot2'*. R package version 0.10.0.001.

Burri, R. *et al*. (2015). Linked selection and recombination rate variation drive the evolution of the genomic landscape of differentiation across the speciation continuum of ficedula flycatchers. *Genome research*, 25(11):1656–1665.

Carroll, S.B., Grenier, J.K. and Weatherbee, S.D. (2005). *From DNA to diversity, molecular genetics and the evolution of animal design*. Blackwell Publishing Ltd, Oxford, 2 edition.

Csardi, G. and Nepusz, T. (2006). The igraph software package for complex network research. *InterJournal*, Complex Systems:1695.

Danecek, P. *et al*. (2011). The variant call format and VCFtools. *Bioinformatics*, 27(15):2156–2158.

Delaneau, O., Marchini, J. and Zagury, J.F. (2012). A linear complexity phasing method for thousands of genomes. *Nat. Methods*, 9:179.

DePristo, M.A. *et al*. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet*, 43(5):491–498.

Di Tommaso, P. *et al*. (2017). Nextflow enables reproducible computational workflows. *Nature Biotechnology*, 35:316.

Dutton, K.A. *et al*. (2001). Zebrafish colourless encodes sox10 and specifies non-ectomesenchymal neural crest fates. *Development*, 128(21):4113–4125.

Elworthy, S. *et al*. (2003). Transcriptional regulation of mitfa accounts for the sox10 requirement in zebrafish melanophore development. *Development*, 130(12):2809–2818.

Feder, J.L. and Nosil, P. (2010). The efficacy of divergence hitchhiking in generating genomic islands during ecological speciation. *Evolution: International Journal of Organic Evolution*, 64(6):1729–1747.

Fischer, E.A. (1980a). Speciation in the hamlets (hypoplectrus: Serranidae): a continuing enigma. *Copeia*, pages 649–659.

Fischer, E.A. (1980b). The relationship between mating system and simultaneous hermaphroditism in the coral-reef fish, *Hypoplectrus nigricans* (Serranidae). *Anim. Behav.*, 28(May):620–633.

Gao, F. *et al*. (2016). New Software for the Fast Estimation of Population Recombination Rates (FastEPRR) in the Genomic Era. *G3: Genes, Genomes, Genetics*, 6(6):1563–1571.

Gillespie, J.H. and Langley, C.H. (1979). Are evolutionary rates really variable? *Journal of Molecular Evolution*, 13(1):27–34.

Guindon, S. *et al*. (2010). New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Per-

formance of PhyML 3.0. *Systematic Biology*, 59(3):307–321.

Han, F. *et al*. (2017). Gene flow, ancient polymorphism, and ecological adaptation shape the genomic landscape of divergence among darwin's finches. *Genome research*, 27(6):1004–1015.

Hao, Z. and Gao, F. (2019). *FastEPRR: FastEPRR: a Fast Estimator for the Population Recombination Rate*. R package version 2.0.

Hench, K. (2019a). *GenomicOriginsScripts: Provides the R scripts for the paper "The genomic origins of a marine radiation"*. R package version 0.0.0.90015.

Hench, K. (2019b). *geomfactory: Parallel ggplot color scales*. R package version 0.0.3.1.

Hench, K. (2019c). *hypogen: Provides Hypoplectrus PopGen Data and Functions*. R package version 1.0.0.1.

Hench, K. (2019d). *hypoimg: Provides Hypoplectrus image annotations*. R package version 1.0.1.1.

Hench, K. *et al*. (2019). Inter-chromosomal coupling between vision and pigmentation genes during genomic divergence. *Nature Ecology & Evolution*, 3(4):657–667.

Hester, J. and Wickham, H. (2019). *vroom: Read and Write Rectangular Text Data Quickly*. R package version 1.0.2.9000.

Jakovlić, I. and Wang, W.M. (2016). Expression of Hox paralog group 13 genes in adult and developing *Megalobrama amblycephala*. *Gene Expr. Patterns*, 21(2):63–68.

Jeong, S., Rokas, A. and Carroll, S.B. (2006). Regulation of body pigmentation by the abdominal-b hox protein and its gain and loss in drosophila evolution. *Cell*, 125(7):1387–1399.

Kronforst, M.R. *et al*. (2013). Hybridization reveals the evolving genomic architecture of speciation. *Cell reports*, 5(3):666–677.

Landgraf, A.J. and Lee, Y. (2015). Dimensionality reduction for binary data through the projection of natural parameters. Technical Report 890, Department of Statistics, The Ohio State University.

Li, H. and Durbin, R. (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics*, 25(14):1754.

Mallet, J., Besansky, N. and Hahn, M.W. (2016). How reticulated are species? *BioEssays*, 38(2):140–149.

Marques, D.A., Meier, J.I. and Seehausen, O. (2019). A combinatorial view on speciation and adaptive radiation. *Trends in ecology & evolution*.

Martin, A. *et al*. (2014). Multiple recent co-options of optix associated with novel traits in adaptive butterfly wing radiations. *EvoDevo*, 5(1):7.

Martin, S.H. (2016). genomics_general: General tools for genomic analyses; a github repository.

Martin, S.H. and Van Belleghem, S.M. (2017). Exploring Evolutionary Relationships Across the Genome Using Topology Weighting. *Genetics*, 206(1):429–438.

Mattar, P. *et al*. (2015). A conserved regulatory logic controls temporal identity in mouse neural progenitors. *Neuron*, 85(3):497–504.

Mattar, P. *et al*. (2018). Casz1 controls higher-order nuclear organization in rod photoreceptors. *Proceedings of the National Academy of Sciences*, 115(34):E7987–E7996.

McCartney, M.A. *et al*. (2003). Genetic mosaic in a marine species flock. *Mol. Ecol.*, 12(11):2963–2973.

McKenna, A. *et al*. (2010). The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, 20(9):1297–1303.

Merrill, R.M. *et al*. (2019). Genetic dissection of assortative mating behavior. *PLoS biology*, 17(2).

Moran, B.M. *et al*. (2019). The evolution of microendemism in a reef fish (*Hypoplectrus maya*). *Molecular Ecology*, 28(11):2872–2885.

Nei, M. (1987). *Molecular evolutionary genetics*. Columbia University Press, New York.

Paradis, E. and Schliep, K. (2018). ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics*, 35:526–528.

Pedersen, T.L. (2019). *patchwork: The Composer of Plots*. https://patchwork.data-imaginist.com, https://github.com/thomasp85/patchwork.

Poelstra, J.W. *et al*. (2015). Transcriptomics of colour patterning and coloration shifts in crows. *Mol. Ecol.*, 24(18):4617–4628.

Puebla, O., Bermingham, E. and Guichard, F. (2012). Pairing dynamics and the origin of species. *Proc. R. Soc. B-Biol. Sci.*, 279(1731):1085–1092.

Puebla, O. *et al*. (2007). Colour pattern as a single trait driving speciation in *Hypoplectrus* coral reef fishes? *Proc. R. Soc. B-Biol. Sci.*, 274(1615):1265–1271.

Puebla, O. *et al*. (2018). Social-trap or mimicry? An empirical evaluation of the *Hypoplectrus unicolor* - *Chaetodon capistratus* association in Bocas del Toro, Panama. *Coral Reefs*, 37:1127–1137.

Purcell, S. *et al*. (2007). PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *Am. J. Hum. Genet.*, 81(3):559–575.

Randall, J.E. (1967). Food habits of reef fishes of the west indies. *Studies in Tropical Oceanography*.

Randall, J.E. (1968). *Caribbean reef fishes*. TFH Publications, Neptune City (NJ), USA.

Reed, R.D. *et al*. (2011). Optix drives the repeated convergent evolution of butterfly wing pattern mimicry. *Science*, 333(6046):1137–1141.

Renaut, S. *et al*. (2013). Genomic islands of divergence are not affected by geography of speciation in sunflowers. *Nature communications*, 4(1):1–8.

Revell, L.J. (2012). phytools: An r package for phylogenetic comparative biology (and other things). *Methods in Ecology and Evolution*, 3:217–223.

Saenko, S.V., Marialva, M.S. and Beldade, P. (2011). Involvement of the conserved hox

gene antennapedia in the development and evolution of a novel trait. *EvoDevo*, 2(1):9.

Schiffels, S. (2014). msmc-tools: Tools and utilities for msmc and msmc2.

Schiffels, S. and Durbin, R. (2014). Inferring human population size and separation history from multiple genome sequences. *Nature Genetics*, 46(8):919.

Schluter, D. (2000). *The Ecology of Adaptive Radiation (Oxford Series in Ecology and Evolution)*. Oxford University Press.

Slowikowski, K. (2019). *ggrepel: Automatically Position Non-Overlapping Text Labels with 'ggplot2'*. R package version 0.8.1.

Stankowski, S. *et al*. (2019). Widespread selection and gene flow shape the genomic landscape during a radiation of monkeyflowers. *PLoS biology*, 17(7):e3000391.

Theodosiou, L., McMillan, W.O. and Puebla, O. (2016). Recombination in the eggs and sperm in a simultaneously hermaphroditic vertebrate. *Proc. R. Soc. B-Biol. Sci.*, 283(1844).

Thresher, R.E. (1978). Polymorphism, mimicry, and the evolution of the hamlets (*Hypoplectrus*, Serranidae). *Bull. Mar. Sci.*, 28(2):345–353.

Thummel, R. *et al*. (2004). Differences in expression pattern and function between zebrafish hoxc13 orthologs: recruitment of Hoxc13b into an early embryonic role. *Dev. Biol.*, 274(2):318–333.

Turner, T.L., Hahn, M.W. and Nuzhdin, S.V. (2005). Genomic islands of speciation in anopheles gambiae. *PLoS Biol.*, 3(9):e285.

Van Belleghem, S.M. *et al*. (2017). Complex modular architecture around a simple toolkit of wing pattern genes. *Nature ecology & evolution*, 1(3):0052.

Van der Auwera, G.A. *et al*. (2013). From fastQ data to high-confidence variant calls: The genome analysis toolkit best practices pipeline. *Current Protocols in Bioinformatics*, (SUPL.43).

Vaughan, D. and Dancho, M. (2019). *furrr: Apply Mapping Functions in Parallel using Futures*. R package version 0.1.0.9002.

Via, S. (2009). Natural selection in action during speciation. *Proceedings of the National Academy of Sciences*, 106(Supplement 1):9939–9946.

Weir, B.S. and Cockerham, C.C. (1984). Estimating f-statistics for the analysis of population-structure. *Evolution*, 38(6):1358–1370.

Whiteman, E., Côté, I. and Reynolds, J. (2007). Ecological differences between hamlet (hypoplectrus: Serranidae) colour morphs: between-morph variation in diet. *Journal of Fish Biology*, 71(1):235–244.

Wickham, H. *et al*. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, 4(43):1686.

Wu, C.I. (2001). The genic view of the process of speciation. *Journal of evolutionary biology*, 14(6):851–865.

Wu, C.I. and Ting, C.T. (2004). Genes and speciation. *Nature Reviews Genetics*, 5(2):114.

Yu, G. *et al*. (2018). Two methods for mapping and visualizing associated data on

phylogeny using ggtree. *Molecular Biology and Evolution*, 35:3041–3043.

Zhou, X. and Stephens, M. (2012). Genome-wide efficient mixed-model analysis for association studies. *Nature Genet.*, 44:821.

– 6 –
Synthesis and Perspective

6.1. Synthesis

Speciation in the Hamlets

The presented studies show that the transition from genetics to genomics has helped to improve our understanding of the process of speciation in the hamlets. This becomes clear when considering the studies directly preceding **manuscript 2**: Puebla *et al.* (2014) and Picq *et al.* (2016). Compared to these studies (both of which had a very similar sampling design), the spatial context provided by the hamlet reference genome and the increased resolution of whole genome resequencing (compared to RAD sequencing) allowed for much more nuanced conclusions. Based on the distribution of F_{ST} values from Puebla *et al.* (2014), it was expected that the genome wide baseline of differentiation would fall onto the low end of the differentiation continuum and that only a small fraction of the genome would display elevated differentiation levels. Still, the small number of very distinct F_{ST} peaks against a genome wide background of basically no differentiation discovered in **manuscript 2** was striking.

The most pronounced result that emerged from the presented studies was that speciation in hamlets appears to be largely driven by a very small number of genes which are directly linked to color and vision and thus can interfere with the assortative mating systems in hamlets. While **manuscript 2** showed that combinations of vision and pigmentation genes appear to be co-selected leading to linkage disequilibrium across chromosomes, **manuscript 4** found that these exact regions are also phylogenetically discordant with the genome wide background for

a large cross section of the hamlet radiation. The available whole genome data also enabled the detection of several high probability hybrids within our samples, confirming ongoing gene flow between species. This was assumed before (eg. Puebla *et al.* 2012), based on hybrid spawning observations and experimental crosses, but the detection of F1 and F2 hybrids indicates that hybrids are actually viable and gene flow between different species appears to be possible. The population genetic and phylogenetic results demonstrate that throughout the major share of the genome this gene flow is rampant, but it also points to those areas where gene flow seems to be restricted. These area harbour differentiated genes, which apparently act as barrier genes.

The young history of the hamlet radiation and the amount of diversification pose a paradox because of the slow pace of mutation. It thus seems necessary that diversity in the genus *Hypoplectus* is fueled by additional sources. Two mechanisms that both seem to contribute are *standing genetic variation* and *hybridization*. Both result in a large pan-hamlet gene pool except for a small number of genes that are either acting as barrier genes, or that are those segments of the gene pool where lineage sorting has advanced furthest. The results are most pronounced in **manuscript 4**, which shows that those population genetic statistics that are based on allele frequency (like F_{ST} or $G \times P$) show strong signals, while those that rely on changes in nucleotide diversity (like π, d_{XY} or phylogentic approaches like twisst) show either little variation between the different species or are discordant across the genome. Again, this reflects that the different hamlet species have only faced a very limited amount of time (if any) as independent evolutionary lineages. This hypothesis is futher supported by the results of analysis of the hamlets demographic history in **manuscript 3** and **manuscript 4**.

In terms of the ecological differences between the hamlet species (regarding the *adaptedness* of the radiation), **manuscript 1** and **manuscript 3** are both pointing to a potential habitat specificity within the hamlets (contrasting to the assumed ecological uniformity). Differences in hamlet communities across coral reef types, association of particular hamlet species with specific coral species like *Acropora cervicornis*, visual water conditions and local endemism indicate that the different hamlet species might actually not be ecologically uniform. Yet, this remains rather anecdotal and the question of the influence of natural selection within the diversification of *Hypoplectus* still begs for a more thorough and systematic effort.

In a nutshell, the forces that seem to drive hamlet diversification appear to be those that are linked to the hamlets color assortative mating system resulting in genomic signals that bear a strong signature of sexual selection.

The Hamlets in a Broader Context

As mentioned in the introduction, the hamlets represent a somewhat exotic marine genus forming a rare example of a marine species flock. Thus comparable systems within speciation genomics research are found rather outside of the ocean. Three systems that seem to be of particular relevance are the *Heliconius* butterflies (Van Belleghem *et al.*, 2017) and

the avian genera *Lonchura* (Faust Stryjewski and Sorenson, 2017) and *Sporophila* (Campagna *et al.*, 2017). All of these systems are young radiations of roughly comparable size (species numbers) and are characterized by a diversity in color patterns. Furthermore, assortative mating is of relevance in both *Heliconius* (Merrill *et al.*, 2019) and *Sporophila* (Repenning and Fontana, 2019).

The comparison to the bird systems is striking for their similarity in population genetic patterns as well as for the fact that a large fraction of the color genes involved in the diversification of these systems is shared with the hamlet radiation. Both *Lonchura* and *Sporophila* have extreme shallow genome wide background levels of differentiation with only a hand full of very pronounced F_{ST} peaks — a pattern that is remarkably similar to the hamlets. What is even more fascinating is that these peaks among others cover the pigmentation genes *kitl/ kitlg* in *Lonchura* (Faust Stryjewski and Sorenson, 2017) and *kitl/ kitlg* and *sox10* in *Sporophila* (Campagna *et al.*, 2017). The same genes were fond under the differentiation peaks in hamlets in **mauscript 2**, **3** and **4**. The emerging picture is that a similar set of pigmentation genes facilitates rapid phenotypic diversification across distant vertebrate groups through rapid modulation of color patterns.

The comparison to *Heliconius* butterflies is interesting because of the similarity in assortative mating, the coupling of vision and color pattern genes and the modular nature in which pattern gene interactions create phenotypic diversity (Van Belleghem *et al.*, 2017). As an example of how closely the perception and expression of color patterns are interlocked, the *optix* gene in *Heliconius* is know to control prominent elements of wing patterning (red band) while being in close genetic linkage with a major effect quantitative trait locus for mate preference (Merrill *et al.*, 2019). Similar evolutionary mechanisms could be at work in the hamlets, with regard to the genomic area around the gene *casz1*. In **manuscript 2** and **4** we found this gene to be closely associated with a bared phenotype in hamlets and in **manuscript 2** it is also expressed in the hamlet retina. The repurposing of genes in multiple selected traits (or the formation of super-genes by several linked genes under selection) might be a common mechanism to create *magic traits* that can effectively promote speciation (Thibert-Plante and Gavrilets, 2013). In both, the *Heliconius* butterflies and the hamlets, the assortative mating behavior might increase the effectiveness of such a color-based magic trait. Furthermore, both systems share a genomic pattern of few large effect loci which seem to create a modular system that can quickly give rise to a large diversity of phenotypes by shuffling the genetic basis of the individual color pattern elements. The redundancy of individual pattern elements across the hamlet (Figure 1.3) and the *Heliconius* radiation (e.g. Figure 6 in Van Belleghem *et al.* 2017) are a quite visual illustration of this modularity. Genetically, this modularity is recovered in the sharp $G \times P$ association peaks and the accompanying phylogenetic discordances seen in **manuscript 4**. Evolutionary, this modular system could be a conceivable method through which hybrid speciation can boost a radiation by quickly multiplying phenotypic diversity though a rearrangement of discrete phenotypic elements previously evolved in independent lineages (Mérot *et al.*, 2020).

All in all, the comparison of the hamlet radiation with *Heliconius*, *Lonchura* and *Sporophila* is a fascinating demonstration of repeatability in evolution. The comparison with the avian systems shows how individual genes can repeatedly impact phenotypic diversification throughout distant vertebrate lineages. In contrast, the comparison with *Heliconius* illustrates how a similar genomic architecture characterized by the coupling of multiple traits under selection and the arrangement of large effect loci underlying discrete traits within a modular system can facilitate rapid diversification within an evolutionary radiation across the tree of life — ironically involving reticulated and thus less tree-like phylogenetic relationships within these radiations.

6.2. Perspective

While genomic approaches have certainly helped to advance our understanding of the hamlet radiation, much remains to be discovered. In a way the findings of this thesis created the foundation to address new and exciting questions much rather than solving the hamlet speciation mystery.

First of all, the question about the *adaptiveness* of the radiation remains. All studies investigating ecological differences in hamlets (including **manuscript 1**) have so far produced only subtle, suggestive results while lacking enough clarity to really justify solid conclusions (Thresher 1978; Holt *et al.* 2008 vs. Whiteman *et al.* 2007; Aguilar-Perera 2003). More extensive and far reaching studies describing and comparing the ecology of the separate hamlet species are needed to clarify the degree of niche divergence across the radiation. Thus far, the ecological assessment has almost exclusively focused on a group of the most abundant and widely distributed hamlet species (*H. aberrans*, *H. chlorurus*, *H. gummigutta*, *H. indigo*, *H. nigricans*, *H. puella* and *H. unicolor*) and have been restricted to the ecological factors of depth distribution, turbidity and diet (Thresher, 1978; Aguilar-Perera, 2003; Whiteman *et al.*, 2007; Holt *et al.*, 2008; Bejarano and Appeldoorn, 2013). There are many more factors that could be adaptive — the most prominent example in hamlets being the potential for aggressive mimicry (which has also been assessed thoroughly in Randall and Randall 1960; Puebla *et al.* 2007, 2018; Picq *et al.* 2019). Yet, we need a more complete image of the ecology of hamlets including factors like preferences in micro-habitat and territoriality or predation upon hamlets. Most importantly we need an ecological assessment specifically of the rare and endemic hamlet species. Species that are limited to specific locations or habitats might be comparably independent in terms of gene flow and thus they harbor most pronounced niche divergence as well.

Further exploration of ecological divergence might be informed by genetic differentiation found at loci that conceivable to be under natural selection (eg. *tpm4*, discussed in Picq *et al.* 2016). Naturally, this also calls for an extension in sampling both across rare species and geographic ranges. Much of the recent genetic and genomic work in hamlets has focused on three species (*H. nigricans*, *H. puella* and *H. unicolor*) from the continental Caribbean coast. To investigate local adaptation, more remote areas of the Caribbean Antilles as well as from the Gulf of Mexico

need to be sequenced. Previous work based on mtDNA indicates that especially comparisons including the Gulf of Mexico might provide new insights, as hamlets from that region could cover a more advanced stretch of the speciation continuum (Ramon *et al.*, 2003). Such an expansion of sampling effort would also provide important context for the findings of this thesis, since the results of **manuscript 2** indicate that divergence in hamlets is likely both influenced by speciation and local adaptation, as differentiation also accumulates between different populations of the same species.

Another interesting direction for further research is a more structured approach to unravel the connection of hamlet pheno- and genotypes. There are up-and-coming approaches to quantify both coloration and patterning in a more systematic and objective manner (eg. Akkaynak *et al.* 2014; Van Belleghem *et al.* 2017). The $G \times P$ results of **manuscript 4** are already promising, but if whole genome sequencing was combined with a more advanced way of phenotyping than a simple binary scoring, more genetic components of the hamlets diversification toolkit might be exposed. This would also help to investigate the inner-specific variation in phenotypes (Suppl. Fig. 5.9) and maybe uncover cryptic species boundaries that are currently obscured by poor phenotyping. As an example of this, based on the results of **manuscript 2**, it is likely that the individual populations of *H. nigricans* might actually be quite differentiated — a finding that corresponds well with subtle differences in phenotype across different *H. nigricans* populations reported in **manuscript 1**.

Also, it should be acknowledged that to this point the genomic findings are rather descriptive. To ultimately confirm the connection of the differentiated genotypes and their phenotypic effect, the genomic findings should be functionally validated. This could be done in a comparable way to Lin *et al.* (2016), who used a CRISPR–Cas9 knockout approach to demonstrate the effect of the *tbx4* gene loss, which triggered the characteristic loss of the pelvic fin in seahorses. Currently, such an approach would be difficult in hamlets though, as the breeding of hamlets in captivity and especially the rearing until the adult stage (which is when the species specific phenotyes form) remains a major hurdle (Domeier, 1994). Yet, if this hurdle could be overcome, the entire hybridization complex of *Hypoplectrus* could be dissected as well. This could include the inheritance of phenotypes, dominance effects and fertility of hybrids, which could be assessed much more thoroughly using a hamlet rearing program coupled with genomic approaches. Another remaining technical obstacle for functional validation of the population genomic findings is the current state of the annotation of the hamlet reference genome. At this stage, the genome annotation still mainly resembles the output of a highly automatized annotation pipeline. While this is sufficient for broad statements about the genes behind differentiation peaks, functional validation asks for a more precise representation of the underlying genes. Thus, much manual work in curating the hamlet genome annotation would be needed before any functional validation is imaginable (Yandell and Ence, 2012).

Finally, the exceptional position of the hamlets as a rare marine species flock should be considered (Bowen *et al.*, 2020). This somewhat exotic setting limits the extent to which speci-

ation in the hamlets can stand as a representative for marine speciation in general. Other marine evolutionary genomic model species are currently still rare and dominated by temperate and cold water fishes like the Antarctic notothenioid ice fishes (another marine species flock, Ceballos *et al.* 2019), the Atlantic cod (*Gadus morhua*, Kess *et al.* 2020) or the European seabass (*Dicentrarchus labrax*, Duranton *et al.* 2018), while tropic model systems are scarce. Hence, to put the population genomic findings from this thesis into perspective and to get a more general understanding of marine speciation, other tropical, non-radiating marine speciation systems are needed. Only in the light of this context can be determined, which signatures of the speciation witnessed in hamlets are a marine phenomenon, which ones are the signature of the radiation within species flocks and which elements are rather typical of speciation driven by sexual selection and assortative mating. At the current state, comparisons across the available marine model systems vary over too many factors to produce distinct conclusions.

6.3. Concluding Remarks

The work presented within this thesis has refined our understanding of the hamlet radiation. It has shown in a conceivable way how the hamlets' mating system is linked to genomic constraints in inter-specific gene flow within the genus *Hypoplectrus*. In uncovering this connection, parallels to other model systems emerged, highlighting that genomic modularity can quickly generate phenotypic diversity by shuffling of discrete phenotypic traits. We also learned that the suite of genes underlying phenotypic diversity has a large overlap among distant vertebrate systems. This work lays the foundation for further research on the connection of the genotype and the phenotype in a reef fish genus which is characterized by a stunning diversity of color patterns — a characteristic that is shared among many reef fishes, that is iconic for biodiversity on coral reef systems and that also marks one of the major reasons these fishes are regarded as interesting and fascinating by researchers (and many others) around the world.

Yet, it is clear that transition of speciation research in hamlets towards genomics can serve only as an entry point for the study of marine speciation by providing a first exemplary case study of the genomic signature of a young marine species flock. In future research, this can serve as a reference or contrast for other marine non-radiating speciation processes. As population genomic approaches mature and become more commonplace in other marine systems, we can expect to uncover more thoroughly the genomic mechanisms of speciation within the largest and oldest habitat on earth. Many fundamental conditions influencing evolutionary dynamics between marine and terrestrial systems differ — maybe the peculiar position of the hamlets as marine species flock might then also serve as a bridge, given its similarities with terrestrial systems which have been uncovered.

Synthesis References

Aguilar-Perera, A. (2003). Abundance and distribution of hamlets (Teleostei : *Hypoplectrus*) in coral reefs off southwestern Puerto Rico: Support for the multiple-

species hypothesis. *Caribbean Journal of Science*, 39(1):147–151.

Akkaynak, D. *et al*. (2014). Use of commercial off-the-shelf digital cameras for scientific data acquisition and scene-specific color calibration. *Journal of the Optical Society of America A*, 31(2):312–321.

Bejarano, I. and Appeldoorn, R.S. (2013). Seawater turbidity and fish communities on coral reefs of Puerto Rico. *Marine Ecology Progress Series*, 474:217–226.

Bowen, B.W. *et al*. (2020). Species Radiations in the Sea: What the Flock? *Journal of Heredity*, 111(1):70–83.

Campagna, L. *et al*. (2017). Repeated divergent selection on pigmentation genes in a rapid finch radiation. *Science Advances*, 3(5):e1602404.

Ceballos, S.G. *et al*. (2019). Phylogenomics of an extra-Antarctic notothenioid radiation reveals a previously unrecognized lineage and diffuse species boundaries. *BMC Evolutionary Biology*, 19(1):13.

Domeier, M.L. (1994). Speciation in the serranid fish *Hypoplectrus*. *Bulletin of Marine Science*, 54(1):103–141.

Duranton, M. *et al*. (2018). The origin and remolding of genomic islands of differentiation in the European sea bass. *Nature Communications*, 9(1):2518.

Faust Stryjewski, K. and Sorenson, M.D. (2017). Mosaic genome evolution in a recent and rapid avian radiation. *Nature Ecology & Evolution*, 1(12):1912–1922.

Holt, B.G. *et al*. (2008). Stable isotope analysis of the *Hypoplectrus* species complex reveals no evidence for dietary niche divergence. *Marine Ecology Progress Series*, 357:283–289.

Kess, T. *et al*. (2020). Modular chromosome rearrangements reveal parallel and nonparallel adaptation in a marine fish. *Ecology and Evolution*, 10(2):638–653.

Lin, Q. *et al*. (2016). The seahorse genome and the evolution of its specialized morphology. *Nature*, 540:395.

Mérot, C. *et al*. (2020). Hybridization and transgressive exploration of colour pattern and wing morphology in *Heliconius* butterflies. *Journal of Evolutionary Biology*, n/a(n/a).

Merrill, R.M. *et al*. (2019). Genetic dissection of assortative mating behavior. 17(2):1–21.

Picq, S., McMillan, W.O. and Puebla, O. (2016). Population genomics of local adaptation versus speciation in coral reef fishes (*Hypoplectrus* spp, Serranidae). *Ecology and Evolution*, 6(7):2109–2124.

Picq, S., Scotti, M. and Puebla, O. (2019). Behavioural syndromes as a link between ecology and mate choice: a field study in a reef fish population. *Animal Behaviour*, 150:219–237.

Puebla, O., Bermingham, E. and Guichard, F. (2012). Pairing dynamics and the origin of species. *Proceedings of the Royal Society B: Biological Sciences*, 279(1731):1085–1092.

Puebla, O. *et al*. (2007). Colour pattern as a single trait driving speciation in *Hypoplectrus* coral reef fishes? *Proceedings of the Royal Society B: Biological Sciences*, 274(1615):1265–1271.

Puebla, O., Bermingham, E. and McMillan, W.O. (2014). Genomic atolls of differentiation in coral reef fishes (*Hypoplectrus* spp., Serranidae). *Molecular Ecology*, 23(21):5291–5303.

Puebla, O. *et al*. (2018). Social-trap or mimicry? An empirical evaluation of the *Hypoplectrus unicolor – Chaetodon capistratus* association in Bocas del Toro, Panama. *Coral Reefs*, 37(4):1127–1137.

Ramon, M.L., Lobel, P.S. and Sorenson, M.D. (2003). Lack of mitochondrial genetic structure in hamlets (*Hypoplectrus* spp.): recent speciation or ongoing hybridization? *Molecular Ecology*, 12(11):2975–2980.

Randall, J.E. and Randall, H.A. (1960). Examples of mimicry and protective resemblance in tropical marine fishes. *Bulletin of Marine Science*, 10(4):444–480.

Repenning, M. and Fontana, C.S. (2019). Distinguishing females of capuchino seedeaters: call repertoires provide evidence for species-level diagnosis. *Ornithology Research*, 27(2):70–78.

Thibert-Plante, X. and Gavrilets, S. (2013). Evolution of mate choice and the so-called magic traits in ecological speciation. *Ecology Letters*, 16(8):1004–1013.

Thresher, R.E. (1978). Polymorphism, mimicry, and the evolution of the hamlets (*Hypoplectrus*, Serranidae). *Bulletin of Marine Science*, 28(2):345–353.

Van Belleghem, S.M. *et al*. (2017). Complex modular architecture around a simple toolkit of wing pattern genes. *Nature Ecology & Evolution*, 1:52.

Whiteman, E.A., Côté, I.M. and Reynolds, J.D. (2007). Ecological differences between hamlet (*Hypoplectrus*: Serranidae) colour morphs: Between-morph variation in diet. *Journal of Fish Biology*, 71(1):235–244.

Yandell, M. and Ence, D. (2012). A beginner's guide to eukaryotic genome annotation. *Nature Reviews Genetics*, 13(5):329–342.

Appendix

List of Figures

1.1 Models of speciation 8
1.2 Overview over all currently described hamlet species 12
1.3 Distribution of the genus *Hypoplectrus* 14
1.4 Overview over all currently described hamlet species 15
1.5 Phylogeny based on cytochrome b 16

2.1 Sampling location and hamlet relative abundances 26
2.2 Non-metric multidimensional scaling (NMDS) 29
2.3 Comparison of the temporal change in the diversity of the first order for inner and outer reefs 30
2.4 Comparison of the temporal change in the diversity of the first order for inner and outer reefs 30

3.1 Sampling design for resequencing 41
3.2 Patterns of genomic differentiation among black (*H. nigricans*), barred (*H. puella*) and butter (*H. unicolor*) hamlets 45
3.3 Close-up on the four major F_{ST} peaks 48
3.4 Long-distance and inter-chromosomal LD among the four major F_{ST} peak regions. 55

4.1 Ranges of the *Hypoplectrus* species considered 89
4.2 Global Divergence 98
4.3 MSMC Demographic History 100
4.4 MSMC Cross-coalescence 105
4.5 Whole-genome $LDNe$ 106

5.1 Genome-wide differentiation (F_{ST}) and divergence (d_{XY}) among pairs of sympatric species. 125
5.2 Progression of genetic differentiation (F_{ST}) across the 28 species pairs. 127
5.3 MSMC and cross coalescence rate. 134
5.4 Close-up on three genomic regions of interest. 136

List of Supplementary Figures

2.1 NMDS outer reefs 29
2.2 NMDS outer reefs 32

3.1 Broad-scale synteny between the hamlet and stickleback genomes 43
3.2 Identification of putative backcrosses and hybrids 44
3.3 Genome wide $G \times P$ association 46
3.4 Large low-recombining region on linkage group 8 (LG08) 47
3.5 Extent of differentiation under various scenarios 49
3.6 Gene expression in the retinal tissue 50
3.7 Genomic intervals above the 99.90^{th} F_{ST} percentile 51
3.8 Close-up on all the intervals above the 99.90^{th} F_{ST} percentile 52
3.8 (continued I) Close-up on all the intervals above the 99.90^{th} F_{ST} percentile . . . 53
3.8 (continued II) Close-up on all the intervals above the 99.90^{th} F_{ST} percentile . . 54
3.9 Close-up on the four additional intervals containing candidate vision and pigmentation genes 56
3.10 Genome-wide recombination patterns 57
3.11 Long-distance and inter-chromosomal linkage disequilibrium (LD) among the eight intervals containing candidate vision and pigmentation genes 58
3.12 Patterns of genomic differentiation among hamlets from Belize, Honduras and Panama 59
3.13 Long-distance and inter-chromosomal linkage disequilibrium (LD) among the four candidate intervals for each species pair 60
3.14 Long-distance and inter-chromosomal linkage disequilibrium (LD) among the eight intervals containing vision and pigmentation candidate genes 61
3.15 PCA based on filtered dataset. 66
3.16 Comparison of linear model and linear mixed model results of the genotype by phenotype ($G \times P$) association. 67

4.1 Belize hamlet survey and sampling locations 92
4.2 Florida hamlet survey and sampling locations 93
4.3 Florida hamlet densities from belt transects and FKRVC surveys 95
4.4 NMDS of Florida hamlet community composition 96
4.5 Decay in linkage disequilibrium with physical distance 97
4.6 PCA of genome-wide SNP data in Belizean hamlets 99
4.7 Sliding window F_{ST} between *H. maya* and other hamlets 101

4.8 Nucleotide diversity by hamlet species . . . 102
4.9 MLE and A_{jk} individual pairwise relatedness estimates . . . 102
4.10 MSMC results including all time segments . . . 103
4.11 MSMC without masking F_{ST} peaks . . . 103
4.12 SMC++ estimates with and without F_{ST} peak masking . . . 105

5.1 Genome wide average F_{ST} across radiations . . . 125
5.2 Genome wide pair wise F_{ST} . . . 126
5.3 F_{ST} vs. ρ . . . 128
5.4 Whole-genome patterns . . . 129
5.5 Genome wide pair wise d_{XY} . . . 131
5.6 Genome wide π . . . 132
5.7 Genetic diversity (π) vs. population recombination rate (ρ). . . . 133
5.8 Whole genome phylogeny . . . 135
5.9 Phenotype overview . . . 137

List of Tables

2.1 Hamlet counts . . . 28

4.1 Global estimates of the mean weighted F_{ST} . . . 97

List of Supplementary Tables

3.1 Genomic regions above the 99.90^{th} F_{ST} percentile . . . 71
3.2 Software versions used in this study . . . 72
3a Sample list for re-sequencing . . . 73
3b Sample list for re-sequencing . . . 74
3c Sample list for re-sequencing . . . 75
3.4 Sequencing data generated for the assembly . . . 75
3.5 Sample list for RNA sequencing . . . 76
3.6 Global estimates of the weighted mean F_{ST} . . . 76

4.1 Southern MBRS endemic fish species . . . 112
4.2 Sample list for re-sequencing data . . . 113

4.3 Genes in F_{ST} outlier windows 113
4.4 Sample IDs and coverage from MSMC runs 114
4.5 Software versions used in this study 115

5.1 Global pairwise F_{ST} and d_{XY} 143
5.2 Genes within F_{ST} outlier windows. 144
5.3 Sample list 145

Declaration of Author Contributions

Manuscript 1:
Conceptualization, O.P. and **K.H.**; Methodology, **K.H.** and O.P.; Investigation, **K.H.**, W.O.M. and O.P; Software, **K.H.**; Data Curation, **K.H.**; Visualization, **K.H.**; Writing – Original Draft, O.P. and **K.H.**; Writing – Review & Editing, O.P. and **K.H.**; Funding Acquisition, O.P. and W.O.M.; Resources, O.P. and R.B.; Supervision, O.P.

Manuscript 2:
O.P., W.O.M. and **K.H.** designed the study and participated in the sampling. **K.H.**, M.V. and O.P. contributed to the DNA extractions. M.V. prepared part of the DNA and all of the RNA libraries. **K.H.** assembled the hamlet genome and conducted the data analyses, in collaboration with O.P. and W.O.M. M.P.H. annotated the genome and contributed to the sequencing. **All authors** contributed to the writing, with major input from **K.H.**, O.P. and W.O.M.

Manuscript 3:
B.M. conceived of the study, conducted field work and data analyses, and wrote the manuscript. O.P. conceived of the study, conducted field work, and contributed to the manuscript. R.W.S. contributed to data analyses and the manuscript. **K.H.** contributed to the data analyses. W.O.M. and M.H. contributed to genome sequencing. C.B. contributed to the curation of specimens. **All co-authors** provided feedback on the manuscript.

Manuscript 4:
Conceptualization, **K.H.**, O.P. and W.O.M.; Methodology, **K.H.** and O.P.; Investigation, **K.H.** and O.P; Software, **K.H.**; Data Curation, **K.H.**; Visualization, **K.H.**; Writing – Original Draft, **K.H.** and O.P.; Writing – Review & Editing, **K.H.** and O.P.; Funding Acquisition, O.P. and W.O.M.; Resources, O.P.; Supervision, O.P.

Curriculum Vitae

Name:	Kosmas Benjamin Hench	**Nationality:**	German
Birthday:	10.09.1988	**Orcid-id**	0000-0003-1119-187X
Place of birth:	Erlenbach am Main		

Education in Science

since 2017/03/01	Doctoral candidate at the Puebla Lab GEOMAR Helmholtz Centre for Ocean Research Kiel / Leibniz Centre for Tropical Marine Research (ZMT)
2014/10/01 - 2017/03/31	Master student (M.Sc. Biological Oceanography) GEOMAR Helmholtz Centre for Ocean Research Kiel
2010/10/01 - 2014/09/30	Bachelor student (B.Sc. Environmental Science) Carl von Ossietzky Universität Oldenburg
2008/08/15 - 2009/02/25	Bachelor student (B.A. Educational Science) Phillips Universität Marburg

Practical Experience in Science/ Teaching

2020/ 01-02	Teaching assistant at Three Seas Fishes class Bocas del Toro, Panama. *Northeastern University*
2019/ 02	Teaching assistant at Three Seas Fishes class Bocas del Toro, Panama. *Northeastern University*
2018/ 01-02	Teaching assistant at Three Seas Fishes class Bocas del Toro, Panama. *Northeastern University*
2017/ 03	Scientist/ Diver during Hamlet fieldwork (phenotyping and transects) La Parguera, Puerto Rico. *GEOMAR*
2017/ 01-02	Teaching assistant at Three Seas Fishes class Bocas del Toro, Panama. *Northeastern University*
2015/ 09-10	Research assistant on research cruise M119 (zooplankton sampling) from Mindelo/ Cape Verde to Recife/ Brasil. *RV Meteor, GEOMAR*

2015/ 05	Participant at the KOSMOS 2015 expedition (mesocosm experiment on ocean accidification) Bergen, Norway, *GEOMAR*
2014 / 08-09	Research assistant / scientific diver for the expedition KOL 19 (Succession of benthic communities in polar environments) Ny Ålesund, Svalbard. *Alfred Wegener Institut*
2014 / 07	Participant at the SPACES/OASIS SO234-2 scientific research and training cruise on the RV Sonne (Air-Sea Interactions in the western Indian Ocean) from Durban/ South Africa to Port Louis/ Mauritius. *RV Sonne, Uni Oslo/ GEOMAR*
2014 / 05-06	Intern at the research group Bentho-Pelagic Processes Bremerhaven. *Alfred Wegener Institut*
2012 - 2013	Research assistant at the COMTESS-projekt concerning sustainable coastal land management Oldenburg. *Carl von Ossietzky Universität Oldenburg*
2012	Intern at the Fisheries Ecosystems Advisory Services, Galway, Irland. *Marine Institute, Ireland*

Additional Scientific Training

2019/03/25-29	Workshop: Landscape genetic data analysis using R PR Statistics, Glasgow, Scotland
2013/10/02	Qualification as German scientific diver *ICBM, Carl von Ossietzky Universität Oldenburg*, Oldenburg, Germany

Acknowledgments

Now, at the end of my thesis, there is a long list of people that I owe much to and who all, in their ways, have supported me along the way. I want to thank all those friends who helped me succeeding on this journey:

Of course, first of all **Oscar Puebla** — what a ride those last five years where. Thanks for everything, from making this thesis possible in the first place, to the constant support, constructive feedback and discussions. I could not have hoped for a better supervisor, so thanks! The only thing I regret is not getting those walkie-talkies — they would have helped when out of sight (particularely under water...)

Owen McMillan, although rather remote I want to thank you for your commitment in pushing our science, for the warm welcomes in Panama and also for the excellent cooking in the field.

All co-authors: **Carole Baldwin**, **Ricardo Betancur-R.**, **MarcHöppner**, **Benjamin Moran**, **Marta Vargas** and **Robin Waples**, for the productive discussions, the insightful inputs and constantly pleasant collaborations.

All the *GEOMAR* office mates that have provided much needed input for the success of this thesis in the form excellent coffee and excellent company. I glad I could spend so much of my PhD life sourounded by **Sophie Picq**, **Melanie Heckwolf**, **Véronique Merten**, **Felix Mittermayer**, **Agnes Piecyk**, **Henry Göhlich** and **Jamie Parker**.

The whole **EV section** at *GEOMAR* who helped me to launch this expedition, particularly **Till Bayer** who helped me learning to navigate the wondeful realm of bioinformatics (I you had not introduced me to $< tab > -complete$, I would likely still be working on this thesis for another few years....). But also thanks to **Svend Mees**, **Cornelia Rüther**, **Diana Gill** and **Thorsten Reusch** for dealing with all the day-to-day issues and keeping the science running.

All members of the **Ecology Group** and every one else at *ZMT* who made sure that I could also finish this expedition. Thanks to **Martin Zimmer** for providing an new home and thanks to **Thomas Mann** and **Elisa Casella** to welcome me there. Also, I want to thank **Steve Doo** and **Florian Hierl** who convinced be that excellent coffee and company is also to be found in Bremen.

Special thanks to the members of the *Three Seas Programme*, particularly **Liz Bentley Magee**, **Andrea Jerabek**, **Kelsey Tuminelli**, **Erin Sayre** who made sure that I could actually visit the hamlets live and to whom I owe four incedible trips to Panama. Also, thanks **Jeanne Bloomberg**, **Michelle Chen**, **Grace McKenna Marcus Drymon** who made these trips even more fun and thanks to all awesome students from *Three Seas* and the nice people from *STRIs* Research Station at Bocas.

To all the friends who have joined the *Puebla Lab* over the years: **Floriane Coulmance**, **Ben-**

jamin Moran and **Cammeron Walsh** — thanks for all the discussions and input on this collaborative quest of making sense of the coolest fishes on the reef.

To all the friends who reminded me that there might also be important and interesting aspects of life besides hamlets. **Lara Schmittmann**, **Serra Örey**, **Verena Kalter**, **Moritz Ehrlich**, **Bastian Kimmel**, **Felix Milke**, **Jan Greiwe**, **Sebastian Storey**, **Mara Heinrichs**, **Maurits Halbach**, **Jana Nau** and **Sebastian Spatz** — all of you have made my life more balanced and rich, everyone in their own way.

To my Family: **Ulli** & **Bernhard**, **Joschka** & **Kati**, **Hanna** & **Fritz**, **Lotta** and all the members of the **Hench-Clan**, thanks for all the support, bustle, the good bad jokes and for keeping me grounded.

Finally, **Nora** for holding on — I know this took forever...

Eidesstattliche Erklärung

Hiermit bestätige ich, Kosmas Hench, dass die folgende Dissertation

Genomics and the Origin of Marine Species

von mir, unter Beratung meines Betreuers, selbstständig verfasst wurde, nach Inhalt und Form meine eigene Arbeit ist und keine weiteren Quellen und Hilfsmittel als die angegebenen verwendet wurden.

Die vorliegende Arbeit ist unter Einhaltung der Regeln guter wissenschaftlicher Praxis der Deutschen Forschungsgemeinschaft entstanden und wurde nicht im Rahmen eines Prüfungsverfahrens an anderer Stelle vorgelegt. Veröffentlichte oder zur Veröffentlichung eingereichte Manuskripte wurden kenntlich gemacht. Mir wurde kein akadamischer Grad entzogen

Bremen, 12.06.2020

Kosmas Hench