Du supergène à l'écosystème à travers le prisme de la génomique : différentes nuances de diversité génétique chez des organismes marins

Pierre Lesturgie

MUSEUM NATIONAL D'HISTOIRE NATURELLE

Ecole Doctorale Sciences de la nature et de l'Homme – ED 227

Année 2022-2023 **N°attribué par la bibliothèque**

|_|_|_|_|_|_|_|_|_|_|_|_|

THESE

Pour obtenir le grade de

DOCTEUR DU MUSEUM NATIONAL D'HISTOIRE NATURELLE

Spécialité : Biologie Évolutive

Présentée et soutenue publiquement par

Pierre Lesturgie

Le 13/12/2023

From (super)genes to ecosystems through the lens of genomics: different shades of genetic diversity in marine organisms

---

Sous la direction de **: Monsieur Mona, Stefano, Directeur d'Étude (EPHE)** et **Madame Lopez-Villavicencio Manuela, Maitre de Conférence (MNHN)**

JURY

Mme. Giraud, Tatiana	Directrice de Recherche, CNRS	**Présidente**
M. Mona, Stefano	Directeur d'Etude, EPHE	**Directeur de Thèse**
Mme. Lopez-Villavicencio, Manuela	Maitre de Conférence, MNHN	**Directrice de Thèse**
Mme. Tenaillon, Maud	Directrice de Recherche, CNRS	**Rapporteure**
M. Boitard, Simon	Chargé de Recherche, INRAE	**Rapporteur**
Mme. Arnaud-Haond, Sophie	Chercheur IFREMER	**Examinatrice**
M. Naylor, Gavin	Full Professor, University of Florida	**Examinateur**

Acknowledgements

First of all, I would like to thank Stefano, who trusted me to carry out this PhD, which would have been very different without his sincere care and support. I'm grateful for everything he taught me in population genetics (but not only!) and for his inspiring vision of science. I am very appreciative that he always expected the best from me and was always willing to help me achieve my goals. I'm also grateful to him for always ensuring that I thrived not only in research, where he gave me great freedom in carrying out my projects, but also and above all in life.

I would also like to thank Manuela, for giving me the opportunity to do this PhD, but also for her support, advice and for the nice and broad discussions during coffee breaks. I'm very grateful to Gavin, for his profound kindness, for teaching me so much about sharks and rays, and for his warm welcome to Gainesville, where I had the chance to discover many new things. I'm also thankful to Valeriano for his valuable insights in ecology and for his encouragements.

I would like to express my sincere thanks to Maud Tenaillon and Simon Boitard for agreeing to be my thesis referees, and to Sophie Arnaud-Haond and Tatiana Giraud for agreeing to be part of my defense jury. I am also very grateful to Pierre-Alexandre Gagnaire, Raphaël Leblois and Violaine Llaurens for their invaluable advice and support during my thesis committees.

Many thanks to everyone at the lab who helped make this journey so enjoyable. In particular, thanks to Hubert, who was there from the start, for his generosity and for all the things we've shared over the years. I'm also grateful to Elise and Romuald, for their invaluable sharing and help, and for the great discussions we've had over (always a few) beers… I'm also thankful to Pascaline, Alba, Nisha, as well as the former trainees who all brought me a lot, in one way or another, over the last few years.

Finally, I would like to deeply thank my friends, parents, sisters, Pepito and Zoé for the unconditional support they have always given me.

Abstract

The evolutionary history of species is shaped by demographic and selective processes, the understanding of which requires complex modeling of genetic diversity in order to grasp insights into phenomena ranging from the local to the ecosystem-scale. Yet this is a challenging exercise in population genetics, as it requires both a good understanding of processes shaping genetic diversity and to investigating meaningful demographic scenarios using carefully designed frameworks tailored to the question and the biology of the organism(s) under study. In this context, the aim of my PhD thesis is twofold. Firstly, it aims to show how the extensive reconstruction of demographic processes provides valuable insights for developing evolutionary hypotheses and conservation strategies, and notably to characterize the interplay between neutral and selective processes. Secondly, it seeks to improve our understanding of how species- and community-level processes influence historical demography.

To that end, I first investigated how population structure (and, more generally, any historical event) influences the distribution of coalescence times and therefore the demographic reconstruction inferred through coalescent-based models assuming random mating (*unstructured* models). To do this, I coupled theoretical insights to empirical test-cases based on widely distributed shark species. Ultimately, I showed how *unstructured* models are extremely useful in inferring the variation of the coalescence rate through time, which is directly linked to the true demography of a species, hence remaining a fundamental exploratory tool to gain insights into species' history, if interpreted under the light of complex scenario rather than panmictic ones.

Secondly, I report the discovery of a size-determining supergene in the Thorny Skate (*Amblyraja radiata*). I then provided insights into its origin and role in the steep decline trajectory of a vulnerable population by extensively reconstructing the demographic history of the species at the scale of its range distribution. This emphasized how crucial demographic modelling is to understand local selective processes, especially when coupled with conservation implications.

Finally, I investigated ecological determinants of genetic diversity using a unique panel of genomic data from 43 species of coral reef fishes. This study allowed to demonstrate that trophic niche width is positively associated to demographic stability, revealing the direct effect of a community-level process on the historical demography of species. While population genetics studies are usually species-centered, this work is one of the first to actively try to evaluate the influence of species interactions over their evolutionary history. Ultimately, it provides insights into how multi-

species population genetics datasets will be key to elucidating the genomic signatures of large-scale processes in the future.

Overall, my thesis highlights the fundamental role of robust demographic reconstruction to answer questions related to both micro (such as adaptation) and macro (such as the ecosystem functioning) evolutionary processes, through case studies of marine species. Notably, it increases our understanding of the evolutionary and ecological underpinnings of genetic diversity and how they influence the coalescent history of a sample of lineages (and hence, the demographic inferences we made out of them). Finally, it highlights the significance and potential power of multi-species studies, that are quite novel in population genetics, and which in the future will make it possible to answer questions at different scales of study with evolutionary, ecological, and conservation implications.

Keywords: Population Genetics, Coalescence, Demographic Modelling, Supergenes, Meta-populations, Genomics.

Résumé

L'histoire évolutive des espèces est façonnée par des processus démographiques et sélectifs, dont la compréhension nécessite la modélisation complexe de la diversité génétique afin de comprendre des phénomènes allant de l'échelle locale à l'échelle de l'écosystème. Il s'agit cependant d'un exercice difficile en génétique des populations, car il nécessite une bonne compréhension des processus façonnant la diversité génétique et l'étude de scénarios complexes et utiles au moyen de cadres inférentiels soigneusement conçus et adaptés à la question et à la biologie du modèle d'étude. Dans ce cadre, ma thèse vise à montrer comment la reconstruction détaillée des processus démographiques fournit des informations précieuses pour élaborer des hypothèses évolutives et des stratégies de conservation, et notamment pour mieux caractériser l'interaction entre processus neutres et sélectifs. Également, elle cherche à améliorer notre compréhension de la façon dont les processus, au niveau des espèces et des communautés, influencent la démographie historique.

Pour cela, j'ai d'abord étudié comment la structuration génétique des populations (et plus généralement, tout événement historique) influence les patrons démographiques inférés par des modèles basés sur la théorie de la coalescence supposant un accouplement aléatoire (modèles *non structurés*). En couplant arguments théoriques à des cas empiriques basés sur des requins à large distribution, j'ai pu montrer comment les modèles *non structurés* sont utiles pour inférer la variation du taux de coalescence dans le temps, qui est directement liée à la vraie démographie de l'espèce. Ceci a permis de mettre en avant que ces modèles restent un outil exploratoire fondamental pour recueillir des éléments sur l'histoire évolutive des espèces, à condition qu'ils soient interprétés à la lumière de scénarios complexes plutôt que panmictiques.

Ensuite, je rapporte la découverte d'un supergène déterminant la taille chez une espèce de raie. Je fournis alors des évidences sur son origine et son rôle dans le déclin abrupt d'une population vulnérable grâce à la reconstruction de l'histoire démographique de l'espèce à l'échelle de son aire de distribution. Cette étude souligne l'importance de la modélisation démographique pour comprendre des processus locaux de sélection, en particulier lorsqu'ils impliquent des enjeux de conservation.

Enfin, j'ai examiné certains déterminants écologiques de la diversité génétique à travers un panel unique de données génomiques provenant de 43 espèces de poissons récifaux. Ceci a permis de montrer une relation positive entre largeur de niche trophique et stabilité démographique, révélant l'effet d'un processus à l'échelle de la communauté sur l'histoire démographique. Les études en

génétique des populations s'articulant en général sur une espèce, ce travail est l'un des premiers à tenter d'évaluer directement la relation entre la diversité des interactions des espèces et leur histoire démographique. Plus généralement, cette étude suggère que les jeux de données multi-espèces pourraient se révéler importants à l'avenir pour détecter les signatures génomiques laissées par des processus à grande échelle.

Dans l'ensemble, ma thèse souligne le rôle fondamental d'une reconstruction démographique robuste pour comprendre des processus micro (tels que l'adaptation) et macro (tels que le fonctionnement des écosystèmes) évolutifs à travers l'étude d'espèces marines. Elle permet aussi de mieux comprendre certains déterminants évolutifs et écologiques de la diversité génétique et la manière dont ils influencent les processus de coalescence (et donc les inférences démographiques en découlant). Enfin, elle souligne l'importance et la puissance potentielle des études multi-espèces, relativement nouvelles en génétique des populations, qui permettront à l'avenir de répondre à des questions à des échelles d'étude différentes avec des implications en évolution, en écologie et en conservation.

Mots-Clés : Génétique des Populations, Coalescence, Modélisation Démographique, Supergènes, Meta-populations, Génomique.

Table of Contents

List of figures

List of tables

Chapter 1. Introduction

1.1. The Complex Evolutionary Processes Shaping Populations and Species

1.1.1. From gene to ecosystem processes

The evolutionary history of species is shaped by different processes (demographic and selective) that can be tackled by modelling genetic diversity at various scales of study. At a local scale, a population is a collection of individuals that share a common genetic pool and have the ability to exchange genetic material during reproduction. Despite this shared genetic background, individuals can locally display strong phenotypic differences, whose maintenance is often driven by selective forces (Hedrick, 2007; Llaurens et al., 2017; Marchinko et al., 2014; Wittmann et al., 2017) and whose determinism can be explained by genetic mechanisms such as single (Abbott & Fairbanks, 2016) or multi-locus polymorphisms (Bouwman et al., 2018; Boyle et al., 2017; Fisher, 1918; Mather, 1941; Wood et al., 2014; Zimmerman et al., 2000), supergenes (Schwander et al., 2014; Thompson & Jiggins, 2014; Wellenreuther & Bernatchez, 2018), or even complex gene-environment interplays (Rutter et al., 2006). Uncovering the origin, determinism and consequences of these phenotypes is essential for grasping the evolutionary history of species. Yet their characterization first requires a good understanding of the (neutral) demographic processes occurring within and between populations.

In the simplest case, any two individuals in a population are equally likely to mate (random mating or panmixia), and the effective size (N_e) of the panmictic population can vary through time as a consequence of changing environmental conditions. Nonetheless, in broadly distributed species, individuals can have a greater tendency to mate with nearby conspecifics, leading to geographical structuration. This phenomenon gives rise to a variety of genetic structure models, such as the totally continuous genetic differentiation of individuals across the entire range (i.e., continuous model), or the organization in set of subpopulations (or demes) spread across a geographic area exchanging migrants with each other (i.e., meta-population model). Such organization may vary in space and time (e.g., (Corrigan et al., *in prep*)) and can reach different levels of complexity going from meta-populations with multiple levels of connectivity in different geographic areas (Baeza & Fuentes, 2013; Maisano Delser et al., 2016, 2019; Pazmiño et al., 2017) to random mating species even at large geographic scale (Corrigan et al., 2018; Pirog et al., 2019). The ability to model genetic diversity across a species' range offers the potential to capture key elements of

its demographic history, such as variations in population sizes or connectivity, migration and colonization processes or population divergence, each of which is expected to leave specific signatures in the genome (Arenas et al., 2012; Excoffier, 2004; Excoffier et al., 2009; Hudson et al., 1992; Kimura & Weiss, 1964; Mona et al., 2014; Nielsen & Slatkin, 2013; Peter & Slatkin, 2013, 2015; Slatkin, 1993; Slatkin & Excoffier, 2012). At a higher resolution scale, species are assembled in communities in a given habitat where their persistence is conditioned by inter-specific interactions (Soule & Stewart, 1970; Vandermeer, 1972) as well as by biogeographic features (Gravel et al., 2011; MacArthur & Wilson, 1967). These community-level processes – that need to be studied using multi-species datasets – have traditionally been neglected in population genetics which rather tends to model one species at a time, although they should leave signatures in the genome of species (Overcast et al., 2023). Characterizing their role in shaping genetic diversity could thus expand our understanding of ecosystem functioning which is all the more important in the light of the biodiversity crisis (Ceballos et al., 2015).

1.1.2. A Need for Complex Models

Understanding the evolutionary history of species is a challenging task whose complexity arises from the need to consider spatiotemporal demographic processes operating at various study scales, as well as their interaction with potential selective processes. This highlights a need both to characterize more how specific features, ranging from meta-population to community-level processes, influence genetic diversity, as well as the development of complex and realistic models tailored to the specific research question and organism(s) under study. These are pivotal not only in fundamental science (e.g., for gaining more insight on the direct interplay between demography and selection, or what drives species organization in space and time), but also in more applied research questions involving the design of coherent conservation plans.

Devising a *good* set of demographic scenarios is therefore complicated and thus requires an accurate understanding of the processes outlined above. Moreover, a framework for building and analyzing complex demographic models is needed, as well as the *right* data on which to test them, the latter being eased by the wealth and increasing affordability of genomic datasets. In this context, the coalescent theory proposes a theoretical basis upon which to infer the complex history of populations (Hudson, 1991; Hudson et al., 1992; Kingman, 1982). However, coalescent-based modelling necessitates a careful design and investigation of scenarios to avoid strong mis-

interpretations (Chikhi et al., 2010; Heller et al., 2013; Maisano Delser et al., 2019; Mazet et al., 2015, 2016). At the same time, classical population genetics models (see below) are usually too simple to harness the complexity of the history of populations while inquiries often necessitate investigating complex scenarios (eventually only investigable using simulation-based methods). This makes the reconstruction of historical demography a challenging exercise in empirical studies as it requires an understanding of the biology of the species, how it is organized in space (and time), a basic understanding of coalescent theory, its assumptions and its related inferential frameworks, as well as the computational resources to handle genomic datasets.

In the following sections of this introduction, I give a brief overview of coalescent theory and how it offers an insightful perspective for investigating the demographic history of species and its limits, including the need to be interpreted carefully to avoid misinterpretations when assumptions are not met. Next, I propose an intuitive inferential approach based on preliminary tests and coalescent-based modeling to be able to study complex but meaningful demographic scenarios and briefly introduce how NGS datasets integrate into this framework. I finally present the main objectives of my PhD in this context and how the different chapters of my PhD will illustrate this.

1.2. Coalescent theory

The way in which the genetic variability of species is shaped by evolutionary processes can be understood (at least partially) by using population genetics theories. The genome of an individual encompasses the whole set of genetic material in the form of DNA sequences, and, in diploid species, each individual receives two sets of nuclear genetic material (i.e., chromosomes) inherited from each parent. In this context, a locus is any kind of genomic region, going from a single base pair to any sequence up to a full chromosome. At a given locus, an individual possesses an allele on each chromosome, and can thus have either a homozygote or heterozygote genotype. At the population scale, a locus can be monomorphic (only one allele) or polymorphic, the latter directly relating to the concept of genetic diversity. In populations, allele frequencies will be influenced in space and time by various evolutionary forces that can increase genetic diversity (mutation), decrease it (selection, genetic drift), or maintain it (gene flow, selection, recombination). Mutations are simply alterations of the DNA sequence that can spread in populations when occurring in germ-cells. Most alterations result in a nucleotide change at a single base (Single Nucleotide Polymorphism, SNP) or in the insertion/deletion of a DNA sequence (i.e., indels). However, alterations can also result in chromosomal rearrangements, such as the rupture of a chromosome fragment that reattaches to the same chromosome but in the opposite direction (inversion) or to another chromosome (translocation). These alterations (or markers) lead to different kind of genetic polymorphism, and are thus used to investigate a huge load of evolutionary processes, including genetic diversity in population genetics. The recombination process happens during the meiosis: the copies of homologous chromosomes exchange material through crossing overs, thus potentially creating new combinations of DNA sequences independently from the mutational process. Mutation and recombination are crucial processes originating the variability in populations. The other evolutionary forces will act on such variants, affecting the trajectories of these polymorphisms.

When a mutation happens in a coding region or in a region having a more or less direct effect in gene expression it can lead to a new phenotype (including lethal) in a population which can be under selection. Selection can be seen as the transmission bias associated to a (set of) genotypes: under a given environmental condition some genotypes may be fitter than others, determining an (almost) predictable variation in allele frequencies. As a result, individual bearing specific genotypes will have more chances to produce viable and successful offspring. Based on the type

of selection, the outcome can be very different. Generally speaking, selection can act against genetic diversity, by favoring the fixation of an allele (e.g., directional selection) or purging deleterious alleles (background/negative selection), or in favor of the maintenance of diversity (e.g., balancing selection, over-dominance, etc.). Some of these processes are more detailed in the introduction of *Chapter 3*. However, most novel alleles have no effect on fitness (Kimura, 1968), arising in a non-coding or coding region not under selection, and are thus considered *neutral*. The fate of a neutral marker will solely depend on the stochastic *genetic drift* and *gene flow* process, which I further jointly refer to as demographic processes. Contrary to selection, which is locus-specific, demographic processes affect the whole genome. In this thesis, I am therefore mostly interested in the study of neutral markers, as they represent the resource on which to reconstruct the demographic history of species, which in turn is necessary to better uncover non-neutral processes.

1.2.1. The WF and the coalescent models

In 1930-31, Wright and Fisher introduced a genetic drift model depicting how the random sampling of alleles at each generation can lead to a change of allele frequencies over time (Fisher, 1930; Wright, 1931). In essence, the Wright-Fisher (WF) model describes changes in genetic variability over time within a population through a binomial sampling of ancestral alleles each generation. WF model depicts an idealized population with multiple assumptions, such as i) a constant size over time, (ii) random mating (i.e., panmixia); (iii) no selection; (iv) no recombination; and (v) no overlapping generations. However, many of these assumptions have been relaxed over the years, leading to elegant mathematical frameworks accounting for selection, gene flow, variation in effective size and non-random mating, making the WF model one of the most commonly used models. However, WF-frameworks are based on whole populations inferences and has some computational limits, in addition to being limited to very simple models. In this context, the coalescent is a probabilistic process that proposes a very efficient framework to understanding and inferring complex demographic processes directly from a sample of lineages rather than whole populations.

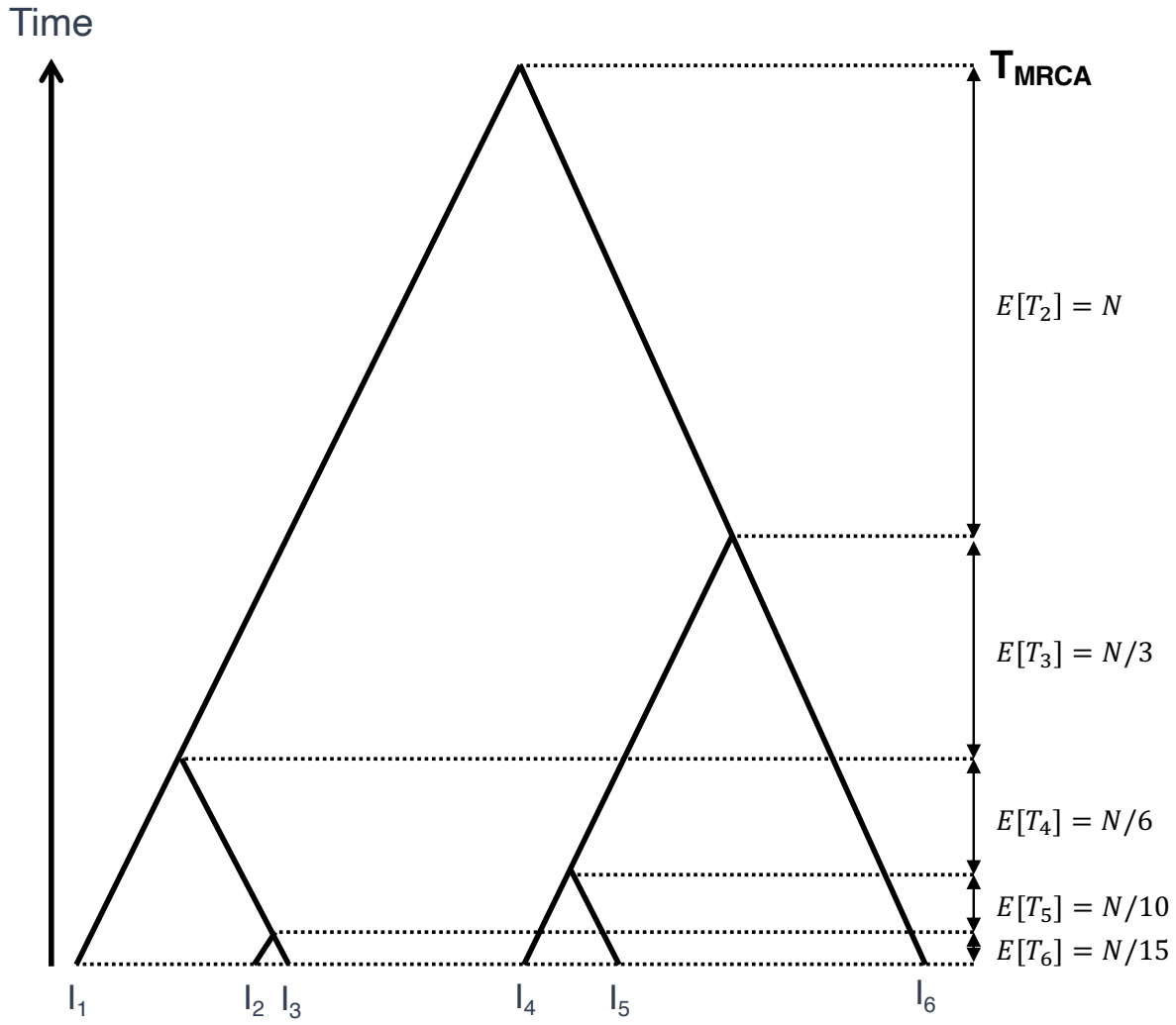

Figure 1.1. Conceptual figure representing a genealogical tree. Going backwards in time, each leaf is represented by a lineage (l_1 to l_6) that coalesce two-by-two at a rate depending on the effective size (N). Waiting times are scaled by the number of remaining active lineages (the expected times to coalesce indicated on the right) until the two last lineages merge into the most recent common ancestor of the sample.

The coalescent was formally introduced in the seminal works of (Kingman, 1982) as an approximation of the ancestral process of an idealized WF population. Hereafter, the description follows the introductions of Hudson (1991), Kingman (1982) and Wakeley (2009). To coalesce means to merge, and to that respect, coalescent theory is about tracing back in time the fate of the sampled lineages, until only one remains. The time of the last coalescence is called the *Most Recent Common Ancestor* (MRCA) of the sample of lineages (individuals when haploid). A graphical representation of this process is a phylogenetic tree, which in this context is called a "gene

genealogy" of a sample of lineages (Figure 1.1). The leaves represent the sampled lineages (i.e., individuals if considering a haploid organism), and each node represent a coalescence event between two active lineages (those that have not yet coalesced). The coalescent has been shown to well approximate the ancestral process of many famous models, including WF model, provided that the number of samples (n) is (very much) smaller than the population size (N), i.e., n << N. The standard genealogical process thus strives on strong assumptions highly similar to those defining a WF population: (1) no fitness related to genetic variation (no selection), (2) No population subdivision (i.e., panmixia), which includes both geographical structures, as well as sex-ratio disequilibrium; (3) no changes in population size over time; and (4) non overlapping generations (Kingman, 1982; Wakeley, 2009).

Briefly, the *genealogical* process describes the distribution of coalescence times – a series of times at which coalescence events happen – in an ideal WF population of size N chromosomes or lineages (Kingman, 1982; Wakeley, 2009; Hudson, 1991), which corresponds to $N_{Haploid}$ or $2*N_{Diploid}$ individuals. In this configuration, the probability that a lineage coalesces with any other lineage at the previous generation is $\frac{1}{N}$, and conversely, the probability that two lineages do not coalesce is $1-\frac{1}{N}$. In a sample of size $n \ll N$ lineages, the probability that no coalescence event happened at the previous generation is therefore the product of probabilities that none of the n sampled lineages coalesce with any of the other $n-1$ lineages:

$$P(n) = \prod_{i=1}^{n-1}\left(1-\frac{i}{N}\right) \approx 1-\frac{n(n-1)}{2N} \quad (1)$$

The probability that a coalescence event occurs at a time t is therefore the probability that it did not occur during $t-1$ generations and that it occurred at time t:

$$P(t) = P(n)^{t-1}[1-P(n)] = \frac{n(n-1)}{2N}\left(1-\frac{n(n-1)}{2N}\right)^{t-1} \quad (2)$$

which, when N is large, can be approximated in by an exponential distribution (making time continuous):

$$P(t) \approx \frac{n(n-1)}{2N}\, e^{-\frac{n(n-1)}{2N}t-1} \quad (3)$$

of parameter $\frac{n(n-1)}{2N}$ and expectation $\frac{2N}{n(n-1)}$.

Because only $n-1$ lineages will remain once the first coalescent event happened, the average time to the next event is $\frac{2N}{(n-1)(n-2)}$, and so on. The waiting time T_n at which n lineages remain can then also be approximated by an exponential distribution with expectation:

$$E[T_n] = \frac{2N}{n(n-1)} \tag{4}$$

This represents the distribution of coalescence times of a sample of size n in a population of size N lineages (i.e., again, corresponding to N haploid or $\frac{N}{2}$ diploid individuals). Note that the variance associated is $Var[T_n] = \left(\frac{2N}{n(n-1)}\right)^2$, which displays one important property of the genealogical process (on which I will come back later): the variance in coalescence times is very large, which means that different trials of the process can lead with different distribution of coalescent times. The fact that coalescence times are exponentially distributed makes the time between coalescence events longer as lineages coalesce together (which is visually evident in Figure 1.1). In fact, the time to the Most Recent Common Ancestor (T_{MRCA}) is a simple function of the distribution of coalescence times:

$$T_{MRCA} = \sum_{i=2}^{n} T_i \tag{5}$$

with expectation:

$$E[T_{MRCA}] = 2N\left(1 - \frac{1}{n}\right) \tag{6}$$

This shows two key properties of the coalescent:

(1) $\lim_{n\to\infty} E[T_{MRCA}] = 2N$, i.e., the T_{MRCA} tend to $2N$ when sample size increases, although this can be reached with a fairly low sample size (e.g., n=10, (Wakeley, 2009)), meaning that we can get a satisfying estimate of the T_{MRCA}without requiring a very large sample.

(2) When n=2, $E[T_{MRCA}] = N$, clearly showing that the last coalescent event takes half the time of all coalescent events. In other words, half the gene genealogy in this model is represented by deep coalescence times, which is evident in Figure 1.1.

Finally, the total length of the tree, $T_{tot} = \sum_{i=2}^{n} iT_i$ with expectation $E[T_{tot}] = 2N\sum_{i=1}^{n-1}\frac{1}{i}$, will always increase with increasing number of sampled lineages, unlike the expectation of the T_{MRCA}. The *genealogical* process described above is only one of the two processes of the standard coalescent, the second being the *mutational* process. The two processes are independent, i.e., the

topology of the tree does not impact the mutational process (and vice-versa). The mutational process follows a Poisson distribution, i.e., a series of Bernoulli success with probability μ, the mutation rate expressed per site and per generation. Under the infinite site and allele models (Kimura, 1969; Kimura & Crow, 1964), the number of sites and alleles are infinite so that a new mutation only occurs in a new site and always gives rise to a new allele. In this case, the expected number of mutations, or segregating sites (S), is simply the product between the mutation rate and the total length of the tree:

$$E[S] = \mu E[T_{tot}] = \mu 2N \sum_{i=1}^{n-1} \frac{1}{i} = \theta \sum_{i=1}^{n-1} \frac{1}{i} \quad (7)$$

with $\theta = 2N\mu$ representing the population mutation rate, or genetic diversity (Watterson, 1975). The mutational process makes the bridge between the genealogical process and *real* data such as DNA sequences as we can reach the *genealogical* process from a sample through estimates of genetic diversity from observed DNA sequences. Two famous estimators of genetic diversity are (i) Watterson's estimate of genetic diversity θ_W which standardizes the number of segregating sites by the total length of the tree (see formula (8); Watterson, 1975) and (ii) the mean pairwise difference θ_π based on the expected number of differences in segregating sites between two individuals as $\theta = E[S]$ (Tajima, 1983). Considering only bi-allelic loci, another famous description of genetic diversity is the *Site Frequency Spectrum* (SFS), which represents the distribution of the frequency of segregating sites in a sample (and from which θ_π and θ_W can directly be computed). The SFS can be assessed in two ways, depending on the available information on the ancestral state of alleles. If alleles are polarized, i.e., we know which one is the ancestral, the SFS represents the distribution of derived alleles frequency, and is referred to as the *unfolded*-SFS, or more intuitively, the *derived* allele frequency spectrum. Conversely, if the ancestral state is unknown, one can calculate the *folded*-SFS, which represents the distribution of the minor allele frequency (i.e., the occurrence of the least frequent of the two alleles) and can be referred to as the *minor* allele frequency spectrum. When the gene genealogy is that of the standard coalescent as described above, the two estimates of genetic diversity (θ_π and θ_W) have the same expectation, and the shape of the SFS can be transformed to a horizontal line (Lapierre et al., 2017) (e.g., dashed lines in any panels of Figure 1.2). However, these summary statistics will change accordingly to the shape of the gene genealogy, which is in turn impacted by the demographic history of the populations/species from which the lineages have been sampled. Any demographic

history different from a constant panmictic population represents a deviation from Kingman's coalescent, affecting the gene genealogy and so the computed summary statistics, as we will see below. For this reason, several neutrality tests based on the SFS (but not exclusively) have been devised to capture the departure from the neutral model. One of the most famous is the *Tajima's D* (Tajima, 1989) which is based on the scaled difference d between θ_π and θ_W (i.e., $d = \theta_\pi - \theta_W$):

$$TD = \frac{d}{\sqrt{Var(d)}} \tag{8}$$

and its response to deviation from some assumptions will be illustrated below.

1.2.2. Changes in effective size and Meta-population structure

The ancestry of samples of most *real* populations is likely poorly represented by a constant-size, non-geographically structured model, and not accounting for these processes can be misleading when implementing the coalescent process. At this point, the most trivial assumption to relax is that of constant effective size, as a sense of how behaves the genealogical process when the effective size changes can be straightforward. The key point to understand it is that a larger N will trivially result in greater waiting times between coalescence events (see equation (4)). In the history of our sample, if N shifts to N_k at T_k generations ago, then, going backwards in time, if $N_k > N$ (i.e., bottkeneck in forwards) coalescence times between T_k and T_{MRCA} are scaled by a larger effective size, vice-versa for $N_k < N$ (expansion). In the bottleneck scenario, waiting times will be longer for the remaining coalescence events, largely increasing the length of branches close to the roots and decreasing the length of branches close to the leaves (Figure 1.2). The total length of the tree will be shorter than if N_k was the effective size during the whole history of the sample. On the contrary, waiting times in the expansion scenario will become comparatively shorter in the deep genealogy than close to the leaves. In the most extreme scenario this will result to what it is usually referred to as a "star-like" genealogy, and the total length of the tree will be longer that if N_k was the effective size during the whole history of the sample. As mutations are randomly Poisson distributed along the genealogy, summary statistics will be impacted by the shape of the gene genealogy. In the bottleneck scenario, there will be a reduction in polymorphic sites belonging to low frequency classes (i.e., singletons, doubletons), due to the longer internal vs external branches when compared to the gene genealogy produced under the neutral model. This

will lead to a SFS with a deficit in low frequency variants (hence a shrinking curve in the normalized SFS in rare frequency classes), skewing the TD to positive values. This is because θ_W is directly impacted by the loss in polymorphic sites, but low frequency variants do not contribute much to θ_π. Conversely, the expansion scenario will lead to an increase of polymorphic sites in terminal branches of the genealogy, hence comparatively increasing low frequency variants, which can be visually displayed by the transformed SFS. For the same reason as above, θ_W is much more impacted than θ_π, resulting in this case in a negative TD value (Figure 1.2).

One of the strongest assumptions of the standard neutral model beside the constant population size is the lack of population structuration. However, many (if not all) species are structured: to be as much general as possible, they are not panmictic over their whole range (independently of its size). In other words, population structure means that if we take a random sample of lineages from our species, they are not equally likely to coalesce with one another. Intuitively, structuration can be geographic: samples located close geographically to one another (e.g., in the same sub-population, or deme) are more likely to have a recent ancestor in common, but not only (i.e., behavioral, cultural, etc.). Classic structured demographic models go from the Isolation with Migration model (IM) where two (or a few) populations are either isolated or connected through migration after divergence (Nielsen & Wakeley, 2001) to equilibrium meta-population models. In the latter, meta-populations are subdivided in arrays of demes exchanging migrants with any deme (e.g., finite island model (FIM)) or only with the closest neighbor's (stepping-stone model, SST) (Kimura & Weiss, 1964) whose migration dynamics is illustrated in the panel D of Figure 1.2.

How does geographic structure affect the gene genealogy, and thus coalescence times? One way to get an insight into this is to follow the elegant work of Wakeley (1999, 2000), which demonstrated that the history of a meta-population could be decomposed in two phases: the *scattering* and the *collecting* phase. Going backwards in time, the *scattering* phase starts at the sampling time (usually the present) of individuals and is very fast in the history of the meta-population (almost instantaneous). Coalescence events happen in deme i at a rate $\frac{n_i(n_i-1)}{2N_D}$ and migrations at a rate $N_D m$, with i being the i^{th} deme of the matrix of sampled demes $(1,2,\dots,d)$ with $d \ll D$ (the number of sampled demes is very much lower than the real number of demes). When each lineage has either coalesce or has been placed in a separate deme by migration, the *scattering* phase ends, marking the start of the *collecting* phase, which will last during most of the history of the sample. At this point, the coalescence events will only happen when a lineage has

managed to migrate into the *right* deme, i.e., a deme within which it can coalesce with another sampled lineage. The *collecting* phase behave as the standard coalescent process in which waiting times are rescaled according to $\frac{n(n-1)}{2N_D\left(1+\frac{1}{4N_D m}\right)}$ when D is large. When sampling within a deme, the coalescence times in the two phases is very different: during the *scattering* phase, the coalescence rate will be faster as it is scaled by N_D, with migration events placing sampled lineages in unsampled demes. The rate of coalescence in the *collective* phase will be slower, because additionally scaled by the number of migrants exchanged each generation, and because most migration events will place a lineage in an unsampled deme (because $d \ll D$) or in a sampled deme with no sampled lineage (because $n \ll N$). In consequence, coalescence times drastically shift downwards between the *scattering* and the *collecting* phase, from a very high to a very slow rate. This shift in coalescence rates is exactly similar to that observed under the bottleneck scenario introduced above, which means that the two processes, while highly different (i.e., the meta-population is constant) will yield a comparable signature in the genome characterized as a deficit in low frequency variants leading to a shrinking SFS in low frequent classes and a positive *Tajima's D* (Figure 1.2). This is a very important result because it means that if one does not account for population structure, and assumes that the population is panmictic, the interpretation will be that of a bottleneck. This is an artifact and it has been discussed widely, supported both by theoretical, simulations and empirical arguments (Chikhi et al., 2010; Heller et al., 2013; Mazet et al., 2015, 2016; Wakeley, 1998, 1999). Note that as explained above, the *scattering* phase is very fast in the history of the sample, which means that the burst in coalescence rate will always be mis-interpreted as a recent bottleneck of the sampled population. This has likely contributed to interpreting recent decrease in genetic diversity due to anthropogenic perturbation in different species (some endangered), thus drastically impeding our ability to design proper conservation plans and our understanding of the evolutionary history of the species.

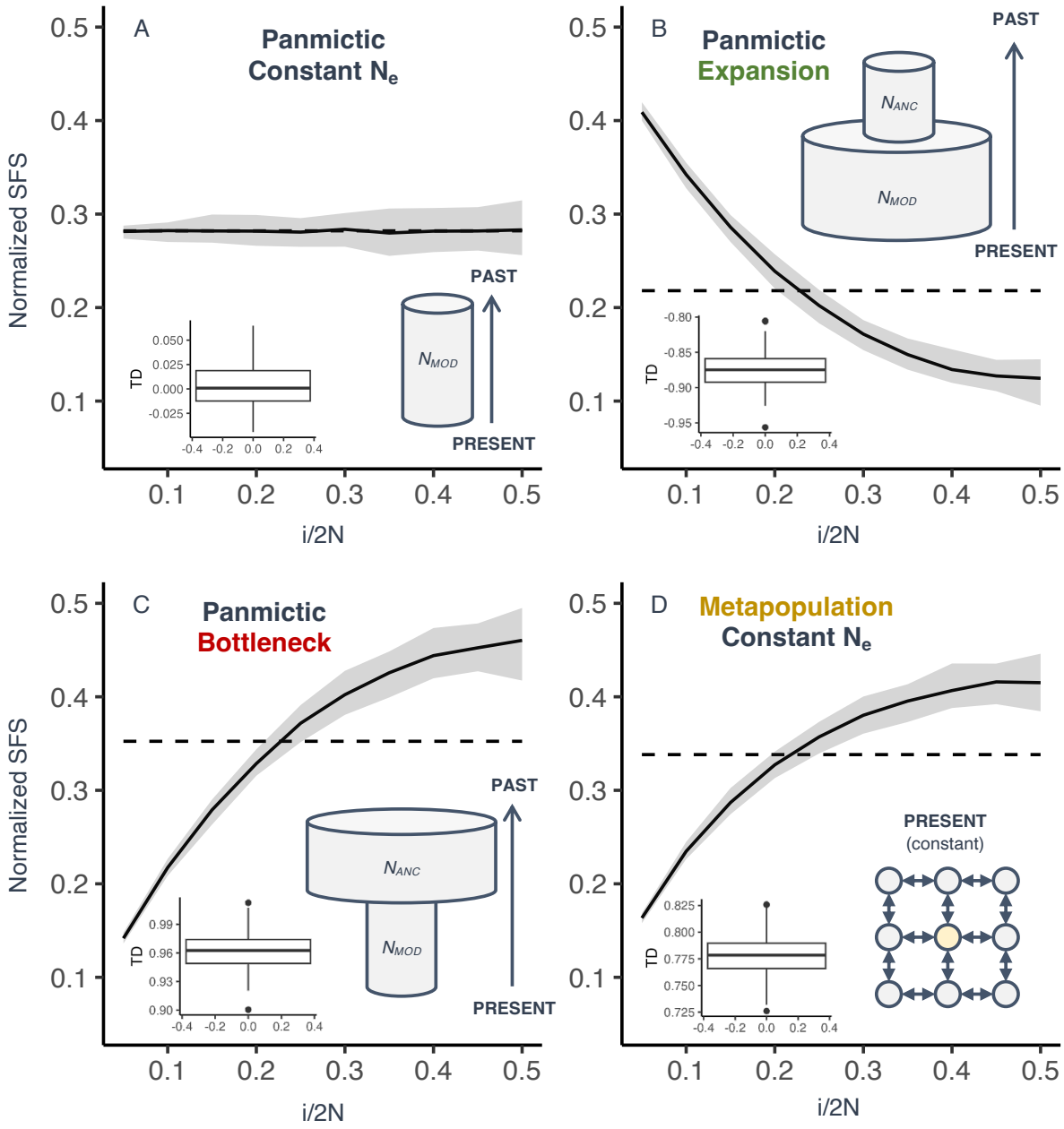

Figure 1.2. Median Normalized Site Frequency Spectrum (SFS) and associated 95% confidence interval and boxplot of Tajima's D (TD) values computed from a sampled of N=20 lineages. 10000 loci are simulated under different demographic scenarios and replicated 100 times: (A) a panmictic and constant population of modern effective size N_{MOD}=100000 individuals; (B) an ancestral panmictic population of size N_{ANC}=100000 which undergoes a 10x expansion 50000 generations ago, (C) an ancestral panmictic population of size N_{ANC} =100000 which undergoes a 10x bottleneck 50000 generations ago and (D) an equilibrium constant stepping-stone meta-population model with 100 demes, each of size N_D=5000 individuals and exchanging with the four closest neighbors *Nm* = 1 migrant per generation following a two-dimensional stepping-stone migration matrix. Mutation rate used was 1.93e-8 per site per generation (generation time of 1 year) and each locus was 115 base pair long.

These examples of deviations from standard assumptions (i.e., non-constant size and population structure) displayed how demographic processes influence the gene genealogy, emphasizing the benefits of studying the coalescence rate trajectory of a sample. As it will be discussed further below, many methods infer with high accuracy changes in Ne under a panmictic population model (hereafter called *unstructured* models). The reason is, as explained above, that changes in Ne are just about rescaling branches length when a change occur, which makes the model efficient to implement. But more broadly, *unstructured* models actually reconstruct the variation of coalescence rate trajectory through time from a sample of lineages, which indeed directly relates to Ne if taken from a panmictic population. As developed with the population structure model example, the coalescence rate reconstructed from a sample of lineages varies accordingly to the specific model: this highlights the interest of investigating *unstructured* models as they convey resourceful information about the true demographic history of the species.

1.2.3. Key Considerations in Coalescent Modeling

The standard coalescent model (including the extensions for changes in N_e and population structure) extract coalescence times simply by tracing the coalescence history of a sample of N, which has been represented through a simple tree structure. However, the structure of the genome makes the relationship between alleles much more complex due to the recombination process, which exchanges homologous genomic regions between two different parental chromosomes by means of crossing overs during meiosis. Recombination has several evolutionary consequences and impact considerably demographic inferences. For example, recombination determines the independent inheritance of sites along the same chromosome copy when far enough. When recombination is considered, the coalescent process becomes much more complex as the process can break up ancestral lineages and lead to a more complicated genealogy (Hudson, 1983, 1991). In the coalescent with recombination, the model takes into account both the coalescent process and the effects of recombination on the genealogical tree (Hudson, 1983, 1991). One intuitive view to understand it is to allow, going backwards in time, the merging of two lineages due to a coalescence event or a recombination event resulting in the split of a lineage in two ancestral lineages. These two ancestral lineages thus represent two loci, each with a different genealogy (called *marginal* genealogies) whose ancestral lineages eventually coalesce together in the history of the sample. This process is referred to as *Ancestral Recombination Graph* (ARG), which simply

represents the sets of marginal genealogies occurring through the process of recombination across a chromosome (Griffiths & Marjoram, 1997). However, the state space of all possible ARG under different sets of genetic parameters is infinite, making the model complicated to explore in empirical studies. To that end, some approximations to the coalescent with recombination have been developed such as the *Sequentially Markovian Coalescent* (SMC) algorithms family (Marjoram & Wall, 2006; McVean & Cardin, 2005) which has been particularly implemented in a lot of software. In brief, The SMC implements coalescence and recombination events along a sequence of a chromosome and allows a floating lineage (i.e., a novel ancestral lineage born by recombination) to coalesce only with a lineage from the previous marginal genealogy (i.e., hence the Markovian nature) unlike the standard model where a floating lineage can coalesce with any lineage from the ARG of all the previous points on the sequence (Griffiths & Marjoram, 1997; McVean & Cardin, 2005; Wilton et al., 2015). SMC was extended to SMC' by simply allowing two new lineages to coalesce back together (which thus results in the absence of changes in coalescence times) which have been shown to better approximate the ARG than SMC (Marjoram & Wall, 2006; Wilton et al., 2015).

The coalescent provides an elegant framework to tracing the ancestry process of a sample of lineages, which bear the signature of the evolutionary forces acting on populations and species. This very simple introduction to the coalescent, from the standard model to the effect of relaxing some of its assumptions and the inclusion of recombination, displays different layers of complexity. Four key points have been emphasized in this brief introduction and are important in the following to understand demographic inferences:

(1) **Coalescence times are determined by demography processes**: investigating variation in coalescence rates through time from a sample using *unstructured* models, whether it directly relates to effective size in a panmictic population or not, ultimately relates to the demographic history of the sample of lineages and is in that respect very resourceful. In addition, there exist many statistics summarizing the shape of the genealogy (i.e., SFS, θ, TD) that can allow to tract demographic processes from empirical datasets when we cannot directly infer or investigates coalescence times.

(2) **Different demographic processes can impact the gene genealogy in a similar way**: different processes can leave the same signature on the gene genealogy. If not coupled with other evidence (e.g., other population genetics inferences or ecological insights), it can lead

to strong misinterpretations of coalescence times as illustrated in the case of population structure *vs.* bottleneck in a panmictic population.

(3) **The coalescent process is associated with great variance**: accuracy in its investigation will therefore be scaled with the quantity and quality of data.

(4) **Complexifying models makes it hard to describe the distribution of coalescent times**: Increasing the complexity of a model (i.e., increasing the number of parameters) can lead to intractable analytical formulas for the density distribution of coalescence times under the specific model, which means that complex scenarios will need to be investigated using alternative approaches than likelihood-based methods.

These four key points are at the core of the development of inferential framework for investigating complex scenarios. They highlight how powerful coalescent-based approaches are, the difficulties they can yield as well as the necessity for extensive, high-quality datasets. In the following part I develop how all these points integrate into an inferential framework in order to investigate the complex evolutionary history of species.

1.3. Historical demography inferences

The coalescent proposes an elegant framework to retrace the history of a sample of lineages taken from a population – or a species – from which we want to deduce demographic processes. However, given the four key points highlighted above, choosing the right tool for the right question (and even the right data) can be complex, but it is mandatory both to avoid highly misleading inferences and to have the ability to answer a specific problem. In this section, therefore, I present what I believe should always be a first step before building complex models, namely gathering information on genetic structure that will provide key elements for building a complex scenario. I then briefly present how coalescence simulators (and simulation-based tools in general) can enable complex model design, and finally highlight how Next-Generation Sequencing (NGS) datasets provide a necessary wealth to meet these challenges.

1.3.1. Towards an "educated" choice of scenarios: the diagnosis step

1.3.1.1. How to correctly investigate demographic history?

Before starting a (more or less) extensive coalescent modelling, gathering enough knowledge of the biology and genetic structure of the species under investigation is warranted. In that respect, the primary step in population genetics inferences and demographic modelling should be to understand the extent of genetic structure underlying the studied sample (i.e., is it better described by a panmictic population or a deme organized in a population?). As outlined above, coalescence times depends on the demographic history of the species and a major dichotomy is between panmictic *vs.* structured models. Moreover, testing for population structure is also important to design meaningful models in regard to the biology of the species (as demonstrated in part 1.2). To that end, one solution is to realize a battery of analyses at both within and between sampling locations scale (hereafter referred to as *descriptive* methods), which, when interpreted altogether, should provide a fertile basis on which designing realistic demographic scenarios.

1.3.1.2. Structured, or not structured?

The extent and diversity of *descriptive* methods applicable obviously depend on the type of genetic dataset at hand (see part 1.3.3), but classical population structure analyses can be performed on any genomic dataset. The question is whether or not the species under investigation is panmictic

and at which scale. In this context, individual-based clustering algorithms such as Principal Components Analysis (PCA), Discriminant Analysis of Principal Components (DAPC) or structure-like methods (Alexander et al., 2009; Frichot et al., 2014; Frichot & François, 2015; Pritchard et al., 2000) coupled to estimation of F_{ST} (Hudson et al., 1992; Reynolds et al., 1983; Weir & Cockerham, 1984) are a powerful diagnosis framework to characterize the extent of genetic disparities across the range distribution of a species. This will inform whether large scale panmixia, divergence model, meta-population(s), or a complex combination of them, are more pertinent to studying the species under investigation. In addition, population structure can be assessed by directly performing demographic modelling. This will be detailed in the 1.3.2 section, but the overall rationale of this strategy is to test whether the gene genealogy of a sample is best depicted by a panmictic population or a deme structured into a meta-population (Maisano Delser et al., 2019). The major advantage of this is that it can be performed on allele frequency data from a single sampling location. Thus, while demographic modeling at this stage can be challenging (e.g., lots of computational resources needed), it can prove useful when the sampling scheme do not cover the whole range of the species, or when it is reduced to a single location, in which case clustering or F_{ST}-based methods cannot assess geographical structure.

1.3.1.3. Gathering additional evolutionary insights

Depending on the questions or hypotheses to be tested and the degree of population structure detected in the first step, one can aim at grasping more information related to the evolutionary processes. Structure analyses may be coupled with spatial analyses: for example, genetic differences may increase between demes (i.e., sub-populations) along with their geographical distance, which is referred to as *Isolation by Distance* (IBD). IBD can be directly tested by using a Mantel test (Mantel, 1967), which computes a correlation between a physical distance matrix and a genetic distance matrix. The genetic distance is usually calculated using one of the estimators of pairwise F_{ST}, but individual-based genetic distances can also be used (Rousset, 2000). Additionally, the extent of genetic isolation/differentiation can be further characterized by accounting for environmental features in IBD, increasing our understanding in terms of connectivity and barriers to gene flow in widely distributed and structured species (Maisano Delser et al., 2019; Mcrae, 2006). Furthermore, most, if not all, species have undergone range contraction, shifts and range expansions (RE) in their history, all leaving specific genomic signatures that could

be detected when studying widely distributed species (Excoffier et al., 2009; Mona et al., 2014). The RE process is a particular case of stepping stone meta-population model where the array of demes is colonized from a starting point representing the origin of RE. The colonization process occurs by series of founding effects (i.e., bottlenecks), leading to a decrease in genetic diversity as well as the fixation of novel alleles (allele surfing) as the colonization goes by (Excoffier, 2004; Excoffier et al., 2009; Mona et al., 2014; Peter & Slatkin, 2013; Slatkin & Excoffier, 2012).These genomic signatures can then be investigated using tools allowing to detect RE occurrence and origin(s), based on the frequency of fixed derived alleles (Peter & Slatkin, 2013, 2015) or the decay of genetic diversity (Ramachandran et al., 2005). These additional steps can allow to identify important biogeographic features such as barriers or corridors to dispersal, as well as the ancestral range distribution or regions more recently colonized, which are all features that can be include in the downstream tested scenario, or that can ease its interpretations *a posteriori*.

1.3.1.4. *Unstructured* coalescence-based models

The final group of descriptive methods utilizes modeling based on coalescent theory. In section 1.2, I highlighted that reconstructing the variation of effective size (N_e) using a panmictic – or *unstructured* – model is quite straightforward as it simply requires to rescale branches length when size changes. And indeed, there are plenty of software implementing *unstructured* models, the most famous being the PSMC (H. Li & Durbin, 2011), MSMC (Schiffels & Durbin, 2014), SMC++ (Terhorst et al., 2017), the skyline plot (Drummond et al., 2005; Pybus et al., 2000) or the stairwayplot (Liu & Fu, 2020, 2015). Yet in a broader sense, these methods reconstruct the variation of the coalescence rate through time, which only relates to N_e if the studied sample is that of a panmictic population. In structured populations, however, the coalescence rate does not relate directly to N_e (as explained in part 1.2), which has been widely discussed before (Heller et al., 2013; Maisano Delser et al., 2016, 2019; Mazet et al., 2015, 2016). This is the reason why (Mazet et al., 2016) proposed to interpret the output of *unstructured* models as the *Inverse Instantaneous Coalescence Rate* (IICR) instead of Ne. Beyond this, the coalescence rate, or the IICR, is expected to hold information about all the parameters of a meta-populations (i.e., effective size, migration, divergence, colonization) which are not necessarily constant over time. This is why *unstructured* models are worth investigated in the set of descriptive analyses: they allow to accurately reconstruct the IICR trajectory which depends on the underlying *true* demographic

model. As such, the IICR can be considered as a summary statistic of demographic history just as estimates of genetic diversity (among others). However, disentangling the effect of historical demography parameters on coalescence times remains a challenge to date as our comprehension of how specific parameter impact coalescence rates need to be refined with extensive investigation. A (beautiful) challenge in population genetics is thus understand what is the most likely combination of demographic parameters that can result in the trajectory of the coalescence rate through time, eventually requiring extensive demographic modelling, as I will outline in the next paragraph.

1.3.2. Investigating complex scenarios using simulations

1.3.2.1. The issue of complex models: how simulations can help

Once the *diagnostic* step is done, one can investigate the demographic history of a population (or a set of populations) using various kind of *coalescent*-based (parametric) models (but not only, see for example *δaδi*, based on diffusion approximations). These models directly aim at interfering the demographic parameters of interests under specific demographic scenario. However, as highlighted above, a strong limitation of the coalescent is that analytical results (i.e., when the expected distribution of coalescent times is known) are limited to only a few simple demographic models (e.g., Isolation-Migration model (Nielsen & Wakeley, 2001), or the n-island model (Beerli & Felsenstein, 2001)) which drastically limits our inferential abilities. Yet, the history of species is very complex, and while simple models might sometimes be enough to answer some questions, investigating more complex scenarios might be mandatory to understand how genetic diversity has been shaped by historical events and spatial features. However, even though deriving the likelihood of the data is possible for few models only, coalescent theory also allows to simulate models of virtually any complexity. In result, a very commonly-used approach in population genetics is to use a coalescent simulator to perform simulation-based inferences to assess demographic parameters. Simulation-based inferences can be done in a variety of ways, such as through approximating likelihood distributions (Excoffier & Foll, 2011) or through Approximate Bayesian Computation (ABC) frameworks (Beaumont, 2019; Beaumont et al., 2002; Bertorelle et al., 2010; Csilléry et al., 2010) that are further explained in the next section.

There are many different kinds of simulators (Hoban et al., 2012), some aiming at inferring with strong accuracy the full ARG despite strong computational burden (e.g., ms (Hudson, 2002) or

msprime (Kelleher et al., 2016)), and other making approximations of the ARG notably through the SMC algorithm to much more efficiently simulate scenarios (e.g., MaCS (Marjoram & Wall, 2006)). One simulator I have been mostly using during my PhD is *fastsimcoal2* (Excoffier et al., 2021; Excoffier & Foll, 2011) which falls into the second category. *fastsimcoal2* is the continuous-time version of simcoal2 and is based on an extension of the SMC' algorithm. As such, it allows the very efficient investigation of any kind of demographic scenario by tracing, going backwards in time, lineages that can coalesce or migrate with possible changes given specific historical events (i.e., colonization times, divergence times, etc.). *fastsimcoal2* has many advantages, encompassing the fact that any kind of scenario can in principle be designed, that any kind of data can be simulated under these scenarios, including complete chromosomes, and that it includes an algorithm to estimates parameters under any simulated scenarios, which I will explain below.

1.3.2.2. Simulation-based inferences using *faststimcoal2*

Approximating the likelihood distribution. To estimate demographic parameters, *fastsimcoal2* uses an approach based on the observed SFS. As it has been explained in part 1.2, the SFS is the distribution of the frequency of mutations, which varies accordingly to the demography. When sampling multiple populations, a multi-dimension SFS (most often, 2D-SFS) can also be computed, which will display the distribution of mutations co-jointly to these sampling locations, thus allowing to grasp insights into the relationship (e.g., migration and divergence) between them. The SFS can thus be modelled using *fastsimcoal2* and demographic parameters are estimated using a *conditional maximization procedure* (i.e., parameters are maximized one at time, leaving the others to their last estimated value). The algorithm search for the combination of parameters producing an SFS (or pairwise 2D-SFS) most closely matching the observed one(s). SNPs used in the analyses are considered independent, producing a composite likelihood estimation. This procedure allows also to compare models, and uncertainty in parameter estimation can be performed by means of non-parametric or parametric bootstrap. It is worth noting that this method, solely based on the SFS in which SNPs are considered independents, does not model recombination, which has the advantage of being faster, but at the same time does not take into account any information related to Linkage Disequilibrium (LD), which is informative about demographic history (especially recent events, see Boitard et al. (2016)).

Approximate Bayesian Computation. Another commonly used strategy is to use *Approximate Bayesian Computation* (ABC) (Beaumont, 2019; Beaumont et al., 2002; Bertorelle et al., 2010; Csilléry et al., 2010). The ABC framework is roughly similar to the procedure above as it relies on simulating summary statistics related to a specific demographic scenario. The major advantage of the ABC framework is its flexibility: it presents the possibility to include any kind of summary statistics, such as SFS-based (i.e., SFS, θ_π, θ_w, TD, F_{ST}), or Linkage Disequilibrium (LD)-based (i.e., LD, Runs of Homozygosity), provided that we can compute it on both the observed and simulated datasets. These set of expected summary statistics will then be compared to the observed summary statistics using various approaches (e.g., rejection algorithms (Csilléry et al., 2012), random forests (Pudlo et al., 2016; Pudlo & Robert, 2019; Raynal et al., 2019), or neural networks (Csilléry et al., 2012; Mondal et al., 2019)) allowing both to choose the most likely scenario and to have a posterior probability distribution of demographic parameters. One approach I have been using is ABC with random forests (ABC-rf), as it proposes many advantages compared to the classic rejection method of Beaumont et al. (2002). Indeed, computationally it requires much less simulations and any summary statistics can be used without having to choose, since the algorithm can account for correlated variables (Pudlo et al., 2016; Pudlo & Robert, 2019; Raynal et al., 2019).

1.3.3. A word on genomic datasets

Coalescent-based demographic inferences require as much loci as possible to avoid the reconstructed signal being impacted by markers under selection, but more importantly to account for the great variance associated to the coalescent process (i.e., as introduced in section 1.2.1, $Var[T_n] = \left(\frac{2N}{n(n-1)}\right)^2$ where T_n represents the coalescent time with n active lineages). Population genetics studies have then been looking for a trade-off between enough sampled individual to accurately calculate statistics (e.g., genetic diversity) and the number of loci to account for this stochastic variance. However, a good estimate of the T_{MRCA}, and thus genetic diversity, can be made with very few samples (e.g., 5 diploids, (Wakeley, 2009)), which put the emphasis on collecting the most loci possible. This was directly stressed by (Felsenstein, 2006) who showed that a considerable number of independent markers rather than samples is mandatory for accurate inferences to help harnessing the high variability of the coalescent process. This became even more striking when Li & Durbin (2011) demonstrated that the coalescence rate through time could be reconstructed from the whole genome of a single diploid individual. In this context, the past

decades have seen the increase in the amount of production of genomic data, which is still getting cheaper and cheaper, allowing the study of demographic and selective processes in model and non-model species through the application of next-generation sequencing (NGS) technologies. NGS datasets can allow to grasp thousands to hundreds of thousands to sometimes millions of Single Nucleotide Polymorphisms (SNPs) and are thus a prime choice for reconstructing historical demography of species. However, the choice of NGS technique is mainly driven by their cost and the nature of the species(s) under investigation (i.e., model vs. non-model species), therefore all are associated with different pros and cons.

When investigating non-model species (i.e., species for which a reference genome has not been established), *Reduced-Representation Libraries* (RRL) techniques offer cost-effective NGS datasets. One of the most famous approaches is *Restriction-associated DNA sequencing* (RADseq) (and associated protocols such as *doubled digested RAD-seq* and *Genotype-By-Sequencing*), which makes use of restriction enzymes to reduce the complexity of the genome (Baird et al., 2008; Miller et al., 2007; Peterson et al., 2012). RADseq is a versatile approach allowing for the sequencing of numerous loci in any species and thus applicable to various research fields like population genetics (Kebaïli et al., 2022; Khimoun et al., 2020; Maier et al., 2022; Mastretta-Yanes et al., 2015), species delimitation (Aurelle et al., 2022; Pante et al., 2015), phylogenomics (Brandrud et al., 2020), and even in aquaculture selective techniques (Robledo et al., 2018). However, challenges arise from limited knowledge of sequenced regions, the non-replicable nature of the protocol, the absence of knowledge of the genome size in non-model species. Despite these issues, when carefully analyzed, RADseq datasets comprising thousands of independent loci are valuable for modeling demographic history and population structure in non-model organisms. Another RRL approach, Target Gene Capture (Jones & Good, 2016), selectively captures and sequences regions of interest in the genome. This approach allows to investigate the same set of homologous regions specifically targeted in different organisms which has been beneficial in phylogenomics (Atta et al., 2022; Bragg et al., 2016) as well as in population genetics (Maisano Delser et al., 2016, 2019). The knowledge associated to the set of regions sequenced is a great advantage over RADseq studies, but it is costlier (except with very large sample size, e.g., > 500) and provides fewer markers.

Whole Genome Sequencing (WGS) aims to sequence the entire genome of an organism, providing a comprehensive view of its genetic makeup. Short-reads WGS are particularly powerful for

identifying as extensively as possible SNPs, but can also help identifying and investigating insertions, deletions, and chromosomal rearrangements. Yet it is worth noting that, to that respect, the 3rd generation of NGS, i.e., long-read WGS sequencing is by far better to identify structural variants. Today, WGS is the tool accounting for genetic and genomic variability in the most extensive way, and is thus by far the ideal tool to studying evolution and biodiversity, including selective processes that are usually hardly investigable with RRL. However, it is more resource-intensive, and while its affordability is always increasing, WGS (and particularly long reads) is extremely expensive compared to other NGS approaches. In addition, it requires a reference genome for the species under study or a closely relative species, which sometimes needs to be established and can be challenging (e.g., quality of the tissues available, abilities to build libraries for long read sequencing, the cost, etc.). To summarize, both WGS and RRL genomic techniques provide valuable information for reconstructing species history, with evidently WGS being by far the best choice in terms of information collected despite its cost. In any case, analyzing thousands of independent loci improves accuracy in coalescence-based inferences over non-genomic data (e.g., mtDNA or microsatellites), although processing such a large amount of information also requires careful bio-informatic processing and can be time and resource consuming, especially for large datasets.

1.4. Overview of the PhD

1.4.1. Main objectives

The ability to model genetic diversity within a species' geographical range presents an opportunity to uncover crucial aspects of its demographic history, whose understanding is in turn essential for gaining insights into localized selective mechanisms. When examined at the community scale, modeling genetic diversity should also facilitates an exploration of how inter-species interactions and biogeographic characteristics collaboratively influence a species' evolutionary trajectory. However, as developed in this introduction, designing and investigating complex models is a challenging task but undeniably helps understanding all of these processes, provided that it is coherent with the question and the organism(s) under investigation. At the same time, it is also crucial to increase our comprehension of how specific processes influence the genome of species(s), which might necessitate to go further the traditional boundaries in population genetics and integrate modeling from multiple species to gain large-scale ecological information.

In this context, the research objectives of my PhD can be distilled into two primary goals. First, it aims to demonstrate how a meticulous examination of the neutral processes impacting a species' genome, achieved through thoughtful model selection, not only yields crucial insights for formulating evolutionary hypotheses and conservation strategies, but can also be essential for uncovering selective processes. Second it aims to extend the conventional single-species approach in population genetics to a broader multi-species perspective, in order to increase our understanding of how large-scale processes might impact the gene genealogy of sampled populations. To answer these objectives, I studied different marine systems and investigated historical demography processes through genetic diversity modelling at different scales, from supergenes to ecosystems, and organized my PhD thesis in the following chapters:

- **Meta-populations, Models and Conservation:** determinants of coalescence times in structured species and its importance in the investigation of widely distributed species. The global aim of this chapter is to investigate the influence of population structure on the demographic history reconstructed using *unstructured* models, with a particular interest in what these models reveal about the evolutionary history of structured species. At the same time, specific life-history traits can impact coalescent patterns. To understand this, I study the evolutionary history of two shark species with very different life history traits but with large distribution.

- **Supergenes, Demography and Conservation**: reconstructing the demographic history to understand the origin and consequences of a supergene determining the size in the vulnerable thorny skate (*Amblyraja radiata*). In this chapter, I present the discovery of an introgressed size-determining supergene using whole-genome sequencing data. I then explain how the careful reconstruction of demographic history is essential for understanding the strong conservation consequences of the supergene, as well as providing key insights into its origin. Finally, I stress that further characterization of the supergene's history will only be possible through a multi-species study.
- **Genetic Signatures of Ecosystem Functioning**: Extending population genetics boundaries to multi-species inferences. The global aim of this chapter is to understand how interactions at the community scale as well as biogeographic features shape genetic diversity through joint modelling of genetic variability and trophic niche size of 43 reef fish species from Moorea. This study, relying on a unique dataset, provides new insights on how ecosystem scale processes influence the genome of species, leading to a discussion over major ecological and biogeographical theories.

Chapter 2. Meta-populations, Models and Conservation

Figure 2.1. Grey reef sharks (*Carcharhinus amblyrhynchos*) swimming with a school of yellowfin goatfish (*Mulloidichthys vanicolensis*) in Fakarava, French Polynesia.

2.1. Context

2.1.1. Meta-populations and *unstructured* models

Many species, if not all, have undergone contractions, expansions and shifts in their range distribution during their history (Arenas et al., 2012; Excoffier et al., 2009; Klopfstein et al., 2006; Slatkin & Excoffier, 2012). This is usually assumed to result from global changes such as glacial-interglacial successions (Arenas et al., 2012; Excoffier et al., 2009; Lee-Yaw et al., 2008), where the range of species tend to contract in refugia during less favorable times after which they can (re)colonize new habitats. However, these can also be induced by more rapid perturbations such as habitat destruction, overfishing causing local extinctions of populations (Pacifici et al., 2020; Worm & Tittensor, 2011; Yan et al., 2021) or invasion of new habitat (Lopes et al., 2023). Genomic signatures of contractions and shifts can be hard to detect empirically, hence resulting in little attention despite an established complex influence on genetic diversity, strongly related to the biology of the species and the velocity of the process (Arenas et al., 2012). On the other hand, range expansions (RE) have been much more documented. REs occur by series of founder effects leaving specific signatures in the genome and for which we have solid theoretical expectations (Excoffier et al., 2009; Mona et al., 2014). In the case of a meta-population – a set of demes (or sub-populations) exchanging migrants with each other – REs leave specific signatures in the gene genealogy of lineages sampled from a deme belonging to a meta-population (Ray et al., 2003). Many species, especially when widely distributed, might be organized in meta-populations and could display signatures of REs. In the introduction, I highlighted how *unstructured* models were powerful tools to reconstruct the variation in coalescence rate through time, as they directly relate to the true demographic history of the species under investigation. In this chapter, I investigate practically the usefulness of unstructured models in the case of a RE.

2.1.2. A test-case on sharks

In this chapter, I aim at increasing our understanding of how meta-population structure impacts the coalescence rate through time from a conceptual point of view, but also to characterize better determinants of population structure in practice, by studying various species of sharks. Sharks represent a rich group of more than 500 species, and are found in all oceans and seas but also in rivers, displaying a vast spectrum of Life History Traits (LHT) and biological features (Compagno,

1984, 2001). Meta-population structure is probably frequent in sharks, whose dispersal ability can be conditioned by their movement range (Smith & Weissman, 2020; Trakhtenbrot et al., 2005), size (Parsons, 1990), or behavioural traits such as residency and philopatry (Chapman et al., 2014). However, while meta-population structure has often been suggested in reef sharks (Gledhill et al., 2015; Maisano Delser et al., 2016, 2019; Momigliano et al., 2015, 2017; Whitney et al., 2012), vagile species can be panmictic at large geographic scale (Corrigan et al., 2018; Karl et al., 2010; Pirog et al., 2019; Vignaud, Maynard, et al., 2014). In result, they represent a great test-case for investigating how LHT and ecological features contribute to shape the genetic structure. At the same time, this group is highly vulnerable: more than 37% of shark species face a risk of extinction (Dulvy et al., 2021) and fewer than 30% of these are experiencing stable or increasing population trends according to the International Union for Conservation of Nature (IUCN) Red List of threatened species, mostly due to overfishing (Dulvy et al., 2014). This emphasizes the need for coherent shark conservation plans which would benefit from additional population genetic studies at the scale of their range. This is all the more important since as meso or apex predators in their communities, sharks play key roles in their ecosystems (Bornatowski et al., 2014), and their decline has already shown to have consequences on the ecosystems they live in (Friedlander & DeMartini, 2002; Myers et al., 2007).

2.2. Objectives

This chapter aims first to increase our understanding of the effect of meta-population structure on the coalescence rate as reconstructed using *unstructured* models. In a second time, it aims to investigate the influence of life history traits (LHT) on the degree of population structure by studying multiple shark species. Finally, it aims to wrap up the former findings through two large scale studies of endangered shark species with contrasting LHT, specifically highlighting the importance of testing for population structure before making further inferences and/or interpreting demographic signals. This chapter is thus composed of three articles:

(1) Coalescence times, Life history Traits and conservation concerns: an example from four coastal shark species from the Indo-Pacific

(2) Ecological and biogeographic features shaped the complex evolutionary history of an iconic apex predator (*Galeocerdo cuvier*)

(3) Like a rolling stone: Colonization and migration dynamics of the gray reef shark (*Carcharhinus amblyrhynchos*)

2.3. Coalescence times, Life history Traits and conservation concerns: an example from four coastal shark species from the Indo-Pacific

This article has been published in *Molecular Ecology Resources*.

Authors:

Pierre Lesturgie, Serge Planes, Stefano Mona

2.3.1. Abstract

Dispersal abilities play a crucial role in shaping the extent of population genetic structure, with more mobile species being panmictic over large geographic ranges and less mobile ones organized in meta-populations exchanging migrants to different degrees. In turn, population structure directly influences the coalescence pattern of the sampled lineages, but the consequences on the estimated variation of the effective population size (Ne) over time obtained by means of unstructured demographic models remain poorly understood. However, this knowledge is crucial for biologically interpreting the observed Ne trajectory and further devising conservation strategies in endangered species. Here we investigated the demographic history of four shark species (Carharhinus melanopterus, Carharhinus limbatus, Carharhinus amblyrhynchos, Galeocerdo cuvier) with different degrees of endangered status and life history traits related to dispersal distributed in the Indo-Pacific and sampled off New Caledonia. We compared several evolutionary scenarios representing both structured (meta-population) and unstructured models and then inferred the Ne variation through time. By performing extensive coalescent simulations, we provided a general framework relating the underlying population structure and the observed Ne dynamics. On this basis, we concluded that the recent decline observed in three out of the four considered species when assuming unstructured demographic models can be explained by the presence of population structure. Furthermore, we also demonstrated the limits of the inferences based on the sole site frequency spectrum and warn that statistics based on linkage disequilibrium will be needed to exclude recent demographic events affecting meta-populations.

Keywords: coalescence, life history traits, meta-population, population genomics, sharks

2.3.2. Introduction

Reconstructing the evolutionary history of a species is a challenging exercise only partially eased by the growing size of genetic data available. Indeed, larger amounts of data will provide more precision but not more accuracy if the model(s) chosen to infer demographic parameters is distant from the true one. Species are dynamic entities whose geographic range has often changed in time through range expansions, contractions and shifts (Arenas et al., 2012; Excoffier et al., 2009; Mona et al., 2014). As a consequence, many species are most likely organized in meta-populations (i.e. groups of demes or sub-populations exchanging migrants to some extent), even though the more vagile ones might be panmictic at a large scale (Corrigan et al., 2018; Karl et al., 2010). Neglecting the meta-population structure (i.e., performing demographic inferences under *unstructured* models) may lead to spurious inference of population size change (Chikhi et al., 2010; Maisano Delser et al., 2016, 2019; Mazet et al., 2015), which is particularly worrisome for species of conservation concern. Unfortunately, the link between the inferred temporal trajectory of the effective population size (*Ne*) and the real demographic history of the meta-population remains largely under explored. However, the role of connectivity, particularly the number of migrants *Nm* exchanged each generation and the migration matrix, has been put forward as a key actor in shaping the gene genealogy of lineages sampled from a deme belonging to a meta-population (Chikhi et al., 2010; Mona et al., 2014; Ray, Currat, & Excoffier, 2003; Städler, Haubold, Merino, Stephan, & Pfaffelhuber, 2009).

Understanding the relations between meta-population structure, the inferred *Ne* variation under *unstructured* models, and species dispersal abilities, is crucial to correctly interpret the pattern of genetic variability and to establish conservation priorities. To search for general rules describing such relations, we followed an inductive approach investigating species: i) with large distribution (which in principle should guarantee an organization in meta-populations); ii) with different life history traits (LHT) related to dispersal; iii) of conservation concerns. In this spirit, we selected for our genomic study four shark species (*Carcharhinus amblyrhynchos*, *Carcharhinus limbatus*, *Carcharhinus melanopterus,* and *Galeocerdo cuvier*) from New Caledonia. These species have a large and overlapping distribution in the Indo-Pacific (https://sharksrays.org/) and they differ for LHT features such as size (which is positively correlated with the capacity for long distance swimming and oceanic migration (Parsons, 1990)), residency pattern, and long-distance dispersal ability as measured by tagging data (Supp. Table 2.3). Moreover, the IUCN red list reported that

the black-tip shark (*C. limbatus*) and the tiger shark (*G. cuvier*) are Near Threatened (with a decreasing trend in the tiger shark), the black-tip reef shark (*C. melanopterus*) is Vulnerable with decreasing trend, and the grey reef shark (*C. amblyrhynchos*) is Endangered with decreasing trend as well. We first compared several population genetics models by means of coalescent simulations coupled with an *approximate Bayesian computation* framework (Bertorelle et al., 2010) to detect whether panmixia or a meta-population model best describe the genomic variation of each species. Then, we inferred the demographic parameters under the most likely model and applied the *stairwayplot*, which assumes a panmictic unstructured population (Liu & Fu, 2015), to detect the *Ne* variation through time in each species. We finally run extensive coalescent simulations under the tested meta-population models with parameters compatible to those observed in real data. The simulated datasets were in turn analysed with the *stairwayplot* to: i) help interpreting the observed data in the four shark species; ii) providing general coalescence arguments relating the demographic history of a meta-population and the reconstructed variation in *Ne* through time by means of *unstructured* models.

2.3.3. Material & Methods

2.3.3.1. Sampling

Eight specimens of tiger shark (*G. cuvier*), 13 black tip shark (*C. limbatus*), and 12 grey reef shark (*C. amblyrhynchos*) were collected off New Caledonia. Total genomic DNA was extracted from muscle tissue or fin clips, and preserved in 96% ethanol using QIAGEN DNeasy Blood and Tissue purification kit (Qiagen, Hilden, Germany) according to the manufacturer's protocols. Double-digest restriction-associated DNA (ddRAD) libraries were prepared following (Peterson et al., 2012) using EcoRI and MspI restriction enzymes and a 400-bp size selection. The genomic libraries obtained were sequenced with a HiSeq 2500 Illumina sequencer (single-end, 125 bp). Exon capture data of eight *C. melanopterus*) from New Caledonia (Maisano Delser et al., 2019) were included in this study for comparative purposes.

2.3.3.2. De novo assembly and data filtering (dd-RADseq samples)

Raw reads were first demultiplexed and quality filtered through the *process_radtags.pl* pipeline in *Stacks* v.2.5 (Rochette et al., 2019). In the absence of a reference genome for any of the three species, RAD-seq loci were *de novo* assembled independently in each species under the *denovo_map.pl* pipeline in *Stacks*. We used the following assembly parameters: *m*=3 (minimum

read depth to create a stack), *M*=4 (number of mismatches allowed between loci within individuals), and *n=4* (number of mismatches allowed between loci within catalogue). We found an average coverage per species of ~10x (see results). A consensus on the threshold below which SNP calling may be considered unreliable is still lacking. However, genotype free estimation of allele frequency is generally recommended with low to medium coverage (Korneliussen et al., 2014). This approach, implemented in the software *Angsd* v.0.923 (Korneliussen et al., 2014), has been rarely applied to Rad-seq data (however, see (Warmuth & Ellegren, 2019) for an exception) and, to our knowledge, never to Rad-seq data from non-model organisms, probably due to the need of a reference sequence for the software to work. Here, we followed the approach of (Heller et al., 2021; Khimoun et al., 2020) by creating an artificial reference sequence. First, we used the *population* script in *Stacks* to assemble loci present in at least 80% of the individuals (using the flag *r=0.8*); then, we concatenated the consensus sequences of the retrieved loci spaced by a stretch of 120 N (unknown) characters (the same length of the Rad-loci) to facilitate the subsequent mapping. Raw reads were then mapped back to the novel reference sequence by means of the *bwa-mem* algorithm with default parameters (H. Li & Durbin, 2009). Using custom bash scripts coupled with *Angsd*, we applied a number of filters to the aligned data and eliminated: *i)* sites with coverage <3 (*-minIndDepth=3* flag), *ii)* bad quality bases and poorly aligning reads (*-minQ* and *-minMapQ* and *-C* flags with default values); *iii)* poor quality sites based on the per base alignment quality (*-baq=1* flag); *iv)* SNPs in the last 5 bp of each locus; *v)* SNPs heterozygote in at least 80% of individuals; *vi)* loci with more than 5 SNPs that could potentially be paralogous; *vii*) sites with missing data by setting the *-minInd* flag to the total number of individuals retained in each species. The filtered dataset was then used to generate a site allele frequency likelihood file, with the genotype likelihoods computed with the SAMtools method (*-GL=1* flag), further optimised to compute a folded *site frequency spectrum* (SFS) with no missing data for downstream analyses. An alternative (and simpler) approach would have been to augment *m* to achieve an higher coverage (Paris et al., 2017). However, beside the considerable loss in the number of assembled loci (and hence of retrieved SNPs), we found by extensive simulation of *in silico* Rad experiments that selecting high coverage loci biases the SFS towards low frequency variants (Mona et al., 2023). The SFS for C. melanopterus was estimated directly from the high coverage exon-capture dataset of Maisano Delser et al. (2019).

2.3.3.3. Genetic diversity and demographic inferences

Nucleotide diversity (θ_π), θ_w (Watterson's theta, based on segregating sites (Watterson, 1975)) and Tajima's *D (TD,* (Tajima, 1989)) were computed from the SFS for each species with custom scripts. Significance of *TD* was evaluated after 1,000 coalescent simulations of a constant population model with scaled size θ_π. To test whether sampled demes are isolated or belong to a structured meta-population and to eventually estimate connectivity, we devised three alternative evolutionary models for each species (Figure 2.2) within an *approximate bayesian computation* (ABC) framework. Model NS (non-structured) defined an isolated population characterized by a modern effective population size (N_{MOD}) switching instantaneously into an ancestral population size (N_{ANC}) at *Tc* generations before present. Model FIM specifies a non-equilibrium finite island model defined by *d*=100 demes exchanging *Nm* migrants each generation under a symmetric migration matrix. The array of demes is instantaneously colonized T_{COL} generations before present from a population with an ancestral size (N_{ANC}). Model SST is similar to FIM but demes exchange migrants only with their four neighbours (or less, if they are at the border of the array), in a steppingstone fashion. We performed 50,000 coalescent simulations from prior distributions using *fastsimcoal* v.2.6.0.3 (Excoffier et al., 2013), reproducing the exact number of individuals and loci for each species (Table 2.1). We first performed model selection through the random forest (RF) classification method implemented in the abcRF R package (Pudlo et al., 2016). We then performed 50,000 additional simulations under the most supported model in order to estimate demographic parameters with the abcRF regression method (Raynal et al., 2019). Both model selection and parameter estimation were computed with the following set of summary statistics: the SFS, θ_π, θ_w and *TD.* The first two axes of a Linear Discriminate Analysis performed on the previous statistics were also included for model selection in order to increase the accuracy of the estimates (Pudlo et al., 2016). Even though θ_π, θ_w and *TD* are function of the SFS, they convey additional information by the non-linear feature of the functions. Information redundancy among the considered summary statistics is accounted for by the RF algorithm. Model selection and parameter estimation were run twice on each set of simulations to check the consistency of the analyses, and cross validation (or confusion matrix for the model selection) was performed on the first of the two runs. The number of trees in each RF algorithm was chosen by monitoring the evolution of the out-of-bag error (Pudlo et al., 2016).

We investigated the variation in the effective population size (*Ne*) through time by running the composite likelihood approach implemented in the *stairwayplot* v.0.2 software (Liu & Fu, 2015). We set the generation time to seven years for *C. melanopterus* (Maisano Delser et al., 2016) and to 10 years for the other species (Cortés, 2002; Pirog et al., 2019) for all demographic inferences. We applied a mutation rate per generation per site of 8.4×10^{-9} to the exon capture data of *C. melanopterus* (Maisano Delser et al., 2016) and of 1.93×10^{-8} to the RADseq data for the remaining three species. This mutation rate was determined by scaling genetic diversity between ddRAD (obtained under the same protocol of this study) and Exon Capture data from 12 *C. melanopterus* individuals from Moorea, French Polynesia (Supplementary Material).

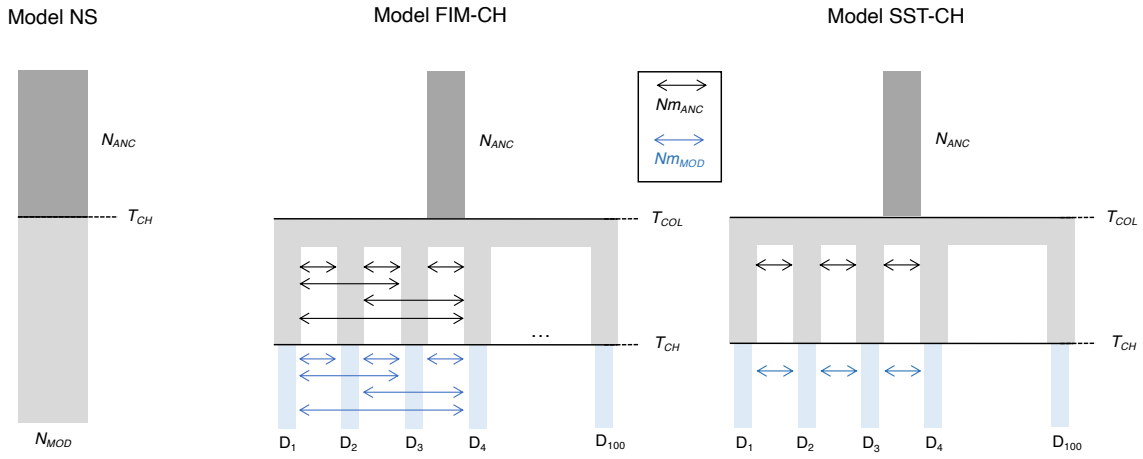

Figure 2.2. Evolutionary scenarios considered in this study (to both infer parameters in real data under an ABC framework and to perform coalescent simulations). SST (FIM) model is a simplified version of SST-CH (FIM-CH) in which connectivity *Nm* is constant after T_{COL}. Details on each parameter are presented in the main text.

2.3.3.4. Simulation study

We ran coalescent simulations under FIM, SST and their modified version FIM-CH and SST-CH, where the *Nm* parameter is changed at T_{CH} generations B.P. (Figure 2.2), to first inspect the shape of the SFS and to further uncover the variation of *Ne* over time assuming a panmictic population by means of the *stairwayplot*. We investigated in total 288 demographic scenarios under the four meta-population models (Table 2.2, Supp. Tables 2.4, 2.5, 2.6, 2.7 and 2.8). Similarly to the analyses performed on the real data, all scenarios were represented by d=100 demes exchanging migrants. We sampled 10 diploid individuals either from a randomly selected deme in the case of FIM/FIM-CH (since all demes have the same coalescence history) or from the central deme of the

array in the case of SST/SST-CH (to avoid border effects). Deme size was fixed to N_{DEME} =5000 with *m* varying accordingly to obtain a long-term *Nm* of 1, 5, 10, and 15 in order to encompass the range of the estimated values (see results). T_{COL} was fixed to 5,000, 15,000 and 50,000 generations B.P. or to ∞ (i.e., equilibrium model), and the ancestral effective size was fixed to N_{ANC} = 50,000. Change of connectivity occurred at T_{CH} =10 or 50 generations B.P., to mimic human induced effects due to overfishing and/or habitat modifications (i.e., climate changes). Looking forward in time, we modelled the change in connectivity by instantaneously decreasing *m* or N_{DEME} by a factor 10 or 100 with respect to the long-term *Nm* (Supp. Tables 2.5, 2.6, 2.7 and 2.8). For each combination of parameters, we performed 100 coalescent simulations of 50,000 Rad-like loci of 115 bp. Mutation rate per site per generations was set to 1.93×10^{-8} and the generation time to 10 years. We computed for each scenario (averaged over the 100 replicates): a) summary statistics (θ_π, θ_w, and *TD*); b) the normalised SFS as in (Lapierre et al., 2017); c) the *stairwayplot*, to reconstruct the apparent variation of *Ne* through time. We note that the number of diploid individuals and simulated loci were chosen to be consistent with our data (preliminary analyses conducted on a subsample of 5,000 loci produced consistent results).

2.3.4. Results

Summary statistics (number of assembled loci, SNPs, genetic diversity and Tajima's *D*) are presented in Table 2.1. Mean coverage (and standard deviation) per sample was 9.02 (± 2.62), 7.93 (± 0.48), 8.39 (± 0.81) for *G. cuvier, C. limbatus* and *C. amblyrhynchos* respectively.

We compared the models NS, FIM, and SST (Figure 2.2) in the four species by means of an ABC-RF algorithm and estimated demographic parameters for the most supported model. After checking for the evolution of the out-of-bag error of the RF, model selection and parameter estimation were computed using respectively 500 and 1,000 trees in each species. We found that NS had the higher posterior probability (*p*=0.84) for *G. cuvier* (Table 2.1 and Supp. Table 2.9). In contrast, demographic histories of the three other species were best described by SST, with a posterior probability ranging from 0.53 to 0.88 (Table 2.1 and Supp. Table 2.9). The estimated median number of migrants per generation *Nm* was 1.8 (95% CI: 0.7-3.0) for *C. melanopterus*, 6.6 (95% CI: 1.5-15.4) for *C. limbatus,* and 11.5 (95% CI: 3.0-22.0) for *C. amblyrhynchos* (Figure 2.3 and Table 2.1). The posterior distribution of *Nm* strongly differed from the prior distribution and showed a clear unimodal peak with small credible intervals, and low mean square error (SME) and

mean root square error (SMRE) in all three species (Figure 2.3 and Supp. Table 2.10), suggesting that these estimates are highly reliable. Conversely, both T_{COL} and N_{ANC} had larger SME and SMRE errors in all species (Table 2.10), but it was only in *C. melanopterus* where posterior and prior distribution could not be distinguished (Figure 2.3). T_{COL} has a clear unimodal distribution in *C. amblyrhynchos* but a more disperse one (and with wider credible intervals) in *C. limbatus* (Figure 2.3, Table 2.1).

Table 2.1. Summary statistics and ABC estimation. Number of loci and SNPs after filtering, mean pairwise difference (θ_π), Watterson theta (θ_w), Tajima's D (TD), posterior probability of the most supported model and its parameters (median value and 95% credible interval in parentheses).

	N° Loci	N° SNP	θ_π	θ_w	TD	Model (probability)[†]	*Nm*	T_{COL} [‡]	N_{ANC}
G. cuvier[§] (N=8)	117976	25785	0.00057	0.00051	-0.03	NS (0.84)	-	-	-
C. amblyrhynchos (N=12)	69490	68355	0.00216	0.00229	-0.23[*]	SST (0.85)	11.5 (3.0-22.0)	20456 (12567-75649)	40961 (1315-49276)
C. limbatus (N=13)	60812	43449	0.00180	0.00166	0.43[*]	SST (0.55)	6.6 (1.5-15.4)	50198 (475-245440)	25521 (1913-52820)
C. melanopterus[¶] (N=8)	926	784	0.00040	0.00030	0.691[*]	SST (0.89)	1.8 (0.7-3.0)	91719 (5000-291341)	34607 (2760-95380)
						Priors [a]	U: 0.001 - 100	U: 1 - 300000	U: 100 – 100000.

[*] Tajima's *D* values are significant (p<0.001).
[†] Most supported model and its posterior probability.
[‡] T_{COL} is expressed in generations.
[§] *G. cuvier* is best represented by the NS model: its demography is depicted through the *stairwayplot* algorithm (see discussion).
[¶] Data from (Maisano Delser et al., 2019)
[a] Uniform prior distribution. The prior distribution of *Nm* is the product of two uniforms (one for *N* and one for *m*).

The *stairwayplot* showed a nearly similar dynamic for *C. amblyrhynchos* and *C. limbatus*, characterized by a strong ancestral expansion (Figure 2.4). When approaching T=0, both species underwent a bottleneck but of distinct strength. This is consistent with the shape of the normalized SFS, which clearly shows a stronger deficit in low frequency variants for *C. limbatus* compared to *C. amblyrhynchos* (Figure 2.4). Similarly to *C. limbatus*, *C. melanopterus* experienced a recent

10-fold population collapse around 20,000 years B.P. starting from a long term constant *Ne*. However, *C. melanopterus* showed no signature of ancestral expansion, consistent to the results obtained by (Maisano Delser et al., 2019) using *abc-skyline* method. Finally, *G. cuvier* displayed an ancestral expansion around 100,000 years B.P. with *Ne* reaching ~12,000 before dropping to ~3000 at T ~1,600 years B.P. Remarkably, the ancestral expansion retrieved by the *stairwayplot* (Figure 2.4) for both *C. amblyrhynchos* and *C. limbatus* overlap with the posterior distribution of T_{COL} estimated by the SST model (Table 2.1). This analogy holds too for *C. melanopterus,* where T_{COL} could not be properly estimated under the structured model (we obtained a flat posterior distribution, Figure 2.3) and there was no signature of ancestral expansion in the *stairwayplot* (Figure 2.4).

Table 2.2. Coalescent simulations of 50,000 Rad-loci under SST model, with mutation rate fixed to $1.93*10^{-8}$ per site per generation and N_{ANC} fixed to 50,000. Mean pairwise difference (θ_π), Watterson theta (θ_w), Tajima's D (TD), and number of segregating sites (S) are averaged over 100 replicates.

Nm	T_{COL}	θ_π ‡	θ_s ‡	**TD**	**S**
1	5000	0.0013	0.0011	0.531	23599
	15000	0.0013	0.0012	0.405	24094
	50000	0.0017	0.0016	0.406	32201
	∞†	0.0161	0.0139	0.669	283564
5	5000	0.0017	0.0016	0.361	32443
	15000	0.0019	0.0018	0.191	37712
	50000	0.0028	0.0028	0.035	56474
	∞	0.0177	0.0150	0.749	306786
10	5000	0.0019	0.0018	0.180	36561
	15000	0.0021	0.0022	-0.087	44380
	50000	0.0031	0.0034	-0.364	69436
	∞	0.0180	0.0158	0.585	321619
15	5000	0.0019	0.0019	0.048	38919
	15000	0.0022	0.0024	-0.274	48479
	50000	0.0032	0.0038	-0.608	77391
	∞	0.0181	0.0163	0.465	331816

† Equilibrium model obtained by simulating $T_{COL}=\infty$.
‡ Theta values are expressed per site per generation.

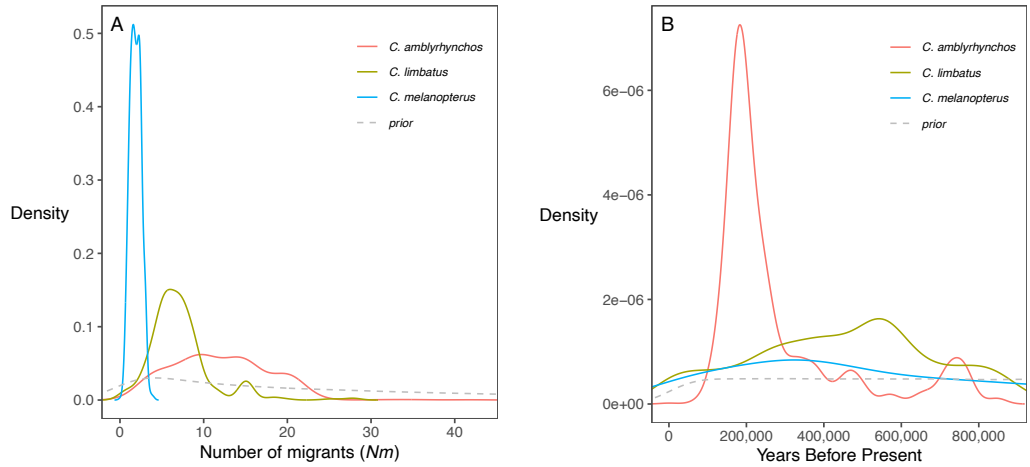

Figure 2.3. Posterior distribution of the number of migrants per generation *Nm* (panel A) and of the colonisation time of the array of deme T_{COL} (panel B) estimated under the stepping stone model (SST) for *Carcharhinus amblyrhynchos* (red), *Carcharhinus limbatus* (green) and *Carcharhinus melanopterus* (blue).

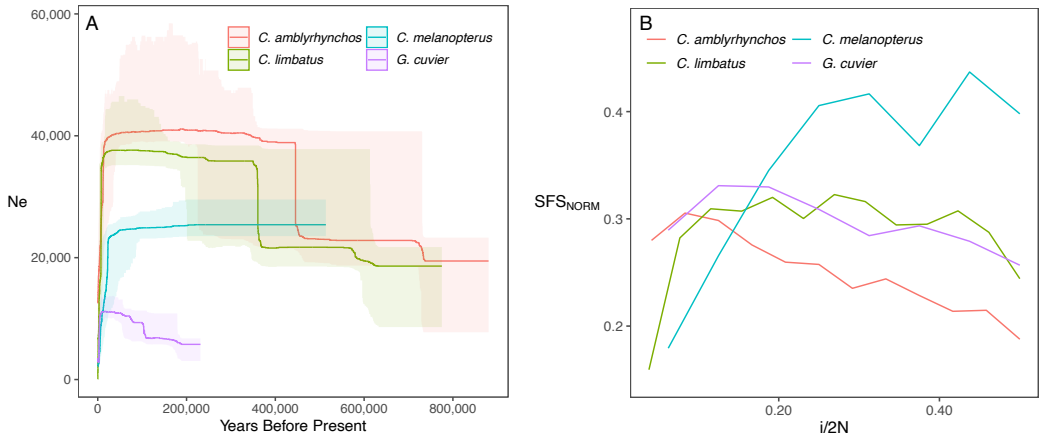

Figure 2.4. Panel A: variation of the effective population size (*Ne*) through time and its 75% confidence interval estimated by the *stairwayplot*. Panel B: normalized SFS computed as in (Lapierre et al., 2017). *Carcharhinus amblyrhynchos* is represented in red, *Carcharhinus limbatus* in green, *Carcharhinus melanopterus* in blue, and *Galeocerdo cuvier* in purple.

The first set of coalescent simulations was run under FIM and SST only (Table 2.2 and Supp. Table 2.4) to check if simulated data could reproduce the pattern of genetic variability (both θ estimators and *TD*) observed for *C. melanopterus*, *C.limbatus*, and *C. amblyrynchos*. The simulated θ values (excluding the equilibrium model) ranged between 0.001 and 0.003 per site, in line with the observed values (Tables 2.1 and 2.2). *TD* follows a U-shaped distribution for each *Nm* value as a function of T_{COL}, being more positive at recent T_{COL} and at equilibrium and less

positive (or negative for higher *Nm*) at intermediate values. Therefore, species demography with *Nm* ~10 (and higher) and T_{COL} within 15k and 50k generations B.P. will have negative *TD* values. In contrast, species with lower *Nm* and very recent or very ancient T_{COL} will have positive *TD*. This matches strikingly the *TD* observed for the three shark species and their estimated demographic parameters under SST (Table 2.1). We plot the normalized SFS and the *stairwayplot* for all scenarios presented in Table 2.2 (Figures 2.5, 2.6 & Supp. Figures 2.9, 2.10 and 2.11). First, we note that none of our scenarios, even those at equilibrium and with no variation in *Nm* through time, showed a normalized SFS compatible with a constant size population (Figures 2.5, 2.6 & Supp. Figures 2.9, 2.10 and 2.11). The normalized SFS and the reconstructed *stairwayplot* depend generally on the interaction between *Nm* and T_{COL} with a dynamic strikingly similar to *TD* (which is indeed a summary of the SFS). For *Nm=1* we observed the signature of a recent decrease in *Ne* for all scenarios and independently of T_{COL} (Figure 2.5). The normalized SFS showed consistently a strong deficit of low frequency variants, typical of a demographic bottleneck and in agreement with the positive *TD* (Figure 2.5 and Table 2.1). Furthermore, the *stairwayplot* could never detect the ancestral expansion for any T_{COL}. For growing *Nm,* the interplay with T_{COL} becomes more complex. A general result is that, once again, all scenarios were characterized by a recent decrease of *Ne* when looking at the *stairwayplot* and a deficit of singletons compared to the other low frequency classes when looking at the normalized SFS (Figure 2.6 & Supp. Figures 2.9 and 2.10). However, a strong signature of ancestral expansion appeared for *Nm* >10 and T_{COL} between 15k and 50k generations B.P., mirroring the results of *TD* for which most of these scenarios displayed a negative value. Remarkably, the *stairwayplot* retrieved the ancestral expansion only slightly overestimating the simulated T_{COL} (Figure 2.6 & Supp. Figures 2.9 and 2.10). Similar results were obtained for FIM (Supp. Figure 2.12 and 2.13).

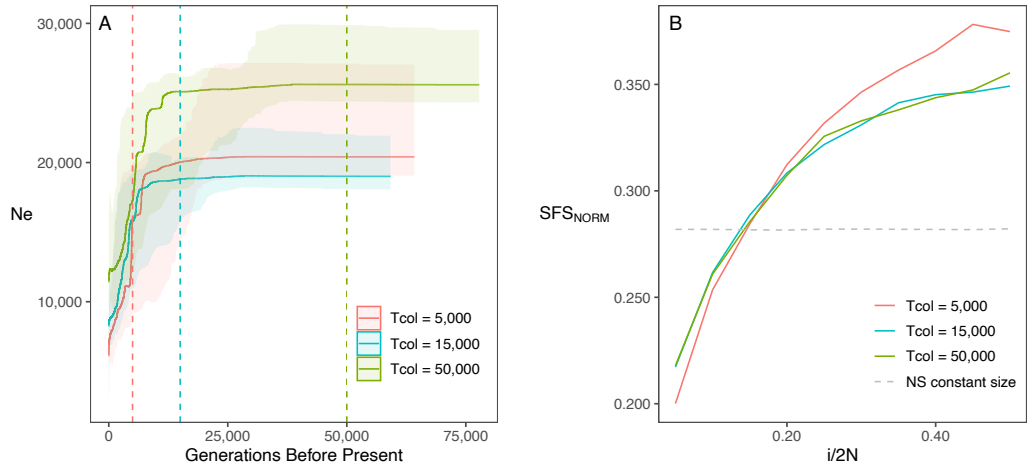

Figure 2.5. *stairwayplot* (maximum likelihood *Ne* and 75% confidence interval) (panel A) and normalized SFS (panel B) computed in simulated non-equilibrium SST scenarios with *Nm*=1, averaged over 100 replicates. Colonisation time of the array of deme T_{COL} occurred 5,000 (red), 15,000 (blue), and 50,000 (green) generations B.P., visually represented by the vertical dashed lines in panel A. The normalized SFS expected under a constant size non-structured model (NS constant size) is also shown (grey dashed line in panel B).

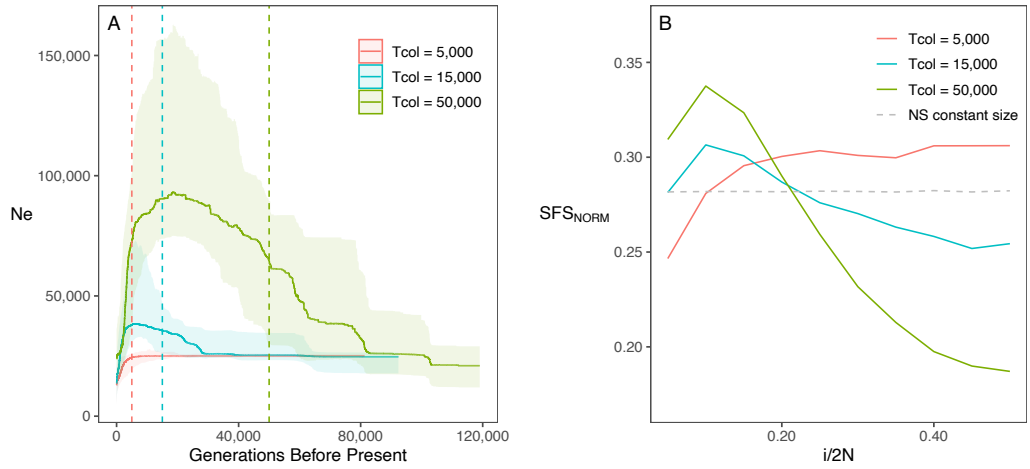

Figure 2.6. *stairwayplot* (maximum likelihood *Ne* and 75% confidence interval) (panel A) and normalized SFS (panel B) computed in simulated non-equilibrium SST scenarios with *Nm*=10, averaged over 100 replicates. Colonisation time of the array of deme T_{COL} occurred 5,000 (red), 15,000 (blue), and 50,000 (green) generations B.P., visually represented by the vertical dashed lines in panel A. The normalized SFS expected under a constant size non-structured model (NS constant size) is also shown (grey dashed line in panel B).

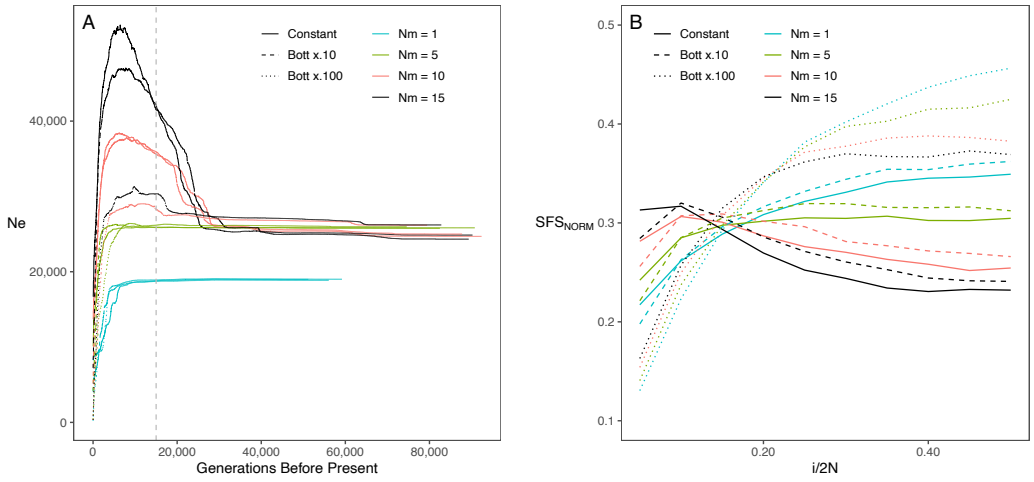

Figure 2.7. *stairwayplot* (maximum likelihood *Ne*) (panel A) and normalized SFS (panel B) computed in simulated non-equilibrium SST scenarios with T_{COL} =15,000 generations B.P. and an instantaneous decrease of the deme size (N_{DEME}) forward in time at T_{CH} =10 generations B.P. Colours represent the long-term connectivity values: *Nm*=1 (blue), *Nm*=5 (green), *Nm*=10 (red), *Nm*=15 (black). Line style represents the 10-fold (small dashes) or 100-fold (dots) reduction of N_{DEME}, or constant *Nm* (continuous line). The vertical grey dashed line in panel A represents the simulated colonisation time T_{COL}.

We compared SST vs SST-CH model (Figure 2.2) by means of the same ABC-RF model selection framework previously adopted. The two models cannot be clearly distinguished in any of the three structured species since: *i*) they showed similar posterior probability (~0.50); *ii*) the prior error rates are large ~0.40 (Supp. Table 2.11); *iii*) posterior distributions of *Nm* before and after T_{CH} are wide and largely overlapping (Supp. Table 2.12); *iv*) the normalized SFS closest to the observed data retrieved under the two models are very similar (Supp. Figure 2.14). We ran a second set of coalescent simulations focusing on the consequences of a recent change in connectivity on the observed SFS and the reconstructed *stairwayplot* (Supp. Tables 2.5 and 2.6). The decrease in connectivity was simulated by reducing either *m* (the migration rate per generation) or N_{DEME} (the effective population size of each deme). As expected, we found a signature of recent population decline in all simulated scenarios, with its intensity only slightly affected by the change in *Nm* (Fig. 2.7 & Supp. Figure 2.15, 2.16 and 2.17). However, the drop in N_{DEME} (Figure 2.7 and Supp. Figure 2.17) had larger effect compared to the drop in *m* (Supp. Figure 2.15 and 2.16) on both the normalized SFS and the expansion time estimated by the *stairwayplot*. In scenarios with 100x reduction in N_{DEME}, the *stairwayplot* could not retrieve the ancestral expansion even for large long-term *Nm* (Figure 2.7). FIM-CH models displayed a behaviour similar to SST-CH models but more pronounced (Supp. Figures 2.18, 2.19, 2.20 and 2.21, Supp. Tables 2.7 and 2.8). While at T_{CH} =

10 a decrease in *Nm* slightly affected the SFS and the reconstructed *stairwayplot*, the consequence of the change in connectivity are more substantial at $T_{CH} = 50$, with a stronger deficit in singletons and a more pronounced recent decline in *Ne* particularly in scenarios with a 100-fold reduction of N_{DEME} (Supp. Figures 2.19 and 2.21).

2.3.5. Discussion

2.3.5.1. Life history traits and demographic history of the four shark species

Discriminating whether the most appropriate model to reconstruct the demographic history of a species is *structured* or *unstructured* should be the first step in empirical population genetics investigations, particularly when targeting species of conservation concerns. Even when an extensive spatial sampling is lacking, an ABC model selection approach can actually distinguish whether the sampled deme belongs or not to a meta-population (similarly to previous studies (Maisano Delser et al., 2019; Peter et al., 2010)). Among the four species considered here, the tiger shark is the only panmictic. The three other species conversely are best described by the SST model, i.e., the sampled populations belong to a meta-population exchanging migrants following a stepping stone matrix. Our results reflect the tight link between the level of meta-population structure (or its absence) and life history traits. The panmictic *G. cuvier* unsurprisingly can accomplish transoceanic movements and has the largest body size among the sharks here considered (Supp. Table 2.3). In the three other species, the estimated number of migrants (*Nm*) remarkably follows the increase of movement range (Table 2.1 and Supp. Table 2.3) and it is consistent with their behaviour and habitat use. Indeed, *C. melanopterus*, a strongly lagoon dependent species, displays the lowest level of connectivity among the studied species (Table 2.1 and Supp. Table 2.3). These results bring meaningful hints about the influence of life history traits on population structure in sharks, but more studies addressing this topic will be needed to accurately detect which traits best predict its extent.

2.3.5.2. Gene genealogies in the four shark species and simulated scenarios

While it may seem counterintuitive to apply *unstructured* models to demes belonging to a meta-population, we further investigated the demographic history of the four species by means of the *stairwayplot*. When enough data is available, non-parametric *unstructured* models (such as the PSMC (H. Li & Durbin, 2011), the extended Bayesian skyline plot (Heled & Drummond, 2008)

and the *stairwayplot* among others) provide a careful description of the distribution of coalescence times of the gene genealogy, which ultimately depends from the "true" demographic history (whether it is known or not) of the sampled lineages. If panmixia is the most likely scenario, the distribution of coalescence times is directly related to the variation of *Ne* through time and can therefore have a direct biological interpretation. This is the case for *G. cuvier* (Table 2.1), whose reconstructed *stairwayplot* suggests that this species experienced a mild ancestral expansion and a recent ~4-fold bottleneck around 2,000 years B.P. (consistent with the results of (Pirog et al., 2019), Figure 2.4). Conversely, signals detected by the *stairwayplot* in the remaining three species, better described by the SST model (Table 2.1), cannot be directly interpreted as changes in *Ne* over time. In this light, we ran coalescence simulations to provide helpful and general insights into the understanding of the relation between the inferences performed under *unstructured* and *structured* models.

We first focus on scenarios simulated under the SST, with parameters close to those estimated in real data. The first and most striking result is that we systematically observed a recent bottleneck under all simulated scenarios (Table 2.2, Figures 2.5, 2.6, Supp. Figures 2.9, 2.10 and 2.11). This result could seem at a first glance surprising and due to an artefact. However, this is not the case, as: i) the signal does not depend on the inferential algorithm chosen to analyse the data (i.e., the *stairwayplot*), since the normalized spectra showed a deficit in singletons compared to the other low frequency classes (Figures 2.5, 2.6, Supp. Figures 2.9 and 2.10), which is typical of a recent population decline; ii) it is consistent with the distribution of the Inverse Instantaneous Coalescence Rate (IICR) computed in one diploid individual, which shows a signature of decline under similar meta-population models (Chikhi et al., 2018; Mazet et al., 2016; Rodríguez et al., 2018). The results of our simulations are consistent with the recent bottleneck observed in the three shark species (Figure 2.4), with its intensity inversely correlated to the estimated *Nm* (i.e., stronger for *C. melanopterus* and *C. limbatus* than for *C. amblyrhynchos*). In our SST model there is an instantaneous colonization of the array of demes at T_{COL}, which corresponds also to a demographic expansion (i.e., the total number of individuals in the array of deme is larger than those in the ancestral deme). However, this signature is detected only for $Nm \geq 5$ when T_{COL} is neither too recent nor too old (at equilibrium) (Figures 2.6, Supp. Figures 2.9, 2.10 and 2.11). In these scenarios, the beginning of the expansion retrieved by the *stairwayplot* broadly corresponds to the simulated T_{COL}. This again corroborates the results obtained for the three shark species, since the two species

with higher *Nm* displayed indeed an ancestral expansion in the *stairwayplot* with a timing consistent with the estimated T_{COL} (Table 2.1, Figures 2.3 and 2.4). Similarly, it explains why we could not retrieve the ancestral expansion for *C. melanopterus* nor estimate T_{COL} under the SST model: this appears to be a property of the coalescence pattern and it is not related to the amount of data available (see below).

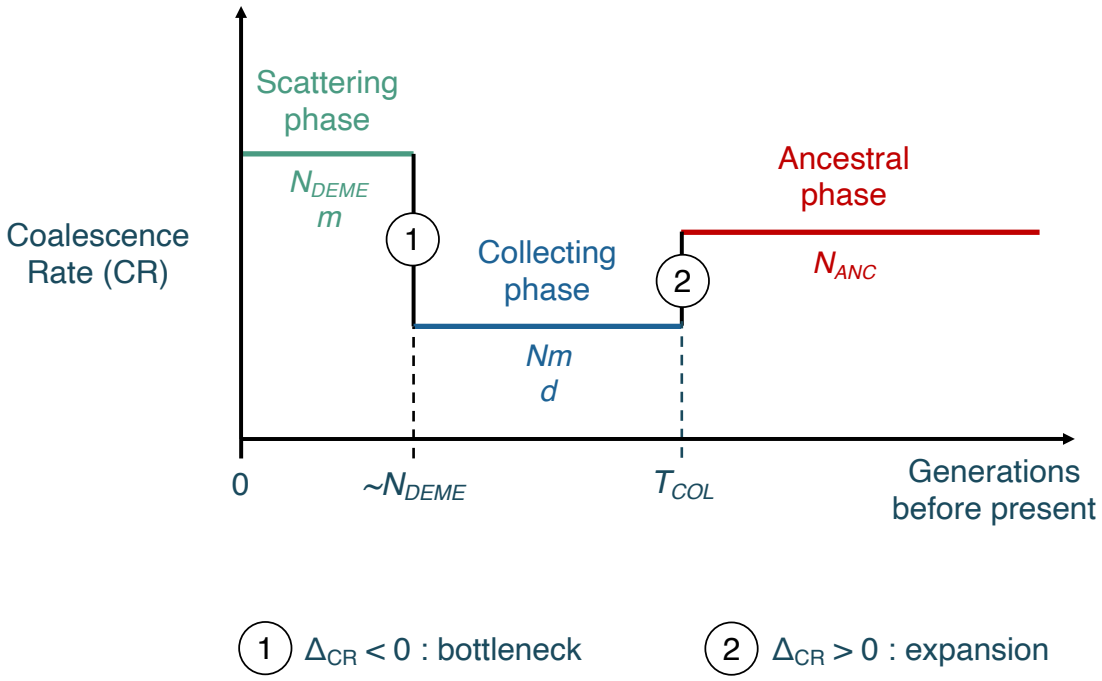

Figure 2.8. Schematic diagram representing the different coalescence phases in the history of lineages sampled from a deme belonging to a non-equilibrium meta-population. Each phase and related parameters are represented by a colour. Parameters influencing the coalescence rate in each phase are: the effective size of the deme (N_{DEME}) and the migration rate (m) for the *scattering* phase; the number of migrants exchanged per generation (Nm) and the number of demes (d) for the *collecting* phase; and the ancestral effective size (N_{ANC}) for the *ancestral* phase.

2.3.5.3. Coalescence phases in structured models

It is now straightforward to frame all these findings under the coalescence perspective. The coalescence history of the lineages sampled from a single deme in an SST (or FIM) model can be separated for simplicity into three phases: the *scattering*, the *collecting* and the *ancestral* phase (Figure 2.8). Going backward in time, lineages will coalesce in the sampled deme with a rate according to both Nm and N_{DEME} until all lineages either have coalesced or migrated to another

deme. This is the *scattering* phase described in the seminal works of Wakeley (1998, 1999). The *scattering* phase was considered instantaneous for mathematical tractability, with its outcome dependent on *Nm* only, but later works could disentangle the effect of N_{DEME} and *m* on the shape of the gene genealogy (Mona, 2017). The *collecting* phase starts when the lineages which did not coalesce have migrated to other demes of the array: they will then coalesce according to a Kingman process with a rate scaled by *Nm* and the number of demes *d* of the array (Wakeley, 1999) (Figure 2.8). Finally, all surviving lineages (in non-equilibrium model) will reach the ancestral deme at T_{COL}, where they will coalesce at a rate depending only on the N_{ANC} parameter (Figure 2.8). The interplay between the demographic parameters ($N_{DEME,}$ *Nm*, N_{ANC}, *d*) and the historical events (T_{COL} and T_{CH}) determines the length of each coalescence phase and the resulting shape of the gene genealogy of the sampled lineages (Figure 2.8).

In species with low *Nm*, the rate of coalescence during the *scattering* phase is very fast since lineages have low probability of emigrating from the sampled deme and high probability of coalescence due to the small *N*. Once all the lineages are dispersed in the array of demes, there will be two possible outcomes: i) in equilibrium model, we shift to the *collecting* phase, where the rate of coalescence drops since lineages will hardly fall in the same deme again; ii) in non-equilibrium model, with the parameters we have simulated here, there will be very few (if any) coalescence events during the collecting phase and the transition will be directly from the *scattering* to the *ancestral* phase. Both the *collecting* and the *ancestral* phases have a rate of coalescence lower than the *scattering* phase, which determines the observed recent drop in *Ne* for all simulated scenarios. Remarkably, the decline in *Ne* is much stronger in equilibrium model, since the rate of coalescence is much lower in the *collecting* than in the *ancestral* phase (Figures 2.5, 2.6, Supp. Figures 2.9, 2.10 and 2.11). Low *Nm* species will therefore have only two coalescence phases, the *scattering* and either the *collecting* (in equilibrium model) or the *ancestral* (in non-equilibrium model) which is why the signature of the ancestral expansion is lost.

For growing *Nm*, in equilibrium model there will be again only two coalescence phases, namely the *scattering* and *collecting*, with the latter having a lower rate of coalescence than the former independently of the simulated parameters. This is why we observed always a strong bottleneck consistent with the distribution of the IICR statistics in any equilibrium model (Chikhi et al., 2018; Mazet et al., 2015; Rodríguez et al., 2018). In non-equilibrium model, there will be two different situations: a) T_{COL} (in generations) is of the same order of the deme size N_{DEME}. In this setting,

going backward in time few lineages would have escaped the sampled demes before T_{COL}. This corresponds to a shift in the coalescence rate directly from the *scattering* to the *ancestral* phase, resulting in a bottleneck of lower intensity compared to an equilibrium model (Figures 2.5, 2.6, Supp. Figures 2.9 and 2.10), for the same reasons as above; b) T_{COL} (in generations) is larger than N_{DEME}. In this setting, some coalescence events may occur during the *collecting* phase, at a rate much slower than the two other phases. This determines the hump observed in the *stairwayplot* (Figures 2.5, 2.6, Supp. Figure 2.9 and 2.10) and explains why in this window of parameters it is also possible to correctly estimate T_{COL} using our ABC framework. Further simulations under the FIM model confirmed those patterns even though the ancestral expansion could be detected for lower long-term *Nm* than the corresponding SST scenario (Supp. Figure 2.12). This is probably due to a higher apparent connectivity underlined the by FIM, where lineages can move more freely during the collecting phase in comparison to SST where migrants only come from the closest neighbours. If many coalescence events occur during the *collecting* phase, the change in coalescence rate will affect the resulting gene genealogy and it will be detected by the *stairwayplot* (or any other *unstructured* method based on coalescent theory).

2.3.5.4. Changes in connectivity

Using coalescence arguments, we clarified why simple meta-population models with constant connectivity generate a gene genealogy harbouring a signature of a recent decline for any parameters' combination. The signature of bottleneck detected by the *stairwayplot* in the three shark species best described by SST can be therefore interpreted as a consequence of the underlying structure. However, connectivity likely changes through time. For instance, human activities have likely impacted the evolutionary history of a large number of species either by decreasing their effective population size and/or by fragmenting their habitat (i.e., reducing migration rates between demes). This intuitively should exacerbate the signature of population decline in the resulting gene genealogy. However, it remains to be shown whether this signature is qualitatively and quantitatively distinguishable from models with constant connectivity. This is a question of fundamental importance to understand whether it is possible to detect recent bottleneck in structured populations. To this end, we further investigated by coalescent simulations the expected gene genealogy in SST-CH (and FIM-CH) models with a change in connectivity 10 or 50 generations B.P., which matches the beginning of extensive anthropogenic influence on biodiversity considering our species' generation time (Ceballos et al., 2015). The resulting gene

genealogies were poorly affected by the recent drop in connectivity, with both the normalized SFS and the inferred *Ne* dynamic following the same trajectory of the corresponding scenario with the same long-term *Nm* and T_{COL} (Figure 2.7, Supp. Figures 2.15, 2.16, 2.17, 2.18, 2.19, 2.20 and 2.21). We noticed the drop in N_{DEME} (Figure 2.7, Supp. Figures 2.17, 2.20 and 2.21) had stronger influence than the drop in *m* (Supp. Figures 2.15, 2.16, 2.18 and 2.19), consistent with previous finding showing that the distribution of coalescence events depends not only by the *Nm* compound parameter but also by their individual values (Mona, 2017). This can be explained once again in the light of the length of the coalescence phases (Figure 2.8). Reducing N_{DEME} will increase exponentially the number of coalescence events, drastically shortening the *scattering* phase and the number of surviving lineages. Reducing *m* will only linearly reduce the probability of migrations outside the deme, marginally affecting the length of the *scattering* phase and the number of surviving lineages compared to constant *Nm* scenarios. This is why a 100-fold reduction in N_{DEME} significantly reduces the number of lineages entering in the *collecting* phase, almost hiding the ancestral expansion in high long-term *Nm* scenarios (Figure 2.7, Supp. Figures 2.17, 2.20 and 2.21), while a 100-fold reduction in *m* is barely detectable (Supp. Figures 2.15, 2.16, 2.18 and 2.19). Similarly, the recent reduction in either N_{DEME} or *m* cannot be detected for lower long-term *Nm* scenarios, where the *collecting* phase is already missing. This explains why the general pattern is strikingly similar between SST-CH and SST simulations, which implies that the simulated change in connectivity is too recent to significantly alter the pattern of coalescence events and that a recent drop can be hardly detected on the basis of the SFS only. Our empirical data are consistent with these findings: when we compared SST vs. SST-CH models in the three shark species using the ABC framework, we failed to clearly distinguish the two models (Supp. Tables 2.11 and 2.12, Supp. Figure 2.14). This seems to be a paradox: we observed a recent bottleneck in species of conservation concern using *unstructured* model, but we cannot exclude that this is just the consequence of population structure.

2.3.5.5. Practical recommendations and conservation concerns

This study highlights once more the importance to explicitly test for meta-population structure before interpreting the demographic signals detected by *unstructured* models, similarly to what advocated previously by (Maisano Delser et al., 2019; Rodríguez et al., 2018). If the meta-population structure hypothesis is rejected, the variation of *Ne* through time can be directly interpreted as the demographic history of the population under investigation, such as the case of

tiger shark. Otherwise, this variation is still related to demographic events, but it has to be explained in the light of population structure and its consequence on the rate of coalescence events. We showed by coalescent simulations how to interpret such variation: the recent bottleneck detected by the *stairwayplot* in demes belonging to a meta-population is a consequence of the coalescence process. In other words, any inferential method implementing an *unstructured* model will detect such decline (if enough data is available) since it is a property of the gene genealogy. Importantly, the gene genealogy is only slightly affected by recent changes in connectivity if the time of this change in generations is of the same order of the size of the deme.

Our study underscores a key issue in conservation genetics as a recent decline inferred by an *unstructured* model can be mis-interpreted as a consequence of recent anthropic pressures (Ceballos et al., 2015) when it actually results from meta-population structure. This is all the more alarming since the majority of species is likely organised in meta-populations across their range rather than panmictic at a large scale. We therefore stress the necessity for an educated choice of tools to correctly uncover the recent trend of a species and design proper conservation programs. For instance, detecting a recent bottleneck in meta-populations will require summary statistics measuring the linkage disequilibrium (Boitard et al., 2016; Kerdoncuff et al., 2020) and/or the inferential framework based on the IICR (Chikhi et al., 2018; Rodríguez et al., 2018) coupled with whole genome data. On a positive note, we showed that the colonization time of the array of demes can be estimated to some extent (and under some combinations of parameters) by *unstructured* models. We believe that this is particularly important because it has been shown that the simple instantaneous colonization process we used here behaves similarly to a spatial explicit range expansion (Hamilton et al., 2005; Mona, 2017), which is certainly a more realistic model but more difficult to investigate. We are aware that the meta-population models here tested are simple and the parameters chosen are specific of the three shark species we focused on. Nevertheless, the time-scale separation of the coalescence process is general, and it allows explaining intuitively any structured models. The four shark species here used as an example has the merit to cover a large spectrum of LHT and consequently a large spectrum of demographic scenarios, going from a highly structured to a panmictic population: this has strong implications on the distribution of coalescence times and therefore on the interpretation of the observed data.

2.3.5.6. Conclusion

In this study we found that population structure, independently from the degree of connectivity between demes and the migration matrix relating them, intrinsically determines a variation in the rate of coalescence events through time. We showed that the intensity and the direction(s) of such variation related to the demographic parameters of the meta-population in a predictable way. Our results highlight the importance of detecting population structure (which depends on LHT among other factors) before performing any demographic inferences but, at the same time, they reveal the utility of *unstructured* models to describe the shape of the gene genealogy, which is the final product of the evolutionary history of a species. A combination of structured and *unstructured* models (better if non-parametric) is therefore the key to best characterize the evolutionary history of a species. We call for a change in perspective when investigating the demographic history of a species: the focus should be put in the reconstruction of the variation of both N and m through time, which requires certainly new methodological development and probably more data.

2.3.6. Supplementary information

2.3.6.1. Acknowledgement

We are grateful to the Genotoul bioinformatics platform Toulouse Midi-Pyrenees (Bioinfo Genotoul; http://bioinfo.genotoul.fr/) for providing computing resources. We are indebted to Oscar Lao for fruitful discussions and careful reading of the manuscript. This work was supported by two ATM grants (2016 and 2017) from the Muséum National d'Histoire Naturelle to S.M.

2.3.6.2. Data availability statement

Fastq sequence files, SFS and scripts are available from the Dryad Digital Repository: https://doi.org/10.5061/dryad.b8gtht7d1.

2.3.6.3. Authors contribution

S.M. and P.L. conceived the project. S.P. provided reagents and samples. S.M. and P.L analysed the data and wrote the manuscript with input from S.P.

2.3.6.4. Supplementary Methods

Calibrating the molecular clock of an orthologous genomic region is traditionally troublesome, requiring either fossil data, serial sampling, or available pedigree genealogy. Even when massive data are available, which is the case for some model species like humans, a consensus rarely emerge either because of different statistical methods used to infer the rate of evolution or because of the inherent time-dependency of the molecular clock, whose estimate depends on the time window selected for the calibration process ((Ho et al., 2015) and references therein). Calibrating genome wide data on non-model species is therefore by no means a simple task and it necessarily comes with some approximations. We previously estimated the molecular clock of an exon-capture dataset in *C. melanopterus* using fossil time calibration (Maisano Delser et al., 2016). However, exon-capture data may be biased towards conserved regions while Rad-seq should target the genome more randomly. To obtain a more reliable genome wide estimate of the molecular clock for the three new species under investigation (*G. cuvier*, *C. limbatus* and *C. amblyrhynchos*), we selected 12 individuals of *C. melanopterus* from Moo'rea (French Polynesia) for which exon-capture data were available (Maisano Delser et al., 2019) and sequenced them under the same ddRADseq protocol used in this study. We followed the same bioinformatics pipeline, combining

STACKS 2.5 then ANGSD, and the same filters used for the other three species to compute the SFS and the genetic diversity of the Moo'rea population. As expected, $\theta_{\pi\text{-CAPTURE}}$ was lower than $\theta_{\pi\text{-RAD}}$, but reassuringly, the SFSs computed from exon capture and Rad-seq were highly similar (once standardized) and gave the same signature of recent bottleneck using the *stairwayplot*. We finally compared $\theta_{\pi\text{-RAD}}$ and $\theta_{\pi\text{-CAPTURE}}$ to derive the mutation rate for the ddRad data: we found a value of 1.93e-8 per site per generation (the exon capture having a rate of 8.4e-9 per site per generation (Maisano Delser et al., 2016)). which we used in the demographic analyses for the Rad data of all other species.

2.3.6.5. Supplementary Tables

Table 2.3. Life history traits and FST values in the Indo-Pacific of the four species studied.

	Size (cm)	Behavioural traits	Movement range	Habitat	F_{ST}	Comments
C. melanopterus	131-134	Philopatry, reef fidelity	Low daily activity (~10km). Max dispersal of 50km	Lagoon	0.64	F_{ST} computed between the East Indian Ocean (Western Australia) and the western Pacific Ocean (New-Caledonia) with genomic data (Exon Capture)
C. amblyrhynchos	~190	Philopatry, residency	Large scale (Up to 700km)	Fringing and barrier reef	~0.4	F_{ST} computed between the central Indian Ocean (Chagos) and the western Pacific Ocean (Eastern Australia) with genomic data (RADseq).
C. limbatus	226-255	Philopatry, massive seasonal aggregations (in the Atlantic Ocean)	Large scale, >190km in Caribe	Fringing and barrier reef	NA	No data available in the Indo-Pacific region for structure.
G. cuvier	370-430	Opportunistic	Transoceanic (1,000 km)	Ocean and coast	~0.001	F_{ST} computed between the east Indian Ocean (Western Australia) and central Pacific Ocean (Hawaii) with microsatellite data.

References: (Almojil et al., 2018; Barnett et al., 2012; Boissin et al., 2019; Bonnin et al., 2019; Compagno, 2001; Espinoza et al., 2014; Field et al., 2011; Holmes et al., 2017; Kajiura & Tellman, 2016; Keeney et al., 2003, 2005; Lea et al., 2015; Maisano Delser et al., 2019; Meyer et al., 2018; Momigliano et al., 2015, 2017; Mourier et al., 2012, 2013; Swinsburg et al., 2012; Vignaud et al., 2014; Werry et al., 2014)

Table 2.4. Coalescent simulations of 50,000 Rad-loci under FIM model, with mutation rate fixed to 1.93*10-8 per site per generation and NANC fixed to 50,000. Mean pairwise difference (θπ), Watterson theta (θw), Tajima's D (TD), and number of segregating sites (S) are averaged over 100 replicates.

Nm	T_{COL}	θ_π ‡	θ_s ‡	**TD**	**S**
1	5000	0.0015	0.0013	0.4101	27109
	15000	0.0017	0.0016	0.2367	31946
	50000	0.0025	0.0024	0.0790	49142
	∞†	0.0182	0.0159	0.5887	325096
5	5000	0.0019	0.0019	0.0572	38367
	15000	0.0022	0.0024	-0.2953	48762
	50000	0.0032	0.0039	-0.6713	78719
	∞	0.0183	0.0166	0.4099	339181
10	5000	0.0020	0.0021	-0.1345	41929
	15000	0.0023	0.0027	-0.5653	54552
	50000	0.0034	0.0044	-0.9542	89159
	∞	0.0182	0.0171	0.2677	349655
15	5000	0.0020	0.0021	-0.2405	43743
	15000	0.0024	0.0028	-0.6828	57270
	50000	0.0034	0.0046	-1.0720	94050
	∞	0.0182	0.0174	0.2044	354988

†Equilibrium model obtained by simulating T_{COL}=∞

‡Theta values are expressed per site per generation.

Table 2.5. Coalescent simulations of 50,000 Rad-loci under SST-CH model with reduction in m at T_{CH}, with mutation rate fixed to 1.93*10-8 per site per generation and N_{ANC} fixed to 50,000. Mean pairwise difference ($\theta\pi$), Watterson theta (θw), Tajima's D (TD), and number of segregating sites (S) are averaged over 100 replicates.

Nm	T_{CH}[§]	*m reduction*	T_{COL}[§]	θ_π[‡]	θ_s[‡]	TD	S
1	10	10x	5000	0.00130	0.00115	0.53905	23507
	10	10x	15000	0.00130	0.00118	0.41408	24139
	10	10x	50000	0.00173	0.00157	0.42420	32028
	10	10x	∞[†]	0.01606	0.01384	0.67526	282257
5	10	10x	5000	0.00172	0.00159	0.36874	32346
	10	10x	15000	0.00193	0.00184	0.20181	37597
	10	10x	50000	0.00279	0.00276	0.04584	56260
	10	10x	∞	0.01767	0.01497	0.75639	305365
10	10	10x	5000	0.00186	0.00178	0.18989	36356
	10	10x	15000	0.00213	0.00217	-0.07747	44263
	10	10x	50000	0.00310	0.00338	-0.35253	68998
	10	10x	∞	0.01797	0.01573	0.59693	320875
15	10	10x	5000	0.00192	0.00189	0.05422	38640
	10	10x	15000	0.00222	0.00237	-0.26258	48245
	10	10x	50000	0.00324	0.00376	-0.58824	76783
	10	10x	∞	0.01803	0.01620	0.47420	330517
1	10	100x	5000	0.00130	0.00115	0.53657	23536
	10	100x	15000	0.00129	0.00117	0.42174	23916
	10	100x	50000	0.00173	0.00157	0.42828	32102
	10	100x	∞[a]	0.01608	0.01386	0.67397	282641
5	10	100x	5000	0.00172	0.00158	0.37189	32277
	10	100x	15000	0.00194	0.00184	0.20593	37635
	10	100x	50000	0.00278	0.00275	0.04378	56196
	10	100x	∞	0.01769	0.01497	0.76118	305417
10	10	100x	5000	0.00186	0.00178	0.19072	36225
	10	100x	15000	0.00212	0.00216	-0.07871	44126
	10	100x	50000	0.00310	0.00339	-0.35126	69067
	10	100x	∞	0.01797	0.01573	0.59577	320982
15	10	100x	5000	0.00192	0.00189	0.06721	38606
	10	100x	15000	0.00221	0.00235	-0.25984	48020
	10	100x	50000	0.00323	0.00376	-0.59422	76691
	10	100x	∞	0.01805	0.01621	0.47736	330705
1	50	10x	5000	0.00130	0.00115	0.55332	23450
	50	10x	15000	0.00129	0.00118	0.42353	23984
	50	10x	50000	0.00173	0.00157	0.42434	32120
	50	10x	∞[a]	0.01607	0.01385	0.67164	282543

Nm	T_{CH}[§]	*m reduction*	T_{COL}[§]	θ_π[‡]	θ_s[‡]	TD	S
5	50	10x	5000	0.00172	0.00159	0.36487	32350
	50	10x	15000	0.00193	0.00184	0.21241	37524
	50	10x	50000	0.00278	0.00275	0.04190	56142
	50	10x	∞	0.01773	0.01501	0.76033	306174
10	50	10x	5000	0.00186	0.00178	0.18804	36308
	50	10x	15000	0.00213	0.00217	-0.07766	44288
	50	10x	50000	0.00310	0.00338	-0.34965	68967
	50	10x	∞	0.01797	0.01574	0.59500	321111
15	50	10x	5000	0.00192	0.00189	0.05718	38619
	50	10x	15000	0.00221	0.00236	-0.26552	48214
	50	10x	50000	0.00324	0.00377	-0.58699	76896
	50	10x	∞	0.01803	0.01619	0.47549	330357
1	50	100x	5000	0.00130	0.00115	0.54173	23525
	50	100x	15000	0.00130	0.00118	0.42190	24085
	50	100x	50000	0.00173	0.00157	0.41987	32087
	50	100x	∞[a]	0.01606	0.01386	0.66660	282769
5	50	100x	5000	0.00173	0.00159	0.37725	32421
	50	100x	15000	0.00194	0.00184	0.20788	37634
	50	100x	50000	0.00278	0.00276	0.03982	56253
	50	100x	∞	0.01765	0.01496	0.75395	305150
10	50	100x	5000	0.00186	0.00178	0.18310	36283
	50	100x	15000	0.00212	0.00216	-0.07626	44098
	50	100x	50000	0.00311	0.00339	-0.34704	69105
	50	100x	∞	0.01797	0.01575	0.59198	321266
15	50	100x	5000	0.00192	0.00189	0.06383	38637
	50	100x	15000	0.00221	0.00236	-0.26121	48105
	50	100x	50000	0.00324	0.00376	-0.58707	76768
	50	100x	∞	0.01804	0.01622	0.47210	330871

[†]Equilibrium model obtained by simulating $T_{COL}=\infty$.
[‡]Theta values are expressed per site per generation.
[§]Time parameters are in generations.

Table 2.6. Coalescent simulations of 50,000 Rad-loci under SST-CH model with reduction in N_{DEME} at T_{CH}, with mutation rate fixed to 1.93*10-8 per site per generation and N_{ANC} fixed to 50,000. Mean pairwise difference (θ_π), Watterson theta (θ_w), Tajima's D (TD), and number of segregating sites (S) are averaged over 100 replicates.

Nm	T_{CH}[§]	*N_{DEME} reduction*	T_{COL}[§]	θ_π[‡]	θ_s[‡]	TD	S
1	10	10x	5000	0.00128	0.00111	0.61198	22721
	10	10x	15000	0.00128	0.00114	0.50115	23258
	10	10x	50000	0.00170	0.00152	0.49629	31045
	10	10x	∞[a]	0.01582	0.01348	0.72556	275086
5	10	10x	5000	0.00170	0.00153	0.44718	31276
	10	10x	15000	0.00190	0.00178	0.29521	36260
	10	10x	50000	0.00274	0.00265	0.15186	54018
	10	10x	∞	0.01743	0.01463	0.80213	298453
10	10	10x	5000	0.00183	0.00172	0.28488	35053
	10	10x	15000	0.00210	0.00208	0.03809	42418
	10	10x	50000	0.00305	0.00323	-0.22768	65876
	10	10x	∞	0.01769	0.01531	0.65256	312263
15	10	10x	5000	0.00189	0.00182	0.15950	37096
	10	10x	15000	0.00218	0.00225	-0.14263	46000
	10	10x	50000	0.00319	0.00358	-0.45708	72982
	10	10x	∞	0.01777	0.01573	0.54436	320951
1	10	100x	5000	0.00108	0.00086	1.02670	17632
	10	100x	15000	0.00107	0.00086	0.98420	17620
	10	100x	50000	0.00143	0.00116	0.98726	23612
	10	100x	∞[a]	0.01324	0.01056	1.06626	215457
5	10	100x	5000	0.00142	0.00116	0.95740	23630
	10	100x	15000	0.00159	0.00132	0.87800	26889
	10	100x	50000	0.00230	0.00194	0.79873	39479
	10	100x	∞	0.01470	0.01166	1.09248	237887
10	10	100x	5000	0.00153	0.00127	0.85589	25849
	10	100x	15000	0.00176	0.00149	0.75054	30434
	10	100x	50000	0.00257	0.00225	0.60559	45805
	10	100x	∞	0.01491	0.01199	1.02366	244582
15	10	100x	5000	0.00159	0.00133	0.80794	27164
	10	100x	15000	0.00183	0.00158	0.66845	32239
	10	100x	50000	0.00268	0.00240	0.48462	49042
	10	100x	∞	0.01503	0.01221	0.97041	249060
1	50	10x	5000	0.00119	0.00100	0.81029	20359
	50	10x	15000	0.00118	0.00101	0.73048	20538
	50	10x	50000	0.00159	0.00135	0.73448	27585
	50	10x	∞[a]	0.01479	0.01220	0.89094	248820
5	50	10x	5000	0.00159	0.00137	0.68787	27936

Nm	T_{CH}[§]	N_{DEME} *reduction*	T_{COL}[§]	θ_π[‡]	θ_s[‡]	TD	S
	50	10x	15000	0.00178	0.00156	0.57630	31893
	50	10x	50000	0.00256	0.00232	0.44223	47240
	50	10x	∞	0.01631	0.01334	0.93703	272031
10	50	10x	5000	0.00172	0.00152	0.55186	30996
	50	10x	15000	0.00196	0.00181	0.36396	36842
	50	10x	50000	0.00287	0.00278	0.13690	56656
	50	10x	∞	0.01661	0.01392	0.81114	283886
15	50	10x	5000	0.00178	0.00161	0.44742	32826
	50	10x	15000	0.00205	0.00195	0.21416	39816
	50	10x	50000	0.00300	0.00304	-0.05967	62031
	50	10x	∞	0.01673	0.01427	0.72334	291199
1	50	100x	5000	0.00049	0.00038	1.21625	7783
	50	100x	15000	0.00049	0.00038	1.20808	7819
	50	100x	50000	0.00066	0.00051	1.22692	10383
	50	100x	∞[a]	0.00613	0.00475	1.21477	96923
5	50	100x	5000	0.00068	0.00054	1.11300	10924
	50	100x	15000	0.00077	0.00061	1.12791	12383
	50	100x	50000	0.00110	0.00087	1.11920	17706
	50	100x	∞	0.00711	0.00559	1.14274	113933
10	50	100x	5000	0.00076	0.00061	1.04387	12465
	50	100x	15000	0.00088	0.00070	1.05086	14317
	50	100x	50000	0.00128	0.00103	1.02564	20994
	50	100x	∞	0.00749	0.00599	1.05312	122161
15	50	100x	5000	0.00082	0.00067	0.97960	13622
	50	100x	15000	0.00095	0.00078	0.96079	15818
	50	100x	50000	0.00140	0.00114	0.95698	23217
	50	100x	∞	0.00786	0.00636	0.98999	129717

[†]Equilibrium model obtained by simulating $T_{COL}=\infty$.
[‡]Theta values are expressed per site per generation.
[§]Time parameters are in generations.

Table 2.7. Coalescent simulations of 50,000 Rad-loci under FIM-CH model with reduction in m at T_{CH}, with mutation rate fixed to 1.93*10-8 per site per generation and N_{ANC} fixed to 50,000. Mean pairwise difference ($\theta\pi$), Watterson theta (θw), Tajima's D (TD), and number of segregating sites (S) are averaged over 100 replicates.

Nm	T_{CH}[§]	*m reduction*	T_{COL}[§]	θ_π[‡]	θ_s[‡]	TD	S
1	10	10x	5000	0.00145	0.00132	0.41122	26929
	10	10x	15000	0.00166	0.00156	0.25078	31871
	10	10x	50000	0.00244	0.00239	0.08899	48841
	10	10x	∞[a]	0.01817	0.01591	0.59801	324456
5	10	10x	5000	0.00190	0.00187	0.06930	38087
	10	10x	15000	0.00221	0.00238	-0.28671	48452
	10	10x	50000	0.00324	0.00384	-0.65235	78360
	10	10x	∞	0.01823	0.01658	0.41710	338240
10	10	10x	5000	0.00198	0.00205	-0.12507	41736
	10	10x	15000	0.00231	0.00265	-0.54503	54149
	10	10x	50000	0.00338	0.00434	-0.92946	88555
	10	10x	∞	0.01823	0.01707	0.28361	348274
15	10	10x	5000	0.00202	0.00213	-0.22220	43454
	10	10x	15000	0.00235	0.00278	-0.65669	56766
	10	10x	50000	0.00342	0.00457	-1.05578	93317
	10	10x	∞	0.01822	0.01732	0.21787	353346
1	10	100x	5000	0.00145	0.00132	0.41503	26965
	10	100x	15000	0.00165	0.00156	0.24537	31812
	10	100x	50000	0.00244	0.00239	0.08486	48785
	10	100x	∞[a]	0.01816	0.01591	0.59565	324464
5	10	100x	5000	0.00190	0.00187	0.06944	38068
	10	100x	15000	0.00221	0.00237	-0.28743	48317
	10	100x	50000	0.00325	0.00383	-0.64425	78220
	10	100x	∞	0.01822	0.01657	0.41858	337978
10	10	100x	5000	0.00198	0.00204	-0.12793	41694
	10	100x	15000	0.00231	0.00265	-0.54354	54150
	10	100x	50000	0.00338	0.00434	-0.92977	88534
	10	100x	∞	0.01820	0.01706	0.27953	348109
15	10	100x	5000	0.00201	0.00213	-0.21978	43368
	10	100x	15000	0.00235	0.00279	-0.65661	56833
	10	100x	50000	0.00342	0.00457	-1.05168	93215
	10	100x	∞	0.01820	0.01732	0.21491	353254
1	50	10x	5000	0.00145	0.00131	0.44880	26633
	50	10x	15000	0.00164	0.00154	0.28906	31341
	50	10x	50000	0.00243	0.00236	0.13199	48136
	50	10x	∞[a]	0.01805	0.01570	0.62619	320363

Nm	T_{CH}[§]	*m reduction*	T_{COL}[§]	θ_π[‡]	θ_s[‡]	TD	S
5	50	10x	5000	0.00189	0.00184	0.11281	37513
	50	10x	15000	0.00220	0.00233	-0.22866	47452
	50	10x	50000	0.00323	0.00375	-0.58620	76506
	50	10x	∞	0.01811	0.01635	0.45215	333456
10	50	10x	5000	0.00198	0.00201	-0.07296	41023
	50	10x	15000	0.00230	0.00259	-0.47505	52808
	50	10x	50000	0.00335	0.00422	-0.86461	86112
	50	10x	∞	0.01813	0.01684	0.32349	343461
15	50	10x	5000	0.00201	0.00209	-0.15797	42666
	50	10x	15000	0.00233	0.00271	-0.59177	55361
	50	10x	50000	0.00341	0.00445	-0.98262	90704
	50	10x	∞	0.01811	0.01705	0.26006	347796
1	50	100x	5000	0.00144	0.00130	0.43650	26592
	50	100x	15000	0.00164	0.00154	0.29003	31319
	50	100x	50000	0.00243	0.00235	0.13260	48020
	50	100x	∞[a]	0.01802	0.01567	0.62890	319726
5	50	100x	5000	0.00190	0.00184	0.12343	37581
	50	100x	15000	0.00219	0.00232	-0.22107	47233
	50	100x	50000	0.00322	0.00373	-0.57931	76116
	50	100x	∞	0.01811	0.01633	0.45961	333022
10	50	100x	5000	0.00197	0.00200	-0.06106	40854
	50	100x	15000	0.00229	0.00258	-0.47116	52651
	50	100x	50000	0.00336	0.00421	-0.85074	85880
	50	100x	∞	0.01808	0.01677	0.32752	342129
15	50	100x	5000	0.00200	0.00208	-0.15069	42413
	50	100x	15000	0.00233	0.00270	-0.57952	55113
	50	100x	50000	0.00340	0.00442	-0.96552	90154
	50	100x	∞	0.01809	0.01703	0.26117	347426

[†]Equilibrium model obtained by simulating $T_{COL}=\infty$.
[‡]Theta values are expressed per site per generation.
[§]Time parameters are in generations.

Table 2.8. Coalescent simulations of 50,000 Rad-loci under FIM-CH model with reduction in N_{DEME} at T_{CH}, with mutation rate fixed to 1.93*10-8 per site per generation and N_{ANC} fixed to 50,000. Mean pairwise difference (θπ), Watterson theta (θw), Tajima's D (TD), and number of segregating sites (S) are averaged over 100 replicates.

Nm	T_{CH}[§]	N_{DEME} *reduction*	T_{COL}[§]	θ_π[‡]	θ_s[‡]	TD	S
1	10	10x	5000	0.00143	0.00128	0.48788	26140
	10	10x	15000	0.00163	0.00151	0.34517	30744
	10	10x	50000	0.00241	0.00230	0.19613	46957
	10	10x	∞[a]	0.01787	0.01544	0.65988	315055
5	10	10x	5000	0.00187	0.00179	0.18543	36583
	10	10x	15000	0.00217	0.00225	-0.16018	45973
	10	10x	50000	0.00319	0.00363	-0.50787	74054
	10	10x	∞	0.01796	0.01607	0.49162	327893
10	10	10x	5000	0.00195	0.00195	-0.00206	39880
	10	10x	15000	0.00228	0.00252	-0.39809	51348
	10	10x	50000	0.00333	0.00408	-0.77863	83301
	10	10x	∞	0.01796	0.01651	0.36849	336717
15	10	10x	5000	0.00199	0.00204	-0.09059	41541
	10	10x	15000	0.00231	0.00262	-0.50877	53521
	10	10x	50000	0.00338	0.00429	-0.89274	87572
	10	10x	∞	0.01797	0.01675	0.30452	341733
1	10	100x	5000	0.00119	0.00096	0.97814	19671
	10	100x	15000	0.00136	0.00112	0.91054	22777
	10	100x	50000	0.00202	0.00168	0.84075	34271
	10	100x	∞[a]	0.01505	0.01206	1.04118	245965
5	10	100x	5000	0.00157	0.00131	0.82823	26722
	10	100x	15000	0.00183	0.00158	0.67710	32148
	10	100x	50000	0.00268	0.00240	0.49106	49021
	10	100x	∞	0.01518	0.01234	0.96438	251826
10	10	100x	5000	0.00164	0.00139	0.74837	28333
	10	100x	15000	0.00191	0.00169	0.55675	34402
	10	100x	50000	0.00280	0.00260	0.33385	52989
	10	100x	∞	0.01520	0.01252	0.90118	255333
15	10	100x	5000	0.00168	0.00144	0.69787	29377
	10	100x	15000	0.00194	0.00173	0.50432	35362
	10	100x	50000	0.00285	0.00268	0.25734	54720
	10	100x	∞	0.01524	0.01265	0.86148	257979
1	50	10x	5000	0.00132	0.00113	0.72479	23022
	50	10x	15000	0.00152	0.00132	0.62679	26915
	50	10x	50000	0.00225	0.00200	0.51158	40859

Nm	T_{CH}[§]	N_{DEME} *reduction*	T_{COL}[§]	θ_π[‡]	θ_s[‡]	TD	S
	50	10x	∞[a]	0.01672	0.01392	0.84635	283891
5	50	10x	5000	0.00175	0.00157	0.48096	32080
	50	10x	15000	0.00203	0.00193	0.23401	39292
	50	10x	50000	0.00299	0.00303	-0.06211	61867
	50	10x	∞	0.01683	0.01441	0.70484	293983
10	50	10x	5000	0.00184	0.00171	0.33290	34848
	50	10x	15000	0.00214	0.00212	0.02648	43298
	50	10x	50000	0.00313	0.00338	-0.30279	68856
	50	10x	∞	0.01694	0.01482	0.60062	302349
15	50	10x	5000	0.00187	0.00177	0.24425	36122
	50	10x	15000	0.00217	0.00221	-0.08514	45184
	50	10x	50000	0.00319	0.00355	-0.42044	72349
	50	10x	∞	0.01699	0.01506	0.53810	307130
1	50	100x	5000	0.00055	0.00043	1.20585	8754
	50	100x	15000	0.00063	0.00049	1.20710	10022
	50	100x	50000	0.00093	0.00072	1.20537	14714
	50	100x	∞[a]	0.00698	0.00543	1.20041	110718
5	50	100x	5000	0.00076	0.00060	1.09153	12331
	50	100x	15000	0.00089	0.00071	1.07177	14400
	50	100x	50000	0.00130	0.00104	1.06655	21220
	50	100x	∞	0.00748	0.00593	1.09777	120969
10	50	100x	5000	0.00084	0.00068	0.97258	13964
	50	100x	15000	0.00099	0.00080	0.96360	16352
	50	100x	50000	0.00144	0.00118	0.93072	24078
	50	100x	∞	0.00789	0.00640	0.98087	130542
15	50	100x	5000	0.00090	0.00074	0.88569	15159
	50	100x	15000	0.00105	0.00087	0.86287	17746
	50	100x	50000	0.00154	0.00128	0.83388	26196
	50	100x	∞	0.00832	0.00684	0.90923	139481

[†]Equilibrium model obtained by simulating $T_{COL}=\infty$.
[‡]Theta values are expressed per site per generation.
[§]Time parameters are in generations.

Table 2.9. Confusion matrix of the model selection procedure: rows indicate the simulated models and columns the votes (in %) attributed by the ABC-RF algorithm to each of them. Classification error and prior error rate are based on the first of the two runs.

		Attributed votes (%)			Class. error	Prior error rate	Run N°1	Run N°2
		FIM	NS	SST				
G. cuvier	FIM	85.6	3.2	11.2	0.14	0.1	NS : 0.84	NS : 0.85
	NS	1.1	97.8	1.1	0.02			
	SST	12.5	2.3	85.2	0.15			
C. amblyrhynchos	FIM	88.6	2.8	8.6	0.11	0.08	SST : 0.85	SST : 0.85
	NS	1.1	98.0	0.9	0.02			
	SST	10.2	1.7	88.1	0.12			
C. limbatus	FIM	88.6	3.0	8.4	0.11	0.08	SST : 0.55	SST : 0.59
	NS	1.0	97.9	1.0	0.02			
	SST	9.8	1.7	88.5	0.12			
C. melanopterus	FIM	68.6	6.3	25.0	0.31	0.23	SST : 0.89	SST : 0.89
	NS	2.0	96.2	1.8	0.04			
	SST	27.1	6.1	66.8	0.33			

Table 2.10. Cross Validation of parameter estimation based on the first run of random forests. Mean Squared Error (SME), Mean Root Squared Error (SRMSE) and 95% coverage of the median value of each parameter computed on 999 pseudo-observed datasets (pods) simulated under the SST model.

		Nm	T_{COL}	N_{ANC}
C. amblyrhynchos (N=12)	SME	0.0042	0.0206	0.0698
	SRMSE	0.0445	0.2362	0.8394
	Coverage	0.994	0.993	0.994
C. limbatus (N=13)	SME	0.0014	0.0442	0.1042
	SRMSE	0.0235	0.7498	0.7655
	Coverage	0.996	0.994	0.993
C. melanopterus (N=8)	SME	0.0102	0.0713	0.0676
	SRMSE	0.0759	1.2819	0.4855
	Coverage	0.998	0.996	0.997

Table 2.11. Confusion matrix of the model selection procedure: rows indicate the simulated models and columns the votes (in %) attributed by the ABC-RF algorithm to each of them. Classification error and prior error rate are based on the first of the two runs.

		Attributed votes (%)		Class Error	Prior error rate	Run N°1	Run N°2
		SST	SST-CH				
C. amblyrhynchos	SST	59.3	40.7	0.41	0.40	SST: 0.60	SST: 0.57
	SST-CHGT	39.4	60.6	0.39			
C. limbatus	SST	58.8	41.2	0.41	0.40	SST: 0.52	SST: 0.52
	SST-CHGT	39.1	60.9	0.39			
C. melanopterus	SST	55.0	45.0	0.45	0.42	SST: 0.54	SST: 0.62
	SST-CHGT	38.8	61.2	0.39			

Table 2.12. Parameters estimation and cross validation under model SST-CH for C. amblyrhynchos, C. limbatus and C. melanopterus. Mean Squared Error (SME), Mean Root Squared Error (SRMSE) and 95% coverage of the median value of each parameter computed on 999 pseudo-observed datasets (pods) simulated under the SST model.

		Nm_{MOD}	Nm_{ANC}	T_{CH}	T_{COL}	N_{ANC}
C. amblyrhynchos	Median (95% CI)	10.9 (3.1 - 28.2)	12.9 (0.3 - 70.7)	11187 (838 - 61407)	40330 (12300 - 172305)	33454 (2546 - 84491)
	SME	0.002	0.223	0.122	0.01	0.229
	SRMSE	0.032	1.754	1.501	0.196	3.874
	Coverage	0.999	0.992	0.995	0.99	0.989
C. limbatus	Median (95% CI)	5.9 (1.7 - 11.4)	26.8 (0.3- 87.3)	31544 (2890 - 136017)	66639 (10993 - 201122)	19682 (746 - 83032)
	SME	0.002	0.308	0.689	0.003	0.067
	SRMSE	0.029	3.749	14.477	0.096	0.471
	Coverage	0.999	0.993	0.99	0.981	0.993
C. melanopterus	Median (95% CI)	1.8 (0.5-2.4)	15.8 (0.6-59.2)	105144 (27412-148715)	186658 (57259-264140)	52870 (1612-96759)
	SME	0.007	0.406	0.651	0.013	0.139
	SRMSE	0.067	4.403	11.45	0.301	1.494
	Coverage	0.987	0.984	0.99	0.989	0.987

2.3.6.6. Supplementary Figures

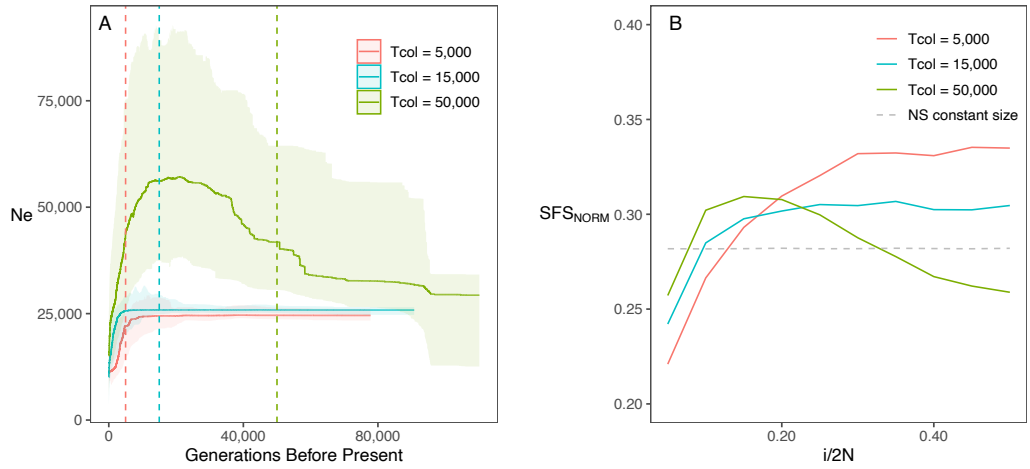

Figure 2.9. *stairwayplot* (maximum likelihood *Ne* and 75% confidence interval) (panel A) and normalized SFS (panel B) computed in simulated non-equilibrium SST scenarios with *Nm*=5, averaged over 100 replicates. Colonisation time of the array of deme T_{COL} occurred 5,000 (red), 15,000 (blue), and 50,000 (green) generations B.P., visually represented by the vertical dashed lines in panel A. The normalized SFS expected under a constant size non-structured model (NS constant size) is also shown (grey dashed line in panel B).

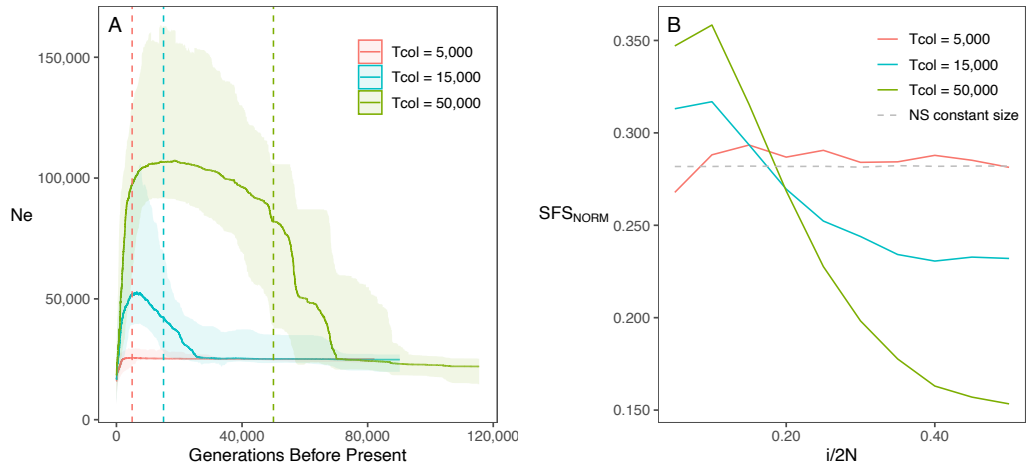

Figure 2.10. *stairwayplot* (maximum likelihood Ne and 75% confidence interval) (panel A) and normalized SFS (panel B) computed in simulated non-equilibrium SST scenarios with Nm=15, averaged over 100 replicates. Colonisation time of the array of deme T_{COL} occurred 5,000 (red), 15,000 (blue), and 50,000 (green) generations B.P., visually represented by the vertical dashed lines in panel A. The normalized SFS expected under a constant size non-structured model (NS constant size) is also shown (grey dashed line in panel B).

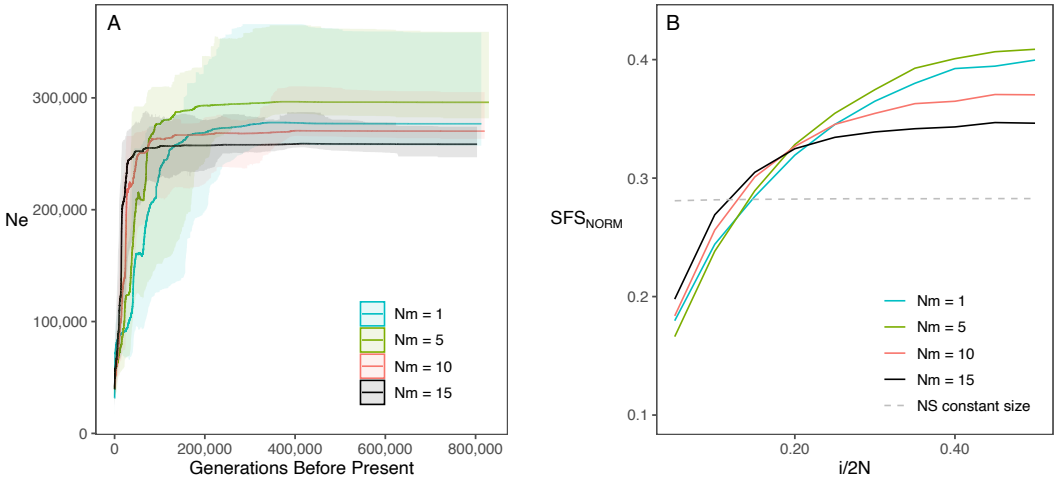

Figure 2.11. *stairwayplot* (maximum likelihood Ne and 75% confidence interval) (panel A) and normalized SFS (panel B) computed in equilibrium SST scenarios, averaged over 100 replicates. Colours represent the long-term connectivity values: Nm=1 (blue), Nm=5 (green), Nm=10 (red), Nm=15 (black). The normalized SFS expected under a constant size non-structured model (NS constant size) is also shown (grey dashed line in panel B).

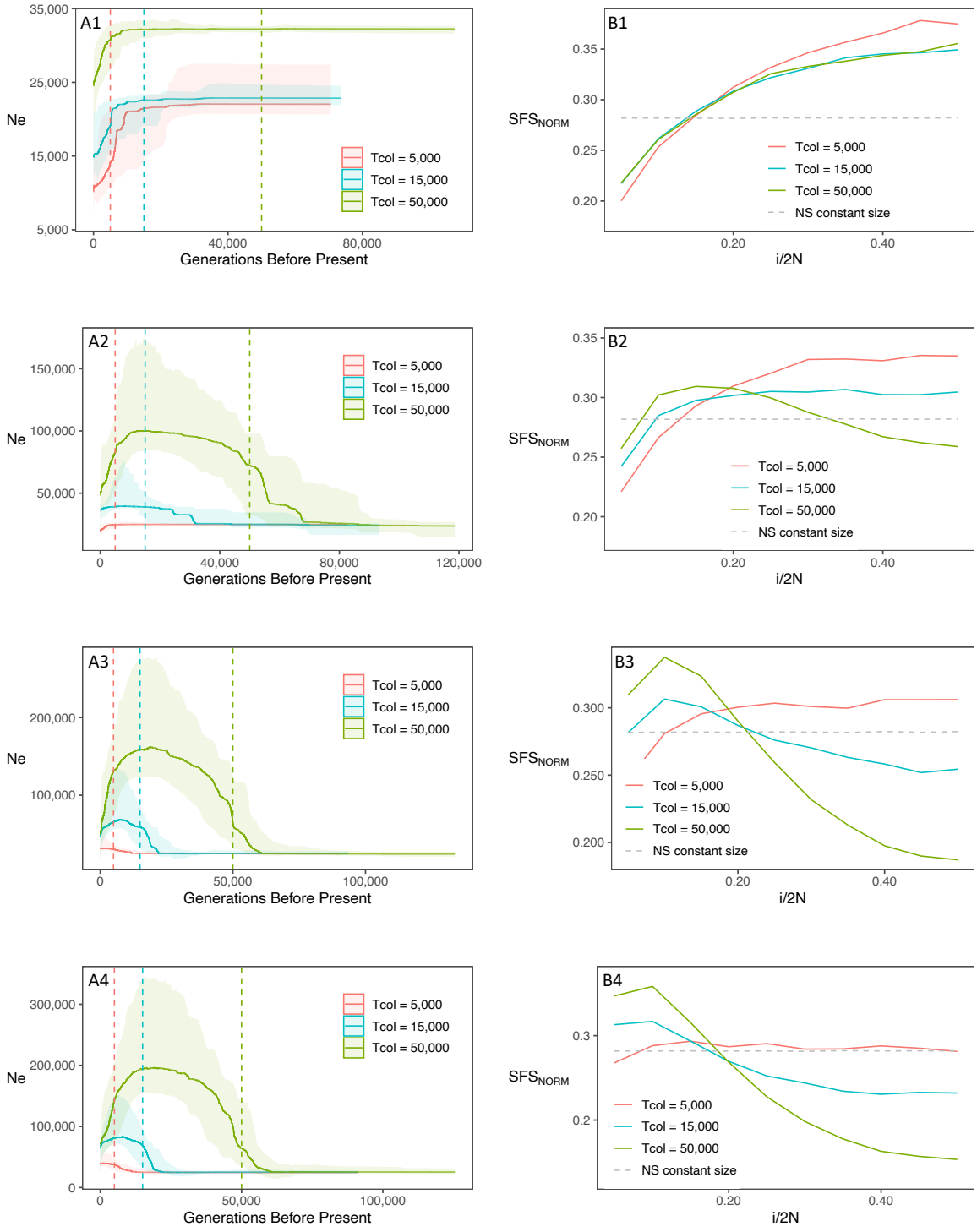

Figure 2.12. *stairwayplot* (maximum likelihood Ne and 75% confidence interval) (panel A) and normalized SFS (panel B) computed in simulated non-equilibrium FIM scenarios. Each row represents a different long-term connectivity: Nm=1 (1), Nm=5 (2), Nm=10 (3), Nm=15 (4). Colours represent the colonisation time of the array of deme T_{COL} 5,000 (red), 15,000 ky (blue), and 50,000 (green) generations B.P.. The dashed lines in panels A indicate the colonisation time Tcol and the grey dashed line in panels B represent the expected normalized SFS under a constant size non-structured model (NS constant size).

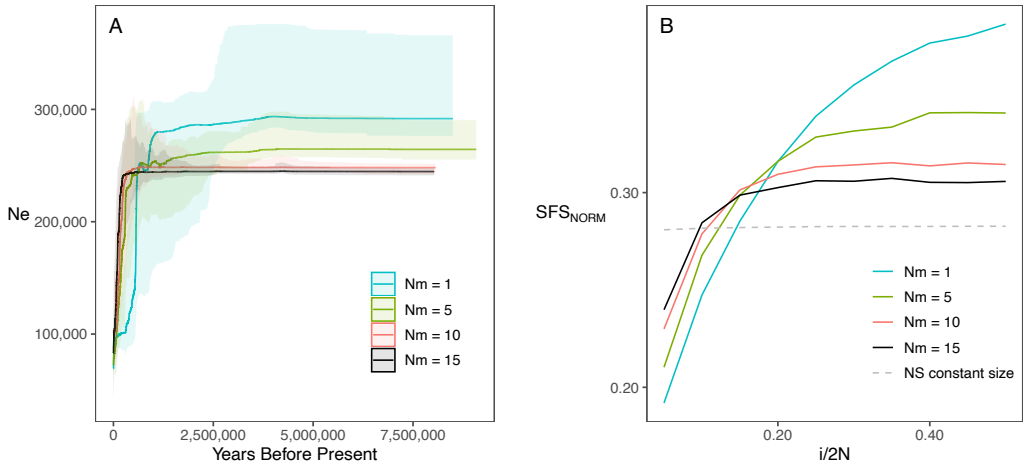

Figure 2.13. *stairwayplot* (maximum likelihood Ne and 75% confidence interval) (panel A) and normalized SFS (panel B) computed in equilibrium FIM scenarios, averaged over 100 replicates. Colours represent the long-term connectivity values: Nm=1 (blue), Nm=5 (green), Nm=10 (red), Nm=15 (black). The normalized SFS expected under a constant size non-structured model (NS constant size) is also shown (grey dashed line in panel B).

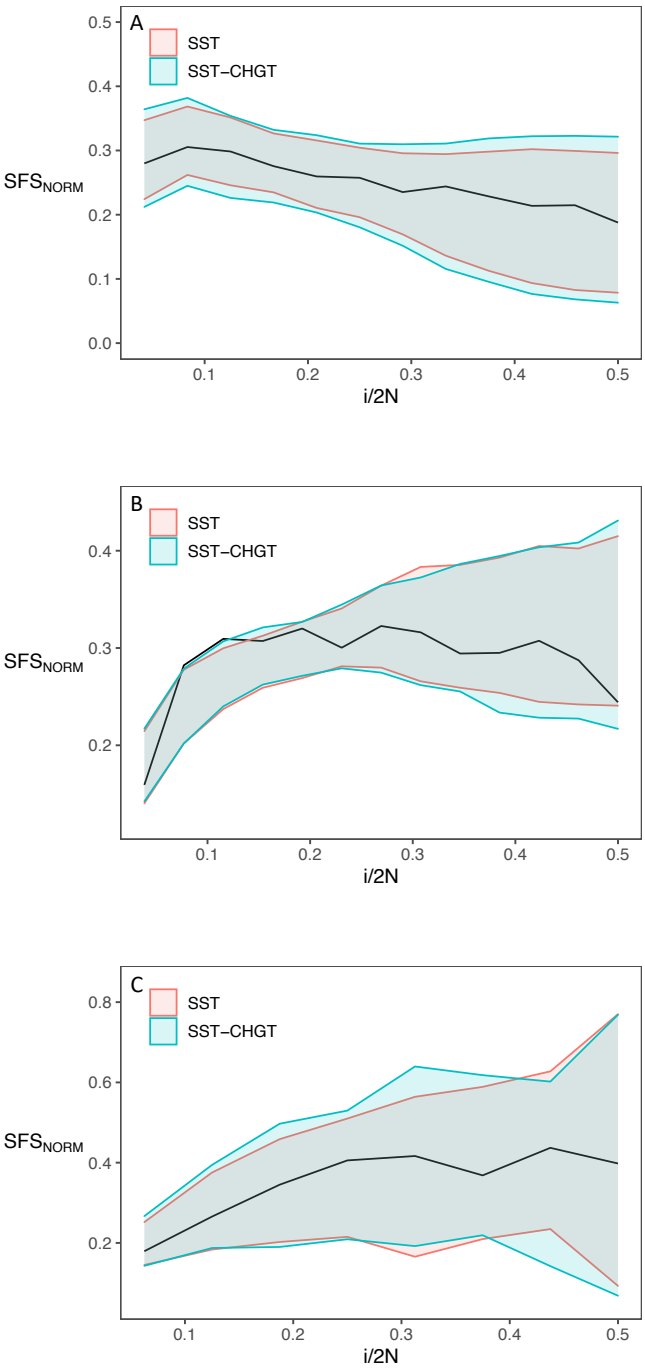

Figure 2.14. Distribution of the 1000 closest normalized SFS retained by the ABC random forest algorithm for models SST (red) and SST-CH (blue) in C. amblyrhynchos (panel A), C. limbatus (panel B) and C. melanopterus (panel C). The black line represents the observed normalized SFS for each species.

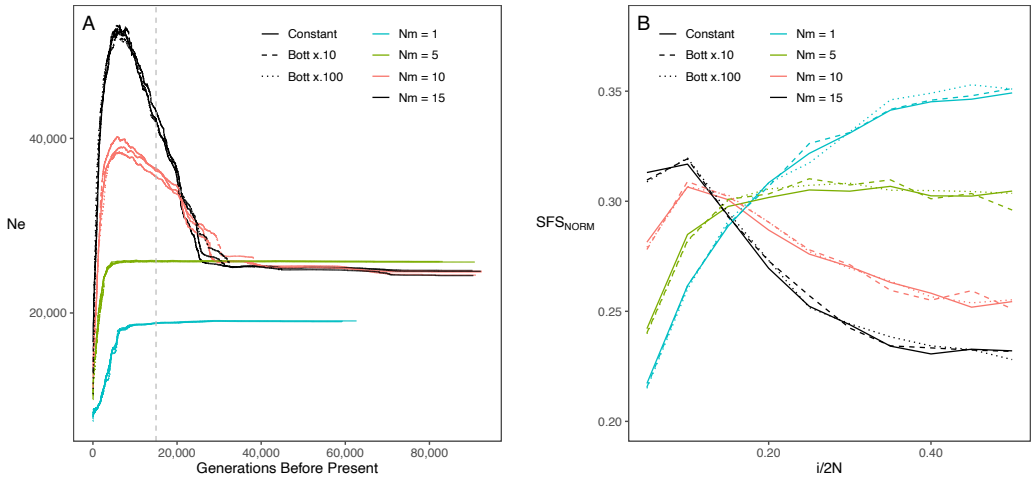

Figure 2.15. *stairwayplot* (maximum likelihood Ne) (panel A) and normalized SFS (panel B) computed in simulated non-equilibrium SST scenarios with T_{COL} =15,000 generations B.P. and an instantaneous decrease of the migration rate (m) forward in time at T_{CH} =10 generations B.P.. Colours represent the long-term connectivity values: Nm=1 (blue), Nm=5 (green), Nm=10 (red), Nm=15 (black). Line style represents the 10-fold (small dashes) or 100-fold (dots) reduction of m, or constant Nm (continuous line). The vertical grey dashed line in panel A represents the simulated colonisation time T_{COL}.

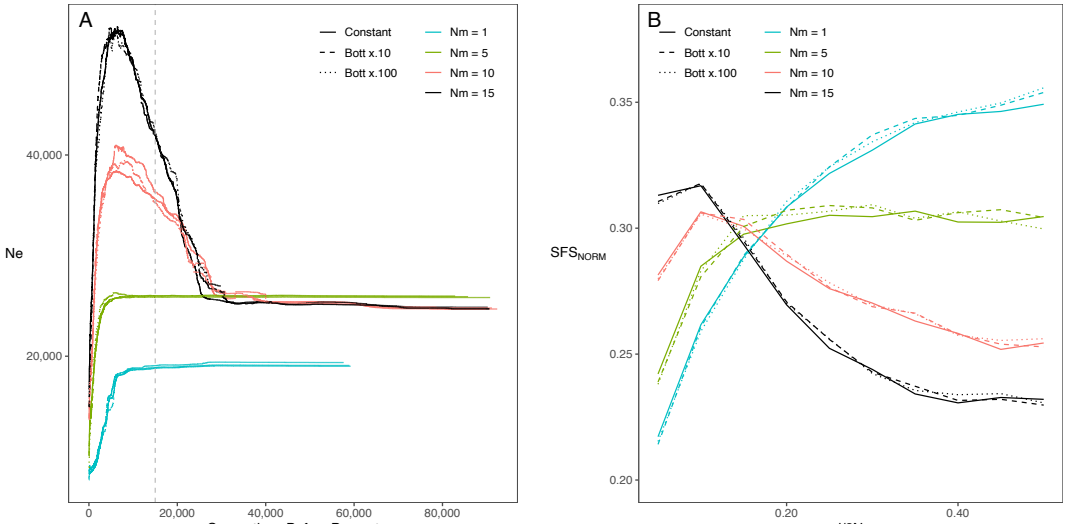

Figure 2.16. *stairwayplot* (maximum likelihood Ne) (panel A) and normalized SFS (panel B) computed in simulated non-equilibrium SST scenarios with T_{COL}=15,000 generations B.P. and an instantaneous decrease of the migration rate (m) forward in time at T_{CH}=50 generations B.P.. Colours represent the long-term connectivity values: Nm=1 (blue), Nm=5 (green), Nm=10 (red), Nm=15 (black). Line style represents the 10-fold (small dashes) or 100-fold (dots) reduction of m, or constant Nm (continuous line). The vertical grey dashed line in panel A represents the simulated colonisation time T_{COL}.

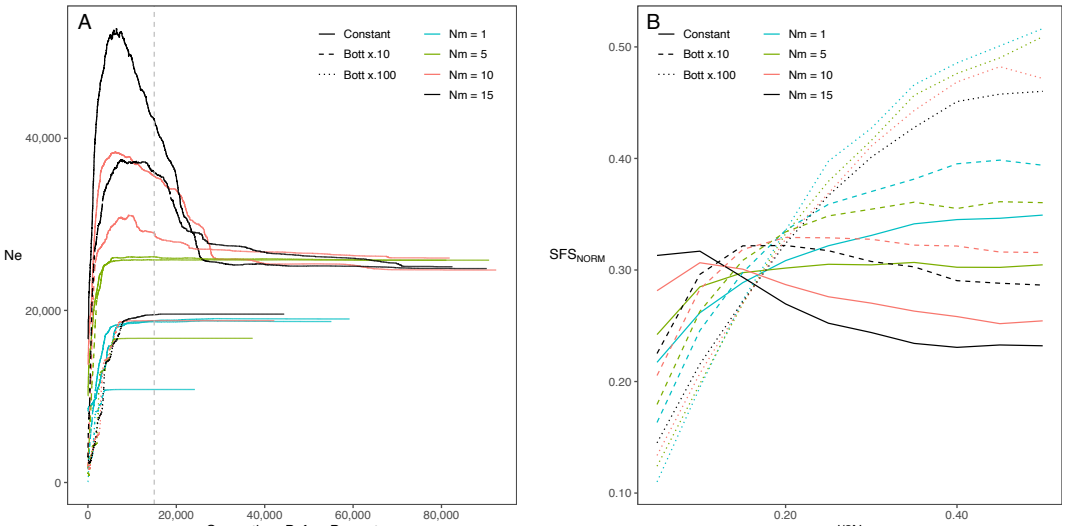

Figure 2.17. *stairwayplot* (maximum likelihood Ne) (panel A) and normalized SFS (panel B) computed in simulated non-equilibrium SST scenarios with T_{COL} =15,000 generations B.P. and an instantaneous decrease of the deme size (N_{DEME}) forward in time at T_{CH} =50 generations B.P.. Colours represent the long-term connectivity values: Nm=1 (blue), Nm=5 (green), Nm=10 (red), Nm=15 (black). Line style represents the 10-fold (small dashes) or 100-fold (dots) reduction of N_{DEME}, or constant Nm (continuous line). The vertical grey dashed line in panel A represents the simulated colonisation time T_{COL}.

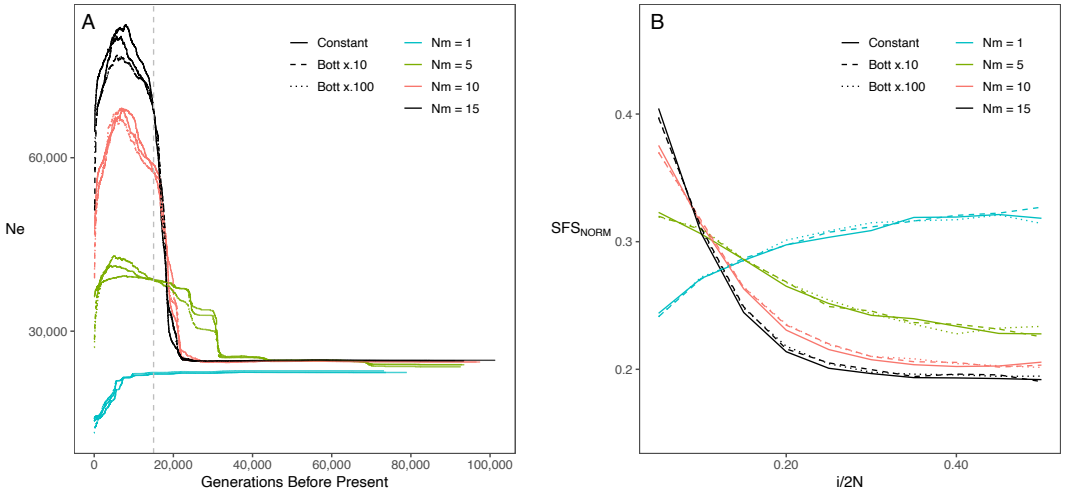

Figure 2.18. *stairwayplot* (maximum likelihood *Ne*) (panel A) and normalized SFS (panel B) computed in simulated non-equilibrium FIM scenarios with T_{COL} =15,000 generations B.P. and an instantaneous decrease of the migration rate (*m*) forward in time at T_{CH} =10 generations B.P. Colours represent the long-term connectivity values: *Nm*=1 (blue), *Nm*=5 (green), *Nm*=10 (red), *Nm*=15 (black). Line style represents the 10-fold (small dashes) or 100-fold (dots) reduction of *m*, or constant *Nm* (continuous line). The vertical grey dashed line in panel A represents the simulated colonisation time T_{COL}.

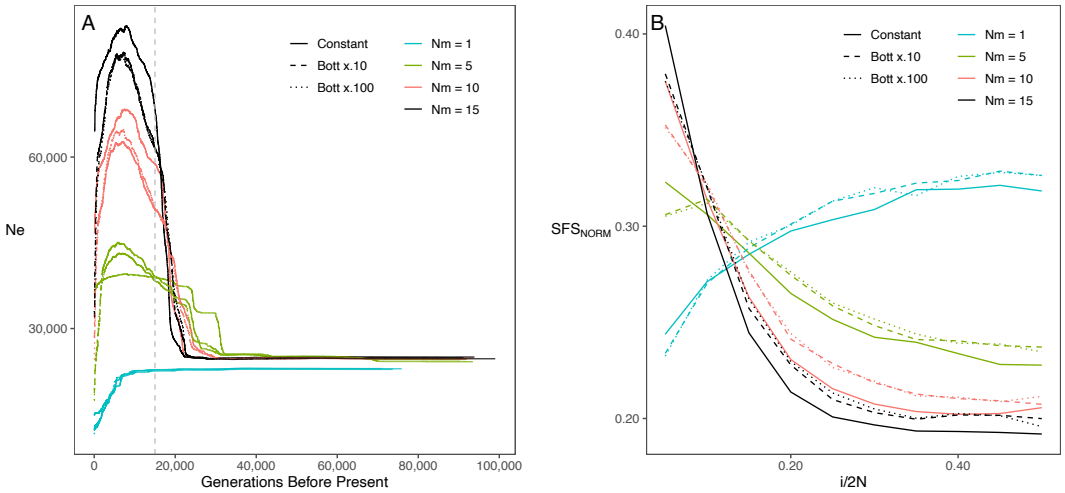

Figure 2.19. *stairwayplot* (maximum likelihood Ne) (panel A) and normalized SFS (panel B) computed in simulated non-equilibrium FIM scenarios with T_{COL} =15,000 generations B.P. and an instantaneous decrease of the migration rate (m) forward in time at T_{CH} =50 generations B.P.. Colours represent the long-term connectivity values: Nm=1 (blue), Nm=5 (green), Nm=10 (red), Nm=15 (black). Line style represents the 10-fold (small dashes) or 100-fold (dots) reduction of m, or constant Nm (continuous line). The vertical grey dashed line in panel A represents the simulated colonisation time T_{COL}.

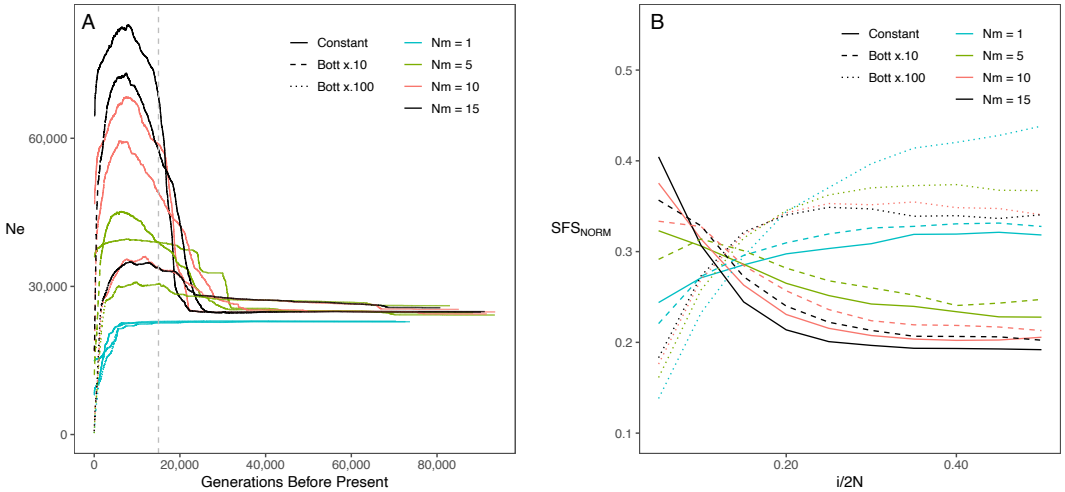

Figure 2.20. *stairwayplot* (maximum likelihood Ne) (panel A) and normalized SFS (panel B) computed in simulated non-equilibrium FIM scenarios with T_{COL} =15,000 generations B.P. and an instantaneous decrease of the deme size (N_{DEME}) forward in time at T_{CH} =10 generations B.P.. Colours represent the long-term connectivity values: Nm=1 (blue), Nm=5 (green), Nm=10 (red), Nm=15 (black). Line style represents the 10-fold (small dashes) or 100-fold (dots) reduction of N_{DEME}, or constant Nm (continuous line). The vertical grey dashed line in panel A represents the simulated colonisation time T_{COL}.

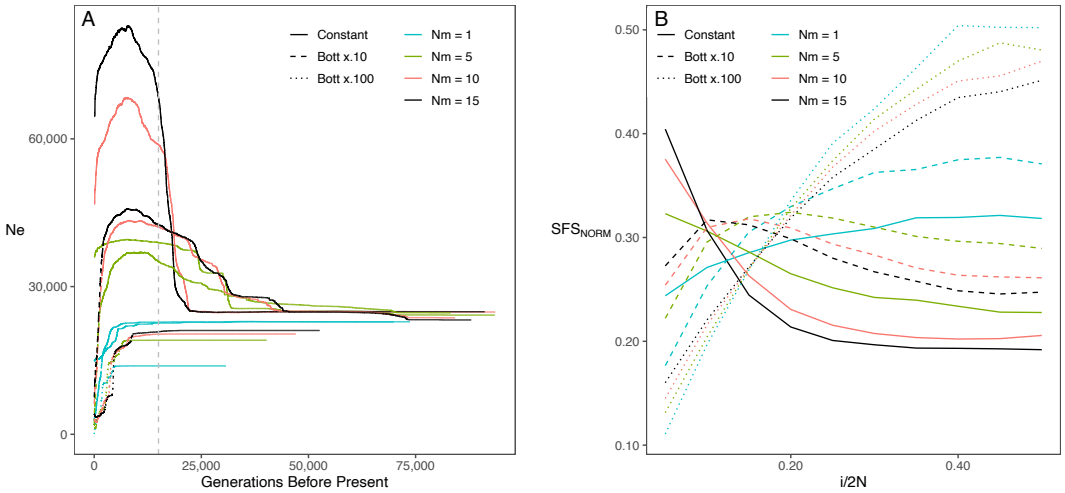

Figure 2.21. *stairwayplot* (maximum likelihood Ne) (panel A) and normalized SFS (panel B) computed in simulated non-equilibrium FIM scenarios with T_{COL} =15,000 generations B.P. and an instantaneous decrease of the deme size (N_{DEME}) forward in time at T_{CH} =50 generations B.P.. Colours represent the long-term connectivity values: Nm=1 (blue), Nm=5 (green), Nm=10 (red), Nm=15 (black). Line style represents the 10-fold (small dashes) or 100-fold (dots) reduction of N_{DEME}, or constant Nm (continuous line). The vertical grey dashed line in panel A represents the simulated colonisation time T_{COL}.

2.4. Ecological and biogeographic features shaped the complex evolutionary history of an iconic apex predator (*Galeocerdo cuvier*)

This article has been published in *BMC Ecology and Evolution*.

Authors:

Pierre Lesturgie, Hugo Lainé, Arnaud Suwalski, Pascaline Chifflet-Belle, Pierpaolo Maisano Delser, Eric Clua, Sébastien Jaquemet, Hélène Magalon and Stefano Mona

2.4.1. Abstract

Background: The tiger shark (*Galeocerdo cuvier*) is a large iconic marine predator inhabiting worldwide tropical and subtropical waters. So far, only mitochondrial markers and microsatellites studies have investigated its worldwide his- torical demography with inconclusive outcomes. Here, we assessed for the first time the genomic variability of tiger shark based on RAD-seq data for 50 individuals from five sampling sites in the Indo-Pacific (IP) and one in the Atlantic Ocean (AO) to decipher the extent of the species' global connectivity and its demographic history.

Results: Clustering algorithms (PCA and NMF), F_{ST} and an approximate Bayesian computation framework revealed the presence of two clusters corresponding to the two oceanic basins. By modelling the two-dimensional site frequency spectrum, we tested alternative isolation/migration scenarios between these two identified populations. We found the highest support for a divergence time between the two ocean basins of ~ 193,000 years before present (B.P) and an ongoing but limited asymmetric migration ~ 176 times larger from the IP to the AO (Nm ~ 3.9) than vice versa (Nm ~ 0.02).

Conclusions: The two oceanic regions are isolated by a strong barrier to dispersal more permeable from the IP to the AO through the Agulhas leakage. We finally emphasized contrasting recent demographic histories for the two regions, with the IP characterized by a recent bottleneck around 2000 years B.P. and the AO by an expansion starting 6000 years B.P. The large differentiation between the two oceanic regions and the absence of population structure within each ocean basin highlight the need for two large management units and call for future conservation pro- grams at the oceanic rather than local scale, particularly in the Indo-Pacific where the population is declining.

Keywords: Agulhas leakage, Coalescent modelling, Demographic history, Population genomics, RAD-seq, Tiger shark

2.4.2. Background

Predation plays a fundamental role in the top-down regulation of ecosystem dynamics, with apex predators being key actors in promoting species diversity (Terborgh, 2015). However, in marine ecosystems, many predatory species have declined across their ranges (Myers & Worm, 2003). Efforts to develop conservation programs need be tailored to the appropriately scaled units of managements for the species under investigation (Palsbøll et al., 2007). Recent advances in DNA sequencing technologies allow the characterization of thousands of independent loci giving the power to assess the genetic diversity of any target model or non-model species, which can inform management policies. However, genetic diversity assessment should be complemented by the reconstruction of species connectivity and historical demography to better establish conservation priorities. For instance, understanding how populations are spatially connected as well as the divergence time between lineages is essential to decipher at which geographic scale a species should be managed. Reconstructing the evolutionary history of a species is often a complex task that requires an educated choice of the most likely model of evolution, often selected among a reduced selection of biologically meaningful models. Unfortunately, selection of an inappropriate model can yield misleading estimates of critically important parameters as more data are collected. This has important implications for conservation genetic applications that rely on accurate estimates of genetic diversity and changes in effective population size through time (Chikhi et al., 2010; Lesturgie, Planes, et al., 2022; Mazet et al., 2015; Mona et al., 2014).

The tiger shark (*Galeocerdo cuvier,* Péron & Lesueur, 1822) is a large and iconic apex marine predator, that is considered "Near Threatened" by the International Union for Conservation of Nature (IUCN). Though the tiger shark is a predominantly coastal species, its distribution includes tropical and subtropical waters worldwide (Compagno, 1984). The species is heavily impacted by fisheries (Temple et al., 2018) and shark control programs in the Indo-Pacific (Sumpton et al., 2011). Indirect estimates have suggested an annual number of tiger shark catches between 50,000 and 300,000 individuals (S. C. Clarke et al., 2006), raising conservation concerns. Though not directly endangered by global climate change, the species is likely to extend its habitat range poleward as a consequence of the rising in annual sea surface temperatures (Payne et al., 2018), which may increase the potential for greater trans-oceanic movements (Holland et al., 2019). Despite being found predominantly along the coast, tiger sharks spend considerable time in pelagic waters and telemetry studies have shown that they can cross oceanic expanses (Lea et al., 2015;

Meyer et al., 2018; Werry et al., 2014), but no evidence of contemporary migration between the Indo-Pacific and the Atlantic Ocean has yet been found.

Knowledge of these ecological traits is important to devise meaningful evolutionary models, but it is not sufficient. Large marine predators with continuous distributions can be intuitively considered as capable of high dispersal due to the absence of clear physical barriers (Palumbi, 1994). Nevertheless, there are both examples of panmictic species, such as the blue shark *Prionace glauca* (18) or the mako shark *Isurus oxyrinchus* (Corrigan et al., 2018), and examples of species structured according to ocean basins such as the Galapagos shark *Carcharhinus galapagensis* and the dusky shark *Carcharhinus obscurus* (Corrigan et al., 2017). In the tiger shark, there has been contrasting evidences about the degree of population structure, the extent of genetic diversity and particularly about the historical demography (Andrade et al., 2021; Bernard et al., 2016, 2021; Carmo et al., 2019; Holmes et al., 2017; Naylor et al., 2012; Pirog et al., 2019). All studies recognize the existence of a clear separation between the Indo-Pacific (IP) and the Atlantic Ocean (AO), with genome-wide data supporting low to now population structure within each basin (Bernard et al., 2021). However, it remains unclear whether the two basins hold two allopatric species as originally proposed by Naylor et. al (2012) or two divergent lineages as proposed by Bernard et. al (2016). An accurate characterization of divergence and migration is additionally still lacking: Bernard et. al (2016) provided a divergence time computed on mtDNA using a non-equilibrium model implemented in MDIV (Nielsen & Wakeley, 2001) but no confidence interval could be determined. At the same time, migration rates were estimated using the equilibrium model implemented in MIGRATE (Beerli & Felsenstein, 2001). Yet a global analysis estimating simultaneously all parameters is warranted. Furthermore, an in-depth analysis of the historical demography in both regions is still lacking, with only (Pirog et al., 2019) supporting a recent decrease in effective population size in two sampling sites from the IP. Using the wealth of data provided by RAD-seq, we sequenced a total of 50 sharks from six sites in the IP and one in the AO (Figure 2.22) in order to shed light on the complex evolutionary history of the tiger shark. We first investigate the extent of genetic diversity, the level of population structure and historical demography in all sampling sites, and finally tested alternative evolutionary scenarios to model the divergence and migration between IP and AO by fitting the observed two-dimensional site frequency spectrum (2D-SFS) with coalescent simulations. These analyses are necessary not only

to reconstruct the evolutionary history of the tiger shark but also to better inform conservation strategies.

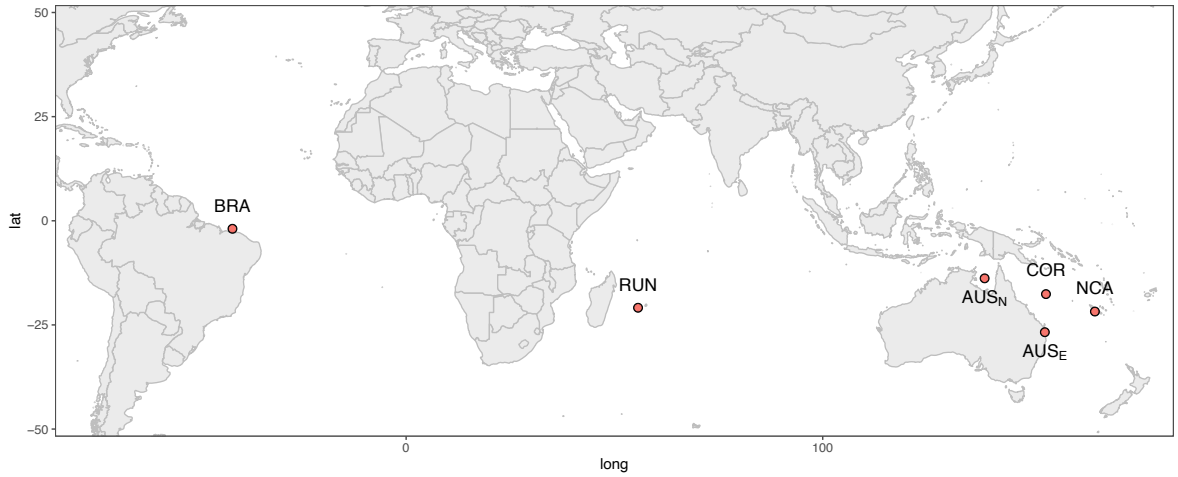

Figure 2.22. Map of the sampling sites. From west to east: Brazil (BRA: n = 7), Reunion Island (RUN; n = 15), North Coast of Australia (AUS_N; n = 8), Coral Sea (COR; n = 5), East Coast of Australia (AUS_E; n = 7) and New Caledonia (NCA; n = 8).

Table 2.13. Sample size (n), mean pairwise difference (θπ), Watterson theta (θw), Tajima's D (TD), and total number of loci (monomorphic included) (nloci) and SNPs (nSNP) without missing data for all sampling sites (ranged from west to east). AUSE: East Coast of Australia; AUSN: North Coast of Australia; BRA: Brazil; COR: Coral Sea; NCA: New Caledonia; RUN: Reunion Island.

	n	θ_π (10^{-3})	θ_w (10^{-3})	TD[a]	n_{loci}	n_{SNP}
BRA	7	0.97	1.16	**0.14**	16,953	5,868
RUN	15	0.59	0.73	**-0.19**	71,214	19,971
AUS_N	8	0.76	0.97	**-0.19**	38,420	11,627
COR	5	0.58	0.72	0.04	97,736	18,407
AUS_E	7	0.63	0.77	0.02	49,380	11,385
NCA	8	0.57	0.7	-0.03	118,591	26,075

[a] Tajima's *D* values in bold are significantly different from 0 ($P < 0.001$).

2.4.3. Results

2.4.3.1. RAD-seq sequencing

The average number of reads retained per individual after the quality filtering and demultiplexing step was 4,011,430 (± 1,314,894 s.d.). After a first round of *de novo* assembly and filtering using STACKS v.2.5, the depth of coverage was low with a mean of 12.65 (± 6.36 s.d.), which motivated

the use of the genotype-free allele frequency estimation pipeline implemented in ANGSD (30) rather than the direct call. The final number of loci (variable and fixed) was highly variable between sampling locations (Table 2.13) ranging from 16,953 to 118,591 in Brazil (BRA) and New Caledonia (NCA) sampling sites respectively, with a number of SNPs with no missing data following a similar trend (from 5,868 to 26,075 for BRA and NCA, respectively).

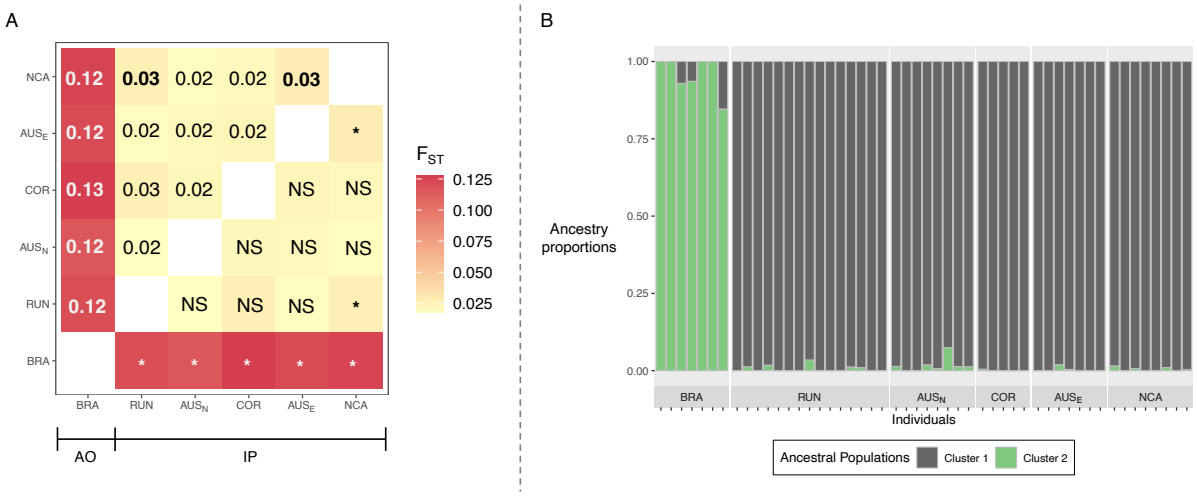

Figure 2.23. Heat map representing the pairwise Reynold's F_{ST} values between sampling sites (A) and ancestry proportions retrieved using the nmf algorithm with K=2 ancestral populations (B). Both analyses were performed with PCANGSD. Values in the upper triangle of the heat map are the pairwise FST values and significance is displayed on the lower triangle: non-significant (NS) or p<0.001 (*).

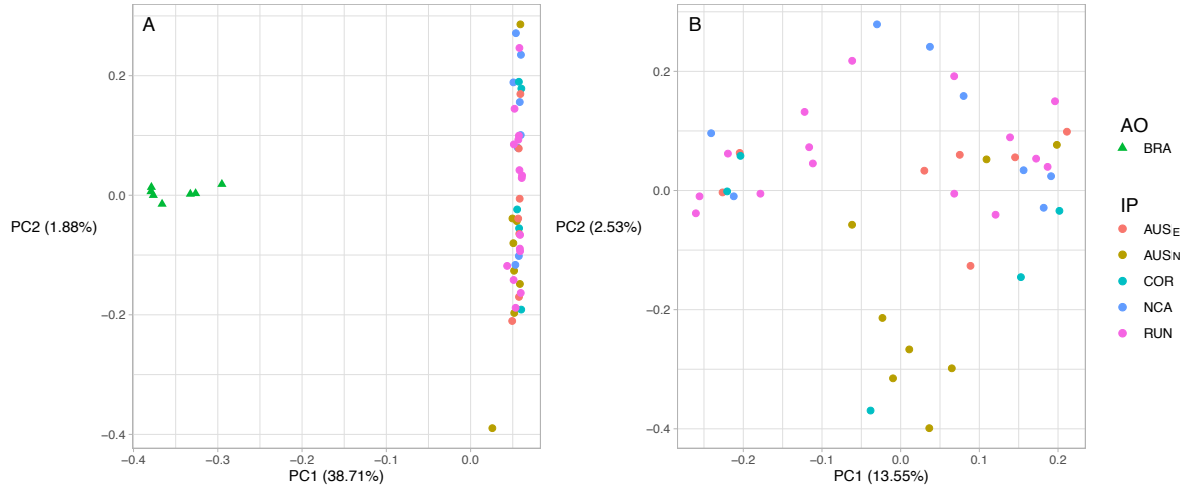

Figure 2.24. Principal Component Analysis (PCA) computed with: (A) all individuals (n = 50) and (B) Indo-Pacific individuals only (n = 43).

2.4.3.2. Population structure

Population structure was investigated using datasets allowing up to 20% of missing data per SNP. Thus, after filtering, the remaining number of SNP was 24,454 for the Principal Component Analysis (PCA) and the *non-negative matrix factorization* (*nmf*) inference, and ranged from 8,785 to 15,977 per population pair for the F_{ST} computation. Pairwise F_{ST} highlighted a moderate differentiation between Indo-Pacific (IP) and Atlantic Ocean (AO) sampling sites, with values ranging from 0.117 to 0.129 and systematically significant ($P \leq 0.001$, Figure 2.23-A and Supp. Table 2.15). Conversely, the average F_{ST} between IP sites was 0.023 (ranging from 0.018 to 0.029) and not statistically significant for the majority of pairwise comparisons (Supp. Table 2.15). The Mantel test, computed between IP sampling sites only (given the evidence of a clear genetic discontinuity between AO and IP), showed no correlation between genetic and geographic distances (r = 0.005, $P = 0.62$, Supp. Figure 2.27). Clustering analyses were consistent with the observed pattern of differentiation. First, the *nmf* algorithm selected K=2 ancestral populations corresponding to IP and AO (Figure 2.23). Individuals from IP had a probability ancestry to cluster 1 ranging from 92.6% to 100% whereas individuals from AO showed a probability ancestry to cluster 2 ranging from 84.6 to 100%. Average ancestry of cluster 2 in IP individuals was only 0.7% while average ancestry of cluster 1 in AO individuals was 4.4%. Second, the PCA clearly segregated AO from IP individuals, with 38.71% of the total variance explained by the first axis (Figure 2.24-A). Individuals from Reunion Island (RUN), the IP site closest to the AO, did not show more proximity to the AO in the PCA or a higher contribution from cluster 2 than other IP individuals, nor did they show a lower pairwise F_{ST} with the AO than the other IP sites. (Figure 2.23, 2.24-A and Supp. Figure 2.28-A). When computed on IP individuals only, the PCA identified a single cluster (Figure 2.24-B and 2.28-B) and the *nmf* did not show any meaningful geographic clusters with $K=2$ (Supp. Figure 2.29). We further applied an Approximate Bayesian Computation (ABC) framework using a 500 trees random forest for all sampling sites after checking for the evolution of the out-of-bag error rate. This coalescent framework allows to detect genetic structure using a single sampling location by testing whether its gene genealogy yield signatures of a *Stepping Stone* (SS), *Finite Island* (FIM) or *Non-Structured* (NS) model (Supp. Figure 2.30). The model selection (Supp. Table 2.16) highlighted NS as the most supported model with a posterior probability ranging from 0.48 to 0.89 in the IP sampling sites and of 0.62 for BRA (Supp. Table 2.17).

2.4.3.3. Genetic diversity and variation of Ne

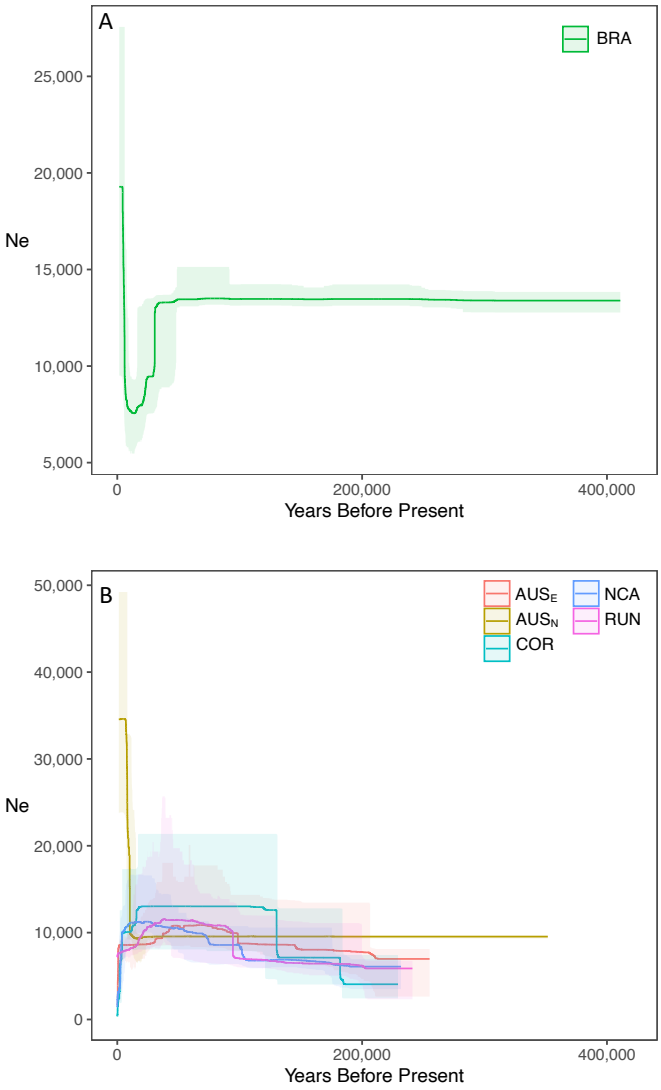

Figure 2.25. Variation of the effective population size (N_e) through time and its 75% confidence interval estimated by the STAIRWAYPLOT for the AO (panel A) and IP (panel B) sampling sites. AUS_E: East Coast of Australia; AUS_N: North Coast of Australia; BRA: Brazil; COR: Coral Sea; NCA: New Caledonia; RUN: Reunion Island.

Genetic diversity values were very similar among sampling sites, with BRA being slightly more variable than the IP counterpart (Table 2.13). Tajima's D (*TD*) was significantly positive for BRA (*TD* = 0.137; $P \leq 0.001$), while significantly negative ($P \leq 0.001$) for Northern Australia (AUS_N) and Reunion Island (RUN) and not significantly different from zero for the other populations (Table 2.13). Except for the AUS_N population, the reconstructions of the effective size (N_e) through time by the STAIRWAYPLOT were very similar among IP locations: an ancestral expansion occurred

between ~100,000 and ~200,000 years before present (BP). bringing the median N_e to ~10,000 followed by a very recent bottleneck ~2,000 to ~4,000 years BP (Figure 2.25). The STAIRWAYPLOT for AUS_N displayed a different signal, with an ancestral N_e median value similar to the one retrieved in the other sampling sites (~10,000) followed by a strong and recent expansion that raised the modern N_e to ~35,000, contrasting with the recent decrease observed for the other IP sampling sites. The demographic history reconstructed for BRA was slightly more complex with the ancestral N_e of ~12,000 first decreasing to ~9,000 at ~40,000 years BP and then increasing (between ~4,000 and 6,000 years BP) to a modern N_e of ~20,000 (Figure 2.25).

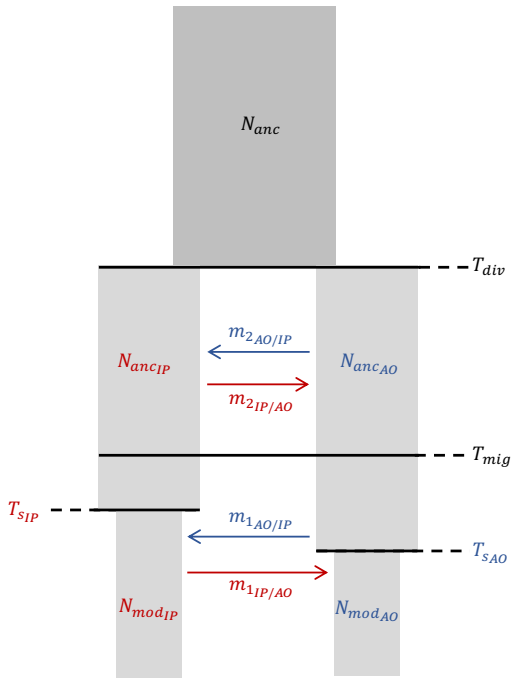

Figure 2.26. Model IM-full, the most parameter-rich model (13 parameters) representing two populations from each ocean basin with an effective size that changed $T_{s_{IP}}$ and $T_{s_{AO}}$ years ago from a modern effective size ($N_{mod_{IP}}$ and $N_{mod_{AO}}$) to an ancestral effective size ($N_{anc_{IP}}$ and $N_{anc_{AO}}$). The two populations are connected by an asymmetrical migration rate allowed to change T_{mig} years ago (respectively from $m_{1_{IP/AO}}$ and $m_{1_{AO/IP}}$ to $m_{2_{IP/AO}}$ and $m_{2_{AO/IP}}$) and diverged T_{div} years ago from an ancestral population of size N_{anc}. The remaining four models are nested within IM-full, having less migration rate parameters: IM-anc is similar to IM-full but only the ancestral migration occurs (i.e., between T_{mig} and T_{div}); IM-rec is similar to IM-full but only the recent migration occurs (i.e., between 0 and T_{mig}); IM-bsc considers the migration constant from 0 to T_{div}; and IM-div is a strict divergence model with no migration.

Table 2.14. Maximum Likelihood (ML), 90% confidence interval (5% lower bound and 95% upper bound) and search bounds for the parameters estimated by FASTSIMCOAL under the IM-bsc model.

Parameter	ML	5% lower bound	95% upper bound	Parameter bounds
N_{anc}	9,274	1,310	80,372	U¶:{100;100,000}
$N_{mod_{IP}}$	17,591	16,643	31,915	U:{100;100,000}
$N_{mod_{AO}}$	16,810	12885	40,036	U:{100;100,000}
$T_{s_{IP}}$*	153,090	27,460	769,750	U:{100;1,000,000}
$T_{s_{AO}}$*	45,980	14,030	53,090	U:{100;1,000,000}
$N_{anc_{IP}}$	48,823	3,103	107,816	U:{100;100,000}
$N_{anc_{AO}}$	1,406	1178	4,781	U:{100;100,000}
T_{div}*	193,850	77,110	355,910	U:{100;1,000,000}
$Nm_{AO/IP}$§	0.022	0.008	0.124	U:{0;50}
$Nm_{IP/AO}$§	3.873	3.171	8.919	U:{0;50}

*Times are expressed in years using a mutation rate of 1.93 x 10^{-8} per generation per site and a generation time of 10 years
§ Number of migrants per generation are expressed in forward.
¶ U: uniform distribution.

2.4.3.4. Population divergence and migration rate estimation

In the light of the absence of population structure signatures within the IP and the AO and the genetic distinctness between them, we tested five Isolation-Migration (IM) models to determine the divergence time and the migration pattern between the two oceanic regions (Figure 2.26). The likelihood distribution computed over 100 replicates was similar for models IM-bsc and IM-full, but the AIC values supported IM-bsc as the model with the highest probability (Supp. Figure 2.31). The distribution of the likelihood evaluated in each model under the Maximum Likelihood (ML) parameters proved that they can be distinguished based on the available data (Supp. Figure 2.31). The two oceanic regions appeared connected, though the migration rate is limited and strongly asymmetric, being ~176 times higher from IP to AO (*Nm*~3.9) than *vice versa* (*Nm*~0.02) (Table 2.14 and Supp. Figure 2.32). Going backward in time the populations from the two regions merged ~193,000 years BP (90% CI: [77,000; 355,000]). The ancestral population size was almost half of those estimated in both IP and AO derived populations (Table 2.14 and Supp. Figure 2.32), indicating an ancestral expansion, consistent with the observed STAIRWAYPLOT dynamics for IP

populations. An increase in effective size was estimated in the AO starting ~ 45,000 years BP (90% CI: [14,000; 53,000]), bringing the effective size from 1,406 (90% CI: [1,178; 4,781]) to 16,810 (90% CI: [12,885; 40,036]) which is consistent with the observed expansion in the STAIRWAYPLOT of BRA. We observed a decrease in N_e in the IP from 49,000 to 17,000, but the timing was poorly estimated. Moreover, ancient and modern N_e in the IP showed largely overlapping confidence intervals (Table 2.14).

2.4.4. Discussion

To shed light on the evolutionary history of the tiger shark, we sequenced thousands of loci in 50 individuals following a double digest RAD-seq protocol. We handled low coverage issues by applying an appropriate framework based on genotype free likelihood estimation of allele frequencies (Korneliussen et al., 2014). The first result is the unambiguous presence of two highly divergent genetic clusters, corresponding to the Indo-Pacific (IP) region and the Atlantic Ocean (AO), and the signature of very weak population structure within each of them. Despite the large panel of SNPs that could potentially detect fine spatial structure compared to previous work based on microsatellites, we did not find any barrier to gene flow within the IP, but rather a signature consistent with a large panmictic population or a meta-population characterized by very large amount of gene flow (Figures 2.23, 2.24, Supp. Figures 2.27, 2.28 and 2.29). This strongly confirms previous findings (Bernard et al., 2021; Holmes et al., 2017; Pirog et al., 2019) and contradicts the conclusions of (Bernard et al., 2016), who found significant evidence for population structure. Using a larger amount of genomic information, we provide evidence for the presence of a single mating population in the IP based on the following observations: (1) the PCA and *nmf* analyses display one single cluster in the IP; (2) the F_{ST} values computed between sampling sites did not exceed 0.029 (as a comparison, an average value of ~0.124 was found between IP and AO) with no signature of isolation by distance (Supp. Figure 2.27); (3) all IP sharks showed a similar amount of genetic contribution from the AO cluster in the *nmf* analysis (Figure 2.23). These results are consistent with one of two explanations: either (1) tiger sharks randomly mate within IP or (2) the number of migrants exchanged each generation between sampling sites is so large to erase any signature of genetic structure. The absence of multiple sampling sites in AO prevented us to perform similar analyses in AO. To test the presence of a single panmictic population, we therefore followed an ABC strategy based on coalescence simulations to reconstruct the evolutionary history

of BRA population and assess whether the patterns of genetic variation within the BRA samples are better described by unstructured or meta-population models. This approach has been successfully applied in both empirical and simulation-based works (Lesturgie, Planes, et al., 2022; Maisano Delser et al., 2019; Peter et al., 2010) and represents an alternative (or complementary, when possible) method to infer the presence of population structure. As found for the IP, the NS (No Structure) model had the most support within the AO (Supp. Table 2.17). Even though more sampling sites and SNPs would likely have provided tighter estimates, this result is most consistent with a single mating population in the AO, despite our single sampling site does not allow us to infer the geographic extent of that population. We note that population structure has previously been reported within the AO based on mitochondrial markers (Andrade et al., 2021; Carmo et al., 2019). However, results in the current study are consistent with the genome wide study of (Bernard et al., 2021) which suggested low to no population structure in the AO. The differences between the inferred mitochondrial structure and genomic DNA signals could be due to sex-biased dispersal, as female philopatry has been proposed for some shark species (Keeney et al., 2005; Mourier & Planes, 2013; Pardini et al., 2001; Tillett et al., 2012). However, inconsistencies between mitochondrial and autosomal data are widespread in nature and it has been suggested to cautiously interpret the mitochondrial variation in the light of the demographic history of a species, as other evolutionary forces such selection may act as confounding factors (Ballard & Whitlock, 2004). Furthermore, the result found herein (i.e., low to no population structure in each of the two oceanic regions) is consistent with the fact that tiger sharks have been documented to move large distances across oceanic basins. However, it does not explain why a large predator that is capable of covering distances of several thousands of kilometers (Holland et al., 2019; Lea et al., 2015; Werry et al., 2014) could not erase the genetic differentiation between the two regions.

One possible explanation was proposed by (Naylor et al., 2012), who suggested the presence of two allopatric subspecies in IP and AO. This hypothesis was later refuted by (Bernard et al., 2016), who still agreed on a long-term genetic isolation between the two oceanic regions but proposed some genetic exchanges. By harnessing the power of RAD-seq genome wide data, our coalescent modelling could not only disentangle the two hypotheses, but also provide quantitative estimates of the tempo and mode of divergence between the two populations. Comparing five Isolation/Migration (IM) scenarios, we found that the most supported model included a divergence around ~193,000 years BP (90% CI: [77,000; 356,000]) between the two regions (Table 2.14),

which nevertheless remained in contact since then through a very limited (3.9 individuals per generation; 90% CI: [3.2; 8.9]) and asymmetric gene flow ~176 times higher from IP to AO than the opposite direction. The low number of migrants *Nm* exchanged each generation (Table 2.14) and the asymmetric exchange are consistent with the clustering results and the F_{ST} values (Figure 2.23), which differentiated the two regions and clearly highlighted a higher, but weak, genetic contribution from IP to AO than *vice versa* (Figure 2.23). These results support the idea that populations from the AO and from the IP represent two lineages (Bernard et al., 2016), rather than two allopatric species (Naylor et al., 2012). A permeable barrier to gene flow between the two oceanic regions is therefore responsible for the observed pattern of divergence and asymmetric migration. The presence of this barrier can be explained by the ecology of the tiger shark and by the environmental conditions governing the Indian-Atlantic water exchange, the so-called Agulhas leakage. As a tropical to sub-tropical species, tiger sharks prefer warm water and they show the peak of swimming activities at ~22°C (Payne et al., 2018) so that their movement from the Atlantic to the Indo-Pacific is hampered by the upwelling of cold water off South-Western Africa (the Benguela current). However, the AO receives warm water from the IP through the Agulhas leakage (Beal et al., 2011), which can account for the asymmetric migration reported, consistent with the pattern observed in other tropical sharks, bony fishes and turtles (Gaither et al., 2016; Maduna et al., 2017; Reid et al., 2019; van der Zee et al., 2021). The Agulhas leakage has not been constant through time, with an increasing intensity in the Holocene, preceded by a period of stasis and a strong peak in the late Pleistocene around 130,000 years BP (Caley et al., 2012, 2014). This variation in Agulhas leakage intensity could have influenced the relation between the two basins. However, we could not distinguish pulses of migrations in our data since the model IM-bsc was preferred to those accounting for variation in migration rate through time. Model selection procedure robustly selected the IM-bsc model (Supp. Figure 2.31), which is neither the most nor the least parameters rich, supporting the idea that our results are not an artefact of incorrect modelling (see also below). In the future there will still be space to improve our estimates: more data will help refining the confidence interval of each parameter and ameliorating the calibration of the molecular clock.

We found divergent demographic histories in the two oceanic regions examined (Figure 2.25). First, we note that since both regions are most likely described by non-structured models and the migration rates between them are very low, it is possible to directly interpret the results of

unstructured model such as the STAIRWAYPLOT (Liu & Fu, 2020). It is important to stress this point, because population structure, if not accounted for, can generate signatures that erroneously look like changes in of population size (Chikhi et al., 2018; Maisano Delser et al., 2019; Mazet et al., 2016). All the Indo-Pacific populations (except AUS_N) underwent a recent bottleneck between ~2,000 and 4,000 years BP, which was robust to the pooling of all IP sampling points but AUS_N (but with larger incertitude in recent times due to the lower number of SNPs, see Supp. Figure 2.33). Despite the caveat regarding the interpretation of N_e variation in recent times due to the inclusion of singletons, this is barely consistent with what was recently proposed by (Pirog et al., 2019) based on 25 microsatellites combined with mitochondrial DNA. Here we refine their estimates and better characterized the intensity of the bottleneck with a non-parametric approach (the STAIRWAYPLOT) exploring a large parameter space to genome wide data. The inferred demographic history of the AO is strikingly different from that of the IP. The estimated N_e was ~20,000 following a population expansion occurring between ~4,000 and 6,000 years BP (Figure 2.25). These values are consistent with the estimates retrieved by the IM-bsc model (Table 2.14). The strong signature of population expansion recovered implies that the tiger shark is profiting from recent environmental changes in AO, in contrast to the IP population. Consistently, (Andrade et al., 2021; C. D. Peterson et al., 2017) found a recent demographic trend suggesting an expansion rather than a decrease in AO, while the recent demographic trends in the IP appear to have been the result of intense pressure from fisheries and shark-control programs (S. C. Clarke et al., 2006; Sumpton et al., 2011; Temple et al., 2018). More investigations are needed to determine the origin of the difference between the two oceans. The applications of methods based on linkage disequilibrium applied to whole genome data will help detect more recent events (Kerdoncuff et al., 2020), which will be important for planning conservation strategies. Given the very low genetic structure within IP, we would expect AUS_N to have the same demographic history than the other sampling sites in the IP. We cannot exclude a scenario where AUS_N represents an isolated population experiencing its own demographic history that separated from the rest of the IP too recently to accumulate divergence. If confirmed, this would highlight the presence of independent lineages in the IP, with important consequence for conservation programs. However, more data is needed to shed light on this topic, both in terms of individuals and in terms of genomic coverage: ultimately only whole genome sequencing will give the opportunity to confidently resolve this issue.

2.4.5. Conclusions

Reconstructing the evolutionary history of a species relies on the application of a realistic demographic model, which is mostly unknown in empirical studies. Here, we investigated the evolutionary history of the tiger shark and found that it is characterized by an asymmetrical migration between the AO and the IP and a signature of random mating within each region. These findings let us model each oceanic region as a single unstructured population and evaluate competing demographic scenarios to investigate their divergence time and migration rates. The two regions are separated by the Benguela barrier, but our estimates strongly suggest that the Agulhas leakage allows an overwhelmingly asymmetric migration between them, by far stronger from IP to AO than in the opposite direction. While we confirmed that the tiger shark is likely undergoing a reduction in *Ne* in the Indo-Pacific, we show that it probably underwent a strong expansion in the Atlantic Ocean. Even if a better calibration of the molecular clock and full genome analyses would still be needed to confirm our results, our findings support the existence of two management units. This implies that local conservation or shark control programs will have very limited impact on the dynamics of the species, which needs to be managed at the ocean basin level, demanding considerable communication efforts among different countries and coordination as suggested for other megafaunal organisms (Barkley et al., 2019).

2.4.6. Material and Methods

2.4.6.1. Sampling, library preparation and sequencing

A total of 50 tiger shark individuals (*Galeocerdo cuvier*) from both the Indo-Pacific (IP) and the Atlantic Ocean (AO) were sampled off (from west to east) Brazil (BRA), Reunion Island (RUN), North Coast of Australia (AUS_N; North Territory), East Cost of Australia (AUS_E; Sunshine Coast), Coral Sea (COR) and New Caledonia (NCA). Sharks were grouped into six populations based on their sampling site (Figure 2.22; Table 2.13). Total genomic DNA was extracted from muscle tissue or fin clips preserved in 96% ethanol using QIAGEN DNeasy Blood and Tissue kit (Qiagen, Hilden, Germany) according to the manufacturer's protocols. Double-digest restriction-associated DNA (ddRAD) libraries were prepared following (Peterson et al., 2012) using EcoRI and MspI restriction enzymes, a 400-bp size selection, and a combination of two indexes and 24 barcodes to pool 48 individuals per lane. The genomic libraries obtained were sequenced with a HiSeq 2500 Illumina sequencer (single-end, 125 bp).

2.4.6.2. RAD-seq de novo assembly

Raw reads were first demultiplexed and quality filtered through the *process_radtags.pl* pipeline in STACKS v.2.5 (Rochette et al., 2019). In the absence of a reference genome of *G. cuvier* or of closely related species, RAD-seq loci (125 bp sequences) were *de novo* assembled under the *denovo_map.pl* pipeline in STACKS. Preliminary results based on parameters *m=3* (minimum read depth to create a stack), *M=3* (number of mismatches allowed between loci within individuals), and *n=3* (number of mismatches allowed between loci within catalogue) found an average depth of ~10x (see Results). Despite the absence of a clear cut-off indicating an acceptable coverage value above which genotype calling may be considered reliable, simulation results suggest that ~10x may produce inconsistent calling under different algorithms (Fountain et al., 2016). To prevent this, we used a genotype free estimation of allele frequencies implemented in the software ANGSD v.0.923 (Analysis of Next Generation Sequencing Data; (Korneliussen et al., 2014)), which has been proven to be a more efficient method for low to medium coverage next-generation sequencing (NGS) data than SNPs calling algorithms (Korneliussen et al., 2014). We describe below the bioinformatics steps required to apply ANGSD to RAD-seq data from a non-model organism and the downstream population genetic analyses applied to the filtered datasets.

2.4.6.3. Assembly pipeline and filtering

ANGSD requires a reference sequence to work, which we were lacking. To circumvent this issue, we followed the approach described in (Khimoun et al., 2020) by creating an artificial reference sequence from loci previously assembled by STACKS under the parameter *m*=3, *N*=3, *M*=3 (based on the results of (Mona et al., 2023)). To this end, we concatenated the consensus sequences of each locus spaced by a stretch of Ns and then map reads back from individual *fastq* files using the *bwa-mem* algorithm with default parameters (H. Li & Durbin, 2009). Using custom bash scripts coupled with ANGSD, we then discarded: (i) sites with coverage <3x (*-minIndDepth* = 3, corresponding to *m* in the first assembly performed by STACKS) and/or of low quality (based on the per base alignment score, *-baq =1* flag); (ii) low quality bases and poorly aligned reads (*-minQ* and *-minMapQ* and *-C* flags with default values); (iii) SNPs present in the last 5 bp of each locus and SNPs genotyped as heterozygous in 80% or more of the individuals; (iv) loci with more than 5 SNPs that might be the result of paralog RAD loci alignment on the reference. Specific filters were further added according to the downstream analyses performed.

2.4.6.4. Population structure

A single reference sequence was created for all populations and we retained sites shared by at least 80% of the samples. The PCA was computed with PCANGSD v.0.97 based on genotype likelihood (Meisner & Albrechtsen, 2018). Admixture was then investigated by running the *non-negative matrix factorization* algorithm (*nmf*) implemented in PCANGSD which is based on the same covariance matrix inferred for the PCA. The number of ancestral populations (*K*) was automatically chosen by PCANGSD to be $e + 1$, where e is the optimal number of significant principal components depicting population structure, resulting from the Velicier's minimum average partial test run on the covariance matrix. The sparseness regularization parameter α (used to reduce the noise in low depth NGS data) that best fitted the data was tested between 0 and 100 and it was chosen by comparing the resulting likelihood following (Meisner & Albrechtsen, 2018). We generated pairwise *site allele frequency likelihood* files and then computed F_{ST} with the realSFS program in ANGSD (Nielsen et al., 2012) using SNPs with a minor allele frequency ≥ 0.05 (*-minMaf* flag). The significance of each pairwise F_{ST} comparison was evaluated with 1,000 permutations by randomly allocating individuals to one of the two populations. We finally tested isolation by distance (IBD) using a Mantel test (Mantel, 1967) and plotted the relationship between genetic *vs.* geographic distances.

We applied an *approximate Bayesian computation* (ABC) approach similar to previous studies (Lesturgie, Planes, et al., 2022; Maisano Delser et al., 2019; Peter et al., 2010) in all sampling sites to further investigate the presence of population structure. This approach is particularly helpful in the Atlantic Ocean (AO) where only one locality was sampled (Brazil; BRA population). Briefly, we designed three demographic models (Supp. Figure 2.30): (1) NS (No Structure) which represents a panmictic population where N_{mod}, the modern effective size instantly changes to N_{anc}, the ancestral effective size, at T_s (time shift) generations; (2) FIM (Finite Island Meta-population) which represents a finite island meta-population model composed of 100 demes exchanging symmetrically Nm migrants per generation with each other. All demes were instantaneously colonised, T_{col} generations ago, from an ancestral population of size N_{anc}. (3) SS (Stepping-Stone) which represents a stepping-stone model where the 100 demes are arranged in a two-dimensional grid and where migration is only allowed symmetrically in both directions between the four nearest neighbouring demes. We performed 50,000 coalescent simulations under each model using FASTSIMCOAL v.2.6.0.3 (Excoffier et al., 2013) extracting parameters from the

prior distributions displayed in Supp. Table 2.15. Model selection was evaluated by the random forest classification method implemented in the *abcRF* package in R (Pudlo et al., 2016). We used the SFS, θ_π and *TD* as summary statistics and further added the first two axes of the Linear Discriminant Analysis in the dataset as suggested by (Pudlo et al., 2016) to increase the classification method accuracy. The number of trees was chosen by checking the evolution of the out-of-bag error.

2.4.6.5. Genetic diversity and effective population size variation

We created one reference sequence per population in order to maximise the number of loci assembled. We filtered the sites with missing data by setting the *-minInd* flag in ANGSD to the total number of individuals in each population. The filtered dataset was then used to generate a *site allele frequency likelihood* (*saf*) file, where genotype likelihoods were computed using the SAMtools method (*-GL=1* flag). The folded site frequency spectrum (SFS) was directly computed from the filtered *saf* datasets through the realSFS program (Nielsen et al., 2012). Nucleotide diversity (θ_π), Watterson's theta based on segregating sites (θ_w;(G. A. A. Watterson, 1975)) and Tajima's *D* (*TD;* (Tajima, 1989)) were computed with custom script from the SFS. Significance of *TD* was evaluated after 1,000 coalescent simulations of a constant population model with size θ_w. We reconstructed the variation in the effective population size (N_e) through time by running the STAIRWAYPLOT v.0.2 software (Liu & Fu, 2020) with singletons, where the composite likelihood is evaluated as the difference between the observed (folded) SFS and its expectation under a specific demographic history.

2.4.6.6. Population divergence and migration rate estimation

Based on the results of the previous analyses, we devised five alternative Isolation/Migration (IM) models of divergence between IP and AO regions using the composite likelihood method implemented in FASTSIMCOAL (Excoffier et al., 2013). We presented in Figure 3 the model richest in parameters, the remaining four representing simplified versions nested within it. Hereafter, a brief description of the five models going from the most complex to the simplest: (a) **IM-full**: the two ocean regions with their respective modern effective population sizes, $N_{mod_{IP}}$ and $N_{mod_{AO}}$, diverged at T_{div} from an ancestral population of effective population size N_{anc}. Due to the STAIRWAYPLOT results, we allowed the two modern effective population sizes $N_{mod_{IP}}$ and $N_{mod_{AO}}$ to change to $N_{anc_{IP}}$ and $N_{anc_{AO}}$ following an exponential dynamic in $T_{s_{IP}}$ and $T_{s_{AO}}$ years respectively. Migration is defined by two time periods: m_1 representing the migration rate

occurring between time 0 until T_{mig} and m_2 between time T_{mig} until T_{div}. The migration matrix in each time period is asymmetric: for instance, $m_{1_{AO/IP}}$ represents the forward migration rate from AO to IP and $m_{1_{IP/AO}}$ from IP to AO. In summary, the model is defined by thirteen parameters: five effective population sizes, four migration rates and four historical events; (b) **IM-anc**: same as IM-full with ancestral migration only between T_{mig} and T_{div}. The model is defined by eleven parameters, the two m_1 migration rates being removed; (c) **IM-rec**: same as IM-anc, but with recent migration only occurring between time 0 until T_{mig}, keeping only the two m_1 migration rates; (d) **IM-bsc**: the classic model where migration is constant from time 0 until T_{div} (i.e. $m_1 = m_2 = m$ (28)). We modelled the variation in effective size of the two regions similarly to the other models, for a total of ten parameters; (e) **IM-div**: a pure divergence model with no migration. This is defined by eight parameters: the five effective population sizes and three historical events ($T_{div}, T_{s_{IP}}, T_{s_{AO}}$). The analyses are based on the folded 2D-SFS computed by ANGSD between six individuals from Brazil (representing the AO) and six from the Indo-Pacific (IP). This sample size was chosen to obtain a balanced design and to maximise the number of SNPs shared among the two ocean basins. Similarly, in each basin, we selected the individuals presenting the smaller proportion of missing data to further increase the number of joint SNPs. To maximize the observed 2D-SFS we applied the following options in FASTSIMCOAL: -N 300,000 (number of coalescent simulations), -L 40 (number of expectation-maximization (EM) cycles), and -C 10 (minimum observed SFS entry count considered for parameter estimation). For all model parameters we used wide search ranges with uniform distributions (Table 2.14). We ran each model 100 times in order to determine the maximum likelihood parameters and to perform model selection using the Akaike's information criterion comparing the best run of each model (Excoffier et al., 2013). To check the robustness of the model selection procedure and to take into account the presence of linked sites in our dataset, we further examined the likelihood distribution obtained based on 100 expected 2D-SFS simulated under the parameters estimated in the best run of each model, each approximated with 10^6 coalescent simulations. This procedure is needed to take into account the variance in the likelihood estimation given our dataset: if the distributions obtained by the various models do overlap, the difference in the estimated likelihoods of our models is not significant (Meier, Sousa, et al., 2017). Finally, we determined the confidence interval of the parameter estimated under the best run of our best model by parametric bootstrapping. The 2D-SFS was

bootstrapped 100 times using FASTSIMCOAL and each of these datasets was analysed under the same conditions as the original data (one hundred independent runs for each dataset). Calibrating the molecular clock is crucial to obtain accurate estimates of demographic parameters and historical events, but it is challenging when fossil records and/or orthologous loci from an outgroup are lacking. Here, all demographic inferences were performed using the RAD-seq mutation rate of $\mu = 1.93 \times 10^{-8}$ per site and per generation previously used for the tiger shark (Lesturgie, Planes, et al., 2022), and the generation time was set to 10 years (Cortés, 2002; Pirog et al., 2019).

2.4.7. Supplementary Information

2.4.7.1. Acknowledgements

We thank Serge Planes for providing DNA samples of all specimens but Reunion Island, Gavin Naylor for critical reading of the manuscript, and Thibaut Caley for discussion above the Agulhas current. We thank Andrea Benazzo for help with Angsd. We are grateful to the Genotoul bioinformatics platform Toulouse Midi-Pyrenees (Bioinfo Genotoul; http://bioinfo.genotoul.fr) for providing computing resources. All data and analyses from Reunion Island were produced with the help of A. Pirog and financially supported by DEAL Réunion under the project ECoReCo-Run.

2.4.7.2. Author contributions

S.M. and H.M. conceived the project. S.M., P.L and H.L. analysed the data with input from P.M.D. E.C., H.M. and S.J. provided reagents and samples. A.S. and P.C.B performed the molecular lab work. P.L., S.M. and H.M. wrote the manu- script with input from all the others. All authors read and approved the final manuscript.

2.4.7.3. Funding

This work was supported by two ATM grants (2016 and 2017) from the Muséum National d'Histoire Naturelle to Stefano Mona.

2.4.7.4. Availability of data and materials

Fastq sequence files are available from the GenBank at the National Center for Biotechnology Information short-read archive database (BioProject ID: PRJNA887936).

2.4.7.5. Supplementary tables

Table 2.15. Matrix of pairwise F_{ST} values (lower triangle) and significancy (upper triangle). F_{ST} values in bold are significantly different from 0 ($P \leq 0.001$).

	BRA	RUN	AUS_N	COR	AUS_E	NCA
BRA		$P \leq 0.001$	$P \leq 0.001$	$P \leq 0.001$	$P \leq 0.001$	$P \leq 0.001$
RUN	**0.12**		NS[1]	NS	NS	$P \leq 0.001$
AUS_N	**0.12**	0.02		NS	NS	NS
COR	**0.13**	0.03	0.02		NS	NS
AUS_E	**0.12**	0.02	0.02	0.02		$P \leq 0.001$
NCA	**0.12**	**0.03**	0.02	0.02	**0.03**	

[1]NS: Not Significant

Table 2.16. Prior distribution of the parameters of the Finite Island (FIM), Stepping Stone model (SS) and Non-Structured (NS) models. Nm represents the number of migrants exchanged per generation either with the four closest neighbouring demes (SS) or with any deme in the matrix (FIM). N_{mod} represents the modern effective population size of the NS model. N_{anc} represents the ancestral effective population size either of the founding deme (in the structured models) or in the panmictic population (NS model). T_{col} is the colonization time of the array of deme (FIM and SS only) and T_c is the time when a change in effective population size happened in the panmictic population (NS only). Time parameters are in generations.

FIM	Nm^*	T_{col}§	N_{anc}
	P*: 0.001 - 100	U¶: 1 – 300,000	U: 100 – 100,000
SS	Nm^*	T_{col}§	N_{anc}
	P: 0.001 - 100	U: 1 – 300,000	U: 100 – 100,000
NS	N_{mod}	T_s§	N_{anc}
	U: 1 – 100,000	U: 1 – 300,000	U: 1 – 100,000

* P: the prior distribution of *Nm* is the product of two uniforms (one for *N* and one for *m*).
¶ U: uniform distribution.

Table 2.17. Confusion matrix of the model selection procedure and posterior probability for the most likely model explaining the structuring: rows indicate the simulated models and columns the votes (in %) attributed by the ABC-RF algorithm to each of them.

		Attributed votes (%)			Class.	Posterior
		FIM	NS	SS	error	Probability
BRA	FIM	75.848	4.438	19.714	0.24152	0.63
	NS	1.466	97.158	1.376	0.02842	
	SS	20.584	4.088	75.328	0.24672	
RUN	FIM	44885	1376	3739	0.1023	0.79
	NS	466	49095	439	0.0181	
	SS	4443	781	44776	0.10448	
AUS_N	FIM	40806	1867	7327	0.18388	0.48
	NS	598	48738	664	0.02524	
	SS	7591	1466	40943	0.18114	
COR	FIM	37878	2121	10001	0.24244	0.69
	NS	755	48450	795	0.031	
	SS	10344	1917	37739	0.24522	
AUS_E	FIM	40162	1849	7989	0.19676	0.86
	NS	623	48720	657	0.0256	
	SS	8334	1562	40104	0.19792	
NCA	FIM	42620	1584	5796	0.1476	0.89
	NS	543	48872	585	0.02256	
	SS	6123	1184	42693	0.14614	

2.4.7.6. Supplementary figures

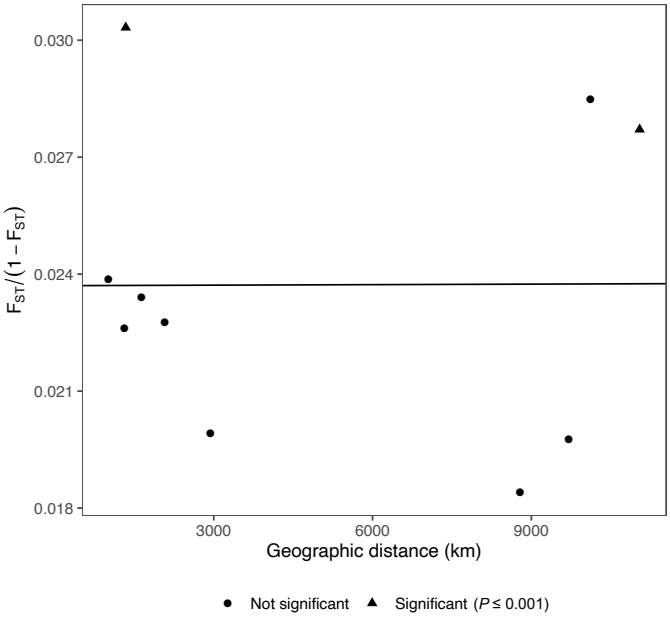

Figure 2.27. Isolation by distance (IBD) plot within the Indo-Pacific. Pairwise genetic distances ($F_{ST}/(1-F_{ST})$) are plotted against geographic distances between Indo-Pacific sampling sites.

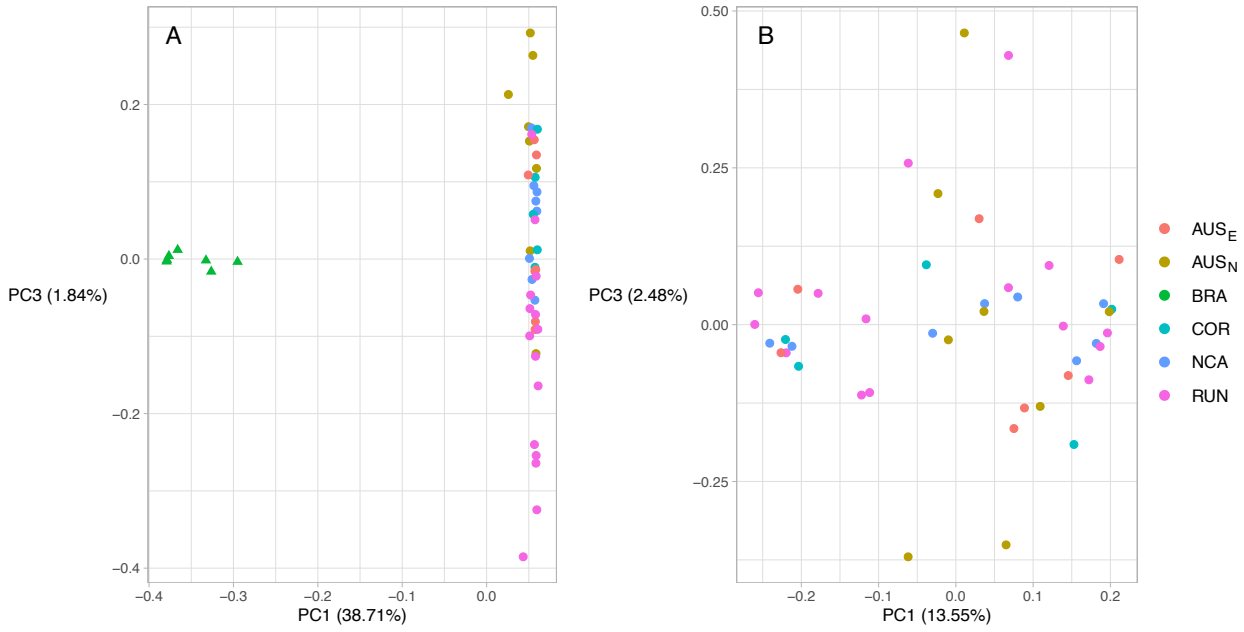

Figure 2.28. Principal Component Analysis (PCA) computed with: (A) all individuals (n = 50) and (B) Indo-Pacific individuals only (n = 43). The axes represented in both panels are the first and the third component.

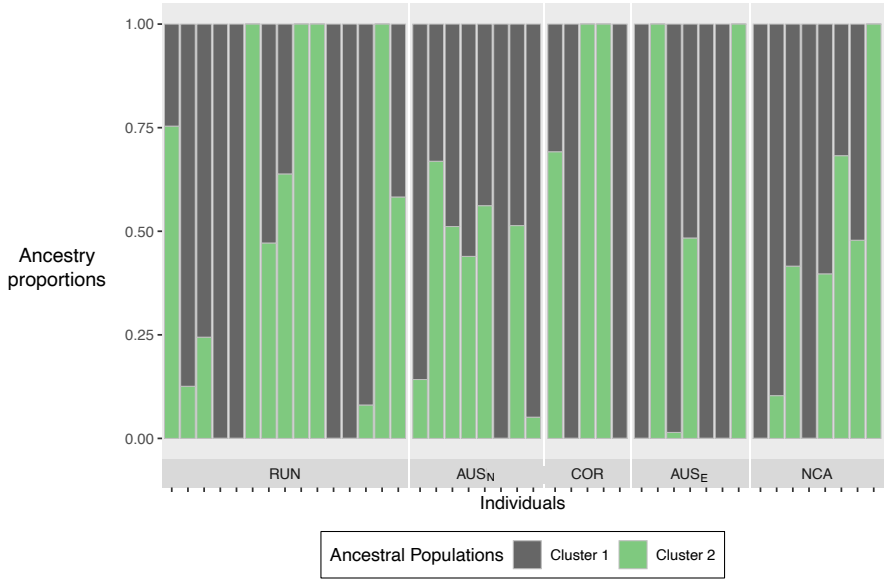

Figure 2.29. Ancestry proportions retrieved using the nmf algorithm with K=2 ancestral populations for Indo-Pacific samples performed with PCANGSD.

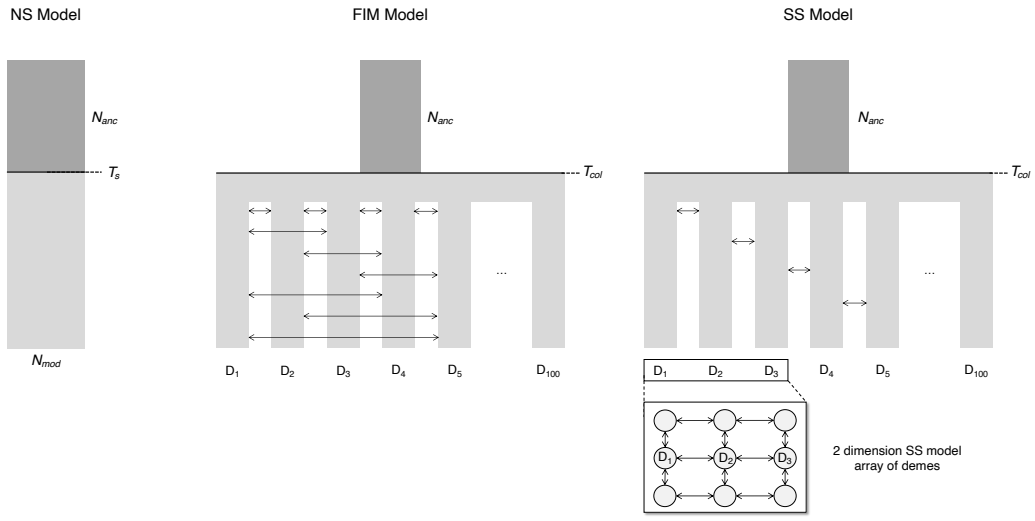

Figure 2.30. Evolutionary scenarios used to investigate the population structure of the Atlantic Ocean based on data from Brazil population through an Approximate Bayesian Computation (ABC) framework. NS (No Structure) is an unstructured model where the modern effective size (N_{mod}) instantaneously changes to N_{anc}, at time shift T_s generations. FIM (Finite Island Meta-population) represents a finite island meta-population model with 100 demes that have been instantaneously colonised T_{col} generations ago, from an ancestral population of size N_{anc}. Demes are allowed to exchange migrants with any other. SS (Stepping-Stone) is similar to FIM but the migrants are only exchanged between the four nearest neighbours in a two-dimensional grid.

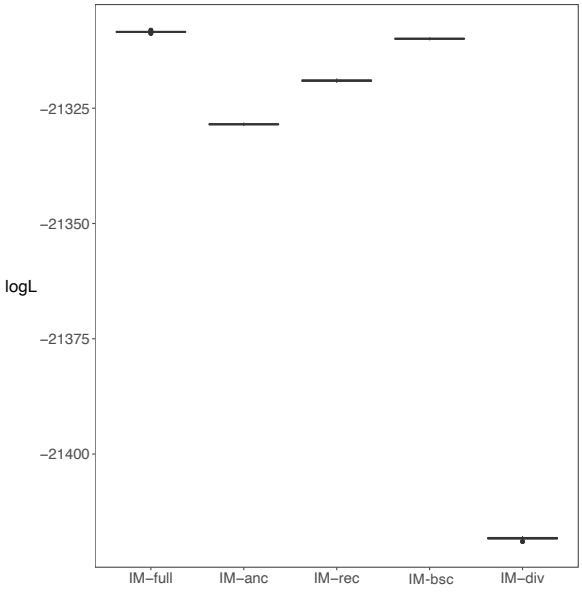

Figure 2.31. Akaike Information Criterion (AIC) values for the five isolation/migration models and the associated ranking on the x-axis. Boxplots represent the likelihood distribution of the data evaluated under the best parameter estimates for each of the five models (presented in Figure 2) after 100 replicates. The models are presented from the richest in parameters (IM-full, 13 parameters) to the poorest (IM-div, 8 parameters).

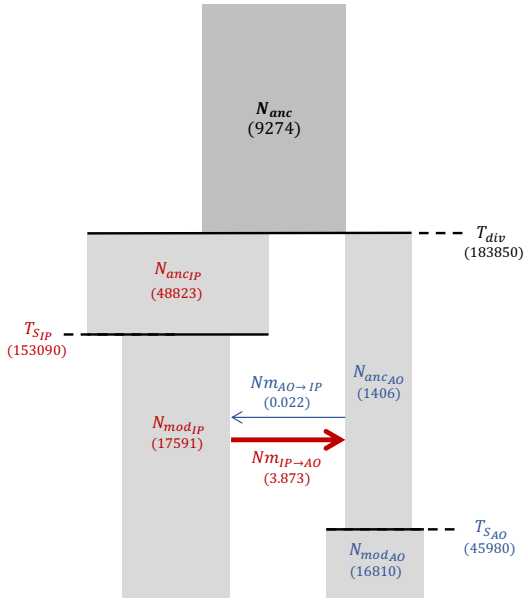

Figure 2.32. Maximum likelihood for the parameter estimated by fastsimcoal under model IM-bsc, representing two populations from each ocean basin with an effective size that changed $T_{s_{IP}}$ and $T_{s_{AO}}$ years ago from a modern effective size ($N_{mod_{IP}}$ and $N_{mod_{AO}}$) to an ancestral effective size ($N_{anc_{IP}}$ and $N_{anc_{AO}}$). The two populations are connected by an asymmetrical number of migrants constant from 0 to T_{div} ($Nm_{IP\rightarrow AO}$ and $Nm_{AO\rightarrow IP}$) and diverged T_{div} years ago from an ancestral population of size N_{anc}.

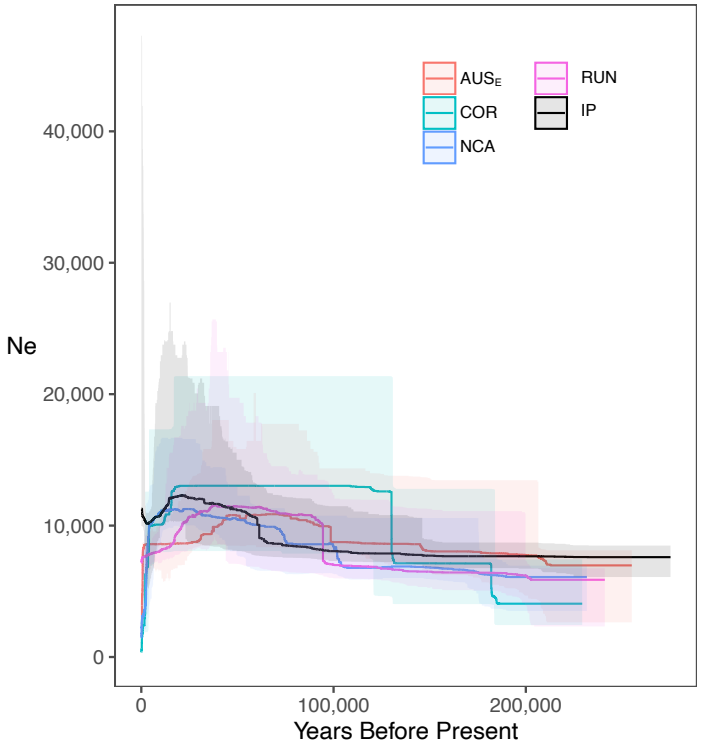

Figure 2.33. Variation of the effective population size (N_e) through time and its 75% confidence interval estimated by the STAIRWAYPLOT for all sampling sites. AUS_E: East Coast of Australia; AUS_N: North Coast of Australia; BRA: Brazil; COR: Coral Sea; NCA: New Caledonia; RUN: Reunion Island; IP: pooled individuals from AUS_E, COR, NCA and RUN sampling locations.

2.5. Like a rolling stone: Colonization and migration dynamics of the gray reef shark (*Carcharhinus amblyrhynchos*)

This article has been published in *Ecology and Evolution.*

Authors:

Pierre Lesturgie, Camrin D. Braun, Eric Clua, Johann Mourier, Simon R. Thorrold, Thomas Vignaud, Serge Planes, Stefano Mona

2.5.1. Abstract

Designing appropriate management plans requires knowledge of both the dispersal ability and what has shaped the current distribution of the species under consideration. Here we investigated the evolutionary history of the endangered grey reef shark (*Carcharhinus amblyrhynchos*) across its range by sequencing thousands of RAD-seq loci in 173 individuals in the Indo-Pacific (IP). We first bring evidence of the occurrence of a range expansion (RE) originating close to the Indo-Australian Archipelago (IAA) where two stepping-stone waves (east and westward) colonized almost the entire IP. Coalescent modeling additionally highlighted a homogenous connectivity (Nm~10 per generation) throughout the range, and an isolation by distance model suggested the absence of barriers to dispersal despite the affinity of *C. amblyrhynchos* to coral reefs. This coincides with long-distance swims previously recorded, suggesting that the strong genetic structure at the IP scale (F_{ST} ~ 0.56 between its ends) is the consequence of its broad current distribution and organization in a large number of demes. Our results strongly suggest that management plans for the grey reef shark should be designed on a range-wide rather than a local scale due to its continuous genetic structure. We further contrasted these results with those obtained previously for the sympatric but strictly lagoon-associated *Carcharhinus melanopterus*, known for its restricted dispersal ability. *C. melanopterus* exhibits similar RE dynamic, but is characterized by stronger genetic structure and a non-homogeneous connectivity largely dependent on local coral reefs availability. This sheds new light on shark evolution, emphasizing the roles of IAA as source of biodiversity and of life history traits in shaping the extent of genetic structure and diversity.

Keywords: Meta-population, Rad-seq, demographic history, range expansion, *Carcharhinus amblyrhynchos*, *Carcharhinus melanopterus*.

2.5.2. Introduction

More than 37% of shark species are currently threatened with extinction (Dulvy et al., 2021) and less than 30% are on stable or increasing population trend according to the International Union for Conservation of Nature (IUCN) Red List of threatened species. As meso or apex predators, they hold important roles in their ecosystems (Bornatowski et al., 2014) and their decline has already shown negative cascading effects on food web structure (Friedlander & DeMartini, 2002; Myers et al., 2007). Although local-scale conservation programs have been established, their efficiency has been questioned for some species of sharks (Robbins et al., 2006; Speed et al., 2016). For instance, local-scale management might not always be consistent with the home range size and the dispersal ability of sharks (but see (Dwyer et al. 2020)). Genetics and ecological evidence have identified both species with very restricted home ranges (Whitney et al. 2012; Mourier et al. 2013) and species capable of crossing large expanses of ocean (Bailleul et al., 2018; Corrigan et al., 2018; Pirog et al., 2019). Designing appropriate management actions is therefore a difficult task requiring the knowledge of both the dispersal ability of the species under investigation and the existence of barriers to gene flow, which are often hard to identify in the marine realm.

Population genomics is becoming increasingly important in this context, particularly because of the large amount of data provided by the emergence of next generation sequencing approaches (NGS). It is now possible to assess the genetic diversity of model or non-model species at an unprecedented level of accuracy (Benazzo et al., 2017; Steiner et al., 2013). However, genetic diversity alone does not provide clues on the evolutionary trajectory of a species and a careful modelling is required to fully understand its demographic history as well as the conservation challenges to be faced. Unfortunately, for computational reasons, many commonly used software implement, under different algorithms, *unstructured* models, i.e., models that consider the population under investigation as isolated or panmictic (Heled & Drummond, 2008; Heller et al., 2013; H. Li & Durbin, 2011; Liu & Fu, 2015). Except for highly vagile species which are panmictic at a large scale (Corrigan et al., 2018; Lesturgie, Planes, et al., 2022; Pirog et al., 2019), broadly distributed shark species are more likely organized in meta-population(s) throughout their range (Maisano Delser et al., 2016, 2019; Momigliano et al., 2017; Pazmiño et al., 2018). The application of *unstructured* models to species organised in meta-populations yield spurious signatures of effective populations size (N_e) changes through time (Chikhi et al., 2010; Maisano Delser et al., 2019; Mazet et al., 2015, 2016), with potentially dangerous consequences in terms of conservation

policies. However, recent studies have highlighted the usefulness of such models to characterize the gene genealogy of the sampled lineages which in turn reveals important features of the meta-population (Arredondo et al., 2021; Lesturgie, Planes, et al., 2022; Rodríguez et al., 2018) This emphasizes the necessity to couple complex meta-population models and *unstructured* models when uncovering the demographic history of a species.

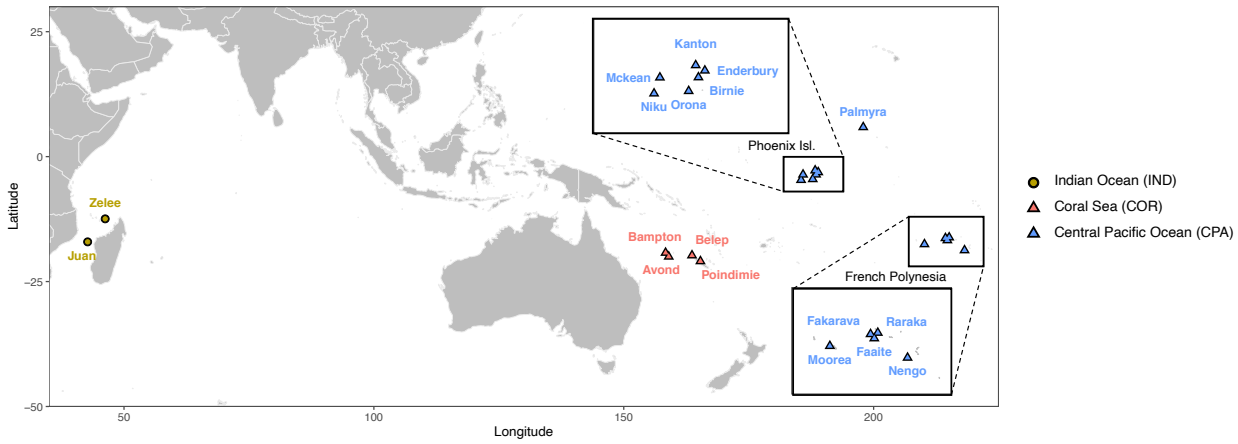

Figure 2.34. Map of the sampling sites. From west to east, Indian Ocean (IND): Juan (n = 13) and Zelee (n = 6); Chesterfield islands (CHE): Bampton (n = 10) and Avond (n = 5), New Caledonia (NCA): Belep (n = 7) and Poindimie (n = 5); Phoenix islands (PHO): Niku (n = 21), Mckean (n = 7), Orona (n = 11), Kanton (n = 10), Birnie (n = 2) and Enderbury (n = 13); Palmyra (PAL, n = 38); French Polynesia (POL): Moorea (n = 5), Fakarava (n = 17), Faaite (n = 1), Raraka (n = 1), and Nengo (n = 1). Colours represent the region of origin of the sampling sites: Indian Ocean (IND, yellow), Coral Sea (COR, red) and Central Pacific Ocean (CPA, blue).

Here we investigated the evolutionary history of the grey reef shark *Carcharhinus amblyrhynchos*, a coral reef-associated shark inhabiting the tropical Indo-Pacific. While *C. amblyrhynchos* is considered one of the most abundant reef sharks in the Indo-Pacific, it is listed as Endangered on the IUCN red list of threatened species. With a mean size of ~190 cm (Compagno, 2001), *C. amblyrhynchos* inhabits either fringing or barrier reefs and displays patterns of reef fidelity (Barnett et al., 2012; Espinoza et al., 2014) as well as philopatry (Field et al., 2011). Tagging studies have indicated long range movement up to ~900 km (Barnett et al., 2012; Bonnin et al., 2019), which raise questions about the extent of residency patterns for this species. Previous molecular studies using both microsatellites and Rad-sequencing did not find signatures of genetic structure at a low geographic scale such as the Great Barrier Reef (Momigliano et al., 2015, 2017), eastern Australia and Indonesia (Boussarie et al., 2022) and the Phoenix Islands archipelago (Boissin et al., 2019). Conversely, isolation by distance patterns have been found at larger scale

and some evidence suggests that coastal abundance of reef can fuel genetic exchanges, while oceanic expanses are barriers to gene flow (Boissin et al., 2019; Boussarie et al., 2022; Momigliano et al., 2017).

To shed light on these contrasting findings, we sequenced DNA from 203 individuals of *C. amblyrhynchos* sampled at 18 sites from the eastern Indian Ocean to French Polynesia (Figure 2.34) following a double digest restriction site associated DNA protocol (Peterson et al., 2012). The large panel of assembled loci was used to: (*i*) detect the occurrence and origin location of a range expansion (RE); (*ii*) investigate its demographic history by implementing both meta-population and *unstructured* models; (*iii*) reassess the population structure of the grey reef shark in the Indo-Pacific. We finally compared the results here obtained with those previously found in the blacktip reef shark (*Carcharhinus melanopterus* (Maisano Delser et al., 2016, 2019)). The two species share a very similar distribution in the Indo-Pacific but are characterized by different habitat preferences and life-history traits, providing an excellent opportunity to improve our knowledge on the biology of sharks.

2.5.3. Material and Methods

2.5.3.1. Sampling and Rad sequencing

We collected 203 samples of *C. amblyrhynchos* that covered most of its longitudinal distribution range (Figure 2.34), with two sampling sites in the Mozambique Channel in the western Indian Ocean (IND – Juan de Nova and Zélée bank) and 16 in the Pacific Ocean (PAC). Among the PAC sampling sites, four were chosen in the Coral Sea (COR): two in the Chesterfield Islands (Bampton and Avond) and two in New Caledonia (Belep and Poindimie). The remaining samples came from the Central and Easter Pacific (CPA): six in the Phoenix Islands (Enderbury, Kanton, McKean, Niku, Orona and Birnie) one in Palmyra Island and five in French Polynesia (Fakarava, Moorea, Faaite, Raraga and Nengo) (Figure 2.34, Table 2.18). Total genomic DNA has been extracted and conserved in 96% ethanol using QIAGEN DNeasy Blood and Tissue purification kit (Qiagen, Hilden, Germany) according to the manufacturer's protocols. We followed the double digest restriction site associated DNA (dd-RADseq) protocol of (Peterson et al., 2012) to create a genomic library, using EcoRI and MSFI as restriction enzymes. We selected fragments of ~400 bp length and sequenced with Illumina HiSeq 2500 machine (single-end, 125 bp).

In the absence of a reference genome, we assembled loci *de novo* using *Stacks* v.2.5 (Rochette et al., 2019). Briefly, we demultiplexed the reads through the *process_radtags.pl* script and assembled the loci using the *denovo_map.pl* pipeline with the parameters *m=3* (minimum read depth to create a stack), *M=3* (number of mismatches allowed between loci within individuals) and *n=3* (number of mismatches allowed between loci within catalogue). We found a mean depth of coverage (over individuals and loci) of ~10x (see Results). Previous work suggested that such low-coverage value may bias a correct genotype calling under the algorithm implemented in *Stacks* v.1, *Stacks* v.2 and *PyRAD* by skewing the site frequency spectrum (SFS) towards an excess of low frequency variants ((Mona et al., 2023); see supplementary materials for details). For this reason, we followed two different bioinformatics pipelines: the first to obtain a dataset to perform analyses based on the SFS (genetic diversity, range expansion and historical demographic inferences) and the second to investigate population structure, for which low frequency variants are not informative and are removed before the downstream analyses.

2.5.3.2. Genetic diversity

We followed the genotype free estimation of allele frequencies pipeline implemented in the software *ANGSD* v.0.923 (Korneliussen et al., 2014). This approach has been suggested to be more efficient for low to medium coverage NGS data than SNPs calling algorithms (Korneliussen et al., 2014). *ANGSD* requires a reference sequence to work. To this end, we followed the framework proposed by (Heller et al., 2021; Khimoun et al., 2020) which we applied to each sampling site separately to maximise the number of loci: i) we assembled Rad loci present in at least 80% of the sampled individuals using *Stacks* with the same parameters as above (i.e., *m=M=n=3*); ii) we concatenated the consensus sequences for each locus, to which we added a stretch of 120 "N" in order to facilitate mapping, to create an artificial reference sequence; iii) we mapped raw reads from individual *fastq* files using the *bwa-mem* algorithm with default parameters (H. Li & Durbin, 2009) against the artificial reference sequence. Using *ANGSD* filters, we discarded *(1)* sites with a coverage < 3 (using the flag *-minIndDepth 3*) *(2)* poor quality and mis-aligned reads (with default parameters and flags *-minQ20* and *-minMapQ 20*), *(3)* poor quality bases (with default parameters and flags *-baq 1* and *-C 50*). We further removed the last 5bp of each locus, SNPs heterozygous in at least 80% individuals, and loci with more than 5 SNPs. We finally filtered all missing data by applying the *-minInd* filter equal to the total number of individual present in each sampling site (Table 2.18). We then created a *site allele frequency likelihood (saf)* file by using the SAMtools

genotype likelihood computation method with the *-GL=1* flag (H. Li & Durbin, 2009) and finally computed the folded *site frequency spectrum* (SFS) from the *saf* files using the *RealSFS* program implemented in *ANGSD*. We computed the mean pairwise difference (θ_π), the number of segregating sites (Watterson's Theta, θ_w) and Tajima's D (*TD*) directly from the SFS. θ_π and θ_w were standardized per site (i.e., by taking into account both monomorphic and polymorphic loci) and significance of *TD* was evaluated under 1,000 coalescent simulations of a constant population model with size θ_π.

Table 2.18. Summary Statistics. Sample size (n), total number of loci (monomorphic included) (n_{loci}) and SNPs (n_{SNP}), mean pairwise difference (θ_π), Watterson theta (θ_w), Tajima's D (TD) for all sampling sites (ranged from west to east).

Region	Group	Sampling site	n	n_{loci}	n_{SNP}	θ_π†	θ_w†	TD‡
IND	IND	Juan	13	95027	45635	1.18	1.09	**0.32**
		Zelee	6	146858	62674	1.30	1.23	**0.26**
COR§	CHE	Bampton	10	89958	82869	2.14	2.26	**-0.22**
		Avond	5	125710	87817	2.10	2.15	**-0.12**
	NCA	Belep	7	120038	103258	2.30	2.35	**-0.11**
		Poindimie	5	107464	72995	2.07	2.09	**-0.05**
CPA§	PHO	Niku	21	49922	53349	2.02	2.16	**-0.25**
		McKean	7	112711	88258	2.13	2.14	-0.01
		Orona	11	81725	75423	2.15	2.20	**-0.09**
		Kanton	10	99720	87202	2.12	2.14	**-0.05**
		Birnie¶	2	-	-	-	-	-
		Enderbury	13	76314	72221	2.09	2.16	**-0.12**
	PAL	Palmyra	38	35594	36982	1.66	1.84	**-0.35**
	POL	Moorea	5	104050	68380	2.03	2.02	0.02
		Fakarava	17	71715	66559	2.01	1.97	**0.08**
		Faaite¶	1	-	-	-	-	-
		Raraka¶	1	-	-	-	-	-
		Nengo¶	1	-	-	-	-	-

† Mean pairwise difference and Watterson theta are expressed per site and are multiplied by a 10^3 factor.
‡ Tajima's D values in bold are significant (P <0.001).
§ COR and CPA regions are from the Pacific Ocean (PAC).
¶ Summary statistics were not computed in sampling sites with n < 5.

2.5.3.3. Range Expansion

Genetic diversity, here measured in each sampling site as θ_π, is expected to decay as a function of the distance from the origin of the range expansion (Ramachandran et al., 2005). Geographic

distances were computed in order to take into account ecological features as it may better represent the capacity of individuals to move between two points than linear distances. To that end, we constructed a raster of 67894 cells using the R package *raster* (Hijmans, 2020) where each cell corresponds either to land, open sea, seamount or reef habitat. Permeability coefficients were fixed respectively to 0 and 1 for land and open sea, whereas coefficients for coral reefs and seamounts were varied between 1 and 100. We applied two constraints: coral reefs should always have the maximum relative permeability value (since they represent the only habitat for *C. amblyrhynchos*) and seamounts have permeability bounded within 1 and coral reefs' value. The most likely values were searched using a custom R script by maximising the correlation between the geographic and genetic distances between the sampled sites. Geographic distances were computed with the *gdistance* R package under the *Least Cost* (LC) criterion algorithm (van Etten, 2017) and genetic distances were measured by the F_{ST} (see below). After this step, we considered each marine cells of the raster to be a potential source of origin of the range expansion (RE) and computed its distance from the sampled sites under the LC criterion with the most likely permeability values previously estimated. We correlated these distances with the genetic diversity of each sampling site to identify areas with more negative values, which are likely associated with the origin of the RE (Ramachandran et al., 2005). We limited these analyses to the PAC sites to avoid possible bias due to the gap in our sampling distribution (i.e., the lack of samples between IND and the westernmost PAC site). Nevertheless, we verified the robustness of our results to the inclusion of IND sites.

2.5.3.4. Historical demographic inferences

To account and test for meta-population structure, we performed model selection as well as parameters estimation using an Approximate Bayesian Computation (ABC) framework (Bertorelle et al., 2010). We tested three demographic scenarios (Figure 2.35) for each sampling site, namely NS, FIM, and SST. *Model NS (no structure)*: going backward in time, NS represents a panmictic population where the effective population size switches instantaneously at T_c generations from N_{mod} to N_{anc}. *Model FIM (Finite Island Model)*: FIM represents a meta-population composed of a two-dimensional array of 10x10 demes (D_i), each of the same size N that exchanges Nm migrants with any other deme each generation. Going backward in time all demes merge into a single population of size N_{anc} at T_{col} generations. *Model SST (Stepping STone)*: SST is similar to FIM but demes exchange migrants only with their four closest neighbours. We performed 50000

simulations under each scenario and for each sampling site independently using *fastsimcoal2* (Excoffier & Foll, 2011). We run the model selection with the Random Forest classification method implemented in the package *abcRF* (Pudlo et al., 2016) using the SFS, θ_π, θ_w and *TD* as summary statistics, to which we added the first two components of the Linear Discriminant Analysis performed on the previous summary statistics as suggested by (Pudlo et al., 2016) to increase accuracy. We performed 50000 additional simulations under the most supported scenario in order to estimate the demographic parameters using the *abcRF* regression method (Raynal et al., 2019) with the same summary statistics as for the model selection. For all analyses, we performed the estimation twice to check for the consistency of the inferences. The number of trees was chosen by checking the out-of-bag error rate (OOB), and cross validation was performed for both parameter inference and model selection (hereafter, the confusion matrix) procedures. We finally modelled the variation of effective population size (N_e) through time in each sampling site with the *stairwayplot* (Liu & Fu, 2015). The *stairwayplot* assumes that the sampled lineages come from an isolated (panmictic) population (i.e., *unstructured*), which is not true in our case (see results). However, this method allows a powerful investigation of the underlying gene genealogy which provides useful elements for interpreting the evolutionary history of a meta-population (Lesturgie, Planes, et al., 2022). All demographic inferences were performed using a generation time of 10 years and a mutation rate of 1.93e-8 per generation and per site following (Lesturgie, Planes, et al., 2022).

2.5.3.5. Population structure

Population structure inferences were performed on the dataset obtained following the assembly pipeline implemented in *Stacks 2.5* as described above. After the *de novo* assembly step, the *population* script was called to keep loci present in at least 80% of the individuals per sampling site ($r = 0.8$) and with a *minor allele frequency* of 0.05, hence removing low frequency variants. We finally retained one random SNP per locus. Using a custom R script, we further filtered: *(i)* SNPs heterozygotes in more than 80% of the sample; *(ii)* loci with coverage higher than ~30x (which corresponds to the mean coverage plus twice the standard deviation); *(iii)* SNPs in the last 5bp of the assembled locus; and *(iv)* loci containing more than five SNPs, after visual inspection of the distribution of segregating sites per locus. The resulting dataset was used for the following analyses. 1) *sNMF* implemented in the R package *LEA* (Frichot & François, 2015): we investigated the number of ancestral clusters *K* by running the algorithm 10 times, with values of *K* ranging

from 1 to 8. We chose the most likely *K* using the cross-entropy criterion and displayed the admixture coefficients under the best run. 2) *DAPC* implemented in the R package *Adegenet* (Jombart et al., 2010): we varied *K* from 1 to 8 and chose the best values based on the BIC criterion. Linear discriminant functions were used to test whether individuals were correctly reassigned to the inferred clusters. 3) F_{ST}: we computed overall and pairwise F_{ST} between sampling sites with more than 5 individuals using the *PopGenome* (flag *nucleotide.F_ST*) library in R (Pfeifer et al., 2014) and tested its significance with 1000 permutations using a custom R script. Isolation by distance (IBD) was computed with a Mantel Test (Mantel, 1967) between the genetic ($F_{ST}/(1-F_{ST})$) and the geographic or LC distance matrices and tested by 1000 permutations with the *ade4* R package (Thioulouse & Dray, 2007). The Mantel test, similarly as before, was limited to PAC sites. To check for IBD in the Indian Ocean, we fit a linear model to the pairwise F_{ST} values computed between the PAC and IND sites and their respective geographic distances.

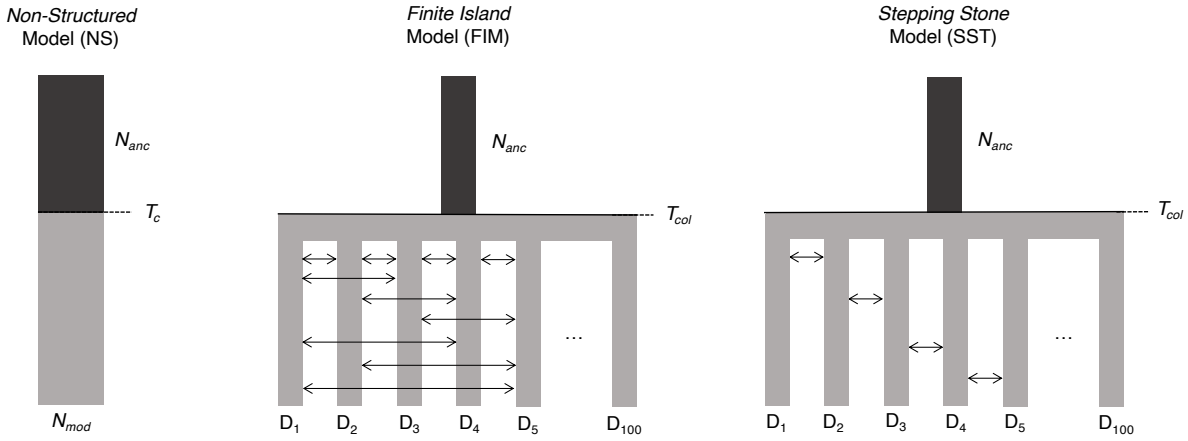

Figure 2.35. Demographic scenarios investigated in all populations with Nind ≥ 7 through an Approximate Bayesian Computation (ABC) framework. N_{anc}: ancestral effective population size; T_c: time of effective population size change (NS only); N_{mod}: modern effective population size (NS only); T_{col}: colonization time of the array of demes (FIM and SST); $D_{1\text{-}100}$: demes (FIM and SST). Arrows represent the migrants exchanged (Nm) between demes. Details on each scenario are presented in the main text.

2.5.4. Results

2.5.4.1. Genetic diversity

We discarded 30 individuals based on an excess of missing data after an initial *de novo* assembly. We found a mean depth of coverage of 10.77x (s.d. = 2.32) for the whole dataset. Summary statistics for all sampling sites are displayed in Table 2.18. The number of loci (monomorphic

included) and SNPs with no missing data ranged from 35594 to 146858 and from 36982 to 103258 respectively across sampling sites (Table 2.18). Genetic diversity (θ_π and θ_w) was lower in IND sampling sites than in PAC (Table 2.18). Tajima's *D* values were positive in IND sampling sites and in Fakarava, suggesting an excess of high frequency variants when compared to the standard neutral model. Conversely, we found negative and significant Tajima's *D* values in all other PAC locations (except for Moorea and Mckean), suggesting an excess of low frequency variants compared to the standard neutral model (Table 2.18).

Table 2.19. ABC estimation. Posterior probability (PP) of the Stepping Stone model (SST) and its parameters (median value and 95% credible interval in parentheses).

Region	Group	Sampling site	PP	*Nm*	T_{col}	N_{anc}
IND	IND	Juan	0.67	5.7 (1.77 - 17.72)	257800 (8086 - 658471)	21086 (399 - 52652)
COR§	CHE	Bampton	0.73	11.41 (3.97 - 19.03)	188782 (127761 - 577503)	45965 (27556 - 49856)
	NCA	Belep	0.51	7.8 (2.84 - 20.82)	241218 (112840 - 843171)	49239 (7346 - 56316)
CPA§	PHO	Enderbury	0.65	8.36 (2.9 - 20.9)	197070 (95260 - 678828)	43602 (14665 - 51030)
		Kanton	0.7	8.16 (2.84 - 16.55)	257718 (118094 - 789320)	41236 (2534 - 52613)
		McKean	0.6	7.09 (2.98 - 15.25)	621535 (158650 - 836223)	18881 (4968 - 51387)
		Niku	0.59	14.1 (3 - 30.55)	152035 (66928 - 598129)	43495 (9184 - 48625)
		Orona	0.48	7.7 (2.93 - 15.31)	269621 (137304 - 799518)	41680 (4575 - 51152)
	PAL	Palmyra	0.73	13.39 (4.16 - 27.22)	142756 (62402 - 445380)	32542 (9502 - 37524)
	POL	Fakarava	0.72	10.2 (2.68 - 15.34)	256744 (110875 - 780150)	40502 (3091 - 49533)
			Priors	**U* [0.0001 ; 100]	*U* [100 ; 1500000]	*U* [100 ; 100000]

***** The prior distribution of *Nm* is the product of two uniforms (one for *N* and one for *m*)
§ COR and CPA regions are from the Pacific Ocean (PAC).

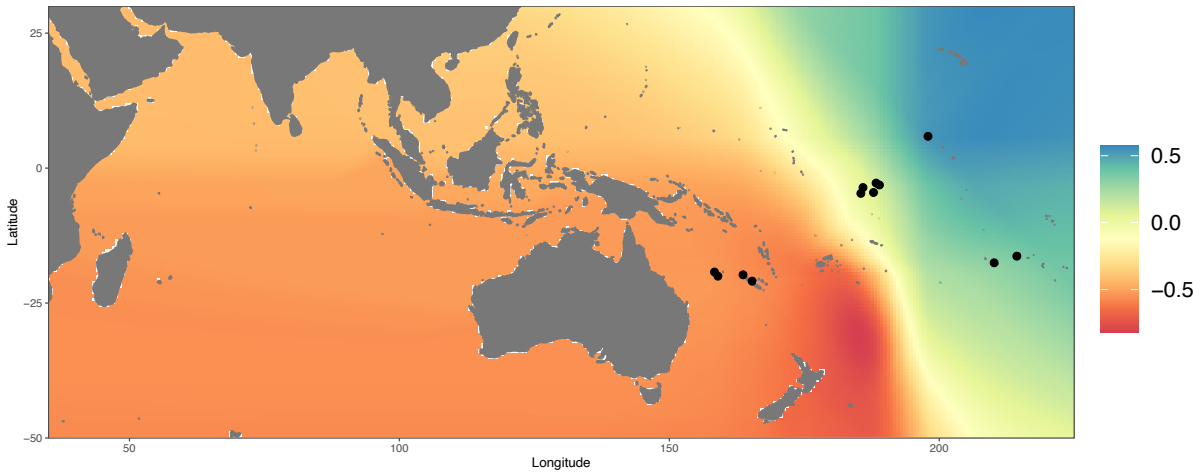

Figure 2.36. Correlation map between genetic diversity (θ_π) and Least Cost (LC) distances when considering Pacific Ocean sampling sites only. Each cell is coloured according to the correlation coefficient value computed between θ_π and the LC distance from the putative origin of the range expansion (RE). Black dots represent the sampling sites considered.

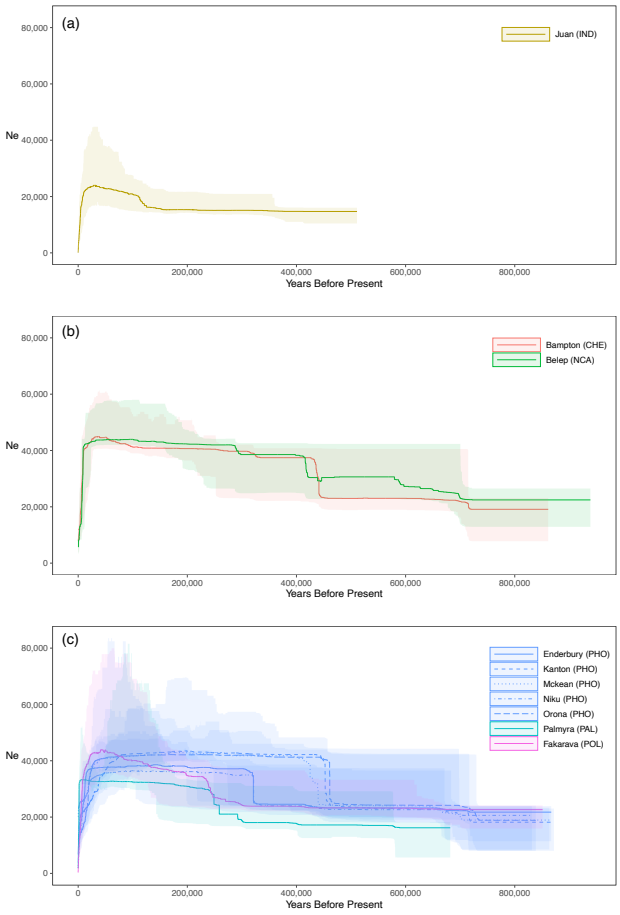

Figure 2.37. Variation of the effective population size (Ne) through time and its 75% confidence interval estimated by the *stairwayplot* for sampling sites of $n \geq 7$ in IND (a), COR (b) and CPA (c) regions.

2.5.4.2. Range Expansion

The permeability coefficients maximising the correlation between genetic and the LC distances were very similar between the three habitat types. Indeed, we estimated the values of 1:1.02:1.02 for open sea, coral reef habitat and seamounts respectively. These values were retained for the following RE and IBD analyses. We plotted the correlation map computed using PAC sites only in Figure 2.36. The most negative correlation coefficients are concentrated close to the COR sampling sites, suggesting that the most likely origin of the RE is slightly east to the IAA region (Figure 2.36). We found consistent results when adding IND sites to the analysis (Supp. Figure 2.40), despite the geographic unbalanced distribution of our samples.

2.5.4.3. Historical demographic inferences

We investigated the demographic history for all sampling sites with $n \geq 7$. We first used an ABC-RF framework to compare demographic scenarios (Figure 2.35). SST was the most supported scenario in all locations, with posterior probabilities ranging from 0.48 to 0.78 and similar classification error rate among locations (Table 2.19 and Supp. Table 2.20). The median *Nm* ranged from ~6 to ~14 (Table 2.19). Posterior distributions of *Nm* were overlapping and clearly distinct from the prior distribution (Supp. Figure 2.41), and both the squared mean error (SME) and the mean root squared error (MRSE) were small among locations, suggesting reliable estimates (Supp. Table 2.21). Posterior distributions of T_{col} overlapped among locations (Supp. Figure 2.41). Juan de Nova displayed a lower N_{anc} median value (~21k) than PAC sampling sites (ranging from ~34k to ~50k) although all credible intervals overlapped (Figure S2 and Table 2.19). Surprisingly, the ABC estimates of T_{col} and N_{anc} for the Mckean sampling site were inconsistent with any other PHO sampling sites (Supp. Figure 2.41and Table 2.19). However, both SME and the MRSE for these two parameters were generally one order of magnitude larger than those estimated for *Nm* in all sampling sites (Supp. Table 2.21), suggesting less accurate estimates for T_{col} and N_{anc}.

We further investigated the variation of *Ne* through time using the *stairwayplot* algorithm (Figure 2.37). We detected a broadly similar *Ne* dynamic across sampling sites that we summarized for simplicity in three time periods: looking forward in time we observed an ancestral expansion followed by a constant phase and a final systematic strong decrease in recent times (Figure 2.37). However, we found three main differences between IND and PAC sampling sites: i) the expansion time was around twice as recent in IND than in PAC (~180ky B.P. vs. ~400ky B.P); ii) the strength

of the expansion is much stronger in PAC sampling sites; iii) *Ne* during the constant period reached a value of ~40000 in PAC sampling sites and of only ~20000 in IND, consistent with the computed θ (Table 2.18). The PAC sampling sites showed a remarkably homogeneous *stairwayplot*, with only the peripheral sites (Fakarava and Palmyra) having a slightly more recent ancestral expansion (Figure 2.37).

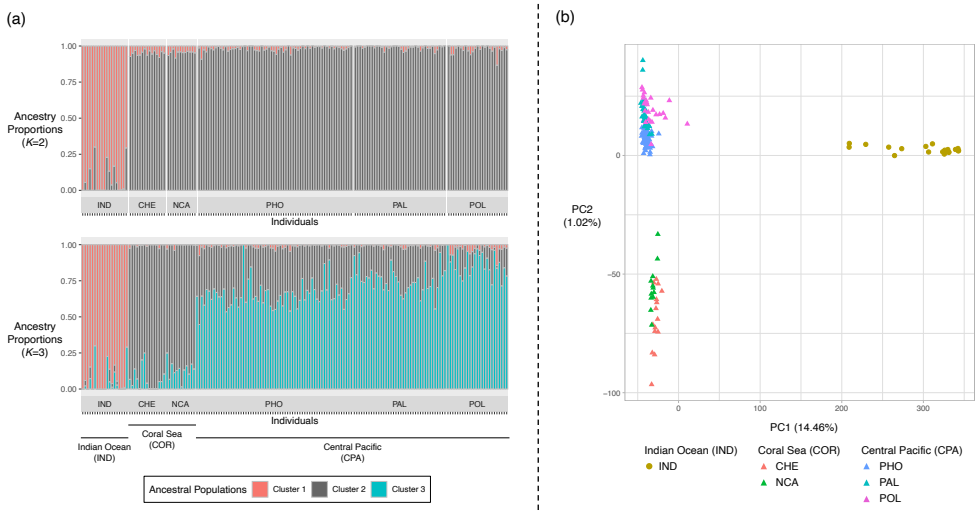

Figure 2.38. Individual-based population structure analyses. Ancestry proportions retrieved using the *sNMF* algorithm with K=2 and K=3 ancestral populations (a) and Principal Component Analysis (b).

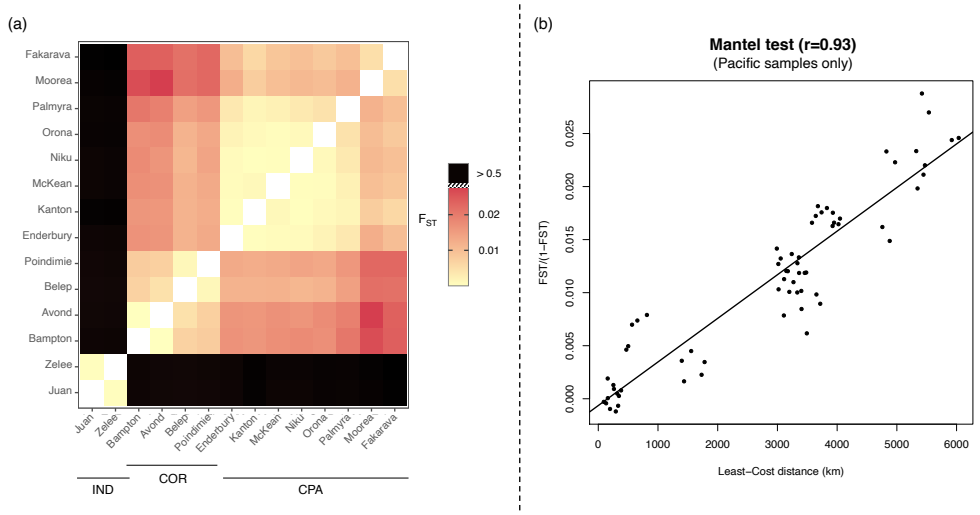

Figure 2.39. Population-based population structure analyses computed with populations of n ≥ 5. Heat map representing the pairwise F_{ST} values between sampling sites (a) and Isolation by distance (IBD) plot considering Pacific sampling sites only (b).

2.5.4.4. Population structure

After filtering, 88276 variable loci were retained to perform individual based structure analyses. Both *sNMF* and the *DAPC* clustering algorithms found *K*=2 as the most likely number of ancestral populations or clusters (Supp. Figures 2.42 and 2.43-a). The ancestral populations inferred by *sNMF* perfectly matched the two oceanic regions, namely the Indian and the Pacific Ocean: the ancestry proportion of *cluster 1* in IND samples ranged from 70% to 100% while the ancestry proportion of *cluster 2* in PAC samples ranged from 87% to 100% (Figure 2.38-a). This highlights slightly more admixture in IND than in PAC samples. We retained one LD function in the *DAPC* which correctly re-assigned all individuals from IND and PAC to *cluster 1* and *cluster 2* respectively (Supp. Figure 2.43-b). We further investigated *K*=3 under both algorithms and found three main results: i) the ancestral populations or clusters clearly identify three geographic areas corresponding to IND, COR, and CPA regions (Figure 2.38-a and Supp. Figure 2.44); ii) the ancestry proportion of *cluster 3* follows a clinal distribution, steadily increasing in frequency from West (Indian Ocean) to East (French Polynesia) (Figure 2.38-a); iii) all individuals belonging to the three areas are correctly re-assigned to the three clusters by the *DAPC* computed with two LD functions (Supp. Figure 2.43-b). We then computed a PCA which showed similar results, with the first principal component explaining ~14.5% of the total variance and clearly separating individuals coming from the two oceans (Figure 5b). The second axis segregated CPA from COR samples. In agreement with the cluster analyses, CPA and COR are only slightly differentiated as the second principal component explains only ~1% of the total variance. The second axis also suggested a clinal differentiation between the two clusters (Figure 2.38-b).

Population based analyses were performed on a reduced dataset excluding sampling sites with less than n=5 individuals. We therefore retained 14 sampling sites, n=168 individuals, and 88824 variable loci and obtained an overall F_{ST} = 0.25 (*p-value*< 0.001). The pairwise F_{ST} highlighted a strong differentiation between Indian and Pacific sampling sites with values ranging from 0.53 to 0.56 (and always significant, *p-value* ≤ 0.001, Supp. Table 2.22). In contrast, comparisons within oceanic regions never exceed 0.03 (Figure 2.39-a) with values not always statistically significant. Consistently with clustering results, a heatmap displaying pairwise F_{ST} values visually suggest the existence of the three clusters previously identified (Figure 6a). However, the average differentiation between COR and CPA is only slightly higher than within group comparisons (Figure 2.39-a). Moreover, we found a strong signature of isolation by distance (IBD) within the

Pacific Ocean (using PAC sites only), since the correlation between the F_{ST} and geographic or LC distance matrices was high and significant (Mantel test: r = 0.93; *p-value* < 0.001 in both cases, Figure 2.39-b). The correlation between genetic and geographic distances by considering only IND *vs*. PAC pairwise distances was also considerable although lower than in PAC region only (r = 0.77, Supp. Figure 2.44).

2.5.5. Discussion

2.5.5.1. Range expansion

Range expansions (RE) occur by a series of founder effects leading to the fixation of novel alleles and the decay in genetic diversity as colonization progresses (Excoffier et al., 2009). They also leave specific signatures in the gene genealogy of lineages sampled from a deme of the meta-population (Maisano Delser et al., 2016; Ray et al., 2003) and in the extent of population structure (Mona, 2017; Mona et al., 2014). Testing for the occurrence of a RE is therefore fundamental to understanding the evolutionary history of a species. Here, the spatial distribution of genetic diversity suggested the occurrence of a RE most likely starting east of the Indo-Australian Archipelago (IAA). The inferred origin area was large (Figure 2.36), likely due to low differences in θ_π between Pacific sampling sites (Table 2.18), but robust to the inclusion of samples from the Indian Ocean (Supp. Figure 2.40). The scenario of a RE was corroborated by other evidence. First, the strong and significant correlation coefficient between genetic and geographic distances in the Pacific Ocean (r=0.93; Mantel *p*-value < 0.001, Figure 2.39-b and Supp. Figure 2.44). This result alone would not be conclusive, since a similar pattern is also expected under an equilibrium isolation by distance, but it strengthens our previous findings. Second, the historical demography inferences performed in each sampled deme showed that the pattern of genetic variability was most likely the outcome of a non-equilibrium meta-population structured according to a stepping stone migration matrix (Table 2.19). In this context, both the colonization times of the meta-population estimated by the ABC (Supp. Figure 2.41) and the expansion times retrieved by the *stairwayplot* (Figure 2.37) harbour the signature of the RE process (Lesturgie, Planes, et al., 2022): the oldest times are expected to be close to the centre of origin of the RE, while the more recent ones are likely associated to the edge of the colonization wave(s). While the large variance in T_{col} estimated by ABC does not allow for an accurate interpretation of the temporal dynamics of colonisation through the Indo-Pacific, the expansion times highlighted by the *stairwayplot* are consistent with

the RE scenario. Indeed, all sampling sites display a simultaneous expansion time around ~400 ky B.P. (Figure 2.37) except for Palmyra, Fakarava and Juan de Nova, which are the sites respectively further east (Palmyra and Fakarava) and west (Juan de Nova) to the inferred origin of the RE. In summary, all the evidence presented thus far point to an origin of *C. amblyrhynchos* east of IAA (particularly, east of New Caledonia), from which two migration waves took place, one to the East Pacific and the other to the Indian Ocean, with the Mozambique Channel being probably one of the last areas to have been colonized.

Our hypothesis is in line with the recent results of (Walsh et al. 2022), but they detected the origin of the RE within rather than eastward the IAA, using a similar genetic diversity decay approach. This discrepancy may be mostly due to the sensibility of this algorithm to the spatial distribution of the sampled populations (Peter & Slatkin, 2013), which differs considerably between the two studies. Another source of discrepancy may lie in the different bioinformatics pipelines. (Walsh et al. 2022) assembled loci with *PyRAD* (Eaton, 2014), whose calling algorithm requires high coverage data to correctly identify genotypes (Rochette et al., 2019). Here, we used the genotype-free approaches implemented in *ANGSD* to avoid possible skew towards low frequency variants in Rad-seq experiment with low to medium coverage (Heller et al., 2021; Mona et al., 2023) To shed more light on this issue, we carefully compared our results (obtained with *ANGSD*) to those obtained by three assembly and calling pipelines (namely, *PyRAD* (Eaton, 2014), *Stacks* v.1.48 (Catchen et al., 2013) and *Stacks* v.2.5 (Rochette et al., 2019), see Supplementary Methods) using the Bampton sampling site as a test case. All three SFS displayed an excess of singletons in comparison to the one inferred by *ANGSD* (Figure S6b), clearly determining not only a stronger ancestral expansion but also the absence of the recent bottleneck when fed to the *stairwayplot* algorithm (Supp. Figure 2.45-a). These results are consistent with (Heller et al., 2021), as we found an excess of low frequency variants when using the *Stacks* pipeline compared to the genotype likelihood approach implemented in *ANGSD*. Consequently, we highlight that the SFS reported by (Walsh et al., 2022) could be slightly biased toward an excess of low frequency variants.

The RE scenario, characterized by a centre of origin and two independent colonization waves, is similar to the one inferred for *C. melanopterus* by (Maisano Delser et al., 2019), a species whose range distribution overlaps with that of the grey reef shark. However, the most likely origin of the RE was located within the IAA for *C. melanopterus*, a well-known centre of origin for many teleost fishes (Cowman & Bellwood, 2013), and a biodiversity hotspot (Allen, 2008). The difference

observed between *C. amblyrhynchos* and *C. melanopterus* could result from the more balanced sampling scheme of (Maisano Delser et al., 2019), who could cover more homogeneously the Indo-Pacific. More samples from the IAA will be needed to refine our estimates. More generally, it will be interesting in the next future to explicitly investigate the role of the IAA for coral reef biodiversity fauna and to reconstruct the colonisations routes in the Indo-Pacific, using population genetics modelling applied to genomics data on multiple marine species to extract more general patterns (see for example (Delrieu-Trottin et al., 2020)).

2.5.5.2. Historical demography

The ABC framework not only provided another evidence in favour of a non-equilibrium meta-population scenario through the model selection analysis, but also allowed us to further refine our understanding of the evolutionary history of the grey reef shark. By analysing each deme separately, we found an overlapping posterior distribution of *Nm* with an average mode of ~10 (Table 2.19 and Supp. Figure 2.41). *C. amblyrhynchos*, similarly to *C. melanopterus*, is strongly dependent on reefs, whose distribution is not homogenous in the Indo-Pacific (Supp. Figure 2.46). We would have expected the connectivity in each sampled deme to be highly correlated to the distribution of coral reef in its neighbourhood, as it was previously observed in *C. melanopterus* (Maisano Delser et al., 2019). However, the two species differ in their dispersal behaviours: while grey reef sharks perform long-distance movements of at least ~900 km (Barnett et al., 2012; Bonnin et al., 2019; T. D. White et al., 2017), the blacktip reef shark exhibits a range of movement not exceeding ~50 km (Mourier & Planes, 2013). Our results reinforce the idea that the neighbourhood size in the two species is very different, with *C. amblyrhynchos* being able to cross expanses of open ocean and therefore being less sensitive to coral reef concentration than *C. melanopterus*.

The homogeneity in the signature of genetic variation in each deme was confirmed by the *stairwayplot* analyses (Figure 2.37), contrasting with the heterogeneity previously described for *C. melanopterus* (Maisano Delser et al., 2019). All demes showed an ancestral expansion followed by a period of stasis and a strong bottleneck in recent times. We recently showed that these three time periods are the typical signature of the variation in the coalescence rate through time due to the meta-population structure, with the slight differences observed between sites being only due to their specific colonization time (Lesturgie, Planes, et al., 2022). This result confirms the similarity of dispersal pattern throughout the Indo-Pacific. Similarly, the signature of bottleneck

observed in recent times for all demes (Figure 2.37) is also the expected consequence of population structure (Chikhi et al., 2018; Lesturgie, Planes, et al., 2022; Mazet et al., 2015; Rodríguez et al., 2018). This is true even when explicitly modelling spatial expansion with low *Nm* and colonization time of the same order as the one estimated in the grey reef sharks (as shown by the TD distribution, (Mona, 2017)). Unfortunately, population structure and demographic decline affect the SFS in a similar fashion making impossible to quantitatively disentangle the contribution of both to the observed bottleneck estimated using RAD-seq data (Lesturgie, Planes, et al., 2022). We stress that investigating local recent changes in connectivity or demographic events will clearly requires whole genome sequencing coupled with inferential methods based on the IICR (Arredondo et al., 2021) and/or linkage disequilibrium (Boitard et al., 2016). More generally, the next challenge will be to perform a full modelling of species structured in many demes as the grey reef shark. Here we took a simplified approach by considering each sampling site separately and by modelling the unsampled demes to estimate local migration rates. We are aware that in the future more data will be needed to explore complex demographic scenarios integrating RE that include both all sampled demes and the unsampled ones.

2.5.5.3. Population structure

The results presented so far suggest that dispersal abilities of *C. amblyrhynchos* are similar throughout the Indo-Pacific and independent of the availability of coral reefs. However, this cannot exclude the presence of barriers to gene flow which may have shaped the connectivity between demes. For widely distributed marine species, detecting such barriers may help to delineate management units and to take proper conservation measures in relation to fisheries (Dudgeon et al., 2012). Several evidence point to an absence of barriers to gene flow in the grey reef shark. First of all, we found a strong IBD pattern with a significant correlation between genetic and geographic distances of > 0.9 when considering only PAC samples (Figure 2.39-b) and a linear relation of smaller intensity between IND and PAC samples (Supp. Figure 2.44). Remarkably, these values are not affected by computing geographic distances between sampling sites under an LC approach. Indeed, the permeability values maximizing the correlation are (almost) the same for the different type of habitats. This suggest that different geographic features do not affect the direction of grey reef shark migrations, indicating, albeit indirectly, the absence of barriers to dispersal, consistently with the occasional long-distance swims detected across the open ocean (Barnett et al., 2012; Bonnin et al., 2019; T. D. White et al., 2017). When strong IBD is present, it

is difficult to attribute a biological meaning to groups identified by clustering algorithms (Meirmans, 2012). Both the *sNMF* and PCA analyses suggested a strong separation between IND and PAC samples (Figure 2.38), with the latter subdivided into two weakly divergent clusters (Figure 2.38 and Supp. Figure 2.47). The IND ancestral components diminished remarkably continuously eastward, once again supporting an IBD structure (Figure 2.38-a) rather than the presence of barriers to gene flow. This is consistent with the pairwise F_{ST} matrix, where intra Pacific comparisons did not exceed ~0.03 while the inter-oceanic comparisons have an average F_{ST} of ~0.54 (Figure 2.39-a). Defining management units within the PAC seems therefore inappropriate in the case of the grey reef shark, as genetic variations are rather continuous. This contrasts with what was previously suggested by (Boissin et al. 2019) at the Pacific scale: however, their results were based on a small number of microsatellites and they did not model IBD between the sampling points.

The pitfall of our study is to extrapolate the dynamic of the grey reef shark at the scale of its whole range by focusing mostly on the Pacific Ocean. Indeed, even if the species seems to follow an IBD pattern also from Chagos to Eastern Australia (Boussarie et al., 2022; Momigliano et al., 2017), the level of population differentiation appears to be higher than what we found in the Pacific for similar geographic distances. However, while the distribution of coral reef in the Pacific Ocean is scattered due to the presence of many archipelagos, coral reefs in the Indian Ocean are more concentrated on the coastal edge of the Asian and African continents (Supp. Figure 2.46). The effective distance between sampling sites within the Indian Ocean would therefore be larger than in the Pacific Ocean, where coral reefs would act as stepping stones to facilitate the colonization process and further migrations. This could also account for the different linear relationship estimated in the Pacific *vs.* the one estimated between Pacific and Indian sampling sites (Supp. Figure 2.44).

2.5.5.4. Conclusion

We explored the evolutionary history of the grey reef shark throughout most of its range in the Indo-Pacific and contrasted the results with those previously obtained for the blacktip reef shark (Maisano Delser et al., 2019). The two species are among the most abundant reef sharks (MacNeil et al., 2020), share an almost overlapping distribution in the Indo-Pacific and are both strictly coral reef-dependent species. Despite similarities in the RE dynamic, patterns of genetic diversity and population structure are very different between the two species. First, *C. melanopterus* is

significantly more structured than *C. amblyrhynchos* at similar spatial distances (for comparison, F_{ST} values are ~30 times higher when comparing French Polynesia vs New Caledonia, see Table S5 of (Maisano Delser et al., 2019) and our Supp. Table 2.22). Second, *C. amblyrhynchos* shows homogeneous migration rates and demographic signals throughout its whole distribution whereas *C. melanopterus* is more sensitive to the spatial distribution of coral reef with a connectivity largely dependent on the short scale reef-availability (Maisano Delser et al., 2019). Indeed, migration rates estimated in areas with extensive coral reefs coverage (e.g., the Great Barrier Reef) are much higher compared to those estimated in isolated islands/atolls in the Indo-Pacific (Maisano Delser et al., 2019), something that we did not observe for *C. amblyrhynchos*. All these differences can be explained by the life history traits related to dispersal abilities of the two species, with *C. amblyrhynchos* moving more freely in open sea expanses compared to *C. melanopterus*, lowering the impact of coral density on the observed genetic diversity. However, it will be important in the next future to precisely characterize the extent of the neighbourhood size for both species. To this end, ecological and genomic data need to be coupled: this will help to carefully decipher how many management units are necessary for species conservation and at which scale they should be established.

2.5.5.5. Comparison to Walsh et al. 2022 results

A reviewer raised some concerns about our claims related to the discrepancies between (Walsh et al. 2022) results and ours. The reviewer first strongly stated that (Walsh et al. 2022) results are not biased because of the coverage. Mean and median values reported for each individual (obtained setting the minimum read depth assembly parameter to 6) are between 15x and 20x: we argue that this value may not be high enough to obtain unbiased results given the variant calling algorithm they use (the one implemented in Pyrad, which is the same as *Stacks* v.1: see (Rochette et al. 2019) for a discussion on this topic). More generally, it has been shown that genotype-free pipelines (such *ANGSD*, which we applied here) perform better than the direct calling approaches in Rad experiments (Warmuth & Ellegren, 2019) and that the direct calling could skew the SFS towards an increase of singletons (Heller et al., 2021). Here, we do not claim that (Walsh et al. 2022) results are all biased – we simply stress that i) theirs SFSs show an increase of singletons when compared to our data (this is particular striking when comparing the Bampton sites, present in both studies); ii) when applying their pipeline to our data (which are low coverage) we found an excess of low frequency variants compared to the results obtained by *ANGSD* (Supp. Figure 2.45-b). These

considerations suggest that Walsh et al. 2022 data could suffer from a slight skew to an excess of low frequency variants, which, in turn, would explain the detection of an ancestral expansion signal and the lack of a recent decrease of effective population size in their *stairwayplot* results (which we observed in our data, compare their Figure 3 with our Figure 2.37).

The Reviewer raised a second point concerning our results: if a RE occurred (as both studies suggest more or less explicitly) then we should not observe a recent bottleneck in the sampled demes. This, according to the Reviewer, would suggest that our results are biased (while Walsh et al. 2022 are correct). This claim is unjustified for two main reasons: i) a recent bottleneck at local or global scale and/or a decrease in connectivity would inflate SNPs with average frequency variants affecting the reconstructed *Ne* trajectory particularly in recent times in any meta-population model (i.e., also in RE); ii) in line with this, and more generally, the behaviour of a sample of lineages from a deme depends specifically from the parameters of the RE: in other words, any possible SFS (and so the coalescence rate or *Ne* trajectory through time estimated out of it) can be obtained by varying these parameters. Similarly, an unstructured model can mimic the SFS produced under any meta-population model simply varying the function of *Ne* variation through time (Chikhi et al., 2018; Mazet et al., 2016). Observing a deficit of low frequency variants in a deme is therefore not at all inconsistent with a species experiencing a RE (see (Mona et al., 2014; Mona, 2017; Ray et al., 2003; Wegmann et al., 2006), among others). Moreover, the estimated time of the ancestral expansion in the grey reef shark is of the order of tens of thousands of generations and the exchanged migrants $Nm \sim 10$ per generation. Spatial explicit RE simulations already proved that under these parameters' combination *TD* can be positive (Mona, 2017) and instantaneous colonization models (lacking the spatial components) SST show signature of recent declines (Lesturgie, Planes, et al., 2022) in agreement with theoretical predictions (Chikhi et al., 2010, 2018; Mazet et al., 2016; Rodríguez et al., 2018).

2.5.6. Supplementary information

2.5.6.1. Author contributions

Pierre Lesturgie: Conceptualization (equal); data curation (lead); formal analysis (lead); investigation (lead); methodology (lead); soft- ware (lead); validation (equal); visualization (lead); writing – original draft (equal); writing – review and editing (equal). **Camrin D. Braun**: Resources (equal); writing – review and editing (supporting). **Eric Clua**: Resources (equal); writing – review and editing (supporting). **Johann Mourier**: Resources (equal); writing – review and editing (supporting). **Simon R. Thorrold**: Resources (equal); writing – re- view and editing (supporting). **Thomas Vignaud**: Resources (equal); writing – review and editing (supporting). **Serge Planes:** Resources (equal); writing – review and editing (supporting). **Stefano Mona**: Conceptualization (equal); data curation (supporting); formal analysis (supporting); funding acquisition (lead); investigation (supporting); methodology (supporting); project administration (lead); resources (equal); software (supporting); supervision (lead); validation (equal); visualization (supporting); writing – original draft (equal); writing – review and editing (equal).

2.5.6.2. Acknowledgements

We are grateful to the Genotoul bioinformatics platform Toulouse Midi-Pyrenees (Bioinfo Genotoul; http://bioinfo.genotoul.fr/) for providing computing resources. We are thankful to Valeriano Parravicini for his input and for providing resources on coral reef distribution in the Indo-Pacific and Romuald Laso-Jadart for criti- cal reading. We thank Jenn Caselle and Darcy Bradley for providing samples from the Phoenix archipelago and Jeremy Kiszka for provid- ing samples from Juan de Nova and Zélée bank. This work was sup- ported by two ATM grants (2016 and 2017) granted by the Muséum National d'Histoire Naturelle to Stefano Mona.

2.5.6.3. Data Availability Statement

VCF files, SFS, and scripts are available from the Dryad Digital Repository: doi:10.5061/dryad.547d7wm9b. Fastq sequence files are available from the GenBank at the National Center for Biotechnology Information short-read archive database (BioProject ID: PRJNA917473).

2.5.6.4. Supplementary Methods

Comparison of site frequency spectrum using different assembly and variant calling pipelines

To empirically investigate the influence of low coverage on variant calling, we investigated the site frequency spectrum (SFS) reconstructed in the Bampton sampling site (N=10) using four assembly and variant calling pipelines:

(1) **S1:** This pipeline is based on *Stacks* v.1.48 (Catchen et al., 2013). We implemented the same assembly parameters as those applied with *Stacks* v.2.5 and detailed in the main text (namely, m=3, n=3, and N=3). *Stacks* v.1.48 uses the calling algorithm of (Lynch, 2009) which requires high coverage data for accurate genotype inference (Rochette *et al.*, 2019). Using a custom R script, we filtered: *(i)* SNPs heterozygotes in more than 80% of the sample; *(ii)* loci with coverage higher than the mean coverage plus twice the standard deviation; *(iii)* SNPs in the last 5bp of the assembled locus; and *(iv)* loci containing more than five SNPs, after visual inspection of the distribution of segregating sites per locus.

(2) **PY:** This pipeline uses the assembly algorithm implemented in *PyRAD* (Eaton, 2014). We applied the same parameters of (Walsh et al., 2022), in order to compare their results to ours. The clustering threshold (level of similarity between sequences to be considered homologous) was set to 0.9, reads with more than 5 low quality bases were discarded, the minimum read depth for base calling was set to 6 and the maximum to 1000. The calling algorithm of *PyRAD* is the same as *Stacks* v.1.48. Loci with more than 5 SNP and sites with higher heterozygosity than 0.5 were also discarded using *PyRAD* pipeline. Using a custom R script, we additionally filtered depth by retaining only sites in the 90% core of the distribution of depth following (Walsh et al., 2022).

(3) **S2:** This pipeline is based on *Stacks* v.2.5 (Rochette et al., 2019) and mainly differs from S1 and PY in the calling algorithm. *Stacks* v.2.5 implements the population-based bayesian framework of (Maruki and Lynch, 2017) for variant calling, which is supposed to be more accurate for low coverage data (Rochette et al., 2019). The assembly step and filters were performed similarly to S1 above.

(4) **AN:** This pipeline is the one we used in the main text for all sampling sites. It is based on a first assembly of a pseudo-reference sequence (as in (Heller et al., 2021; Khimoun et al., 2020)) against which raw reads are mapped back, before using *ANGSD* (Korneliussen et al., 2014) for the genotype free allele frequency estimation. This pipeline has been

previously described and successfully applied to low-coverage RAD-seq data (Heller et al., 2021; Khimoun et al., 2020; Lesturgie, Planes, et al., 2022) and it is detailed in the main text.

We retained only loci with no missing data (monomorphic loci were used to properly scale the genetic diversity). The folded SFS was then computed by using a custom R script except for the folded SFS produced through the AN pipeline which was directly inferred using the *RealSFS* program implemented within the *ANGSD* framework. We computed the normalized SFS as in (Lapierre et al., 2017) to compare the distribution of alleles frequency between the four pipelines. The expectation of the normalized SFS is a horizontal line in a panmictic population of constant *Ne* (the standard coalescent model). The normalized SFS allows an immediate and qualitative description of the excess or deficit of low frequency variants compared to the standard coalescent model. We then inferred the variation of effective size (*Ne*) through time by modelling the SFS with the *stairwayplot* software (Liu & Fu, 2020). To be correctly scaled, the *stairwayplot* needs the total number of sites without missing data (monomorphic sites included) which were either directly extracted from the variant calling output (AN, S1 and S2) or estimated from the missing data rate detected in variant sites and the total number of sites assembled (PY). To compare the inferred *stairwayplot* with (Walsh et al., 2022), we used their same generation time of 16.4 years and mutation rate μ=1.9434e-08 per site per generation. This mutation rate was taken from (Maisano Delser et al., 2016), who estimated it based on the exon capture data of the black tip reef shark *C. melanopterus*. This value was later adjusted to represent a true genomic average, since exon capture represent a genomic sample enriched in conserved regions (Lesturgie, Lainé, et al., 2022; Lesturgie, Planes, et al., 2022). Therefore, in the main text we used the corrected mutation rate of 1.93e-8 per site per generation and a generation time of 10 years as in (Lesturgie, Planes, et al., 2022).

2.5.6.5. Supplementary Tables

Table 2.20. Confusion matrix of the model selection procedure: rows indicate the simulated models and columns the votes (in %) attributed by the ABC-RF algorithm to each of them.

		Attributed votes (%)			Class. error
		FIM	NS	SST	
Juan	FIM	41396	2027	6577	0.17
	NS	663	48706	631	0.03
	SST	5813	1059	43128	0.14
Bampton	FIM	40105	2251	7644	0.2
	NS	719	48590	691	0.03
	SST	6949	1432	41619	0.17
Belep	FIM	38085	2537	9378	0.24
	NS	822	48275	903	0.03
	SST	8790	1745	39465	0.21
Enderbury	FIM	41148	2141	6711	0.18
	NS	662	48637	701	0.03
	SST	5939	1091	42970	0.14
Kanton	FIM	40262	2269	7469	0.19
	NS	758	48507	735	0.03
	SST	6821	1366	41813	0.16
McKean	FIM	37984	2613	9403	0.24
	NS	837	48299	864	0.03
	SST	8825	1752	39423	0.21
Niku	FIM	42260	1851	5889	0.15
	NS	566	48925	509	0.02
	SST	4988	795	44217	0.12
Orona	FIM	40395	2203	7402	0.19
	NS	705	48595	700	0.03
	SST	6538	1284	42178	0.16
Palmyra	FIM	43083	1611	5306	0.14
	NS	483	49113	404	0.02
	SST	4344	561	45095	0.1
Fakarava	FIM	42156	1887	5957	0.16
	NS	607	48829	564	0.02
	SST	5140	870	43990	0.12

Table 2.21. Cross-validation of the ABC-RF procedure of the SST model: Mean Squared Error (SME), Mean Root Squared Error (RMSE) and 95% coverage of the median value for each parameter computed on 999 pseudo-observed datasets (pods).

		Nm	T_{col}	N_{anc}
Juan	Coverage	0.997	1	0.996
	SME	0.00448424	0.01473625	0.01679217
	MRSE	0.06559084	0.15917352	0.20927024
Bampton	Coverage	0.996	0.993	0.996
	SME	0.00418188	0.14437886	0.03688622
	MRSE	0.03547396	2.89427943	0.37985902
Belep	Coverage	0.997	0.994	0.998
	SME	0.00607304	0.06137466	0.0207203
	MRSE	0.08473424	0.78386441	0.17434075
Enderbury	Coverage	0.996	0.996	0.997
	SME	0.00179036	0.13584242	0.04827846
	MRSE	0.02448663	2.86874802	0.73693169
Kanton	Coverage	0.99	0.993	0.997
	SME	0.00153847	0.06132533	0.08600365
	MRSE	0.03178348	0.63525529	1.30039719
Mckean	Coverage	0.992	0.994	0.99
	SME	0.00563147	0.11468267	0.06404819
	MRSE	0.06642286	1.13569438	0.92112944
Niku	Coverage	0.999	0.999	0.999
	SME	0.00220155	0.03579402	0.01821552
	MRSE	0.03032898	0.47453067	0.25503461
Orona	Coverage	0.998	0.997	0.996
	SME	0.00489788	0.02427145	0.04237548
	MRSE	0.04727987	0.27677012	0.38470955
Palmyra	Coverage	0.999	0.998	0.998
	SME	0.00012809	0.01689414	0.02611148
	MRSE	0.01099299	0.27849299	0.45547058
Fakarava	Coverage	0.999	0.997	0.997
	SME	0.00304558	0.10564623	0.0475489
	MRSE	0.02769755	2.11211995	0.55497854

Table 2.22. Matrix of pairwise F_{ST} values (lower triangle) and associated p-value (upper triangle). Color represents the region of origin: Indian ocean (yellow), Chesterfield islands (red), New Caledonia (green), Phoenix islands (blue), Palmyra (cyan) and French Polynesia (pink).

	Juan	Zelee	Bampton	Avond	Belep	Poindimie	Niku	Mckean	Orona	Kanton	Enderbury	Palmyra	Moorea	Fakarava
Juan		NS	$p \leq 0.001$	$p \leq 0.001$	$p \leq 0.001$	$p \leq 0.001$	$p \leq 0.001$	$p \leq 0.001$	$p \leq 0.001$	$p \leq 0.001$	$p \leq 0.001$	$p \leq 0.001$	$p \leq 0.001$	$p \leq 0.001$
Zelee	0.0003		$p \leq 0.001$	$p \leq 0.001$	$p \leq 0.001$	$p \leq 0.001$	$p \leq 0.001$	$p \leq 0.001$	$p \leq 0.001$	$p \leq 0.001$	$p \leq 0.001$	$p \leq 0.001$	$p \leq 0.001$	$p \leq 0.001$
Bampton	0.5351	0.5373		NS	$p \leq 0.001$	$p \leq 0.001$	$p \leq 0.001$	$p \leq 0.001$	$p \leq 0.001$	$p \leq 0.001$	$p \leq 0.001$	$p \leq 0.001$	$p \leq 0.001$	$p \leq 0.001$
Avond	0.5298	0.5324	0.0001		$p \leq 0.001$	NS	$p \leq 0.001$	$p \leq 0.001$	$p \leq 0.001$	$p \leq 0.001$	$p \leq 0.001$	$p \leq 0.001$	$p \leq 0.001$	$p \leq 0.001$
Belep	0.5271	0.5296	0.0069	0.0046		NS	$p \leq 0.001$	$p \leq 0.001$	$p \leq 0.001$	$p \leq 0.001$	$p \leq 0.001$	$p \leq 0.001$	$p \leq 0.001$	$p \leq 0.001$
Poindimie	0.5302	0.5327	0.0078	0.0073	0.0013		$p \leq 0.001$	$p \leq 0.001$	$p \leq 0.001$	$p \leq 0.001$	$p \leq 0.001$	$p \leq 0.001$	$p \leq 0.001$	$p \leq 0.001$
Niku	0.5378	0.5402	0.0178	0.0163	0.0111	0.014		NS	NS	NS	NS	$p \leq 0.001$	$p \leq 0.001$	$p \leq 0.001$
Mckean	0.5379	0.5404	0.0173	0.0169	0.0119	0.013	0.0004		NS	NS	NS	NS	$p \leq 0.001$	$p \leq 0.001$
Orona	0.5434	0.5459	0.0172	0.0177	0.0117	0.0135	0.0009	0.0005		NS	NS	$p \leq 0.001$	$p \leq 0.001$	$p \leq 0.001$
Kanton	0.5493	0.5519	0.0162	0.016	0.0117	0.0126	0.0003	0.0012	0.001		NS	NS	$p \leq 0.001$	$p \leq 0.001$
Enderbury	0.5352	0.538	0.0167	0.0163	0.0118	0.0131	0.0008	0.0007	0.0019	0.0003		$p \leq 0.001$	NS	$p \leq 0.001$
Palmyra	0.5419	0.5444	0.0207	0.0194	0.0147	0.0159	0.0034	0.0023	0.0045	0.0016	0.0036		$p \leq 0.001$	$p \leq 0.001$
Moorea	0.5473	0.5499	0.0263	0.028	0.0218	0.0228	0.0109	0.0099	0.0102	0.0078	0.0125	0.0119		NS
Fakarava	0.5549	0.5575	0.024	0.0238	0.0215	0.0228	0.0097	0.0089	0.0084	0.0061	0.0101	0.01	0.0049	

2.5.6.6. Supplementary Figures

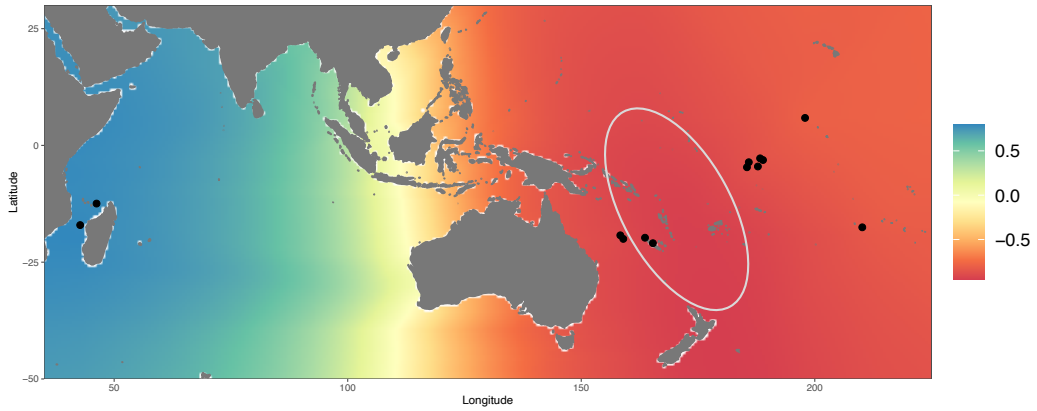

Figure 2.40. Correlation map between genetic diversity (θ_π) and Least Cost (LC) distances when considering all sampling sites. Each cell is coloured according to the correlation coefficient value computed between θ_π and the LC distance from the putative origin of the range expansion (RE). Black dots represent the sampling sites considered.

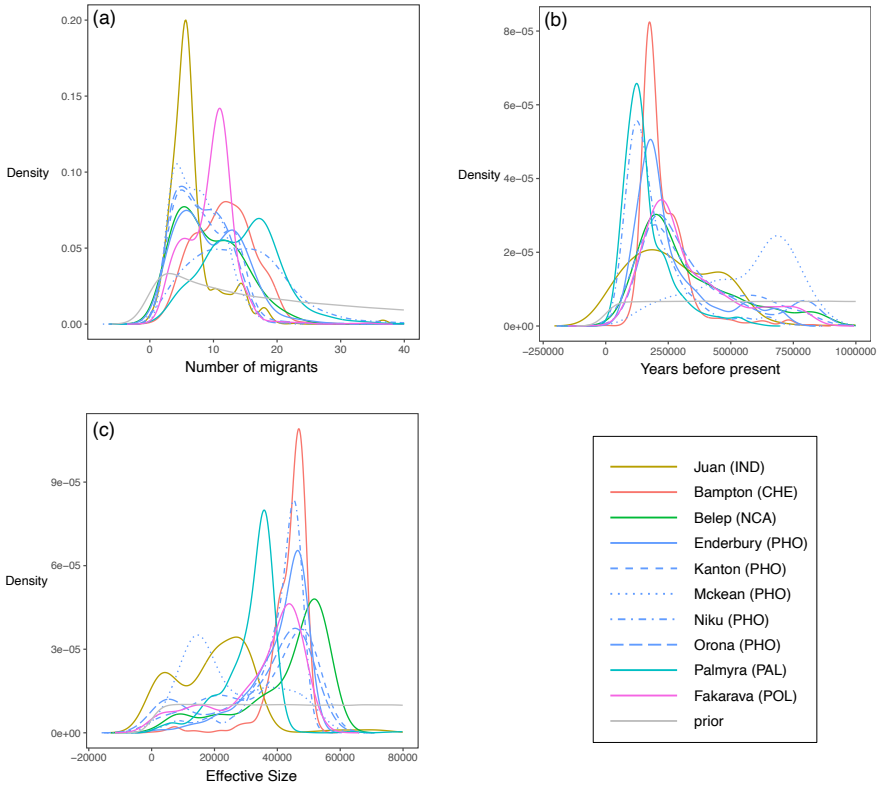

Figure 2.41. Posterior distribution of the number of migrants per generation Nm (a), the colonisation time of the array of deme T_{col} (b) and of the ancestral effective size N_{anc} (c) estimated under the stepping stone

model (SST) for all sampling sites with $N_{ind} \geq 7$. Colours represent the origin of the populations: Indian Ocean (yellow), Chesterfield islands (red), New Caledonia (green), Phoenix islands (blue), Palmyra (cyan) and Polynesia (purple). Line types represent the different populations from the Phoenix islands: Enderbury (solid), Kanton (dashes), Mckean (dots), Niku (dot-dashes) and Orona (long-dashes). The prior distribution is coloured in grey.

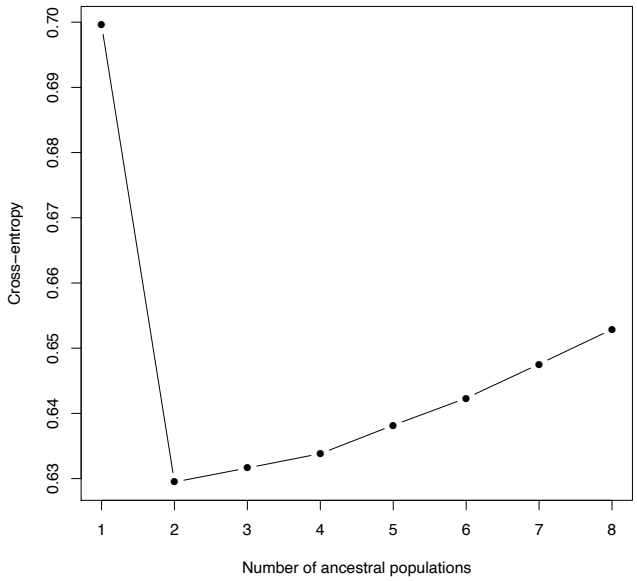

Figure 2.42. Cross entropy criterion of the sNMF algorithm computed for K=1 to K=8 ancestral populations.

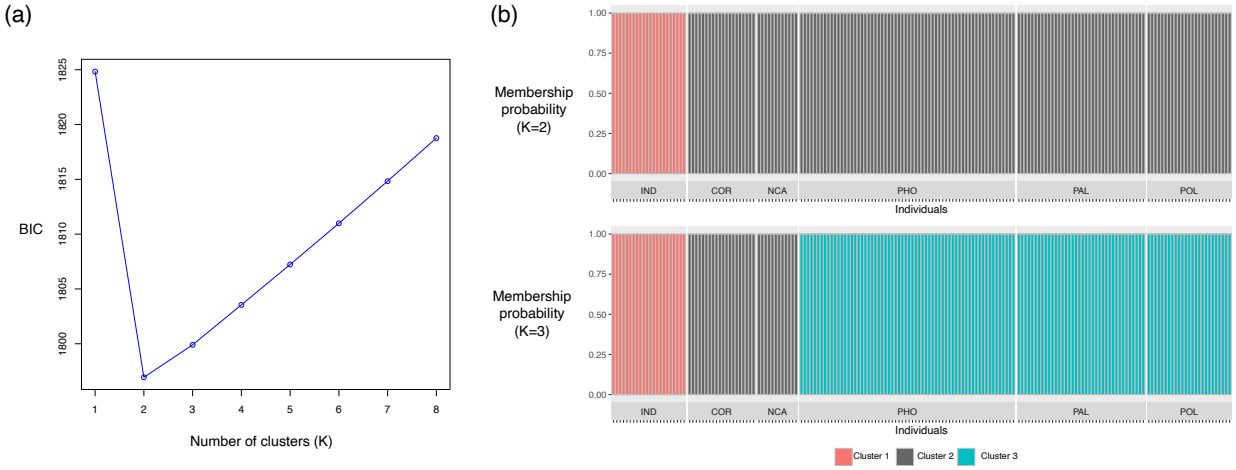

Figure 2.43. Results of the Discriminant Analysis of Principal Components. Bayesian Information Criterion (BIC) computed from K=1 to K=8 clusters (a) and membership probability of each individual to the clusters when considering K = 2 or K = 3 (b).

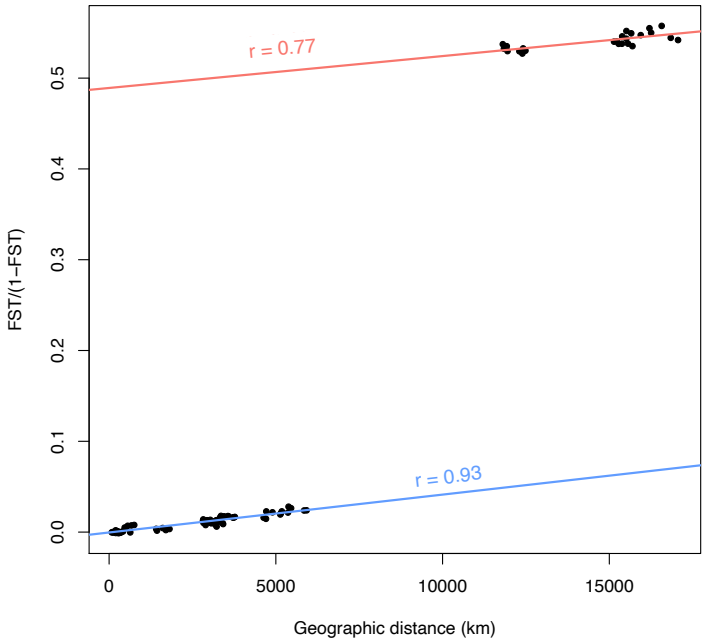

Figure 2.44. Isolation by distance (IBD) plot with all sampling sites. Correlation value and regression line computed between genetic and geographic distances when considering only Indian *vs.* Pacific sampling sites (red) or when considering only Pacific sampling sites (blue).

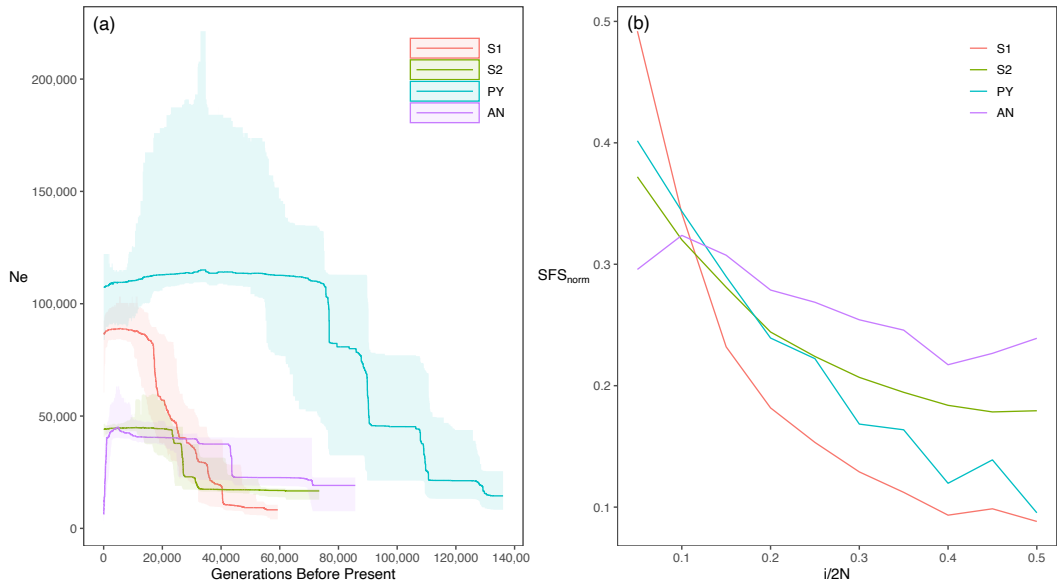

Figure 2.45. Variation of the effective population size (Ne) through time and its 75% confidence interval estimated by the stairwayplot (a) and Normalized Site Frequency Spectrum (b) of Bampton site (N=10) computed from data assembled using the different variant calling pipelines: ANGSD (AN, purple), STACKS v.2.5 (S2, green), STACKS v.1.48 (S1, red) and Pyrad (PY, blue). The stairwayplot was computed using the mutation rate μ=1.9434e-08 per site per generation and a generation time of 16.4 years as in (Walsh et al., 2022).

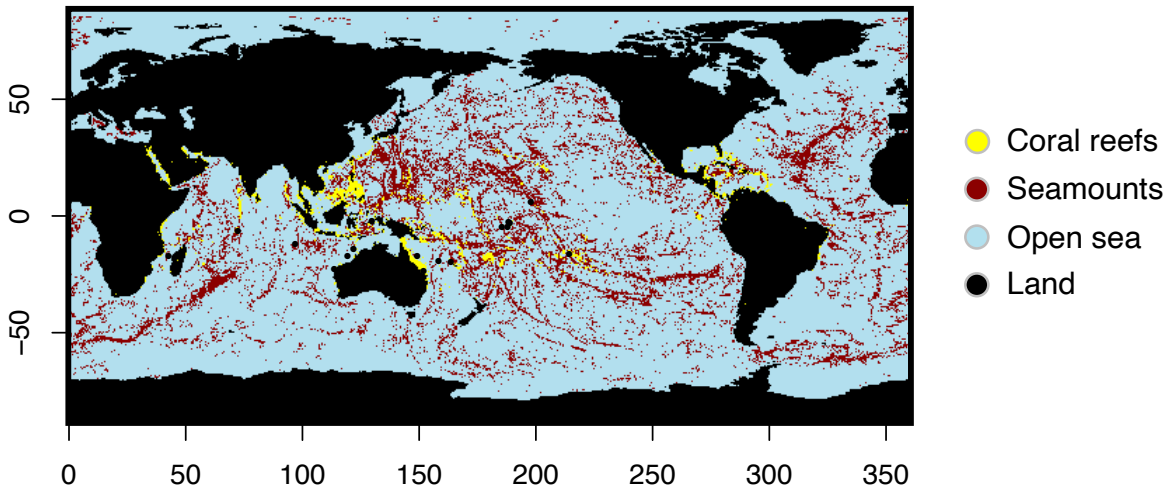

Figure 2.46. Distribution of corals and seamounts in the Indo-Pacific oceans. Cells are coloured according to their habitat type: coral reefs (yellow), seamounts (red), open sea (blue) and land (black).

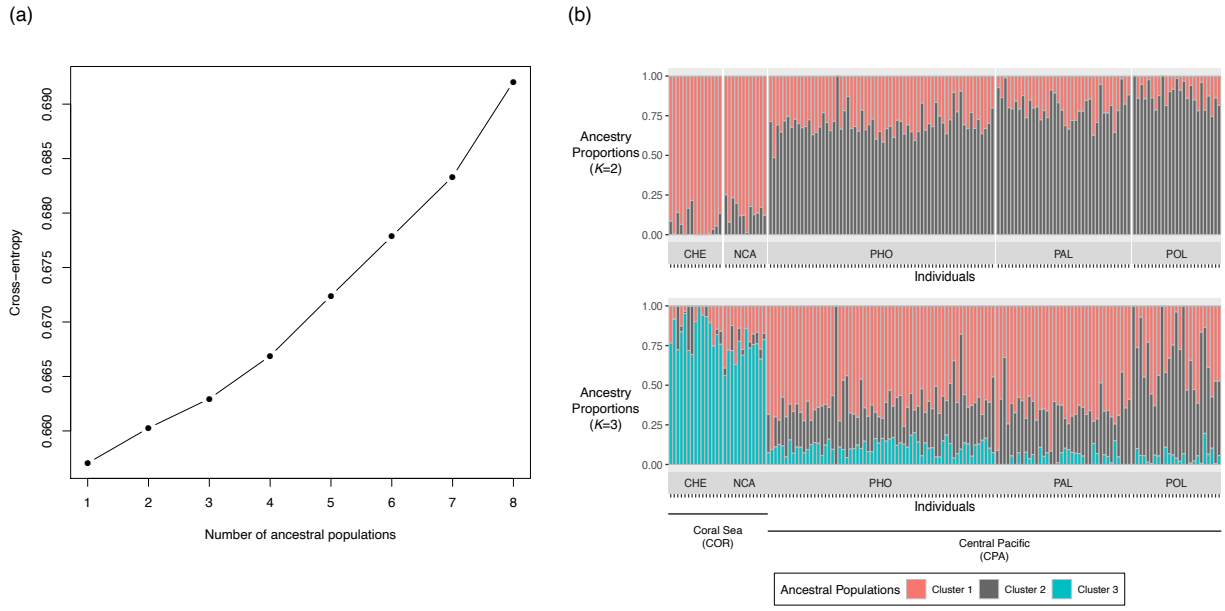

Figure 2.47. *sNMF* algorithm computed on Pacific samples only. Cross entropy criterion of the sNMF algorithm computed for K=1 to K=8 ancestral populations (a) and ancestry proportions retrieved with K=2 and K=3 ancestral populations (b).

2.6. Conclusion and perspectives

2.6.1. Coalescence Times and *Unstructured* Models in Meta-populations

In this chapter I investigated the relations between life history traits (LHT), population structure and *unstructured* models, and their implications in the reconstruction of the evolutionary history of species. In section **2.3**, I first explored through simulations of RADseq loci the reconstructed *Inverse Instantaneous Coalescence Rate* (IICR) from a sample of lineages from a deme belonging to a meta-population. To that end, I used the *stairwayplot*, an SFS-based *unstructured* method, but the IICR could have been similarly reconstructed using methods such as the PSMC (Li & Durbin, 2011) if the whole genome was simulated. For mathematical tractability reasons, such investigations have traditionally been restricted to equilibrium models, i.e., where the meta-population has always (i.e., for a very long time) been established in the range of the species (e.g., (Arredondo et al., 2021; Chikhi et al., 2018; Mazet et al., 2015, 2016; Rodríguez et al., 2018)). In result, non-equilibrium models investigations, where the array of demes gets colonized from an ancestral deme in history, remain scarce, despite it is expected to be more realistic as most widespread species have likely undergone range expansions (Excoffier et al., 2009). To this end, I investigated by simulating thousands of RAD-seq loci the effect of genetic structure on the gene genealogy in a non-equilibrium meta-population model when connectivity changes or not. Simulations were performed under two meta-populations scenarios: the *Finite Island Meta-population* (FIM) and the *Stepping Stone* (SST) models (Figure 2.48, panels B and C). Both scenarios depict an ancestral deme that instantly colonized an array of 100 demes T_{COL} generations ago (hence the non-equilibrium nature). Since then, each deme exchanged a number of migrants (*Nm*) either with any other deme of the array (FIM) or only with the closest neighbours in a 2D-grid (SST). Modified scenarios of SST and FIM included a possible instantaneous decrease in connectivity T_{CH} generations ago (called FIM-CH and SST-CH). Various values of Nm, T_{COL}, T_{CH} and of the strength of decrease in connectivity were investigated, and the *stairwayplot* was run to characterize their effect on the shape of the gene genealogy.

This general framework provided three general findings. First, it confirmed the artificial signature of bottleneck typical of lineages sampled from a deme belonging to a meta-population (Chikhi et al., 2018; Maisano Delser et al., 2016, 2019; Mazet et al., 2015), as predicted by theoretical arguments (Wakeley, 1998, 1999), no matter the scenario simulated. Second, it demonstrated that a decrease in connectivity (i.e., a *true* bottleneck in a deme), produces the same gene genealogy

than the population structure artificial bottleneck. In consequence, it does not change the trajectory reconstructed though *unstructured* models. Third, the detection of an ancestral expansion by the stairwayplot happening at a time roughly consistent with the simulated T_{COL}, but only in scenarios with high enough Nm and/or T_{COL}. This suggests the ability of *unstructured* models to recover the time at which the species colonized the habitat, an important event of the history of a meta-population, although its detection depends on the interplay between the level of connectivity and the colonization time of the deme.

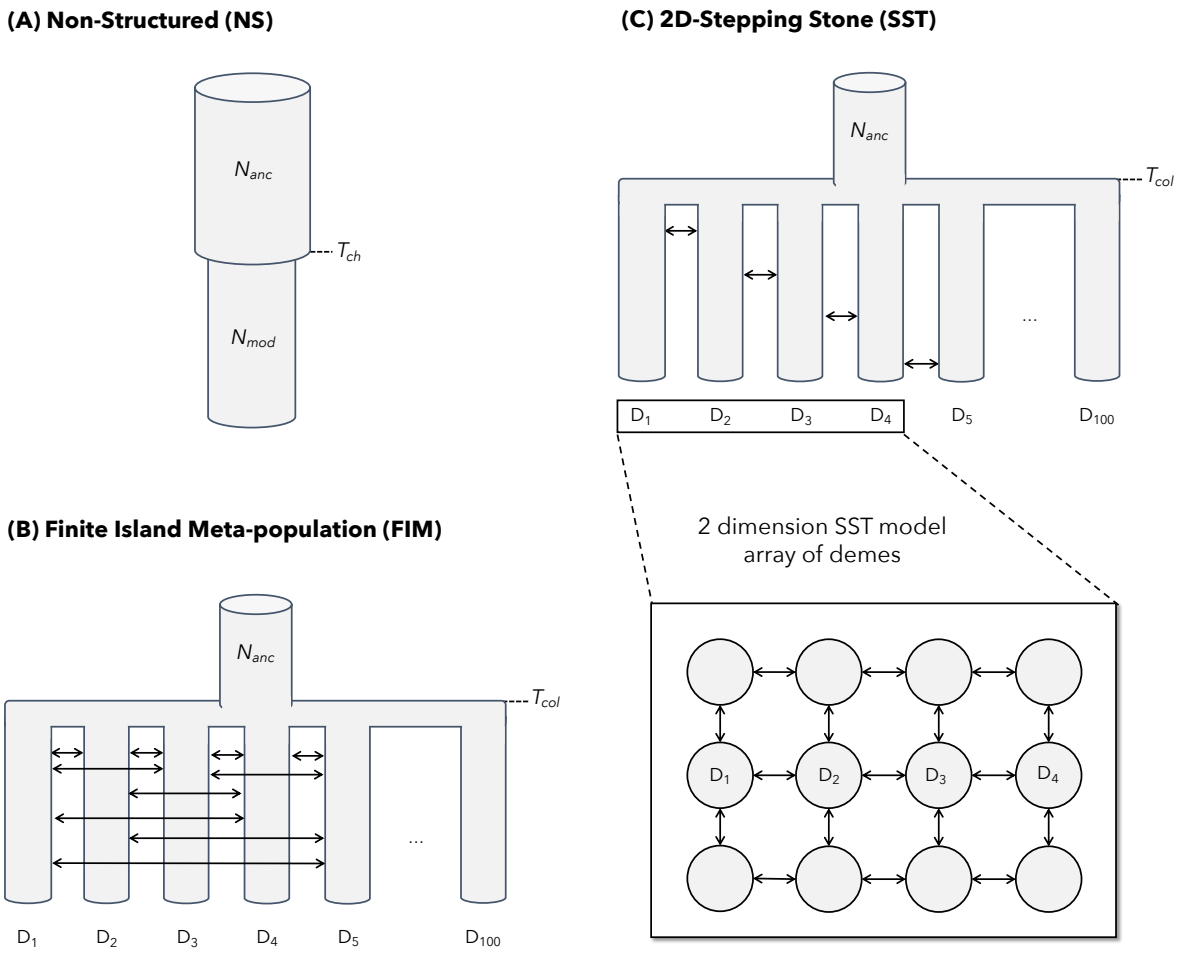

Figure 2.48. Simulated demographic scenarios in the different sections of this chapter. Panel A: NS (No Structure) is a non-structured model where the modern effective size (N_{mod}) instantaneously changes to N_{anc}, at time shift T_{ch} generations ago. Panel B represents a FIM (Finite Island Meta-population) model with 100 demes that have been instantaneously colonised T_{col} generations ago, from an ancestral population of size N_{anc}. Demes are allowed to exchange migrants with any other. Panel C represents a SST (Stepping-Stone) model. It is similar to FIM but the migrants are only exchanged between the four nearest neighbours in a two-dimensional grid (displayed below the scenario).

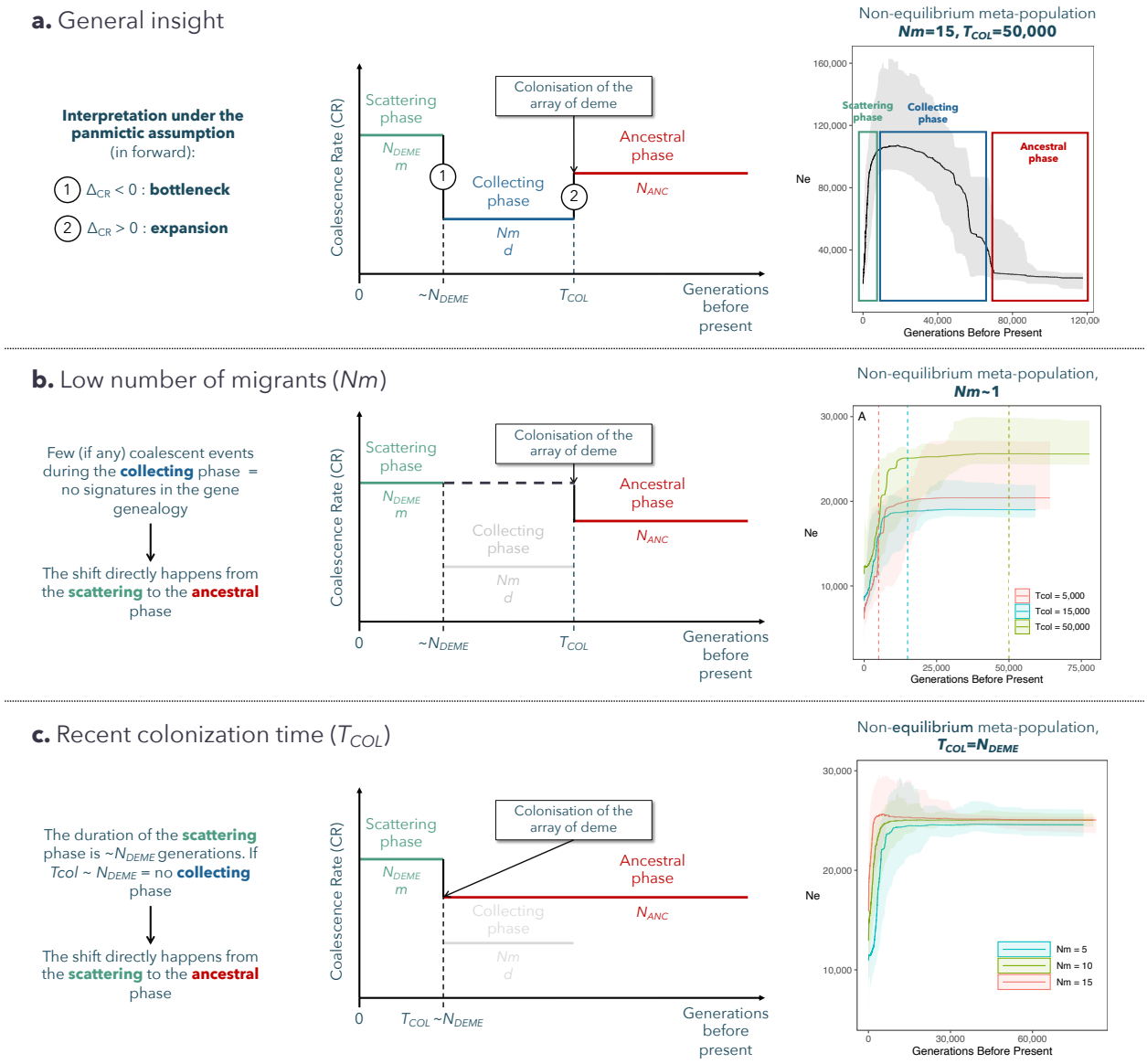

Figure 2.49. Coalescence phases in different meta-population scenarios with associated examples of reconstructed dynamics through *unstructured* models. **Row a.** General insight on the three different coalescence phases in the history of lineages sampled from a deme belonging to a stepping-stone non-equilibrium meta-population described in this chapter. On the left, description of the interpretation of the shifts in coalescence rate under the panmictic assumption. In the middle, schematic diagram representing the coalescence rate in each phase. Each phase and related parameters are represented by a color. Parameters influencing the coalescence rate in each phase are the effective size of the deme (N_{DEME}) and the migration rate (m) for the *scattering* phase (green); the number of migrants exchanged per generation (Nm) and the number of demes (d) for the *collecting* phase (blue); and the ancestral effective size (N_{ANC}) for the *ancestral* phase (red). In the right panel, the different phases are illustrated in practice by the reconstruction of the coalescence rate with the *stairwayplot*. **Row b.** Similar to row a. but in the specific case of a low Nm where the *collecting* phase cannot be reconstructed due to the absence of coalescence events in this time frame as schematized in the middle panel and illustrated by the stairwayplot in the right panel. **Row c.** Similar to A but in the specific scenario of a too recent T_{COL} (i.e., roughly similar to the value of N_{DEME}) where the *collecting* phase do not happen in the history of the sample as schematized

in the middle panel and illustrated by the stairwayplot in the right panel. Both scenarios displayed in rows b. and c. will lead to a similar signature on the gene genealogy, but with a different underlying process (i.e., no signature of *collecting* phase despite it is present in row b., and total absence of the *collecting* phase in the history in row c.).

These findings can be explained using coalescence arguments. In summary, Wakeley (1998, 1999) described the history of a meta-population in two phases (going backwards in time): the *scattering* phase, where lineages can freely merge or migrate to other demes, and the *collecting* phase happening at the end of the *scattering* phase, where each lineage has reached a separate deme. The different dynamic of these phases will impact the coalescence rates with the *scattering* phase having a much higher coalescence rate than in the *collecting* phase. During the transition from the *scattering* to the *collecting* phase, the coalescence rate will then decrease, resulting in an apparent increase of IICR (or Ne) in the *unstructured* model, or, going forward in time, a bottleneck such as detected in all our scenarios (e.g., Figure 2.49-a). So far, the two phases described by Wakeley do not explain why an ancestral expansion was detected: this is because the meta-population model described is at the equilibrium, i.e., demes have always been established in the range of a species or for a very long time. Here, we modelled the colonization of the array of deme: this means that at some point during the *collecting* phase, lineages will instantly merge into the ancestral deme. In this *ancestral* phase, coalescence events happen at a rate N_{ANC} which will be higher than in the *collecting* phase, resulting in the detection of a decrease in Ne during the transition between the two phases, or, going forward in time, an ancestral expansion (Figure 2.49-a). Finally, this general insight in three clear ancestral phases can be hampered by the demographic parameters of the meta-population. For instance, only two phases were detected in scenarios with low *Nm* or too recent T_{COL}. While the scattering phase was detected in all scenarios, the detection and/or presence of the ancestral or collecting phase depends on the parameters of the model. The *ancestral* phase can be undetected in two cases: (1) T_{COL} could be too old (i.e., tending to an equilibrium model), putting the *ancestral* phase older to the T_{MRCA} of the sample and therefore removing its signature from the gene genealogy, and (2) the *ancestral* phase is not detectable if the coalescence rate is roughly similar between the *collecting* and the *ancestral* phases, although it is realistic to assume it is different as the number of demes is likely very large in species organized in meta-populations. In addition, the *collecting* phase might not leave a signature in the gene genealogy depending on an interplay between parameters of the meta-population model: (1) when the number of migrants is too low the *collecting* phase is not detectable because too few coalescent events occurred during

the phase to leave a signature in the gene genealogy (low Nm, Figure 2.49-b), and (2), when T_{COL} is too recent, i.e., close to N_{DEME} generations, the *collective* phase is simply absent from the history of the sample as the *scattering* phase roughly lasts N_{DEME} generations (recent T_{COL} 2.49-c).
This paper considerably increased our understanding of how population structure influences the gene genealogy, and how *unstructured* models can be helpful in inferring the demographic history of a deme belonging to a meta-population. Specifically, it shows how some processes can impact coalescence rate and can thus be detected using *unstructured* methods. However, it also confirms that an un-educated use of these models will lead to a mis-interpretation of recent trends, which can pose strong issues when investigating widespread endangered species. This highlights the importance of investigating meta-population structure before performing any inferences using *unstructured* models, as they will allow to recover the colonization time of the habitat, an important feature of the evolutionary history of most widely distributed species (which is exemplified in section 1.6.3).

2.6.2. Life History Traits and Population Structure

Different parameters of a meta-population model will influence the reconstructed coalescent rate through time as demonstrated in section **2.3**. But how does it integrate with empirical inquiries? Are all widespread species genetically structured, and if so, how do Life History Traits (LHT) influence the degree of structure? To understand this, we investigated the correlation between LHT and demography in sharks by sampling four species in New-Caledonia that displayed a dispersal capacity gradient related to their LHT. The blacktip reef shark (*Carcharhinus melanopterus*) is the least mobile species as it is highly infeudated to lagoons. The blacktip shark (*C. limbatus*) and the grey reef shark (*C. amblyrhynchos*) are both fringing reefs associated species and display higher mobility than the blacktip reef shark, with movements reaching respectively ~150km and ~900km. Finally, the tiger shark (*Galeocerdo cuvier*) differs a lot from the three other species as it is significantly bigger, is capable of transoceanic movements and has a semi-oceanic habitat. By using an Approximate Bayesian Computation coupled to Random Forests (ABC-rf; Pudlo, 2018; Raynal et al., 2019) framework, I tested for meta-population structure in the four species. To that end I investigated the previously described SST and FIM models but also a non-structured model (NS) to account for the absence of structure (Figure 2.48), corresponding to the scenario of a panmictic population undergoing a change in its effective size.

Among the four species, only the tiger shark was best depicted by the NS model, the remaining three species being best explained by the SST scenario, where the estimated *Nm* increased consistently as the known movement range ability increased. The influence of LHT on population structure could then be summarize as follow: when movement abilities increase, the connectivity between demes increases until population structure fades into panmixia such as in the tiger shark. I then reconstructed the variation in coalescence rate through time using the *stairwayplot* in the three species organized in meta-populations. The expected recent bottleneck signal was detected, but also an ancestral expansion signal in *C. limbatus* and *C. amblyrhynchos* which was not present in *C. melanopterus*, associated with the lowest value of *Nm*. This corroborated the previously developed theoretical findings: *Nm* might be too small in *C. melanopterus* to detect the *collecting* coalescence phase. Conversely, the higher connectivity in both *C. amblyrhynchos* and *limbatus* determines a substantial amount of coalescence events in the *collecting* phase, allowing to the detection of the shift between the *collecting* and *ancestral* phases through an expansion in the IICR. Finally, this expansion signal is likely a signature of the colonization of the habitat, but extensive studies of these species at their range scale is further needed to confirm such intuition.
The investigation of the relationship between LHT and historical demography has allowed to characterize some determinants of genetic structure in sharks, which is important both for multi-species conservation planning and for understanding general drivers of population structure. In the future, more species should be included to precisely account for determinants of genetic diversity and structuration.

2.6.3. Descriptive Methods: A Baseline for Demographic Inferences

Testing for meta-population structure should be the first step in studying the evolutionary history of a species through its range distribution (as introduced in section **1.3**). Above, I emphasized two reasons why this is crucial: (1) not accounting for population structure can strongly bias our interpretations of the observed variation in coalescence rate inferred under *unstructured* models; and (2) species can display various degrees of genetic structure depending on their LHT as exemplified in sharks (section **2.3**). At the same time, this work highlights how useful *unstructured* models can be to detect signature of the colonization process in structured species, if interpreted correctly. In this chapter, I exemplify the necessity and benefits of the *descriptive* methods step (see **Chapter 1**) to assess the degree of population structure through the investigation of the

evolutionary history of two species of contrasted genetic structure: the tiger shark (*G. cuvier*), and the grey reef shark (*C. amblyrhynchos*). The two species were introduced earlier (in **1.6.2**): ABC-rf investigations in a single location suggested either panmixia (*G. cuvier*) or meta-population structure (*C. amblyrhynchos*). In both cases, I run a similar set of descriptive analyses composed of: clustering and F_{ST}-based analyses; model selection between NS, FIM and SST models in each sampling locations following the ABC-RF framework detailed above (Figure 2.48); and *stairwayplot* inference of the variation in coalescence rate in each sampling location.

The case of the tiger shark. 50 individuals were sampled from five Indo-Pacific (IP) sampling sites and one Atlantic Ocean (AO) sampling location and sequenced following a dd-RADseq protocol. By coupling clustering algorithms, F_{ST} and the ABC-RF framework, I highlighted high genetic differentiation between the AO and the IP but also provided robust evidences of panmixia at the oceanic scale (i.e., in AO and IP). When panmixia is confirmed, *unstructured* models have a direct biological meaning as signals inferred can be interpreted as Ne variation through time. Moreover, having two panmictic populations provide the opportunity to design relatively simple models (in comparison to meta-population models) to better understand the evolutionary history of the species. To illustrate that, I investigated patterns of migration and divergence between the two panmictic populations. Multiple Isolation-Migration (IM) nested models were investigated using *fastsimcoal* approximating likelihood framework, which models the two-dimensional SFS to estimate the most likely set of parameters under a user-defined model and is also able to compare the likelihood of different models. The most likely scenario displayed an ancestral divergence ~193,000 years ago and an ongoing but limited asymmetric migration ~176 times larger from the Indo-Pacific to the Atlantic Ocean than vice versa. Given the preference of the tiger shark for warm waters (Payne et al., 2018), I explained the limited migration by the occurrence of the cold Benguela current off south Africa, acting as a barrier to gene flow between the two basins. Nevertheless, the barrier is made permeable by the Agulhas leakage flowing warm water from the IP to the AO (Beal et al., 2011), which can account for the asymmetric migration consistently with observations in other sharks and bony fishes (Gaither et al., 2016; Maduna et al., 2017). The two populations thus likely remain connected, although they separated a *long* time ago and effective migrants are rare. I furthermore highlight contrasted demographic trends in the two basins using the stairwayplot, with a recent bottleneck in the IP and a recent expansion in the AO. Overall, given the absence of population structure at a large scale, we were able to precisely investigate the

divergence between populations and their respective demographic trends, which allowed to disentangle previous hypothesis about the tiger shark evolutionary history.

The case of the grey reef shark. I investigated the demographic history of the grey reef shark with the ultimate goal of comparing it to the sympatric and congeneric species *C. melanopterus* whose history was investigated in (Maisano Delser et al., 2019). To that end, 175 individuals were sampled within its range (the Indo-Pacific), and sequenced following a dd-RADseq protocol. Clustering algorithms and pairwise-F_{ST} first highlighted strong genetic differentiation between the Indian and the Pacific Ocean (F_{ST}~0.56). The ABC-rf framework further demonstrated that the species is structured in an SST meta-population with a homogeneous level of connectivity through its range. Consistently with meta-population structure, the species displayed strong isolation by distance (IBD) but low genetic differentiation at the oceanic scale, and further landscape genetics analyses suggested the absence of clear barrier to dispersal in the Pacific. This confirmed the ability of *C. amblyrhynchos* to perform long distance dispersal (LDD), consistently with long range movements monitored up to 926km (Barnett et al., 2012; Bonnin et al., 2019). In addition, the species displayed a range expansion signature, which was shown by coupling the investigation of the frequency of derived alleles and the decay of genetic diversity with distance to a putative origin (Peter & Slatkin, 2013; Ramachandran et al., 2005). This indicated a likely origin of RE east of the Indo-Australian Archipelago (IAA), also known as the Coral Triangle. Unlike the tiger shark, panmictic at a large scale, the *stairwayplot* does not convey (directly) variations of effective sizes in demes for the grey reef shark. To understand better the history of the species, one strategy would be to device more complex SST scenarios, for example introducing a change in connectivity through time. Detecting changes in connectivity might only be possible in the future larger amount of data coupled to additional statistics (notably LD-based). However, the *stairwayplot* is expected to provide additional details about the colonization dynamics as demonstrated in section **2.3**, notably about the tempo of the process. In fact, we detected the most recent ancestral expansions for the two populations located close to the ends of the range distribution, with the most recent expansion being detected the Indian Ocean. Coupled to RE results, it provides strong evidences of two colonization waves taking place from the IAA to the Pacific and likely more recently to the Indian Ocean. This RE dynamics is highly congruent with the one inferred in the sympatric *C. melanopterus* (Maisano Delser et al., 2019), reinforcing the hypothesis around the role of the IAA, a current biodiversity hotspot (Allen, 2008), as a centre of origin for many teleost fishes (Cowman

& Bellwood, 2013). The two species share the same coral habitat and showed similar amount of genetic structure between the edge of their range distribution and an organization in meta-population, however, they display several contrasting demographic features. Notably, the LDD suggested in *C. amblyrhynchos* contrasts with *C. melanopterus* which is strongly structured at a low scale (Maisano Delser et al., 2016, 2019; Mourier & Planes, 2013). However, this is not surprising given their respective LHT, with *C. amblyrhynchos* being less dependent on coral reef distribution, bigger, and capable of wider movements. This is suggestive of differences in the ecology of the species, and possibly in their genetic resilience. For instance, the heterogeneous estimated *Nm* across the range of *C. melanopterus* (Maisano Delser et al., 2019) likely emphasizes the strong dependence of *C. melanopterus* to the coral habitat whose cover is heterogeneous the Indo-Pacific. This contrasts with the homogeneous estimate of *Nm* through the range of *C. amblyrhynchos*, suggesting that it is capable of large migrations and less dependent on coral cover.

2.6.4. General conclusions

In this chapter, I first highlighted the utility of *unstructured* models to uncover important elements of the history of a meta-population by underlining the influence of meta-population organization and its history on the shape of the gene genealogy. I then provide evidences of the influence of LHT on the degree of population structure. Overall, I stress the importance to always test for population structure using descriptive approaches, and then to perform and interpret demographic inferences accordingly to the underlying genetic structure. This was illustrated by studying the evolutionary history of species with contrasting genetic structure. In *G. cuvier*, panmictic at the oceanic scale I was able to characterize migration and divergence between two oceanic basins, ultimately allowing to disentangle previous hypothesis about the evolutionary history of the species. Conversely, descriptive analyses in *C. amblyrhynchos* displayed meta-population structure within its range, and more specifically, I could detect range expansions (RE) signatures. This context allowed me to show empirically the utility of *unstructured* models to describe the shape of the gene genealogy, in agreement to what proposed in section **2.3.** Indeed, the combination of structured and *unstructured* models allowed to characterized its evolutionary history and to compared it to *C. melanopterus*. In the end, we highlighted the strong genetic differences between the two reef species, likely driven by their biology, which emphasizes once more the influence of LHT in shaping the evolutionary history of species.

Chapter 3. Supergenes, Demography and Conservation

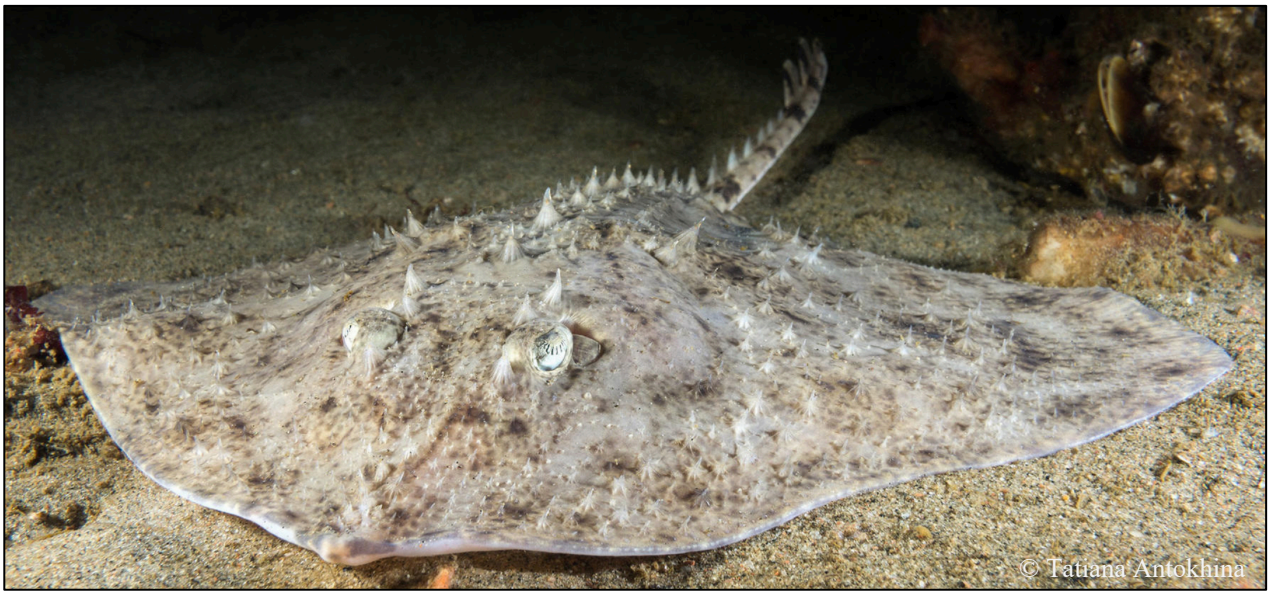

Figure 3.1. Thorny Skate (*Amblyraja radiata*).

3.1. Context

3.1.1. The thorny skate: an endangered species with a striking size polymorphism

The evolutionary history of species is shaped by demographic and selective processes. Today, studies have not always extensively set out to understand their interplay, which should be eased by demographic modelling. In this chapter, I investigate the case of the Thorny Skate, a Vulnerable species (Kulka et al., 2020; Sosebee et al., 2016) that displays a size polymorphism in part of its range (Mcphie & Campana, 2009; Sosebee, 2004; Sulikowski et al., 2005; E. G. Templeman, 1984; W. Templeman, 1987), which is believed to be associated to decreasing demographic trends (Kulka et al., 2020; Sosebee et al., 2016). Yet do date, this association is not clear, and the determinism behind the discrete size polymorphism remains unknown despite some population genetics attempts to characterize it (Denton et al., *in prep*; Lynghammar et al., 2016). One of the main objectives of this chapter was to understand the origin of the polymorphism. This was revealed to be determined by a supergene, a complex system that I introduce below. Thus, the main focus of this chapter is to show how the extensive reconstruction of demographic history is crucial to elaborate conservation hypotheses as well as to understand the origin of such peculiar system.

3.1.2. A word on Supergenes

Supergenes are genomic regions that encompass genes all inherited together as a single gene in a Mendelian fashion (Schwander et al., 2014; Thompson & Jiggins, 2014; Wellenreuther & Bernatchez, 2018), hence the term “supergene”. Supergenes are generally considered to arise from chromosomal inversions (Gutiérrez-Valencia et al., 2021; Schwander et al., 2014), which are chromosome fragments that, during DNA replication in meiosis, separate from the replicating DNA segment and reattach to the chromosome in the opposite direction. The recombination process is then suppressed between chromosomes whose set of genes is in the opposite direction because a crossing over would lead to gametes with non-even genetic material (i.e., non-even chromosome sizes), except in the case of occasional double crossing-over. However, recombination remains active between chromosomes keeping the same genetic order. As direct consequence of their nature, chromosomal inversions can lead to structural issues for

heterokaryotype individuals (i.e., individuals having one inverted and one non-inverted chromatid) during meiosis (especially when the inversion encompass the centromere), or the disruption of a coding sequence with consequences for gene expression. In this context, supergenes are series of co-adapted alleles remain stacked together in one haplotype by (large) inversions. The co-adaptation of such alleles is deemed to be the major force leading to the maintenance of polymorphic inversions in populations (Wellenreuther & Bernatchez, 2018).

The nature of the many different genes inherited in a single block associates supergenes with complex phenotypes (and sometimes pleiotropy). One of the most famous examples of supergene are sex chromosomes (Branco et al., 2018; B. Charlesworth, 1996; B. Charlesworth & Charlesworth, 1978; D. Charlesworth, 2016), as they enable the efficient inheritance of genes involved in the determinism of sex. However, supergenes are also involved in the determinism of very unique, non-sex related, traits in different organisms (Schwander et al., 2014; Thompson & Jiggins, 2014). In animals, the most documented examples are supergenes determining mimicry and wing morph in butterflies (Clarke & Sheppard, 1960; Ford, 1966; Joron et al., 2006, 2011), and social behavior polymorphism in ants (Avril et al., 2019; Brelsford et al., 2020; Chapuisat, 2023; Kay et al., 2022; Stolle et al., 2022). While supergene regions have been well characterized and strongly associated with one (or more) phenotypes in few model system, this is not the case in less documented species where associations with phenotypes remain less robust to date. For instance, several supergenes have been found in Cod species (Barney et al., 2017; Matschiner et al., 2022), Atlantic salmon (Stenløkk et al., 2022), or Rainbow trout (Pearse et al., 2019), but their role remains debated. Many supergene-related determinisms have yet to be characterized, which will eventually be increasingly possible thanks to the possibilities offered by whole-genome sequencing data and their growing affordability.

The peculiar nature of supergenes and associated complex phenotypes poses many questions as to how they can spread in space and be maintained in (Schwander et al., 2014; Thompson & Jiggins, 2014). How is it possible that supergenes, sometimes very large and millions of generations old (Wellenreuther & Bernatchez, 2018), persist through time in the face of the lack of recombination? The recombination process is complex in that respect, as it can prevent the fixation of maladapted alleles and favors the spread of advantageous alleles (Felsenstein, 1974) but its absence can also favor the maintenance of co-adapted alleles which has been shown to be extremely useful in the case of sex chromosomes (Branco et al., 2018). This complex interplay between the advantage or

disadvantage of alleles trapped in supergenes, and the benefit of co-adapted alleles result in complex selective processes of different kind (Wellenreuther & Bernatchez, 2018).

A huge diversity of selective processes has been described across the years, with definitions varying depending on the context. One of the most important features is that selection can either act against polymorphism (or genetic diversity), or in favor of it, therefore leading to the maintenance of alleles. In the first case, supergenes can be under *directional* selection (i.e., one haplotype/allele is associated with better fit), which can lead to the fixation of one allele of a supergene (Lee et al., 2017; Schaeffer, 2008). Additionally, a deficit of heterozygotes can be observed when associated with less fitted phenotype in the case of *divergent* selection (Barth et al., 2017; Jones et al., 2012; Kozak et al., 2017) or in the case of *positive assortative mating* (Ayala et al., 2013). These two processes (i.e., *divergent* selection and *positive assortative mating*) are known to lead to sympatric speciation and therefore do not act in favor of the maintenance of polymorphic supergenes in populations. Supergene maintenance is usually believed to happen through *balancing* selection (Wellenreuther & Bernatchez, 2018) which has been documented through three main processes: (1) *Overdominance*, which occurs when the heterozygote genotype is advantageous (Lindtke et al., 2017), and can happen rapidly when dominant beneficial alleles arise in the two haplotypes of an inversion (Kim et al., 2017); (2) *Varying selection in time and space*, which occurs when fitness of supergene alleles changes with environment conditions, the latter can also change across the range distribution (e.g., a gradient) and/or over time (Cheng et al., 2012; Wallberg et al., 2017; White et al., 2007); or (3) *negative frequency-dependent* selection, which is close to *varying* selection, but occurs when the fittest phenotype is the least frequent and can be mediated by sexual selection (Chouteau et al., 2017).

The long-term persistence of supergenes is therefore often explained by balancing selection mechanisms. However, short-term consequences of supergenes have been poorly (if ever) documented, despite the established accumulation of deleterious alleles (Berdan, Blanckaert, et al., 2022; Berdan et al., 2021) that could have rapid consequences, especially in a context of intensifying global change (Roesti et al., 2022) with possible worrisome effects for endangered species. At the same time, while both long and short-term selective processes remain key to understanding the trajectories of populations, comprehending the origin and maintenance of supergenes is also dependent on demographic processes occurring through the range of the species (Jay et al., 2020; Schaal et al., 2022; Thompson & Jiggins, 2014). Overall, a robust and thorough

characterization of the origin and maintenance of supergenes thus requires a deep understanding of demographic processes which can be tackled by modelling genetic diversity at the scale of whole distribution of a species.

3.2. Objectives

This chapter has two main objectives. First, I aim to characterize the determinism behind the size polymorphism in the Thorny Skate and its possible relation with conservation. Second, as I report the discovery of a size-determining supergene to answer part of the first objective, and given the conservation implications, I demonstrate how the thorough reconstruction of demographic history provides key elements to understand both the origin of the supergene and increase our understanding of the conservation implications.

3.3. A Size-determining Supergene Hampers a Vulnerable Population Recovery

Article in preparation for *Nature*.

Authors:

Pierre Lesturgie, John Denton, Lei Yang, Shannon Corrigan, Jeff Kneebone, Romuald Laso-Jadart, Arve Lynghammar, Olivier Fedrigo, Stefano Mona, Gavin Naylor

3.3.1. Abstract

The Thorny skate (*Amblyraja radiata*) population of the NW Atlantic is endangered (Kulka et al., 2020). It also exhibits a unique discrete size polymorphism not seen in other parts of its range. In 2003 US federal protections were put in place to limit harvesting of all 7 skate species in the region. All but the Thorny skate have shown signs of successful population recovery (Kulka et al., 2020). We conducted a genomic screen of the Thorny skate and discovered a 31 megabase "supergene" contained in an inversion, suppressing recombination, that is associated with the size polymorphism. This "supergene" is inherited in a mendelian fashion (Schwander et al., 2014; Thompson & Jiggins, 2014). In Canadian waters, where there are signs of population recovery (Kulka et al., 2020), the supergene genotypes are in Hardy Weinberg equilibrium. However, in the Gulf of Maine, where population non-recovery is most acute, there was a deficit of heterozygotes, consistent either with assortative mating or selection against heterozygotes. Demographic modelling indicates that the large allele (HB) originally introgressed into the ancestral Thorny skate population in the last 160k years from a congeneric species. The HB allele subsequently spread through much of the NW part of the range where it now appears to have context dependent sub-regional effects on fitness, despite high regional gene flow in the recombining genome prevents speciation and replenish genetic diversity in the Gulf of Maine. The study highlights a rarely considered role for context dependent genetic compatibilities in the conservation and management of endangered populations.

3.3.2. Background

Chromosomal inversions prevent recombination, preserving the integrity and coherence of the linked genes they contain (Dobzhansky & Sturtevant, 1938; Faria et al., 2019; Hoffmann & Rieseberg, 2008; Kirkpatrick, 2010; Wellenreuther & Bernatchez, 2018). Suites of genes in inversions are often referred to as supergenes, which can lead to the Mendelian inheritance of complex phenotypes (Schwander et al., 2014; Thompson & Jiggins, 2014). While the existence of supergene systems has long been acknowledged (C. A. Clarke & Sheppard, 1960), the accessibility of whole genome sequencing (WGS) data has significantly amplified their detection. The presence of supergene-associated traits has been shown in several systems i.e: sociality in ants (Avril et al., 2019; Brelsford et al., 2020; Chapuisat, 2023; Kay et al., 2022; Lagunas-Robles et al., 2021); migratory behavior and adaptation to salinity and temperature in cod (Barney et al., 2017; Berg et al., 2015, 2016), and wing morphology and pattern coloration in butterflies (Joron et al., 1999, 2011). Supergenes are maintained in populations by an interplay between demographic and selective processes (Thompson & Jiggins, 2014) the relative contributions of which can be difficult to disentangle without careful reconstruction of the demographic history of populations based on neutral markers. While varying selection in space and time is often invoked to explain long-term persistence of supergenes (Wellenreuther & Bernatchez, 2018), the absence of variability resulting from limited recombination alongside intricate phenotypes impede swift adaptation responses to rapid environmental shifts. This could critically impact endangered species, underscoring the importance of comprehending the potential effect of supergenes on the short-term evolutionary dynamics of species.

The Thorny skate (*Amblyraja radiata*) is a vulnerable species inhabiting the coastal shelves from South Carolina in the Northwest Atlantic (NWA) to the Barent Sea and the British islands in the Northeast Atlantic (NEA, Figure 3.2) (Kulka et al., 2020). The species used to be intensively fished in NWA which leads to a steep decline in US stocks from which they have not yet recovered, in spite of stringent conservation measures designed to protect them (Kulka et al., 2020; Sosebee et al., 2016). The NWA is home to two morphs of different size each displaying characteristic growth curves (Mcphie & Campana, 2009; Sosebee, 2004; Sulikowski et al., 2005; E. G. Templeman, 1984; W. Templeman, 1987), hereafter referred to as large and small morphs. The large morph, which is restricted to NWA, reaches a maximum size of 104 cm Total Length (TL) while the small morph, occurring over the whole range of the species, reaches a maximum size 72 cm TL (W.

Templeman, 1987). To date the genetic underpinnings of this morphological polymorphism have eluded detection, as neither microsatellite nor mitochondrial data show genetic differentiation between large and small forms (Denton et al., n.d.; Lynghammar et al., 2016). At the same time, genetic diversity patterns across the species range remain poorly understood. Mitochondrial data suggest weak population genetic structure and isolation by distance across the entire range (Chevolot et al., 2007) while microsatellite data show genetic differentiation between the NWA and NEA regions (Lynghammar et al., 2016).

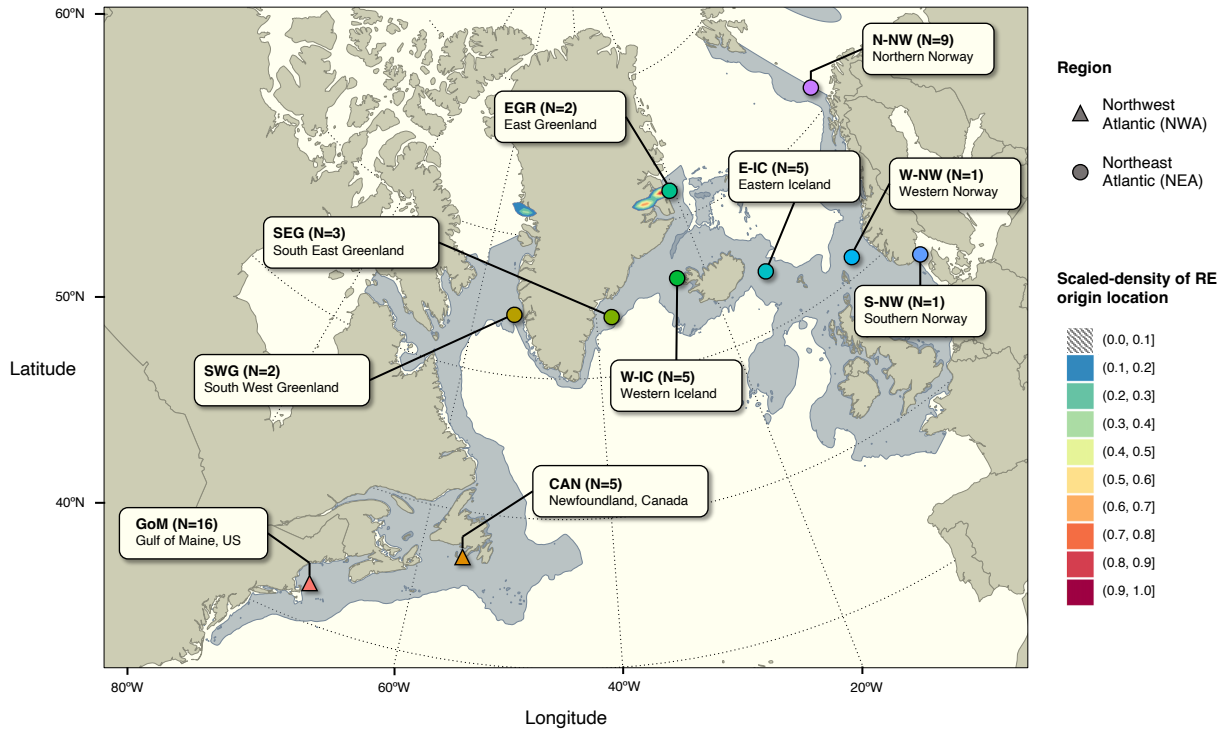

Figure 3.2. Whole Genome sample scheme of thorny skates. The range distribution of the Thorny skate is filled in blue. Map is displayed using a central conic projection at latitude 60ºN. Shape of the sampling location point represents the geographical region: circle = Northeast Atlantic (NEA); triangle = Northwest Atlantic (NWA). The sequential colored areas represent the scaled density of the range expansion origin inferred using the TDOA algorithm computed 100 times (see results).

Here we set out to understand the genomic and/or environmental origins of the size polymorphism in the northwest Atlantic and its implications for the conservation of the species. We first established a high-quality reference genome for the Thorny skate based on a combination of long read (PacBio), Hi-C and Bionano and Illumina short read sequencing in collaboration with the Vertebrate Genomes Project. We subsequently collected whole-genome ~17x Illumina sequence data for a sample of 49 individuals spanning the species' distribution range (Figure 3.2). This

approach allowed us to discover a size-determining ~31 Mb supergene contained in a chromosome inversion on chromosome 2, that was polymorphic in the NW Atlantic region. We further screened 470 individuals across the range of the species to characterize the distribution of the supergene's alleles and confirmed its association with size. To better understand the origin, maintenance and allelic distribution of the supergene, we reconstructed the demographic history of the species based on analysis of millions of neutral genome-wide SNPs. PacBio sequencing of an individual of the sister species *A. hyperborea* revealed the supergene to be present in at least one of the congeners. When this information was combined with the extant geographic distribution of both alleles and the historical reconstruction of demography, we were able to infer that the supergene was most likely transmitted to *A. radiata* through introgression from a congener. Our findings further show that the supergene is hampering the recovery of the highly vulnerable US stocks, presenting a particular challenge for Thorny skate conservation and management in the NW Atlantic.

3.3.3. Results

3.3.3.1. Population structure

We used Principal Component Analysis (PCA) of SNP variation to explore population structure. The first axis (~14% of total variance) revealed two clusters, corresponding to the North Eastern (NEA) and the North Western Atlantic (NWA) regions respectively (Figure 3.3-A). The second axis (~2% of total variance) separated individuals sampled from the Gulf of Maine (GoM) with those from Newfoundland (CAN), the two NWA sampling sites (Figure 3.3-A). We estimated the individual ancestry coefficients and the most likely number of ancestral populations using the sNMF algorithm (Frichot & François, 2015). The cross-entropy criterion identified K=2 as the most likely number of clusters (Supp. Figure 3.7-A-B), perfectly matching those detected by the PCA (Figure 3.3-A). We further run both the PCA and sNMF within each cluster separately. The first two PC axes explained as low as ~5% and ~4% of the total variance in NWA and NEA respectively, and in both datasets K=1 was the most likely number ancestral populations. However, both algorithms harbored signatures of fine scale population structure as suggested by clinal distribution of genetic variation within both regions (Figure 3.3-B-C and Supp. Figure 3.7-C-F). All pairwise F_{ST} comparisons were statistically significant ($p \leq 0.001$) and generally consistent with the results provided by the clustering algorithms. Values ranged from 0.002 to 0.004 in intra-

cluster comparisons (i.e., within NEA and within NWA) and from 0.173 to 0.189 when comparing NEA vs NWA sampling sites (Figure 3.3-D).

Table 3.1. Summary statistics for each sampling location. Number of individuals sampled for the whole genome study (N_{WG}) and the screening study (N_{SC}). For each sampling location with $N_{WG} \geq 5$: total number of SNPs (N_{SNPs}), total number of sites (N_{sites}), mean pairwise difference (θ_π) and Watterson's estimator of genetic diversity (θ_w) both scaled N_{sites} and Tajima's D (TD). For sampling sites included in the screening study ($N_{SC}>0$), total number of individuals carrying each genotype (N_{HBHB}, N_{HBHS}, N_{HSHS}) and HB allele frequency (f_{HB}). Number of individuals carrying each genotype in bold are not in Hardy-Weinberg equilibrium (HW exact-test: $p<0.001$).

		N_{WG}	N_{SC}	N_{SNPs}	N_{sites}	θ_π	θ_w	TD	$N_{HB/HB}$	$N_{HB/HS}$	$N_{HS/HS}$	f_{HB}
NWA	GoM	16	284	13,926,040	439,435,032	0.0063	0.0079	-0.7903	**78 (f=0.27)**	**10 (f=0.04)**	**196 (f=0.69)**	0.29
	CAN	5	40	11,872,562	604,725,402	0.0063	0.0069	-0.4877	4 (*f*=0.09)	23 (*f*=0.58)	13 (*f*=0.33)	0.39
NEA	SWG	2	7	-	-	-	-	-	0 (*f*=0)	0 (*f*=0)	7 (*f*=1)	0
	SEG	3	0	-	-	-	-	-	-	-	-	-
	E-GR	2	2	-	-	-	-	-	0 (*f*=0)	0 (*f*=0)	2 (*f*=1)	0
	W-IC	5	50	10,275,797	558,262,795	0.0058	0.0065	-0.5318	0 (*f*=0)	0 (*f*=0)	50 (*f*=1)	0
	E-IC	5	34	9,982,473	535,940,022	0.0059	0.0066	-0.5212	0 (*f*=0)	0 (*f*=0)	34 (*f*=1)	0
	W-NW	1	0	-	-	-	-	-	-	-	-	-
	S-NW	1	5	-	-	-	-	-	0 (*f*=0)	0 (*f*=0)	5 (*f*=1)	0
	N-NW	9	47	12,708,852	538,100,423	0.0057	0.0069	-0.6983	0 (*f*=0)	0 (*f*=0)	47 (*f*=1)	0

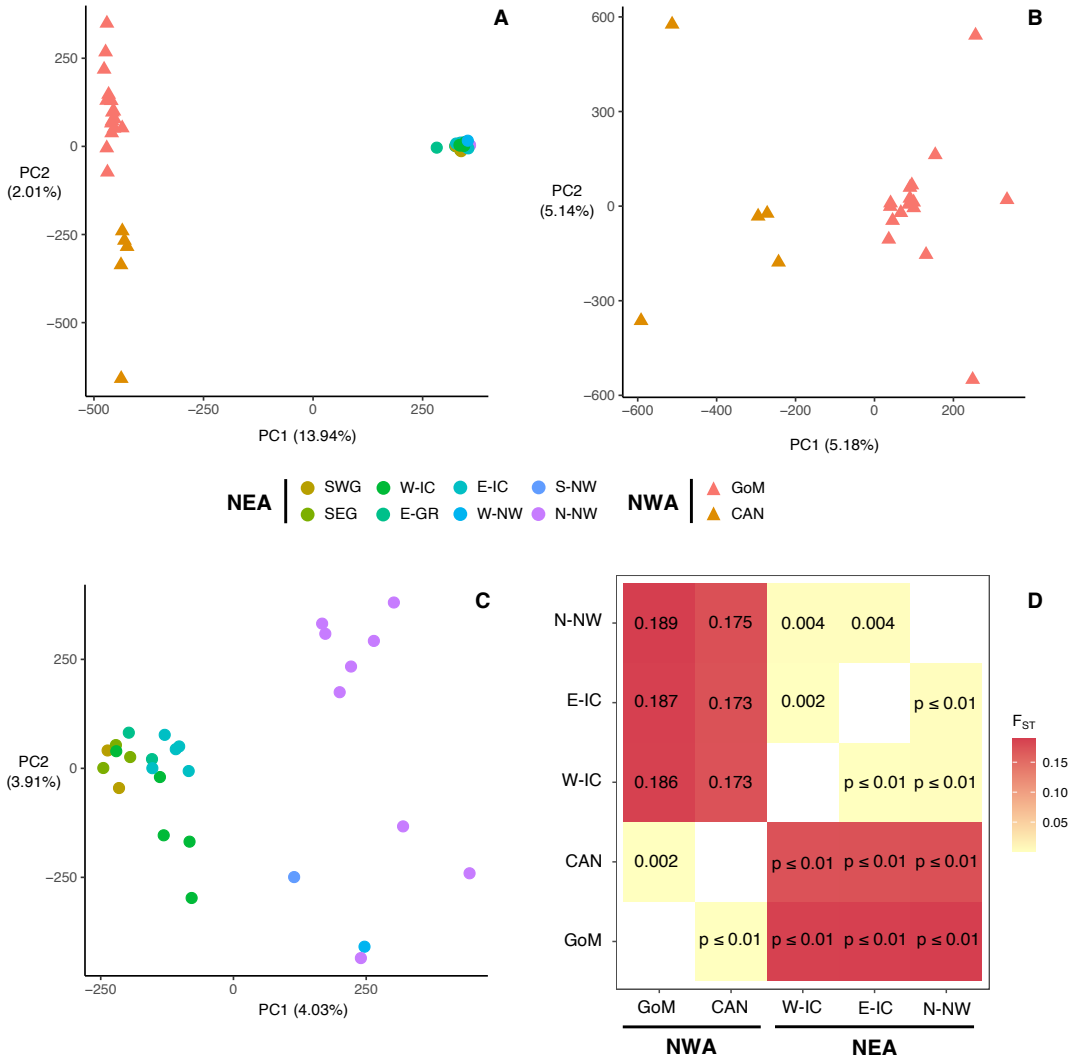

Figure 3.3. Population structure of the thorny skate. Panels A-C: Principal Component Analyses (PCA) using all individuals (panel A), only GoM and CAN individuals (cluster NWA, panel B) and only individuals from SWG, SEG, W-IC, E-GR, E-IC, W-NW, S-NW and N-NW (cluster NEA, panel C). Panel D: Heatmap of pairwise F_{ST} values between sampling locations with $N \geq 5$ (upper left) and associated significancy evaluated using 1000 permutations for each pairwise comparison (lower right).

3.3.3.2. Detection of a supergene

The two size morphs only occur in the Northwestern part of the species range. Genome wide SNP analyses of individuals from this region suggest weak geographic population structure but no genetic association related to size (Supp. Figure 3.7-G-H). However, when we used a genomic sliding windows analyses of PCA (to compute the % of total variance explained by the first axis within each window) over the pooled NWA sample, we identified a ~31Mb region in chromosome 2 (from ~17Mb to 12 48Mb; Figure 3.4-A) that was strikingly different from the surrounding parts

of the genome and the genome-wide average. A comparable sliding window analysis using *Tajima's D* (TD) showed a similar pattern in which TD values were 3-times more positive than it was outside the 31MB region, suggesting stronger population structure and an excess of high frequency variants (Figure 3.4-B). Local PCA computed within this region displayed three clusters segregated by the first axis (Figure 3.4-C). The two most distant clusters on the first axis are characterized by an excess of the two alternative homozygous genotypes, while individuals in the middle of the first axis displayed an excess of heterozygous genotypes (Figure 3.4-D). This result was corroborated by the sNMF which found K=2 as the most likely number of ancestral populations, corresponding to those found by the local PCA, with individuals showing an excess of heterozygotes being almost exactly half admixed between them (Supp. Figure 3.8-A-B). Finally, we investigated Linkage Disequilibrium (LD) in both the pooled sample and in the two clusters separately (Figure 3.4-E and Supp. Figure 3.8-E-F): the region is characterized by strong LD in the pooled sample when compared to the rest of chromosome 2. Conversely, LD values are similar to the rest of the genome (or lower) when computed within each previously identified cluster. Additionally, F_{ST} values between the two clusters characterized by an excess of homozygosity reached up to ~1 in the region (suggestive of total divergence) while remaining distributed around ~0 outside (Figure 3.4-F). All these results suggest that recombination in this region has been suppressed. Given the occurrence of 226 annotated genes in the 17-48Mb region (NCBI:txid386614), we refer hereafter to this region as a supergene, characterized by two haplotypes (HB and HS) inherited in a Mendelian fashion (Thompson & Jiggins, 2014). Individuals enriched for alternative homozygous genotypes are HB/HB, individuals enriched for reference homozygous genotypes are HS/HS, and individuals enriched for heterozygotes are HB/HS. Preliminary results suggested that the size of individuals was different between the two homozygous genotypes: HB/HB had an average size of ~71.7cm and HS/HS of ~53.9cm. However, sample sizes are too low (N=9 and N=10 for HB/HB and HS/HS respectively) to model confounding factors such as sex and maturity. When a local PCA was run including the NEA samples, the first axis segregated NEA and HS individuals from HB before the NEA-NWA geographical divergence detected by the genome-wide structure analyses (Figure 3.4-G). The same pattern was observed when computing the individual ancestry with the sNMF (Supp. Figure 3.8-C-D), which implies that NEA individuals are all HS/HS and the divergence between HB and HS alleles predates the split between NEA and NWA regions.

3.3.3.3. Genotype screening and size association

To further investigate the association between the supergene genotypes and size, we first selected two regions each with > 4 SNPs discriminating HB and HS alleles within the supergene and further screened by PCR and Sanger sequencing 501 individuals (470 after filtering, see supplementary results) throughout the range of the species (Table 3.1). HB was absent in NEA, consistent with the lack of size polymorphism in this part of the range. Conversely, HB reached a frequency of 0.29 and 0.39 in GoM and CAN respectively (Table 3.1). However, the distribution of genotypes in the two sampling sites was strikingly different: GoM displayed a strong deficit in heterozygotes (only 10 out of 284 individuals, Hardy-Weinberg exact-test: $p < 0.001$), while CAN was in Hardy-Weinberg equilibrium. We then investigated the relationship between Size and Haplotype controlling for Maturity and/or Sex using linear models in a bayesian framework (Supp. Table 3.2) in the 243 GoM individuals with no missing information on any trait. The Leave-One-Out cross validation indicated the model including Size and Maturity only as the most accurate (see supplementary results). Posterior distribution of size for HB/HB and HB/HS individuals largely overlapped, with median values and 95% credibility intervals (averaged over the levels of Maturity) of 66.95 cm (95% CI [64.73, 69.17] cm) for the former and 65.61 cm (95% CI [60.81, 70.50] cm) for the latter. Conversely, size for HS/HS individuals was strikingly lower (median value of 50.68 cm, 95% CI [49.28, 52.07] cm) and associated with disjunct posterior distribution from HB carriers (Figure 3.4-H).

3.3.3.4. Historical demography

The restricted distribution of HB might be the consequence of neutral and/or selective forces. To better understand the origin, maintenance and historical demography of the size polymorphism we ran the Pairwise-Sequentially Markovian Coalescent (PSMC) algorithm on each individual. PSMC curves were (almost) identical for every individual at the regional scale but the dynamics differed between NEA and NWA, whose trajectory started to diverge ~1My ago (Figure 3.5-A and Supp. Figure 3.9-A). While the exact date of divergence between the two trajectories may be inaccurate (Lesturgie, Planes, et al., 2022), there is a clear separation of the evolutionary trajectories between NEA and NWA in ancient times, consistent with population structure (Figure 3.3). We further examined the genome for signatures of range expansion (RE) to understand the colonization history. PacBio sequencing data of one *A. hyperborea* individual was used to polarize SNPs found in the 49 *A. radiata* individuals. We further computed the directionality index (ψ) between each

pair of individuals and fit the time difference of arrival (TDOA)(Peter & Slatkin, 2013) location algorithm to test for the occurrence of a range expansion and find its geographic origin. To have a balanced sampling design, we randomly extracted one individual per sampling location and repeated the process 100 times to obtain the density distribution of the center of origin of the range expansion, which was always found in NEA region, off the coast of Greenland, with more than 80% runs located on the eastern coast of Greenland (Figure 3.2). We additionally investigated Runs of Homozygosity (ROH) in sampling locations with N≥5 individuals. ROH were arbitrarily classified in different length categories (Supp. Figure 3.10). The number (N_{ROH}) and genomic coverage (SUM_{ROH}) were always the lowest in the two Iceland sampling sites (W-IC and E-IC) and strikingly higher in GoM followed by CAN and N-NW (Supp. Figure 3.10). This further supports the idea that Iceland/Greenland lies in the center of the ancestral distribution of the species while GoM, CAN and N-NW are the more derived populations.

Based on all these findings (Population structure, PSMC, RE, and ROH distribution), we investigated five demographic scenarios describing patterns of migration and divergence between and within the two meta-populations (NEA and NWA; Figure 3.5-C and Supp. Figure 3.11-A-E). To this end, we applied the maximum likelihood approximation approach of *fastsimcoal* (Excoffier et al., 2021) by modeling the set of two-dimensional Site Frequency Spectrum (2D-SFS) calculated between locations of N ≥ 5 (Table 3.1). The AIC criterion computed after choosing the best out of 10 replicates of each model indicated IMM-5-NM-STOP as the most likely (Fig. S5F). IMM-5-NM-STOP depicts two meta-populations composed (in this order) of GoM and CAN sampling locations (NWA meta-population) and W-IC, E-IC and N-NW sampling locations (NEA meta-population). Deme effective sizes were highly similar between NEA and NWA demes (N_E ~ 80000, 95% CI [74595, 81106] and N_W ~ 82000, 95% CI [78482, 90962], Supp. Table 3.3). However, demes were twice as connected in NEA than in NWA despite largely overlapping confidence intervals (Nm_E~118 95% CI [76.74, 144.91] and Nm_W~61 95% CI [49.67, 124.05], respectively, Figure 3.3) suggestive of high local connectivity within both meta-populations. Going backward in time, the two meta-populations were isolated until T_{CH}~160000 years (95% CI [47000, 163000]) when started an asymmetrical exchange of migrants three times greater from NEA to NWA than otherwise ($Nm_{E\rightarrow W}$~5.1, 95% CI [4.88, 13.73] and $Nm_{W\rightarrow E}$~1.5, 95% CI [1.69, 5.56] per generation). All lineages finally merged into an ancestral population of N_{ANC}~101000

(95% CI [99116, 106634]) at T_{DIV}~891000 (95% CI [800000, 920000]) years ago (i.e., the NEA-NWA divergence time).

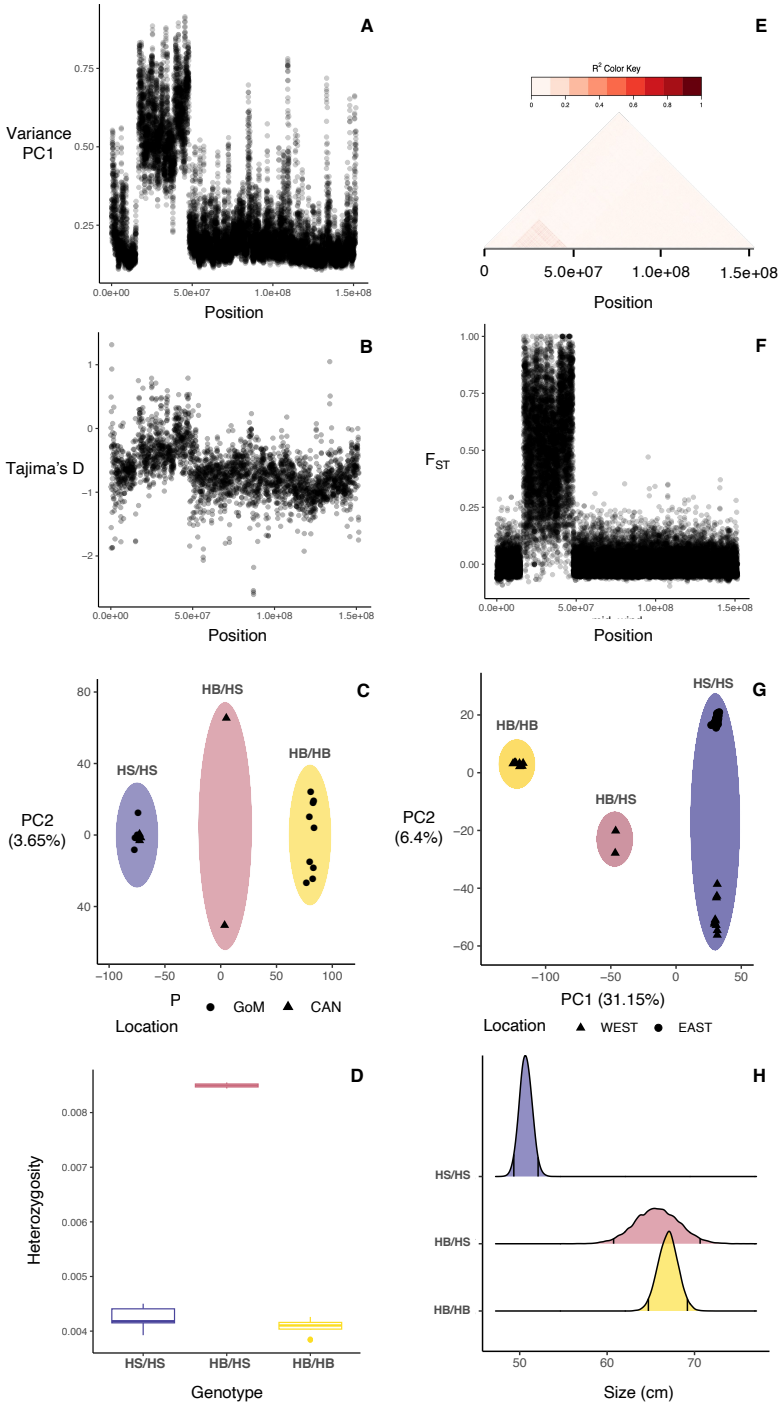

Figure 3.4. Size-determining supergene in chromosome 2. Panels A-B: Sliding windows analyses of the proportion of variance explained by the first axis of a PCA (panel A) and of Tajima's D (panel B) computed in NWA on chromosome 2. Panel C: Local PCA within the 17000000-48000000 region of chromosome 2 (supergene region) including only NWA individuals. Dot shape represents the sampling location and color the genotype at the supergene: HS/HS (purple), HB/HS (red) and HB/HB (yellow).

Panel D: Proportion of heterozygotes within the supergene region for each genotype. Panel E: Heatmap of the pairwise linkage disequilibrium between SNPs. Color gradient represent the value of the R^2 correlation between SNPs. Panel F: Sliding window F_{ST} between HB/HB and HS/HS NEA individuals. Panel G: Local PCA within the supergene region including both NWA and NEA individuals. Panel H: Posterior distribution of the size as estimated by model HaploMat for each genotype: HS/HS (purple), HB/HS (Red) and HB/HB (yellow). 2.5% and 97.5% quantiles are represented in each distribution by vertical bars.

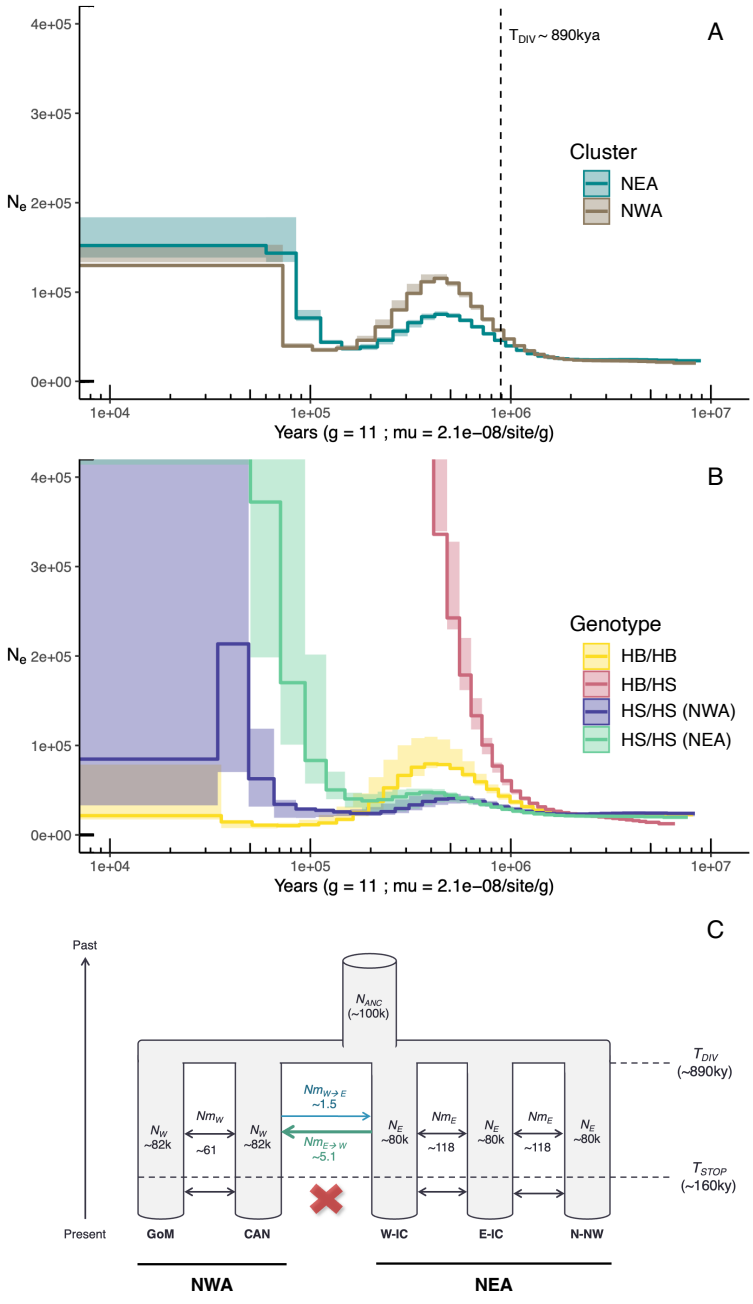

Figure 3.5. Global and within supergene historical demography. Panel A-B: PSMC computed using the whole genome data in two individuals representative of NEA (turquoise) and NWA (brown) (A) and within the chromosome 2 supergene in four individuals: HB/HB (Yellow), HB/HS (Red), HS/HS for

NWA (Blue), HS/HS for NEA (Green) (B). Shaded areas are computed after 100 bootstraps. The vertical dotted line in panel A represents T_{DIV} (divergence between NEA and NWA) as estimated by fastsimcoal under IMM-5-NM-STOP model. Panel C: Demographic model IMM-5-NM-STOP with maximum likelihood estimates for each parameter.

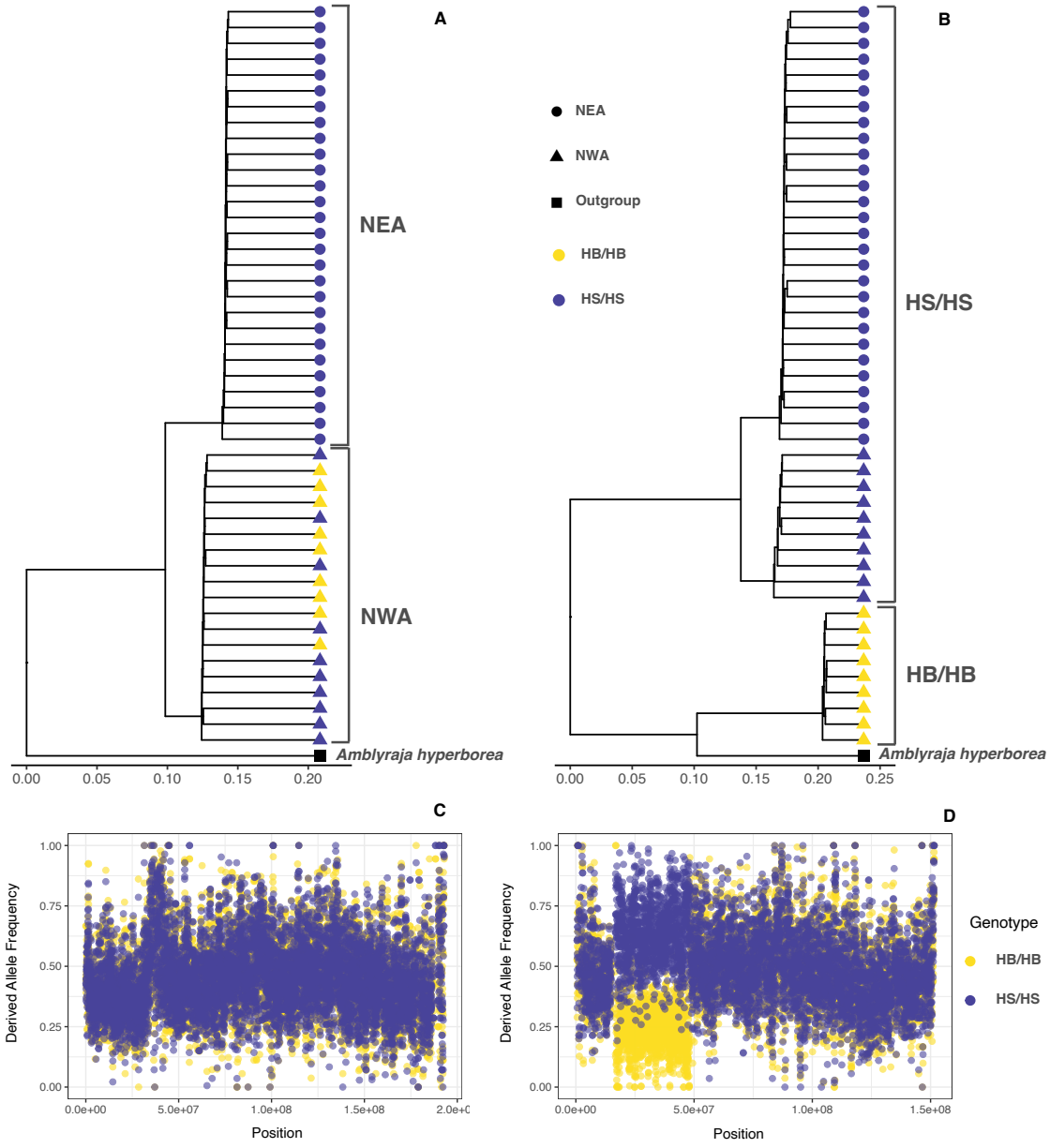

Figure 3.6. Characterization of the supergene's introgression. Panel A-B: UPGMA trees based on genetic distance computed in chromosome 1 (A) or in the supergene region (B) using all individuals but the two heterozygotes (HB/HS). Dot shape represent the geographic cluster of origin (circle: NEA; triangle: NWA, square: outgroup) and color the genotype at the inversion (purple: HS/HS, yellow: HB/HB, black: outgroup). Panel C-D: Sliding windows of the average derived allele frequency in chromosome 1 and 2 for HB/HB (yellow) and HS/HS (blue) groups in GoM.

3.3.3.5. Origin of the supergene

We computed Nei's pairwise genetic distance both in the supergene and in the first chromosome between all sampled individuals (all *A. radiata* and one *A. hyperborea*). The UPGMA tree computed on chromosome 1 confirmed the genetic structure previously observed with clustering algorithms in *A. radiata* (i.e., the separation between NEA and NWA) and *A. hyperborea* as an outgroup. Conversely, the UPGMA tree computed in the supergene clearly suggested that the HB allele is more closely related to *A. hyperborea* than to the HS allele (Figure 3.6-A-B). This result was corroborated by contrasting the frequency of derived alleles (f_{DER}) computed within and outside the supergene: f_{DER} was much higher in HB/HB individuals than in HS/HS individuals within the supergene, while no difference was observed in the rest of the genome (Figure 3.6-C-D). Following (Cahill et al., 2016) we further estimated the divergence time between alleles HB and HS at ~1.5M by computing the PSMC in heterozygote individuals in the supergene region (Figure 3.5-B and Supp. Figure 3.12-B). The PSMC computed in both HS/HS and HB/HB individuals within the supergene was strikingly different to the trajectory estimated over the whole genome (Figure 3.5-A). Similarly, the one hundred PSMC curves obtained by randomly sampling each time a 31Mb region in the genome from both HB/HB and HS/HS individuals were incongruent with the supergene's PSMC curves but consistent with the genome-wide estimates (Supp. Figure 3.12-A).

3.3.4. Discussion

The striking size polymorphism in the vulnerable Thorny skate *A. radiata* (Kulka et al., 2020; Mcphie & Campana, 2009; Sosebee, 2004; Sulikowski et al., 2005; E. G. Templeman, 1984; W. Templeman, 1984) offers a rare opportunity to dissect the genetic basis of size variation and to improve our understanding of how it affects the trajectory of endangered populations. Using Whole Genome Sequencing data, we identified a ~31 Mb size-determining supergene characterized by two alleles HB and HS between which recombination is hampered (Figure 3.4). HB was only found in NWA (Table 3.1) where genotype distribution was strikingly different with a significant deficit of heterozygotes detected in GoM but not in CAN (Table 3.1). Association with size was subsequently investigated through a Bayesian linear model in GoM (the site with the largest sample size) accounting for maturity stage only as Sex did not explain substantial variability in size (Supp. Table 3.2). Individuals carrying at least one copy of HB had a size of ~67

cm while homozygotes for HS had a size of ~51 cm (Figure 3.4), suggesting the dominance of HB. Supergenes are known to determine a wide variety of traits (Avril et al., 2019; Barney et al., 2017; Berg et al., 2015, 2016; Brelsford et al., 2020; Chapuisat, 2023; Kay et al., 2022; Lagunas-Robles et al., 2021), but, to our knowledge, this is the first study to directly highlight a clear association with a continuous quantitative phenotype in a group of individuals living in sympatry. Notably, size is a continuous trait with polygenic determinism across various species (Bouwman et al., 2018; Boyle et al., 2017; Wood et al., 2014). It is likely that the near-discrete size-determinism revealed in our study involves several genes spanning the supergene region but we cannot exclude the interplay between them and other genes across the genome, similarly to what has been described in others supergene systems (Errbii et al., 2023; Jones et al., 2012). Given the substantial length of the supergene (~31Mb) and the presence of numerous genes (~226), it is also likely that this region controls multiple phenotypes, as exemplified in ants, cods or butterflies (Avril et al., 2019; Barney et al., 2017; Berg et al., 2015, 2016; Brelsford et al., 2020; Chapuisat, 2023; Joron et al., 2011; Kay et al., 2022; Lagunas-Robles et al., 2021). More data will be required to better characterize the phenotypic consequences of this supergene.

Supergenes are maintained in space and time through a combination of demographic (neutral) and selective processes (Thompson & Jiggins, 2014). To shed light on the distribution of the two supergene alleles and of the genotypes in NWA, we investigated the historical demography of *A. radiata* through its whole range. Previous studies using mitochondrial and microsatellite markers were inconclusive, suggesting respectively low (Chevolot et al., 2007) to high levels of genetic differentiation across the whole range (Lynghammar et al., 2016). Defining a comprehensive demographic scenario explaining the whole history of a species is a challenging task, but whole genome data provide an informed basis upon which to test competing models of demographic history. First, clustering algorithms, F_{ST} and PSMC analyses supported the unambiguous signature of long-term divergence between NEA and NWA and of a weak but spatially continuous genetic differentiation within each region (Figures 3.3, 3.5, Supp. Figures 3.7 and 3.9), consistent with recent mitogenomic findings (Denton et al., n.d.). We further detected signatures of range expansion (RE) by investigating the distribution of shared derived allele frequencies (Excoffier et al., 2009; Peter & Slatkin, 2013): NEA region represents the ancestral range of the species, which most likely expanded from Greenland or Iceland first eastward to the European coasts, and then westward, colonizing NEA (Figure 3.2). This is consistent with the distribution of ROH, which

were never larger than ~1.41Mb and were more numerous and longer in peripheral populations than in Greenland (Supp. Figure 3.10). REs are indeed characterized by a series of founding events determining more genetic drift in populations far away from the origin of the expansion (Slatkin & Excoffier, 2012), supporting the finding of GoM, CAN and N-NW as the more derived and E-IC and W-IC as the more ancestral populations. Based on these results, we compared a set of five demographic scenarios under the inferential framework of *fastsimcoal* (Excoffier et al., 2013). The most likely scenario highlights a divergence between NWA-NEA occurring ~900ky ago (consistently with the time at which PSMC individual curves of NWA and NEA start to differ, Figure 3.5 and Supp. Figure 3.9). The separation between the two clusters corresponds to the colonization time of the NWA regions (Figure 3.5 and Supp. Table 3.3). After colonization, the NWA and NEA metapopulations remained first connected by an asymmetrical exchange of migrants ~3 times higher from NEA to NWA than otherwise, becoming isolated ~160ky, possibly in consequence of the spread ice sheets during the last ice age. The modelling of the genome wide diversity would therefore suggest that HB either originated or introgressed in the NWA regions more recently than ~160 ky, as this allele is absent in NEA. To disentangle the two hypotheses, we further investigated the supergene regions in *Amblyraja hyperborea* (Figure 3.6), which provided strong evidence that HB did not originate in *A. radiata*, but rather introgressed from a donor species. Indeed, i) *A. hyperborea* variants were more frequent in HB than in HS (Figure 3.6), which was confirmed by a phylogenetic tree clustering HB/HB individuals with *A. hyperborea* in the supergene region (Figure 3.6); and ii) clustering algorithms indicate that HB/HB individuals are separated by all HS/HS independently of their geographic origin (NWA or NEA), in stark contrast with the genome wide results, suggesting that the HB-HS divergence predates the NEA-NWA divergence (Figure 3.4). Indeed, the PSMC estimated the divergence between HB and HS at ~1.5 My (Figure 3.5), likely corresponding to the separation between *A. radiata* and the donor species (which is not necessarily *A. hyperborea*, though *A. hyperborea* contains the HB allele). Given the demographic scenario, the time separation between HB and HS as well as their present-day spatial distribution, and the divergence between *A. radiata* and *A. hyberborea*, we believe that the time when migration stopped between NEA and NWA provides a reasonable upper limit to the HB introgression into NEA individuals from a donor species yet to be identified. We notice that none of the congeneric species of *A. radiata* show size polymorphism and their size distribution seem more similar to HB carriers except for *A. doellojuradoi*, seemingly reaching a

maximum size closer to that of HS/HS (Last et al., 2016). This suggests that the dynamics of the evolution of this supergene will need multi-species investigations to be elucidated, similarly to what has been done to uncover the history of the social supergene in fire ants and timing the introgression events (Helleu et al., 2022; Stolle et al., 2022; Z. Yan et al., 2020).

Demographic modelling of genome wide data highlighted high connectivity between GoM and CAN (Nm~61, Figure 3.5 and Supp. Table 3.3). This suggests that the strongly different distribution of supergene genotypes between GoM and CAN cannot (at least totally) be explained by demographic processes. This is further highlighted by the PSMC curves within the supergene: neither HB/HB nor HS/HS individuals in NWA show a coalescence rate dynamic over time consistent with the genome-wide pattern (Figure 3.5 and Supp. Figure 3.9). Supergenes can promote local adaptation even when gene flow is high (Schaal et al., 2022), and, more generally, are usually maintained by various selective pressures (Berdan, Flatt, et al., 2022; Thompson & Jiggins, 2014; Wellenreuther & Bernatchez, 2018). Here, we argue that positive assortative mating could explain the observed deficit in heterozygotes in GoM. This hypothesis is driven by the previously discussed physical incompatibility in mating between larger and smaller skates in the Gulf of Maine (Denton et al., *in prep*; Lynghammar et al., 2016) due to evident differences in maximum size and size-at-maturity (Sosebee, 2004; Sulikowski et al., 2005; W. Templeman, 1987). These differences tend to disappear northwards: skates sampled off the coast of Newfoundland (i.e., CAN sampling site) do not show a bi-modal distribution of size at first maturity as in GoM, but rather a unimodal distribution associated with larger variance (W. Templeman, 1987). Maturity can covary with the environment (Martin & Leberg, 2011) and could be a key factor in explaining possible mating in CAN but not in GoM. This would have considerable implications for the conservation of the NWA population as a whole in the context of global warming as increasing sea temperature can alter age and size-at-maturity (Niu et al., 2023). We further note that whilst assortative mating itself is a process leading to sympatric speciation (Straw, 1955), recombination between HB and HS carriers in GoM could be maintained through the very high connectivity with CAN (Figures 3.3, 3.5 and Supp. Table 3.3), thus explaining the absence of neutral divergence between small and large individuals in GoM (Figure 3.3 and Supp. Figure 3.7). Finally, we note that it is possible that the supergene controls other traits than size, which may be under negative selection against heterozygotes in GoM.

We discovered a size-determining supergene, and highlight, for the first time, the major short-term implications such a system may have in the trajectory of vulnerable and non-vulnerable populations. Our study first demonstrates (once more) the importance of reconstructing the neutral evolutionary history of a species, an essential background needed to uncover complex non-neutral processes. The inferred demographic scenario was of paramount importance not only to interpret the spatial distribution of the two alleles of the supergene but also to issuing a hypothesis over the genotype distribution in NWA, which in turn carry profound implications for the conservation of the vulnerable Thorny skate (Kulka et al., 2020). This is striking in light of the different population trends across NWA, with the Gulf of Maine still showing significant declines long after the end of fishing in the region while CAN is recovering (Kulka et al., 2020). It is likely that the observed deficit of heterozygotes at the supergene region in GoM, likely linked to positive assortative mating, is hindering population recovery with large and small skates probably competing for the same resources while failing to mate. This is not happening in CAN where heterozygotes are common and no barriers to reproduction seem to exist between the two morphs, which is consistent with the recovering population trends. The high gene flow across the region suggests that northern demes could play a fundamental role in preventing the speciation of two distinctly sized species in the GoM and contribute to replenish the genetic variability of the region, despite the lack of census size recovery. This complex pattern suggests for the first time the responsibility of a supergene on the fate of an endangered population, and strongly emphasizes the necessity for comprehensive region-wide conservation plans and highlight the crucial contribution of evolutionary biology to more applied research fields. Our study also demonstrates that this supergene has been introgressed in *A. radiata* in the last ~160000 years from a donor species which will have to be characterized in the future to better understand the supergene's implications at higher taxonomic level. We emphasize two additional implications in evolutionary biology: i) supergene system evolved independently in Chondrichthyes, a clade of vertebrates which remains understudied at the genomic level; ii) this is, to our knowledge, the first direct example of a continuous quantitative trait whose distribution is largely explained by a simple Mendelian inheritance, hence providing the opportunity to dissect the genetic determinism of size. This will a have tremendous impact in fields such as population genetics, quantitative genetics, macro-evolution and ecology by providing an opportunity to understand better a complex determinism and both the short and long-term dynamics of a supergene associated with conservation issues.

3.3.5. Material & Methods

3.3.5.1. Whole Genome Sequencing

49 Amblyraja radiata individuals were sampled from west to east off the Gulf of Maine in US (GoM, N=16), Newfoundland in Canada (CAN, N = 5), South West Greenland (SWG, N=2), South East Greenland (SEG, N=3), East Greenland (EAG, N=2), West Iceland (W-ICE, N=5), East Iceland (E-ICE, N=5), West Norway (W-NOR, N=1), South Norway (S-NOR, N=1) and North Norway (N-NOR, N=9). Genomic DNA was extracted using the E.Z.N.A. Tissue DNA Kit (Omega Bio-Tek, Inc., Norcross, GA, USA) following the manufacturer's instructions. The extracted DNAs were then sent to the Next-Generation Sequencing (NGS) Core of the University of Florida's Interdisciplinary Center for Biotechnology Research (UF ICBR) for QC. After that, libraries were prepared, pooled, and loaded on the Illumina NovaSeq 6000 platform for whole genome sequencing with S4 flow cell and 2x151 setup.

3.3.5.2. Repeat annotation and masking

We downloaded the reference genome of the Thorny Skate from the NCBI website (BioProject number: PRJNA591369). The genome was first masked using the Chondrichthyes database in a first run of *RepeatMasker* v.4.1.0 (Smit et al., 2015). We then created a de novo database for the *A. radiata* by using *RepeatModeler* v.2.0.3 (Smit & Hubley, 2015) on the genome masked at the first step. Finally, we masked the repeated elements annotated in the de novo database by running *RepeatMasker* a second time on the initially masked genome. We finally extracted a bed-file of the masked regions further used in downstream bio-informatic analyses.

3.3.5.3. Main bioinformatics pipeline

Reads were trimmed for adapter and quality using *bbduk* from *bbmap* v.38.44 suite (sourceforge.net/projects/bbmap/). After checking for quality using *FastQC* v0.11.7 (Andrews, 2010), reads were mapped against the reference genome using *bwa mem* algorithm v.0.7.17 (H. Li, 2013) with -M option. Mapped reads were sorted and indexed using *samtools* v.1.10 (Danecek et al., 2021) and then marked for duplicates using *Picard* v.2.21.2 MarkDuplicates (Broad Institute, 2019). Except for the *PSMC* analysis (see below), indexed reads were fed to *GATK* v.4.1.9.0 (McKenna et al., 2010) haplotypecaller algorithm for variant discovery using the -gvcf option to obtain individual variant calling fil e (VCF). Individual VCFs were then combined together using CombineGVCF to build datasets with different number of individuals according to the downstream analysis (see below). Joint calling was then performed for each dataset using GenotypeGVCF by

including both monomorphic and polymorphic sites (all-sites argument) which are necessary for scaling correctly genetic diversity. We then selected the 49 identified autosomes and removed the regions annotated as repeats using the bed-file produced by the repeat masking step. By combining VariantFiltration and SelectVariant *GATK's* scripts, we then filtered out sites with Mapping Quality < 40 and marked genotypes as missing if genotypic depth (i.e., depth per individual and per site) was below 6 or over 50. We further removed chromosome 2 and 8 for all genome-wide historical demographic analyses after genomic scans identified two potential large chromosomal inversions. Additional filters were applied on the resulting VCF depending on the analysis.

3.3.5.4. Population structure

Population structure datasets were filtered using a combination of *vcftools* v.0.1.16 (Danecek et al., 2011), *bcftools* v.1.15 and custom python scripts, keeping only bi-allelic SNPs with a missing data rate of less than 20% and discarding SNPs heterozygous in more than 80% individuals and with a minor allele frequency < 0.05. VCFs were binned by only selecting SNPs that were at least 1kb apart to each other to account for linkage disequilibrium. Depending on the analysis and on the scale of investigation, we built different datasets. We first built a dataset including all individuals: ALL (N=49). Based on global population structure (see results), two additional datasets were created to investigate fine scale structuration: NWA dataset (N=21), which only included individuals from CAN and GoM sampling locations, and NEA dataset (N=28) including all the remaining individuals (Figure 3.2 and Table 3.1). The three datasets were used to investigate individual-based population structure. We first performed a PCA and then ran the sNMF algorithm, both implemented in the R package *LEA* (Frichot & François, 2015) on each dataset separately. The sNMF algorithm is a clustering algorithm allowing to find the most likely number of K ancestral populations best describing the genomic variability and to infer the individual admixture proportions under the selected model. The algorithm was run 10 times with values of K ranging from 1 to 6, and we chose the most likely model as the one associated with the lowest cross-entropy value. We finally built a final dataset to quantify genetic differentiation between sampling locations (FST, N=40) including only individuals from sampling locations with N≥5 (Figure 3.2, Table 3.1). We computed Hudson's estimator of pairwise-F_{ST} (Hudson, 1983) between each location using a custom R script and evaluated significancy by randomly permuting individuals 1000 times for each comparison.

3.3.5.5. Genetic diversity

Genetic diversity datasets were filtered using custom bash and python scripts, keeping only biallelic sites with no missing data and removing indels and SNPs heterozygous in more than 80% individuals. We built one dataset per each sampling location of $N \geq 5$ from which we computed the folded site frequency spectrum (SFS) using a custom python script. Using custom R scripts, we then computed Tajima's D (Tajima, 1989), and two estimators of θ, namely the mean pairwise difference θ_p and Watterson's θ_s, both standardized by the total number of called sites (i.e., monomorphic sites included). We investigated the influence of binning the dataset on the reconstructed Site Frequency Spectrum and genetic diversity estimates by sampling regions of 100 bp (to account for monomorphic sites) apart from 1kb, 10kb, 50kb or 100kb in GoM (the site with the more samples). All statistics remained similar (see supplementary results, Supp. Figure 3.12) and since accuracy in demographic inferences is improved by having mode data (Felsenstein, 2006), we decided not to bin datasets for historical demographic reconstructions (see below).

3.3.5.6. Detection of a supergene

In the light of the high degree of genetic differentiation between NEA and NWA but low within NWA region (see results), we performed genomic scans at the NWA scale (including GoM and CAN sampling locations) in order to find regions putatively related to the size polymorphism. We first scanned the genome by using an approach coupling PCA and Tajima's D which do not require any prior information on phenotypes. PCA can detect genomic regions of more than average population structure without any a priori individuals' grouping, being particularly useful when looking for the association with a complex trait which would need a large sample size for a robust characterization. For the PCA scan, we run the algorithm implemented by *Hierfstat* package (Goudet, 2005) on each chromosome in sliding windows of 100kb with a jump of 10kb on the NWA dataset. The proportion of variance explained by PC1 was extracted for windows with more than 50 SNPs and plotted against the location on the chromosomes. For TD scan, we built a dataset including both GoM and CAN individuals that was processed using the same filters applied for genetic diversity analyses (see above). We computed the folded SFS in sliding windows of 100kb with a jump of 50kb and computed Tajima's D from each window using a custom R script. These analyses revealed a region of high divergence on chromosome 2 (see results) in which we further computed linkage disequilibrium as r^2 correlation values between SNPs 50kb apart (to avoid computational burden) using *LDheatmap* R package (Shin et al., 2006). Local PCA, sNMF, and

the analysis of genotype distribution highlighted the occurrence of a bi-allelic supergene (HB and HS) in which all the three possible genotypes (HB/HB, HB/HS, HS/HS) were present (Fig 3). Additionally, we computed sliding windows F_{ST} between the two clusters of homozygous at the supergene in the NWA dataset (see Fig 3) using windows of 10kb with a jump of 5kb. Finally, we ran the local PCA and sNMF in the supergene region using the ALL dataset.

3.3.5.7. Genotype screening and linear modeling

Preliminary assessment of the relationship between size and haplotypes suggested an effect of the supergene genotypes on size (see results). To directly test the relation between size and the supergene, we identified two regions: from 25075452 to 25075619 (167 bp) and from 41404405 to 41404539 (134 bp) with respectively five and four SNPs discriminating the two supergene alleles. Primers were designed on 500bp flanking our target regions on each side and used to amplify 501 individuals sampled from the whole range (Table 3.1). The regions with five and four discriminating SNPs were hereafter referred to as “Region 051” and “Region 034”, respectively. The primers designed for “Region 051” (051F: 5'- CGG CAG TTS ACC ATC TTA GA -3'; 051R: 5'- GCT TGT AAC CAC ACT GCT -3') are targeting a fragment of ~280bp in length. The primers designed for “Region 034” (034F: 5'- GTA TGG AGT ACC ACC TTG AAT G -3'; 034R: 5'- GGT TGA TGT ATC TGC TGT AAG -3') are targeting a fragment of ~760bp in length. PCR reactions were carried out in 25 µL tubes by adding 14.775 µL of PCR grade water, 2.5 µL of PCR buffer, 2.0 µL of MgCL2 (25 mM), 2.0 µL of dNTP mix (2.5 mM each), 0.8 µL of each primer (10 µM), 0.125 µL of GoTaq® Hot Start Polymerase (Promega, Madison, WI, USA; 5 U/µL) and 2 µL of DNA template. The reaction mix was denatured at 94 °C for 2 min, followed by 35 cycles of denaturation at 94°C for 30 sec, annealing at 52°C (Region 051) or 52°C (Region 034) for 30 sec and extension at 72°C for 60 sec. PCR products were sent off to Retrogen Inc. (San Diego, CA, USA) for purification and sequencing. Genotypes for the 9 discriminating SNPs were attributed by visually assessment of base sequencing peaks. Genotypes were attributed a NA value when the peak was ambiguous. Individuals with missing genotype or for which the supergene genotype could not be determined throughout the 9 SNPs were discarded. We then tested whether the genotypes at the supergene were at the Hardy Weinberg equilibrium using a Chi2 test using *HardyWeinberg* R package (Graffelman, 2015) in sampling locations where the supergene was polymorphic.

We tested for the association between size and haplotype by accounting for maturity stage and sex. We filtered out individuals with missing information for any of these traits, resulting in individuals from GoM only (N=243). We designed three linear models using the bayesian framework implemented in the R package *brms* (Bürkner, 2021). The richest model in parameters, model HaploMatSex, included Haplotype, Maturity and Sex traits as determining variables ("Size ~ Haplotype + Maturity + Sex"). The two other models were nested within HaploMatSex, removing the variable "Sex" for model HaploMat and both "Sex" and "Maturity" variables for model Haplo. Four MCMC runs, each of 10000 iterations with 1000 warmup samples and a thinning of 4 were performed for each model, using flat (non-informative) priors. We assessed which model was the most accurate by performing the approximate leave-one-out (LOO) cross validation implemented in *brms*. Median values and posterior distributions of size under the best model were averaged on the levels of the other variables by using the package *emmeans* and summarized by 95% quantiles.

3.3.5.8. Ancestral range distribution

Range expansions (RE) occur by series of founding events leading to the fixation of derived alleles along the colonization process (Peter & Slatkin, 2013). Areas located further away from the origin of RE are therefore expected to display stronger linkage disequilibrium and higher frequency of fixed derived alleles, which are patterns that can afterwards be used to infer the colonization dynamics of a species. To investigate this, we followed two approaches. First, we investigated Runs of Homozygosity (ROH) signatures using the HMM model implemented in *bcftools-ROH* (Narasimhan et al., 2016). The analysis was run for each sampling locations with N≥5 separately by specifying to *bcftools-ROH* to estimate allele frequencies from genotypes. ROH were classified into three arbitrarily chosen length categories of 10kb to investigate changes in signals related to the length of ROH: ROH shorter than 10kb, ROH of length between 10kb and 20kb, and ROH larger than 20kb. We then plotted both the number (N_{ROH}) and the sum of ROH (SUM_{ROH}) for each class and for each sampling location. In a second time, we investigated the signatures of shared derived alleles across the whole range. However, such analysis requires polarizing the allelic state. To that end, we sequenced one individual of a closely related species Amblyraja hyperborea using the PacBio Sequel IIe System at the NGS Core of UF ICBR. A total of two SMRT cell runs (each generates 3-5 million reads) have been performed. Hifi long reads were then mapped using *pbmm2* v.1.3.0 align subcommand (https://github.com/PacificBiosciences/pbmm2) with the HIFI preset. Mapped reads were then sorted and indexed using *samtools*. Variants were

called using *deepvariant* v.1.4.0 (Poplin et al., 2018) by applying PacBio model and by specifying to output both polymorphic and monomorphic sites. Using *bcftools*, we then merged the long-read VCF to a dataset including all the short-read *A. radiata* individuals previously processed for genetic diversity analyses. We then filtered out non-bi-allelic sites and polarized the remaining variable sites based on *A. hyperborea* individual state (i.e., the outgroup individual, hereafter referred to as OG) by: (1) discarding sites heterozygous in OG; (2) recoding all allele(s) as ANC (ancestral, coded as "0") when corresponding to the allele for which OG is homozygote and as DER (derived, coded as "1") otherwise. The derived allele frequency was calculated per individual and the average value for each sampling location was reported. Based on the number of derived alleles per individual and per site, we calculated the directionality index (ψ) (Peter & Slatkin, 2013, 2015), which is the pairwise difference between shared derived alleles and is expected to be different from 0 if there is a signature of range expansion. The TDOA location algorithm of (Peter & Slatkin, 2013, 2015) was run on the pairwise ψ matrix to identify the RE origin. We ran the algorithm 100 times using one random individual from each location and displayed results as a density of the location of RE.

3.3.5.9. Historical demography

We first investigated the variation of the coalescence rate through time by using the *Pairwise Sequentially Markovian Coalescent (PSMC)* model on each individual. We followed the recommended bioinformatic pipeline: we called SNPs from each bam file using *bcftools* to obtain one VCF per individual. Each VCF was masked for the repeats using *bedtools* (Quinlan & Hall, 2010). Using scripts provided with the PSMC (H. Li & Durbin, 2011), we filtered for the depth of coverage using the parameters -d 6 and -D 50 and extracted the consensus sequence that was fed to the PSMC algorithm using the following parameters: -t15 -N25 -r5 -p "6+30*2+4+6". Because PSMC curves were highly similar in each cluster (see results), we computed 100 bootstraps for one individual in each cluster (i.e., one for NEA and one for NWA).

Devising a set of meaningful historical demographic scenarios to be quantitatively tested is traditionally challenging and sometimes arbitrary. Here we investigated demographic scenarios (Figure 3.3 and Supp. Figure 3.11) by using the composite likelihood approach of *fastsimcoal* 2.7 (Excoffier et al., 2021) to further investigate migration and divergence patterns at the scale of the range distribution. Model IMM-5 depicts an Isolation-Migration Metapopulation scenario with 5 demes connected in a one-dimensional stepping-stone fashion (i.e., migrants are only exchanged

with direct neighbors). The five demes refer to the sampling locations with N ≥ 5 (Table 3.1) and are spread into two meta-populations corresponding to NWA (GoM and CAN) and NEA (W-ICE, E-ICE and NOR) genetic clusters (see results). In NWA, the two demes have a size of N_W and exchange Nm_W migrants per generation with each other. Similarly, demes have a size of N_E in NEA and exchange Nm_E migrants per generation with the closest neighbor. The two regions are connected by an asymmetrical exchange of migrant of $Nm_{W\rightarrow E}$ from CAN to W-ICE and $Nm_{E\rightarrow W}$ from W-ICE to CAN (Supp. Figure 3.11). Going backwards in time, all demes merge into an ancestral population of size N_{ANC} at T_{DIV} generations ago. Two scenarios based on IMM-5 topology were additionally tested. Going back in time, the IMM-5-NM-CH model describes a change in connectivity between NWA and NEA happening T_{CH} generations ago, going from $Nm_{W\rightarrow E\text{-}MOD}$ and $Nm_{E\rightarrow W\text{-}MOD}$ from the present to T_{CH} to $Nm_{W\rightarrow E\text{-}ANC}$ to $Nm_{E\rightarrow W\text{-}ANC}$ from T_{CH} to T_{DIV}. IMM-5-NM-STOP is similar to IMM-5-NM-CH but NWA and NEA are isolated from the present to T_{CH}, with the two regions being then connected by $Nm_{W\rightarrow E}$ and $Nm_{E\rightarrow W}$ from T_{CH} to T_{DIV}. Additional scenarios were tested to investigate whether adding unsampled demes better depicted the genetic variability due to meta-population structure because the five sampled demes do not cover the whole range distribution of the species. IMM-20 is similar to IMM-5 but includes unsampled demes so that each of the two regions are composed of D=10 demes, resulting in a 20-demes 1D-stepping-stone matrix. From west to east, GoM and CAN are respectively sampled at demes 2 and 8 in NWA and W-ICE, E-ICE and NOR at demes 4, 5 and 8 in NEA, and the asymmetrical gene flow from NWA to NEA occurs from demes 10 and 11 and vice-versa. IMM-30 is similar to IMM-20, but NEA is composed of D=20 demes, with W-ICE, E-ICE and N-NW respectively sampled at deme 8, 10 and 17. This model was investigated to account for the likely different number of demes in each meta-population given the larger geographical area covered by cluster NEA (i.e., from Greenland to Norway, see results). fastsimcoal algorithm is based on the modelling of a set of two-dimensional SFS (2D-SFS) between sampled locations. To that end, we built a dataset with all individuals from CAN, E-ICE and W-ICE and a random subset of 5 individuals from N-NW and GoM to get a balanced sampling scheme. Using a custom R script, we processed the dataset using the same filters than for genetic diversity datasets and computed the set of 2D-SFS between each sampling location. The set of observed SFS were maximized using 100000 coalescent simulations (-n 100000), 40 expectation-maximization cycles (-L 40) and by considering at least 10 entry counts in the SFS to perform parameter estimation (-C 10). We

performed 10 independent runs for each scenario and used the best run (i.e., with the highest likelihood) to compute the AIC in order to perform model selection. Moreover, we computed the likelihood distribution for each scenario by simulating 100 replicates under the previously estimated best set of parameters: this procedure is necessary to determine whether the different scenarios are distinguishable or not. We then computed a confidence interval for parameter values for the model with the lowest AIC. To that end, we calculated 100 non-parametric bootstrapped 2D-SFS by randomly sampling blocks of 10,000 bp with replacement using a custom R script. Each set of bootstrapped 2D-SFS were maximized following the same procedure applied to the observed set of 2D-SFS. The 95% confidence intervals were calculated from the distribution of the best ML estimates for each bootstrap set. All historical demography inferences were performed using a mutation rate μ=2.01e-8 per site and per generation following a generation time of 11 years (average time at maturity, COSEPAC, 2012) and the genomic mutation rate estimated for the chondrichthyan species *Carcharhinus melanopterus* (Lesturgie, Planes, et al., 2022).

3.3.5.10. Origin of the supergene

We ran the PSMC algorithm within the supergene region for one HB/HB, one HS/HS and one HS/HB individuals that were randomly selected to detect when the divergence between the two haplotypes occurred. The PSMC estimates the distribution of coalescence times along the genome between two chromosomes: in a non-recombining block such as our supergene, this amounts to compute the divergence between the two alleles, which graphically corresponds to the time when the Ne suddenly increase to infinite in heterozygotes individuals (Cahill et al., 2016). To investigate whether the PSMC run in the supergene region reproduced the signal in the whole genome datasets, we randomly sampled 100 regions of 31Mb (the size of the supergene) spread in the genome on which we run the PSMC. We used the dataset polarized with *A. hyperborea* to perform sliding windows analyses of the ancestral allele frequency for HB/HB and HS/HS NWA individuals in chromosome 2. We then computed an UPGMA tree with Nei's distance using *poppr* R package (Kamvar et al., 2015) within the chromosome 2 supergene region and in chromosome 1 to investigate discrepancies between the supergene and the genome-wide trees. This was performed on the merging the ALL dataset with *A. hyperborea* individual.

3.3.6. Supplementary information

3.3.6.1. Acknowledgements

We are grateful to the genotoul bioinformatics platform Toulouse Occitanie (Bioinfo Genotoul, https://doi.org/10.15454/1.5572369328961167E12), to the HPC resources of the SACADO MeSU platform at Sorbonne Université (https://sacado.sorbonne-universite.fr/mesu/), the University of Florida Research Computing (http://www.rc.ufl.edu) and the Plateforme de Calcul Intensif et Algorithmique (PCIA), Muséum national d'histoire naturelle, Centre national de la recherche scientifique (http://uar2700.mnhn.fr/fr/pcia-9024) for providing computational and storage resources that have contributed to the research results reported in this publication.

3.3.6.2. Supplementary results

Summary statistics

Depth of coverage was on average ~17x. After filtering and binning, we performed population structure analyses using ~1.15 to ~1.19 millions of SNPs. Genetic diversity estimates were computed from the Site Frequency Spectrum (SFS) in sampling locations with $N \geq 5$, based on ~9.98 to ~13.93 millions of SNPs after filtering (Table 3.1). We investigated the effect of binning on genetic diversity estimates and the shape of the SFS by computing the normalized SFS as in 1. To that end, we sampled genomic regions of 100 bp (to account for monomorphic sites) apart from 1kb, 10kb, 50kb or 100kb in GoM (the site with the largest sample size). Genetic diversity estimates as well as the shape of the SFS were similar at different levels of binning (Supp. Figure 3.13), thus we decided not to bin data to keep as much information as possible, as accuracy in demographic inferences is influenced by the number of SNPs. Genetic diversity estimates (θ_π, θ_w) were highly similar in the whole range of the Thorny skate, with θ_π ranging from 0.0057 to 0. 0063 and θ_w from 0.0065 to 0.0079. Tajima's D ranged from -0.49 to -0.79 (Table 3.1).

Haplotype screening and linear modeling

501 individuals were screened by PCR and Sanger sequencing in two regions with ≥ 4 SNPs discriminating the two alleles (HB and HS) of the supergene. 31 individuals for which genotypes were ambiguous (i.e., not possible to determine genotypes at all discriminating SNPs) were discarded. Linear modelling was performed using a Bayesian framework implemented in the R library brms 2. We tested three models. *GenoMatSex*, the richest model in terms of variables included Maturity and Sex along Genotype as dependent variables: "Size ~ Genotype + Maturity + Sex". Two nested models were investigated: *GenoMat* ("Size ~ Genotype + Maturity") and Geno

("Size ~ Genotype"). Modeling was performed on a subset of individuals for which Maturity and Sex were available. To avoid bias due to population structure and maximize the sample size, we only retained N=243 individuals from GoM. After 10000 MCMC iterations with 10% burn-in and a thinning of 4, all models converged for all parameters in all four runs (Rhat=1) with effective sample sizes (ESS) ≥ 8000 for any estimate (with all four runs pooled). The Leave-One-Out cross validation displayed model *GenoMat* as the most accurate, even though the expected log pointwise predictive density (ELPD) for *GenoMatSex*, the richest model, was highly similar (difference of -0.3). The posterior predictive check was assessed by using 100 posterior draws from *HaploMat* model and suggested high adequacy between observed and predicted data.

Historical demographic modelling

Five demographic scenarios were investigated (Supp. Figure 3.11). A first set of three models was tested, investigating specifically patterns of migration and divergence between the two meta-populations: IMM-5, IMM-5-NM-STOP and IMM-5-NM-CH, the last being the richest in terms of number of parameters (Supp. Figure 3.11). We note that while IMM-5-NM-STOP is the model with the lowest AIC, its likelihood distribution computed for the set of ML parameters' value slightly overlaps with that of IMM-5-NM-CH (Supp. Figure 3.11). This would suggest that the two models cannot be statistically distinguished (Meier, Marques, et al., 2017). However, the ML values estimated under the two models support the same biological scenario. Going backward in time, the migration rate between the two meta-populations estimated under IMM-5-NM-CH is very close to 0 until T_{CH} (~141ky), suggesting non-significant exchange of migrants in this time frame which overlaps that of IMM-5-NM-STOP (where T_{CH} ~160ky). Similarly, migration rates sharply increase between T_{CH} and T_{DV} to values similar to those estimated under IMM-5-NM-STOP in a similar time range (Supp. Table 3.3). Finally, we note that the AIC values for IMM-5 was larger, suggesting that modelling a change in connectivity significantly improves our understanding of the demographic dynamics of the Thorny skate. Remarkably, all three scenarios were highly consistent in the estimates of the divergence time between the two metapopulations NEA and NWA and in the intra-region estimates of connectivity (Supp. Table 3.3). We further run a second set of scenarios including *ghost* demes in order to account for the unsampled demes in both metapopulations. Two scenarios were investigated based on IMM-5 topology: IMM-20, with two one-dimensional-matrices of 10 demes exchanging migrants in a Stepping-Stone fashion (one matrix per region) and IMM-30 in which NEA region was represented by D=20 demes and NWA

by D=10 demes. The two scenarios were strikingly least likely than the IMM-5-*like* models (Supp. Figure 3.11) suggesting that introducing *ghost* demes do not improve our understanding of *A. radiata* historical demography.

3.3.6.3. Supplementary figures

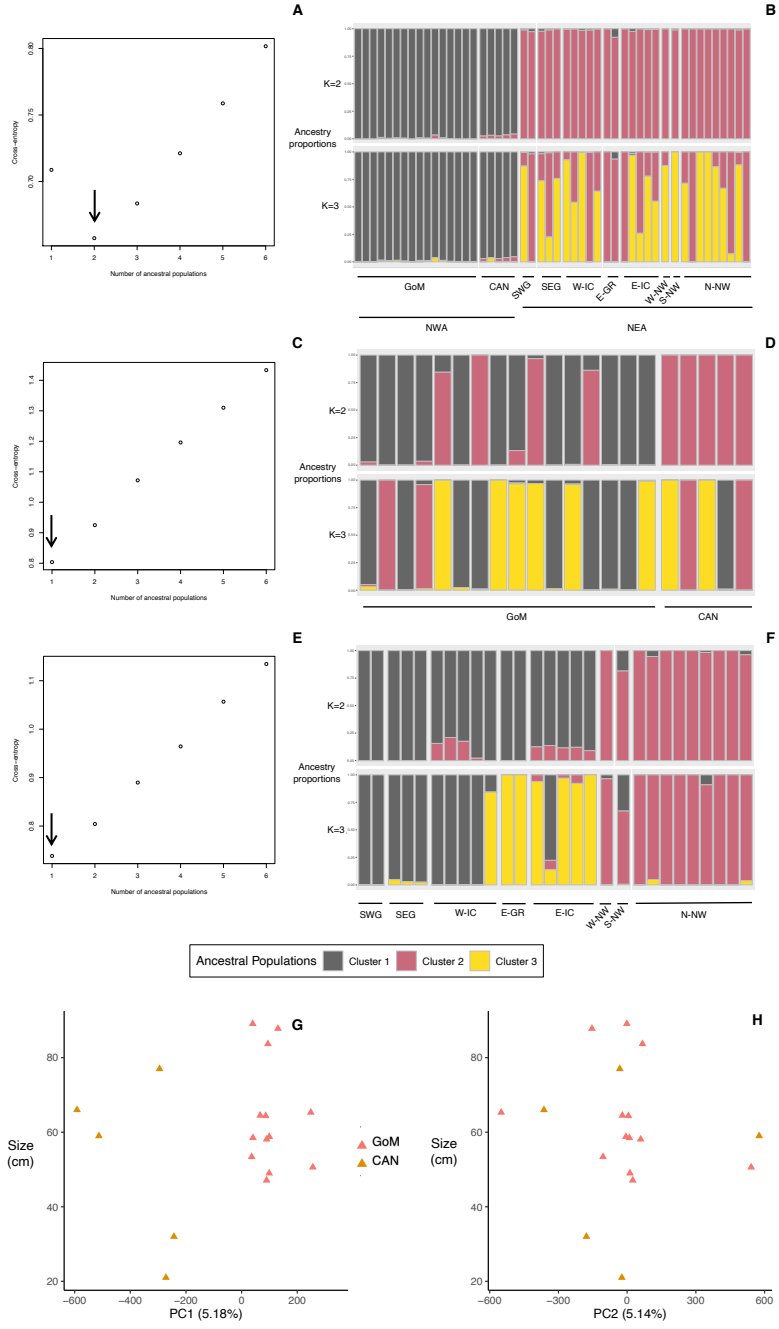

Figure 3.7. Panels A-F: Cross entropy criterion with an arrow indicating the most likely number of ancestral populations K when using all individuals (A), only GoM and CAN individuals (Cluster WEST,

C) or only SWG, SEG, W-IC, E-GR, E-IC, W-NW, S-NW and N-NW individuals (Cluster EAST, E) and corresponding admixture proportions for each individual estimated for K=2 and K=3 ancestral populations when using all individuals (B), WEST Cluster individuals (D) or EAST Cluster individuals (F). Panels G-H: distribution of Size (in cm) along the PC1 axis (G) and PC2 axis (H) within NWA.

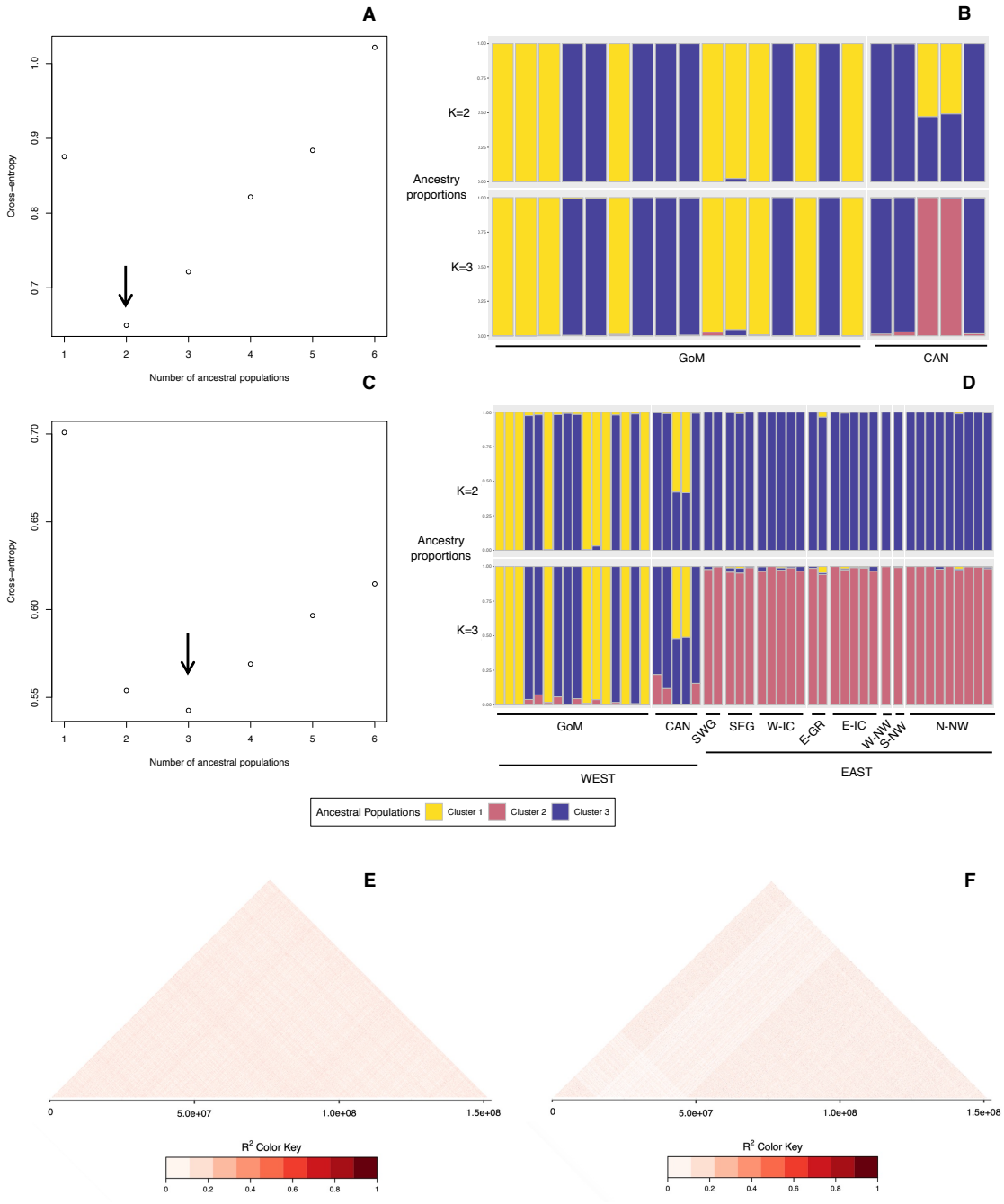

Figure 3.8. Genetic structure within the S2 inversion. Panel A: Local PCA including all individuals (both EAST and WEST) within the 17000000-48000000 region of SUPER 2 contig (S2 region). Dot shape represents the sampling location and color the attributed genotype for the inversion: HS/HS (blue), HB/HS (Red) and HB/HB (yellow). Panel F: Sliding windows of the average ancestral allele frequency in SUPER 2 for HB/HB (yellow) and HS/HS (blue) groups in GoM sampling location. Panels B & D: Cross

entropy criterion with an arrow indicating the most likely number of ancestral populations K when using only GoM and CAN individuals (panel B) or all individuals (panel D). Panels C & D: admixture proportions for each individual estimated for K=2 and K=3 ancestral populations when using only GoM and CAN (C) or all individuals. Panels E-F: Heatmaps of the pairwise linkage disequilibrium between SNPs for HS/HS individuals (E) or HB/HB individuals (F). Color gradient represent the value of the R^2 correlation between SNPs.

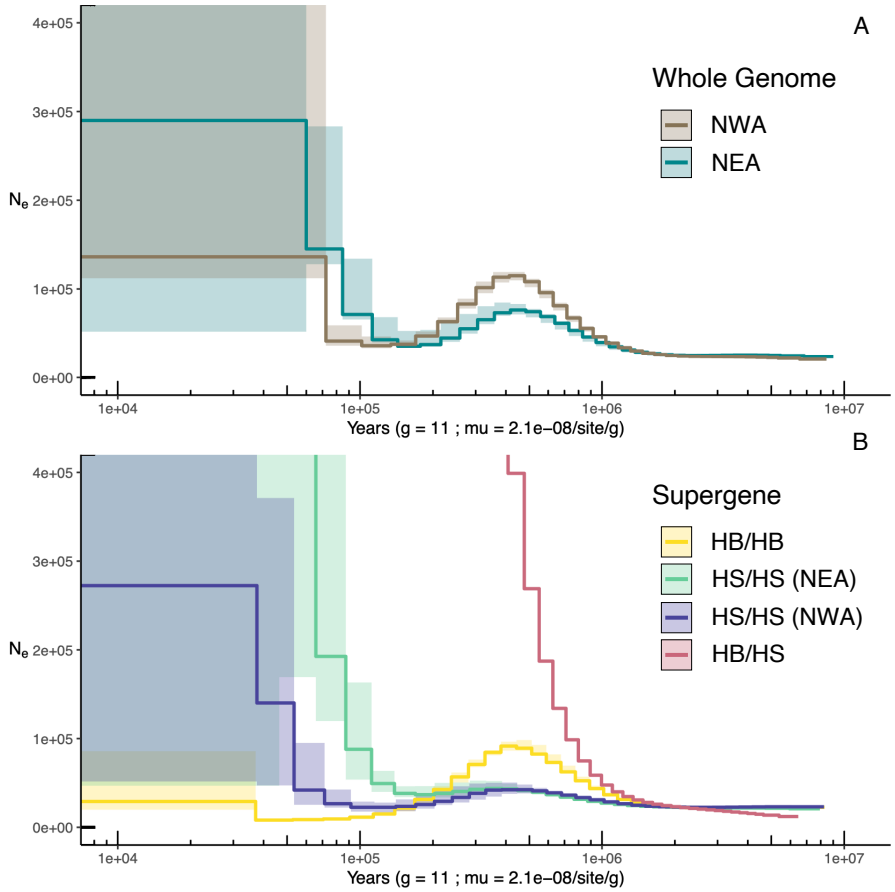

Figure 3.9. Variation of the coalescence rate through time as estimated by the PSMC algorithm. Panels A and B: inference on the whole genome with NWA (Brown) and NEA (Turquoise) individuals (A) and within the chromosome 2 supergene region (B) for HB/HB (Yellow), HS/HS for NEA (Green), HS/HS for NWA (Blue) and HB/HS (Yellow). The shaded areas represent the distribution of effective sizes (N_e) covered by the 49 individuals at each interval and the curve the median value of the distribution of N_e.

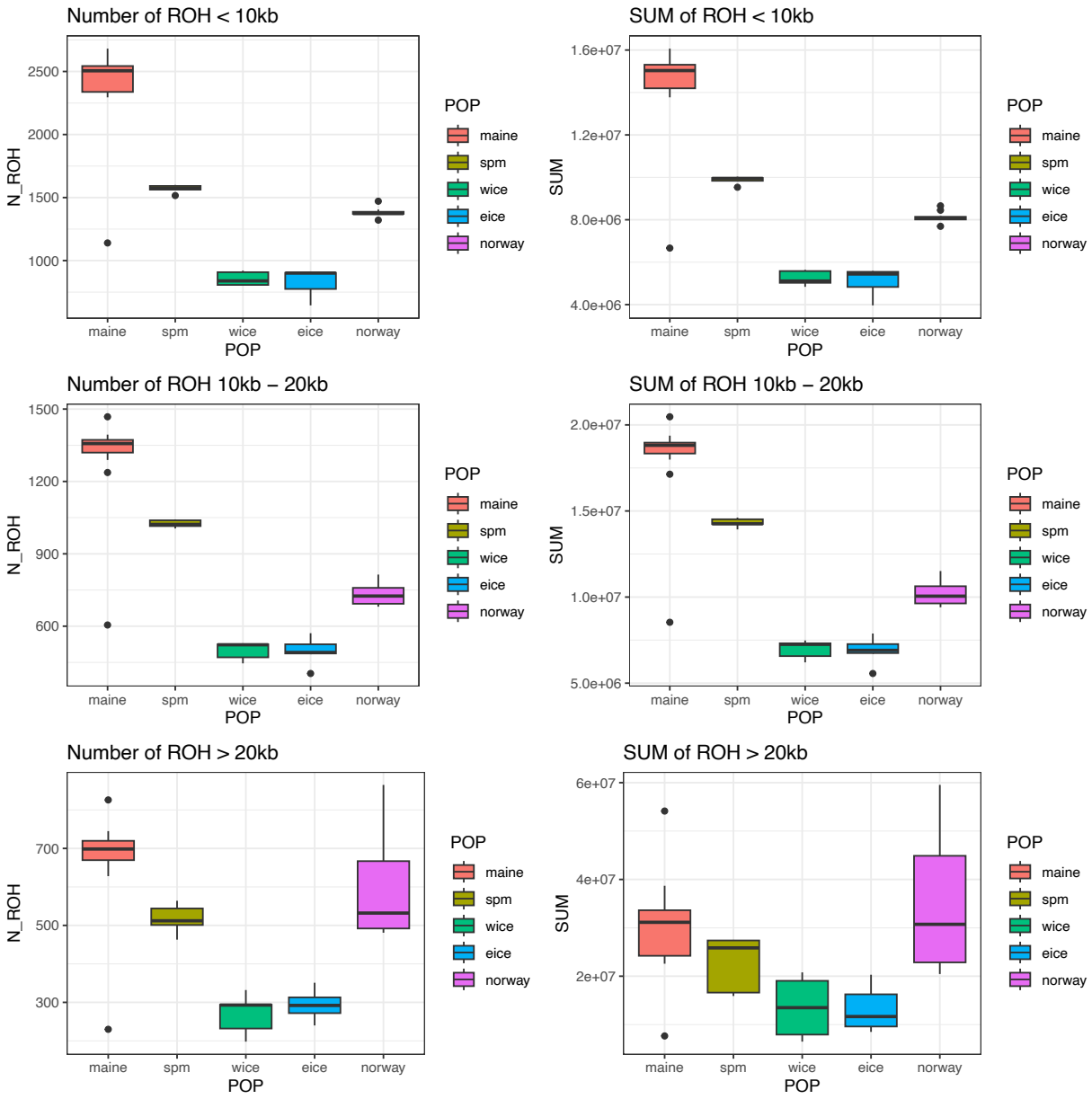

Figure 3.10. Distribution of Runs of Homozygosity (ROH) in sampling locations with N≥5. Number of ROH (panels A1-A3) and sum of the ROH (panels B1-B3) for different ROH size classes: below 10kb (A1 & B1), between 10kb and 20kb (A2 & B2) and over 20kb (A3 & B3).

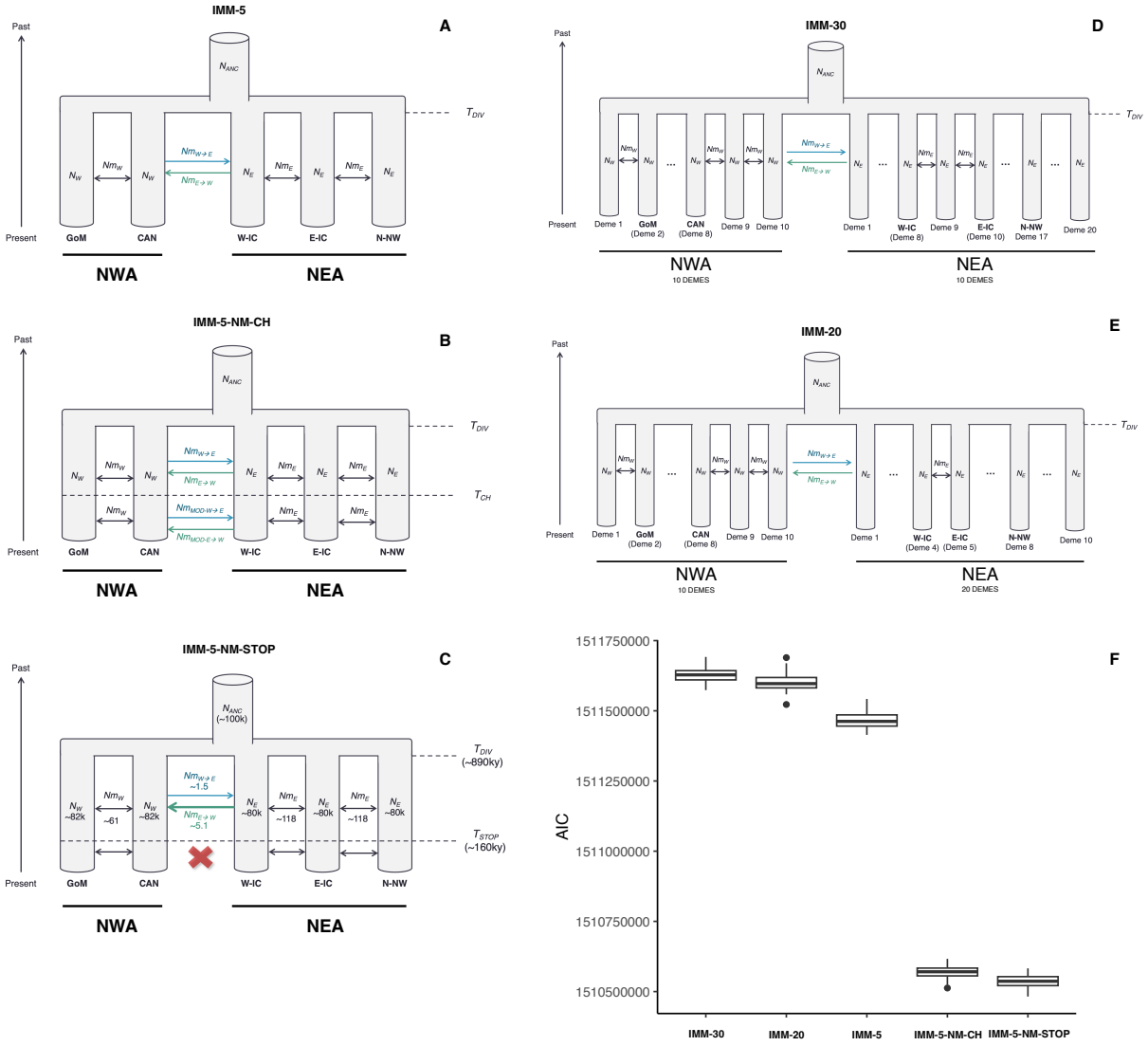

Figure 3.11. Demographic scenarios tested (panels A-E) and associated AIC values (Panel F).

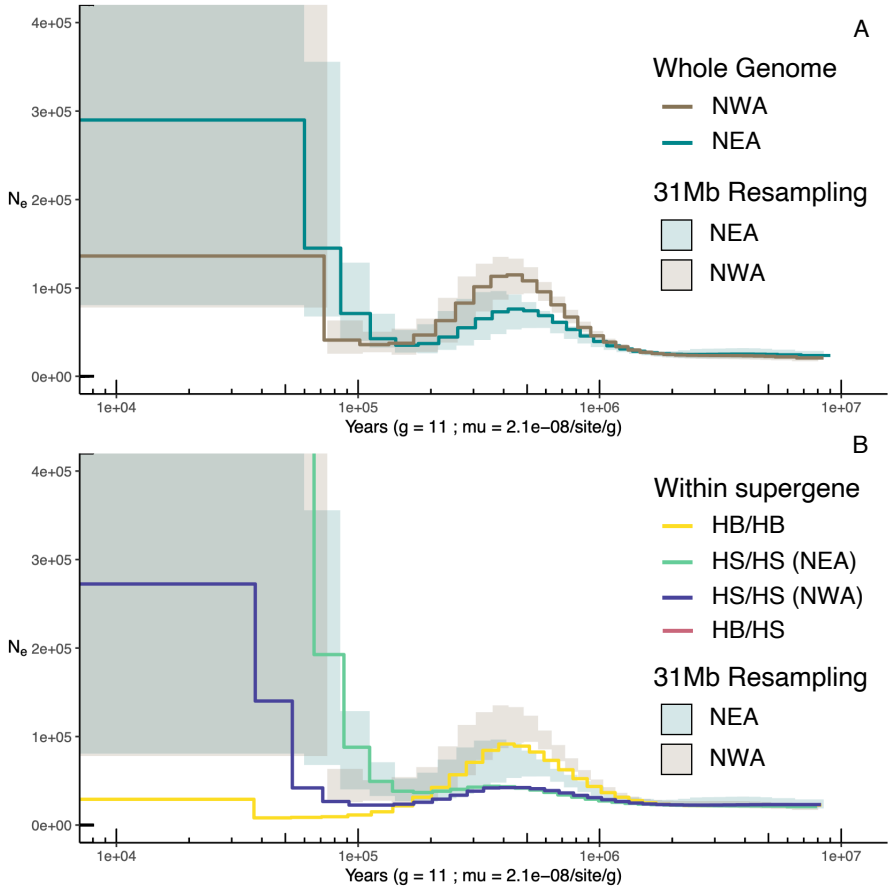

Figure 3.12. Variation of the coalescence rate through time in random resampled regions of 31Mb. Panels A and B: Median of the distributions of inferences (see Fig. S3) on the whole genome with NWA (Brown) and NEA (Turquoise) individuals (A) and within the chromosome 2 supergene region (B) for HB/HB (Yellow), HS/HS for NEA (Green), HS/HS for NWA (Blue) and HB/HS (Yellow). The shaded areas represent the 95% quantiles of the distribution of random resampling of 31Mb regions across the genome in an NEA (Turquoise) and NWA (Brown) individual.

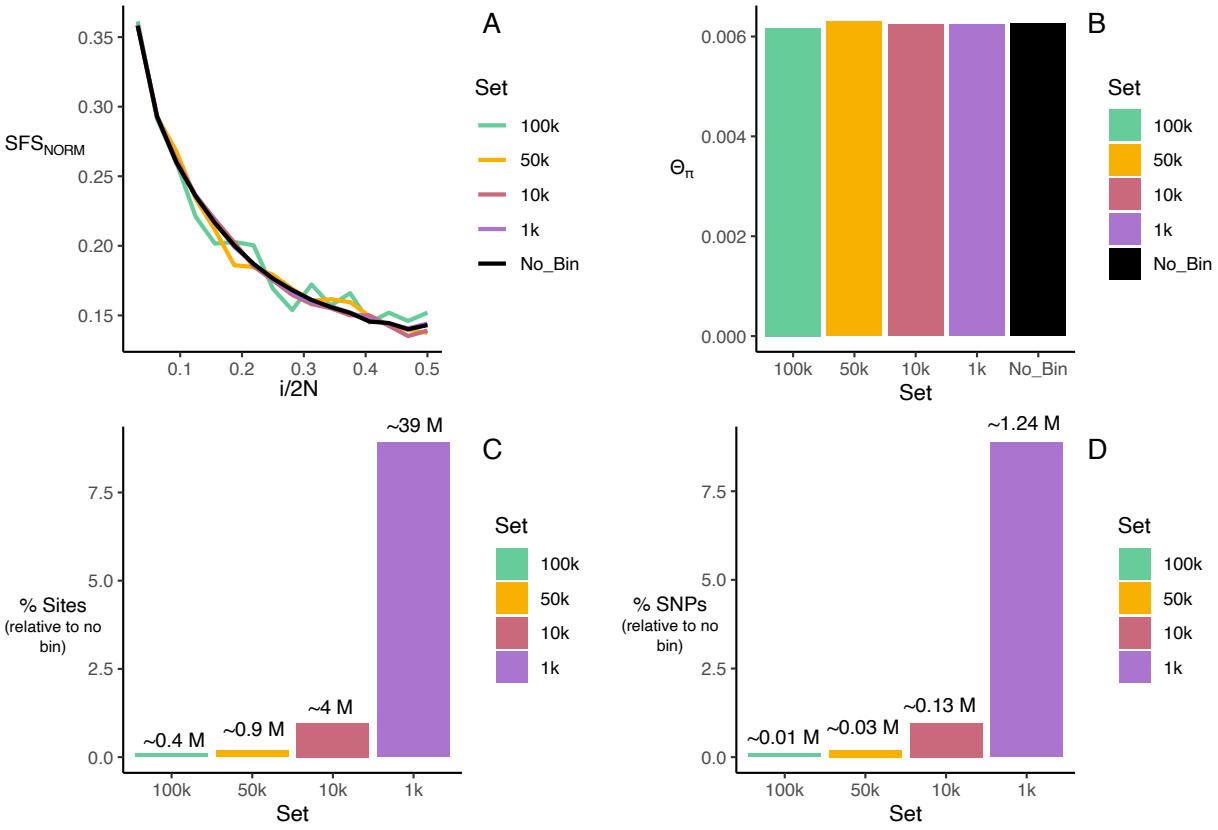

Figure 3.13. Influence of binning on summary statistics computed in GoM (N=16). Panel A: normalized SFS. Panel B: mean pairwise difference. Panels C and D: barplots of percentage of Sites (C) and SNPs (D) relative to the reference dataset (no binning) with the observed number of Sites and SNPs indicated above each bar. Each color represents a different level of binning: regions separated by 100kb (green), 50kb (orange), 10kb (red), 1kb (purple). Reference dataset (no binning) is presented in black.

3.3.6.4. Supplementary tables

Table 3.2. Median values and associated 95% credibility intervals averaged over the levels of other factors of size estimates (in cm) for the three linear models tested: Haplo ("Size ~ Genotype"), HaploMat ("Size ~ Genotype + Maturity") and HaploMatSex ("Size ~ Genotype + Maturity + Sex"). The Expected Log Pointwise Predictive Density resulting from the Leave-One-Out cross-validation step is indicated for each model.

Parameter	Level	Haplo	HaploMat	HaploMatSex
Genotype	HB/HB	60.31 [57.93; 62.73]	66.95 [64.81; 69.28]	67.15 [64.99; 69.49]
	HB/HS	57.23 [51.34; 63.04]	65.64 [60.70; 70.55]	65.43 [60.74; 70.47]
	HS/HS	47.59 [46.10; 49.98]	50.69 [49.28; 52.10]	50.81 [49.35; 52.18]
Maturity	1	-	51.54 [49.78; 53.40]	51.49 [49.56; 53.25]
	2	-	58.9 [55.67; 62.16]	58.75 [55.47; 61.95]
	3	-	62.76 [60.34; 65.31]	62.61 [60.15; 65.15]
	4	-	71.19 [67.00; 75.36]	71.67 [67.41; 75.83]
Sex	Female	-	-	61.69 [59.33; 63.89]
	Male	-	-	60.56 [58.26; 62.75]
ELPD[1]		-1454.96	-1397.70	-1397.95

[1]Expected Log Pointwise predictive Density (ELPD) by the Leave-One-Out (LOO) cross-validation step.

Table 3.3. Maximum Likelihood (ML) value of parameters estimated for the five demographic models and 95% confidence interval associated for the IMM-5-CH-STOP. AIC value of the best run under each model is reported.

	IMM-20	IMM-30	IMM-5	IMM-5-NM-CH	IMM-5-NM-STOP
N_{ANC}	103270	104630	99492	100752	101095 (99116-106634)
N_{EAST}	16316	16587	79910	77862	79941 (74595-81106)
N_{WEST}	24996	12459	82112	78452	81621 (78482-90962)
Nm_{E-W}	2.77	2.65	2.35	4.69	5.09 (4.88-13.73)
Nm_{W-E}	0.78	0.88	0.86	1.97	1.54 (1.69-5.56)
$Nm_{E-W-MOD}$	-	-	-	0.5	-
$Nm_{W-E-MOD}$	-	-	-	0.005	-
Nm_{W}	121.92	161.05	44.86	65.31	60.64 (49.67-124.05)
Nm_{E}	165.29	251.83	101.49	120.07	117.69 (76.74-144.91)
T_{CH}[1]	-	-	-	141174	160677 (46871-163108)
T_{DIV}[1]	872344	861399	928092	931645	890714 (799964-919941)
AIC	1511534769	1511569382	1511382039	1510516765	1510483219

[1]Times are expressed in year (converted using a generation time of 11 years).

3.4. Conclusions and perspectives

In this chapter, I bring light to the presence of a supergene in the genome of the Thorny Skate and I show its statistical association with size. Furthermore, by using extensive coalescent modelling, I show that i) this supergene was introgressed from a yet unidentified skate species; ii) it drives the recent demography of the Thorny Skate, being also responsible of the non-recovery of a population in the Gulf of Maine. This is the first time a supergene is found in a chondrichthyan species, probably due to the fact that these taxa are unrepresented in genomic studies. Beyond this, the chapter brings multiple perspectives for the study of supergenes and the related selective processes, notably because:

(1) A single (though) very large genomic region is significantly involved in the determinism of a known polygenic and environmentally modulated trait;
(2) Historical demography modelling was shown to be mandatory for extracting information about the supergene's maintenance and origin;
(3) The chapter emphasizes how using multi-species population genetics inferences will be of paramount importance to better characterize its origin in the future.

3.4.1. Size: A Polygenic Trait "Discretized" by a Supergene

I this chapter, I provided evidences that the size polymorphism observed in the Northwest Atlantic part of the range distribution (Mcphie & Campana, 2009; Sosebee, 2004; Sulikowski et al., 2005; E. G. Templeman, 1984; W. Templeman, 1987) was be determined by a large supergene. For the first time (to my knowledge) I directly bring to light a supergene system involved in the determinism of a known continuous trait with polygenic determinism (Bouwman et al., 2018; Boyle et al., 2017; Wood et al., 2014) as well as likely affected by environmental conditions. Polymorphism at the supergene results in two distributions of size, with individuals having at least on HB allele being significantly larger than individuals homozygous for HS. It is likely that the supergene determines other phenotype(s) than size as they have been shown to determine multiple phenotypes before (Errbii et al., 2023; Jones et al., 2012). The "two distributions" of size fashion strongly suggest that several genes involved in such determinism might span the region. Therefore, this supergene, with its considerable length (~31Mb) and number of genes (~226), offers an unprecedented opportunity to dissect the genetic basis of size variation. Beyond the Thorny Skate

and its conservation, this will therefore have broader implications in quantitative and population genetics as well as in fields related to functional genetics.

3.4.2. Historical Demography Inferences to Understand a Supergene's Evolution

In this chapter, I detected that the supergene was only polymorphic in northwest Atlantic (NWA). Moreover, the Gulf of Maine (GoM) sampling location displayed a significant deficit in heterozygotes that was not observed in Canada (CAN). This finding is likely explained by *positive assortative mating* occurring in GoM, which follows previous suggestions of mating incompatibility between large and small skates in GoM (Denton et al., *in prep*; Lynghammar et al., 2016). The difference with CAN could result from different maturity properties (Sosebee, 2004; Sulikowski et al., 2005; W. Templeman, 1987), which is known to covary with environment (Martin & Leberg, 2011). This positive assortative mating situation could be at the origin of the non-recovery of the GoM population, where two groups of individuals would still compete for the same resources but would not be able to form a single gene pool. At the same time, as described in the introduction of this chapter, positive assortative mating is known to lead to sympatric speciation (Straw, 1955). Yet, there is no genomic divergence – except within the supergene region – between large and small individuals in GoM, thus raising the question of why sympatric speciation is not happening. Here, I demonstrate how reconstructing the history of the species using an unprecedent amount of marker has increased our understanding on both how to maintain this polymorphism despite positive assortative mating system in a population and why one allele of the supergene is totally absent from one part of the range distribution (i.e., the northeast Atlantic, NEA). As introduced in **Chapter 1**, and further emphasized in **Chapter 2**, I first followed the crucial *diagnosis* step based on descriptive methods, required to device a meaningful set of demographic scenarios. For instance, I investigated population structure descriptive analyses and range expansion inferences prior to design a set of scenarios depicting the colonization history of the species and connectivity patterns within and between the NEA and NWA regions (hereafter referred to as meta-populations). In result, I was able to estimate a very high connectivity within each meta-population which could explain why the process of sympatric speciation is impeded in the GoM: while mating is impossible between large and small GoM individuals, it is possible in CAN where genomes of small skates homogenize with large ones. In result, a high gene flow from

CAN (or other northern demes) provides a mean for small and large individuals to maintain a common gene pool in GoM (except in the supergene region) despite being unable to mate. This has considerable implications, as this means that it is only for the high migration rate that a speciation process might be hampered. Additionally, I was able to infer that the NWA was colonized a long time ago (~900,000 years ago) and has remained isolated from the NEA for the past ~160,000 years: this allowed me to make the parsimonious hypothesis that the supergene is absent from the NEA because it introgressed the Thorny Skate after the migration between the two regions stopped. These two interpretations, on the origin and maintenance of the supergene, are key to understand better the evolutionary history of the Thorny Skate and of the supergene, and would not have been possible without extensive demographic modelling, further emphasizing its importance for comprehending evolutionary processes.

3.4.3. Towards a Multi-Species Framework to Date Supergenes Origins

The two results developed above (i.e., migration hampers sympatric speciation and the origin of supergene ≤ 160,000 years ago) are key to understanding the peculiar nature of this supergene. However, the time window during which the supergene originated remains very large (≤160,000 years!), which raises the question: can we date the time when the supergene either formed or introgressed in this species? Additional sequencing of a closely related species (*Amblyraja hyperborea*) provided evidence that one allele, HB, has resulted from an introgression of a donor species – not necessarily *A. hyperborea*. Dating intogressed regions is unfortunately troublesome as it would require first to identify the donor species, possibly by comparing the supergene-tree with a species-tree as it has been done before (Helleu et al., 2022; Stolle et al., 2022; Yan et al., 2020), thus requiring extensive sequencing of multi-species population samples. Not only this would allow to determine the donor, but also to gain more knowledge on how the supergene is distributed in multiple species, as supergenes have been shown to promote adaptation and selection across groups of taxa (Chouteau et al., 2017; Joron et al., 2006; Purcell et al., 2014; Stolle et al., 2022; Z. Yan et al., 2020). Once the donor species is identified, time of origin could possibly be obtained by dating the divergence between HB_{DONOR} and $HB_{RADIATA}$. However, it is notable that the dating process of supergenes has not been thoroughly investigated so far. Dating methods usually assume neutrality while supergenes are subject to different selective pressures, hence complicating enormously historical demographic inferences. For instance, in the case of the thorny

skate, the IICR curves reconstructed by the PSMC within the supergene region clearly differed from those on the whole genome, strongly suggesting that the supergene is under selection, which likely changed in time. Dating the apparition of supergenes – introgressed or not – will be in the future a key question to understand precisely the origin of such systems, which might require theoretical developments. It is interesting to note that investigating these perspectives will be only be made possible in the future by building multi-species datasets eventually studied using population genetics tools. This highlights how multi-species frameworks will be key in the future to study highly specific and niche evolutionary processes, such as the origin of introgressed supergenes, but also large-scale processes as it will be highlighted in the following chapter.

Chapter 4. Genetic Signatures of Ecosystem Functioning

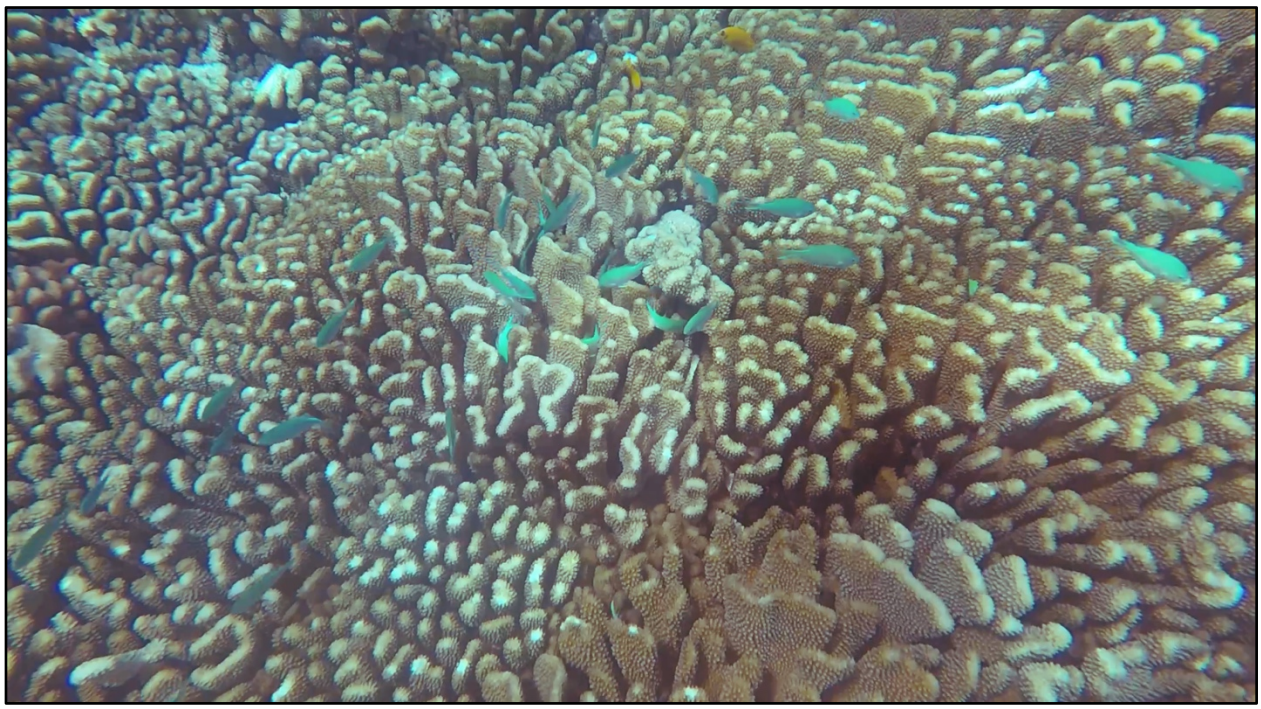

Figure 4.1. Reef from Fakarava, French Polynesia.

4.1. General context

4.1.1. Beyond the Species-centered Population Genetics Paradigm

Chapters 2 and **3** both treated a population genetics question from a species-centered point of view, yet they both conveyed a multi-species inferences perspective. In **Chapter 2**, I investigated how the ecology of a small group of species influences their genetic diversity and degree of population structure, further suggesting larger-scale studies to find ecological determinant of historical demography. In **Chapter 3**, I showed that species interaction, specifically an introgression, shaped the evolution of a supergene, and emphasized that further comprehension of this system will thus require multiple species investigations. This shows how multi-species inferences are clearly crucial to getting a further understanding of *traditional* population genetics issues. Beyond that, multi-species inferences can elucidate the impact of large-scale phenomena on the genome, from the determinism of genetic diversity by certain LHTs (Ellegren & Galtier, 2016), to the study of biogeographical processes (Overcast et al., 2023). In addition, transcending the traditional focus of empirical population genetics studies to the community level would facilitate the development of conservation strategies at the ecosystem level.

To date, few studies have considered, directly or indirectly, multi-species population genetics inferences (but see (Chan et al., 2014; Delrieu-Trottin et al., 2020; Hickerson & Meyer, 2008; Xue & Hickerson, 2015)). Building multi-species datasets for population genetics inference is a complex task. Genetic diversity needs to be directly compared between species, i.e., one has to target similar or comparable regions in the genome of a set of species. This means that RAD-sequencing is hardly usable in this respect. As detailed in the introductive chapter, Whole Genome Sequencing (WGS) is usually the most ideal tool for any genomic investigation. Yet it is likely to be extremely expensive if needed in several species for which reference genomes are also required. Reference genomes, despite the growing number of international consortium (such as the Vertebrate Genome Project and the European Reference Genome Atlas), are still available for a limited number of species. The lack of reference genomes is therefore an issue when investigating a community rather than a group of congeneric species. An interesting approach in this context is Target Gene Capture (C. Li et al., 2013), which specifically targets a set of homologous genes in all species investigated thus allowing unbiased comparison of genetic diversity. Such protocol however necessitates designing baits to capture previously established homologous regions for the set of species. In this chapter, I used such an approach to study coral reef fishes sampled off

Moorea, French Polynesia. This project aimed at increasing our understanding of how community-scale level impacted demographic history, which I introduce below.

4.1.2. Linking ecological theories to population genetics modelling

Species are organized in communities (i.e., sympatric populations of different species) and their evolutionary history should harbor signatures of inter-specific interactions as well as biogeographic features inherent to the community (Overcast et al., 2023). Species living in the same communities interact with each other and their environment, and the diversity of species composing it is a key factor as it increases the whole functioning (i.e., the overall energy fluxes) of the ecosystem (Tilman et al., 2014). Species occupy different ecological niche in the ecosystem, which can represent both the requirements for them to live in the habitat (i.e., environmental conditions, such as temperature or salinity), and their role in the community (i.e., the resources they consume and represent in a trophic network; Polechová & Storch, 2008; Sexton et al., 2017). Consequently, species interact with one another based on the ecological niche they occupy, engaging in either competition for similar resources or direct interactions, such as predator-prey relationships. However, species are not equal in terms of the *width* or *breadth* of the ecological niche they occupy, which refers to the diversity of conditions under which a species can strive or of resources that a species can consume (Sexton et al., 2017; Vandermeer, 1972). Typically, this niche breadth spectrum is best depicted by defining species in its two extremes: *generalist* species, which are capable of using a great variety of resources and have thence a wide ecological niche and *specialist* species which exploiting a single (or few) resource are thus associated with a narrow niche (Vandermeer, 1972). The concept of niche breadth was related to genetic variability early in history, as generalists were expected to display more morphological and thus genetic polymorphism than specialists following the *Niche Variation Hypothesis* (NVH) (Levene et al., 1966; Soule & Stewart, 1970; Van Valen, 1965). Consequently, it has been hypothesized that generalist species are more prone to adaptation in response to changing environments and thus less *vulnerable* compared to their specialist counterparts.

As a direct consequence of the NVH, interactions should leave signature in the genome of the species conditioned to the niche width: generalist species are expected to be more resilient in time than specialists as less prone to (local) extinction due to their ability to use a large number of resources and conditions and thus to acclimate (Carscadden et al., 2020; Colles et al., 2009). Direct

population genetics expectations have been conceptualized in the *Specialist-Generalist Variation Hypothesis* (SGVH) (S. Li et al., 2014): (1) specialists should display lower levels of genetic diversity because of higher stochastic fluctuation events in their history (following expectations of (Kimura & Crow, 1963)); and (2) population structure should be stronger for specialists because of a reduced gene flow due to the necessity to colonize habitats with the required (scarce) resource. To date, empirical evidence has both confirmed (S. Li et al., 2014; Matthee, 2020; Pasinelli, 2022) and discarded the SGVH, particularly in the marine realm (S. Li et al., 2014; Matthee et al., 2018; Titus & Daly, 2017). The latter might be explained by the fact that specialists could counter the expected vulnerability risk and thus demographic instability by targeting abundant species (Colles et al., 2009; Strona et al., 2013) and/or by being better adapted to their resources than generalists, hence leading to better resource assimilation (Colles et al., 2009). However, the debate remains open, since no study has (to my knowledge) directly and formally tested the relationship niche width – historical demography: studies always attempted to explain patterns of genetic diversity in light of "known" or hypothetical niche width, or direct diet observation and always using a small sample of species and genetic markers (most often, limited to mitochondrial DNA). Additionally, the relationship between demographic stability and genetic diversity might be complex and might benefit from the direct demographic modelling using the amount of data provided by genomic data. Indeed, formally testing this hypothesis would therefore require both high quality data from multiple species to model genetic diversity and an accurate characterization of the niche width for a set of species, which remains a technical and expensive challenge.

4.2. Objectives

In this chapter, I aim at directly investigating the relationship between historical demography and niche width, in the specific case of trophic interactions (i.e., trophic niche width). To that end, I coupled meta-barcoding of gut content data to Target Gene Capture data in ~40 species of coral reef fishes from Moorea, French Polynesia. I compute indices of genetic diversity as well as design and implement demographic stability indices notably computed from the reconstructed coalescence rate (or IICR) through time. I use reef fishes as test group as they display a large diversity of interactions and trophic guilds in coral reefs making this ecosystem a great model to test the SGVH hypothesis. Additionally, coral reefs possess the most important marine biodiversity (Tittensor et al., 2010): although covering less than 0.1% of the ocean surface (Spalding & Grenfell, 1997), they host nearly 25% of its global biodiversity (Allsopp et al., 2008). However, the coral cover decreased of 1 to 2% per year in the Indo-Pacific since 1970 (Bruno & Selig, 2007). This resulted in a decrease of species, functional and phylogenetic diversity (D'Agata et al., 2014), as well as genetic (Pini et al., 2011; Pinsky & Palumbi, 2014). Maintaining great diversity drive the resilience of these ecosystems: knowing the evolution of their state is therefore essential in conservation biology. This means that this chapter, additionally to conceptually testing for the first time a relation between niche width and genetic diversity, brings more knowledge about biodiversity and its resilience in the endangered coral reef ecosystems.

4.3. Larger trophic niche increases stability along evolutionary times

Article in preparation for *Nature Ecology & Evolution* (Brief Communication)

Authors:

Pierre Lesturgie, Maël Le Gouellec, Simon J. Brandl, Jordan M. Casey, Valeriano Parravicini & Stefano Mona

4.3.1. Abstract

Understanding the underpinnings of species vulnerability is key in the context of intensifying global changes. According to the *Specialist-Generalist Variation Hypothesis*, trophic niche width determines species success, with generalist species being more stable through time. We followed for the first time a population genetics perspective to test this hypothesis, estimating demographic stability by genomic sequencing 38 fish species and assessed their trophic niche width by gut-content meta-barcoding. Demographic stability was significantly positively associated with niche width, underscoring that generalists are less prone to local extinction, participating to the stability of the community. Our innovative framework will contribute deciphering ecosystem functioning.

4.3.2. Main

Understanding the underpinnings of species vulnerability is essential to grasping their response to rapid perturbations and, ultimately, to predicting the future resilience of biodiversity in a context of global crisis (Ceballos et al., 2015). Niche width, i.e., the extent of resources used and viable conditions for a species (Sexton et al., 2017; Vandermeer, 1972), has been questioned as a major determinant of vulnerability (Colles et al., 2009). Specialist species (i.e., with narrow niche width) should exhibit pronounced population fluctuations due to their exclusive dependence on the availability of a few resources or because of limited viable environmental conditions (Gravel et al., 2011). Generalists, on the other hand, are expected to exhibit greater morphological and genetic variation (Levene et al., 1966; Soule & Stewart, 1970; Van Valen, 1965) and should therefore be more prone to adaptation in response to changing environments (Carscadden et al., 2020; Colles et al., 2009). Accordingly, the *Specialist-Generalist Variation Hypothesis* (SGVH) predicts that specialists should display i) lower genetic diversity due to greater stochastic fluctuation in population size (Kimura & Crow, 1963); and ii) reduced gene flow due to more scattered distribution (Gravel et al., 2011; S. Li et al., 2014; Pasinelli, 2022) than generalist species. This conjecture was either corroborated (S. Li et al., 2014; Matthee, 2020; Pasinelli, 2022) and rejected (S. Li et al., 2014; Matthee et al., 2018; Titus & Daly, 2017). Empirical arguments suggest that rejection could be due to the ability of specialists to counterbalance challenges posed by narrow niche width by interacting with abundant species and by better assimilating resources than generalists (Colles et al., 2009; Strona et al., 2013). However, the SGVH has never been tested extensively to date. Notably, the relationship between niche width and historical demography as inferred by genetic data has never been quantitatively evaluated (i.e., with a large sample of species).

Here, we coupled a large nuclear genomic dataset to trophic niche width data assessed by meta-barcoding of gut contents in a coral reef fish fauna. Specifically, we selected 43 species of coral reef fish (540 individuals) sampled off Moorea, French Polynesia (Table S1). Nuclear genomic DNA was sequenced using a Target Gene Capture protocol (C. Li et al., 2013) that amplifies and sequence homologous loci across species. Based on hundreds to thousands of Single Nucleotide Polymorphisms (SNPs, Table S1) in each species, we computed genetic diversity and historical demographic indices. We assessed trophic niche width by calculating the number of Exact

Sequence Variants (ESV), species, genus and families detected in the gut content of the sampled species with COI and 18S markers from the meta-barcoding data of (Casey et al., *in prep*).

We performed Bayesian linear modelling to relate genetic indices and the number of resources consumed as estimated with COI or 18S marker. At the species level (Figure 1), both genetic diversity estimates (θ_π and θ_w) decreased with the number of consumed resources, although support ranged from weak (posterior probability of the slope $P = 0.75$ for θ_π for COI marker) to very strong ($P = 0.96$ for θ_w for 18S marker). Conversely, we found a robust and positive correlation between the N° of consumed resources and Tajima's *D* (*TD*, P=0.97 and P=0.99 for 18S and COI respectively): generalists tend to have TD closer to 0, which is suggestive of a constant historical demography (Tajima, 1989). To investigate more precisely this relationship, we devised four demographic stability indices based on (1) the distance of the observed *Site Frequency Spectrum* (SFS) to the expected one under a constant population scenario (d_{SFS}); (2) the distance of the genetic diversity trajectory inferred by the *stairwayplot* to the observed θ_π (d_{STAIR}); (3) the ratio of modern to ancestral θ reconstructed by the *stairwayplot* (R_{STAIR}); and (4) the absolute sum of slopes between time intervals of the trajectory inferred by the *stairwayplot* (f_{STAIR}). Strikingly, all demographic stability indices were robustly and positively correlated with the number of consumed resources (P ranging from 0.89 to 1) strongly suggesting that generalist species are more stable along historical times. To account for phylogenetic proximity, which may potentially bias our results, we inferred a phylogenetic tree (Figure S1) (see supplementary material for additional results) which was added as random effect in the Bayesian linear modelling. Results remained very similar, confirming the robustness of our analyses (Table S2, Figure S2). Finally, we considered the number of consumed resources at the ESV, genus and family levels: all the models confirmed the trend previously observed (despite lower support for d_{STAIR} and f_{STAIR} at the ESV scale), either including or not the phylogenetic tree (Table S2, Figures S3-S8). This, coupled to the similar signals obtained when considering COI or 18S at any scale of study, further suggest the robustness of our results both to the marker used and to the chosen taxonomic resolution.

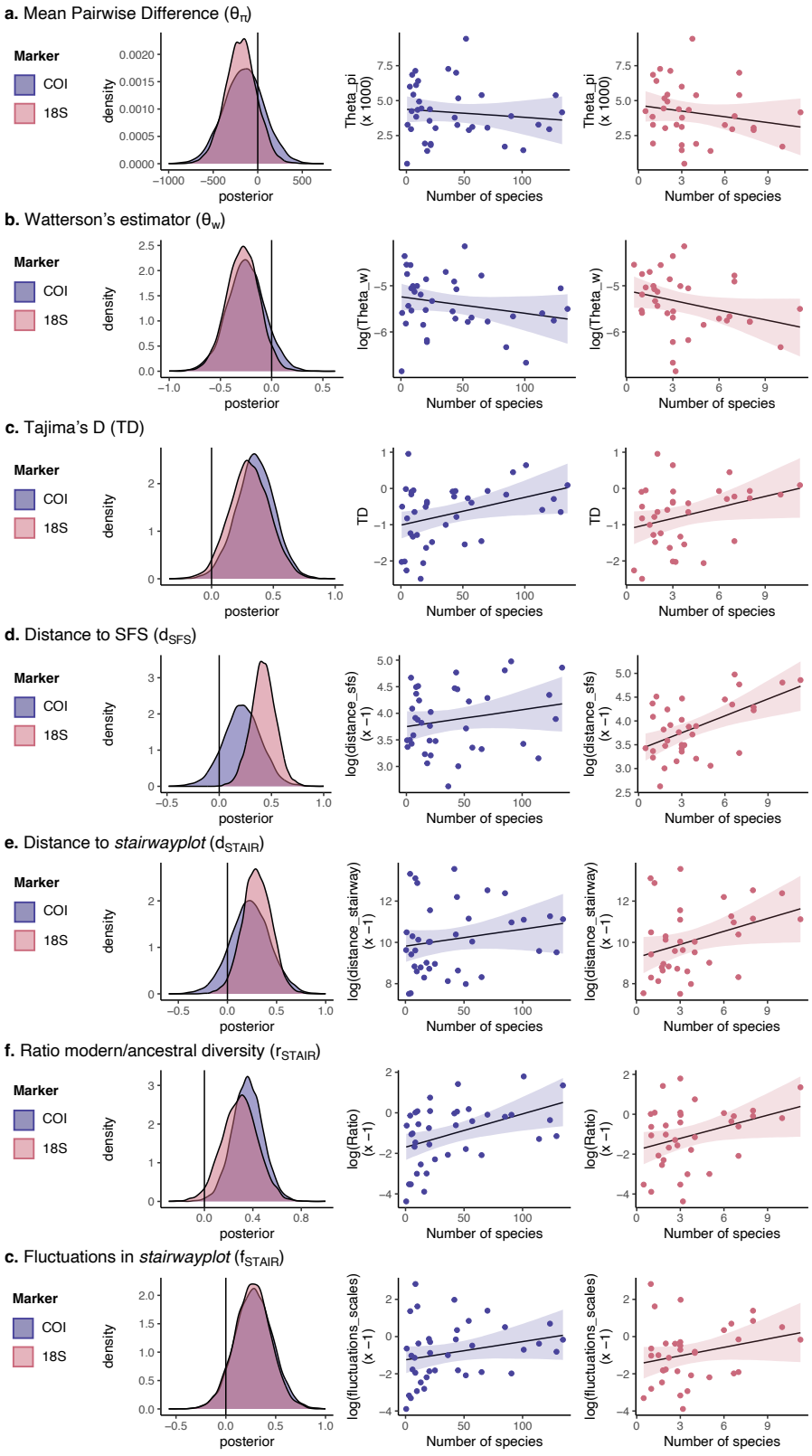

Figure 4.2. Bayesian Linear models relating genetic indices and the number of consumed species. Each row represents a genetic index. For each genetic index, the three panels represent: i) the posterior distribution of the slope when the N° of consumed species is estimated by the COI (purple) or 18S (red); ii) the regression line relating genetic indices and N° of consumed species estimated by the COI and iii) 18S marker. The posterior distributions of the slopes are scaled by their maximum absolute value for comparative purposes.

Our study unambiguously highlights a positive relationship between historical demography stability and trophic niche width, corroborating theoretical expectations of (Carscadden et al., 2020; Colles et al., 2009; Gravel et al., 2011; S. Li et al., 2014). Under the SGVH, genetic diversity is expected to be greater for generalists (S. Li et al., 2014) as a direct consequence of their lower demographic fluctuations (Kimura & Crow, 1963) than expected in specialists (Gravel et al., 2011). However, here, θ_π and θ_w hardly correlated with niche width (Figure 1). We argue that the relationship between genetic diversity and demographic fluctuations can be tricky to evaluate, as a burst in genetic diversity can be observed when two populations are reconnected after isolation (Alcala et al., 2013). This could be common in specialist species: rather than undergoing the expected extinction-recolonization cycle (Gravel et al., 2011; S. Li et al., 2014), specialists could undergo frequent strong local bottlenecks followed by re-connection, which should be further investigated in the future. All in all, d_{SFS} and stairwayplot-based indices are more directly related to demographic trends than θ estimates, and the fact that they correlate with niche width gives us confidence in our interpretation. More importantly, it prompts (1) caution when directly interpreting genetic diversity estimates in relation to ecological traits; and (2) the necessity of coupling diversity estimators with more detailed indicators of population demographic history, such as the SFS and SFS-based models (e.g., *stairwayplot*).

Our approach is novel regarding previous investigations – which conveyed or not similar conclusions (S. Li et al., 2014; Matthee, 2020; Matthee et al., 2018; Pasinelli, 2022; Titus & Daly, 2017) – as it is the first to our knowledge, to i) quantitatively test the relationship between historical demography and niche width; ii) integrate indices of demographic stability to classical estimates of genetic diversity; and to iii) use the strong accuracy provided by a multi-species genomic dataset coupled with precise meta-barcoding assessment of trophic niche width, thus going beyond previous qualitative assessment of niche width in the context of the SGVH (S. Li et al., 2014; Matthee, 2020; Matthee et al., 2018; Pasinelli, 2022; Titus & Daly, 2017). This dataset is a first example of the power of combining population genetics and functional ecology approaches to the study of a fish reef fauna. Such approach can be clearly extended to any ecosystem, providing a powerful tool to interpret and eventually forecast its stability. Our study might also be extended to additional sampling location in the Pacific to test whether the migration and colonization dynamics is related to the niche width (S. Li et al., 2014). In addition, it could be used to test the expectations of the *Trophic Theory of Island Biogeography* (Gravel et al., 2011) which predicts the variation of

spatial demographic stability, due to the extent of niche width, along a gradient of biodiversity. Coral reef fish fauna from the Pacific Ocean represent a perfect setting to test this hypothesis, as biodiversity decreases moving away from the Coral Triangle (Roberts et al., 2002).

Our study provides strong evidence of greater demographic stability related to the larger trophic niche width in a coral reef fish fauna, which directly shows for the first time that generalist species might be undergo fewer demographic fluctuations along evolutionary times. This suggests that generalists might be less prone to local extinctions and thus less vulnerable than specialists, providing a quantitative approach to define the stability of an ecosystem. We emphasize the power of coupling population genomics to ecological datasets to unravel the determinants of ecosystem functioning and of the vulnerability of species, which is crucial to better understand the challenges for biodiversity in the current global crisis (Ceballos et al., 2015).

4.3.3. Material and Methods

4.3.3.1. Sampling, Sequencing and De novo assembly

540 individuals belonging to 43 species were sampled off Moorea, French Polynesia. DNA was extracted from fin clips using QIAGEN Dneasy blood and tissue kit (QIAGEN, Germany) following the manufacturer's indications. We then followed a Target Gene Capture protocol (C. Li et al., 2013; H. Li et al., 2018). DNA was first PCR-amplified and libraries were prepared following a capture protocol realized using myBaits hybridization capture kit (Arbor Bioscience, USA), targeting a set of 4434 autosomal regions homologous in ray finned fishes (Jiang et al., 2019). Libraries were then re-amplified and sequenced using a paired end approach (150bp reads) on an Illumina NovaSeq 6000 sequencer.

The Target Gene Capture protocol not only allows to sequence the targeted regions (i.e., the exons) but also the flanking regions (i.e., introns). Therefore, to increase the number of Single Nucleotide Polymorphisms (SNPs) for downstream analyses, we developed a pipeline to perform a species-specific *de novo* reference assembly from the sequenced reads. After quality checking *fastq* files, reads were trimmed using TrimGalore-0.6.5 (M. Martin, 2011). We randomly selected two individuals per species to assemble reference contigs and to check for consistency a posteriori. Each assembly was realized using SPAdes-3.15.4 algorithm (Bankevich et al., 2012), using the reference baits as guide for the assembly (*--trusted-contigs* argument). The following parameters were used: *--careful* (limit the number of mismatches and short indels in the assemblies); *--cov-*

cutoff 10 (coverage > 10 for the assembled sites to ensure sufficient coverage for the assembled sites and thus increase the robustness of the assemblies). Each assembly was then aligned to the reference set of baits used for the capture experiment using *nucmer* function in the mummer-4.0.0beta2 package (Marçais et al., 2018) in order to keep only known targeted loci. The minimum size threshold for considering two sequences identical was set to *–mincluster=40* in order to match homologous loci even in divergent species to the one used for baits design (*Oreochromis niloticus*). Reads were mapped against the reference contigs obtained for each species[22] using bwa-mem algorithm (H. Li, 2013). Duplicates were removed using *MarkDuplicates* in Picard (Broad Institute, 2019). Variant calling was performed for each species separately using GATK (McKenna et al., 2010) by keeping all sites (using the –all-sites flag in GenotypeGVCFs). We applied filters following GATK's hard filtering best practices (https://gatk.broadinstitute.org/) and discarded indels. Depth was additionally filtered at the genotype level using custom R scripts, with a lower bound of always *dp*=10 and a higher bound depending on the per genotype distribution of coverage extracted using VCFtools (--geno-depth flag). Genotypes that were outside the boundaries of the filter value were attributed a missing value.

4.3.3.2. Genetic diversity and demographic indices

One species was removed from the dataset because of a too low sample size (*Epibulus insidiator*, N=4). For the remaining 42 species, all sites with missing data were removed, resulting in per-species variant calling files (VCFs) including monomorphic sites and Single Nucleotide Polymorphism (SNPs). We computed genetic summary statistics using custom R scripts from the resulting filtered datasets. We first calculated the folded *site frequency spectrum* (SFS). The mean pairwise difference (θ_π), Watterson's estimator of genetic diversity (θ_w) (Watterson, 1975) and the Tajima's D (TD) (Tajima, 1989) were directly computed from the SFS, and θ_π and θ_w were both standardized by the total number of sites (i.e., monomorphic sites included). We then used the SFS as input into the *stairwayplot* software (Liu & Fu, 2020), a non-parametric model inferring variations in the coalescence rate through time. We used a mutation rate of μ = 2e-9 mutations per site per generation for all species (in the range of [30,31]). Generation time was not available in the literature for all species, so when missing it was inferred by building a regression model using the maximum size as predictor (see Supplementary material).

Under the standard coalescent model (a panmictic population of constant size), the SFS is expected to be flat when normalized as in (Lapierre et al., 2017) (hereafter referred to as *norm-SFS*).

Similarly, the *stairwayplot* reconstructed from such *SFS* is expected to display no variation of coalescence rate through time. In consequence, and in addition to classical summary statistics, we computed four indices expected to reflect deviations from the constant expectations of the *norm-SFS* and the *stairwayplot*:

(1) The standardized Euclidian distance between the observed *norm-SFS* and the expected *norm-SFS* under the constant demographic model: $d_{SFS} = \frac{\sqrt{\sum_{i=1}^{n}\left(\zeta_{OBS_i} - \zeta_{EXP_i}\right)^2}}{n}$, where ζ_{OBS_i} and ζ_{EXP_i} respectively represent the observed and expected normalized values in class i of a sample of size n;

(2) The standardized Euclidian distance between the value of θ through time reconstructed by the *stairwayplot* and the observed θ_π: $d_{STAIR} = \frac{\sqrt{\sum_{i}^{l}\left(\theta_{OBS_i} - \theta_\pi\right)^2}}{l}$ where θ_{OBS_i} represents the value of θ in each time interval (l);

(3) Ratio between the ancestral (i.e., at the time to the most recent common ancestor of the sample, t_{MRCA}) and modern θ estimated by the *stairwayplot*: $R_{STAIR} = \frac{\theta_{t_{MRCA}}}{\theta_{MOD}}$;

(4) Fluctuations in the *stairwayplot*, as assessed by computing the sum of absolute slopes between each time interval: $S_{STAIR} = \frac{\sum_{i=2}^{l}\left|\frac{\theta_i - \theta_{i-1}}{T_i - T_{i-1}}\right|}{l}$, where time was discretized in l intervals of 10 generations (a trade-off between computational time and resolution) and θ values were averaged within each of them.

Because of the reduced accuracy of the *stairwayplot* in reconstructing coalescent rate in recent or very old times (Liu & Fu, 2015; Reid & Pinsky, 2022) the most recent 100 generations as well as those older than 85% of the estimated t_{MRCA} were removed from the analyses.

4.3.3.3. Phylogeny

For each of the 43 species, we sampled one individual with lower rate of missing data. To obtain a joint variant calling, trimmed reads were mapped against the baits used in the capture protocol. The joint variant calling was performed following the same workflow used on the individual species in which the reference was built with the SPAdes-3.15.4 algorithm. We removed i) sites with a depth of coverage below 10 or above 200; ii) sites with more than 20% of missing data; iii) loci with less than 10 SNPs. Finally, we wrote a Phylip file for each locus (i.e., partition) using vcf2phylip.py software (Ortiz, 2019). To infer the phylogeny, we used the IQ-TREE v.2 (Minh et

al., 2020) pipeline. We allowed each locus to have its own phylogenetic tree, but restrained the number of independent substitution rate parameters by applying the same values to loci selected for the identical most likely substitution model. (using -p and -model TESTMERGE options, resulting in 41 different partitions). A consensus tree was then estimated and branch support evaluated using 10,000 *ultra-fast* bootstraps.

4.3.3.4. Metabarcoding curation and Niche Breadth

COI and 18S metabarcoding data of gut content was subsetted for all available species from the dataset of (Casey et al., in prep). Four species did not have gut content data (*Cantherines sandwichiensis, Cephalopholis argus, Cephalopholis urodeta* and *Chlorus spirulus*). Analyses for the remaining 38 species were performed at different resource levels: ESV, species, genus and family. For each dataset, we used the following general workflow in three steps: (1) data was pooled over similar resource taxa (i.e., ESV, species, genus family); (2) for each consumer species, the number of sequenced reads per resource taxa was summed over all sampled individuals; and (3) the number of taxa consumed was computed by counting the number of resources taxa for which there was at least one sequence. Different sample size in consumer species can however influence the total number of resource taxa retrieved. To overcome this, we either subsampled to N=10 individuals (in consumer species with N>10) or projected the expected number of consumed taxa for N=10 individuals (in consumer species with N<10 and N>1). For the subsampling process, the final number of consumed resources was the average value over 100 random resampling runs. For the projection (i.e., N<10), we performed 100 rarefaction curves between the number of consumed resources and the sample size, and extracted the average number of interactions for each sample size. We then performed a linear model and used as number of consumed resources the value predicted for N=10.

4.3.3.5. Linear modelling

We performed a set of Bayesian linear models using the R library brms (Bürkner, 2021) to test for the direct effect of the number of consumed resources on genetic indices, *i.e.,* θ_π, θ_w, TD, d_{SFS}, d_{STAIR}, R_θ, S_{STAIR}. To test whether the signals were biased by shared ancestry, we also tested the same set of models including a phylogenetic variance-covariance matrix as random effect on the intercept. This matrix was estimated by means of the *ape* R package to which we fed the previously computed phylogeny. All these models were performed considering consumed resources at the ESV, species, genus and family levels for all three (COI, 18S, 23S) markers. For each model, we

performed 10,000 MCMC total iterations, with a burn-in of 2,000 and a thinning of 4. The analysis was repeated four times to check for convergence leading to an effective sample size of almost 8,000. Variables θ_w, d_{SFS}, d_{STAIR}, R_θ and S_{STAIR} were log-transformed prior to modeling to fit a gaussian distribution. We note that the results of demographic indices were multiplied by (-1) to correspond to a stability index instead of an instability index.

4.3.4. Supplementary Material

4.3.4.1. Supplementary Results

Estimation of generation times

As the generation time was not available in all species (i.e., only N=17), we estimated missing values by projection on the regression curve obtained through a linear model between the generation time (predictive variable) and the maximum age of the species ($p<0.05$, Adj. $R^2=0.29$). Note that five individuals had missing maximum age values: those were estimated by a projection on the regression curve obtained through a linear model between maximum age (predictive variable) and the maximum size ($p<0.001$, Adj. $R^2=0.43$).

Phylogeny

The phylogeny (here displayed with 20% of missing data) was performed on 1239 loci that were merged into 41 partitions after model selection and merging step performed using IQTREE v2. All branches were supported by more than 78% bootstrapped values and, interestingly, *Eviota* species formed a paraphyletic group with *Eviota infulata* estimated as an outgroup to *Paragobiodon modestus*, *Pleurosicya labiata* and the three remained sampled *Eviota* species (Figure S1).

4.3.4.2. Supplementary Tables

Table 4.1. Genetic Summary Statistics per species: number of sampled individuals (N), number of SNPs, Mean Pairwise Difference (θ_π), Watterson's estimator of genetic diversity (θ_w) and Tajima's D (TD).

Species	N	SNPs	θ_π	θ_w	TD
Abudefduf sexfasciatus	7	3667	0.00145324	0.00127321	0.6417675
Acanthurus triostegus	9	8546	0.00295615	0.00317118	-0.29061
Caracanthus maculatus	15	26188	0.00685149	0.00867827	-0.8253215
Centropyge bispinosa	12	25722	0.0053974	0.0084041	-1.453252
Chaetodon auriga	6	10215	0.00325639	0.00330588	-0.0708874
Chaetodon citrinellus	8	6989	0.00289369	0.00308267	-0.2695122
Chromis iomelas	18	1976	0.00193298	0.00557242	-2.4921097
Cirripectes variolosus	13	817	0.00140029	0.00288904	-2.0595693
Ctenochaetus striatus	9	18896	0.00699565	0.00738178	-0.2242177
Dascyllus flavicaudus	8	9469	0.00416333	0.00407928	0.0905871
Dascyllus trimaculatus	6	6956	0.00305922	0.00310359	-0.0677079
Enneapterygius pyramis	20	15162	0.00438627	0.00775826	-1.6376676

Species	N	SNPs	θ_π	θ_w	TD
Eviota afelei	18	1604	0.00378158	0.00386783	-0.0850551
Eviota albolineata	18	498	0.00444007	0.00670237	-1.2824561
Eviota distigma	16	1512	0.00543099	0.00435766	0.95502828
Eviota infulata	21	2410	0.00640856	0.00650005	-0.0526517
Fusigobius neophytus	14	1713	0.00384503	0.00393777	-0.0931764
Glyptoparus delicatulus	20	1212	0.00048788	0.00105869	-2.0269845
Gnatholepis cauerensis	19	2136	0.00434018	0.00668806	-1.3300267
Myripristis berndti	7	10085	0.00516948	0.00624918	-0.7844138
Myripristis kuntee	5	13935	0.00726968	0.00908755	-1.0050481
Myripristis violacea	6	6468	0.00355744	0.00398411	-0.5071049
Naso lituratus	5	6812	0.00311676	0.0033817	-0.3935952
Neocirrhites armatus	15	6097	0.00493581	0.00591438	-0.6484911
Neoniphon sammara	5	3440	0.00190846	0.00207726	-0.4081514
Ostorhinchus angustatus	10	2898	0.00181787	0.00199197	-0.3666312
Paracirrhites arcatus	14	12038	0.00303369	0.00484143	-1.4791819
Paragobiodon modestus	13	1594	0.0032709	0.00374009	-0.5021346
Plectranthias nanus	18	25901	0.00942652	0.01582091	-1.5442789
Pleurosicya labiata	16	1542	0.00610982	0.00637898	-0.1636114
Priolepis semidoliata	14	1130	0.00296025	0.00297264	-0.0164784
Pseudocheilinus hexataenia	18	16871	0.00712995	0.01054111	-1.236369
Pseudogramma polyacanthum	20	27625	0.00424859	0.01062537	-2.2615175
Rhinecanthus aculeatus	11	3339	0.00170866	0.00178212	-0.1698934
Sebastapistes fowleri	19	18707	0.0060021	0.01279912	-2.0144227
Stegastes nigricans	23	12917	0.00389195	0.00347187	0.44804371
Sufflamen bursa	5	6484	0.00327873	0.00371479	-0.5897021
Zebrasoma scopas	10	16892	0.00538594	0.0063735	-0.6503832

Table 4.2. Posterior summary of linear models: Median, 95% Confidence Interval, hypothesis tested and associated posterior probability for the slope of the different Genetic Indices when modelled with the number of interactions calculated at different phylogenetic scales using COI or 18S marker and with or without the phylogenetic variance covariance matrix as random effect on the intercept (Phylo column).

Scale	Phylo	Gen. Index	Marker	Median	CI (Lw)	CI (Up)	Hypothesis	Post.Prob
Species	No	d_{SFS}	18S	0.11866	0.17294	0.06246	$d_{SFS} > 0$	0.999625
			COI	0.00318	0.00734	-0.00107		0.89275
		d_{STAIR}	18S	0.20902	0.37783	0.03937	$d_{STAIR} > 0$	0.977
			COI	0.00819	0.02005	-0.00363		0.877625
		f_{STAIR}	18S	0.15068	0.31581	-0.01710	$f_{STAIR} > 0$	0.929875
			COI	0.00987	0.02084	-0.00083		0.934
		R_{STAIR}	18S	0.19597	0.35367	0.03558	$R_{STAIR} > 0$	0.976625
			COI	0.01633	0.02598	0.00675		0.99575
		TD	18S	0.10022	0.01459	0.18869	TD > 0	0.97375
			COI	0.00768	0.00197	0.01316		0.987125
		θ_π	18S	-0.00014	-0.00035	0.00008	$\theta_\pi < 0$	0.861375
			COI	-0.00001	-0.00002	0.00001		0.7545
		θ_w	18S	-0.06989	-0.13656	-0.00571	$\theta_w < 0$	0.96225
			COI	-0.00359	-0.00790	0.00085		0.91225
	Yes	d_{SFS}	18S	0.14374	0.20017	0.08584	$d_{SFS} > 0$	1
			COI	0.00384	0.00839	-0.00072		0.918875
		d_{STAIR}	18S	0.27132	0.43311	0.11136	$d_{STAIR} > 0$	0.995375
			COI	0.01237	0.02449	0.00026		0.953
		f_{STAIR}	18S	0.19109	0.34993	0.03100	$f_{STAIR} > 0$	0.974125
			COI	0.01259	0.02432	0.00117		0.963125
		R_{STAIR}	18S	0.22560	0.38536	0.06355	$R_{STAIR} > 0$	0.987875
			COI	0.01911	0.02969	0.00870		0.9985
		TD	18S	0.13042	0.04425	0.21870	TD > 0	0.992125
			COI	0.01039	0.00465	0.01638		0.9975
		θ_π	18S	-0.00004	-0.00027	0.00017	$\theta_\pi < 0$	0.62425
			COI	0.00000	-0.00002	0.00001		0.55075
		θ_w	18S	-0.04785	-0.11433	0.01265	$\theta_w < 0$	0.909125
			COI	-0.00300	-0.00769	0.00140		0.8685
Genus	No	d_{SFS}	18S	0.02074	0.03578	0.00542	$d_{SFS} > 0$	0.9855
			COI	0.00476	0.00997	-0.00038		0.9365
		d_{STAIR}	18S	0.04078	0.08277	-0.00256	$d_{STAIR} > 0$	0.9395
			COI	0.01292	0.02725	-0.00190		0.926

Scale	Phylo	Gen. Index	Marker	Median	CI (Lw)	CI (Up)	Hypothesis	Post.Prob
		f_{STAIR}	18S	0.02987	0.07128	-0.01262	$f_{STAIR} > 0$	0.8765
			COI	0.01402	0.02773	0.00039		0.954625
		R_{STAIR}	18S	0.03685	0.07707	-0.00365	$R_{STAIR} > 0$	0.93425
			COI	0.02140	0.03344	0.01014		0.99825
		TD	18S	0.02517	0.00263	0.04706	TD > 0	0.9665
			COI	0.01020	0.00338	0.01703		0.993
		θ_π	18S	-0.00004	-0.00009	0.00002	$\theta_\pi < 0$	0.880875
			COI	-0.00001	-0.00002	0.00001		0.772625
		θ_w	18S	-0.01786	-0.03460	-0.00161	$\theta_w < 0$	0.965
			COI	-0.00476	-0.01007	0.00063		0.931875
	Yes	d_{SFS}	18S	0.02550	0.04354	0.00901	$d_{SFS} > 0$	0.992625
			COI	0.00605	0.01186	0.00043		0.961375
		d_{STAIR}	18S	0.05627	0.10202	0.00994	$d_{STAIR} > 0$	0.9785
			COI	0.01970	0.03463	0.00456		0.981
		f_{STAIR}	18S	0.03685	0.08097	-0.00869	$f_{STAIR} > 0$	0.91075
			COI	0.01937	0.03372	0.00476		0.9845
		R_{STAIR}	18S	0.04395	0.08911	0.00096	$R_{STAIR} > 0$	0.954375
			COI	0.02625	0.03980	0.01322		0.999375
		TD	18S	0.03627	0.01204	0.06022	TD > 0	0.992625
			COI	0.01466	0.00726	0.02188		0.99925
		θ_π	18S	-0.00002	-0.00008	0.00004	$\theta_\pi < 0$	0.741625
			COI	0.00000	-0.00002	0.00002		0.522375
		θ_w	18S	-0.01715	-0.03442	0.00039	$\theta_w < 0$	0.9455
			COI	-0.00433	-0.01027	0.00163		0.88625
Family	No	d_{SFS}	18S	0.02679	0.04539	0.00827	$d_{SFS} > 0$	0.990125
			COI	0.00569	0.01193	-0.00060		0.931875
		d_{STAIR}	18S	0.05447	0.10714	0.00318	$d_{STAIR} > 0$	0.95975
			COI	0.01629	0.03352	-0.00068		0.942
		f_{STAIR}	18S	0.04109	0.09195	-0.00771	$f_{STAIR} > 0$	0.916
			COI	0.01588	0.03233	-0.00039		0.945375
		R_{STAIR}	18S	0.05005	0.09953	0.00124	$R_{STAIR} > 0$	0.95425
			COI	0.02635	0.04065	0.01247		0.99875
		TD	18S	0.03119	0.00480	0.05786	TD > 0	0.973375
			COI	0.01270	0.00477	0.02081		0.993625
		θ_π	18S	-0.00005	-0.00011	0.00002	$\theta_\pi < 0$	0.895125
			COI	-0.00001	-0.00003	0.00001		0.795125

Scale	Phylo	Gen. Index	Marker	Median	CI (Lw)	CI (Up)	Hypothesis	Post.Prob
ESV		θ_w	18S	-0.02286	-0.04253	-0.00302	$\theta_w < 0$	0.967875
			COI	-0.00617	-0.01252	0.00011		0.9465
	Yes	d_{SFS}	18S	0.03565	0.05672	0.01451	$d_{SFS} > 0$	0.997125
			COI	0.00767	0.01470	0.00081		0.96475
		d_{STAIR}	18S	0.07860	0.14464	0.02287	$d_{STAIR} > 0$	0.98625
			COI	0.02606	0.04462	0.00773		0.990125
		f_{STAIR}	18S	0.05134	0.10696	-0.00148	$f_{STAIR} > 0$	0.944375
			COI	0.02311	0.04100	0.00517		0.983625
		R_{STAIR}	18S	0.05926	0.11314	0.00726	$R_{STAIR} > 0$	0.966875
			COI	0.03436	0.05123	0.01786		0.999375
		TD	18S	0.04646	0.01790	0.07556	TD > 0	0.994875
			COI	0.01943	0.01063	0.02840		0.999875
		θ_π	18S	-0.00003	-0.00010	0.00005	$\theta_\pi < 0$	0.708
			COI	0.00000	-0.00003	0.00003		0.463875
		θ_w	18S	-0.02062	-0.04243	0.00041	$\theta_w < 0$	0.946625
			COI	-0.00519	-0.01278	0.00235		0.871875
	No	d_{SFS}	18S	0.00174	0.00359	-0.00002	$d_{SFS} > 0$	0.948
			COI	0.00113	0.00254	-0.00033		0.902875
		d_{STAIR}	18S	0.00176	0.00682	-0.00335	$d_{STAIR} > 0$	0.71475
			COI	0.00259	0.00659	-0.00143		0.86
		f_{STAIR}	18S	0.00167	0.00635	-0.00317	$f_{STAIR} > 0$	0.719375
			COI	0.00308	0.00668	-0.00052		0.91675
		R_{STAIR}	18S	0.00358	0.00815	-0.00079	$R_{STAIR} > 0$	0.91
			COI	0.00463	0.00809	0.00138		0.988
		TD	18S	0.00194	-0.00061	0.00436	TD > 0	0.895
			COI	0.00231	0.00043	0.00418		0.975375
		θ_π	18S	0.00000	-0.00001	0.00000	$\theta_\pi < 0$	0.79425
			COI	0.00000	-0.00001	0.00000		0.738375
		θ_w	18S	-0.00117	-0.00305	0.00066	$\theta_w < 0$	0.852375
			COI	-0.00109	-0.00254	0.00034		0.895375
	Yes	d_{SFS}	18S	0.00214	0.00403	0.00024	$d_{SFS} > 0$	0.96725
			COI	0.00125	0.00276	-0.00027		0.9165
		d_{STAIR}	18S	0.00335	0.00867	-0.00203	$d_{STAIR} > 0$	0.851125
			COI	0.00363	0.00777	-0.00043		0.928375
		f_{STAIR}	18S	0.00268	0.00779	-0.00229	$f_{STAIR} > 0$	0.809375
			COI	0.00377	0.00765	-0.00010		0.945
		R_{STAIR}	18S	0.00443	0.00923	-0.00042	$R_{STAIR} > 0$	0.937

Scale	Phylo	Gen. Index	Marker	Median	CI (Lw)	CI (Up)	Hypothesis	Post.Prob
			COI	0.00531	0.00898	0.00169		0.9895
		TD	18S	0.00295	0.00028	0.00556	TD > 0	0.964625
			COI	0.00299	0.00109	0.00498		0.994125
		θ_π	18S	0.00000	-0.00001	0.00001	$\theta_\pi < 0$	0.641875
			COI	0.00000	-0.00001	0.00000		0.6425
		θ_w	18S	-0.00089	-0.00292	0.00107	$\theta_w < 0$	0.77025
			COI	-0.00104	-0.00254	0.00042		0.88375

4.3.4.3. Supplementary Figures

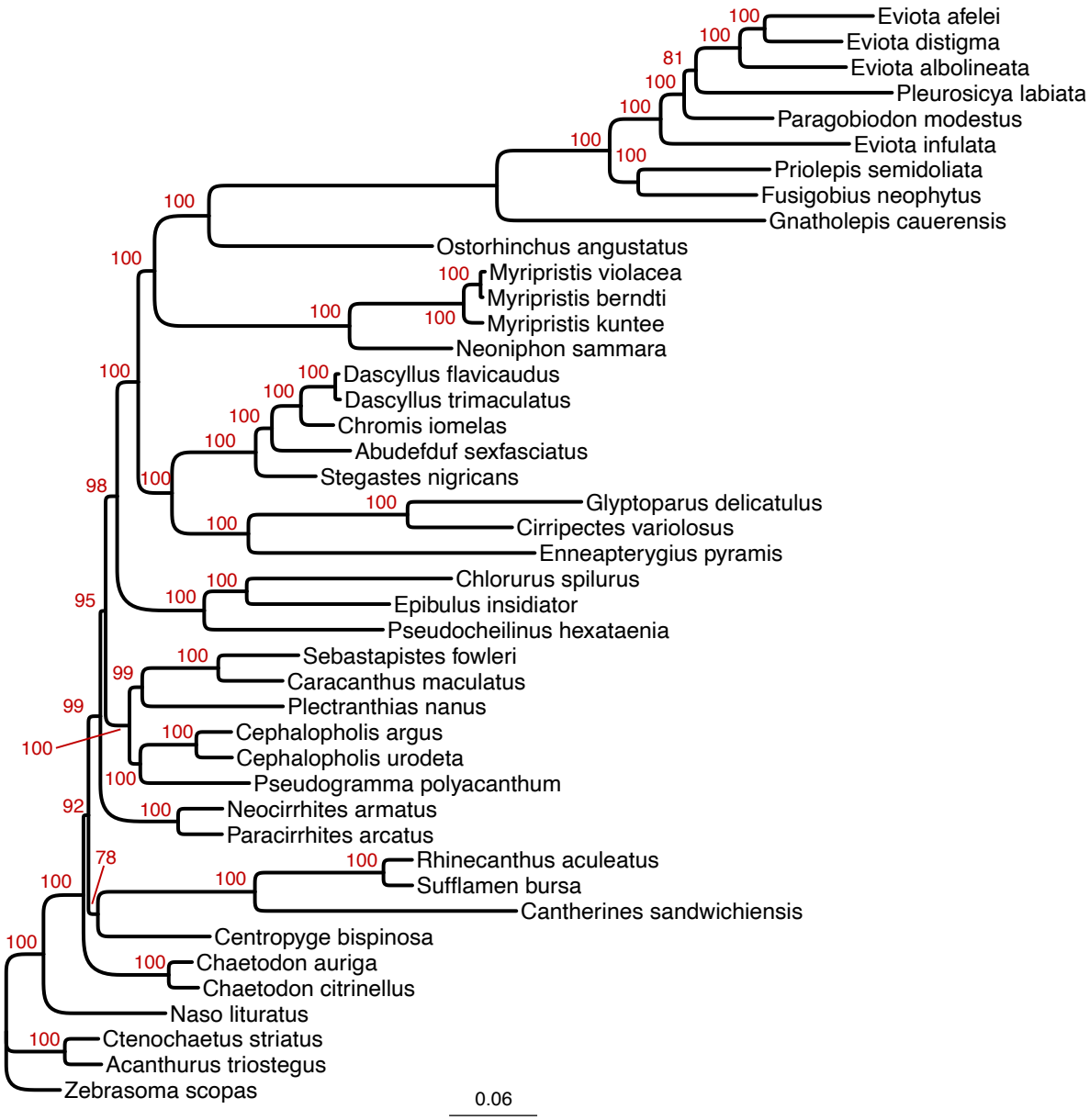

Figure 4.3. Phylogenetic tree of the 43 species. Values in red represent the percentage of branch support calculated from 10,000 ultra-fast bootstrap iterations.

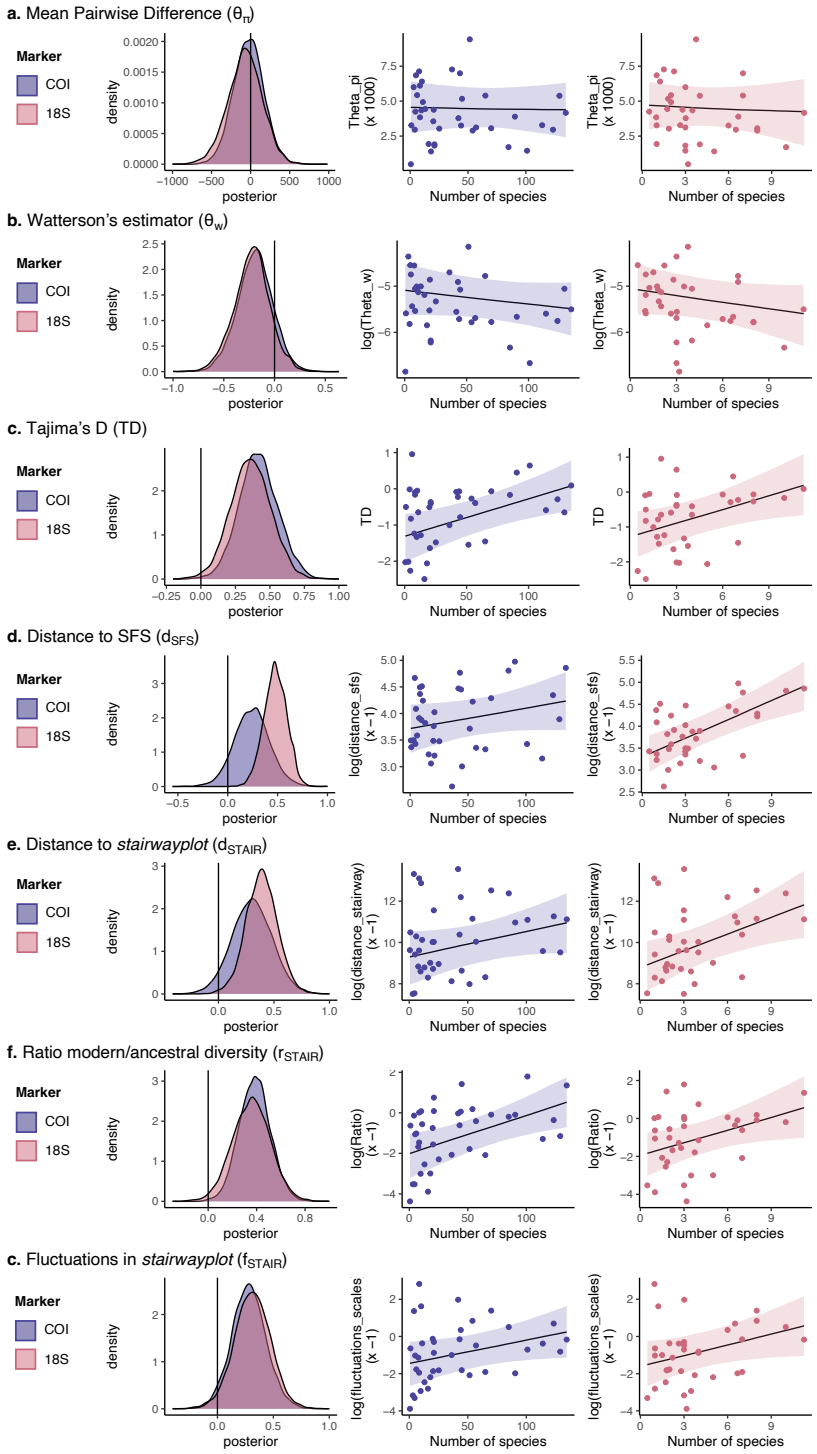

Figure 4.4. Bayesian Linear models relating genetic indices and the number of consumed species with the phylogeny as random effect around the intercept. Each row represents a genetic index. For each genetic index, the three panels represent: i) the posterior distribution of the slope when the N° of consumed species is estimated by the COI (purple) or 18S (red); ii) the regression line relating genetic indices and N° of consumed species estimated by the COI and iii) 18S marker. The posterior distributions of the slopes are scaled by their maximum absolute value for comparison purposes.

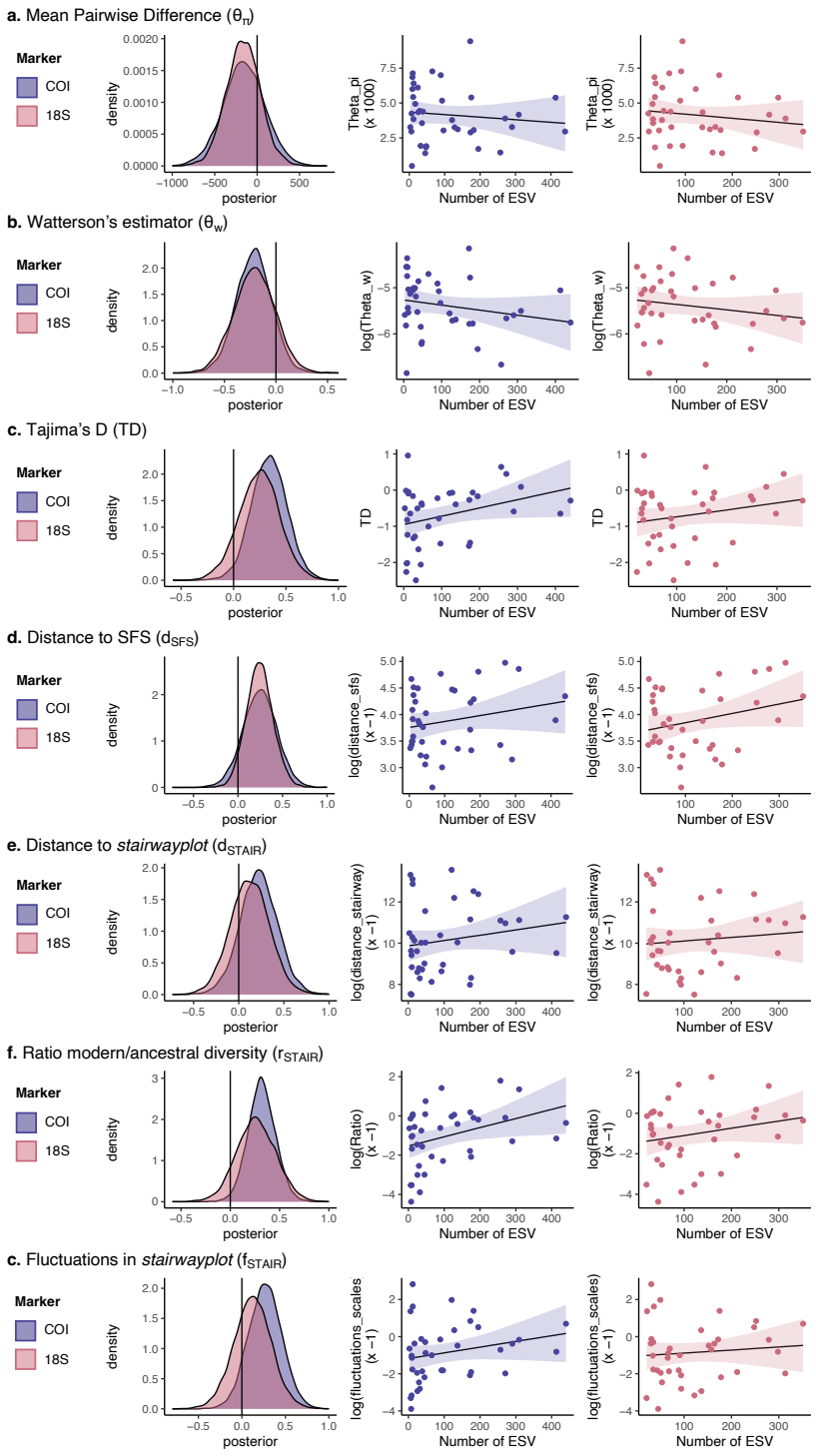

Figure 4.5. Bayesian Linear models relating genetic indices and the number of consumed ESV. Each row represents a genetic index. For each genetic index, the three panels represent: i) the posterior distribution of the slope when the N° of consumed ESV is estimated by the COI (purple) or 18S (red); ii) the regression line relating genetic indices and N° of consumed ESV estimated by the COI and iii) 18S marker. The posterior distributions of the slopes are scaled by their maximum absolute value for comparison purposes.

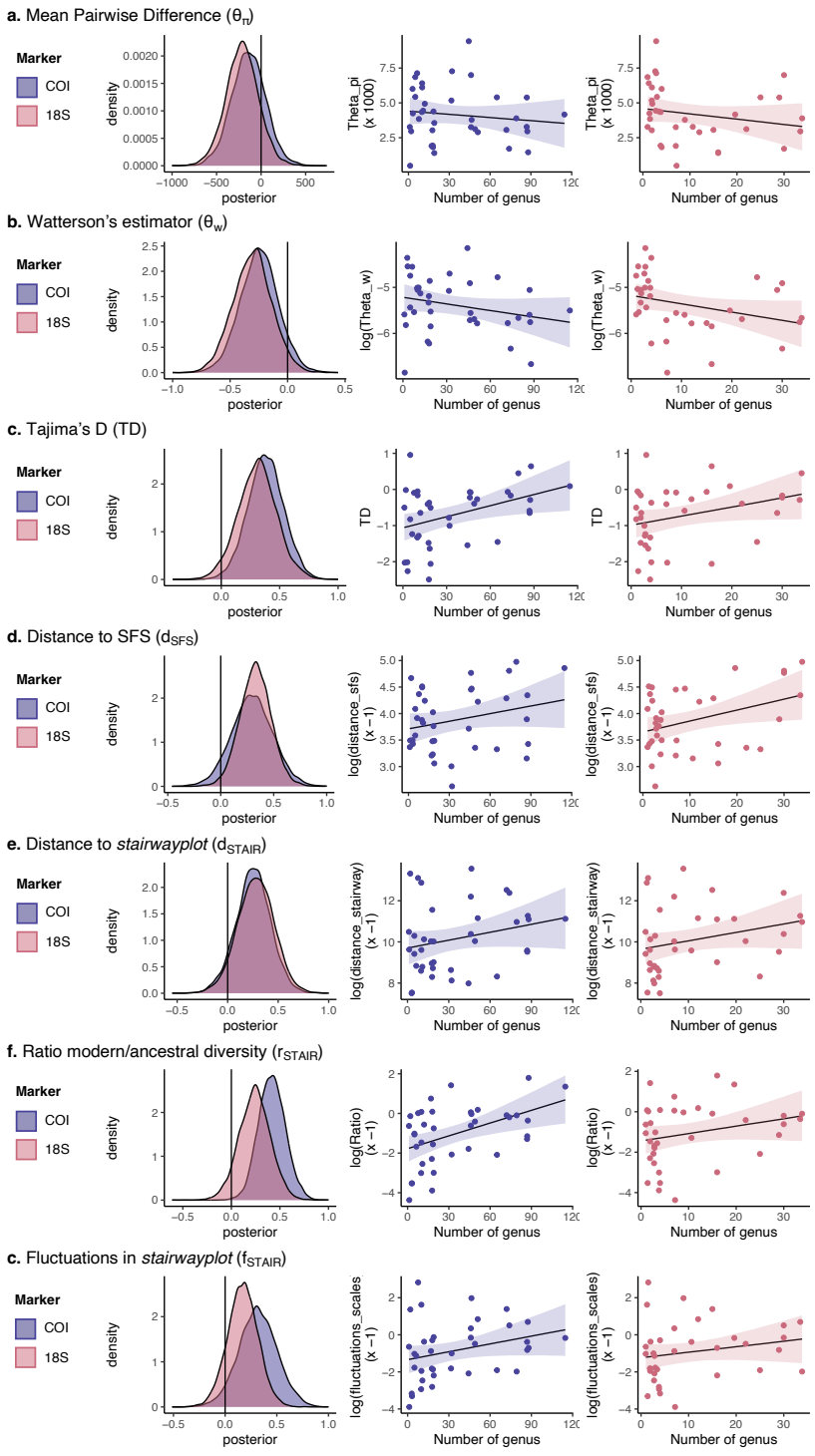

Figure 4.6. Bayesian Linear models relating genetic indices and the number of consumed genera. Each row represents a genetic index. For each genetic index, the three panels represent: i) the posterior distribution of the slope when the N° of consumed genera is estimated by the COI (purple) or 18S (red); ii) the regression line relating genetic indices and N° of consumed genera estimated by the COI and iii) 18S marker. The posterior distributions of the slopes are scaled by their maximum absolute value for comparison purposes.

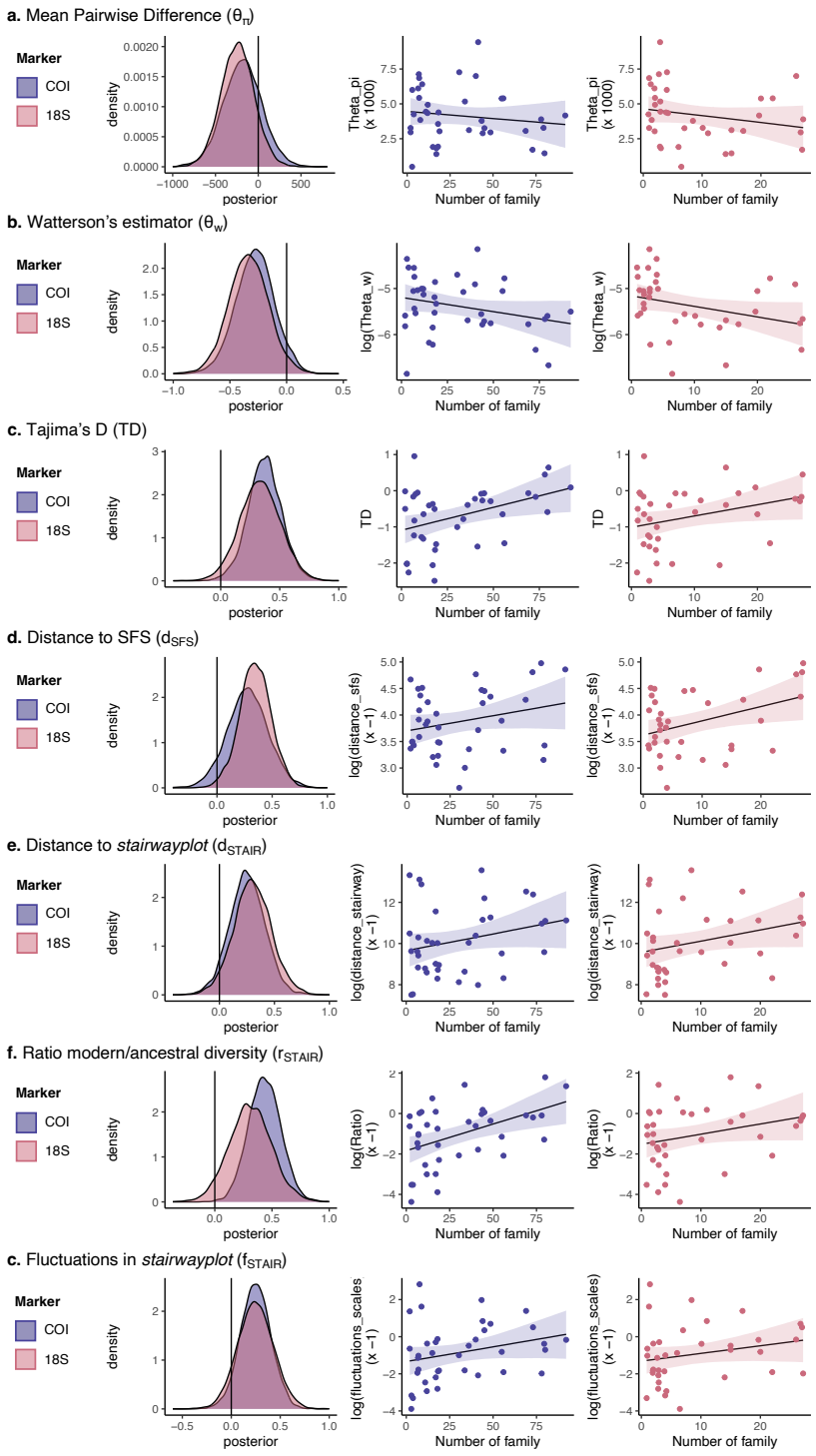

Figure 4.7. Bayesian Linear models relating genetic indices and the number of consumed families. Each row represents a genetic index. For each genetic index, the three panels represent: i) the posterior distribution of the slope when the N° of consumed families is estimated by the COI (purple) or 18S (red); ii) the regression line relating genetic indices and N° of consumed families estimated by the COI and iii) 18S marker. The posterior distributions of the slopes are scaled by their maximum absolute value for comparison purposes.

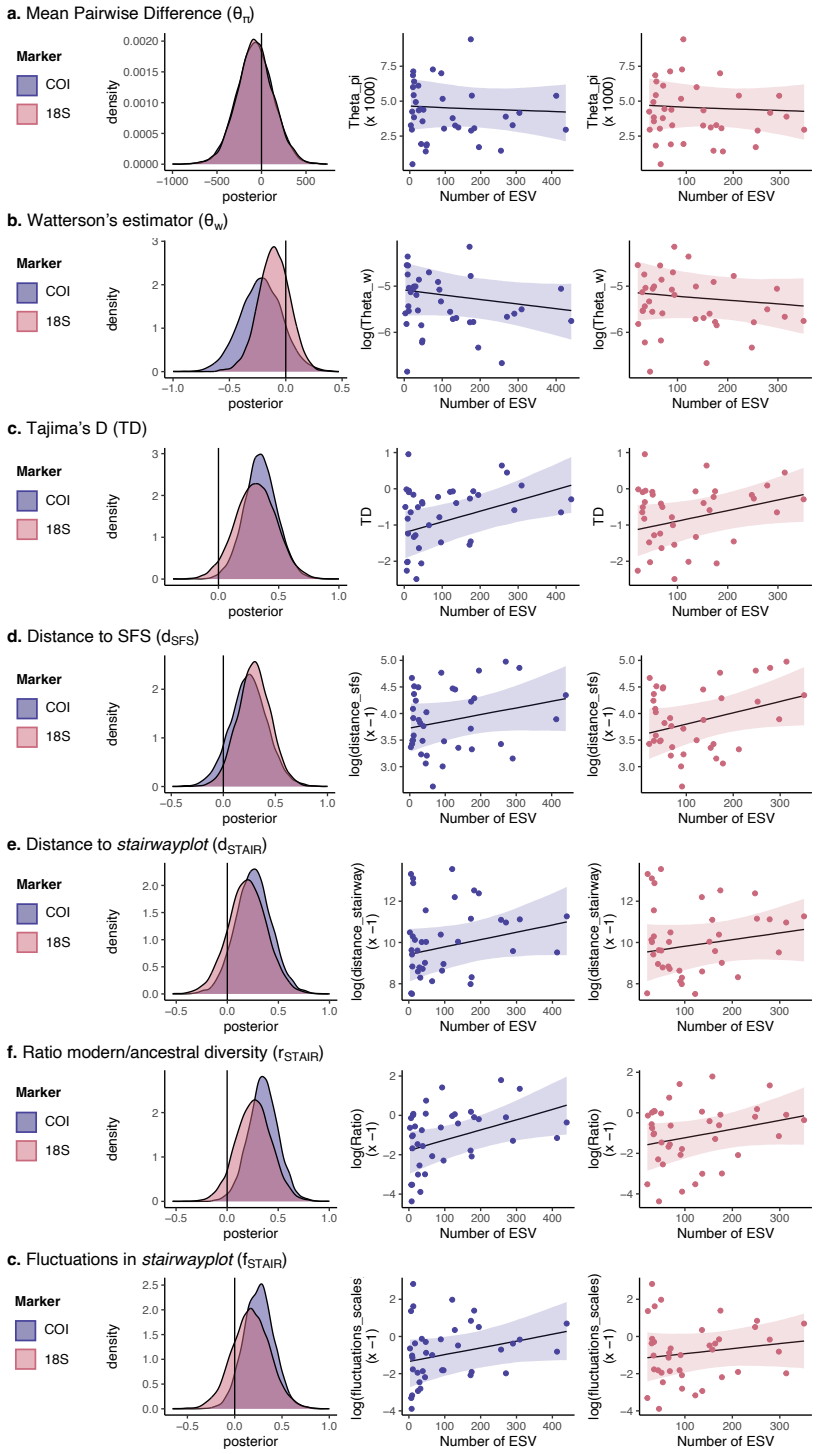

Figure 4.8. Bayesian Linear models relating genetic indices and the number of consumed ESV with the phylogeny as random effect around the intercept. Each row represents a genetic index. For each genetic index, the three panels represent: i) the posterior distribution of the slope when the N° of consumed ESV is estimated by the COI (purple) or 18S (red); ii) the regression line relating genetic indices and N° of consumed ESV estimated by the COI and iii) 18S marker. The posterior distributions of the slopes are scaled by their maximum absolute value for comparison purposes.

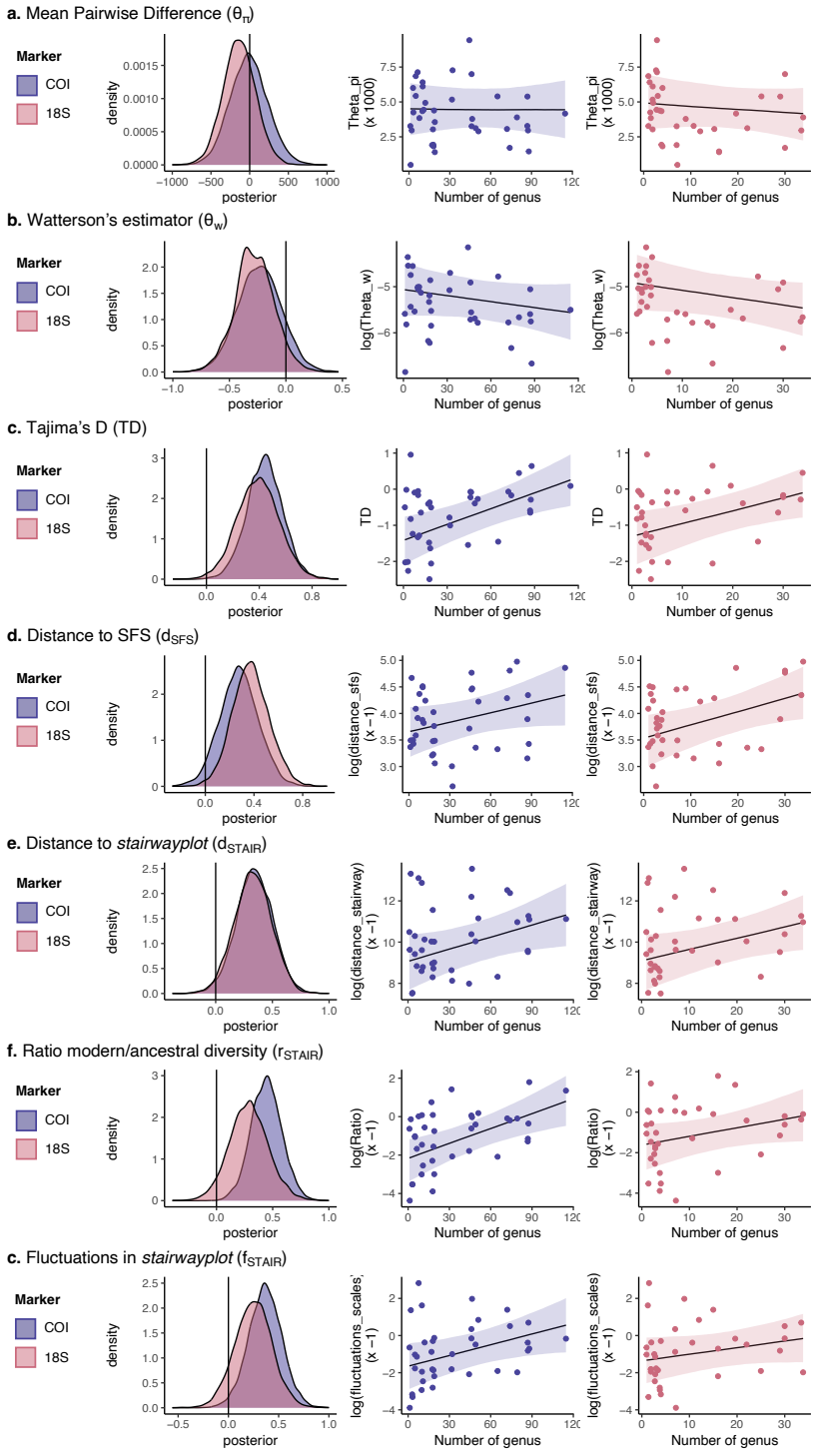

Figure 4.9. Bayesian Linear models relating genetic indices and the number of consumed genera with the phylogeny as random effect around the intercept. Each row represents a genetic index. For each genetic index, the three panels represent: i) the posterior distribution of the slope when the N° of consumed genera is estimated by the COI (purple) or 18S (red); ii) the regression line relating genetic indices and N° of consumed genera estimated by the COI and iii) 18S marker. The posterior distributions of the slopes are scaled by their maximum absolute value for comparison purposes.

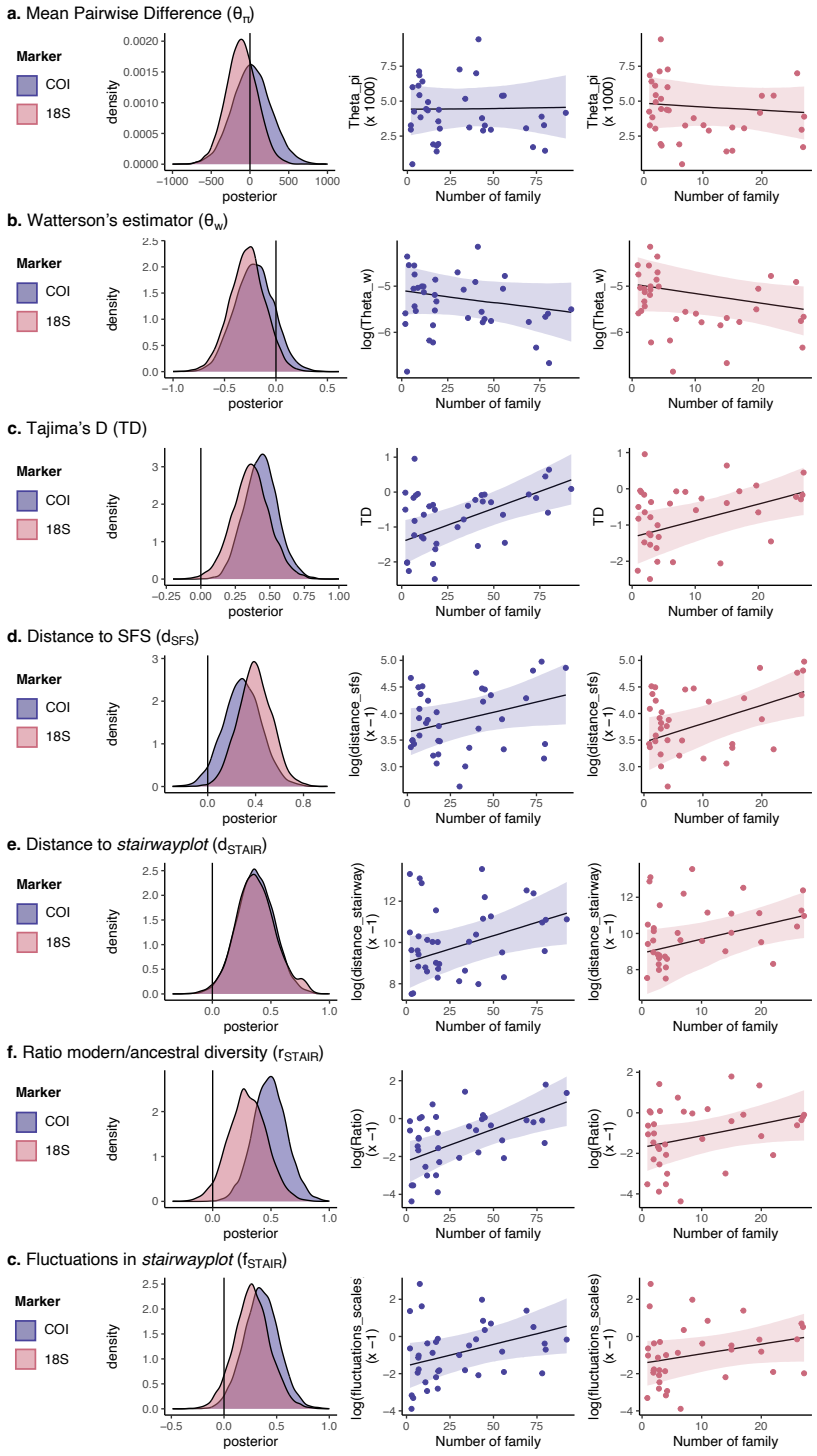

Figure 4.10. Bayesian Linear models relating genetic indices and the number of consumed families with the phylogeny as random effect around the intercept. Each row represents a genetic index. For each genetic index, the three panels represent: i) the posterior distribution of the slope when the N° of consumed families is estimated by the COI (purple) or 18S (red); ii) the regression line relating genetic indices and N° of consumed families estimated by the COI and iii) 18S marker. The posterior distributions of the slopes are scaled by their maximum absolute value for comparison purposes.

4.4. Conclusions and perspectives

4.4.1. A multi-species population genetics dataset to understand ecosystem functioning

In this chapter I set up a unique interdisciplinary approach allowing to establish a relation between ecological (e.g., ecosystem functioning) and population genetics concepts (e.g., historical demographic inferences). Population genetics traditionally focuses on single-species inferences because of a lack of both theoretical developments on the influence of species interactions on genetic diversity and of multi-species population genetics dataset, which have been unaffordable for long time. To accomplish this challenging task, I followed a Target Gene Capture protocol (C. Li et al., 2013) which targets a set of specific loci homologous across species. This provides two major advantages: (1) the ability to compare genetic estimates between species and (2) the ability to, if desired, extend the dataset to other species and/or other locations, directly profiting from the homologous nature of the targeted genes. The overall framework proved powerful, as it enabled us to demonstrate a remarkable positive correlation between trophic niche width and stability in demographic history, consistently with previous hypotheses (Carscadden et al., 2020; Colles et al., 2009; Gravel et al., 2011; S. Li et al., 2014; Pasinelli, 2022), suggesting that generalist species might be less prone to extinction than specialist ones. This framework led to strong conservation implications driven by the conclusion that generalist species should be less vulnerable than specialists. In addition, the framework opens a door to many perspectives and generalizations – some of which I explain below. Ultimately, they will help grasping more knowledge on ecosystem functioning and on the genomic signatures left by large-scale processes, both of which being key to understand the future resilience of biodiversity in the context of a huge biodiversity crisis (Ceballos et al., 2015).

4.4.2. Predictors of historical demography, coalescence rate and beyond

In this chapter, I devised several demographic stability indices, based on both the Site Frequency Spectrum (SFS) and the reconstructed coalescence rates through time by the *stairwayplot*, an SFS-based *unstructured* model. I performed linear models using trophic niche width data estimated from meta-barcoding of gut contents of the same species (extracted from Casey et al., *in prep*) as predictor variables of the genetic indices. Trophic niche was thus found to be a predictor of

demographic history as larger niche width was positively associated with all stability indices. However, it is important to note that species can be generalists in some niche and specialists in another (Poisot et al., 2011). Despite the strength of our result suggests that trophic niche width plays an important role in shaping demographic history, it might be resourceful in the future to understand how other niche components drive historical demography, such as the climatic niche (i.e., the extent of climatic conditions under which species can strive). This would allow to disentangle the relative contribution of each niche component to demographic stability, which might be proven useful to refine predictors of genetic diversity and vulnerability, eventually crucial in order to predict the future of biodiversity.

Interestingly, I noted that trophic niche width hardly correlated with direct estimates of genetic diversity (i.e., the mean pairwise difference, θ_π, (Tajima, 1983), and Watterson's estimate θ_w), which summaries the SFS, already a summary the gene genealogy. This could be due to a more intricate relation between genetic diversity and fluctuations than initially expected by Kimura & Crow (1963). More importantly it suggests that interpretations of genetic diversity estimates must be made carefully and prompts to couple investigation with more detailed indices of genetic diversity, such as the SFS. Additionally, once again, this chapter displays the benefits of studying the reconstruction of the IICR through time using *unstructured* models. For instance, it shows that ecological features such as niche width impact the gene genealogy (as shown by the strong correlations with the SFS) and thus as expected the coalescence rate (as shown by the correlations with the IICR reconstructed by the *stairwayplot*). This increases our understanding of the determinants of coalescence times and shows how powerful such statistics can be to investigate processes up to the scale of the ecosystem.

However, the current modelling (limited to *unstructured* models) is unlikely to be enough to understand the full evolutionary picture of each species studied here, just as it has been developed in **Chapters 2** and **3**. In fact, here I inferred stability as less changes in coalescence rate, but **Chapter 2** (and the introduction, **Chapter 1**) clearly showed how population structure drives variation in the coalescence rate even with no changes in effective size. The model behind should thus be complexified in the future to account for the degree of structure in each species. In addition, as developed in the introduction to this chapter, under the SGVH specialist species should also display lower amounts of connectivity between *populations* (Gravel et al., 2011; S. Li et al., 2014). As already suggested, one strategy to formally test it would be to extend the current dataset to

other sampling sites (which will be more detailed in the following section). I note however that more complex modelling can be complicated in this context for two reasons: (1) the sometimes low and unequal number of SNPs in the set of species, which will condition accuracy in estimates of complex models and (2) the likely non-adequacy of the standard coalescent models to fishes. For instance, fishes display sweepstake reproduction strategies (Hedgecock & Pudovkin, 2011), and their ancestral process might be better described by a multiple-merger coalescent framework (Pitman, 1999; Sagitov, 1999; Tellier & Lemaire, 2014). In this chapter, I assumed that the bias generated by applying Kingman's coalescence should be the same in all species. This is not necessarily true, since the variance in reproductive process is likely to vary between species (therefore, the multiple merger process may not be the same in all sampled species). Moreover, more complex modelling under the standard (Kingman) coalescent could be highly misleading and the originated bias heterogeneous in the set of species here considered (Vendrami et al., 2021). While our result remains robust, it will be pertinent in the future to investigate demographic modelling using a multiple-merger coalescent simulator.

4.4.3. Extending the dataset to test biogeographic hypotheses

As suggested above, the genetic dataset could be extended to test whether connectivity is conditioned by trophic niche width, i.e., to test the second part of the SGVH. In fact, the SGVH is intricately related to the *Trophic Theory of Island Biogeography* (T-TIB) (Gravel et al., 2011). In essence, the T-TIB extends the *Theory of Island Biogeography* (TIB) (MacArthur & Wilson, 1967) to account for trophic interactions. The TIB is a predictive theory of biodiversity at the species scale: the distribution of species in a given habitat depends on the interplay between colonization and extinction rates. Each rate varies in function of the size of the habitat and in function of the distance from a propagule source (i.e., a source of biodiversity), and notably, colonization rate decreases and extinction rate increases with the distance from the source (MacArthur & Wilson, 1967). The T-TIB integrates the influence of trophic niche width onto the colonization/extinction dynamics, as a species must have colonised a habitat for its predator to colonise it, which impedes colonization and increases extinction risk for specialist species in comparison to generalists' ones (Gravel et al., 2011). The T-TIB thus provides a theoretical framework to the expectations of the SGVH related to the higher demographic fluctuations due to higher frequency of extinction rates, but also to the decrease of connectivity in specialists due to the need for the right resources along

the colonization pathway. More interestingly the T-TIB adds another layer of complexity as the demographic stability when distance from source increases is not expected to be consistently conditioned by niche width (Gravel et al., 2011): while generalist species should guarantee greater demographic stability along the biodiversity gradient, and therefore stable genetic diversity between sampled sites along the gradient, specialist species should display a decrease in genetic diversity with the distance from the source, sensitive to the colonisation-extinction cycles of the few prey species consumed and to the absence of specific resources in habitats on the colonization pathway.

The dataset elaborated in this chapter is unfortunately not enough to test the expectations of the TIB and the T-TIB. However, extending it to other sites along a gradient of biodiversity will be a golden opportunity in the future to test whether stability in historical demography remains the same conditioned on niche width. Additionally, the source of biodiversity could represent a refugia for multiple species and thus an origin of Range Expansion (RE). This means that species should display more recent colonization times with increasing distance to the source. Extending the dataset could therefore allow testing whether we can find such signatures of colonisations along the gradient but also to test whether the ability to detect colonization changes between generalist and specialist species due to different rates of fluctuations. In this context, extending our dataset which includes a sampling location in French Polynesia is pertinent, as in the Pacific, coral reefs are characterized by a decrease in species richness away from the Coral Triangle a biodiversity hotspot (Roberts et al., 2002) and past refugia for reef fishes (Cowman & Bellwood, 2013) that would correspond to the source of propagule under in the TIB (MacArthur & Wilson, 1967). Pacific reefs thus represent a perfect region for testing this hypothesis, which will be key in the context of global changes. For instance, studying communities along a biodiversity gradient represents a proxy of the shift away from highly diverse to less rich habitats expected to happen over time due to biodiversity erosion.

Chapter 5. Global Synthesis

5.1. Main findings

In my PhD, I sought to demonstrate the value of complex demographic inferences to formulate evolutionary and conservation hypotheses, and to enhance our comprehension of how species- and community-level features drive historical demography and the reconstruction we may (and should) make of it. To that end, I studied how processes ranging from local structural variations in the genome to trophic interactions have shaped the evolutionary history of marine organisms using population genetics applied to genomic data. In **Chapter 1**, I introduced the coalescent theory and how it provided a resourceful basis upon which to investigate complex demographic scenarios, notably through simulations. I then introduced a demographic framework aiming to correctly infer and interpret the history of species in two steps: (1) investigation of descriptive analyses of the genetic variability in space and time, and (2) design of complex demographic scenarios based on the information obtained in the first step. This general framework rooted the investigations of the following chapters.

In **Chapter 2**, I investigated the relationship between life history traits, population structure, and coalescence rate trajectory through time as reconstructed by *unstructured* models (i.e., models assuming panmixia). Through simulations and empirical investigations, I first highlighted the signatures of the degree of population structure and specific historical events on the reconstructed coalescence rate obtained by means of *unstructured* models. This emphasized (i) how specific parameters of the *true* demographic history of species influences the distribution of coalescence times, (ii) how *unstructured* models are thus powerful tools to detect their signature; but also (iii) how blindly assuming that variation in coalescent rate is only due to variation in effective population size through time lead to strong mis-interpretations and it is necessary to couple this analysis with additional ecological and genetics evidences. This stressed the necessity of testing for structure co-jointly to applying *unstructured* models, as the two sets of analysis can uncover different features of a species' evolutionary history and need to be interpreted in concert. This was further supported in two examples of widespread shark species having a very different set of life history traits, underlining the importance of species-specific thoughtful model setting, to which rigorous investigation of population structure is of paramount importance.

In **chapter 3**, I present the discovery of a supergene responsible for a size polymorphism in the vulnerable Thorny Skate. This is a remarkable discovery which highlights, for the first time unambiguously, how a supergene interacts with the determinism of a known polygenic trait,

suggesting that several genes involved in size determinism are located within the supergene region. The system thus offers perspectives beyond the field of population genetics, notably in quantitative or functional genetics. Furthermore, this chapter focused on the supergene's origin and putative role in the conservation of this species. By thoroughly reconstructing the species' demographic history at the scale of its whole range, I was able to show that this supergene originated through introgression from a sister species, though accurate dating would require genetic data from con-generic species in the future. In addition, I was able to explain why this supergene is limited to a specific area of the Thorny Skate's range. Finally, and more importantly, I uncovered the role of the supergene in the non-recovery of one Thorny Skate population. This chapter thus provided critical insights into the conservation consequences of a size-determining supergene and its origins. In addition, it also shed light on the value of demographic investigation to understand local selective processes, as well as on the need for multi-species investigations to better characterize supergene's origin and evolution.

Finally, in **Chapter 4** I took an interdisciplinary approach coupling genomic to ecological data in order to investigate ecosystem-scale determinants of genetic diversity. To that end, I set up a unique multi-species genomic dataset that targeted homologous regions across 43 reef-associated fish species, eventually allowing for comparisons of genetic diversity between them. Genetic indices of demographic stability were studied in the light of trophic niche width reconstructed by meta-barcoding of gut contents (the latter kindly provided by Valeriano Parravicini research group). This allowed to show a positive relationship between trophic niche width and demographic history stability. This indicated that generalist species may be less vulnerable to extinction than specialists, linking directly for the first time a community-scale process to historical demography. This chapter brought support to the *Specialist-Generalist Variation Hypothesis* (SGVH), allowing to answer a long-standing question in community ecology, and at the same time provided a glance of the power of multi-species population genetics dataset for inter-disciplinary investigations.

5.2. A Framework for robust demographic inferences

Devising meaningful demographic models is challenging. In the introduction (**Chapter 1**), I developed briefly how the coalescent theory allowed to investigate intricate scenarios through simulations. In addition, I stress the need to establish a robust framework in order to study a (set of) coherent scenario(s) and thus avoid misinterpretations based on inappropriate models (e.g., as in the case of neglecting population structure). Notably, I highlighted an intuitive and meaningful demographic set-up consisting of:

(1) Collecting as much evidence on the nature and degree of population structure through descriptive analyses notably composed of F_{ST}-based statistics, clustering algorithms, ABC-RF and *unstructured* models, needed to estimate the variation in coalescence rate through time;

(2) Device the *best* (set of) scenario(s) in relation to the question and the species investigated, bolstered by the information provided by step (1)

(3) Find a computational way to model the chosen scenario(s). This can be done using coalescent-based frameworks, by means of simulations or full likelihood-based models when an analytical solution exists (but other frameworks are available, e.g., such as the diffusion approximations in δaδi).

I note that while non-implemented in my thesis (except to some extent in **Chapter 2**), an additional step should be to simulate data under the most likely scenario investigated in (3) to reproduce the descriptive results of step (1), when possible. This would thus allow to perform a sort of cross-validation of the demographic framework as a whole, as some studies have started to implement (e.g., (Corrigan et al., *in prep*)). Designing such framework is ultimately required to properly reconstruct the history of species, which is necessary to formulate proper evolutionary and conservation hypotheses.

5.3. The coalescence rate: species-to-community insights

5.3.1. Drivers of the (reconstructed) coalescence rate

The gene genealogy of a sample of lineages is determined by the demographic history of the species under investigation, therefore impacting the reconstruction of the coalescence rate achieved under panmictic (or *unstructured*) models. When the sampled lineages are not part of a panmictic species, the reconstructed coalescence rate does not relate directly to the effective size trajectory of the population from which lineages have been sampled and should in turn be referred to as the *Inverse Instantaneous Coalescence Rate* (IICR) (Mazet et al., 2016). In the end, models assuming panmixia are (generally) well performing (with several limits for recent or too old times, not detailed here) in reconstructing the vector of coalescence times underlying the gene genealogy, i.e., the IICR, which in turn yields information about the true demographic history of the sampled lineages. This makes *unstructured* models an incredibly resourceful descriptive statistic of a species history, as providing useful hints on the evolutionary processes shaping the history of the sampled lineages.

As such, it is important to uncover the determinants of the coalescence rate (and therefore the shape of the IICR) beyond the panmictic case, where the IICR is expected to vary co-jointly with changes in effective size through the history of the sampled population. However, this necessitates to couple its inspection to a thorough examination of complex demographic models. In the case of population structure, the trajectory of the IICR has traditionally been investigated in *simple* demographic scenarios, such as equilibrium island or stepping-stone models (e.g., (Mazet et al., 2015, 2016; Rodríguez et al., 2018)), notably for computational reasons. Yet, as highlighted in **Chapter 1**, the use of coalescent simulations and the current computational abilities allow nowadays to investigate more complex scenarios and therefore to better characterize the determinants of the coalescence rate trajectory.

In this context I investigated in **Chapter 2** the signatures of different set of demographic parameters within non-equilibrium meta-population models (i.e., explicitly including the colonization time of the array of the demes) on the IICR reconstructed by an *unstructured* model. Using simulations, I confirmed the presence of a spurious *bottleneck* signal on the IICR, typical of population structure and already identified in equilibrium meta-population models (Heller et al., 2013; Mazet et al., 2016; Rodríguez et al., 2018). In addition, I demonstrated that it was possible to detect the colonization time of the habitat directly from the IICR, even though its detection

remains conditioned by an intricate interplay between demographic parameters, notably the degree of connectivity and the colonization time. This result is important as most structured species are more likely to be better described by a non-equilibrium rather than equilibrium meta-population model, as the latter depicts an established meta-population, i.e., it has always, or for a very long time, been in its current range. Consequently, the ability to detect the colonization directly from the IICR conveys two major insights for the study of structured species:

(1) An ancestral increase in IICR is not (necessarily) an ancestral increase in effective size: in fact, it will be important in the future to check whether a colonization due to an expanding population leaves the same signature as a fragmentation (e.g., from a panmictic population to a meta-population);
(2) The colonization timing inferred over large geographical extent (i.e., investigating the IICR in more demes) can allow to make interpretation about the colonization process of the whole range (e.g., ultimately characterizing the dynamics of a range expansion).

This study slightly increased our understanding of how the IICR holds specific signatures of the history of the meta-population. Nonetheless, whilst the non-equilibrium meta-population model is likely more realistic than equilibrium ones in most cases, this chapter investigated a tiny window of all the processes that could influence the IICR. It will be important in the future to test in similar framework other (more) complex models to increase our understanding of how specific parameters influence the gene genealogy of a sample of lineages, and thus the trajectory of coalescence times reconstructed by *unstructured* models.

Finally, it is worth stressing that the IICR interpreted in real species (as those analyzed within my PhD) is always generated by *unstructured* models and remains associated with its load of uncertainties. It is therefore important to note that in the future, the performance of *unstructured* models for inferring IIRC in complex scenarios will need to be investigated. Typically, once source of variation arises from the (numerous) different methods to reconstruct the IICR (e.g., PSMC (H. Li & Durbin, 2011), MSMC (Schiffels & Durbin, 2014), SMC++ (Terhorst et al., 2017), Skyline Plot (Ho & Shapiro, 2011) or stairwayplot (Liu & Fu, 2020, 2015)). These methods do not infer the IICR with the same accuracy and perform differently depending on the time frame under examination. For instance, PSMC and MSMC uses linkage disequilibrium (LD) to reconstruct the ARG, Skyline and stairwayplot are full likelihood or SFS-based that consider loci as independent.

Finally, the SMC++ combines both the SFS and LD to better approximate the ARG. It is therefore important to keep in mind the source of uncertainty related to these computationally intensive approaches, which has to be accounted for when interpreting the IICR or when using it as a summary statistic.

5.3.2. The IICR as a summary statistic in practice

The reconstructed IICR provides information about the true demographic model of the sampled lineages. In practice it can be used in various ways as a summary statistic of the true gene genealogy, from the inference of effective size trajectory (when assumptions of panmixia are met), to, more broadly, a descriptive tool of the overall genetic variability, which, when analyzed at the multi-species scale, can provide information about the functioning of a community.

Chapter 2 illustrated two ways of using the IICR through the study of two widely shark species. In the tiger shark, panmictic at large scale, the IICR was used to directly interpret trends of effective population size through time. This likely represents a rare case where a species strongly tends to panmixia. In contrast, the grey reef shark is genetically structured in its whole range (the Indo-Pacific) with an organization close to bi-dimensional stepping-stone meta-population. Interestingly, this species displayed signatures of a range expansion originating close to the Coral Triangle. Investigations of the IICR at different sampling locations across its range corroborated the RE process, as colonization time was the youngest at the borders of the range distribution, and followed an incrementation when going towards the center of origin. This showed how the sole use of *unstructured* models allowed to detect an important event of the history of the species, and thus nicely illustrated the previous theoretical findings of a relationship between an ancestral increase in IICR and the colonization of the habitat. Interestingly, the Thorny Skate (**Chapter 3**) displayed an organization in two isolated meta-populations with a very high connectivity. In this case, the IICR conveyed clear signatures of the divergence between the two meta-populations, which was further confirmed by explicit demographic modelling, but demes within each meta-population displayed similar IICR, according to the extremely high migration rate inferred. Further, I used the IICR to shed more light on the supergene evolution. Indeed, the IICR trajectory is different in the supergene region compared to the genome-wide one, which suggests for the presence of selection – though its nature remains to be characterized – acting on the supergene.

This example highlights another use of the IICR as a descriptive statistic to investigate (and detect) local genomic signatures.

In **Chapter 4**, I extended the use of the IICR to a multi-species setting. I investigated 43 coral reef fish species to test whether species with large niche width are indeed less vulnerable than narrower niche width species, following the *Specialist-Generalist Variation Hypothesis* (SGVH; S. Li et al., 2014). The SGVH is based on the idea that specialists are more subject to demographic fluctuations over time due to the low availability of specific resources than generalists, who should in turn have greater demographic stability and greater overall genetic diversity (Gravel et al., 2011). To understand this, I investigated demographic stability indices as fluctuations in the IICR trajectory (among other indices) for each species and performed linear models to find whether it was related to the extent of their trophic interactions. I showed that flatness in the IICR significantly increased with the number of resources consumed and interpreted this signal as an increase in demographic stability with larger trophic niche, corroborating the expected lower vulnerability of generalists. This original approach presents another example of the usefulness of the signatures retrieved by the IICR reconstructed by *unstructured* models. It showed that even when we do not have a clue of the true demographic scenario behind the IICR trajectory, its computation is still very useful, in this case to answer an ecological inquiry. Yet this is also the principal limit to the framework used in this chapter, which thus differs from **Chapter 2** investigations where signatures in the IICR were examined using an explicit demographic scenario. Stability in the IICR could be due to a *true* stability in the deme or population considered, but as highlighted many times across this thesis, the IICR can be influenced similarly by many different scenarios and/or parameters. In this light, **Chapter 4** sets up a novel approach (as detailed more below) which represents a necessary stone to start understanding from a population genetics point of view ecological processes. In consequence, an important perspective of the chapter will be to refine the investigations notably by explicitly modelling demographic scenarios of the whole reef fish fauna (such as extinction/re-colonization processes or, bottleneck-reconnection processes to be coherent with the biogeographic and ecological hypotheses).

The different examples of use of IICR in this thesis provide two general perspectives. First, the usefulness of *unstructured* models in the *descriptive* step process, and how they can refine the interpretation of the evolutionary history of a species before explicit demographic modelling. Second, the need to implement the IICR as a summary statistic of the true demographic history of

the species under investigation. For instance, the ABC-RF framework allows the inclusion of many summary statistics expected to vary accordingly to the demographic scenario, exactly as the IICR. In addition, finding more determinants of the coalescence rate inferred from a sample of lineages (i.e., evolutionary but also ecological drivers) will be important also for understanding large-scale (i.e., ecosystem) processes. This could be key for conservation plannings, helping to better understand the determinants of vulnerability.

5.4. Multi-species population genetics: a step into the future

5.4.1. Current perspectives are multi-species perspectives

Traditionally, population genetics studies have focused on species-centered inferences. Typically, this is the case for (most of) **Chapter 2** and **Chapter 3**, which roughly aimed at reconstructing the demographic history of species at the scale of their range distribution. Yet, both of these chapters led to multi-species perspectives.

In **Chapter 2**, I investigated how life history traits in species could determine population structure, and how the latter influenced the reconstructed IICR trajectory by using *unstructured* models. Partly through a multi-species investigation, it highlighted that indeed, specific traits were related to the degree of connectivity. However, the dataset was small (i.e., four species), which clearly suggested in the future, the necessity to build extensive multi-species datasets, to uncover determinants of connectivity and genetic diversity. In the case of the Thorny Skate (**Chapter 3**), the perspective was all the more striking as it directly suggested the requirement of multi-species datasets to better characterize a local polymorphic supergene. For instance, I showed that one allele at the supergene had introgressed, but I was not able to date the time of its introgression, which would require knowledge of the *donor* species of the introgressed allele. This could be done in the future by investigating all congeneric species to identify the donor species. This shows that even micro-evolutionary processes (such as the functioning of a supergene system) might require multi-species population genetics modelling to be fully depicted. All in all, it emphasizes how multi-species investigations will become more common in the future even to investigate processes that seem species-centered at first.

5.4.2. A test-case on reef fishes from the Indo-Pacific

Multi-species datasets can be a tool to investigate large-scale processes and to investigate ecological and biogeographic hypotheses. In **Chapter 4**, I set up a 43-species genomic dataset and reconstructed for each of them the coalescence rate trajectory using an *unstructured* model. I then devised genetic indices related to demographic stability (or IICR stability) and modelled them by the number of trophic interactions as measured by meta-barcoding of gut contents. The results suggested a clear correlation between demographic fluctuations and trophic niche width, highlighting for the first time a relationship between demographic stability (especially as

reconstructed using a coalescence-based model) and a community-level process. This helped understanding the resilience of biodiversity, and thus ultimately emphasized the importance of this study for conservation.

This chapter highlighted the power of multi-species investigation to unravel important hypotheses, especially coming from another field such as community ecology. It will be possible in the future to broaden the range of ecological factors studied (i.e., investigating other determinants of the coalescence rates for example). In addition, coupling this dataset with additional sampling locations will allow for the investigation of large-scale biogeographic hypotheses. This shows how multi-species investigation in the context of population genetics will become greatly valuable to answer novel questions related to genetic diversity but also integrating a population genetics perspective to other fields of investigation.

5.4.3. Challenges in multi-species population genetics inference

While multi-species investigations in the context of population genetics is very promising, it can remain, to date, challenging. One reason is that building a dataset can be complex if it is required to compare genetic diversity between the species of interests. Here we built a unique dataset of 43 species by sequencing a set of homologous regions following a Target Gene Capture protocol, but this might not always be possible if no pre-designed baits are available. However, not all multi-species investigations do necessarily need homologous loci (such as, for example, the LHT investigation in **Chapter 2**), and the increasing affordability in *Whole Genome Sequencing* will allow in the future to have more and more model-species, enhancing the multi-species investigation possibilities. However, the main challenge of multi-species investigations resides in the fact that there are no real theoretical expectations of multi-species population genetics models, i.e., considering the effect of species interactions. Clearly, theoretical work will be needed in the next future to extend our species centered vision of population genetics. In addition, when investigating multi-species demographic signatures, the whole set of descriptive analyses discussed several times in this thesis should be perform at a species-scale. This means that the resources required for such studies will scale-up to the number of species investigated, although this could likely be eased in the future by automation methods such as deep-learning-based frameworks.

5.5. General conclusion

My PhD thesis underscores the importance of robust demographic modeling, of multi-species investigations, and stress the potential for coalescence-based statistics to provide valuable insights into species' histories and to test ecological hypothesis. I used marine organisms as test-cases, but my findings go beyond aquatic ecosystems. I emphasize the importance of establishing a meaningful demographic framework involving descriptive analyses and explicit complex demographic modelling. These are necessary for making accurate evolutionary and conservation hypotheses as highlighted by the study of widely distributed and vulnerable shark species. In addition, this was further underlined by the study of a supergene system in a skate species, where detailed demographic reconstructions were mandatory to understand the interplay of a local selective process with demography and how it related to conservation issues. Furthermore, my thesis highlights the value of multi-species population genetics investigations, revealing their potential for answering large-scale problematics: multi-species studies hold the potential to uncover ecological and biogeographic processes, providing a holistic perspective to interpret genetic diversity and its role in the resilience of biodiversity. Ultimately this thesis contributes to a deeper comprehension of how species- and community-level features drive historical demography. It underscores the intricate nature of demographic inferences in population genetics, with a focus on their broader relevance to evolutionary biology, conservation and ecology.

References

Abbott, S., & Fairbanks, D. J. (2016). Experiments on plant hybrids by Gregor Mendel. In *Genetics* (Vol. 204, Issue 2, pp. 407–422). Genetics. https://doi.org/10.1534/genetics.116.195198

Alcala, N., Streit, D., Goudet, J., & Vuilleumier, S. (2013). Peak and persistent excess of genetic diversity following an abrupt migration increase. *Genetics*, *193*(3), 953–971. https://doi.org/10.1534/genetics.112.147785

Alexander, D. H., Novembre, J., & Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome Research*, *19*(9), 1655–1664. https://doi.org/10.1101/gr.094052.109

Allen, G. R. (2008). Conservation hotspots of biodiversity and endemism for Indo-Pacific coral reef fishes. *Aquatic Conservation: Marine and Freshwater Ecosystems*, *18*(5), 541–556. https://doi.org/10.1002/aqc.880

Allsopp, M., Page, Ri., Johnston, P., & Santillo, D. (2008). *State of the World's Oceans*. Springer Science & Business Media.

Almojil, D., Cliff, G., & Spaet, J. L. Y. (2018). Weak population structure of the Spot-tail shark Carcharhinus sorrah and the Blacktip shark C. limbatus along the coasts of the Arabian Peninsula, Pakistan, and South Africa. *Ecology and Evolution*, *8*(18), 9536–9549. https://doi.org/10.1002/ece3.4468

Andrade, F. R. S., Afonso, A. S., Hazin, F. H. V., Mendonça, F. F., & Torres, R. A. (2021). Population genetics reveals global and regional history of the apex predator Galeocerdo cuvier (carcharhiniformes) with comments on mitigating shark attacks in north-eastern brazil. *Marine Ecology*, *42*(2), 1–16. https://doi.org/10.1111/maec.12640

Andrews, S. (2010). *FastQC: A Quality Control Tool for High Throughput Sequence Data* (v. 0.11.7).

Arenas, M., Ray, N., Currat, M., & Excoffier, L. (2012). Consequences of range contractions and range shifts on molecular diversity. *Molecular Biology and Evolution*, *29*(1), 207–218. https://doi.org/10.1093/molbev/msr187

Arredondo, A., Mourato, B., Nguyen, K., Boitard, S., Rodríguez, W., Noûs, C., Mazet, O., & Chikhi, L. (2021). Inferring number of populations and changes in connectivity under the n-island model. *Heredity*, *126*(6), 896–912. https://doi.org/10.1038/s41437-021-00426-9

Atta, C. J., Yuan, H., Li, C., Arcila, D., Betancur-R, R., Hughes, L. C., Ortí, G., & Tornabene, L. (2022). Exon-capture data and locus screening provide new insights into the phylogeny of flatfishes (Pleuronectoidei). *Molecular Phylogenetics and Evolution*, *166*. https://doi.org/10.1016/j.ympev.2021.107315

Aurelle, D., Pratlong, M., Oury, N., Haguenauer, A., Gélin, P., Magalon, H., Adjeroud, M., Romans, P., Vidal-Dupiol, J., Claereboudt, M., Noûs, C., Reynes, L., Toulza, E., Bonhomme, F., Mitta, G., & Pontarotti, P. (2022). Species and population genomic differentiation in Pocillopora corals (Cnidaria, Hexacorallia). *Genetica*, *150*(5), 247–262. https://doi.org/10.1007/s10709-022-00165-7

Avril, A., Purcell, J., Brelsford, A., & Chapuisat, M. (2019). Asymmetric assortative mating and queen polyandry are linked to a supergene controlling ant social organization. *Molecular Ecology*, *28*(6), 1428–1438. https://doi.org/10.1111/mec.14793

Ayala, D., Guerrero, R. F., & Kirkpatrick, M. (2013). Reproductive isolation and local adaptation quantified for a chromosome inversion in a malaria mosquito. *Evolution*, *67*(4), 946–958. https://doi.org/10.1111/j.1558-5646.2012.01836.x

Baeza, J. A., & Fuentes, M. S. (2013). Phylogeography of the shrimp palaemon floridanus (Crustacea: Caridea: Palaemonidae): A partial test of meta-population genetic structure in the wider caribbean. *Marine Ecology*, *34*(4), 381–393. https://doi.org/10.1111/maec.12038

Bailleul, D., Mackenzie, A., Sacchi, O., Poisson, F., Bierne, N., & Arnaud-Haond, S. (2018). Large-scale genetic panmixia in the blue shark (Prionace glauca): A single worldwide population, or a genetic lag-time effect of the "grey zone" of differentiation? *Evolutionary Applications*, *11*(5), 614–630. https://doi.org/10.1111/eva.12591

Baird, N. A., Etter, P. D., Atwood, T. S., Currey, M. C., Shiver, A. L., Lewis, Z. A., Selker, E. U., Cresko, W. A., & Johnson, E. A. (2008). Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS ONE*, *3*(10). https://doi.org/10.1371/journal.pone.0003376

Ballard, J. W. O., & Whitlock, M. C. (2004). The incomplete natural history of mitochondria. *Molecular Ecology*, *13*(4), 729–744. https://doi.org/10.1046/j.1365-294X.2003.02063.x

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., Lesin, V. M., Nikolenko, S. I., Pham, S., Prjibelski, A. D., Pyshkin, A. V., Sirotkin, A. V., Vyahhi, N., Tesler, G., Alekseyev, M. A., & Pevzner, P. A. (2012). SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. *Journal of Computational Biology*, *19*(5), 455–477. https://doi.org/10.1089/cmb.2012.0021

Barkley, A. N., Gollock, M., Samoilys, M., Llewellyn, F., Shivji, M., Wetherbee, B., & Hussey, N. E. (2019). Complex transboundary movements of marine megafauna in the Western Indian Ocean. *Animal Conservation*, *22*(5), 420–431. https://doi.org/10.1111/acv.12493

Barnett, A., Abrantes, K. G., Seymour, J., & Fitzpatrick, R. (2012). Residency and spatial use by reef sharks of an isolated seamount and its implications for conservation. *PLoS ONE*, *7*(5), 1–12. https://doi.org/10.1371/journal.pone.0036574

Barney, B. T., Munkholm, C., Walt, D. R., & Palumbi, S. R. (2017). Highly localized divergence within supergenes in Atlantic cod (Gadus morhua) within the Gulf of Maine. *BMC Genomics*, *18*(1). https://doi.org/10.1186/s12864-017-3660-3

Barth, J. M. I., Berg, P. R., Jonsson, P. R., Bonanomi, S., Corell, H., Hemmer-Hansen, J., Jakobsen, K. S., Johannesson, K., Jorde, P. E., Knutsen, H., Moksnes, P. O., Star, B., Stenseth, N. C., Svedäng, H., Jentoft, S., & André, C. (2017). Genome architecture enables local adaptation of Atlantic cod despite high connectivity. *Molecular Ecology*, *26*(17), 4452–4466. https://doi.org/10.1111/mec.14207

Beal, L. M., Ruijter, W. P. M. De, Biastoch, A., Zahn, R., Wcrp, S., & Working, I. (2011). On the role of the Agulhas system in ocean circulation and climate. *Nature*, *472*(7344), 429–436. https://doi.org/10.1038/nature09983

Beaumont, M. A. (2019). Approximate Bayesian computation. *Annual Review of Statistics and Its Application*, *6*, 379–403. https://doi.org/10.1146/annurev-statistics-030718-105212

Beaumont, M. A., Zhang, W., & Balding, D. J. (2002). Approximate Bayesian computation in population genetics. *Genetics*, *162*(4), 2025–2035. https://doi.org/10.1111/j.1937-2817.2010.tb01236.x

Beerli, P., & Felsenstein, J. (2001). Maximum likelihood estimation of a migration matrix and effective population sizes in n subpopulations by using a coalescent approach. *Proceedings of the National Academy of Sciences of the United States of America*, *98*(8), 4563–4568. https://doi.org/10.1073/pnas.081068098

Benazzo, A., Trucchi, E., Cahill, J. A., Delser, P. M., Mona, S., Fumagalli, M., Bunnefeld, L., Cornetti, L., Ghirotto, S., Girardi, M., Ometto, L., Panziera, A., Rota-Stabelli, O., Zanetti, E., Karamanlidis, A., Groff, C., Paule, L., Gentile, L., Vilà, C., … Bertorelle, G. (2017). Survival and divergence in a small group: The extraordinary genomic history of the endangered Apennine brown bear stragglers. *Proceedings of the National Academy of Sciences of the United States of America*, *114*(45), E9589–E9597. https://doi.org/10.1073/pnas.1707279114

Berdan, E. L., Blanckaert, A., Butlin, R. K., & Bank, C. (2021). Deleterious mutation accumulation and the long-term fate of chromosomal inversions. *PLoS Genetics*, *17*(3). https://doi.org/10.1371/journal.pgen.1009411

Berdan, E. L., Blanckaert, A., Butlin, R. K., Flatt, T., Slotte, T., & Wielstra, B. (2022). Mutation accumulation opposes polymorphism: supergenes and the curious case of balanced lethals. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *377*(1856). https://doi.org/10.1098/rstb.2021.0199

Berdan, E. L., Flatt, T., Kozak, G. M., Lotterhos, K. E., Wielstra, B., & Berdan, E. L. (2022). Genomic architecture of supergenes: Connecting form and function. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *377*(1856). https://doi.org/10.1098/rstb.2021.0192

Berg, P. R., Jentoft, S., Star, B., Ring, K. H., Knutsen, H., Lien, S., Jakobsen, K. S., & André, C. (2015). Adaptation to low salinity promotes genomic divergence in Atlantic Cod (Gadus morhua L.). *Genome Biology and Evolution*, *7*(6), 1644–1663. https://doi.org/10.1093/gbe/evv093

Berg, P. R., Star, B., Pampoulie, C., Sodeland, M., Barth, J. M. I., Knutsen, H., Jakobsen, K. S., & Jentoft, S. (2016). Three chromosomal rearrangements promote genomic divergence between migratory and stationary ecotypes of Atlantic cod. *Scientific Reports*, *6*. https://doi.org/10.1038/srep23246

Bernard, A. M., Feldheim, K. A., Heithaus, M. R., Wintner, S. P., Wetherbee, B. M., & Shivji, M. S. (2016). Global population genetic dynamics of a highly migratory, apex predator shark. *Molecular Ecology*, *25*(21), 5312–5329. https://doi.org/10.1111/mec.13845

Bernard, A. M., Finnegan, K. A., Pavinski Bitar, P., Stanhope, M. J., & Shivji, M. S. (2021). Genomic Assessment of Global Population Structure in a Highly Migratory and Habitat Versatile Apex Predator, the Tiger Shark (Galeocerdo cuvier). *Journal of Heredity*, *112*(6), 497–507. https://doi.org/10.1093/jhered/esab046

Bertorelle, G., Benazzo, A., & Mona, S. (2010). ABC as a flexible framework to estimate demography over space and time: Some cons, many pros. *Molecular Ecology*, *19*(13), 2609–2625. https://doi.org/10.1111/j.1365-294X.2010.04690.x

Boissin, E., Thorrold, S. R., Braun, C. D., Zhou, Y., Clua, E. E., & Planes, S. (2019). Contrasting global, regional and local patterns of genetic structure in gray reef shark populations from the Indo-Pacific region. *Scientific Reports*, *9*(1), 1–9. https://doi.org/10.1038/s41598-019-52221-6

Boitard, S., Rodríguez, W., Jay, F., Mona, S., & Austerlitz, F. (2016). Inferring Population Size History from Large Samples of Genome-Wide Molecular Data - An Approximate Bayesian Computation Approach. *PLoS Genetics*, *12*(3). https://doi.org/10.1371/journal.pgen.1005877

Bonnin, L., Robbins, W. D., Boussarie, G., Kiszka, J. J., Dagorn, L., Mouillot, D., & Vigliola, L. (2019). Repeated long-range migrations of adult males in a common Indo-Pacific reef shark. *Coral Reefs*, *38*(6), 1121–1132. https://doi.org/10.1007/s00338-019-01858-w

Bornatowski, H., Navia, A. F., Braga, R. R., Abilhoa, V., & Corrêa, M. F. M. (2014). Ecological importance of sharks and rays in a structural foodweb analysis in southern Brazil. *ICES Journal of Marine Science*, *71*(7), 1586–1592. https://doi.org/10.1093/icesjms/fsu025

Boussarie, G., Momigliano, P., Robbins, W. D., Bonnin, L., Cornu, J. F., Fauvelot, C., Kiszka, J. J., Manel, S., Mouillot, D., & Vigliola, L. (2022). Identifying barriers to gene flow and hierarchical conservation units from seascape genomics: a modelling framework applied to a marine predator. *Ecography*, *2022*(7), 1–14. https://doi.org/10.1111/ecog.06158

Bouwman, A. C., Daetwyler, H. D., Chamberlain, A. J., Ponce, C. H., Sargolzaei, M., Schenkel, F. S., Sahana, G., Govignon-Gion, A., Boitard, S., Dolezal, M., Pausch, H., Brøndum, R. F., Bowman, P. J., Thomsen, B., Guldbrandtsen, B., Lund, M. S., Servin, B., Garrick, D. J., Reecy, J., … Hayes, B. J. (2018). Meta-analysis of genome-wide association studies for cattle stature identifies common genes that regulate body size in mammals. *Nature Genetics*, *50*(3), 362–367. https://doi.org/10.1038/s41588-018-0056-5

Boyle, E. A., Li, Y. I., & Pritchard, J. K. (2017). An Expanded View of Complex Traits: From Polygenic to Omnigenic. In *Cell* (Vol. 169, Issue 7, pp. 1177–1186). Cell Press. https://doi.org/10.1016/j.cell.2017.05.038

Bragg, J. G., Potter, S., Bi, K., & Moritz, C. (2016). Exon capture phylogenomics: efficacy across scales of divergence. *Molecular Ecology Resources*, *16*(5), 1059–1068. https://doi.org/10.1111/1755-0998.12449

Branco, S., Carpentier, F., De La Vega, R. C. R., Badouin, H., Snirc, A., Le Prieur, S., Coelho, M. A., De Vienne, D. M., Hartmann, F. E., Begerow, D., Hood, M. E., & Giraud, T. (2018). Multiple convergent supergene evolution events in mating-type chromosomes. *Nature Communications*, *9*(1). https://doi.org/10.1038/s41467-018-04380-9

Brandrud, M. K., Baar, J., Lorenzo, M. T., Athanasiadis, A., Bateman, R. M., Chase, M. W., Hedrén, M., & Paun, O. (2020). Phylogenomic relationships of diploids and the origins of allotetraploids in Dactylorhiza (Orchidaceae). *Systematic Biology*, *69*(1), 91–109. https://doi.org/10.1093/sysbio/syz035

Brelsford, A., Purcell, J., Avril, A., Tran Van, P., Zhang, J., Brütsch, T., Sundström, L., Helanterä, H., & Chapuisat, M. (2020). An Ancient and Eroded Social Supergene Is Widespread across Formica Ants. *Current Biology*, *30*(2), 304-311.e4. https://doi.org/10.1016/j.cub.2019.11.032

Broad Institute. (2019). *Picard Toolkit*. GitHub Repository. https://broadinstitute.github.io/picard/

Bruno, J. F., & Selig, E. R. (2007). Regional decline of coral cover in the Indo-Pacific: Timing, extent, and subregional comparisons. *PLoS ONE*, *2*(8). https://doi.org/10.1371/journal.pone.0000711

Bürkner, P. C. (2021). Bayesian Item Response Modeling in R with brms and Stan. *Journal of Statistical Software*, *100*(5). https://doi.org/10.18637/JSS.V100.I05

Cahill, J. A., Soares, A. E. R., Green, R. E., & Shapiro, B. (2016). Inferring species divergence times using pairwise sequential markovian coalescent modelling and low-coverage genomic data. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *371*(1699). https://doi.org/10.1098/rstb.2015.0138

Caley, T., Giraudeau, J., Malaizé, B., Rossignol, L., & Pierre, C. (2012). Agulhas leakage as a key process in the modes of Quaternary climate changes. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(18), 6835–6839. https://doi.org/10.1073/pnas.1115545109

Caley, T., Peeters, F. J. C., Biastoch, A., Rossignol, L., van Sebille, E., Durgadoo, J., Malaizé, B., Giraudeau, J., Arthur, K., & Zahn, R. (2014). Quantitative estimate of the paleo-Agulhas

leakage. *Geophysical Research Letters*, *41*(4), 1238–1246. https://doi.org/10.1002/2014GL059278

Carmo, C. B., Ferrette, B. L. S., Camargo, S. M., Roxo, F. F., Coelho, R., Garla, R. C., Oliveira, C., Piercy, A. N., Bornatowski, H., Foresti, F., Burgess, G. H., & Mendonça, F. F. (2019). A new map of the tiger shark (<scp> *Galeocerdo cuvier* </scp>) genetic population structure in the western Atlantic Ocean: Hypothesis of an equatorial convergence centre. *Aquatic Conservation: Marine and Freshwater Ecosystems*, *29*(5), 760–772. https://doi.org/10.1002/aqc.3029

Carscadden, K. A., Emery, N. C., Arnillas, C. A., & Cadotte, M. W. (2020). NICHE BREADTH: CAUSES AND CONSEQUENCES FOR ECOLOGY, EVOLUTION, AND CONSERVATION. *The Quarterly Review of Biology*, *95*(3).

Catchen, J., Hohenlohe, P. A., Bassham, S., Amores, A., & Cresko, W. A. (2013). Stacks: an analysis tool set for population genomics. *Molecular Ecology*, *22*(11), 3124–3140. https://doi.org/10.1111/mec.12354

Ceballos, G., Ehrlich, P. R., Barnosky, A. D., García, A., Pringle, R. M., & Palmer, T. M. (2015). Accelerated modern human-induced species losses: Entering the sixth mass extinction. *Science Advances*, *1*(5), 9–13. https://doi.org/10.1126/sciadv.1400253

Chan, Y. L., Schanzenbach, D., & Hickerson, M. J. (2014). Detecting concerted demographic response across community assemblages using hierarchical approximate Bayesian computation. *Molecular Biology and Evolution*, *31*(9), 2501–2515. https://doi.org/10.1093/molbev/msu187

Chapman, D. D., Feldheim, K. A., Papastamatiou, Y. P., & Hueter, R. E. (2014). There and Back Again: A Review of Residency and Return Migrations in Sharks, with Implications for Population Structure and Management. *Annual Review of Marine Science*, *7*(1), 547–570. https://doi.org/10.1146/annurev-marine-010814-015730

Chapuisat, M. (2023). Supergenes as drivers of ant evolution. *Myrmecol. News*, *33*, 1–18. https://doi.org/10.25849/myrmecol.news_033:001

Charlesworth, B. (1996). The evolution of chromosomal sex determination and dosage compensation. *Current Biology*, *6*, 149–162.

Charlesworth, B., & Charlesworth, D. (1978). A Model for the Evolution of Dioecy and Gynodioecy. *The American Naturalist*, *112*(988), 975–997. https://about.jstor.org/terms

Charlesworth, D. (2016). The status of supergenes in the 21st century: Recombination suppression in Batesian mimicry and sex chromosomes and other complex adaptations. *Evolutionary Applications*, *9*(1), 74–90. https://doi.org/10.1111/eva.12291

Cheng, C., White, B. J., Kamdem, C., Mockaitis, K., Costantini, C., Hahn, M. W., & Besansky, N. J. (2012). Ecological genomics of anopheles gambiae along a latitudinal cline: A population-resequencing approach. *Genetics*, *190*(4), 1417–1432. https://doi.org/10.1534/genetics.111.137794

Chevolot, M., Wolfs, P. H. J., Pálsson, J., Rijnsdorp, A. D., Stam, W. T., & Olsen, J. L. (2007). Population structure and historical demography of the thorny skate (Amblyraja radiata, Rajidae) in the North Atlantic. *Marine Biology*, *151*(4), 1275–1286. https://doi.org/10.1007/s00227-006-0556-1

Chikhi, L., Rodríguez, W., Grusea, S., Santos, P., Boitard, S., & Mazet, O. (2018). The IICR (inverse instantaneous coalescence rate) as a summary of genomic diversity: Insights into demographic inference and model choice. *Heredity*, *120*(1), 13–24. https://doi.org/10.1038/s41437-017-0005-6

Chikhi, L., Sousa, V. C., Luisi, P., Goossens, B., & Beaumont, M. A. (2010). The Confounding Effects of Population Structure, Genetic Diversity and the Sampling Scheme on the Detection and Quantification of Population Size Changes. *Genetics*, *186*(3), 983–995. https://doi.org/10.1534/genetics.110.118661

Chouteau, M., Llaurens, V., Piron-Prunier, F., & Joron, M. (2017). Polymorphism at a mimicry supergene maintained by opposing frequency-dependent selection pressures. *Proceedings of the National Academy of Sciences of the United States of America*, *114*(31), 8325–8329. https://doi.org/10.1073/pnas.1702482114

Clarke, C. A., & Sheppard, P. M. (1960). SUPER-GENES AND MIMICRY. *Heredity*, *14*, 175–185.

Clarke, S. C., McAllister, M. K., Milner-Gulland, E. J., Kirkwood, G. P., Michielsens, C. G. J., Agnew, D. J., Pikitch, E. K., Nakano, H., & Shivji, M. S. (2006). Global estimates of shark catches using trade records from commercial markets. *Ecology Letters*, *9*(10), 1115–1126. https://doi.org/10.1111/j.1461-0248.2006.00968.x

Colles, A., Liow, L. H., & Prinzing, A. (2009). Are specialists at risk under environmental change? Neoecological, paleoecological and phylogenetic approaches. *Ecology Letters*, *12*(8), 849–863. https://doi.org/10.1111/j.1461-0248.2009.01336.x

Compagno, L. J. V. (1984). FAO species catalogue, Vol. 4: Sharks of the world: An annotated and illustrated catalogue of shark species known to date. Part 2 - Carcharhiniformes. In *FAO Fisheries Synopsis* (pp. 503–512).

Compagno, L. J. V. (2001). Sharks of the world. An annotated and illustrated catalogue of shark species known to date. Volume 2. Bullhead, mackerel and carpet sharks (Heterodontiformes, Lamniformes and Orectolobiformes). *FAO Species Catalogue for Fishery Purposes*, *2*(1), 108–125.

Corrigan, S., Laso-Jadart, R., Yang, L., Gay, E., Lesturgie, P., Lee, A., Fedrigo, O., Hoyos, M., Lowe, C., Lyons, K., Cliff, G., Sato, K., Tomita, T., Mona, S., & Naylor, G. (n.d.). The checkered past of the world's white shark populations. *In Prep*.

Corrigan, S., Lowther, A. D., Beheregaray, L. B., Bruce, B. D., Cliff, G., Duffy, C. A., Foulis, A., Francis, M. P., Goldsworthy, S. D., Hyde, J. R., Jabado, R. W., Kacev, D., Marshall, L., Mucientes, G. R., Naylor, G. J. P., Pepperell, J. G., Queiroz, N., White, W. T., Wintner, S. P., & Rogers, P. J. (2018). Population Connectivity of the Highly Migratory Shortfin Mako (Isurus oxyrinchus Rafinesque 1810) and Implications for Management in the Southern Hemisphere. *Frontiers in Ecology and Evolution*, *6*(NOV), 1–15. https://doi.org/10.3389/fevo.2018.00187

Corrigan, S., Maisano Delser, P., Eddy, C., Duffy, C., Yang, L., Li, C., Bazinet, A. L., Mona, S., & Naylor, G. J. P. (2017). Historical introgression drives pervasive mitochondrial admixture between two species of pelagic sharks. *Molecular Phylogenetics and Evolution*, *110*(December), 122–126. https://doi.org/10.1016/j.ympev.2017.03.011

Cortés, E. (2002). Incorporating uncertainty into demographic modeling: Application to shark populations and their conservation. *Conservation Biology*, *16*(4), 1048–1062. https://doi.org/10.1046/j.1523-1739.2002.00423.x

Cowman, P. F., & Bellwood, D. R. (2013). The historical biogeography of coral reef fishes: Global patterns of origination and dispersal. *Journal of Biogeography*, *40*(2), 209–224. https://doi.org/10.1111/jbi.12003

Csilléry, K., Blum, M. G. B., Gaggiotti, O. E., & François, O. (2010). Approximate Bayesian Computation (ABC) in practice. In *Trends in Ecology and Evolution* (Vol. 25, Issue 7, pp. 410–418). https://doi.org/10.1016/j.tree.2010.04.001

Csilléry, K., François, O., & Blum, M. G. B. (2012). Abc: An R package for approximate Bayesian computation (ABC). *Methods in Ecology and Evolution*, *3*(3), 475–479. https://doi.org/10.1111/j.2041-210X.2011.00179.x

D'Agata, S., Mouillot, D., Kulbicki, M., Andréfouët, S., Bellwood, D. R., Cinner, J. E., Cowman, P. F., Kronen, M., Pinca, S., & Vigliola, L. (2014). Human-mediated loss of phylogenetic and functional diversity in coral reef fishes. *Current Biology*, *24*(5), 555–560. https://doi.org/10.1016/j.cub.2014.01.049

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. (2011). The variant call format and VCFtools. *Bioinformatics*, *27*(15), 2156–2158. https://doi.org/10.1093/bioinformatics/btr330

Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., Whitwham, A., Keane, T., McCarthy, S. A., & Davies, R. M. (2021). Twelve years of SAMtools and BCFtools. *GigaScience*, *10*(2). https://doi.org/10.1093/gigascience/giab008

Delrieu-Trottin, E., Hubert, N., Giles, E. C., Chifflet-Belle, P., Suwalski, A., Neglia, V., Rapu-Edmunds, C., Mona, S., & Saenz-Agudelo, P. (2020). Coping with Pleistocene climatic fluctuations: Demographic responses in remote endemic reef fishes. *Molecular Ecology*, *29*(12), 2218–2233. https://doi.org/10.1111/mec.15478

Denton, J., Kneebone, J., Yang, L., Lynghammar, A., McElroy, D., Corrigan, S., Jakobsdóttir, K., Miri, C., Simpson, M., & Naylor, G. (n.d.). Mitogenomic evidence of population differentiation of thorny skate, Amblyraja radiata (Rajiformes: Rajidae) in the North Atlantic. *(In Prep)*.

Dobzhansky, T., & Sturtevant, A. H. (1938). INVERSIONS IN THE CHROMOSOMES OF DROSOPHILA PSEUDOOBSCURA. *Genetics*, *23*(1), 28–64. https://doi.org/10.1093/genetics/23.1.28

Drummond, A. J., Rambaut, A., Shapiro, B., & Pybus, O. G. (2005). Bayesian coalescent inference of past population dynamics from molecular sequences. *Molecular Biology and Evolution*, *22*(5), 1185–1192. https://doi.org/10.1093/molbev/msi103

Dudgeon, C. L., Blower, D. C., Broderick, D., Giles, J. L., Holmes, B. J., Kashiwagi, T., Krück, N. C., Morgan, J. A. T., Tillett, B. J., & Ovenden, J. R. (2012). A review of the application of molecular genetics for fisheries management and conservation of sharks and rays. *Journal of Fish Biology*, *80*(5), 1789–1843. https://doi.org/10.1111/j.1095-8649.2012.03265.x

Dulvy, N. K., Fowler, S. L., Musick, J. A., Cavanagh, R. D., Kyne, P. M., Harrison, L. R., Carlson, J. K., Davidson, L. N. K., Fordham, S. V., Francis, M. P., Pollock, C. M., Simpfendorfer, C. A., Burgess, G. H., Carpenter, K. E., Compagno, L. J. V., Ebert, D. A., Gibson, C., Heupel, M. R., Livingstone, S. R., … White, W. T. (2014). Extinction risk and conservation of the world's sharks and rays. *ELife*, *2014*(3), 1–34. https://doi.org/10.7554/eLife.00590.001

Dulvy, N. K., Pacoureau, N., Rigby, C. L., Pollom, R. A., Jabado, R. W., Ebert, D. A., Finucci, B., Pollock, C. M., Cheok, J., Derrick, D. H., Herman, K. B., Sherman, C. S., VanderWright, W. J., Lawson, J. M., Walls, R. H. L., Carlson, J. K., Charvet, P., Bineesh, K. K., Fernando, D., … Simpfendorfer, C. A. (2021). Overfishing drives over one-third of all sharks and rays toward a global extinction crisis. *Current Biology*, 1–15. https://doi.org/10.1016/j.cub.2021.08.062

Dwyer, R. G., Krueck, N. C., Udyawer, V., Heupel, M. R., Chapman, D., Pratt, H. L., Garla, R., & Simpfendorfer, C. A. (2020). Individual and Population Benefits of Marine Reserves for Reef Sharks. *Current Biology*, *30*(3), 480-489.e5. https://doi.org/10.1016/j.cub.2019.12.005

Eaton, D. A. R. (2014). PyRAD: Assembly of de novo RADseq loci for phylogenetic analyses. *Bioinformatics*, *30*(13), 1844–1849. https://doi.org/10.1093/bioinformatics/btu121

Ellegren, H., & Galtier, N. (2016). Determinants of genetic diversity. In *Nature Reviews Genetics* (Vol. 17, Issue 7, pp. 422–433). Nature Publishing Group. https://doi.org/10.1038/nrg.2016.58

Errbii, M., Ernst, U. R., Lajmi, A., Privman, E., Gadau, J., & Schrader, L. (2023). Evolutionary genomics of socially polymorphic populations of Pogonomyrmex californicus. *BioRxiv*. https://doi.org/10.1101/2021.03.21.436260

Espinoza, M., Heupel, M. R., Tobin, A. J., & Simpfendorfer, C. A. (2014). Residency patterns and movements of grey reef sharks (Carcharhinus amblyrhynchos) in semi-isolated coral reef habitats. *Marine Biology*, *162*(2), 343–358. https://doi.org/10.1007/s00227-014-2572-x

Excoffier, L. (2004). Patterns of DNA sequence diversity and genetic structure after a range expansion: Lessons from the infinite-island model. *Molecular Ecology*, *13*(4), 853–864. https://doi.org/10.1046/j.1365-294X.2003.02004.x

Excoffier, L., Dupanloup, I., Huerta-Sánchez, E., Sousa, V. C., & Foll, M. (2013). Robust Demographic Inference from Genomic and SNP Data. *PLoS Genetics*, *9*(10), e1003905. https://doi.org/10.1371/journal.pgen.1003905

Excoffier, L., & Foll, M. (2011). fastsimcoal: A continuous-time coalescent simulator of genomic diversity under arbitrarily complex evolutionary scenarios. *Bioinformatics*, *27*(9), 1332–1334. https://doi.org/10.1093/bioinformatics/btr124

Excoffier, L., Foll, M., & Petit, R. J. (2009). Genetic Consequences of Range Expansions. *Annual Review of Ecology, Evolution, and Systematics*, *40*(1), 481–501. https://doi.org/10.1146/annurev.ecolsys.39.110707.173414

Excoffier, L., Marchi, N., Marques, D. A., Matthey-Doret, R., Gouy, A., & Sousa, V. C. (2021). Fastsimcoal2: Demographic inference under complex evolutionary scenarios. *Bioinformatics*, *37*(24), 4882–4885. https://doi.org/10.1093/bioinformatics/btab468

Faria, R., Johannesson, K., Butlin, R. K., & Westram, A. M. (2019). Evolving Inversions. In *Trends in Ecology and Evolution* (Vol. 34, Issue 3, pp. 239–248). Elsevier Ltd. https://doi.org/10.1016/j.tree.2018.12.005

Felsenstein, J. (1974). THE EVOLUTIONARY ADVANTAGE OF RECOMBINATION'. *Genetics*, *78*, 737–756.

Felsenstein, J. (2006). Accuracy of coalescent likelihood estimates: Do we need more sites, more sequences, or more loci? *Molecular Biology and Evolution*, *23*(3), 691–700. https://doi.org/10.1093/molbev/msj079

Field, I. C., Meekan, M. G., Speed, C. W., White, W., & Bradshaw, C. J. A. (2011). Quantifying movement patterns for shark conservation at remote coral atolls in the Indian Ocean. *Coral Reefs*, *30*(1), 61–71. https://doi.org/10.1007/s00338-010-0699-x

Fisher, R. A. (1918). The correlation between relatives on the supposition of Mendelian inheritance. *Earth and Environmental Science Transactions of The Royal Society of Edinburgh*, *52*(2), 399–433.

Ford, E. B. (1966). *A Symposium from Mendel's Factors to the Genetic Code* (Vol. 164, Issue 995). https://about.jstor.org/terms

Fountain, E. D., Pauli, J. N., Reid, B. N., Palsbøll, P. J., & Peery, M. Z. (2016). Finding the right coverage: the impact of coverage and sequence quality on single nucleotide polymorphism genotyping error rates. *Molecular Ecology Resources*, *16*(4), 966–978. https://doi.org/10.1111/1755-0998.12519

Frichot, E., & François, O. (2015). LEA: An R package for landscape and ecological association studies. *Methods in Ecology and Evolution*, *6*(8), 925–929. https://doi.org/10.1111/2041-210X.12382

Frichot, E., Mathieu, F., Trouillon, T., Bouchard, G., & François, O. (2014). Fast and efficient estimation of individual ancestry coefficients. *Genetics*, *196*(4), 973–983. https://doi.org/10.1534/genetics.113.160572

Friedlander, A. M., & DeMartini, E. E. (2002). Contrasts in density, size, and biomass of reef fishes between the northwestern and the main Hawaiian islands: The effects of fishing down apex predators. *Marine Ecology Progress Series*, *230*, 253–264. https://doi.org/10.3354/meps230253

Gaither, M. R., Bowen, B. W., Rocha, L. A., & Briggs, J. C. (2016). Fishes that rule the world: circumtropical distributions revisited. *Fish and Fisheries*, *17*(3), 664–679. https://doi.org/10.1111/faf.12136

Gledhill, K. S., Kessel, S. T., Guttridge, T. L., Hansell, A. C., Bester-van der Merwe, A. E., Feldheim, K. A., Gruber, S. H., & Chapman, D. D. (2015). Genetic structure, population demography and seasonal occurrence of blacktip shark Carcharhinus limbatus in Bimini, the Bahamas. *Journal of Fish Biology*, *87*(6), 1371–1388. https://doi.org/10.1111/jfb.12821

Goudet, J. (2005). HIERFSTAT , a package for R to compute and test hierarchical F-statistics. *Molecular Ecology Notes*, *5*, 184–186. https://doi.org/10.1111/j.1471-8278

Graffelman, J. (2015). *Journal of Statistical Software Exploring Diallelic Genetic Markers: The HardyWeinberg Package* (Vol. 64). http://www.jstatsoft.org/

Gravel, D., Massol, F., Canard, E., Mouillot, D., & Mouquet, N. (2011). Trophic theory of island biogeography. *Ecology Letters*, *14*(10), 1010–1016. https://doi.org/10.1111/j.1461-0248.2011.01667.x

Griffiths, R., & Marjoram, P. (1997). An ancestral recombination graph. *Progress in Population Genetics and Human Evolution* , *87*, 257–270.

Gutiérrez-Valencia, J., Hughes, P. W., Berdan, E. L., & Slotte, T. (2021). The Genomic Architecture and Evolutionary Fates of Supergenes. *Genome Biology and Evolution*, *13*(5). https://doi.org/10.1093/gbe/evab057

Hamilton, G., Stoneking, M., & Excoffier, L. (2005). Molecular analysis reveals tighter social regulation of immigration in patrilocal populations than in matrilocal populations. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(21), 7476–7480. https://doi.org/10.1073/pnas.0409253102

Hedgecock, D., & Pudovkin, A. I. (2011). Sweepstakes reproductive success in highly fecund marine fish and shellfish : A review and commentary. *Bulletin of Marine Science*, *87*(4), 971–1002. https://doi.org/10.5343/bms.2010.1051

Hedrick, P. W. (2007). Balancing selection. *Current Biology*, *17*(7).

Heled, J., & Drummond, A. J. (2008). Bayesian inference of population size history from multiple loci. *BMC Evolutionary Biology*, *8*(1), 289. https://doi.org/10.1186/1471-2148-8-289

Heller, R., Chikhi, L., & Siegismund, H. R. (2013). The Confounding Effect of Population Structure on Bayesian Skyline Plot Inferences of Demographic History. *PLoS ONE*, *8*(5). https://doi.org/10.1371/journal.pone.0062992

Heller, R., Nursyifa, C., Garcia-Erill, G., Salmona, J., Chikhi, L., Meisner, J., Korneliussen, T. S., & Albrechtsen, A. (2021). A reference-free approach to analyse RADseq data using standard next generation sequencing toolkits. *Molecular Ecology Resources*, *21*(4), 1085–1097. https://doi.org/10.1111/1755-0998.13324

Helleu, Q., Roux, C., Ross, K. G., & Keller, L. (2022). *Radiation and hybridization underpin the spread of the fire ant social supergene*. https://doi.org/10.1073/pnas

Hickerson, M. J., & Meyer, C. P. (2008). Testing comparative phylogeographic models of marine vicariance and dispersal using a hierarchical Bayesian approach. *BMC Evolutionary Biology*, *8*(1). https://doi.org/10.1186/1471-2148-8-322

Hijmans, R. J. (2020). *Raster: Geographic Data Analysis and Modeling.* (R package version 3.0-12.).

Ho, S. Y. W., Duchêne, S., Molak, M., & Shapiro, B. (2015). Time-dependent estimates of molecular evolutionary rates: Evidence and causes. *Molecular Ecology*, *24*(24), 6007–6012. https://doi.org/10.1111/mec.13450

Ho, S. Y. W., & Shapiro, B. (2011). Skyline-plot methods for estimating demographic history from nucleotide sequences. In *Molecular Ecology Resources* (Vol. 11, Issue 3, pp. 423–434). https://doi.org/10.1111/j.1755-0998.2011.02988.x

Hoban, S., Bertorelle, G., & Gaggiotti, O. E. (2012). Computer simulations: Tools for population and evolutionary genetics. *Nature Reviews Genetics*, *13*(2), 110–122. https://doi.org/10.1038/nrg3130

Hoffmann, A. A., & Rieseberg, L. H. (2008). Revisiting the impact of inversions in evolution: From population genetic markers to drivers of adaptive shifts and speciation? In *Annual Review of Ecology, Evolution, and Systematics* (Vol. 39, pp. 21–42). https://doi.org/10.1146/annurev.ecolsys.39.110707.173532

Holland, K. N., Anderson, J. M., Coffey, D. M., Holmes, B. J., Meyer, C. G., & Royer, M. A. (2019). A Perspective on Future Tiger Shark Research. *Frontiers in Marine Science*, *6*(FEB), 1–7. https://doi.org/10.3389/fmars.2019.00037

Holmes, B. J., Williams, S. M., Otway, N. M., Nielsen, E. E., Maher, S. L., Bennett, M. B., & Ovenden, J. R. (2017). Population structure and connectivity of tiger sharks (Galeocerdo cuvier) across the Indo-Pacific Ocean basin. *Royal Society Open Science*, *4*(7), 170309. https://doi.org/10.1098/rsos.170309

Hudson, R. R. (1983). Properties of a Neutral Allele Model with lntragenic Recombination. In *POPULATION BIOLOGY* (Vol. 23).

Hudson, R. R. (1991). Gene genealogies and the coalescent process. *Oxford Surveys in Evolutionary Biology*, *7*, 1–44.

Hudson, R. R. (2002). Generating samples under a Wright-Fisher neutral model of genetic variation. In *BIOINFORMATICS APPLICATIONS NOTE* (Vol. 18, Issue 2). http://home.uchicago.edu/~rhudson1/source/mksamples.html.

Hudson, R. R., Slatkin, M., & Maddison, W. P. (1992). Estimation of levels of gene flow from DNA sequence data. *Genetics*, *132*(2), 583–589. https://doi.org/10.1093/genetics/132.2.583

Jay, P., Aubier, T. G., & Joron, M. (2020). Admixture can readily lead to the formation of supergenes. *BioRxiv*. https://doi.org/10.1101/2020.11.19.389577

Jiang, J., Yuan, H., Zheng, X., Wang, Q., Kuang, T., Li, J., Liu, J., Song, S., Wang, W., Cheng, F., Li, H., Huang, J., & Li, C. (2019). Gene markers for exon capture and phylogenomics in ray-finned fishes. *Ecology and Evolution*, *9*(7), 3973–3983. https://doi.org/10.1002/ece3.5026

Jombart, T., Devillard, S., & Balloux, F. (2010). Discriminant analysis of principal components: A new method for the analysis of genetically structured populations. *BMC Genetics*, *11*(1), 94. https://doi.org/10.1186/1471-2156-11-94

Jones, F. C., Grabherr, M. G., Chan, Y. F., Russell, P., Mauceli, E., Johnson, J., Swofford, R., Pirun, M., Zody, M. C., White, S., Birney, E., Searle, S., Schmutz, J., Grimwood, J., Dickson, M. C., Myers, R. M., Miller, C. T., Summers, B. R., Knecht, A. K., … Kingsley, D. M. (2012). The genomic basis of adaptive evolution in threespine sticklebacks. *Nature*, *484*(7392), 55–61. https://doi.org/10.1038/nature10944

Jones, M. R., & Good, J. M. (2016). Targeted capture in evolutionary and ecological genomics. In *Molecular Ecology* (Vol. 25, Issue 1, pp. 185–202). Blackwell Publishing Ltd. https://doi.org/10.1111/mec.13304

Jones, R. T., Salazar, P. A., Ffrench-Constant, R. H., Jiggins, C. D., & Joron, M. (2012). Evolution of a mimicry supergene from a multilocus architecture. *Proceedings of the Royal Society B: Biological Sciences*, *279*(1727), 316–325. https://doi.org/10.1098/rspb.2011.0882

Joron, M., Frezal, L., Jones, R. T., Chamberlain, N. L., Lee, S. F., Haag, C. R., Whibley, A., Becuwe, M., Baxter, S. W., Ferguson, L., Wilkinson, P. A., Salazar, C., Davidson, C., Clark, R., Quail, M. A., Beasley, H., Glithero, R., Lloyd, C., Sims, S., … Ffrench-Constant, R. H. (2011). Chromosomal rearrangements maintain a polymorphic supergene controlling butterfly mimicry. *Nature*, *477*(7363), 203–206. https://doi.org/10.1038/nature10341

Joron, M., Papa, R., Beltrán, M., Chamberlain, N., Mavárez, J., Baxter, S., Abanto, M., Bermingham, E., Humphray, S. J., Rogers, J., Beasley, H., Barlow, K., Ffrench-Constant, R. H., Mallet, J., McMillan, W. O., & Jiggins, C. D. (2006). A conserved supergene locus controls colour pattern diversity in Heliconius butterflies. *PLoS Biology*, *4*(10), 1831–1840. https://doi.org/10.1371/journal.pbio.0040303

Joron, M., Wynne, I. R., Lamas, G., & Mallet, J. (1999). Variable Selection and the Coexistence of Multiple mimetic forms of the Butterfly Heliconius numata. *Evolutionary Ecology*, *13*(7–8), 721–754. https://doi.org/10.1023/A:1010875213123

Kajiura, S. M., & Tellman, S. L. (2016). Quantification of massive seasonal aggregations of blacktip sharks (Carcharhinus limbatus) in southeast Florida. *PLoS ONE*, *11*(3), 1–16. https://doi.org/10.1371/journal.pone.0150911

Kamvar, Z. N., Brooks, J. C., & Grünwald, N. J. (2015). Novel R tools for analysis of genome-wide population genetic data with emphasis on clonality. *Frontiers in Genetics*, *6*(JUN). https://doi.org/10.3389/fgene.2015.00208

Karl, S. A., Motta, P. J., Stewart, B. S., Wilson, S. G., Bowen, B. W., Castro, A. L. F., Meekan, M. G., & Hueter, R. E. (2010). Population genetic structure of Earth's largest fish, the whale shark (Rhincodon typus). *Molecular Ecology*, *16*(24), 5183–5192. https://doi.org/10.1111/j.1365-294x.2007.03597.x

Kay, T., Helleu, Q., & Keller, L. (2022). Iterative evolution of supergene-based social polymorphism in ants. In *Philosophical Transactions of the Royal Society B: Biological Sciences* (Vol. 377, Issue 1856). Royal Society Publishing. https://doi.org/10.1098/rstb.2021.0196

Kebaïli, C., Sherpa, S., Rioux, D., & Després, L. (2022). Demographic inferences and climatic niche modelling shed light on the evolutionary history of the emblematic cold-adapted Apollo butterfly at regional scale. *Molecular Ecology*, *31*(2), 448–466. https://doi.org/10.1111/mec.16244

Keeney, D. B., Heupel, M., Hueter, R. E., & Heist, E. J. (2003). Genetic heterogeneity among blacktip shark, Carcharhinus limbatus, continental nurseries along the U.S. Atlantic and Gulf of Mexico. *Marine Biology*, *143*(6), 1039–1046. https://doi.org/10.1007/s00227-003-1166-9

Keeney, D. B., Heupel, M. R., Hueter, R. E., & Heist, E. J. (2005). Microsatellite and mitochondrial DNA analyses of the genetic structure of blacktip shark (Carcharhinus limbatus) nurseries in the northwestern Atlantic, Gulf of Mexico, and Caribbean Sea. *Molecular Ecology*, *14*(7), 1911–1923. https://doi.org/10.1111/j.1365-294X.2005.02549.x

Kelleher, J., Etheridge, A. M., & McVean, G. (2016). Efficient Coalescent Simulation and Genealogical Analysis for Large Sample Sizes. *PLoS Computational Biology*, *12*(5). https://doi.org/10.1371/journal.pcbi.1004842

Kerdoncuff, E., Lambert, A., & Achaz, G. (2020). Testing for population decline using maximal linkage disequilibrium blocks. *Theoretical Population Biology*, *134*, 171–181. https://doi.org/10.1016/j.tpb.2020.03.004

Khimoun, A., Doums, C., Molet, M., Kaufmann, B., Peronnet, R., Eyer, P. A., & Mona, S. (2020). Urbanization without isolation: The absence of genetic structure among cities and forests in the tiny acorn ant Temnothorax nylanderi. *Biology Letters*, *16*(1). https://doi.org/10.1098/rsbl.2019.0741

Kim, K. W., Bennison, C., Hemmings, N., Brookes, L., Hurley, L. L., Griffith, S. C., Burke, T., Birkhead, T. R., & Slate, J. (2017). A sex-linked supergene controls sperm morphology and swimming speed in a songbird. *Nature Ecology and Evolution*, *1*(8), 1168–1176. https://doi.org/10.1038/s41559-017-0235-2

Kimura, M. (1968). Evolutionary Rate at the Molecular Level. *Nature*, *217*, 624–626.

Kimura, M. (1969). THE NUMBER OF HETEROZYGOUS NUCLEOTIDE SITES MAINTAINED IN A FINITE POPULATION DUE TO STEADY FLUX OF MUTATIONS. *Genetics*, *61*, 893–903.

Kimura, M., & Crow, J. F. (1963). The Measurement of Effective Population Number. In *Source: Evolution* (Vol. 17, Issue 3).

Kimura, M., & Crow, J. F. (1964). THE NUMBER OF ALLELES THAT CAN BE MAINTAINED IN A FINITE POPULATION'. *Genetics*, *49*, 725–738.

Kimura, M., & Weiss, G. H. (1964). THE STEPPING STONE MODEL OF POPULATION STRUCTURE AND THE DECREASE OF GENETIC CORRELATION WITH DISTANCE. *Genetics*, *49*, 561–576.

Kingman, J. F. C. (1982). The coalescent. *Stochastic Processes and Their Applications*, *13*(3), 235–248. https://doi.org/10.1016/0304-4149(82)90011-4

Kirkpatrick, M. (2010). How and why chromosome inversions evolve. *PLoS Biology*, *8*(9). https://doi.org/10.1371/journal.pbio.1000501

Klopfstein, S., Currat, M., & Excoffier, L. (2006). The fate of mutations surfing on the wave of a range expansion. *Molecular Biology and Evolution*, *23*(3), 482–490. https://doi.org/10.1093/molbev/msj057

Korneliussen, T. S., Albrechtsen, A., & Nielsen, R. (2014). ANGSD: Analysis of Next Generation Sequencing Data. *BMC Bioinformatics*, *15*(1), 1–13. https://doi.org/10.1186/s12859-014-0356-4

Kozak, G. M., Wadsworth, C. B., Kahne, S. C., Bogdanowicz, S. M., Harrison, R. G., Coates, B. S., & Dopman, E. B. (2017). A combination of sexual and ecological divergence contributes to rearrangement spread during initial stages of speciation. *Molecular Ecology*, *26*(8), 2331–2347. https://doi.org/10.1111/mec.14036

Kulka, D. W., Ellis, J., Anderson, B., Cotton, C. F., Derrick, D., Pacoureau, N., & Dulvy, N. K. (2020). Amblyraja radiata. *The IUCN Red List of Threatened Species* .

Lagunas-Robles, G., Purcell, J., & Brelsford, A. (2021). Linked supergenes underlie split sex ratio and social organization in an ant. *Proceedings of the National Academy of Sciences*, *118*(46). https://doi.org/10.1073/pnas.2101427118

Lapierre, M., Lambert, A., & Achaz, G. (2017). Accuracy of Demographic Inferences from the Site Frequency Spectrum: The Case of the Yoruba Population. *Genetics*, *206*(1), 439–449. https://doi.org/10.1534/genetics.116.192708

Last, P. R., White, M. de C., Séret, B., Stehmann, M., & Naylor, G. (2016). *Rays of the world*. CSIRO PUBLISHING.

Lea, J. S. E., Wetherbee, B. M., Queiroz, N., Burnie, N., Aming, C., Sousa, L. L., Mucientes, G. R., Humphries, N. E., Harvey, G. M., Sims, D. W., & Shivji, M. S. (2015). Repeated, long-distance migrations by a philopatric predator targeting highly contrasting ecosystems. *Scientific Reports*, *5*(1), 11202. https://doi.org/10.1038/srep11202

Lee, C. R., Wang, B., Mojica, J. P., Mandáková, T., Prasad, K. V. S. K., Goicoechea, J. L., Perera, N., Hellsten, U., Hundley, H. N., Johnson, J., Grimwood, J., Barry, K., Fairclough, T., Jenkins, J. W., Yu, Y., Kudrna, D., Zhang, J., Talag, J., Golser, W., … Mitchell-Olds, T. (2017). Young inversion with multiple linked QTLs under selection in a hybrid zone. *Nature Ecology and Evolution*, *1*(5). https://doi.org/10.1038/s41559-017-0119

Lee-Yaw, J. A., Irwin, J. T., & Green, D. M. (2008). Postglacial range expansion from northern refugia by the wood frog, Rana sylvatica. *Molecular Ecology*, *17*(3), 867–884. https://doi.org/10.1111/j.1365-294X.2007.03611.x

Lesturgie, P., Lainé, H., Asuwalski, A., Chifflet-Belle, P., Maisano Delser, P., Magalon, H., & Mona, S. (2022). Life history traits and biogeographic features shaped the complex evolutionary history of an iconic apex predator (Galeocerdo cuvier). *Research Square*, 1–23. https://doi.org/10.21203/rs.3.rs-1635778/v1

Lesturgie, P., Planes, S., & Mona, S. (2022). Coalescence times, life history traits and conservation concerns: An example from four coastal shark species from the Indo-Pacific. *Molecular Ecology Resources*, *22*(2), 554–566. https://doi.org/10.1111/1755-0998.13487

Levene, H., Levins, R., & Macarthur, R. (1966). The Maintenance of Genetic Polymorphism in a Spatially Heterogeneous Environment: Variations on a Theme by Howard Levene. *The American Naturalist*, *100*(916), 585–589. https://www.jstor.org/stable/2459296

Li, C., Hofreiter, M., Straube, N., Corrigan, S., & Naylor, G. J. P. (2013). Capturing protein-coding genes across highly divergent species. *BioTechniques*, *54*(6), 321–326. https://doi.org/10.2144/000114039

Li, H. (2013). *Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM*. *00*(00), 1–3. http://arxiv.org/abs/1303.3997

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, *25*(14), 1754–1760. https://doi.org/10.1093/bioinformatics/btp324

Li, H., & Durbin, R. (2011). Inference of human population history from individual whole-genome sequences. *Nature*, *475*(7357), 493–496. https://doi.org/10.1038/nature10231

Li, H., He, Y., Jiang, J., Liu, Z., & Li, C. (2018). Molecular systematics and phylogenetic analysis of the Asian endemic freshwater sleepers (Gobiiformes: Odontobutidae). *Molecular Phylogenetics and Evolution*, *121*(August 2017), 1–11. https://doi.org/10.1016/j.ympev.2017.12.026

Li, S., Jovelin, R., Yoshiga, T., Tanaka, R., & Cutter, A. D. (2014). Specialist versus generalist life histories and nucleotide diversity in Caenorhabditis nematodes. *Proceedings of the Royal Society B: Biological Sciences*, *281*(1777). https://doi.org/10.1098/rspb.2013.2858

Lindtke, D., Lucek, K., Soria-Carrasco, V., Villoutreix, R., Farkas, T. E., Riesch, R., Dennis, S. R., Gompert, Z., & Nosil, P. (2017). Long-term balancing selection on chromosomal variants associated with crypsis in a stick insect. *Molecular Ecology*, *26*(22), 6189–6205. https://doi.org/10.1111/mec.14280

Liu, X., & Fu, Y. X. (2020). Stairwayplot 2: demographic history inference with folded SNP frequency spectra. *Genome Biology*, *21*(1), 1–9. https://doi.org/10.1186/s13059-020-02196-9

Liu, X., & Fu, Y.-X. (2015). Exploring population size changes using SNP frequency spectra. *Nature Genetics*, *47*(5), 555–559. https://doi.org/10.1038/ng.3254

Llaurens, V., Whibley, A., & Joron, M. (2017). Genetic architecture and balancing selection: the life and death of differentiated variants. In *Molecular Ecology* (Vol. 26, Issue 9, pp. 2430–2448). Blackwell Publishing Ltd. https://doi.org/10.1111/mec.14051

Lopes, A., Demarchi, L. O., Piedade, M. T. F., Schöngart, J., Wittmann, F., Munhoz, C. B. R., Ferreira, C. S., & Franco, A. C. (2023). Predicting the range expansion of invasive alien grasses under climate change in the Neotropics. *Perspectives in Ecology and Conservation*, *21*(2), 128–135. https://doi.org/10.1016/j.pecon.2023.02.005

Lynch, M. (2009). Estimation of allele frequencies from high-coverage genome-sequencing projects. *Genetics*, *182*(1), 295–301. https://doi.org/10.1534/genetics.109.100479

Lynghammar, A., Præbel, K., Bhat, S., Fevolden, S., & Christiansen, J. (2016). Widespread physical mixing of starry ray from differentiated populations and life histories in the North Atlantic. *Marine Ecology Progress Series*, *562*, 123–134. https://doi.org/10.3354/meps11958

MacArthur, R. H., & Wilson, E. O. (1967). The Theory of Island Biogeography. *Princeton University Press*.

MacNeil, M. A., Chapman, D. D., Heupel, M., Simpfendorfer, C. A., Heithaus, M., Meekan, M., Harvey, E., Goetze, J., Kiszka, J., Bond, M. E., Currey-Randall, L. M., Speed, C. W., Sherman, C. S., Rees, M. J., Udyawer, V., Flowers, K. I., Clementi, G., Valentin-Albanese, J., Gorham, T., … Cinner, J. E. (2020). Global status and conservation potential of reef sharks. *Nature*, *583*(7818), 801–806. https://doi.org/10.1038/s41586-020-2519-y

Maduna, S. N., Rossouw, C., da Silva, C., Soekoe, M., & Bester-van der Merwe, A. E. (2017). Species identification and comparative population genetics of four coastal houndsharks based on novel NGS-mined microsatellites. *Ecology and Evolution*, *7*(5), 1462–1486. https://doi.org/10.1002/ece3.2770

Maier, P. A., Vandergast, A. G., Ostoja, S. M., Aguilar, A., & Bohonak, A. J. (2022). Landscape genetics of a sub-alpine toad: climate change predicted to induce upward range shifts via asymmetrical migration corridors. *Heredity*, *129*(5), 257–272. https://doi.org/10.1038/s41437-022-00561-x

Maisano Delser, P., Corrigan, S., Duckett, D., Suwalski, A., Veuille, M., Planes, S., Naylor, G. J. P., & Mona, S. (2019). Demographic inferences after a range expansion can be biased: the test case of the blacktip reef shark (Carcharhinus melanopterus). *Heredity*, *122*(6), 759–769. https://doi.org/10.1038/s41437-018-0164-0

Maisano Delser, P., Corrigan, S., Hale, M., Li, C., Veuille, M., Planes, S., Naylor, G., & Mona, S. (2016). Population genomics of C. melanopterus using target gene capture data: Demographic

inferences and conservation perspectives. *Scientific Reports*, *6*(April), 1–12. https://doi.org/10.1038/srep33753

Mantel, N. (1967). The Detection of Disease Clustering and a Generalized Regression Approach. *Cancer Research*, *27*(2), 209–220.

Marçais, G., Delcher, A. L., Phillippy, A. M., Coston, R., Salzberg, S. L., & Zimin, A. (2018). MUMmer4: A fast and versatile genome alignment system. *PLoS Computational Biology*, *14*(1). https://doi.org/10.1371/journal.pcbi.1005944

Marchinko, K. B., Matthews, B., Arnegard, M. E., Rogers, S. M., & Schluter, D. (2014). Maintenance of a genetic polymorphism with disruptive natural selection in stickleback. *Current Biology*, *24*(11), 1289–1292. https://doi.org/10.1016/j.cub.2014.04.026

Marjoram, P., & Wall, J. D. (2006). Fast "coalescent" simulation. *BMC Genetics*, *7*. https://doi.org/10.1186/1471-2156-7-16

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.Journal*, *17*(1), 10. https://doi.org/10.14806/ej.17.1.200

Martin, S. B., & Leberg, P. L. (2011). Influence of environmental stress on age- and size-at-maturity: Genetic and plastic responses of coastal marsh fishes to changing salinities. *Canadian Journal of Fisheries and Aquatic Sciences*, *68*(12), 2121–2131. https://doi.org/10.1139/F2011-119

Mastretta-Yanes, A., Arrigo, N., Alvarez, N., Jorgensen, T. H., Piñero, D., & Emerson, B. C. (2015). Restriction site-associated DNA sequencing, genotyping error estimation and de novo assembly optimization for population genetic inference. *Molecular Ecology Resources*, *15*(1), 28–41. https://doi.org/10.1111/1755-0998.12291

Mather, K. (1941). Variation and Selection of Polygenic Characters. *Journal of Genetics*, *41*, 159–193.

Matschiner, M., Barth, J. M. I., Tørresen, O. K., Star, B., Baalsrud, H. T., Brieuc, M. S. O., Pampoulie, C., Bradbury, I., Jakobsen, K. S., & Jentoft, S. (2022). Supergene origin and maintenance in Atlantic cod. *Nature Ecology and Evolution*, *6*(4), 469–481. https://doi.org/10.1038/s41559-022-01661-x

Matthee, C. A. (2020). The Influence of Host Dispersal on the Gene Flow and Genetic Diversity of Generalist and Specialist Ectoparasites. *African Zoology*, *55*(2), 119–126. https://doi.org/10.1080/15627020.2020.1762512

Matthee, C. A., Engelbrecht, A., & Matthee, S. (2018). Comparative phylogeography of parasitic Laelaps mites contribute new insights into the specialist-generalist variation hypothesis (SGVH). *BMC Evolutionary Biology*, *18*(1), 1–11. https://doi.org/10.1186/s12862-018-1245-7

Mazet, O., Rodríguez, W., & Chikhi, L. (2015). Demographic inference using genetic data from a single individual: Separating population size variation from population structure. *Theoretical Population Biology*, *104*, 46–58. https://doi.org/10.1016/j.tpb.2015.06.003

Mazet, O., Rodríguez, W., Grusea, S., Boitard, S., & Chikhi, L. (2016). On the importance of being structured: instantaneous coalescence rates and human evolution—lessons for ancestral population size inference? *Heredity*, *116*(4), 362–371. https://doi.org/10.1038/hdy.2015.104

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., & DePristo, M. A. (2010). The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, *20*(9), 1297–1303. https://doi.org/10.1101/gr.107524.110

Mcphie, R. P., & Campana, S. E. (2009). Reproductive characteristics and population decline of four species of skate (Rajidae) off the eastern coast of Canada. *Journal of Fish Biology*, *75*(1), 223–246. https://doi.org/10.1111/j.1095-8649.2009.02282.x

Mcrae, B. H. (2006). ISOLATION BY RESISTANCE. In *Evolution* (Vol. 60, Issue 8). https://academic.oup.com/evolut/article/60/8/1551/6756311

McVean, G. A. T., & Cardin, N. J. (2005). Approximating the coalescent with recombination. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *360*(1459), 1387–1393. https://doi.org/10.1098/rstb.2005.1673

Meier, J. I., Marques, D. A., Mwaiko, S., Wagner, C. E., Excoffier, L., & Seehausen, O. (2017). Ancient hybridization fuels rapid cichlid fish adaptive radiations. *Nature Communications*, *8*. https://doi.org/10.1038/ncomms14363

Meier, J. I., Sousa, V. C., Marques, D. A., Selz, O. M., Wagner, C. E., Excoffier, L., & Seehausen, O. (2017). Demographic modelling with whole-genome data reveals parallel origin of similar Pundamilia cichlid species after hybridization. *Molecular Ecology*, *26*(1), 123–141. https://doi.org/10.1111/mec.13838

Meirmans, P. G. (2012). The trouble with isolation by distance. *Molecular Ecology*, *21*(12), 2839–2846. https://doi.org/10.1111/j.1365-294X.2012.05578.x

Meisner, J., & Albrechtsen, A. (2018). Inferring Population Structure and Admixture Proportions in Low-Depth NGS Data. *Genetics*, *210*(2), 719–731. https://doi.org/10.1534/genetics.118.301336

Meyer, C. G., Anderson, J. M., Coffey, D. M., Hutchinson, M. R., Royer, M. A., & Holland, K. N. (2018). Habitat geography around Hawaii's oceanic islands influences tiger shark (Galeocerdo cuvier) spatial behaviour and shark bite risk at ocean recreation sites. *Scientific Reports*, *8*(1), 4945. https://doi.org/10.1038/s41598-018-23006-0

Miller, M. R., Dunham, J. P., Amores, A., Cresko, W. A., & Johnson, E. A. (2007). Rapid and cost-effective polymorphism identification and genotyping using restriction site associated DNA (RAD) markers. *Genome Research*, *17*(2), 240–248. https://doi.org/10.1101/gr.5681207

Minh, B. Q., Schmidt, H. A., Chernomor, O., Schrempf, D., Woodhams, M. D., Von Haeseler, A., Lanfear, R., & Teeling, E. (2020). IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Molecular Biology and Evolution*, *37*(5), 1530–1534. https://doi.org/10.1093/molbev/msaa015

Momigliano, P., Harcourt, R., Robbins, W. D., Jaiteh, V., Mahardika, G. N., Sembiring, A., & Stow, A. (2017). Genetic structure and signatures of selection in grey reef sharks (Carcharhinus amblyrhynchos). *Heredity*, *119*(3), 142–153. https://doi.org/10.1038/hdy.2017.21

Momigliano, P., Harcourt, R., Robbins, W. D., & Stow, A. (2015). Connectivity in grey reef sharks (Carcharhinus amblyrhynchos) determined using empirical and simulated genetic data. *Scientific Reports*, *5*(August), 1–9. https://doi.org/10.1038/srep13229

Mona, S. (2017). On the role played by the carrying capacity and the ancestral population size during a range expansion. *Heredity*, *118*(2), 143–153. https://doi.org/10.1038/hdy.2016.73

Mona, S., Benazzo, A., Delrieu-Trottin, E., & Lesturgie, P. (2023). *Population genetics using low coverage RADseq data in non-model organisms: biases and solutions*. https://doi.org/10.22541/au.168252801.19878064/v1

Mona, S., Ray, N., Arenas, M., & Excoffier, L. (2014). Genetic consequences of habitat fragmentation during a range expansion. *Heredity*, *112*(3), 291–299. https://doi.org/10.1038/hdy.2013.105

Mondal, M., Bertranpetit, J., & Lao, O. (2019). Approximate Bayesian computation with deep learning supports a third archaic introgression in Asia and Oceania. *Nature Communications*, *10*(1). https://doi.org/10.1038/s41467-018-08089-7

Mourier, J., Mills, S. C., & Planes, S. (2013). Population structure, spatial distribution and life-history traits of blacktip reef sharks Carcharhinus melanopterus. *Journal of Fish Biology*, *82*(3), 979–993. https://doi.org/10.1111/jfb.12039

Mourier, J., & Planes, S. (2013). Direct genetic evidence for reproductive philopatry and associated fine-scale migrations in female blacktip reef sharks (Carcharhinus melanopterus) in French Polynesia. *Molecular Ecology*, *22*(1), 201–214. https://doi.org/10.1111/mec.12103

Mourier, J., Vercelloni, J., & Planes, S. (2012). Evidence of social communities in a spatially structured network of a free-ranging shark species. *Animal Behaviour*, *83*(2), 389–401. https://doi.org/10.1016/j.anbehav.2011.11.008

Myers, R. A., Baum, J. K., Shepherd, T. D., Powers, S. P., & Peterson, C. H. (2007). Cascading effects of the loss of apex predatory sharks from a coastal ocean. *Science*, *315*(5820), 1846–1850. https://doi.org/10.1126/science.1138657

Myers, R. A., & Worm, B. (2003). Rapid worldwide depletion of predatory fish communities. *Nature*, *423*(6937), 280–283. https://doi.org/10.1038/nature01610

Narasimhan, V., Danecek, P., Scally, A., Xue, Y., Tyler-Smith, C., & Durbin, R. (2016). BCFtools/RoH: A hidden Markov model approach for detecting autozygosity from next-generation sequencing data. *Bioinformatics*, *32*(11), 1749–1751. https://doi.org/10.1093/bioinformatics/btw044

Naylor, G. J. P., Caira, J. N., Jensen, K., Rosana, K. A. M., White, W. T., & Last, P. R. (2012). A DNA sequencebased approach to the identification of shark and ray species and its implications for global elasmobranch diversity and parasitology. In *Bulletin of the American Museum of Natural History* (Vol. 367, Issue 367, pp. 1–262). https://doi.org/10.1206/754.1

Nielsen, R., Korneliussen, T., Albrechtsen, A., Li, Y., & Wang, J. (2012). SNP Calling, Genotype Calling, and Sample Allele Frequency Estimation from New-Generation Sequencing Data. *PLoS ONE*, *7*(7), e37558. https://doi.org/10.1371/journal.pone.0037558

Nielsen, R., & Slatkin, M. (2013). *An Introduction to Population Genetics Theory and Applications* (Oxford Uni).

Nielsen, R., & Wakeley, J. (2001). Distinguishing migration from isolation: A Markov chain Monte Carlo approach. *Genetics*, *158*(2), 885–896. https://doi.org/10.1093/genetics/158.2.885

Niu, J., Huss, M., Vasemägi, A., & Gårdmark, A. (2023). Decades of warming alters maturation and reproductive investment in fish. *Ecosphere*, *14*(1). https://doi.org/10.1002/ecs2.4381

Ortiz, E. M. (2019). *vcf2phylip v2.0: convert a VCF matrix into several matrix formats for phylogenetic analysis.* (v2). Zenodo.

Overcast, I., Achaz, G., Aguilée, R., Andújar, C., Arribas, P., Creedy, T. J., Economo, E. P., Etienne, R. S., Gillespie, R., Jacquet, C., Jay, F., Kennedy, S., Krehenwinkel, H., Lambert, A., Meramveliotakis, E., Noguerales, V., Perez-Lamarque, B., Roderick, G., Rogers, H., … Morlon, H. (2023). Towards a genetic theory of island biogeography: Inferring processes from multidimensional community-scale data. In *Global Ecology and Biogeography* (Vol. 32, Issue 1, pp. 4–23). John Wiley and Sons Inc. https://doi.org/10.1111/geb.13604

Pacifici, M., Rondinini, C., Rhodes, J. R., Burbidge, A. A., Cristiano, A., Watson, J. E. M., Woinarski, J. C. Z., & Di Marco, M. (2020). Global correlates of range contractions and expansions in terrestrial mammals. *Nature Communications*, *11*(1). https://doi.org/10.1038/s41467-020-16684-w

Palsbøll, P. J., Bérubé, M., & Allendorf, F. W. (2007). Identification of management units using population genetic data. *Trends in Ecology and Evolution*, *22*(1), 11–16. https://doi.org/10.1016/j.tree.2006.09.003

Palumbi, S. R. (1994). Genetic divergence, reproductive isolation, and marine speciation. In *Annual Review of Ecology and Systematics* (Vol. 25, Issue 1, pp. 547–572). https://doi.org/10.1146/annurev.es.25.110194.002555

Pante, E., Abdelkrim, J., Viricel, A., Gey, D., France, S. C., Boisselier, M. C., & Samadi, S. (2015). Use of RAD sequencing for delimiting species. *Heredity*, *114*(5), 450–459. https://doi.org/10.1038/hdy.2014.105

Pardini, A. T., Jones, C. S., Noble, L. R., Kreiser, B., Malcolm, H., Bruce, B. D., Stevens, J. D., Cliff, G., Scholl, M. C., Francis, M., Duffy, C. A. J., Martin, A. P., Pardini, A. T., Jones, C. S., Jones, C. S., Noble, L. R., Noble, L. R., Kreiser, B., Kreiser, B., … Martin, A. P. (2001). Sex-biased dispersal of great white sharks. *Nature*, *412*(6843), 139–140. https://doi.org/10.1038/35084125

Paris, J. R., Stevens, J. R., & Catchen, J. M. (2017). Lost in parameter space: a road map for stacks. *Methods in Ecology and Evolution*, *8*(10), 1360–1373. https://doi.org/10.1111/2041-210X.12775

Parsons, G. R. (1990). Metabolism and swimming efficiency of the bonnethead shark Sphyrna tiburo. *Marine Biology*, *104*(3), 363–367. https://doi.org/10.1007/BF01314338

Pasinelli, G. (2022). Genetic diversity and spatial genetic structure support the specialist-generalist variation hypothesis in two sympatric woodpecker species. *Conservation Genetics*, *23*(4), 821–837. https://doi.org/10.1007/s10592-022-01451-9

Payne, N. L., Meyer, C. G., Smith, J. A., Houghton, J. D. R., Barnett, A., Holmes, B. J., Nakamura, I., Papastamatiou, Y. P., Royer, M. A., Coffey, D. M., Anderson, J. M., Hutchinson, M. R., Sato, K., & Halsey, L. G. (2018). Combining abundance and performance data reveals how temperature regulates coastal occurrences and activity of a roaming apex predator. *Global Change Biology*, *24*(5), 1884–1893. https://doi.org/10.1111/gcb.14088

Pazmiño, D. A., Maes, G. E., Green, M. E., Simpfendorfer, C. A., Hoyos-Padilla, E. M., Duffy, C. J. A., Meyer, C. G., Kerwath, S. E., Salinas-De-León, P., & Van Herwerden, L. (2018). Strong trans-Pacific break and local conservation units in the Galapagos shark (Carcharhinus galapagensis) revealed by genome-wide cytonuclear markers. *Heredity*, *120*(5), 407–421. https://doi.org/10.1038/s41437-017-0025-2

Pazmiño, D. A., Maes, G. E., Simpfendorfer, C. A., Salinas-de-León, P., & van Herwerden, L. (2017). Genome-wide SNPs reveal low effective population size within confined management units of the highly vagile Galapagos shark (Carcharhinus galapagensis). *Conservation Genetics*, *18*(5), 1151–1163. https://doi.org/10.1007/s10592-017-0967-1

Pearse, D. E., Barson, N. J., Nome, T., Gao, G., Campbell, M. A., Abadía-Cardoso, A., Anderson, E. C., Rundio, D. E., Williams, T. H., Naish, K. A., Moen, T., Liu, S., Kent, M., Moser, M., Minkley, D. R., Rondeau, E. B., Brieuc, M. S. O., Sandve, S. R., Miller, M. R., … Lien, S. (2019). Sex-dependent dominance maintains migration supergene in rainbow trout. *Nature Ecology and Evolution*, *3*(12), 1731–1742. https://doi.org/10.1038/s41559-019-1044-6

Peter, B. M., & Slatkin, M. (2013). Detecting range expansions from genetic data. *Evolution*, *67*(11), 3274–3289. https://doi.org/10.1111/evo.12202

Peter, B. M., & Slatkin, M. (2015). The effective founder effect in a spatially expanding population. *Evolution*, *69*(3), 721–734. https://doi.org/10.1111/evo.12609

Peter, B. M., Wegmann, D., & Excoffier, L. (2010). Distinguishing between population bottleneck and population subdivision by a Bayesian model choice procedure. *Molecular Ecology*, *19*(21), 4648–4660. https://doi.org/10.1111/j.1365-294X.2010.04783.x

Peterson, B. K., Weber, J. N., Kay, E. H., Fisher, H. S., & Hoekstra, H. E. (2012). Double Digest RADseq: An Inexpensive Method for De Novo SNP Discovery and Genotyping in Model and Non-Model Species. *PLoS ONE*, *7*(5), e37135. https://doi.org/10.1371/journal.pone.0037135

Peterson, C. D., Belcher, C. N., Bethea, D. M., Driggers, W. B., Frazier, B. S., & Latour, R. J. (2017). Preliminary recovery of coastal sharks in the south-east United States. *Fish and Fisheries*, *18*(5), 845–859. https://doi.org/10.1111/faf.12210

Pfeifer, B., Wittelsbürger, U., Ramos-Onsins, S. E., & Lercher, M. J. (2014). PopGenome: An efficient swiss army knife for population genomic analyses in R. *Molecular Biology and Evolution*, *31*(7), 1929–1936. https://doi.org/10.1093/molbev/msu136

Pini, J., Planes, S., Rochel, E., Lecchini, D., & Fauvelot, C. (2011). Genetic diversity loss associated to high mortality and environmental stress during the recruitment stage of a coral reef fish. *Coral Reefs*, *30*(2), 399–404. https://doi.org/10.1007/s00338-011-0718-6

Pinsky, M. L., & Palumbi, S. R. (2014). Meta-analysis reveals lower genetic diversity in overfished populations. *Molecular Ecology*, *23*(1), 29–39. https://doi.org/10.1111/mec.12509

Pirog, A., Jaquemet, S., Ravigné, V., Cliff, G., Clua, E., Holmes, B. J., Hussey, N. E., Nevill, J. E. G., Temple, A. J., Berggren, P., Vigliola, L., & Magalon, H. (2019). Genetic population structure and demography of an apex predator, the tiger shark Galeocerdo cuvier. *Ecology and Evolution*, *9*(10), 5551–5571. https://doi.org/10.1002/ece3.5111

Pitman, J. (1999). Coalescents with Multiple Collisions. *The Annals of Probability*, *27*(4), 1870–1902.

Poisot, T., Bever, J. D., Nemri, A., Thrall, P. H., & Hochberg, M. E. (2011). A conceptual framework for the evolution of ecological specialisation. *Ecology Letters*, *14*(9), 841–851. https://doi.org/10.1111/j.1461-0248.2011.01645.x

Polechová, J., & Storch, D. (2008). Ecological niche. *Encyclopedia of Ecology*.

Poplin, R., Chang, P. C., Alexander, D., Schwartz, S., Colthurst, T., Ku, A., Newburger, D., Dijamco, J., Nguyen, N., Afshar, P. T., Gross, S. S., Dorfman, L., McLean, C. Y., & Depristo, M. A. (2018). A universal snp and small-indel variant caller using deep neural networks. *Nature Biotechnology*, *36*(10), 983. https://doi.org/10.1038/nbt.4235

Pritchard, J. K., Stephens, M., & Donnelly, P. (2000). *Inference of Population Structure Using Multilocus Genotype Data*. http://www.stats.ox.ac.uk/pritch/home.html.

Pudlo, P. (2018). *Approximate Bayesian model choice as a Machine Learning problem*.

Pudlo, P., Marin, J. M., Estoup, A., Cornuet, J. M., Gautier, M., & Robert, C. P. (2016). Reliable ABC model choice via random forests. *Bioinformatics*, *32*(6), 859–866. https://doi.org/10.1093/bioinformatics/btv684

Pudlo, P., & Robert, C. P. (2019). *Package ‘ abcrf .’* https://doi.org/10.1093/bioinformatics/btv684>.Estoup

Purcell, J., Brelsford, A., Wurm, Y., Perrin, N., & Chapuisat, M. (2014). Convergent genetic architecture underlies social organization in ants. *Current Biology*, *24*(22), 2728–2732. https://doi.org/10.1016/j.cub.2014.09.071

Pybus, O. G., Rambaut, A., & Harvey, P. H. (2000). An Integrated Framework for the Inference of Viral Population History From Reconstructed Genealogies. *Genetics*, *155*, 1429–1437. https://academic.oup.com/genetics/article/155/3/1429/6050940

Quinlan, A. R., & Hall, I. M. (2010). BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics*, *26*(6), 841–842. https://doi.org/10.1093/bioinformatics/btq033

Ramachandran, S., Deshpande, O., Roseman, C. C., Rosenberg, N. A., Feldman, M. W., & Cavalli-Sforza, L. L. (2005). Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa. *Proceedings of the National Academy of Sciences*, *102*(44), 15942–15947. https://doi.org/10.1073/pnas.0507611102

Ray, N., Currat, M., & Excoffier, L. (2003). Intra-deme molecular diversity in spatially expanding populations. *Molecular Biology and Evolution*, *20*(1), 76–86. https://doi.org/10.1093/molbev/msg009

Raynal, L., Marin, J. M., Pudlo, P., Ribatet, M., Robert, C. P., & Estoup, A. (2019). ABC random forests for Bayesian parameter inference. *Bioinformatics*, *35*(10), 1720–1728. https://doi.org/10.1093/bioinformatics/bty867

Reid, B. N., Naro-Maciel, E., Hahn, A. T., FitzSimmons, N. N., & Gehara, M. (2019). Geography best explains global patterns of genetic diversity and postglacial co-expansion in marine turtles. *Molecular Ecology*, *28*(14), 3358–3370. https://doi.org/10.1111/mec.15165

Reid, B. N., & Pinsky, M. L. (2022). Simulation-Based Evaluation of Methods, Data Types, and Temporal Sampling Schemes for Detecting Recent Population Declines. In *Integrative and Comparative Biology* (Vol. 62, Issue 6, pp. 1849–1863). Oxford University Press. https://doi.org/10.1093/icb/icac144

Reynolds, J., Weir, B. S., & Cockerham, C. C. (1983). *ESTIMATION O F T H E COANCESTRY COEFFICIENT: BASIS FOR A SHORT-TERM GENETIC DISTANCE*. https://academic.oup.com/genetics/article/105/3/767/5996242

Robbins, W. D., Hisano, M., Connolly, S. R., & Choat, J. H. (2006). Ongoing Collapse of Coral-Reef Shark Populations. *Current Biology*, *16*(23), 2314–2319. https://doi.org/10.1016/j.cub.2006.09.044

Roberts, C. M., McClean, C. J., Veron, J. E. N., Hawkins, J. P., Allen, G. R., McAllister, D. E., Mittermeier, C. G., Schueler, F. W., Spalding, M., Wells, F., Vynne, C., & Werner, T. B. (2002). Marine biodiversity hotspots and conservation priorities for tropical reefs. *Science*, *295*(5558), 1280–1284. https://doi.org/10.1126/science.1067728

Robledo, D., Palaiokostas, C., Bargelloni, L., Martínez, P., & Houston, R. (2018). Applications of genotyping by sequencing in aquaculture breeding and genetics. In *Reviews in Aquaculture* (Vol. 10, Issue 3, pp. 670–682). Wiley-Blackwell. https://doi.org/10.1111/raq.12193

Rochette, N. C., Rivera-Colón, A. G., & Catchen, J. M. (2019). Stacks 2: Analytical methods for paired-end sequencing improve RADseq-based population genomics. *Molecular Ecology*, *28*(21), 4737–4754. https://doi.org/10.1111/mec.15253

Rodríguez, W., Mazet, O., Grusea, S., Arredondo, A., Corujo, J. M., Boitard, S., & Chikhi, L. (2018). The IICR and the non-stationary structured coalescent: towards demographic inference with arbitrary changes in population structure. *Heredity*, *121*(6), 663–678. https://doi.org/10.1038/s41437-018-0148-0

Roesti, M., Gilbert, K. J., & Samuk, K. (2022). Chromosomal inversions can limit adaptation to new environments. *Molecular Ecology*, *31*(17), 4435–4439. https://doi.org/10.1111/mec.16609

Rousset, F. (2000). Genetic differentiation between individuals. *Journal of Evolutionary Biology*, *13*(1), 58–62. https://doi.org/10.1046/j.1420-9101.2000.00137.x

Rutter, M., Moffitt, T. E., & Caspi, A. (2006). Gene-environment interplay and psychopathology: Multiple varieties but real effects. In *Journal of Child Psychology and Psychiatry and Allied Disciplines* (Vol. 47, Issues 3–4, pp. 226–261). https://doi.org/10.1111/j.1469-7610.2005.01557.x

Sagitov, S. (1999). The General Coalescent with Asynchronous Mergers of Ancestral Lines. In *Source: Journal of Applied Probability* (Vol. 36, Issue 4). https://about.jstor.org/terms

Schaal, S. M., Haller, B. C., & Lotterhos, K. E. (2022). Inversion invasions: When the genetic basis of local adaptation is concentrated within inversions in the face of gene flow. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *377*(1856). https://doi.org/10.1098/rstb.2021.0200

Schaeffer, S. W. (2008). Selection in heterogeneous environments maintains the gene arrangement polymorphism of Drosophila pseudoobscura. *Evolution*, *62*(12), 3082–3099. https://doi.org/10.1111/j.1558-5646.2008.00504.x

Schiffels, S., & Durbin, R. (2014). Inferring human population size and separation history from multiple genome sequences. *Nature Genetics*, *46*(8), 919–925. https://doi.org/10.1038/ng.3015

Schwander, T., Libbrecht, R., & Keller, L. (2014). Supergenes and complex phenotypes. In *Current Biology* (Vol. 24, Issue 7). Cell Press. https://doi.org/10.1016/j.cub.2014.01.056

Sexton, J. P., Montiel, J., Shay, J. E., Stephens, M. R., & Slatyer, R. A. (2017). Evolution of Ecological Niche Breadth. *Annu. Rev. Ecol. Evol. Syst*, *48*, 183–206. https://doi.org/10.1146/annurev-ecolsys-110316

Shin, J.-H., Blay, S., Mcneney, B., & Graham, J. (2006). LDheatmap: An R Function for Graphical Display of Pairwise Linkage Disequilibria Between Single Nucleotide Polymorphisms. In *JSS Journal of Statistical Software* (Vol. 16). http://www.jstatsoft.org/

Slatkin, M. (1993). ISOLAnON BY DISTANCE IN EQUILIBRIUM AND NON-EQUILIBRIUM POPULATIONS. In *Evolution* (Vol. 47, Issue 1).

Slatkin, M., & Excoffier, L. (2012). Serial Founder Effects During Range Expansion: A Spatial Analog of Genetic Drift. *Genetics*, *191*(1), 171–181. https://doi.org/10.1534/genetics.112.139022

Smit, A., & Hubley, R. (2015). *RepeatModeler Open-1.0. 2008-2015* (4.1.0).

Smit, A., Hubley, R., & Green, P. (2015). *RepeatMasker Open-4.0. 2013-2015* (4.1.0).

Smith, T., & Weissman, D. B. (2020). Isolation by Distance in Populations with Long-Range Dispersal. *BioRxiv*, 1–36. https://doi.org/10.1101/2020.06.24.168211%0Ahttps://www.biorxiv.org/content/10.1101/2020.06.24.168211v1

Sosebee, K. A. (2004). Maturity of Skates in Northeast United States Waters. *Journal of Northwest Atlantic Fishery Science*, *35*, 141–153. https://doi.org/10.2960/J.v35.m499

Sosebee, K. A., Miller, A., O'brien, L., Mcelroy, D., & Sherman, S. (2016). Update of thorny skate (Amblyraja radiata) commercial and survey data. *Northeast Fisheries Science Center Reference Document ; 16-08*. https://doi.org/10.7289/V5/RD-NEFSC-16-08

Soule, M., & Stewart, B. R. (1970). The "Niche-Variation" Hypothesis: A Test and Alternatives. *The American Naturalist*, *104*(935), 85–97. https://www.jstor.org/stable/2459075

Spalding, M. D., & Grenfell, A. M. (1997). New estimates of global and regional coral reef areas. *Coral Reefs*, *16*(4), 225–230. https://doi.org/10.1007/s003380050078

Speed, C. W., Meekan, M. G., Field, I. C., McMahon, C. R., Harcourt, R. G., Stevens, J. D., Babcock, R. C., Pillans, R. D., & Bradshaw, C. J. A. (2016). Reef shark movements relative to a coastal marine protected area. *Regional Studies in Marine Science*, *3*, 58–66. https://doi.org/10.1016/j.rsma.2015.05.002

Städler, T., Haubold, B., Merino, C., Stephan, W., & Pfaffelhuber, P. (2009). The impact of sampling schemes on the site frequency spectrum in nonequilibrium subdivided populations. *Genetics*, *182*(1), 205–216. https://doi.org/10.1534/genetics.108.094904

Steiner, C. C., Putnam, A. S., Hoeck, P. E. A., & Ryder, O. A. (2013). Conservation Genomics of Threatened Animal Species. *Annual Review of Animal Biosciences*, *1*(1), 261–281. https://doi.org/10.1146/annurev-animal-031412-103636

Stenløkk, K., Saitou, M., Rud-Johansen, L., Nome, T., Moser, M., Árnyasi, M., Kent, M., Barson, N. J., & Lien, S. (2022). The emergence of supergenes from inversions in Atlantic salmon. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *377*(1856). https://doi.org/10.1098/rstb.2021.0195

Stolle, E., Pracana, R., López-Osorio, F., Priebe, M. K., Hernández, G. L., Castillo-Carrillo, C., Arias, M. C., Paris, C. I., Bollazzi, M., Priyam, A., & Wurm, Y. (2022). Recurring adaptive introgression of a supergene variant that determines social organization. *Nature Communications*, *13*(1). https://doi.org/10.1038/s41467-022-28806-7

Straw, R. M. (1955). HYBRIDIZATION, HOMOGAMY, AND SYMPATRIC SPECIATION. *Evolution*, *9*(4), 441–444. https://doi.org/10.1111/j.1558-5646.1955.tb01553.x

Strona, G., Galli, P., & Fattorini, S. (2013). Fish parasites resolve the paradox of missing coextinctions. *Nature Communications*, *4*. https://doi.org/10.1038/ncomms2723

Sulikowski, J. A., Kneebone, J., Elzey, S., Jurek, J., Danley, P. D., & Huntting, W. (2005). Age and growth estimates of the thorny skate (Amblyraja radiata) in the western Gulf of Maine. *Fishery Bulletin*, *4*. https://scholars.unh.edu/biosci_facpub

Sumpton, W., Taylor, S., Gribble, N., McPherson, G., & Ham, T. (2011). Gear selectivity of large-mesh nets and drumlines used to catch sharks in the Queensland Shark Control Program. *African Journal of Marine Science*, *33*(1), 37–43. https://doi.org/10.2989/1814232X.2011.572335

Swinsburg, W., Kohler, N. E., Turner, P. A., & Camilla, T. (2012). *Mark / Recapture Data for the Blacktip Shark , Carcharhinus limbatus , in the Gulf of Mexico from the NEFSC Cooperative Shark Tagging Program SEDAR29-WP-16 Date Submitted : 6 March 2012*. *March*.

Tajima, F. (1983). EVOLUTIONARY RELATIONSHIP OF DNA SEQUENCES IN FINITE POPULATIONS. *Genetics*, *105*, 437–460.

Tajima, F. (1989). Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*, *123*(3), 585–595. https://doi.org/10.1093/genetics/123.3.585

Tellier, A., & Lemaire, C. (2014). Coalescence 2.0: A multiple branching of recent theoretical developments and their applications. *Molecular Ecology*, *23*(11), 2637–2652. https://doi.org/10.1111/mec.12755

Temple, A. J., Kiszka, J. J., Stead, S. M., Wambiji, N., Brito, A., Poonian, C. N. S., Amir, O. A., Jiddawi, N., Fennessy, S. T., Pérez-Jorge, S., & Berggren, P. (2018). Marine megafauna interactions with small-scale fisheries in the southwestern Indian Ocean: a review of status

and challenges for research and management. *Reviews in Fish Biology and Fisheries*, *28*(1), 89–115. https://doi.org/10.1007/s11160-017-9494-x

Templeman, E. G. (1984). Variations in Numbers of Median Dorsal Thorns and Rows of Teeth in Thorny Skate (*Raja radiata*) of the Northwest Atlantic. *Journal of Northwest Atlantic Fishery Science*, *5*, 171–179. https://doi.org/10.2960/J.v5.a21

Templeman, W. (1984). Migrations of Thorny Skate, *Raja radiata* , Tagged in Newfoundland. *Journal of Northwest Atlantic Fishery Science*, *5*, 55–63. https://doi.org/10.2960/J.v5.a6

Templeman, W. (1987). Differences in Sexual Maturity and Related Characteristics Between Populations of Thorny Skate (Raja radiata) in the Northwest Atlantic. In *J. Northw. Atl. Fish. SeL* (Vol. 7). http://journal.nafo.int

Terborgh, J. W. (2015). Toward a trophic theory of species diversity. *Proceedings of the National Academy of Sciences of the United States of America*, *112*(37), 11415–11422. https://doi.org/10.1073/pnas.1501070112

Terhorst, J., Kamm, J. A., & Song, Y. S. (2017). Robust and scalable inference of population history from hundreds of unphased whole genomes. *Nature Genetics*, *49*(2), 303–309. https://doi.org/10.1038/ng.3748

Thioulouse, J., & Dray, S. (2007). Interactive multivariate data analysis in R with the ade4 and ade4TkGUI packages. *Journal of Statistical Software*, *22*(5), 1–14. https://doi.org/10.18637/jss.v022.i05

Thompson, M. J., & Jiggins, C. D. (2014). Supergenes and their role in evolution. In *Heredity* (Vol. 113, Issue 1, pp. 1–8). Nature Publishing Group. https://doi.org/10.1038/hdy.2014.20

Tillett, B. J., Meekan, M. G., Field, I. C., Thorburn, D. C., & Ovenden, J. R. (2012). Evidence for reproductive philopatry in the bull shark Carcharhinus leucas. *Journal of Fish Biology*, *80*(6), 2140–2158. https://doi.org/10.1111/j.1095-8649.2012.03228.x

Tilman, D., Isbell, F., & Cowles, J. M. (2014). Biodiversity and ecosystem functioning. *Annual Review of Ecology, Evolution, and Systematics*, *45*, 471–493. https://doi.org/10.1146/annurev-ecolsys-120213-091917

Tittensor, D. P., Mora, C., Jetz, W., Lotze, H. K., Ricard, D., Berghe, E. Vanden, & Worm, B. (2010). Global patterns and predictors of marine biodiversity across taxa. *Nature*, *466*(7310), 1098–1101. https://doi.org/10.1038/nature09329

Titus, B. M., & Daly, M. (2017). Specialist and generalist symbionts show counterintuitive levels of genetic diversity and discordant demographic histories along the Florida Reef Tract. *Coral Reefs*, *36*(1), 339–354. https://doi.org/10.1007/s00338-016-1515-z

Trakhtenbrot, A., Nathan, R., Perry, G., & Richardson, D. M. (2005). The importance of long-distance dispersal in biodiversity conservation. In *Diversity and Distributions* (Vol. 11, Issue 2, pp. 173–181). https://doi.org/10.1111/j.1366-9516.2005.00156.x

van der Zee, J. P., Christianen, M. J. A., Bérubé, M., Nava, M., Schut, K., Humber, F., Alfaro-Núñez, A., Becking, L. E., & Palsbøll, P. J. (2021). The population genomic structure of green turtles (Chelonia mydas) suggests a warm-water corridor for tropical marine fauna between the Atlantic and Indian oceans during the last interglacial. *Heredity*, *127*(6), 510–521. https://doi.org/10.1038/s41437-021-00475-0

van Etten, J. (2017). R package gdistance: Distances and routes on geographical grids. *Journal of Statistical Software*, *76*(1). https://doi.org/10.18637/jss.v076.i13

Van Valen, L. (1965). Morphological Variation and Width of Ecological Niche. *The American Naturalist*, *99*(908), 377–390. https://about.jstor.org/terms

Vandermeer, J. H. (1972). *NICHE THEORY HISTORICAL ROOTS OF NICHE THEORY*. www.annualreviews.org

Vendrami, D. L. J., Peck, L. S., Clark, M. S., Eldon, B., Meredith, M., & Hoffman, J. I. (2021). Sweepstake reproductive success and collective dispersal produce chaotic genetic patchiness in a broadcast spawner. In *Sci. Adv* (Vol. 7). https://www.science.org

Vignaud, T. M., Maynard, J. A., Leblois, R., Meekan, M. G., Vázquez-Juárez, R., Ramírez-Macías, D., Pierce, S. J., Rowat, D., Berumen, M. L., Beeravolu, C., Baksay, S., & Planes, S. (2014). Genetic structure of populations of whale sharks among ocean basins and evidence for their historic rise and recent decline. *Molecular Ecology*, *23*(10), 2590–2601. https://doi.org/10.1111/mec.12754

Wakeley, J. (1998). Segregating Sites in Wright's Island Model. *Theoretical Population Biology*, *53*, 166–174.

Wakeley, J. (1999). Nonequilibrium migration in human history. *Genetics*, *153*(4), 1863–1871.

Wakeley, J. (2009). *Coalescent theory : an introduction*. Roberts & Co. Publishers.

Wallberg, A., Schöning, C., Webster, M. T., & Hasselmann, M. (2017). Two extended haplotype blocks are associated with adaptation to high altitude habitats in East African honey bees. *PLoS Genetics*, *13*(5). https://doi.org/10.1371/journal.pgen.1006792

Walsh, C. A. J., Momigliano, P., Boussarie, G., Robbins, W. D., Bonnin, L., Fauvelot, C., Kiszka, J. J., Mouillot, D., Vigliola, L., & Manel, S. (2022). *Genomic insights into the historical and contemporary demographics of the grey reef shark*. *February*. https://doi.org/10.1038/s41437-022-00514-4

Warmuth, V. M., & Ellegren, H. (2019). Genotype-free estimation of allele frequencies reduces bias and improves demographic inference from RADSeq data. *Molecular Ecology Resources*, *19*(3), 586–596. https://doi.org/10.1111/1755-0998.12990

Watterson, G. A. A. (1975). On the number of segregating sites in genetical models without recombination. *Theoretical Population Biology*, *7*(2), 256–276. https://doi.org/10.1016/0040-5809(75)90020-9

Wegmann, D., Currat, M., & Excoffier, L. (2006). Molecular diversity after a range expansion in heterogeneous environments. *Genetics*, *174*(4), 2009–2020. https://doi.org/10.1534/genetics.106.062851

Weir, B. S., & Cockerham, C. C. (1984). Estimating F-Statistics for the Analysis of Population Structure. *Evolution*, *38*(6), 1358. https://doi.org/10.2307/2408641

Wellenreuther, M., & Bernatchez, L. (2018). Eco-Evolutionary Genomics of Chromosomal Inversions. In *Trends in Ecology and Evolution* (Vol. 33, Issue 6, pp. 427–440). Elsevier Ltd. https://doi.org/10.1016/j.tree.2018.04.002

Werry, J. M., Planes, S., Berumen, M. L., Lee, K. A., Braun, C. D., & Clua, E. (2014). Reef-Fidelity and Migration of Tiger Sharks, Galeocerdo cuvier, across the Coral Sea. *PLoS ONE*, *9*(1), e83249. https://doi.org/10.1371/journal.pone.0083249

White, B. J., Hahn, M. W., Pombi, M., Cassone, B. J., Lobo, N. F., Simard, F., & Besansky, N. J. (2007). Localization of candidate regions maintaining a common polymorphic inversion (2La) in Anopheles gambiae. *PLoS Genetics*, *3*(12), 2404–2414. https://doi.org/10.1371/journal.pgen.0030217

White, T. D., Carlisle, A. B., Kroodsma, D. A., Block, B. A., Casagrandi, R., De Leo, G. A., Gatto, M., Micheli, F., & McCauley, D. J. (2017). Assessing the effectiveness of a large marine protected area for reef shark conservation. *Biological Conservation*, *207*, 64–71. https://doi.org/10.1016/j.biocon.2017.01.009

Whitney, N. M., Robbins, W. D., Schultz, J. K., Bowen, B. W., & Holland, K. N. (2012). Oceanic dispersal in a sedentary reef shark (Triaenodon obesus): Genetic evidence for extensive connectivity without a pelagic larval stage. *Journal of Biogeography*, *39*(6), 1144–1156. https://doi.org/10.1111/j.1365-2699.2011.02660.x

Wilton, P. R., Carmi, S., & Hobolth, A. (2015). The SMC' is a highly accurate approximation to the ancestral recombination graph. *Genetics*, *200*(1), 343–355. https://doi.org/10.1534/genetics.114.173898

Wittmann, M. J., Bergland, A. O., Feldman, M. W., Schmidt, P. S., & Petrov, D. A. (2017). Seasonally fluctuating selection can maintain polymorphism at many loci via segregation lift. *Proceedings of the National Academy of Sciences*, *114*(46). https://doi.org/10.1073/pnas.1702994114

Wood, A. R., Esko, T., Yang, J., Vedantam, S., Pers, T. H., Gustafsson, S., Chu, A. Y., Estrada, K., Luan, J., Kutalik, Z., Amin, N., Buchkovich, M. L., Croteau-Chonka, D. C., Day, F. R., Duan, Y., Fall, T., Fehrmann, R., Ferreira, T., Jackson, A. U., … Frayling, T. M. (2014). Defining the role of common variation in the genomic and biological architecture of adult human height. *Nature Genetics*, *46*(11), 1173–1186. https://doi.org/10.1038/ng.3097

Worm, B., & Tittensor, D. P. (2011). Range contraction in large pelagic predators. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(29), 11942–11947. https://doi.org/10.1073/pnas.1102353108

Xue, A. T., & Hickerson, M. J. (2015). The aggregate site frequency spectrum for comparative population genomic inference. *Molecular Ecology*, *24*(24), 6223–6240. https://doi.org/10.1111/mec.13447

Yan, H. F., Kyne, P. M., Jabado, R. W., Leeney, R. H., Davidson, L. N. K., Derrick, D. H., Finucci, B., Freckleton, R. P., Fordham, S. V, & Dulvy, N. K. (2021). Overfishing and habitat loss drives range contraction of iconic marine fishes to near extinction. In *Sci. Adv* (Vol. 7). https://www.science.org

Yan, Z., Martin, S. H., Gotzek, D., Arsenault, S. V., Duchen, P., Helleu, Q., Riba-Grognuz, O., Hunt, B. G., Salamin, N., Shoemaker, D. W., Ross, K. G., & Keller, L. (2020). Evolution of a supergene that regulates a trans-species social polymorphism. *Nature Ecology and Evolution*, *4*(2), 240–249. https://doi.org/10.1038/s41559-019-1081-1

Zimmerman, E., Palsson, A., & Gibson, G. (2000). *Quantitative Trait Loci Affecting Components of Wing Shape in Drosophila melanogaster*. www.stonybrook.edu/

Appendix

Appendix 1. List of Abbreviations

ABC: Approximate Bayesian Computation
AIC: Akaike Information Criterion
ARG: Ancestral Recombination Graph
BIC: Bayesian Information Criterion
CR: Coalescence rate
DAPC: Discriminant Analysis of Principal Components
dd-RADseq: double-digested Restriction Associated DNA sequencing
DNA: Deoxyribonucleic acid
ESV: Exact Sequence Variants
IAA: Indo-Australian Archipelago
IBD: Isolation By Distance
IICR: Inverse Instantaneous Coalescence Rate
LC: Least-Cost
LD: Linkage Disequilibrium
LHT: Life History Traits
ML: Maximum Likelihood
MSMC: Multiple Sequentially Markovian Coalescent
MRCA: Most Recent Common Ancestor
NGS: Next-Generation Sequencing
NVH: Niche Variation Hypothesis
PCA: Principal Component Analysis
PSMC: Pairwise Sequentially Markovian Coalescent
RADseq: Restriction Associated DNA sequencing
RE: Range Expansion
RF: Random Forest
RFLP: Restriction fragment length polymorphism
ROH: Runs of Homozygosity
RRL: Reduced-Representation Libraries
SFS: Site Frequency Spectrum
SGVH: Specialist-Generalist Variation Hypothesis
SMC: Sequentially Markovian Coalescent
SMC++: Sequentially Markovian Coalescent Plus Plus
SME: Squared Mean Error
sNMF: sparse non-Negative Matrix Factorization
SNP: Single Nucleotide Polymorphism
SRMSE: Square Root Mean Square Error
TIB: Theory of Island Biogeography
TL: Total Length
T-TIB: Trophic Theory of Island Biogeography
VCF: Variant Call Format
WF: Wright-Fisher
WGS: Whole Genome Sequencing

Appendix 2. Articles

Published articles

- **Lesturgie, P.,** Planes, S., & Mona, S. (2022). Coalescence times, life history traits and conservation concerns: An example from four coastal shark species from the Indo-Pacific. Molecular Ecology Resources, 22(2), 554–566. doi: 10.1111/1755-0998.13487
- **Lesturgie, P.,** Lainé, H., Suwalski, A., Chifflet-Belle, P., Maisano Delser, P., Clua, E., … Mona, S. (2022). Ecological and biogeographic features shaped the complex evolutionary history of an iconic apex predator (*Galeocerdo cuvier*). BMC Ecology and Evolution, 22(1), 147. doi: 10.1186/s12862-022-02100-y
- **Lesturgie, P.,** Braun, C., Clua, E., Mourier, J., Vignaud, T., Planes, S., Mona, S. (2023). Like a rolling stone: colonization and migration dynamics of the gray reef shark (*Carcharhinus amblyrhynchos*). Ecology and Evolution, 13(1), doi: 10.1002/ece3.9746

Articles in preparation

- **Lesturgie, P.,** Denton, J., Kneebone, J., Laso-Jadart, R., Yang, L., Mona, S., Naylor, G. (In prep). A Size-determining Supergene Hampers a Vulnerable Population Recovery.
- **Lesturgie, P.,** Le Gouellec, M., Brandl, S. J., Casey, J.M., Parravicini, V., Mona, S. (In prep). Larger trophic niche increases stability along evolutionary times.

Collaborations

- Corrigan, S., Laso-Jadart, R., Yang, L., Gay, E., **Lesturgie, P.,** Lee, A., Fedrigo, O., Hoyos, M., Lowe, C., Lyons, K., Cliff, G., Sato, K., Tomita, T., Mona, S., & Naylor, G.. The checkered past of the world's white shark populations. *In prep.*
- Mona, S., Benazzo, A., Delrieu-Trottin, E., & **Lesturgie, P**.. Population genetics using low coverage RADseq data in non-model organisms: biases and solutions. Pending resubmission in *Molecular Ecology Resources*. Preprint DOI: https://doi.org/10.22541/au.168252801.19878064/v1
- Postaire, B., Devloo-Delva, F., Brunnschweiler, J.M., Charvet, P., Chen, X., Cliff, G., Daly,R., Drymon, J.M., Espinoza, M., Fernando, D., Glaus, K., Grant, M.I., Hernandez, S., Hyodo, S., Jabado, R.W., Jaquemet, S., Johnson, G., Naylor, G.J.P., Nevill, J.E.G.,

Pathirana, B.M., Pillans, R.D., Smoothey, AF., Tachihara, K., Tillet, B.J., Valerio-Vargas, J.A., **Lesturgie, P.** Magalon, H., Feutry, P., Mona, S. Global genetic diversity and historical demography of the Bull Shark *Carcharhinus leucas.* Accepted in *Journal of Biogeography.*

Appendix 3. Participation to Conferences and Seminars

- **Talk,** ***Mathematical and Computational Evolutionary Biology*** **(MCEB),** Porquerolles, France, **June 2021**

  Coalescence times, life history traits and conservation concerns: an example from four shark species from the Indo-Pacific. Lesturgie, P., Planes, S., & Mona, S.

- **Poster,** ***European Society for Evolutionary Biology*** **(ESEB),** Prague, Czech Republic, **August 2022**

  Influence of trophic niche on the historical demography of a coral reef fish fauna. Lesturgie, P., Brandl, S.J., Casey, J.M., Parravicini, V. & Mona, S.

- **Poster,** ***Joint meeting Alphy & AIEM*****.** Grenoble, France **January 2023**

  Understanding size differences in the Thorny Skate ***(Amblyraja radiata):*** **insights from whole genome data.** Lesturgie, P., Denton, J., Kneebone, J., Laso-Jadart , R., Yang, L., Mona, S., Naylor, G.

- **Talk,** ***Society for Molecular Biology and Evolution*** **(SMBE),** Ferrara, Italy, **July 2023**

  Size matters: An introgressed supergene shaped the evolutionary dynamics of a chondrichthyan species. Lesturgie, P., Denton, J., Kneebone, J., Laso-Jadart, R., Yang, L., Mona, S., Naylor, G.

Appendix 4. Résumé détaillé

L'histoire évolutive des espèces est façonnée par des processus démographiques et sélectifs, nécessitant une modélisation complexe de la diversité génétique pour comprendre des phénomènes allant de l'échelle locale à l'échelle de l'écosystème. Cependant, il s'agit d'un exercice difficile en génétique des populations, car il nécessite une bonne compréhension des processus façonnant la diversité génétique, ainsi que l'exploration de scénarios complexes adaptés à la question et à la biologie de l'espèce étudiée. Dans ce contexte, la théorie du coalescent offre un puissant cadre inférentiel pour explorer des modèles très complexes, en particulier à travers des simulations. Toutefois, il est à coupler avec une méthodologie rigoureuse, notamment en incluant des analyses descriptives, afin de concevoir et interpréter de manière optimale des scénarios démographiques. Dans ce cadre, ma thèse vise à montrer comment la reconstruction détaillée des processus démographiques peut fournir des informations précieuses pour élaborer des hypothèses évolutives et des stratégies de conservation, tout en contribuant à une meilleure caractérisation de l'interaction entre les processus neutres et sélectifs. Également, elle cherche à améliorer notre compréhension de la façon dont les processus, de l'échelle de l'espèce aux communautés, influencent la démographie historique. Pour cela, j'ai étudié différentes questions et processus différents chez des organismes marins, à travers des données génomiques.

Dans un premier temps, j'ai étudié comment la structuration génétique des populations (et plus généralement, tout événement historique) influence les patrons démographiques inférés par des modèles basés sur la théorie de la coalescence supposant un accouplement aléatoire (modèles *non structurés*). En utilisant des arguments théoriques, j'ai d'abord pu montrer comment les modèles *non structurés* sont utiles pour inférer la variation du taux de coalescence dans le temps, qui est directement liée à la vraie démographie de l'espèce. Notamment, dans le cas des espèces structurées, j'ai révélé la signature de la colonisation de l'habitat directement sur la variation du taux de coalescence reconstruite par des modèles *non structurés*. Par la suite, j'ai confirmé empiriquement ce résultat, mais ai également montré la nécessité des analyses descriptives avant d'effectuer des inférences démographiques. Pour cela, j'ai étudié deux espèces de requins à large distribution et aux traits d'histoire de vie très différents : le requin tigre, panmictique à large échelle, et le requin gris, structuré en méta-population avec des signatures de colonisation de son aire de distribution dans la généalogie génique. Ceci a permis de mettre en avant que le modèles non-structurés sont un outil exploratoire fondamental pour recueillir des éléments sur l'histoire

évolutive des espèces, à condition qu'ils soient interprétés à la lumière de scénarios complexes plutôt que panmictiques. Plus généralement, ce chapitre démontre aussi l'utilité d'améliorer notre compréhension des signatures laissées par des paramètres démographiques dans la généalogie génique, ce qui nécessitera d'investiguer à l'avenir des modèles plus complexes.

Ensuite, je rapporte la découverte d'un supergène déterminant la taille chez une espèce de raie. C'est la première documentation directe d'un supergène impliqué dans le déterminisme d'un trait pourtant connu pour avoir un déterminisme polygénique, et potentiellement affecté par l'environnement, ce qui pourrait avoir des implications dépassant la génétique des populations. Au-delà de cela, le supergène n'est polymorphique que dans une partie de l'aire de distribution, et implique un accouplement assortatif positif dans le Golfe du Maine, une sous-population vulnérable connaissant un déclin continu depuis plusieurs décennies. Grâce à la reconstruction de l'histoire démographique de l'espèce à l'échelle de son aire de distribution, j'ai pu montrer qu'une forte connectivité à l'échelle régionale empêche une spéciation sympatrique dans le Golfe du Maine. Également, je montre que le supergène est polymorphique dans une région isolée depuis ~160,000 ans, proposant alors un intervalle de temps pour son origine. Finalement, je montre que l'un des allèles du supergène est introgressé, probablement entre aujourd'hui et 160,000 ans. Ce chapitre souligne à quel point la compréhension de l'origine du supergène nécessitera à l'avenir son investigation à l'aide de jeux de données multi-espèce, et potentiellement des développements théoriques. Surtout, il souligne l'importance de la modélisation démographique pour comprendre des processus locaux de sélection, en particulier lorsqu'ils impliquent des enjeux de conservation.

Enfin, dans un dernier chapitre, j'examine certains déterminants écologiques de la diversité génétique. Pour cela, j'ai mis en place un panel unique de données génomiques provenant de 40 espèces de poissons récifaux, que j'ai couplé à la reconstruction de leur niche trophique par des données de méta-barcoding de contenus stomachaux. Ceci a permis d'investiguer une hypothèse selon laquelle les espèces ne consommant peu de ressources, ou spécialistes (i.e., à faible largeur de niche trophique) devraient avoir une diversité génétique plus faible dû à des fluctuations démographiques plus fréquentes. L'analyse de modèles linéaires entre largeur de niche trophique et des indices de stabilité démographique (notamment construits par des analyses basées sur la théorie de la coalescence) a permis de montrer une relation positive entre largeur de niche trophique et stabilité démographique. Ceci a révélé pour la première fois de manière quantitative l'effet d'un processus à l'échelle de la communauté sur l'histoire démographique, avec des

implications de conservation importantes. Spécifiquement, le chapitre montre que les généralistes sont plus stables, et donc probablement moins vulnérables aux changements rapides, participant donc à la stabilité des écosystèmes. Les études en génétique des populations s'articulant en général sur une espèce, et ce travail est l'un des premiers à tenter d'évaluer directement la relation entre la diversité des interactions des espèces et leur histoire démographique. Plus généralement, cette étude suggère que les jeux de données multi-espèces pourraient se révéler importants à l'avenir pour détecter les signatures génomiques laissées par des processus à grande échelle.

Durant ma thèse, je me suis donc focalisé sur des organismes marins, en utilisant des techniques de génétique des populations et des données génomiques pour étudier comment différents processus, des variations génétiques locales aux interactions trophiques, influencent leur histoire évolutive. Dans l'ensemble, ma thèse souligne le rôle fondamental d'une reconstruction démographique robuste pour comprendre des processus micro (tels que l'adaptation) et macro (tels que le fonctionnement des écosystèmes) évolutifs. Notamment, je mets en avant un cadre inférentiel impliquant des analyses descriptives de la variabilité génétique permettant alors une conception de scénarios démographiques cohérents, et l'utilisation de méthodes computationnelles appropriées pour modéliser les scénarios choisis. De plus, ma thèse permet de mieux comprendre certains déterminants évolutifs et écologiques de la diversité génétique et la manière dont ils influencent la généalogie génique (et donc les inférences démographiques basées sur la coalescence). En particulier, je discute de l'importance du taux de coalescence reconstruit qui dépend de la véritable histoire démographique des lignées échantillonnées, et donc de son utilité comme statistique résumée pour comprendre l'histoire évolutive des espèces. Enfin, elle souligne l'importance et la puissance potentielle des études multi-espèces, relativement nouvelles en génétique des populations, qui peuvent fournir des informations précieuses sur des processus évolutifs. Les jeux de données multi-espèce permettront probablement à l'avenir de répondre à des questions à des échelles d'étude différentes, potentiellement pluridisciplinaires, avec alors des implications en évolution, en écologie et en conservation.