# THÈSE POUR OBTENIR LE GRADE DE DOCTEUR DE L'UNIVERSITÉ DE MONTPELLIER

**En génétique et génomique**

**École doctorale GAIA**

**Unité de recherche ISEM**

# ROLES DES CONTRAINTES GENOMIQUES ET DES TRAITS D'HISTOIRE DE VIE DANS LA SPECIATION : UNE APPROCHE DE GENOMIQUE COMPARATIVE


**Présentée par Pierre BARRY**


**Le 29 avril 2022**


**Sous la direction de Pierre-Alexandre GAGNAIRE et Thomas BROQUET**


**Devant le jury composé de**

| | |
|---|---|
| Guillaume EVANNO, Directeur de recherche INRAe, Université de Rennes, France | Rapporteur et président de jury |
| Kerstin JOHANNESSON, Professeure, Université de Gothenburg, Suède | Rapportrice |
| Christelle FRAÏSSE, Chargée de recherche CNRS, Université de Lille, France | Examinatrice |
| Oscar PUEBLA, Professeur, Université d'Oldenburg, Allemagne | Examinateur |
| Bert VAN BOCXLAER, Chargé de recherche CNRS, Université de Lille | Examinateur |
| Pierre-Alexandre GAGNAIRE, Chargé de recherche, CNRS, Université de Montpellier, France | Co-directeur |
| Thomas BROQUET, Chargé de recherche CNRS, Station Biologique de Roscoff, France | Co-directeur invité |

UNIVERSITÉ DE MONTPELLIER

**Remerciements**

J'aimerais tout d'abord remercier les membres du jury de thèse, Guillaume Evanno, Kerstin Johannesson, Christelle Fraïsse, Bert Van Bocxlaer et Oscar Puebla d'avoir accepté de lire mes travaux de thèse. J'aimerai aussi remercier les membres de mon comité de thèse, Stéphanie Manel, Sébastien Villéger, Mathieu Gautier, Michael Fontaine, Nicolas Galtier et Carole Smadja pour leurs conseils et les discussions enrichissantes. J'aimerais aussi remercier la direction et tous les personnels de l'ISEM pour leur accueil et leur gentillesse qui est la source de la bonne ambiance qu'il règne dans cet institut de recherche.

Un grand merci à mes deux directeurs de thèse Pierre-Alexandre et Thomas sans qui tout ce travail de thèse n'aurait pas été possible. Merci beaucoup de m'avoir fait confiance de m'avoir confié cette thèse: le Triskell se souvient encore de cette première rencontre ! Merci aussi de m'avoir donné confiance à travers votre écoute, votre soutien à tout heure et vos conseils toujours avisés. J'ai adoré les longues discussions qu'on a pu avoir à trois, ça a toujours été des moments stimulants de chauffage de cerveaux. Ca a été vraiment chouette aussi de vadrouiller sur le terrain à la recherche de la cache à blennie coiffée, de trouver la technique pour appâter les gobies noires ou les escapades dans les pêcheries espagnoles pour couper les nageoires aux baudroies (même si le réveil à 4h pour aller à la pêcherie de St Jean de Luz, ça fait tirer sur les adducteurs) ! Merci à Thomas de m'avoir accueilli à Roscoff de m'avoir fait découvrir l'air vivifiant de la Bretagne ! Bref, ces trois années et demi auront vraiment été exceptionnelles tant d'un point de vue de la science que du point de vue humain ! Je pense que je n'aurais pu pas rêvé mieux comme directeurs de thèses.

Merci beaucoup à toute les personnes qui nous ont aidé sur le terrain pour récupérer ces précieux échantillons. Merci à Cristina Mena de l'Asociación Hippocampus pour l'échantillon d'hippocampe et l'information sur le splot de la marina qui était en fait un HLM à hippocampe et à gobies noires; merci à Rita Castilho et à Regina Mena pour nous avoir aidé dans la pêcherie à Olhao et nous avoir fait découvrir le coin à girelle à Sagres; merci à tout les pêcheurs du Golf du Lion, du Golf de Gascogne, espagnols et portugais qui ont pris le temps de nous écouter, de nous conseiller et de nous avoir permis de récupérer des échantillons précieux. Merci à Sébastien Villéger pour les crénilabres à Nico pour les syngnathes de Thau. Merci à Thierry Pastor pour l'aide de terrain et pour les soirées sympathiques dans le AirBNB de la Costa Calida. Thank you to the Speciation Genomics workshop participants in Berlin in November 2018 especially Maximilian it was fun to talk about blennies while strolling in Berlin with you (and I finally found the Lepadogaster!).

Je fais une petite pause dans les remerciements pouir ne **PAS** remercier le COVID. Voilà c'est dit.

Merci beaucoup à la BEM team aux anciens comme aux nouveaux de m'avoir accompagné pendant ces 3 ans et demi. Merci à Nico, Maud et Alexis de m'avoir si gentiment accueilli dans l'équipe à mon arrivée. Merci aux doctorants, stagiaires et post-doc de l'équipe, Maurine, Adrien, Marie, Laura, Fanny, Salomé, Jeanne, Paul, Iago, Arthur, Johan, Emma, pour tous les moments d'échanges. Merci à Marion, ça a été très sympa de travailler avec toi, j'espère que tu t'amuses bien en Suisse. Merci à tout le bâtiment 24, anciens et passés, Khalid, Rémy, Nelly, Emilie, Nelly, Stephen, Jimmy, Mathieu, Paul, Bruno, Fred & Fred, Erick, François, Christine pour votre bonne humeur et gentillesse. Merci à Camille Roux de m'avoir accueilli à Lille et d'avoir pris du temps de m'expliquer l'ABC; merci à Estelle pour l'hébergement à Lille et aux collègues de l'Université, Zoé, Thomas pour leur accueil, c'était très chouette (et la bière était

bonne et pas chère) ! Merci à Jade, Martin, Didier et toute l'équipe de Roscoff pour leur accueil ! Merci aux participants de iMarco à Aveiro et à Paris !

Merci à tous les copains de l'ISEM du 22 ! Même si le COVID (encore lui, quelle sale bête) nous a un peu distancié depuis 1 an et demi, merci pour les apéros, les sorties dans les bars et pour le week-end d'intégration ! Et longue vie à la croziflette ! Merci à tous les copains de Montpellier, les exilés ou pas de Supagro, les doctorants du CEFE, et tous les visiteurs d'un soir, d'un week-end ou plus longtemps, merci pour les moments de franche rigolade et de grosse poilade dans les bars, à la maison, ou n'importe ou d'ailleurs. Petite pensée émue à la Sahel qui nous aura bien accompagné pendant le confinement. Merci à Val pour les 3 and et demi de super ambiance à la coloc et merci à toi pour l'initiative de ces fameux week-ends champêtres ! Merci aux magasins Noz de nous ravir de ses magnifiques perles chaque jour. Merci aux truites, aux saumons, aux flanby, aux boules de neige (riz?), aux pétitions, aux cagnottes, aux huîtres et au vin blanc, aux plaques de marbre, aux sports extrêmes in-door, aux PMU, aux bob, aux véloutés de potiron, aux crêpes salées (fait maison), au goudda et j'en passe. Merci à Fanny pour les gnocchis du réconfort ! Merci à la team Ecossaise pour ces derniers moments sans masque en février 2020 (do you known Antoine Dupont ?) et pour la réunion de poulet au Gigot. Merci à tous les copains de la flore aagronome pour les visites culturelles dans des lieux exotiques tel que Budapest, Bruxelles, Clermont (?) ou Couzon au Mont Dore (???). Un immense merci à Louis la Brocante (ça passe pas).

Merci à toute ma famille, à mes parents et à mon frère tout particulièrement, de m'avoir soutenu et de m'avoir conseillé sur mes choix d'orientations ! Merci particulièrement de m'écouter gentiment quand je vous explique ce que je fait en thèse : promis j'essaye de pas trop mettre du jargon de biologiste dans mes explications, je vous dois bien cela !

Et évidemment merci à Lucie pour m'avoir accompagné durant ces 3 années et demi et depuis bien plus longtemps.

## Résumé

La spéciation est le processus évolutif au cours duquel une espèce se scinde en deux lignées qui divergent en accumulant des barrières reproductives, jusqu'à l'acquisition d'un isolement reproductif total. Durant ce processus, les lignées divergentes peuvent toujours s'échanger des gènes par hybridation, mais le flux génique est progressivement limité par l'accumulation des barrières. Il en résulte une semi-perméabilité des génomes, où certains locus s'échangent librement entre lignées et restent indifférenciés tandis que d'autres n'introgressent pas, contribuant ainsi à l'établissement de régions génomiques divergentes, appelées îlots génomiques de spéciation. Ces locus barrières peuvent être impliqués dans différents types de mécanismes d'isolement, incluant le choix de partenaire, l'adaptation à différents environnements, ou des incompatibilités génétiques entre plusieurs gènes coadaptés. L'étude de l'établissement, l'accumulation, l'érosion et la maintenance de ces barrières et de leurs effets sur la semiperméabilité des génomes de lignées en cours de spéciation permet de comprendre comment de nouvelles espèces se forment. L'avènement des techniques de séquençage à haut débit a permis de caractériser le paysage génomique de divergence chez de multiples lignées en cours de spéciation à travers l'arbre du vivant. Ces études ont permis de mesurer l'influence de l'histoire démographique et de l'architecture génomique comme déterminants majeurs du paysage génomique de divergence. Toutefois, d'autres facteurs pourraient intervenir et expliquer la diversité des trajectoires évolutives pouvant conduire ou non à la spéciation. Le principal objectif de cette thèse est d'évaluer l'impact des traits d'histoire de vie des espèces sur la spéciation. Nous avons choisi d'étudier 20 espèces de poissons marins subdivisées en deux lignées (Atlantique et Méditerranéenne), et présentant une large diversité de niveaux de divergence et de traits d'histoire de vie. Ces traits sont supposés impacter l'intensité de la dérive génétique, les capacités de dispersion et le temps de génération des espèces. Le contrôle par une histoire biogéographique et une architecture génomique commune à toutes les espèces nous permet de tester spécifiquement le rôle des traits d'histoire de vie sur plusieurs mécanismes évolutifs intervenant dans la spéciation. Dans le premier chapitre, nous avons étudié les déterminants de la diversité génétique, substrat sur lequel s'établit la divergence lors de la séparation initiale des lignées. Nous avons observé que la longévité adulte des poissons marins est corrélée négativement à la diversité génétique, et nous avons démontré que cette relation pouvait s'expliquer par une plus grande variance du succès reproducteur chez les espèces longévives à cause de stratégies reproductives particulières aux poissons marins (forte mortalité juvénile, faible mortalité adulte et augmentation de la fécondité avec l'âge). Puis, dans un second chapitre, nous avons détecté une grande diversité d'histoires évolutives entre espèces, caractérisée par un fort gradient de divergence génétique entre lignées atlantiques et méditerranéennes. Ce gradient reflète en partie le niveau de semi-perméabilité des génomes. Les espèces à faible différentiation présentent un isolement reproductif faible, alors que les espèces les plus fortement différenciées montrent un isolement reproductif quasi-complet. Les traits d'histoire de vie des espèces expliquent en partie cette diversité de niveaux d'isolement via différents mécanismes. La durée de vie larvaire influence négativement la différenciation génétique en modulant les capacités de dispersion, l'effet de la taille du corps indique un effet négatif de l'abondance long-terme sur la divergence, et la longévité semble impacter le nombre de générations écoulées depuis la séparation ancestrale. Enfin, dans un dernier chapitre, nous avons montré que les patrons de divergence détectés sur le génome nucléaire se reflétaient en partie sur les génomes mitochondriaux. En conclusion, les 20 espèces étudiées présentent une variabilité surprenante d'histoires évolutives au regard des similitudes de leur histoire biogéographique et leur architecture génomique. Les relations entre traits d'histoire de vie et histoire évolutive des espèces sont complexes, mais nous avons pu éclairer certaines d'entre elles en décomposant l'implication des traits dans les différentes étapes de la spéciation. L'application de l'approche de génomique comparative développée au cours de cette thèse dans d'autres zones de suture permettra d'étendre nos connaissances des déterminants du tempo et du mode de la spéciation.



**Mots-clés:** spéciation, poissons marins, traits d'histoire de vie, diversité et divergence génétique, zone de suture atlantico-méditerranéenne.

**Abstract**

Speciation is the evolutionary process through which a species splits into two lineages that diverge and accumulate reproductive barriers, until complete reproductive isolation is achieved. During this process, the diverging lineages can still exchange genes by hybridisation, but gene flow is progressively restricted by the accumulation of barriers. This results in semi-permeable genomes, whereby some loci exchange freely between lineages and remain undifferentiated while others do not introgress, thus contributing to the establishment of divergent genomic regions, called genomic islands of speciation. These barrier loci may be involved in different types of isolating mechanisms, including mate choice, adaptation to different environments, or genetic incompatibilities between co-adapted genes. The study of the establishment, accumulation, erosion and maintenance of these barriers and their effects on the semipermeability of the genomes of lineages undergoing speciation helps to understand how new species are formed. The advent of high-throughput sequencing techniques has made it possible to characterise the genomic landscape of divergence in multiple lineages undergoing speciation across the tree of life. These studies have shown the influence of the demographic history and genomic architecture as major determinants of the genomic landscape of divergence. However, other factors could intervene and explain the diversity of evolutionary trajectories that may or may not lead to speciation. The main objective of this thesis is to assess the impact of species' life history traits on speciation. We have chosen to study 20 marine fish species subdivided into two lineages (Atlantic and Mediterranean), and presenting a wide diversity of degrees of divergence and life history traits. These traits are thought to impact on the intensity of genetic drift, dispersal abilities and generation time of the species. Controlling for a shared biogeographic history and genomic architecture across species allowed us to specifically test the role of life history traits on several evolutionary mechanisms involved in speciation. In the first chapter, we studied the determinants of genetic diversity, the substrate on which divergence is built during the initial separation of lineages. We observed that adult longevity of marine fishes is negatively correlated with genetic diversity, and we demonstrated that this relationship could be explained by a greater variance in reproductive success in long-lived species due to reproductive strategies specific to marine fishes (high juvenile mortality, low adult mortality and increased fecundity with age). Then, in a second chapter, we discovered a great diversity of evolutionary histories between species, characterised by a strong gradient of genetic divergence between Atlantic and Mediterranean lineages. This gradient partly reflects the level of semi-permeability of the genomes. Species with low differentiation show low reproductive isolation, whereas the most highly differentiated species show almost complete reproductive isolation. Species' life history traits partly explain this diversity in isolation levels via different mechanisms. Larval duration negatively influences genetic differentiation by modulating dispersal capacities, the effect of body size indicates a negative effect of long-term abundance on divergence, while longevity seems to impact the number of generations elapsed since ancestral separation. Finally, in a last chapter, we showed that the divergence patterns detected on the nuclear genome were partly reflected on the mitochondrial genomes. In conclusion, the 20 species studied show a surprising variability of evolutionary histories considering the similarities of their biogeographic history and genomic architecture. The relationships between life-history traits and the evolutionary history of the species proved to be complex, but we were nevertheless able to shed light on some of them by decomposing the involvement of traits in the different stages of speciation. The application of the comparative genomics approach developed in this thesis to other suture zones will further extend our knowledge of the determinants of the tempo and mode of speciation.

**Key words:** speciation, marine fishes, life history traits, genetic diversity and divergence, Atlantic - Mediterranean suture zone.

# General contents

**1 Introduction** **1**

**2 Chapter 1** **72**
2.1 Abstract . . . . . 73
2.2 Article . . . . . 75

**3 Chapter 2** **92**
3.1 Abstract . . . . . 93
3.2 Article . . . . . 97

**4 Chapter 3** **140**
4.1 Abstract . . . . . 141
4.2 Article . . . . . 143

**5 Discussion** **159**

**6 Annex of chapter 1** **183**

**7 Annex of chapter 2** **223**

**8 Annex of chapter 3** **265**

**9 Annex 4** **279**
9.1 Abstract . . . . . 280
9.2 Article . . . . . 283

# INTRODUCTION

# Introduction contents

**1 The species and the classification of living organisms 7**

1.1 The first classifications . . . . . 7

1.2 Species concept in the XX$^{th}$ century: what is a species ? . . . . . 8

**2 Speciation: studying the evolution of reproductive isolation barriers 9**

2.1 The nature of reproductive isolation barriers . . . . . 9

2.2 A molecular model of speciation . . . . . 11

2.3 Biogeographical context . . . . . 13

2.3.1 Allopatric speciation . . . . . 13

2.3.2 Sympatric speciation . . . . . 15

2.4 Genetic architecture . . . . . 15

2.4.1 Recombination rate . . . . . 15

2.4.2 Coding genes and functionally conserved elements . . . . . 17

2.4.3 Sex chromosomes . . . . . 17

2.4.4 Structural rearrangements . . . . . 18

2.4.5 Speciation genes . . . . . 18

**3 Exploring speciation in genome sequences 19**

3.1 Molecular markers and speciation . . . . . 19

3.2 The search for *islands of speciation* . . . . . 20

3.3 A mosaic of different genealogies . . . . . 21

**4 The speciation continuum 22**

**5 The sequential components of speciation 23**

5.1 Ancestral genetic diversity: the substrate of reproductive barriers . . . . . 25

5.2 Linked selection . . . . . 25

5.3 Complete and incomplete lineage sorting: differential fixation of alleles . . . . . 26

5.4 Introgression: the exchange of divergent genetic material . . . . . 26

5.5 Divergence: the accumulation of molecular differences . . . . . 27

**6 Impact of demographic parameters on speciation and their link with life-history traits 27**

6.1 How should effective population size ($N_e$) impact speciation? . . . . . 27

6.1.1 Time to fixation of neutral and advantageous mutations . . . . . 28

6.1.2 Probability of fixation of mutations . . . . . 29

6.1.3 Deleterious mutations accumulation and introgression: the *hybridization load* . . . . . 29

6.1.4 Lineage sorting . . . . . 31

6.1.5 Linked selection . . . . . 32

6.1.6 Life history traits determinants of $N_e$ . . . . . 33

6.2 How $m$ should impact speciation ? . . . . . 34

6.2.1 Genetic homogenization . . . . . 34

6.2.2 Width of a cline in a hybrid zone and migration-selection antagonism . . 36

6.2.3 Life history traits determinants of $m$ . . . . . 36

6.3 How $T$ should impact speciation dynamics ? . . . . . 37

6.3.1 The speciation clock . . . . . 37

6.3.2 Accumulation of genetic incompatibilities . . . . . 37

6.3.3 Divergence . . . . . 38

6.3.4 Life history traits determinants of $T$ . . . . . 39

6.4 Other life-history traits . . . . . 39

6.4.1 Mate choice and assortative mating . . . . . 39

6.4.2 Parental care and the evolution of postzygotic barriers . . . . . 40

**7 Life-history traits and speciation: the current state-of-the-art 41**

**8 This thesis: the impact of life-history traits on speciation in the Atlantic Ocean - Mediterranean Sea suture zone across 20 teleostean marine fish species 42**

8.1 Speciation in marine fishes . . . . . 43

8.2 The Atlantic Ocean - Mediterranean suture zone . . . . . 44

8.3 Relationships between fish life-history traits and demographic parameters . . . . 45

8.4 Methodological plan . . . . . 47

# Glossary

BGS : Background selection
BSC : Biological Species Concept
BDM : Bateson-Dobzhansky-Muller
DNA : Deoxyribonucleic Acid
$d_a$ : net divergence
$d_{XY}$: absolute divergence
$F_{ST}$: fixation index, a measure of population differentiation
HPA: Histidino-phosphate amino-transferase
I: mean number of incompatibilities
K: number of substitutions
$m$ : migration rate
m-RNA : Messenger RNA
$m_e$ : effective migration rate
$N_e$ : effective population size
$N_e m$ : gene flow
$\pi$ : genetic diversity
$p$ : allele frequency
$P_{fixation}$: probability of fixation of a mutation
PLD : Pelagic Larval Duration
$P_d$: probability of fixation of deleterious mutations
$p_T$ : probability of fixation of an advantageous mutations
RAD-seq : Restriction site-associated DNA sequencing
RI : reproductive isolation
$s$ : coefficient of selection
SNPs : Single Nucleotide Polymorphism
$t$ : time
$T_{fixation}$ : time of fixation of a new advantageous mutation
$\theta$ : $4N_e\mu$
$\mu$ : mutation rate
$V_k$ : variance in reproductive success

*Species arise from other species, and during this process, there will be unclear cases, no matter now one define species. [...] In fact, one could consider speciation as the conversion of genotypic clusters species into biological species, a process that is continuous, yielding ever-increasing barriers to gene flow (Coyne and Orr, 2004).*

# Introduction

## 1 The species and the classification of living organisms

### 1.1 The first classifications

A naturalist observation of the variety of living organisms that constitute biodiversity leads to the consideration that all individuals differ from each other in several manners: their size, the ability to move, the habitat they occupy, the sound they make, the way they reproduce and other morphological, behavioral and physiological characters. A closer look leads to the observation that some groups of individuals look more alike with each other than with individuals from another group. This observation has led numerous scientists to propose a classification of the biodiversity. Aristoteles (384-322) was one of the first to propose such a classification in the IV$^{\text{th}}$ century before our era based on a thorough examination of numerous organisms (Aristoteles, 1883). His classification did not rely on a taxonomic choice but rather on morphological criteria that made him split warm-blooded organisms like mammals from cold-blooded like mollusks. If he was the first to classify individuals in species and genera, he thought that there is a continuity between any living organisms. In the XVIII$^{\text{th}}$ century, with the multiplicity of naturalist travels all around the world, Carl Linnaeus (1707 - 1778) published a new classification in this *Systema naturae*: especially he proposed a nested hierarchy with five main ranks (from upper to lower, kingdom, class, order, genera and species). He also proposed the binominal species identification still used today. However, this classification based on morphological characters can lead to errors: for example, Linnaeus first classified the female and male mallards *Anas platyrhynchos* in two separate species based on their different plumage coloration (Diamond, 1992). Moreover, the classification did not rely on evolutionary classifications and no explanation was provided to explain species formations except religious ones.

## 1.2 Species concept in the XX$^{th}$ century: what is a species ?

In this seminal work, Darwin (1859) proposed an evolutionary model to the species problem: because of gradual small changes occurring over time and then favored by natural selection, some populations could slowly diverge and become two isolated species. This led to a change of *paradigm* of species classification in the late XIX$^{th}$ and XX$^{th}$ century among evolutionary biologists and motivated a great breadth of studies; but at the same time, it raised the issue of a new definition of species that relies on this evolutionary model. If, populations accumulate differences gradually, then when are they considered to be *good* species?

Today, there are many dozens of *species concepts* and each of them lies on different aspects of the nature and evolution of species: some rely on the absence of reproduction between different species such as the Biological Species Concept (BSC) defined by Mayr as *"groups of interbreeding natural populations that are reproductively isolated from other such groups"* (Mayr, 1963), and a relaxed version of the BSC defined by Coyne and Orr (2004) which considers species that have substantial but not necessarily complete hampered reproduction. For instance, one $F_1$ hybrid may be formed between two species every 1000 generations: these species will not correspond to the strict BSC definition but as such a low level of hybridization will not greatly impact the genetic differences between them, we can consider these two taxa as *good* species. Others concepts rather rely on the reciprocal monophyly between species such as The Phylogenetical Species Concept proposed by Cracraft (1983) which defines species as *" the smallest diagnosable cluster of individual organisms within which there is a parental pattern of ancestry and descent"*. Some rely on the independence of evolutionary and ecological processes between different species such as the Evolutionary Species Concept defined by Simpson (1961) in which *"a species is a lineage of ancestral descendant populations which maintains its identity from other such lineages and which has its own evolutionary tendencies and historical fate"*. Finally, other concepts rely on ecological niches such as Van Valen (1976) Ecological Species Concept defined as *"a lineage (or a closely related set of lineages) which occupies an adaptive zone minimally different from that of any other lineages in its range and which evolves separately from all lineages outside its range"*. Finding the criterion to classify every living organism in one species appears elusive and might be irrealistic. A complete description and critics of each species concept and can be found in Coyne and Orr (2004) and Hey (2001). A recent survey shows there are substantial differences between the species concepts used between research disciplines (Stankowski and Ravinet, 2021b): while paleontologists prefer the Evolutionary Species Concept, those who study genetics and genomics prefer the Biological Species Concepts.

A common point to all these concepts is that species arise during the course of evolution by the split of one ancestral species into two diverging lineages; and that lineages progressively

become completely isolated from each other while experiencing different evolutionary processes, having different ecological strategies and morphological differences during the course of their divergence and after reproductive isolation is complete. Beyond, the difficult task of classifying species, we might want to address: how species arise? What are the evolutionary mechanisms that split lineages into distinct species?

# 2 Speciation: studying the evolution of reproductive isolation barriers

## 2.1 The nature of reproductive isolation barriers

Speciation is the evolutionary process that separates one species into two (or more) that can no longer exchange genetic materials - i.e. that are or nearly completely reproductively isolated. During this process, lineages progressively accumulate reproductive isolation barriers that impede gene flow between heterospecific individuals. Barriers between species can be categorized in several categories, depending on whether they act before (*prezygotic isolation*) or after *(post-zygotic isolation*) a sperm cell fertilizes an egg cell (*zygote*) between the two nascent species or whether they result from internal determinants (*intrinsic*) or not (*extrinsinc*).

Extrinsic prezygotic barriers include different reproductive behaviors such as differences in host-plants where individuals reproduce in *Rhagelotis* flies (Feder et al., 1994), temporal isolation such as different years of emergence for *Magicicada* periodical cicadas (Marshall and Cooley, 2000), asynchrony of larval release between different forms of the stony octocoral *Heliopora coerulea* (Villanueva, 2015), different life histories of 2-year semelparous pink salmon (*Oncorhynchus gorbuscha*) populations that reproduce on the same place but not on the same years (Limborg et al., 2014); habitat isolation such as different ecological niches occupied by different monkeyflower plant species *Mimulus* (Sobel et al., 2014); mate choice preferences such as in *Xiphophorus* fish (Crapon de Caprona and Ryan, 1990), pollinator specialization in Mediterranean orchids (Scopece et al., 2007); immigrant inviability, such as reduced survival of immigrants in *Agelenopsis aperta* (Riechert and Hall, 2001).

Intrinsic prezygotic barriers include gametic incompatibility in plants, mechanical isolation because of incompatibilities between male and female genitalia such as in *Drosophila* (Tanaka et al., 2018); gametic isolation that occurs between copulation and fertilization such as removal of heterospecific sperm in Trinididan guppies *Poecilia reticulata* (Ludlow and Magurran, 2006) or in flour beetles (*Tribolium castaneum*) (Wade and Johnson, 1994).

Intrinsic postzygotic isolation is either expressed as hybrid inviability, such as higher preva-

lence of embryonic letality in hybrids whitefish *Coregonus clupeaformis* (Rogers and Bernatchez, 2006), reduced number of eggs that develop normaly in the grasshopper *Podisma pedestris* (Barton and Hewitt, 1981); or as hybrid sterility such as in hybrids of the house mouse between *Mus musculus* and *Mus domesticus* (Good et al., 2008), between subspecies of the domesticated rice *Oryza sativa subsp sativa* and *japonica* (Kubo et al., 2011) or between *Heliconius melpomene* butterflies lineages (Jiggins et al., 2001).

Extrinsic postzygotic isolation occurs in case of low performance of hybrids maladapted to environmental conditions such as in the stickleback (Thompson et al., 2022), illustrated also by the lowest survival of immigrants cichlids *Astatotilapia burtoni* (Rajkov et al., 2018) or lowest reproductive success in hybrids of two diverging pine-sawfly *Neodiprion* with diverging host-preference and reproductive trait morphology that do not match in hybrids (Bendall et al., 2017).

Multiple barriers to gene flow are often present at the same time and thus collectively participate to reproductive isolation, especially in the late-stage of speciation such as between two species of snapdragons plants *Antirrhinum* with different flowering phenology, different pollinator species and incompatibilities between heterospecific pistil and pollen (Carrió and Güemes, 2014). Another example is given by the divergence of adult host-preferences combined with lower hybrid larval performance and lowest survival of offspring that migrates on the other host in the beetle *Galerucella nymphaeae* (Pappers et al., 2002). Thus, total reproductive isolation can be defined as the cumulated independent effects of each barrier that reduce overall heterospecific gene flow during the entire life cycle (Coyne and Orr, 2004; Mallet, 2006; Stankowski and Ravinet, 2021a). The relative importance of each type of barrier in speciation remains controversial: some argue that intrinsic barriers are more efficient to promote reduction to gene flow at the beginning of the speciation process (Irwin, 2020) while others argue that they play a role at all stages of speciation (Coughlan and Matute, 2020). Assessing the relative importance and the sequential appearance or different types of barriers is a difficult task as the barriers with contemporary stronger effects on the reduction of the effective migration rate might not be the first barriers that appeared in the genome. Moreover, additional barriers to gene flow can appear after reproductive isolation was completed (Coyne and Orr, 2004).

Hence, multiple kinds of barriers that promote reproductive isolation between diverging lineages have been described in a myriad of taxa; but it does not explain how genetic changes at the molecular level can promote the evolution of barriers and conversely, how the presence of barriers promotes genetic differentiation.

## 2.2 A molecular model of speciation

DNA is the material basis of hereditary changes. It is a macromolecule composed of two antiparallel strands that form a double helix. Each strand is a linear stretch of nucleotides that can contain four possible bases: adenine (A), cytosine (C), guanine (G) and thymine (T). The complete molecular structure of the DNA was found by Waston, Crick and Rosalind Franklin in 1953 (Watson and Crick, 1953), nearly one century after Charles Darwin's *Origins of Species* was published in 1859. DNA molecules are frequently organized in *chromosomes* either linear in eukaryotes or circular in prokaryotes and the complete set of chromosomes constitutes the *genome*. The genome is composed of *genes* that contains the coding information to code proteins and *non-coding* DNA that includes regulatory elements, pseudogenes or selfish elements such as transposons and repeated sequences. Genome structure varies widely between species. Variation in size: genome size varies between 10 000 base pairs for viruses to 3.5 billion bases for modern humans (*Homo sapiens*, and much larger in some species); variation in the number of genes: the majority of prokaryotes have a maximum of 8 000 genes whereas nearly all animals and plants have more than 13 000 genes (Lynch, 2007); variation in the extent of *non-coding* DNA only 1-2% of the human genome is coding DNA (International Human Genome Sequencing Consortium, 2004) while prokaryotes and viruses contain more than 80% of coding DNA. DNA molecules represent a trade-off between stability and alterity; stability because DNA molecules that are present in the germline of individuals are replicated and transmitted nearly identically to the offspring either through asexual or sexual reproduction to ensure normal offsprings development; alterity because there are some differences between adult and offsprings DNA because of *mutations* that change DNA sequence or structure in the adult germline cells and that are transmitted to their offsprings. *Mutation* is an evolutionary processes that creates new genetic compositions, defined as *genotype* (Watson et al., 2013). Mutations can be called substitutions if one base is replaced by another, deletion or insertion if one or several bases are removed or inserted, respectively. They can be induced by environmental conditions but they are mainly caused by the imperfect fidelity of the cell replication machinery. Population genetics study the evolution of frequencies of new mutations across time such as whether they disappear or become fixed within the population or maintained at some evolutionary stable state. For example, an error in the replication that mutates a nucleotide at a given position from A to G will initially be present at a frequency of $\frac{1}{N}$, where $N$ is the number of haploid individuals. The evolution of the frequency of the G mutation and its fate - whether it will disappear or become fixed in the population - across generations will depend on the random sampling of genes, a process called *genetic drit* (Wright, 1931; Charlesworth, 2009) and on *selection*, that is, if the mutation is advantageous of disadvantageous with respect to survival and fecundity (lifetime fitness) of individuals that carry this mutation.

How do these evolutionary forces promote the evolution of reproductive isolation barriers? (Wu, 2001) proposed an evolutionary model called the *genic view of speciation*. First, mu-

tations generate novel genotypes at a small number of loci. These new genotypes can have no impact on phenotypes and can be neutral. However, novel mutations could also cause the emergence of new phenotypes that may participate in any kind of barrier introduced above. Alternatively, the new mutations that appeared in different genetic backgrounds can reveal incompatible once combined in the same genome following hybridization because the change in gene sequences have also changed the corresponding protein sequence in a way that alters the normal function metabolic process and subsequently reduces the fitness of an individual that would carry both mutations. In this situation of incomplete reproductive isolation explained by a few loci, genes between the two populations can be freely exchanged except in the small regions of the genome around these mutations. Thus, it results in a reduced overall gene flow between the two populations that can foster the fixation of further new mutations that contribute to strengthening reproductive isolation. These mutations can emerge in the vicinity of already established barrier loci therefore creating larger contiguous barrier regions but also on other parts of the genome thereby introducing new barrier loci that continues to reduce overall gene flow (i.e., genomes may become *congealed*, Feder et al. (2014)). This evolutionary process continues until barrier locus are widespread enough in the genome that reproductive isolation is complete and that the two populations can be considered as biological *species*. The accumulation of barriers to reproduction continues even after gene flow is completely stopped. This model has two fundamental consequences for speciation research. The genic view of speciation defined the gene as the units of speciation: it recognizes that while species are being formed, the genome is *semi-permable* - i.e., some regions of the genome are genetically differentiated while others remain genetically similar.

A lot of empirical studies has confirmed the semi-permeability nature of incipient species boundary (Turner et al., 2005; Teeter et al., 2008; Ellegren et al., 2012; Andrew and Rieseberg, 2013; Bay and Ruegg, 2017; Ravinet et al., 2018; Duranton et al., 2018). Regions of the genome strongly differentiated were coined speciation islands or differentiation islands, as a metaphor of *islands* of differentiation in a *sea* of free interbreeding, and later on, these "islands" grow until they become *continents*. This results in the variation of the intensity of gene flow caused by variation in the strength of barrier to gene flow along the genome. If two populations exchange $m$ migrants per generation and if we were able to estimate the effective migration rate ($m_e$) along the genome, the strength of a barrier to gene exchange at a particular locus is (Barton and Bengtsson, 1986; Stankowski and Ravinet, 2021a):

$$b = \frac{m}{m_e} \tag{1}$$

Under the genic view of speciation model, studying the causes of speciation amounts to understand why barriers to gene flow appear in the genome and how they accumulate until reproductive isolation is complete. In a first coarse approach, we can classify these causes in

two categories: those that are relevant to the *ecological and biogeographical context* of speciation and those that depend on the species *genomic architecture*.

## 2.3 Biogeographical context

Historical classification of the biogeographical context of speciation has been made using the levels of *gene flow* ($Nem$) when the first reproductive barriers appear: *allopatric speciation* when gene flow is absent due to physical barriers such as geographic distance or *sympatric speciation* where gene flow is unimpeded by such barriers.

### 2.3.1 Allopatric speciation

*Allopatric speciation* occurs when gene flow is absent due to some sort of geographical isolation ($N_e m = 0$). This mode of speciation was historically considered to be prevalent (Mayr, 1963). A simple and famous model that can explain the establishment of reproductive isolation loci in allopatry is the Bateson-Dobhzansky-Muller incompatibilities (BDM) model (Bateson, 1909; Dobzhansky, 1936; Muller, 1942) due to the negative epistatic interaction between alleles from different genetic backgrounds in hybrid individuals. During allopatric speciation, the separated populations accumulate different mutations that are either neutral or advantageous within each population. However, if these populations come in secondary contact, these mutations may reveal incompatible (i.e., they have not been tested for compatibility before) and lead to a reduction of fitness in hybrids that carry both mutations in their genome. This model is remarkably simple and needs few assumptions: the establishment of mutations does not require any adaptation to local environments, only some period of allopatry. Importantly, BDM model offers a simple solution to explain how barrier loci can be established without being purged: they have a deleterious effect only when different mutations are combined within a hybrid individual. Taking the metaphor and mathematical model of *fitness landscapes*, BDM model explains how one population can "move" from one *peak* of fitness to another peak without crossing a *valley* of low fitness that separates these two peaks. This geographical mechanism of speciation seems plausible for two reasons: i) first, various types of physical barriers exist that are able to split two lineages with no gene flow, such as mountains or rivers, ii) the model of accumulation of incompatibilities does not rely on selection and does not require ecological divergence between the two lineages. There are many examples of speciation that seem to have started in the absence of gene flow such as species that splits in different glacial refugia during the Last Glacial Maximum in the Pleistocene and later came back into contact (Avise et al., 1998; Hewitt, 2000) or marine organisms that were separated between the Pacific and the Atlantic Ocean after the closure of the Isthmus of Panama (Knowlton and Weigt, 1998).

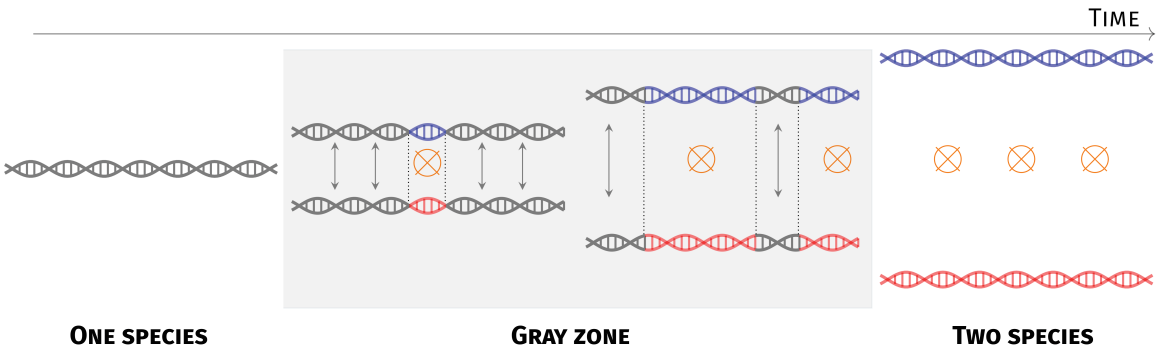



Figure 1: **The evolution of barriers to reproduction during speciation**. Speciation starts from one species represented here on the left by a single panmictic population, with free gene exchange and no differentiation (gray colors). Once the first barriers to reproduction appear (orange circle and crossbar), gene flow is reduced or absent around the barrier loci but is not impeded in the rest of the genome (gray arrows). The first reduction of gene flow drives the widening of the first barrier and new barrier locus on different regions of the genome. Once barriers become widespread, gene flow is totally impeded, reproductive isolation is complete and the two lineages can be considered as two species. During the speciation process, populations start to be differentiated from each other but can continue to reproduce and form hybrids (gray zone of speciation, Roux et al. (2016)).
.

### 2.3.2 Sympatric speciation

A second model called sympatric speciation considers free gene flow between populations at the onset of divergence ($N_e = 0.5$). In this case, the apparition of barriers to gene flow is caused by divergent selection and local adaptation, where alleles are differentially favored between populations because they provided adaptations to different local environmental conditions (Via and West (2008), a process called *ecological speciation* Nosil (2012)). For example, two species of *Howea* palms on Lord Howe islands separated after island formation by evolving differences in habitat preference and flowering time (Savolainen et al., 2006); North-American flies *Rhagoletis* that lays eggs in apple trees diverged from the ancestral populations that reproduce on hawthorn (Feder et al., 1994). Sympatric speciation was also demonstrated in a controlled experiment where *Drosophila melanogaster* individuals started to diverge because of different habitat choice for reproduction as an example of extrinsic prezygotic barrier (Rice and Salt, 1990). Proponents of the sympatric speciation model argue that divergence could be possible if reproductive isolation evolves as a correlated character, for example between habitat choice and mating preference with conspecific individuals. However, sympatric speciation could not be possible if the strength of diverging selection is not sufficiently strong to counteract the effect of gene flow and recombination that break up the association local adaptation and assortment loci (Felsenstein, 1981). Contrary to allopatric speciation, islands of speciation in the case of sympatric speciation as supposed to be limited to few large clusters of loci, which is necessary to alleviate gene flow and recombination (Nosil et al., 2009). Beyond these, *parapatric speciation* represents intermediate levels of gene flow ($0 < N_e < 0.5$).

## 2.4 Genetic architecture

Several characteristics of the species genome can have an impact on speciation: recombination, conserved functional elements coding sequences, structural rearrangements, sex chromosomes and particular gene functions that could play a disproportionate role in speciation.

### 2.4.1 Recombination rate

Recombination is the exchange of genetic material between homologous chromosomes during meiosis that creates novel combinations of alleles (Peñalba and Wolf, 2020). In eukaryotes, recombination rate is highly variable along the genome at both fine (kilobase scale) and broad-scale (megabase). In most plants and animals, broad-scale variation is characterized by a U-shaped distribution along chromosomes: higher near the telomeres and lower in the center (Haenel et al., 2018). Recombination rate variation along the genome can affect speciation in two ways.

In a neutral world, recombination rate variation will have no effect on genomic variation in genetic differentiation beyond the effect of autocorrelated gene genealogies at a local scale. However, the effect of selection on a particular locus will affect the neighborhood of that locus to an extent that depends on linkage. The neutral mutations close to selected loci will experience an attenuated indirect effect of selection through linkage with the selected loci until they become not linked. This effect called *linked selection* acts either with negative selection (background selection) (Charlesworth et al., 1993) (constant removal of deleterious variants) or positive selection (selective sweeps) (Maynard Smith and Haigh, 1974) (adaptive fixation). As selection reduces genetic diversity, a practical consequence of this evolutionary process is a higher depletion of genetic diversity in regions of low recombination rate because of higher linkage between neutral and selected loci. As differentiation will be increased in regions of low genetic diversity (Charlesworth, 1998), regions of low recombination rate will be prone to elevate genetic differentiation between incipient lineages (Burri, 2017). The impact of linked selection can be observed even long after reproductive isolation is complete: notably, phylogenetically closed species should display correlated landscapes of differentiation along their genome, as it has been found in birds (Vijay et al., 2017) - i.e. homologous regions of the genome show higher or lower genetic differentiation. These last results can be observed if three requirements are fulfilled (Burri, 2017): i) variation in recombination rate along the genome is the same between the two species, ii) genome-wide deleterious mutation rate is conserved between the two species, iii) differentiation within each species is mainly driven by faster lineage sorting and higher accumulation of differentiation in regions of low $N_e$. Considering the relatively slow effect of linked selection on genetic diversity, we might expect that the correlation levels between species landscapes will be strong for within-species pairs that have diverged for a certain time. On the other hand, as divergence grows, differentiation determinants might be different than only differential lineage sorting imposed by variation in recombination rate. Hence, correlated landscapes of genetic differentiation between species could be highest for intermediate stages of divergence (Burri, 2017).

Secondly, the local recombination rate is expected to have a role in the differential gene exchange between regions of the genome and the maintenance of reproductive barriers. If barriers to reproduction are polygenic and widespread in the genome, selection will be more efficient to remove foreign ancestry in regions of low recombination because deleterious mutations in foreign non-recombinant blocks will remain unbroken. Thus, under this model, a positive correlation between local recombination rate and local introgression should be observed as for instance in *Heliconius* butterflies (Martin et al., 2019), European sea bass (Duranton et al., 2018) or maize (Calfee et al., 2021). However, the relationship can be reversed in the presence of coadapted variants that are advantageous in the introgressed population if they remain in strong linkage disequilibrium, as observed between wild maize and teosinte (Calfee et al., 2021), or because of local heterosis caused by associate overdominance (Leitwein et al., 2019).

Finally, the effect of linked selection can also vary between species with different aggregate recombination rates in the genomes: species with a low number of chromosomes such as *Drosophila* will experience stronger linked selection because of a lower aggregate recombination rate (Veller et al., 2021).

### 2.4.2 Coding genes and functionally conserved elements

Barriers to gene flow are more frequently found in coding and conserved regions than in the rest of the genome since genetic incompatibilities can appear at a low molecular divergence in these regions and diverging selection acts preferentially on coding sequences as they have a direct effect on phenotypes and biological functions. Gene flow is reduced in genome regions with high density of functional elements between teosinte and maize (Calfee et al., 2021), *Histoplasma* fungi lineages (Maxwell et al., 2018). Other examples come from lower cline widths in regions of high gene density in the house mouse (Teeter et al., 2008) and higher differentiation in high genomic regions of *Ficedula* flycatchers (Burri et al., 2015).

### 2.4.3 Sex chromosomes

Analyzing the results from several studies that span several diverse taxa including insects, mammals and birds, Haldane (1922) observed that some crosses lead to inviability or sterility only in one sex. Comparing these data with the sex determination, he proposed that *"when in the $F_1$ offspring of a cross between two animal species or race one sex is absent, rare, or sterile, that sex is always the heterozygous sex"*, which is known as the *Haldane's rule* and was confirmed and precised later (Schilthuizen et al., 2011). This led researchers to consider that sex chromosomes might play an important role in establishing barriers to reproduction. In species with XX/XY sex determination (where males are hemizygous), such as in mammals and some flies, ZZ/ZW (where females are hemizygous) such as in birds and some butterflies, the Y and W chromosome show respectively reduced gene flow and higher genetic differentiation such as in *Drosophila* (Turissini and Matute, 2017), *Heliconius* butterflies (Martin et al., 2013) and *Anopheles gambiae* species complex (Fontaine et al., 2015). Several hypotheses can explain the important role of sex chromosomes in establishing barriers to reproductive isolation: the presence of X or W-linked genes that generates hybrid sterility, the lower effective population size compared to the autosomes that might increase differentiation or the arrest of recombination in the heterogametic sex that may facilitate the emergence, spread and coupling of barrier loci bearing for instance mate choice or genetic incompatibilities.

### 2.4.4 Structural rearrangements

Aside from single nucleotide substitutions, mutations can also change the structure of the genome such as with insertions, deletions, gene or whole-genome duplications, chromosomal inversions and translocations. All these forms of structural variation can promote the evolution of barriers to gene flow.

For example, the histidino-phospate amino-transferase gene (HPA) codes for a protein that participates in histidine biosynthesis. In *Arabidopsis thaliana*, this gene has been duplicated on separate chromosomes HPA1 and HPA2. In different lineages, one of the two duplicates has been silenced: thus 25% of $F_2$ hybrids between lineages that silenced differents duplicates show severe embryo lethality (Bikard et al., 2009).

Chromosomal inversions are structural rearrangements where one DNA segment is reversed compared to its original orientation of the chromosome (Fuller et al., 2017). They were first discovered in 1921 after discovering that variants appeared in different orders in the chromosomes of *Drosophila melanogaster* and *Drosophila simulans* (Sturtevant, 1921). Homozygous individuals on the inversions do not show reduced fitness: however, heterozygous individuals that carry both the inverted and the non-inverted segments can suffer from sterility or inviability because of problems in chromosome pairing during meiosis. Moreover, crossing over and recombination are suppressed in the inversions. These two characteristics made inversions good candidates to barrier to reproduction: because they reduce fitness in $F_1$ hybrids (hybrids dysgenesis) or prevent selection to eliminate alleles that confer hybrid sterility (the recombination model, (Rieseberg, 2001), (Noor and Bennett, 2009), that can help alleviate the selection-recombination antagonism (Trickett and Butlin, 1994). Inversions are at the basis of the *stasipatric* model of speciation where lineages of small population sizes become rapidly isolated because of underdominant chromosomal inversions with large negative effects, as proposed by (White, 1978) analyses of different karyotypic lineages of *Vandiemenella* grasshoppers. However, inversions are now thought to have a stronger effect on speciation via their effect on recombination ((Rieseberg, 2001),(Faria and Navarro, 2010)). They have been mapped and their role in reproductive isolation has been supported in various organisms such as *Anopheles gambiae* (Coluzzi et al., 1985; Fontaine et al., 2015), *Littorina saxatilis* periwinkle (Faria et al., 2019), sea-weed fly *Coelopa frigida* (Berdan et al., 2021), the marine fish capelin *Mallotus villosus* (Cayuela et al., 2020).

### 2.4.5 Speciation genes

Finally, one can aim to identify particular *speciation genes*, that *"contributes to the splitting of two lineages by reducing the amount of gene flow between them"* (Rieseberg and Blackman, 2010). However, proving that one particular gene has an impact on speciation remains a

daunting task and requires several proofs: i) that the gene has an impact on hybrid fitness which needs genetic transformation for validation and that ii) gene divergence at this gene started during speciation before reproductive isolation is complete (Blackman, 2016). Thus, some speciation genes have been identified, generally in model organisms (Presgraves, 2010), where experimental manipulation is possible. Powell et al. (2020) identifies genetic incompatibilities between the gene *xmrk* and *cd97* that diverged and generate melanoma that reduce defense against predators and survival between swordtail fish species *Xiphophorus*. Another example is found in flowering plants where a loss of function in the *ANTHOCYANIN2* (AN2) transcription factor gene that mediates flower color drives pollinators specialization and prezygotic barriers to reproductive isolation in two *Petunia* plants (Hoballah et al., 2007).

# 3 Exploring speciation in genome sequences

An analysis of nucleotide sequence variation of several individuals from different populations is mandatory to study speciation and the evolution of reproductive barriers. Evolutionary processes that affect each population will leave their footprint on the genome (e.g., Tajima (1989)). Thus, the analysis of genome sequences might help to decipher which evolutionary or demographic forces act on populations based on statistical inferences from observation of genome patterns. DNA sequences can be compared as a *palimpsest*: a manuscript whose writing was washed off in order to be reused. Similarly, demography, drift, selection and recombination constantly shape genetic variation in the genome. Thus, analyzing patterns of empirical nucleotide diversity can help to answer what were the evolutionary forces that acted on these lineages. Molecular markers differ in resolution and genome coverage (Gagnaire, 2020). Thus, the choice of molecular markers has also consequences on genetic statistical estimation.

## 3.1 Molecular markers and speciation

As in other fields in population genetics, the first markers used to study speciation and characterize genetic distance between individuals were *allozymes* - different alleles or forms of a protein of a same gene. These markers were used for instance by Szymura (1983) to characterize the genetic differentiation between two species of fire-bellied toads *Bombina variegata* and *bombina*. Later on, the development of molecular markers such as *mitochondrial DNA* and *microsatellites* (short tandem repeats) has allowed the development of further analysis of the demographic and historical aspects of speciation (Avise et al., 1983; Salzburger et al., 2002; Natoli et al., 2004; Lemaire et al., 2005). Typically, these data have allowed the detection of cryptic differentiation between phylogeographical lineages and measure differentiation and dispersal. However, these methods yield information for a limited number of loci, which limits information about the variation in differentiation, divergence and diversity along the genome.

*RAD-seq* (Restriction-site-associated DNA sequencing) - the genotyping of short DNA fragments adjacent to a restriction site enzyme - and *transcriptome* - the sequence of mRNA transcripts - has allowed to sequenced more than tens of thousands of independent loci throughout the genome capturing genetic diversity with *SNPs* (Single Nucleotide Polymorphism) - a genetic polymorphism that corresponds to a change of a single base. The large amount of genome-wide polymorphism information available with these approaches has allowed a description of the semi-permeability of the genome, distinguishing variation in genetic differentiation, divergence and gene flow along the genome (Hohenlohe et al., 2010; Gagnaire et al., 2013; Malinsky et al., 2015; Papadopulos et al., 2019). This also allows estimating the demographic history of diverging species including the population size, the gene flow and time of split using new emerging methods that can take into account heterogeneity of selection and introgression (Hey and Nielsen, 2007; Gutenkunst et al., 2009; Roux et al., 2013).

Finally, *whole-genome* sequencing has enabled researchers to fully describe the patterns of nucleotide variation nearly all along one individual's genome. For a long time, a main barrier has been the cost of sequencing a whole-genome. Before 2008, Sanger technologies were the main methods to sequence whole-genomes which was a long and very costly task. For example, one high-quality human genome (around 3.5 billion bases) costed around 300 millions dollars in 2000 ; 150 millions in 2003 ; 20 and 25 millions dollars in 2006. At late 2015, this cost was estimated only at 1500 dollars. This severe reduction of cost in a short period of time had given to opportunity to develop population whole-genomics study to describe patterns of reproductive isolation and semi-permability in various taxa such as *Heliconius* butterflies (Martin et al., 2013), *Ficedula* fly catchers (Ellegren et al., 2012), *Helianthus* sunflowers (Todesco et al., 2020), *Littorina* periwinkle (Johannesson et al., 2020a), *Gastroletus* sticklebacks (Ravinet et al., 2018), *Timema* stick insect (Riesch et al., 2017). Moreover, whole-genome sequences allow the analysis of contiguous haplotypes blocks that inform about introgression (Racimo et al., 2015; Duranton et al., 2018), demography (Harris and Nielsen, 2013) and selection (Garud et al., 2021).

## 3.2 The search for *islands of speciation*

The empirical descriptions of genetic differentiation between diverging species allowed to describe the heterogeneous landscapes of differentiation and divergence as well as to map *differentiation* or *divergence islands*. Hence, a first goal in the field of speciation genomics has been to identify these putative islands of speciation, because they can inform us about the type and extent of genetic barriers that drive reproductive isolation - referred also as differentiation islands or divergence islands. They first were described in *Anopheles gambiae*, where three regions of the genome marked sharp genetic differentiation between two forms M and S that live and reproduce in different habitats while the rest of the genome shows no genetic differentiation (Turner et al., 2005). They used $F_{ST}$ (Wright and Wright, 1984), a well-known estimator of

*genetic differentiation* used in population genetics that estimates the variance component of genetic diversity between populations compared to the total variance in genetic diversity. They measured $F_{ST}$ within windows along the genome and found a majority of windows showing $F_{ST}$ nears 0 meaning no differentiation and a few islands of differentiation where $F_{ST}$ reached 1, which means complete differentiation.

However, later studies showed skepticism about the importance of these islands as representing regions truly involved in reproductive isolation: Cruickshank and Hahn (2014) show that islands previously identified in mosquitoes, butterflies and mice were located in regions of reduced diversity within populations that inflated genetic differentiation, coherent with faster lineage sorting in regions of low recombination rate after complete reproductive isolation. They proposed that, to be considered as real speciation islands, they not only have to represent elevated relative differentiation (as measured by the $F_{ST}$) with the rest of the genome but also elevated absolute genetic divergence (measured by $d_{XY}$). Bay and Ruegg (2017) found that putative islands of differentiation are not correlated to local reduction of effective migration rate, but rather to selective sweeps in one population and introgression in the other populations upon secondary contact in two lineages of Swainson's thrush. Also, speciation islands may represent regions of low recombination such as chromosomal rearrangements or centromeric regions with elevated differentiation compared to the rest of the genome because of low genetic diversity (Noor and Bennett, 2009). Moreover, some studies rely on $F_{ST}$ outliers of the genomic distribution but without knowing if these outliers are significant compared to a null model or matching the highest values of the distribution of $F_{ST}$ simply due to the stochastic nature of the coalescent during divergence. Ravinet et al. (2017) summarized the complex interplay of many evolutionary processes that shape differentiation and divergence landscapes such as variation in recombination rate, gene densities and demographic history. The main obstacle is that incomplete lineage sorting will create variation in genetic differentiation and divergence that could mimic variation in gene flow and reproductive isolation along the genome, thus creating many false islands of speciation. A clear and comprehensive view of demographic and selective forces that shape genomic differentiation and divergence landscapes is thus mandatory to assess the location of barriers to gene flow along the genome and to quantify the amount of reproductive isolation between two nascent species (Lohse, 2017).

## 3.3 A mosaic of different genealogies

Recently, new methods have been developed to answer these questions. Hybridization between species leads to the exchange of blocks of ancestry in the foreign genetic background: these blocks are broken down by recombination through time (Liang and Nielsen, 2014). An analysis of the distribution of block length distributions can understand the timing of genetic admixture, the proportion of introgression and the regions of the genome that shows no in-

trogression (Turissini and Matute, 2017; Duranton et al., 2018). Individuals present different mosaics of blocks of ancestry that reflects the levels of admixture represented by the mixture of different genealogies: regions of the genome with foreign local ancestry are more genetically similar to individuals of the other populations while other regions of the genome that represent barriers to gene flow are genetically more related to individuals from the same populations. Recently, new methods have been developed to infer the variation in gene genealogies along the genome such as `ARGweaver` (Rasmussen et al., 2013), `RELATE` (Speidel et al., 2019) and `tsinfer` (Kelleher et al., 2019) from whole-genome sequences. Along with previously developed approaches, the analyses of the inferred mosaic of genealogies offer promising approaches to estimate the multiple facets of speciation. However, beside theses methods, a conceptual framework is also needed to study speciation.

# 4 The speciation continuum

The *speciation continuum* is a verbal model that represent speciation as a continuous gradient of *reproductive isolation* (RI) ranging from no reproductive isolation at the start of the continuum (RI = 0) to complete reproductive isolation on the end of the continuum (RI = 1) (Stankowski and Ravinet, 2021a). This verbal model is embedded in the gradualist view, already acknowledged by Darwin in the late XIX$^{\text{th}}$ century, that species evolves through slow and gradual changes imposed by natural selection. It fits well with the previously defined *genic view of speciation* (Wu, 2001) that states that speciation starts with few loci generating partial reproductive isolation and then larger, more widespread blocks accumulate through the genome during the progress of speciation (Nosil et al., 2009).

During the first stages of the continuum, it is easy to confound early-stage speciation with population-level genetic differentiation (Coyne and Orr, 2004), because any geographical barriers to gene flow can either produce reproductive isolation or merely genetic differentiation that is inversely proportional to gene flow (Wright, 1931). To circumvent these problems, it is necessary to detect, at least, one reproductive barrier to gene flow, even with low genetic divergences and genetic differentiation at the whole-genome scale.

Verbal representations and metaphors in biology can be useful to describe complex processes with simple words, but they can be also misleading because of their over-simplicity, and so does the speciation continuum. It does not imply that every speciation events occur in the same way, i.e. that the reproductive barriers and build-in of barrier loci follow the same path. It does also not imply that the accumulation of reproductive isolation is linear and irreversible, nor that absolute time to complete speciation is the same between speciation events. Finally, it is not a continuum of genetic differentiation nor divergence (Stankowski and Ravinet, 2021a) because species can achieve complete reproductive isolation for varying levels of genetic differentiation

and divergence. Comparing 61 independent divergent lineages that present different levels of reproductive isolation, Roux et al. (2016) set the bounds of the *gray zone of speciation* between 0.5 and 2% of net synonymous divergence: within this interval of divergence there are pairs of species that are strongly isolated while others still experience heterospecific gene flow.

In spite of these limitations, the model of speciation continuum can be useful to understand the factors affecting the path to complete reproductive isolation. A major difficulty in the study of speciation is that it occurs on a long timeframe that meaning that it is impossible to study a complete speciation event from the beginning to complete reproductive isolation. In the speciation continuum framework, one can consider each different sister species as a different step in the speciation continuum that corresponds to a different accumulation of reproductive isolation barriers. By comparing the demographic and selective forces or by measuring the different barriers to reproduction that acted upon different species that represent different stages of the speciation continuum, it is then possible to determine the evolutionary forces that drive speciation.

For example, this approach has been developed to detect that sympatric flies *Drosophila* accumulated more prezygotic barriers to gene flow than allopatric ones (Coyne and Orr, 1989), and to further enhance the role of reinforcement in *Drosophila* speciation (Yukilevich, 2012). It also allowed to demonstrate the pervasive impact of introgression between different species of *Heliconius* butterflies (Martin et al., 2013), a positive association between habitat difference and postmating reproductive isolation (Funk et al., 2006), to assess the importance of intrinsic postzygotic barriers and several ecological prezygotic barriers in 18 diverging *Streptanthus* jewel flowers lineages (Christie and Strauss, 2018). All these studies might lead us to ask: what are the evolutionary forces and processes that explain the evolution of reproductive isolation between species along the speciation continuum?

# 5 The sequential components of speciation

Speciation is a complex process that is affected by a range of demographic and selective forces that act from the first appearance of barriers to reproduction to the maintenance and spread of these barriers until reproductive isolation is complete. Thus, it is necessary to analyze these different evolutionary processes and genetic characteristics that may have an impact on speciation: i) the genetic diversity in the ancestral population, ii) the sorting of the variation in each lineage, iii) the acceleration of lineage sorting through linked selection, iv) the homogenization of the two lineages through introgression or/and admixture and the v) accumulation of molecular divergence.

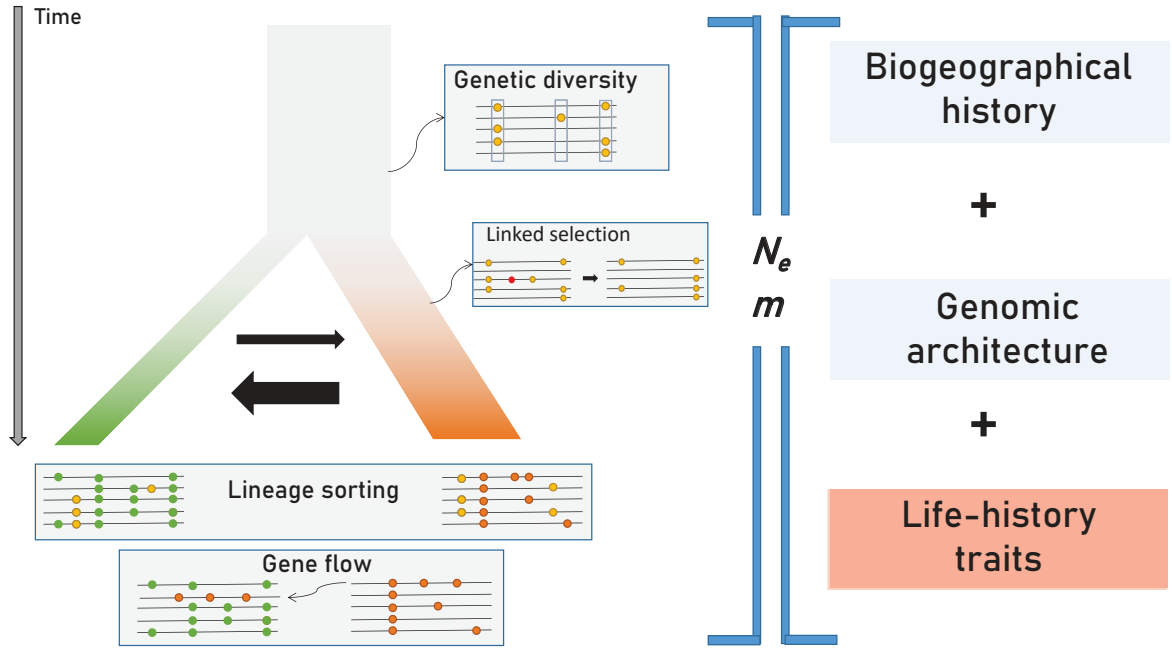



Figure 2: **Multiple evolutionary processes that affect speciation**. Starting from an ancestral population (top, in gray), two populations (green and orange) start to diverge following the appearance of the first barrier effects. The evolution, maintenance or erosion of barriers depend on 4 evolutionary factors: the ancestral genetic diversity, the rate of lineage sorting by neutral and selective forces (linked selection) and gene flow that results in introgression and admixture. Yellow dots represent ancestral shared diversity that is slowly replaced by green and orange new variants that arise in the left or right population respectively, and that can introgress in the other population. Black horizontal arrows represent migration and the size is proportional to gene flow: here migration is strongly asymmetrical from right to left. Red dots in the linked selection panel represent deleterious mutations that are removed by purifying selection (or background selection) which drives the loss of neutral variants in the vicinity of the counter-selected locus. Here, genomes are semi-permeable: some regions of the genome are undifferentiated because of low genetic diversity or similar genotypic composition because of the maintenance of ancestral variants due to incomplete lineage sorting or introgression; some regions show transient new polymorphisms that are specific to each lineage and that contribute to differentiation, and finally some regions some differential substitutions where all individuals in the same population have the same population-specific mutations. Accumulation of differential fixation and molecular divergence will be also higher for a longer time since ancestral split. All these evolutionary processes are dependent on two demographic parameters: the effective population sizes $N_e$, the migration rates $m$ and generation time, $t$, that themselves depend on three factors: the biogeographic history, the genetic architecture and species life-history traits.

## 5.1 Ancestral genetic diversity: the substrate of reproductive barriers

Genetic diversity ($\theta$) is the mean amount of variation in DNA sequences (Ellegren and Galtier, 2016) and is directly linked to effective population size ($N_e$) and mutation rate ($\mu$) as $\theta = 4N_e\mu$. Conceptually, we see that the equilibrium levels of genetic diversity is dependent on both the input of new mutations - proportional to $\mu$ - and the expected time to fixation which is inversely proportional to $N_e$. Thus, species with larger $N_e$ can maintain more transient polymorphism than species with lower $N_e$. Long-term genetic diversity is impacted by any parameters that change $N_e$ such as imbalanced sex ratio, fluctuations of $N_e$ with time, some metapopulation configurations, or the variance in reproductive success among individuals.

Genetic diversity in the ancestral population is the substrate on which divergence between lineages may build-up since it is directly accessible to selection. The average distance between individuals of a panmictic population ($d_{XY}$) is equal to ancestral genetic diversity ($\pi$). So, supposed that genetic incompatibilities already segregate as standing variation in the ancestral population prior to the population split, species with increased ancestral genetic diversity might evolve more reproductive isolation barriers in the early stage of speciation (Cutter, 2012). Moreover, regions of the genome that harbor higher diversity in the ancestral population - because of balancing selection caused by overdominant selection, negative-frequency dependent selection or differential selection across space and time - might contribute to regions of high divergence during speciation if the two different alleles are fixed within each population (Guerrero and Hahn, 2017).

## 5.2 Linked selection

*Linked selection* is the indirect effects of selection on neutral variants linked to positively or negatively selected alleles. In the former case, neutral variants may become fixed through genetic hitchhiking (Maynard Smith and Haigh, 1974) and in the latter case through *background selection* - the constant removal of deleterious mutants (Charlesworth et al., 1993).

Linked selection is supposed to be a major driver of genomic variation in the extent of differentiation of *Ficedula* species where the authors found correlated landscapes of differentiation between pairwise comparisons of *F. albicollis*, *F. hypoleuca*, *F. semitorquata* and *F. speculigera* (Burri et al., 2015). These correlations are thought to reflect the effect of accelerated lineage sorting in low-recombining regions. Indeed, recombination rates correlate negatively with $F_{ST}$ and positively with $\pi$ for all species comparisons.

## 5.3 Complete and incomplete lineage sorting: differential fixation of alleles

*Incomplete lineage sorting* (sometimes coined as *hemiplasy*) is the persistence of ancestral polymorphism after or during speciation events. For example, although humans are more closely related to chimpanzees than gorilla, 15% of genes reveals more proximity between humans and gorillas and 15% between gorillas and chimpanzees (Burgess and Yang, 2008). Incomplete lineage sorting creates discordance between the population or the species tree and a particular gene tree: in the previous examples, 30% of genes are discordant with the overall species ((human,chimpanzee);gorilla) tree that reflects evolutionary history. Incomplete lineage sorting is also widespread in all species of *Neoaves* - a clade that includes nearly all modern birds species: after its radiation 70Ma years ago, there is still around 30% of genes that show incomplete lineage sorting (Suh et al., 2015). At a lower evolutionary scale, this could blur the history of speciation along the genomes and make it difficult to decipher whether gene discordance is caused by incomplete lineage sorting or hybridization.

## 5.4 Introgression: the exchange of divergent genetic material

Introgression is defined as the transfer of genetic variation between related species through hybridization and repeated backcrossing (Martin and Jiggins, 2017). Hybridization does not always lead to introgression as migrating variation may be rapidly eliminated by selection from the gene pool of the recipient species, or negligible if hybridization is not frequent (Mallet et al., 2016). Hybridization and its consequences on introgression have been known for several decades - essentially for plants - but its pervasive presence in natural populations has been underestimated in the second part of the XX$^{\text{th}}$ century and was considered as "unnatural" and mainly caused by human disturbance (Anderson, 1948). This view has completely changed since and introgression is now recognized as a widespread phenomenon in the entire tree of life (Dagilis et al., 2021). As with incomplete lineage sorting, introgression also creates discordance between the species tree and gene trees: local introgression in particular regions of the genome reduces divergence times and can modify the local genealogy between species.

Once genetic variants have been introduced by introgression in the genome of the related species, allelic frequencies are controlled by drift and selection. *Adaptive introgression* can occur if introgressed variants are positively selected in the new species, such as genes encoding mimetism in *Heliconius* (Dasmahapatra et al., 2012), adaptation to altitude in humans (Huerta-Sánchez et al., 2014), some genes conferring adaptation to serpentine soils in *Arabidopsis arenosa* introgressed from *Arabidopsis lyrata* (Arnold et al., 2016). However, in most cases, introgression is expected to be neutral or deleterious. As divergence grows, introgression is likely to be reduced because of the increase of dysfunctional genetic incompatibilities between

the genetic backgrounds of the two species throughout the genome and the increase of their effect on fitness. Hence, local introgression reduces divergence and genetic differentiation at a local site: consequently, an old barrier to gene flow might, on the contrary, both have higher $F_{ST}$ and $d_{XY}$ following genome-wide erosion of divergence by introgression (Cruickshank and Hahn, 2014).

Introgression might be different depending on the spatial distribution of populations: we expect that introgression should be higher for populations in sympatry or, at least, more proximal than populations in allopatry or more distant, because there is much more opportunity to gene flow when individuals of the different lineages are less distant (Grant et al., 2005; Martin et al., 2013).

### 5.5 Divergence: the accumulation of molecular differences

Once barriers to reproduction are established and there is no longer gene flow at a specific locus, molecular divergence - the absolute pairwise nucleotide differences between two distinct populations - will increase. If divergence at a particular site is higher than in the rest of the genome (divergence is equal to genetic diversity in case of no population structure), this may correspond to the footprint of barrier to reproduction. As the net accumulation of divergence ($d_a$) at this locus, which is absolute divergence ($d_{XY}$) minus genetic diversity, is only proportional to mutation rate ($\mu$) and time ($t$), it can also give a clue about the time since these particular barriers impede gene flow.

## 6 Impact of demographic parameters on speciation and their link with life-history traits

So far, we have introduced that speciation is the accumulation of barriers to reproduction; that this accumulation can be explained by the demographic history of lineages separation and by the genetic architecture; and how these two factors affect all the evolutionary processes previously described. Here, we will now see how species-specific demographic parameters such as effective population size $N_e$, migration rate $m$ and generation time $t$ affect speciation and how they are related to specific species life-history traits.

### 6.1 How should effective population size ($N_e$) impact speciation?

Effective population size ($N_e$) has been a central concept in population genetics since its first formulation by Sewall Wright (Wright, 1931). $N_e$ is defined as the number of individuals $N$ in a

Wright-Fisher model (a hermaphroditic population that reproduces in panmixia with discrete generations) that would present the same amount of drift as observed empirically in a given population. It is a central concept that interacts with other neutral and selective evolutionary forces.

### 6.1.1 Time to fixation of neutral and advantageous mutations

Throughout the separation of two lineages, new alleles will constantly appear by mutation, and will change in frequency through time. These mutations can contribute to barriers to gene flow in two intensively studied manners: if i) they are neutral within lineages but are incompatible with the foreign genetic background (e.g., BDM incompatibilities), ii) they are under divergent selection and result in lower individual fitness due to local environmental conditions. In these two cases, the time needed for these mutations to become fixed in one of the two populations is dependent on $N_e$.

For the former case, the time to fixation of neutral mutations, if they are not lost by drift, is directly proportional to $N_e$ following (Crow and Kimura, 1970):

$$t_{fix} = 4N_e \tag{2}$$

In large populations, the intensity of drift is weaker, resulting in a longer time (in generations) needed for a new neutral mutation to become fixed in one population. This results also in higher levels of transient polymorphism in high $N_e$ species.

For the latter case, the probability that an advantageous mutation reaches frequency $p_T$ after $t$ generations follows (Cutter, 2019):

$$p_T = [1 + (2N_e - 1)e^{-st}]^{-1} \tag{3}$$

with $s$ the selection coefficient (Fig 3). Likewise, the expected number of generations needed for a new advantageous mutations to be fixed, conditioned on the fact that the allele does fix, and for very strong selection ($s >> 0$) (Nei and Li, 1973):

$$T_{fixation} = \frac{2ln(2N_e - 1)}{s} \tag{4}$$

We expect that a large population size $N_e$ will increase time to fixation both for neutral mutations that could become involved in BDM incompatibilities after secondary contact or locally advantageous mutations that could become counter-selected in a different environment.

### 6.1.2 Probability of fixation of mutations

As for the time to fixation, the expected probability that a mutation becomes fixed, $P_{fixation}$ depends on whether the mutation is neutral or under selection. For a new neutral mutation, the relation is quite simple because the probability of fixation depends on initial frequency $\frac{1}{4N_e}$. For a selected mutation, the probability of fixation at a given frequency $p$ is (Crow and Kimura, 1970):

$$P_{fixation} = \frac{1 - e^{-4N_e sp}}{1 - e^{-4N_e s}} \tag{5}$$

However, if $N_e$ is sufficiently high, $P_{fixation}$ become independent of $N_e$ as $P_{fixation} \approx 2s$ (Wright, 1931).

Hence, the impact of $N_e$ on the probability of fixation of new mutations is more complex: if the majority of mutations that contribute to speciation are neutral within each population (for example during allopatric speciation), more mutations are expected to be fixed in low $N_e$ species starting with the same amount of ancestral variation; if the majority of these mutations are under positive selection (local adaptation and diverging selection in sympatric speciation), high $N_e$ species should fix more divergent mutations; but if $N_e$ is sufficiently high, it becomes irrelevant to the probability of fixation of new advantageous mutations.

### 6.1.3 Deleterious mutations accumulation and introgression: the *hybridization load*

The majority of non-neutral mutations are weakly deleterious and are thus constantly maintained at low frequencies or removed from the gene pool by background selection. However, as the strength of purifying selection is weaker in species with low population size, the accumulation of deleterious mutations is higher in these populations (Charlesworth, 2009). The probability of fixation of deleterious mutations, $P_d$, follows (Kimura, 1957):

$$P_d \approx \frac{2s}{1 - exp^{-4Ns}} \tag{6}$$

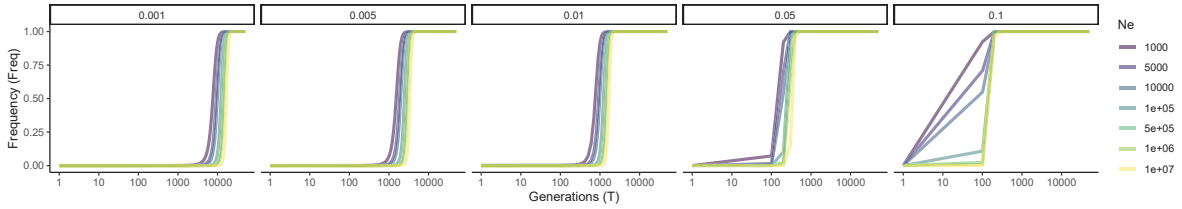



Figure 3: **The probability that an advantageous mutation reaches frequency $p$ at generation $T$.** Higher frequency is reached faster with strong selection coefficients (right panels) and in low $N_e$ populations (lighter to darker shows higher to lower $N_e$ )

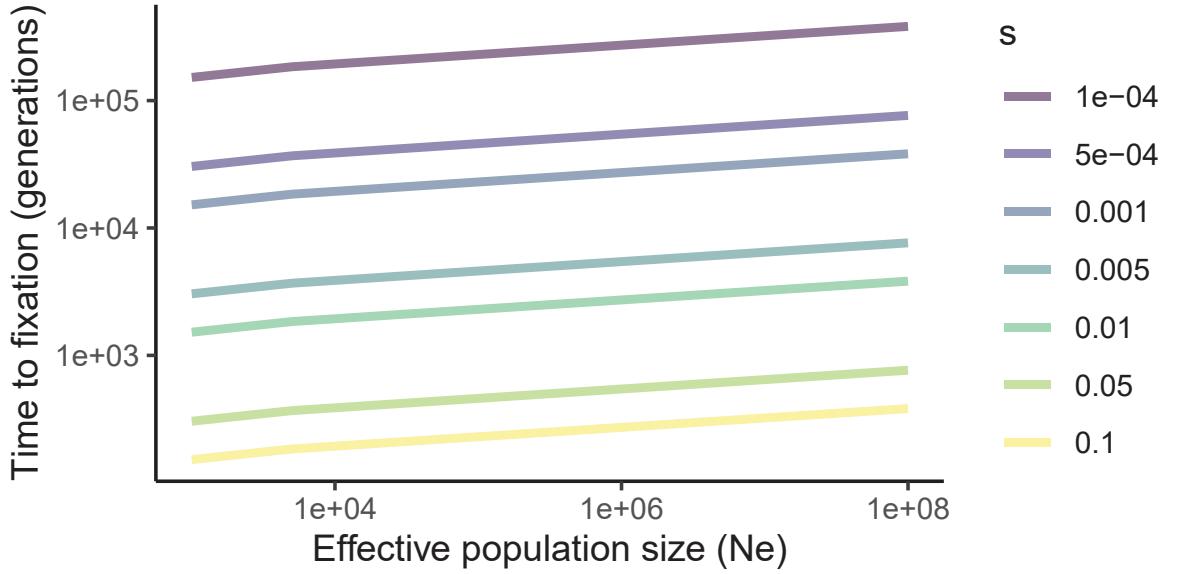



Figure 4: **Time to fixation of an advantageous mutation depending on the effective population size** $N_e$. Time to fixation increases with the log of $N_e$ and the rate of increase is proportional to the inverse of the selection coefficient (lighter to darker colors shows respectively stronger to weaker advantageous selection coefficients, $s$)

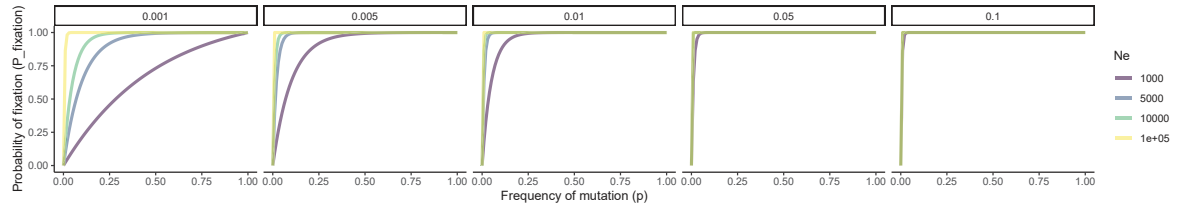



Figure 5: **Probability of fixation of a mutation given its frequency ($p$).** Probability to fixation increases with the frequency of mutation and with $N_e$ (lighter to darker colors shows respectively high to low $N_e$), and with higher selection coefficients ($s$) from left to right.

This relationship means that the probability of fixation $P_d$ mainly depends on the ratio of effective population $N_e$ and selection coefficient $s$: low $N_e$ increase the probability of fixation of deleterious mutations. As two divergent populations fix deleterious mutations in different sites of the genome, this has important consequences on introgression and differentiation. First, introgression may be advantageous if the majority of these mutations are also recessive as heterozygous hybrids mask the effect of deleterious and generally recessive mutations. In regions of high recombination rate, the effect of introgression may be reduced because recombination will break up and unmask the effect of deleterious mutations (Moran et al., 2021). Alternatively, hybrids may experience higher fitness reduction in low $N_e$ species that accumulated more deleterious partially recessive mutations than large $N_e$ species, due to *hybridization load*. These two effects can differentially lineages that have different population sizes, for example in cases of the two lineages that occupy different geographical habitats that do not have the same available niche and the ability to sustain the same number of individuals. This might be the case in the case of island-continent differentiation. The accumulation of deleterious variants is expected to be stronger in populations with lower effective population sizes because of a lower efficacy of selection to remove these variants. In this case, we can expect asymmetrical introgression between the two populations: introgression from the population of high to low $N_e$ may be advantageous for individuals in low $N_e$ populations because foreign haplotype blocks will mask the effect of deleterious load in hybrids. This effect was demonstrated in the brook charr (*Salvelinus fontinalis*) where blocks of foreign ancestry originating from a large $N_e$ domestic strain are found introgressed in regions of low recombination rate in wilds population with low $N_e$ (Leitwein et al., 2019). However, individuals from the high $N_e$ lineage will suffer a higher reduction of fitness because of the introgression of foreign blocks with a high density of deleterious mutations from the low $N_e$ lineages. Consequently, this will decrease introgression from low to high $N_e$ species, especially in around coding sequences (Harris and Nielsen, 2016).

Moreover, an interesting consequence of this model is that it can explain the fixation of underdominant chromosomal inversions that maintain barriers of reproductive isolation. As individuals that are heterozygous for the inversion are strongly selected against, the fixation of these inversions might be possible only in populations with very low $N_e$ to counterbalance the deleterious inversions Walsh (1982). A very low $N_e$ can be found in populations that have been impacted by strong bottlenecks, or in highly selfing plant species with high inbreeding.

### 6.1.4 Lineage sorting

The evolutionary forces that drive lineage sorting can be understood under the coalescent model. Any two haplotypes are related to each other to a unique ancestor at a specific time: they *coalesce* into a single lineage (Hahn, 2018). Given two populations of size $N$ that separated $t$ generations ago from an ancestral source, if two haplotypes from one population coalesce before $t$ at specific variants, lineages have been sorted: there is a concordance between the overall

population and gene trees; if they coalescence after $t$, there is incomplete lineage sorting. The coalescent theory states that the probability that two lineages coalesce before $t$ corresponds to (Hudson, 1983):

$$P[coalescence\ before\ t] = 1 - exp^{\frac{t}{2N_e}} \tag{7}$$

As the probability to coalesce before a given time increases with $N_e$, higher incomplete lineage sorting and lower genetic differentiation due to maintenance of ancestral polymorphism are expected in high $N_e$ species.

Lineage sorting, the differential fixation of ancestral neutral genetic variants between species or lineages is proportional to $N_e$. Under the coalescent model with no gene flow, two lineages diverging in complete isolation (i.e. without gene flow) are expected to share almost no ancestral polymorphism $10N_e$ generations after the split (between 8 and $12N_e$).

Species with higher $N_e$ where drift is weak will sort ancestral variants more slowly than species with low $N_e$ where drift is stronger. Levels of incomplete lineage sorting are mainly dependent on effective population size ($N_e$) and divergence time ($t$): species with high $N_e$ will have a slower rate of lineage sorting among the genome through time as drift is limited. Naturally, levels of lineage sorting are also dependent on $t$: higher divergence times allow more variants to be sorted between species. A convenient way to consider time during lineage sorting is to use demographic time since lineage sorting has a fixed rate in units of $2N_e$ generations.

An interesting consequence is that the probability that two haplotypes from two different populations coalesce before a time $t$ is equal to the mean genetic differentiation in a strict isolation model, where one population splits into two lineages at time $t$ with no gene flow (Wright, 1931). Thus, genetic differentiation is expected to be higher for low $N_e$ species under a strict isolation model at a given time of divergence.

### 6.1.5 Linked selection

The effect of background selection and genetic hitchhiking on genetic diversity under linked selection is complex. The models that formalize these effects differ in the sets of parameters taken into account. One of these models defines linked selection as a reduction of $\theta$ by a parameter $f$, where $\pi_{observed} = 4N_e\mu f$, and (Corbett-Detig et al., 2015):

$$f = \frac{1}{\frac{1}{exp^{-}G} + \alpha\frac{f_d}{r}} \tag{8}$$

where, $G$ is a complex parameter that depends both on the genome-wide deleterious mutation rate, selection against deleterious mutations, functional density, dominance, $r$ the recombination rate per base pair per generation and $\alpha$ which is also a complex parameter that is proportional to $N_e$. Assuming all other parameters are constant, species with high $N_e$ experience higher linked selection and higher reduction of genetic diversity compared to the neutral estimate because of a stronger efficiency of selection. This has two important consequences: i) the reduction of ancestral genetic diversity through linked selection is expected to be stronger in high $N_e$ species as shown (Corbett-Detig et al., 2015) (although high $N_e$ species have already higher neutral genetic diversity), ii) lineage sorting during lineages separation is stronger in high compared to low $N_e$ species.

### 6.1.6 Life history traits determinants of $N_e$

As a first approximation, $N_e$ should be directly correlated to abundance ($N$): species with a high number of individuals (e.g. *Drosophila* species) will have higher $N_e$ compared to species with a low number of individuals (e.g., gray wolf *Canis lupus*). A general law in ecology states that in Metazoans individual body size is negatively correlated with species abundance (White et al., 2007); hence, species with small individual body size such as invertebrates, will have high $N_e$ compared to species with individuals of large body sizes such as mammals or birds. This was observed for instance in European butterflies (Mackintosh et al., 2019), pinnipeds (Peart et al., 2020) and Darwin's finches (Brüniche-Olsen et al., 2019).

However, among 76 Metazoan, Romiguier et al. (2014) found that propagule size and fecundity best explain genetic diversity and long-term $N_e$. Species with lower parental investment (low propagule size and high fecundity) have higher genetic diversity, possibly because only species with large carrying capacity can buffer demographic fluctuations due to environmental stochasticity. In a similar study, Chen et al. (2017) found that genetic diversity and the ratio of non-synonymous to synonymous diversity ($\frac{\pi_n}{\pi_s}$) - the efficacy of selection against deleterious mutations, negatively correlated to $N_e$ - are higher in species with long lifespan among 60 species of animals and plants.

Finally, the actual value of $N_e$ is generally smaller than $N$ and the ratio $N_e/N$ can be correlated to other life-history traits. Variance in reproductive success ($V_k$), the variance in lifetime reproductive success between individuals, can strongly reduce $N_e$ compared to $N$ such as in marine organisms with high fecundity, long lifespan and low juvenile survival (Waples et al., 2018). Comparing 63 species, Waples (2016) found that 2/3 of the variance of $N_e/N$ can be explained by two life-history traits: age at maturity and adult lifespan (i.e., the difference between lifespan and age at maturity). Parental investment, adult lifespan and body size seem to be good predictors of $N_e$ in animal and plants even if the relationship between each of these traits might differ between taxa.

## 6.2 How $m$ should impact speciation ?

The migration rate $m$ is defined as the proportion of individuals received by a population every generation due to immigration from another population. It is directly linked, albeit different, from the concept of dispersal, to what evolutionary biologists define as any movements of individuals that induce gene flow across space (Ronce, 2007) and from gene flow, the number of individuals that migrate from one to another population per generation ($N_e \times m$).

### 6.2.1 Genetic homogenization

Under migration-drift equilibrium and a Wright-Fisher model - an infinite island model where all populations have the same effective population size $N_e$ and migration is equal between all populations and $m$ is large before $\mu$ and there is no selection - we expect that the amount of genetic differentiation between populations equals (Wright, 1931) :

$$F_{ST} = \frac{1}{1 + 4N_e m} \tag{9}$$

The level of genetic differentiation at equilibrium is expected to decrease in an inverse-function fashion with an increasing level of gene flow, $N_e m$: for low levels of genetic differentiation, it decreases faster with little changes in gene flow, but for low genetic differentiation, $F_{ST}$ becomes less sensitive to changes in gene flow. Thus, we can expect an inverse negative relationship between species genetic differentiation and dispersal capacities. However, for many reasons, natural populations do not respect the restrictive assumptions of the Wright-Fisher model (Whitlock and McCauley, 1999). Slatkin (1985) show that the time to reach equilibrium is approximated by $\frac{1}{m}$, thus after a demographic change, species with lower dispersal capacities should take more time to return to migration-drift equilibrium than species with higher dispersal capacities.

Similarly, during an episode of divergence in allopatry, species accumulate genetic differentiation genome-wide. However, once populations come into secondary contact, genetic differentiation at non-barrier loci is eroded at a rate proportional to the migration rate intensity (Barton and Bengtsson, 1986). Using a spatial model of secondary contact, Sedghifar et al. (2016) shows that the extent of individuals with admixed ancestry around the spatial location of secondary contact is proportional to the variance of spatial displacement ($\sigma^2$) as $\pm 2\sigma\sqrt{t_c}$.

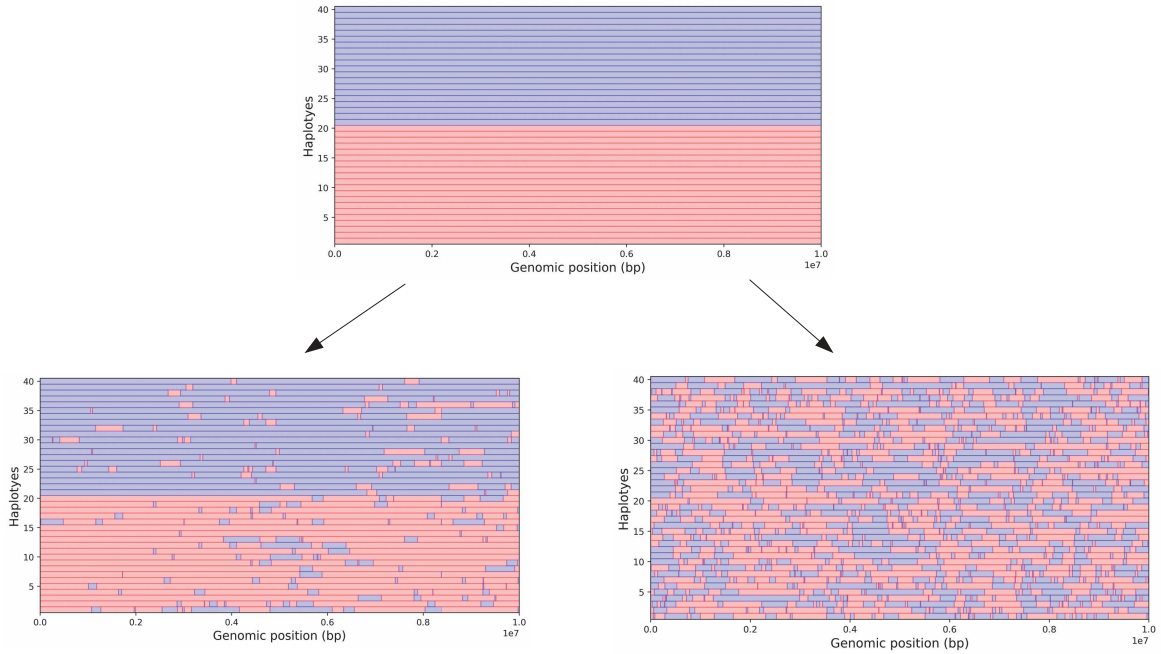



Figure 6: **Theoretical expectations of genomic homogenization following secondary contact between two genetically differentiated populations** - I simulated 20 diploid genomes sampled from two populations that split in complete isolation 100 000 generations ago (haplotypes 0 to 20: population A in red; haplotypes 21 to 40: population B in blue) and came back into secondary contact 10 000 generations ago. Each color represents the ancestry of origin before the time of secondary contact (e.g., a red portion of the genome in population A indicates introgressed ancestry from population B through gene flow). I simulated two scenarios with either low (left) or high (right) intensities of migration after secondary contact, with migration rate, $m$, equals 0.0001 and 0.001, meaning that 0.1‰and 1‰of individuals within each population are migrants, respectively on the left and on the right. All things being equal, species with lower dispersal capacities shows greater genetic differentiation and slower genetic homogenization. I simulated two populations with $N_e = 10000$ and a contiguous locus of 10Mbp under Wright-Fisher model with no selection. Simulations were run with `msprime` (Kelleher et al., 2016).

### 6.2.2 Width of a cline in a hybrid zone and migration-selection antagonism

Hybrid zones are defined as places where divergent taxa meet and exchange migrants. This generates *"interactions between genetically distinct groups of individuals resulting in at least some offspring of mixed ancestry"* (Harrison, 1990). Hybrid zones can follow from a secondary contact when two populations diverged and came back into contact or can evolve primarily such as during range expansions producing a cline of gene and/or phenotype frequency. Hybrid zones are defined as *tension zones* when genetic clines are maintained by a balance between dispersal ($\sigma$) and selection against hybrids ($s$). The resulting clines width is therefore a function of the ratio between dispersal and selection following (Barton and Hewitt, 1985):

$$w = \sqrt{\frac{\sigma^2}{s}} \tag{10}$$

. We might predict that for a given strength of selection against hybrids, species with lower dispersal capacities would show steeper clines.

Similarly, for a given strength selection against hybrids due to their genetic incompatibilities or local adaptations to different environmental conditions, barrier loci might be maintained in low $m$ species as lower migration and gene flow would have a lower counterbalancing effect relative to selection against hybrids. The accumulation of reproductive isolation barriers might thus be higher in species with low dispersal capacities.

### 6.2.3 Life history traits determinants of $m$

Many comparative studies and meta-analyses have compared mean or maximal individual dispersal distance to some life-history traits in various taxa. For example, among 210 plant species, Thomson et al. (2011) shows that plant height is positively correlated to individual dispersal distance, as an evolution to avoid reduced individual fitness when reproducing with related individuals. Moreover, plants that disperse seeds via biotic vectors (ant, seed-caching, ingestion and attachment) have greater dispersal than those which used abiotic vectors (wind, ballistic unassisted and water). Similarly, comparing dispersal distances in 795 species taken across the entire tree of life, Jenkins et al. (2007) found that active dispersers - defined as self-propulsion - show higher individual dispersal distance than passive dispersers; and among active dispersers, species with high propagule mass have higher dispersal distances while it has no effect within passive dispersers. However, the relation between life-history traits and dispersal can be more complex, as different life-history traits can affect different components of dispersal (mean dispersal distance, frequency of long-range dispersal, probability to disperse and gene flow) and that life-history traits that promote dispersal could have limited impact on gene flow as demonstrated by (Stevens et al., 2013), for instance: in butterflies, gene flow

was explained by four life-history traits, voltinism (annual number of generations), fecundity, number of eggs in female abdomen at emergence and female maturation, whereas wing size only affected the frequency of long-range dispersal.

## 6.3 How $T$ should impact speciation dynamics ?

In population genetics, the generation time, $T$, is the unit that scales how fast every evolutionary process act to change gene frequencies with time.

### 6.3.1 The speciation clock

If the accumulation of barrier loci throughout the genome is gradual with time, we might observe a positive correlation between reproductive isolation and any proxy of time (divergence, branch length or any genetic distance). A positive relationship between time and strength of postzygotic reproductive isolation - estimated as the ratio of hybrid to conspecific crosses fitness - has been found in many taxa, such as *Drosophila* flies (Coyne and Orr, 1989), frogs (Sasa et al., 1998), butterflies (Presgraves, 2002), fishes (Russell, 2003; Bolnick and Near, 2005), *Coreopsis* plants (Archibald et al., 2005), Mediterranean orchids (Scopece et al., 2007) and many other organisms (Coughlan and Matute, 2020). All these empirical observations led to formulate the presence of a *speciation clock*, similar to the *molecular clock* that states that substitutions rate accumulates gradually at a rate equal to the mutation rate (Kimura, 1983): a gradual accumulation of barriers that "clocks" at the same rate between independent diverging lineages. This means that we also expect that species with lower generation time should accumulate barriers to reproductive isolation more rapidly (in absolute time). However, across a broad taxonomic scale, divergence correlates poorly with the strength of reproductive isolation (Edmands, 2002).

### 6.3.2 Accumulation of genetic incompatibilities

During the separation of two lineages, genetic incompatibilities (e.g. BDMI) are expected to accumulate progressively through time. Orr (1995) demonstrated that as genetic incompatibilities accumulate in the genome, the mean number of incompatibilities, $I$, follows:

$$I = \frac{K^2 p}{2} \tag{11}$$

with $K$, the number of substitutions that have accumulated between the two lineages, $p$, the probability that new alleles are incompatible with any previously derived alleles. As the

number of substitutions, $K$, is directly correlated to time, the mean number of incompatibilities, $I$, increases with time faster than linear (this relationship is often coined as the snowball effect, (Orr and Turelli, 2001)). The first consequence of this snowball model is that species with shorter generation time might have accumulated more genetic incompatibilities on an absolute timescale. A second consequence is that we expect that reproductive isolation evolves quadratically with any life-history traits related to generation time and even faster if the incompatibilities are more complex and involve higher-order interactions compared to simple pairwise interactions between loci (Orr, 1995).

However, there is controversy about the validity of this model because of the low number of comparative empirical studies that found a square relationship between genetic incompatibilities and divergence (Gourbière and Mallet, 2010). Relaxing the infinite sites model in the Orr model and adding linked selection, Maya-Lastra and Eaton (2021) found that genetic incompatibilities do not accumulate with the square of time but reach a plateau. They also found a negative correlation between accumulation of genetic incompatibilities and $N_e$: under purifying selection, low $N_e$ populations are supposed to fix more deleterious mutations in different regions of the genome that may subsequently have higher chance to become genetic incompatibilities. On the contrary, in large $N_e$ populations, substitutions are likely to be found in the same locus resulting in lower total strength generated by potential incompatibilities.

### 6.3.3 Divergence

Once barriers to gene flow are established and nascent species can no longer exchange genetic material in a particular region of the genome where such incompatibilities map, incompatibility loci start to accumulate absolute genetic divergence ($d_{XY}$) defined as the mean number of molecular differences per site between pairwise haplotypes from distinct populations (Nei, 1975). In a panmictic population, $d_{XY} = \theta = 4N_e\mu$, but once gene flow stops, divergence equals the ancestral genetic diversity plus the number of substitutions that have independently accumulated in the two branches, defined as the net divergence ($d_a$) as:

$$d_{XY} = \theta + 2\mu t = 4N_e\mu + d_a \tag{12}$$

Thus, absolute ($d_{XY}$) and ($d_a$) between species are expected to increase monotonically with time at the barrier loci at a rate equal to twice the mutation rate, $\mu$. Hence, species with long generation time are expected to harbor lower net genetic divergence for a given mutation rate and absolute time of split.

### 6.3.4 Life history traits determinants of $T$

Usually, the units of time in evolutionary models such as those cited in this introduction are in generations as mutation and recombination rates are expressed per base and per generation. If the species generation time is equal to 1 year, there is no difference between genetic parameters (e.g. divergence) expressed in number of generations or absolute time (i.e. number of years). However, for many species, generation time is different from 1. For example, 2 or 3 generations occur in one reproductive season in some species of butterflies, while human generation time typically ranges from 22 to 33 years. Some species reproduce several times during their life, while others are semelparous. Generation time can be approximately defined as the mean age of parents: thus, iteroparity, delayed age at first maturity and longevity should be good predictors of generation time.

A better estimation of generation length can be retrieved from the analyses of life tables, which summarizes age-specific fecundity and survival. When these data are available, $T$ is estimated as (Ricklefs and Miller, 1999):

$$T = \frac{\sum_{x=0}^{L} x l_x m_x}{\sum_{x=0}^{L} l_x m_x} \tag{13}$$

where x ranges from 0 to maximum age at reproduction, $L$, $l_x$ and $m_x$ the age-specific survival and fecundity respectively at age $x$. Basically, the denominator is the reproductive rate - i.e. the average number of offspring produced by an individual across its total lifetime; the numerator scales this reproductive rate by $x$ in a way that if more individuals of high age $x$ reproduces, generation time increases. Thus, as previously shown for $N_e$, life tables can provide a better predictor of a species' generation time because it takes into account the variance of reproductive success with simple life-history traits such as age at maturity and lifespan.

## 6.4 Other life-history traits

Besides life-history traits that are directly impacting the three previously described demographic characteristics, two different kinds of life-history traits might have an impact on the speciation process: *mate choice* and *parental care*.

### 6.4.1 Mate choice and assortative mating

*Mate choice* is the non-random mating of one sex with individuals of the opposite sex because of variation in expressed traits (Edward, 2015). Mate choice can lead to *assortative mating*,

the preferential reproduction between individuals that are genetically or phenotypically more similar. Species with a higher presence of assortative mating behavior can enhance speciation in allopatric or sympatric biogeographic scenarios. For the former, independent evolution of mate preferences and/or sexual traits can reduce heterospecific mating when lineages come into contact. Other models shed light on the increase of premating barriers in response to postzygotic barriers, called *reinforcement* (Servedio, 2004). Individuals who mate with conspecific individuals will be positively selected because their offspring will have higher fitness compared with individuals who mate with heterospecific individuals. If the basis of preference to mate with conspecifics is heritable, prezygotic barriers are selectively advantageous. In sympatry, van Doorn et al. (2009) proposed a model where two lineages are adapted to different environmental conditions with diverging selection and one sex chooses mates of the other sex-based on ornaments that indicate genotype and phenotype quality. Hybrids will show reduced fitness because of their distance to local optima that decrease their chance of mating success because of the low quality of their phenotypic ornaments. Thus, species with behavorial mate choice that rely on the choice of one sex based on phenotypic ornaments might show greater barriers to reproduction due to the evolution of assortative mating or other sexual isolation processes that reduce gene flow.

Species differ widely in mating systems, from species where there is no mate choice to species with strong sexual selection based on complex male choice mechanisms (Andersson, 1994). We can expect that species with known mate discrimination behavior (such as size discrimination) could be more prone to these previously defined evolutionary processes.

### 6.4.2 Parental care and the evolution of postzygotic barriers

Coyne (1974) proposed that hybrid inviability could be advantageous for species with *"substantial parental investment in the production and care of progeny"*. In this verbal model, Coyne formulated several hypotheses: first, it exists some postzygotic barriers that create some levels, not necessarily complete of hybrid inviability fixed between the two populations as a result, e.g., of a long allopatric period (e.g. DBM incompatibilities). Second, parents bring substantial parental care to their offpsring. Third, these parents can deliberately shorten the time devoted to their lesser fit progenies. And fourth, reproduction takes place several times with different partners during the breeding season. Under these conditions and if there is no assortative mating and if individuals from the two populations can freely interbreed, females bearing a new gene conferring higher hybrid inviability will abort their investment in hybrid progeny and compensates by reproducing with conspecific males whereas the other females will have a mixed progeny of hybrids and non-hybrids. The two females will produce the same number of viable offspring but females with no new hybrid inviability genes will produce more hybrids with lower fitness: as a result, increasing hybrid inviability will be positively selected. This model was used by Wade and Johnson (1994) and adapted to within-families development of offspring to

explain the evolution of hybrid sterility in flour-beetle species of *Tribolium* as there is variance in hybrid inviability between several populations of *Tribolium* and that families with hybrid sterility genes early eliminate their hybrid offsprings from the progeny and as a consequence have higher conspecific offsprings.

Species with higher parental care investment which can compensate hybrid offspring inviability with conspecific offsprings through abortion and new fertilization or siblings competition may more easily accumulate postzygotic barriers.

# 7 Life-history traits and speciation: the current state-of-the-art

As presented previously, some life-history traits are expected to affect fundamental demographic parameters ($N_e$, $m$ and $t$) and selective mechanisms that have a potential effect on the evolution of reproductive isolation during speciation. Thus, it might be expected to observe empirical correlations between species' life-history traits and some aspects of speciation dynamics. To test these predictions, three criteria have to be fulfilled: i) comparisons must be done between several species with varied life-history traits, ii) these species must be somewhat in two (or more) populations, lineages, or ecotypes that currently lies somewhere in the speciation continuum (Stankowski and Ravinet, 2021a), and iii) control must be done as much as possible for any other factors that can potentially affect speciation dynamics. For the moment, comparative studies that have met these three criteria remains scarce. So far, two different kinds of studies have been made: the first one compared levels of reproductive isolation between species with different life-history traits controlling for the levels of genetic divergence; the second one inferred genetic characteristics such as genetic differentiation or cline widths in different species and compared these life-history traits.

For the former category, three main studies compared the rate of accumulation of postzygotic barriers through experimental procedures in plants and linked it to life-history traits. Among *Coreopsis* plant species, Archibald et al. (2005) found that pollen inviability evolves faster in annual than in perennial species, controlling for molecular divergence. Similarly, Owens and Rieseberg (2014) found a similar result, among 52 species of *Helianthus* and *Madiinae* species of the *Asteracea* family. Annual plant species exhibit more chromosomal rearrangements than perennial. The authors proposed that annual species may experience lower $N_e$ because of more frequent founder events and subsequent inbreeding and self-fertilization, which may lead to a higher probability of fixation of chromosomal rearrangements. Another explanation relates more frequent meiotic events, because of lower generation time in annual plants, to an increased probability of appearance of chromosomal rearrangements due to errors in the replication machinery. Scopece et al. (2007) found contrasting types of reproductive isolation barriers in

Mediterranean orchids: while sex-deceptive species are mainly isolated by pre-zygotic barriers (pollinators choice), food-deceptive species are isolated by post-zygotic barriers. Sex-deceptive orchids attract pollinators by mimicking potential pollinators congeners with flower visual and olfactive traits specific to each pollinator species: this leads to a great specificity between pollinator and orchid species. On the contrary, food-deceptive species attract pollinators by mimicking nectariferous species without producing nectar. The contrasted evolution of different reproductive barriers can be explained by alternative reproductive strategies.

For the second category, in a meta-analysis of 135 hybrid zones from various amphibian, bird, fish, insect, mammal and invertebrate organisms, McEntee et al. (2020) found that, overall, dispersal capacities correlate positively with cline width, a result found also within birds, insects, mammals and non-avian reptiles but not amphibians. Overall, dispersal explained 33.5% of variation in hybrid zone widths. Particularly, species with high dispersal capacities have a lower limit on cline width, meaning that highly dispersive species might not form narrow clines in these groups. In the Atlantic Ocean - Mediterranean Sea suture zone, looking at genetic summary statistics and life-history trait data for 20 marine species, Patarnello et al. (2007) found no correlation between life-history traits and genetic differentiation, time of population expansion and Tajima's D. Likewise, in a more recent review of another multispecies contact zone between the North Sea and the Baltic Sea, Johannesson et al. (2020b) found no significant association between genetic clines characteristics and species life-history traits, including those related to dispersal. The authors emphasized specific effects (e.g. egg buoyancy in the European flounder *Platichthys flesus*) but they note that general effects of life-history traits might be blurred by variations in the strength of divergent selection, the demographic history, and genomic architecture. Finally, a multispecies comparative analysis of the demographic divergence history inferred in 61 population/species pairs of animals with varying levels of divergence detected no significant effect of species life-history traits on parameters such as the probability of ongoing gene flow (Roux et al., 2016).

Thus, our current knowledge of the possible links between species' life-history traits and the evolutionary dynamics of speciation remains limited.

# 8 This thesis: the impact of life-history traits on speciation in the Atlantic Ocean - Mediterranean Sea suture zone across 20 teleostean marine fish species

The main objective of this thesis is to understand the impact of life-history traits on divergence and speciation. To answer these questions, we will study the evolutionary processes that shape the genomic landscapes of diversity, differentiation, divergence and introgression

of 20 marine teleostean fishes that are subdivided to different extents in two genetic lineages separated by the Atlantic Ocean-Mediterranean Sea suture zone.

## 8.1 Speciation in marine fishes

Actinopterygii has emerged 400 MY years ago in the Devonian and have experienced a radiation around 100 MY years ago essentially within the Percomorpha clade, which contains more than 17000 extant species (Hughes et al., 2018). They suffered intense mass extinctions during the Late Permian when 96% of existing marine species went extinct (Raup, 1979). Many different life-history strategies appeared during the radiation of this clade. However, despite this long evolutionary history, genome architecture is quite conserved across all species, with the number and lengths of chromosomes being roughly similar due to relative karyotype stasis (Vitturi et al., 1998, 1996, 1992; Almeida et al., 2017; Galvão et al., 2011; García-Souto et al., 2015; Martínez-Rodríguez et al., 1989). Moreover, the broad-scale variation in local recombination rate also tends to follow a similar trend across chromosomes and species. Chromosomal recombination landscapes follow a typical U-shape pattern, with generally reduced recombination rates near the central part of the chromosomes and higher rates near the chromosome extremities (Haenel et al., 2018; Tine et al., 2014; Roesti et al., 2013).

Speciation rates in the sea have been historically underestimated for several reasons: i) first, because of difficulty to access and document species diversity in the sea, ii) secondly, because it was long thought that long-range oceanic dispersal helped by the pelagic larval phase in most marine organisms (Di Franco et al., 2015) would result in a "panmictic oceanic soup" (Knowlton, 1993). Finally, iii) because closely related species may differ by different chemical recognition mechanisms, such as that involved in mate choice (Crapon de Caprona and Ryan, 1990; Wong et al., 2005; Fisher and Rosenthal, 2006) which are difficult to observe empirically. Genetics has been a powerful tool to study speciation and the impact of biogeographic barriers in marine organisms. It shed light on the existence of numerous cryptic species with no or little phenotypic differences that are in fact distinct or at least diverging species whose geographic distributions boundaries are sometimes concordant with geographic barriers (Knowlton, 1993; Palumbi, 1996). Today, the paradigm of unlimited dispersion in the ocean has been widely questioned, phylogeographical breaks congruent across many marine organisms have been found in different biogeographic frontiers: such as the Indo-Pacific front (Barber et al., 2000), the North Sea-Baltic Sea transition zone (Johannesson et al., 2020b), the Isthmus of Panama (Knowlton and Weigt, 1998), the Angola-Berguenla frontal zone (Henriques et al., 2014), the Antarctic Polar Front (Shaw et al., 2004), and the Atlantic Ocean - Mediterranean sea suture zone.

## 8.2 The Atlantic Ocean - Mediterranean suture zone

The Mediterranean Sea is a geographically enclosed basin located between Europe and Africa and connected to the Atlantic Ocean by the Strait of Gibraltar, which has a 286 m depth and a 12.9km width (Patarnello et al., 2007). This strait represents the only historical connection to oceanic waters (apart from the human-made Suez Canal which connects the Mediterranean Sea to the Indian Ocean by the Red Sea since 1869). The Strait of Gibraltar experienced variation in the sea level caused by glacial and interglacial cycling during the Pleistocene, which reduced connections between Ocean Atlantic and the Mediterranean Sea multiple times, without interrupting water flow (Patarnello et al., 2007). Sea level was higher during warm periods which occurred approximately around 600 000, 300 000, 200 000, 100 000 and 10 000 years ago (Lambeck et al., 2002). Between all these periods, glacial episodes reduced sea level by around 100m with the lowest sea level reaching 140 000 and 30 0000 years ago (-130 and -120m, respectively) (Patarnello et al., 2007). Finally, 20 000 years ago occurred the Last Glacial Maximum, and 11 500 years ago the beginning of the current interglacial, the Holocene, accompanied by an increase in sea level to the current level.

For various marine organisms, the Atlantic Ocean - Mediterranean Sea transition zone is corresponding to an important phylogeographical break (Borsa et al., 1997; Patarnello et al., 2007). It is true for several vertebrate teleostean fishes such as the European sea bass, *Dicentarchus labrax* (Lemaire et al., 2005; Tine et al., 2014; Duranton et al., 2018), crustaceans such as *Homarus gammarus* (Jenkins et al., 2019), *Palaemon elegans* (Reuschel et al., 2010), mollusks life the blue mussel *Mytilus galloprovincialis* (Gosset and Bierne, 2013), and even plant species with *Zostera noltii* (Coyer et al., 2004). Hence, the Atlantic Ocean - Mediterranean Sea phylogeographical break corresponds to a *suture zone* as defined by Remington (1968) that is a *"band of geographic overlap between major biotic assemblages, including some pairs of species or semispecies which hybridize in the zone"*.

Two explanations have been proposed to explain the congruency of intra-specific differentiation between the Atlantic Ocean and Mediterranean lineages: a past history of vicariance or environmentally-induced limited dispersal. The former states that sea-level fluctuations as a consequence of climate fluctuations during the Pleistocene created a sufficient barrier to gene flow in the vicinity of the Strait of Gibraltar that promoted genetic divergence and differentiation. The latter explanation states that habitat choice and/or fidelity and ocean currents may limit gene flow between the two basins (*isolation by resistance*, McRae (2006)). The constant influx of Atlantic Ocean waters in the Mediterranean Sea creates an anticyclonic gyre that would prevent passive larvae to cross the Almeria-Oran Front (Patarnello et al., 2007). It is, however, often difficult to disentangle these two hypotheses without detailed knowledge on dispersal and genetic differentiation.

Looking into details, however, reveals that all species living in the Northeastern Atlantic - Mediterranean ecosystem do not show the same patterns of genetic differentiation: some show reciprocical monophyly (e.g. Mediterranean rainbow wrasse, *C. julis* Fruciano et al. (2011)) while some show no significant genetic structure (e.g., the chub mackerel, *Scomber japonicus*, Zardoya et al. (2004)). As stated by Patarnello et al. (2007), we might want to answer *"why did vicariance,separating populations across a given boundary, not equally affect all species over the same geographical range?"*. The previous results were ambivalent. For example, Bargelloni et al. (2003) and Bargelloni et al. (2005) found discordant patterns among several species of the *Sparidae* family: while *Lithognathus mormyrus*, *Spondyliosoma cantharus* and *Diplodus puntazzo* showed genetic differentiation between Atlantic Ocean and Mediterranean Sea; *Pagellus bogaraveo* and *Diplodus sargus* did not. Charrier et al. (2006) found also discordant results between two *Lophius* species: while *L. budegassa* showed weak population genetic structure between Atlantic Ocean and Mediterranean Sea, *L. piscatorius* not. Other species like, *Scomber japonicus* show no genetic differentiation at all (Zardoya et al., 2004). On a multispecies scale, Patarnello et al. (2007) found no correlation between life-history traits and genetic differentiation, time of population expansion and Tajima's D obtained from mitochondrial sequences. The discrepancies of genetic differentiation patterns between marine organisms in the Atlantic Ocean - Mediterranean Sea suture zone thus remain to be explained.

## 8.3 Relationships between fish life-history traits and demographic parameters

As previously introduced for other organisms, specific links exist between life-history traits and demographic parameters for marine fishes.

As for many animal species, $N_e$ is negatively correlated to body size. However, marine fishes display intense variance in reproductive success among individuals because of low juvenile survival, high adult fecundity and long adult lifespan. These characteristics cause some individuals, sometimes coined Big Old Fat Fertile Female Fish (BOFFFF), to have disproportionate contributions to the gene pool compared to other individuals, which may create a drastic reduction of effective populations size compared to abundance (i.e. census population size).

The life cycle of many marine organisms included a dispersive planktonic larval phase after hatching and before settling to a suitable habitat. The length of the planktonic phase, so-called pelagic larval duration (PLD), varies between species from a few minutes to 4 months (Shanks, 2009), and even more in freshwater eels. As the larval transport is considered to be passive, larvae are often considered to drift with oceanographic currents: thus, the longer the PLD, the longest the dispersal distance between one individual's birth and settlement place. Thus, PLD

is considered to be a good predictor of species' dispersal capacities in marine organisms (Selkoe et al., 2014). However, empirical observations are mitigated in marine fishes: some studies found a negative relationship between differentiation and PLD (Doherty et al., 1995; Shulman and Bermingham, 1995; Riginos and Victor, 2001; Purcell et al., 2006; Selkoe et al., 2014); Ramon et al. (2008) found that one species of damselfish that reproduces in wave-sheltered lagoons where ocean flows are weaker displays stronger genetic structure than another species that reproduces in open water; Waples (1987) found that species with high fecundity, long PLD and offshore reproduction have stronger genetic structure; Galarza et al. (2009) found no effect of PLD and larvae location distance to shore; Riginos et al. (2011) found no effect of PLD but rather that species with benthic egg are more differentiated. These diverse results can be explained by several reasons: i) there is no or little correlation between PLD and dispersal capacities for pelagic spawners (Macpherson and Raventos, 2006), ii) Despite long PLD, some coral reef species show strong larval retention, where larvae tend to settle near the place where they were born (Jones et al., 1999; Swearer et al., 1999), iii) many marine organisms show reduced dispersal distance ($< 1$ km) despite a PLD longer than one week for some of them (Shanks, 2009), iv) realized dispersal distances under a Lagrangian model of dispersal modeling passive larvae are often greater than actual dispersal distances (Shanks, 2009). Shanks (2009) stated that PLD can be only used as an "indicator of dispersal potential" because species with low PLD ($< 10$ hours) clearly cannot have long dispersal distances. On the other hand, this could lead to a "problematic correlation analysis" as species with long PLD ($> 10$ hours) can display huge variability in observed dispersal distances because of the variability of larval dispersal behavior between species. In fact, no correlation between PLD and observed dispersal distances was found for species with dispersal distances longer than 20 km. Larval behaviors that reduce dispersal includes larval self-recruitment or retention in coral reef fishes (Jones et al., 1999; Swearer et al., 1999), larvae staying in the bottom of the sea where the current flow is weaker, larvae able to sustain swimming against the current for several kilometers (Stobutzki and Bellwood, 1997), and the use of sound as a cue (Tolimieri et al., 2000).

The majority of fish species are iteroparous and they show substantial variation in the age at first maturity and maximum longevity. For instance, longevity can range from a few years for some gobies to several dozens of years in *Sebastes* rockfishes.

Finally, some of our studied species are characterized by life-history traits suggesting intense parental care investment. For example, the long-snouted sea horse (*H. guttulatus*) provides intense parental care through egg protection in a seal-brood pouch in the male. They also engage in multiple spawnings with different mates during the breeding period (Curtis, 2007). A recent genetic study in the long-snouted sea horse showed the existence of four evolutionary lineages that have been separated by geographical barriers and that could have fixed preliminaries genetic incompatibilities causing hybrid breakdown (Riquet et al., 2019). Patterns of parental

care, multiple spawnings in one breeding season and evidence for allopatric differentiation have also been found in the pipefish (*S. typhle*) (Wilson et al., 2003; Rispoli and Wilson, 2008; Wilson and Eigenmann Veraguth, 2010), the grey wrasse (*S. cinereus*) and the black goby (*G. niger*) (Joyeux et al., 1991; Mazzoldi and Rasotto, 2002).

## 8.4 Methodological plan

We chose 20 teleostan marine fish species having nearly similar Northeastern Atlantic Ocean - Mediterranean Sea distributions. We based our selection of species on three criteria: i) previous studies showed either marked or significant genetic differentiation between the Atlantic Ocean and the Mediterranean Sea, or for one case marked phenotypic differentiation between the two basins, ii) they have an Atlantic-Mediterranean distribution and are present in the four locations we selected (see below), iii) they present variable life-history traits that may affect speciation through their impact on effective population size, migration rate, generation time and other demographic and/or evolutionary forces.

We selected pelagic species with either small (European pilchard *Sardina pilchardus*) or large body size (European hake, *Merluccius merluccius*, Black anglermonk *Lophius budegassa*); coastal species (Striped red mullet, *Mullus surmuletus*, Sea comber, *Serranus cabrilla*); lagoonal and estuarine species (Sharp-snouted sea horse, *Hippocampus guttulatus*, Broadnose pipefish, *Syngnathus typhle*). The 20 species span subclades that cover most of the teleostean phylogeny. We selected 3 species of *Clupeidae* (*Sardina pilchardus*, *Engraulis encrasicolus*, *Alosa fallax*) positioned near the root of the teleostan tree; 3 species of Syngnathiformes (*Mullus surmuletus*, *Syngnathus typhle* and *Hippocampus guttulatus*); one species of the Scombridae family (*Sarda sarda*); 9 species of *Percomopha* order, one of the most diversified order in Vertebrates. All these species have experienced a similar biogeographic history (although not necessarily the same detailed histories), which has been marked by the climatic and sea level fluctuations of the Pleiostocene.

In order to decipher whether shared alleles between Atlantic and Mediterranean populations are caused by introgression or incomplete lineage sorting, we sampled each species in two different locations in each basin: one in the supposed suture zone and one in a remote location. In the Mediterranean Sea, we sampled one location in the Gulf of Lion, South of France (away from the suture zone) and one location in the Costa Calida region, South-East of Spain (in the suture zone); in the Atlantic Ocean, we sampled one location in the Bay of Biscay, South-West of France (away from the suture zone) and one in Algarve, South of Portugal (in the Atlantic side of the suture zone).

We sampled 20 individuals for each species, 5 in each of these four localities. As the list of

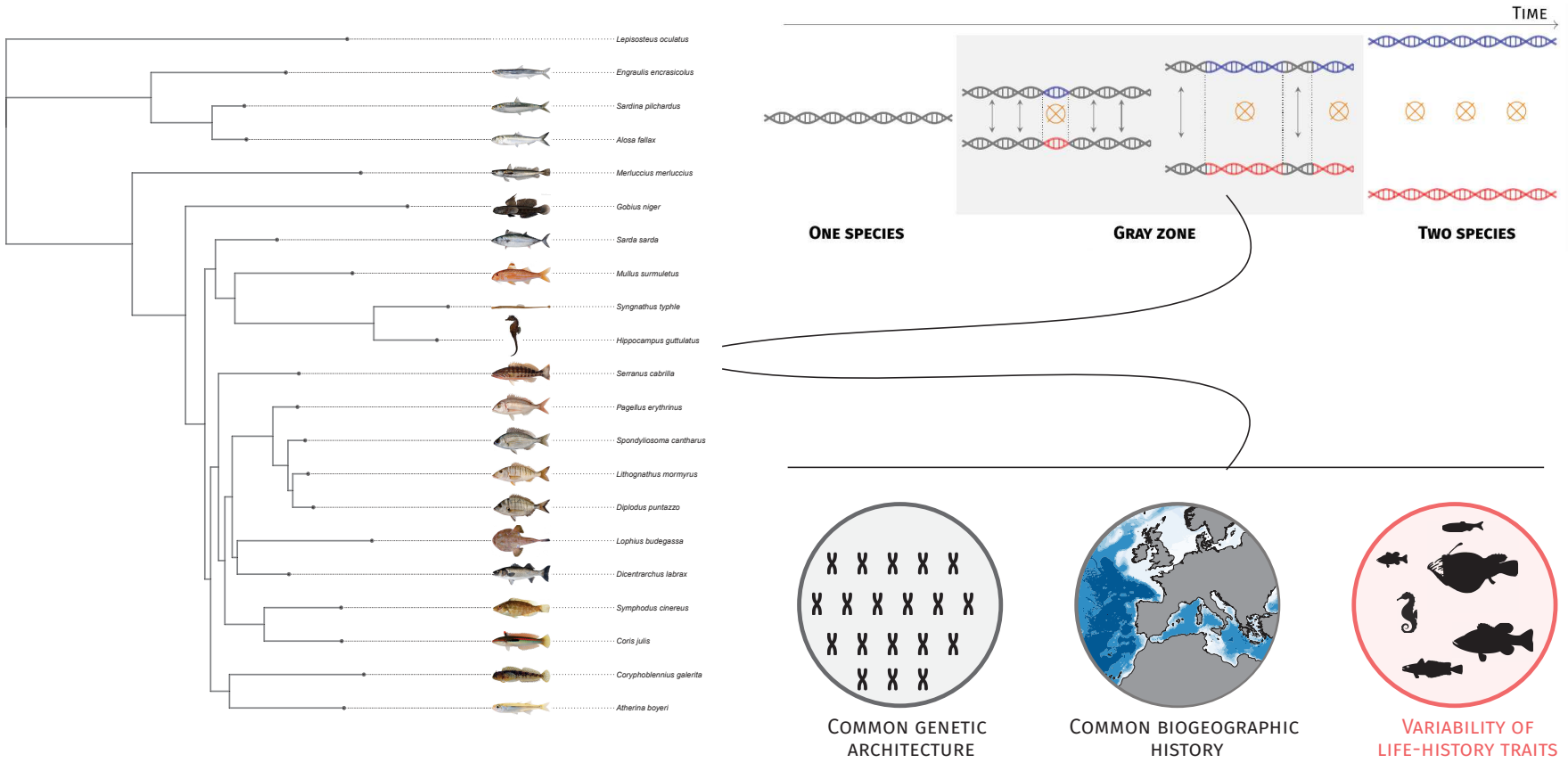



Figure 7: **Methodological plan of the thesis.** We selected 20 species spanning the phylogenetic diversity of teleostan marine fish that showed evidence of more or less pronounced genetic subdivisions between the Atlantic Ocean and the Mediterranean Sea populations. Most species population pairs across the Atlantic-Mediterranean suture zone are hypothesized to lie in the so-called gray zone of speciation, where the different populations have started to accumulate reproductive isolation barriers but can still hybridize. The genome architecture is conserved in teleostan fish landscapes with most species sharing similar numbers and roughly length distribution of chromosomes (although total genome size may vary among species) and recombination landscapes along chromosomes. Finally, these 20 species present a large diversity of life-history traits (body size, adult lifespan, fecundity, pelagic larval duration), ecological strategies (trophic level) and habitat type (lagoons, coastal shores, open waters) that might affect species demographic parameters, such as effective population size ($N_e$), migration rate ($m$) and generation time ($t$), and ultimately speciation dynamics.

selected species use various habitats, we used several methods to sample individuals: landing from industrial fisheries (*Sardina pilchardus*, *Lophius budegassa*, *Merluccius merluccius*, *Pagellus erythrinus*, *Sarda sarda*) or landings from traditional and local fisheries (*Dicentarchus labrax*, *Alosa fallax*, *Diplodus puntazzo*, *Lithognathus mormyrus*, *Serranus cabrilla*, *Spondyliosoma cantharus*), spearfishing (*Coris julis*, *Mullus surmuletus*) or hand nets (*Atherina boyeri*, *Coryphoblennius galerita*, *Hippocampus guttulatus*, *Gobius niger*, *Symphodus cinereus*, *Syngnathus typhle*). We sequenced the whole-genome of each individual to an average coverage depth of 20X and analyzed species patterns of nucleotide diversity, differentiation, divergence and introgression. Finally we compared each different facet of speciation to corresponding life-history traits and ecological strategies.

Moreover, contrary to previous comparative studies (Coyne and Orr, 1989; Burri et al., 2015; Martin et al., 2013), the 20 sister species we compared are phylogenetically distant and can therefore be used as independent speciation histories. As a rule of thumb, complete lineage sorting is expected to occur after $10N_e$ generations, and limited amounts of shared ancestral polymorphism is expected to be found after this threshold. Assuming an upper range of $N_e$ of 5 million individuals, complete lineage sorting is expected to occur after 50Ma. Since nearly all times to the most recent common ancestor between any two of our pairs of the 20 species are at least 100MYa (Hughes et al., 2018), we can conclude that each species' speciation dynamics is independent for other species in our list. Thus, we are confident that any correlations between life-history traits and genetic features we might observe should not result from correlated histories but rather perhaps reflect the effect of life-history traits.

Finally, all these species display a large variety of life-history traits that might affect effective population sizes such as body size, trophic level, parental care; generation time such as age at maturity and adult lifespan; migration rates such as pelagic larval duration and habitat types such as coastal lagoons, coastal shores, benthic or pelagic zones.

The objective of this thesis is to address the following questions:

- What are the determinants of genetic diversity in marine fishes and in particular what is the impact of life-history traits?
- What is the extent of Atlantic Ocean - Mediterranean Sea genetic differentiation and divergence across marine fishes species?
- What is the variability in the rate of introgression across the genome and of semi-permeability levels across species?
- What is the impact of life-history traits on speciation in these taxa?

- Are the patterns of mitochondrial differentiation related to nuclear patterns of diversity and divergence across all species?

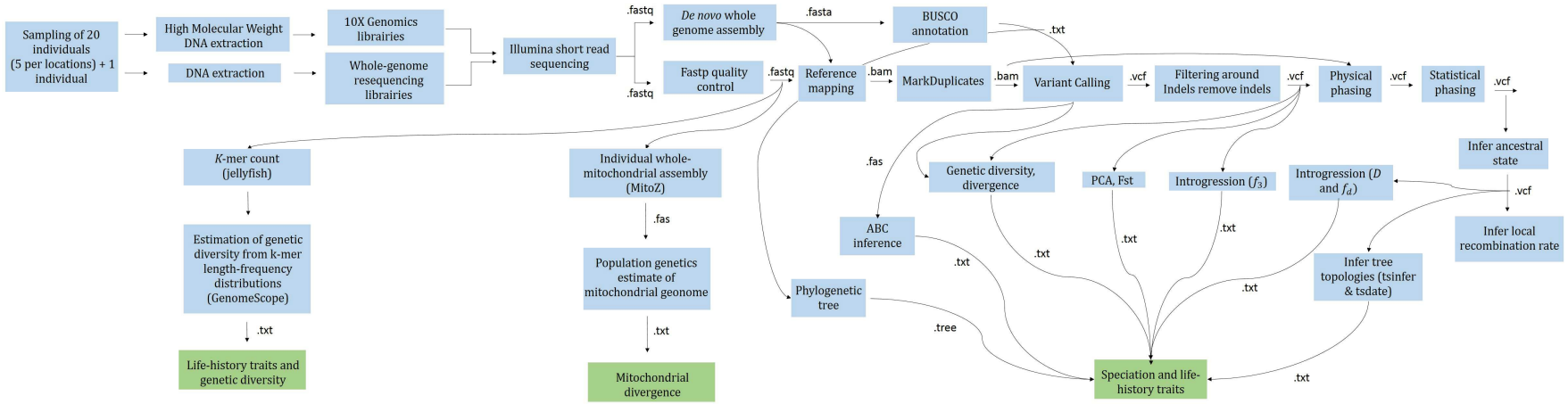



Figure 8: **The complete description of the methodological pipeline implemented in the thesis**. The blue rectangles show fieldwork, labwork, bioinformatics or data analysis. The green rectangles correspond to the three chapters presented in this thesis. The methodology procedure starts from the left and follows arrows in a way that a given output is strictly dependent on all previous blue boxes (e.g., $k$-mer analysis depends on fastp quality control, short-read sequencing, library construction and sampling). The format of the output of each procedure is shown on the top or on the bottom of the arrows (.txt = text file that shows simple data such as table; . fastq = biological sequences and their corresponding sequence quality; .fas = FASTA file, with the name of samples and scaffolds with corresponding haplotype nucleotide sequences; vcf= VCF file format that contains polymorphism data with corresponding variant quality and individual genotypes; bam = binary alignment sequences information; tree = phylogeny output with branch length.

# References

Almeida, L. A. H., Nunes, L. A., Bitencourt, J. A., Molina, W. F., and Affonso, P. R. A. M. (2017). Chromosomal Evolution and Cytotaxonomy in Wrasses (Perciformes; Labridae). *Journal of Heredity*, 108(3):239–253.

Anderson, E. (1948). Hybridization of the Habitat. *Evolution*, 2(1):1–9.

Andersson, M. (1994). *Sexual Selection*. Princeton University Press.

Andrew, R. L. and Rieseberg, L. H. (2013). Divergence Is Focused on Few Genomic Regions Early in Speciation: Incipient Speciation of Sunflower Ecotypes. *Evolution*, 67(9):2468–2482.

Archibald, J. K., Mort, M. E., Crawford, D. J., and Kelly, J. K. (2005). Life history affects the evolution of reproductive isolation among species of Coreopsis (Asteraceae). *Evolution; International Journal of Organic Evolution*, 59(11):2362–2369.

Aristoteles (1883). *Histoire des animaux*. Hachette et cie.

Arnold, B. J., Lahner, B., DaCosta, J. M., Weisman, C. M., Hollister, J. D., Salt, D. E., Bomblies, K., and Yant, L. (2016). Borrowed alleles and convergence in serpentine adaptation. *Proceedings of the National Academy of Sciences of the United States of America*, 113(29):8320–8325.

Avise, J. C., Shapira, J. F., Daniel, S. W., Aquadro, C. F., and Lansman, R. A. (1983). Mitochondrial DNA differentiation during the speciation process in Peromyscus. *Molecular Biology and Evolution*, 1(1):38–56.

Avise, J. C., Walker, D., and Johns, G. C. (1998). Speciation durations and Pleistocene effects on vertebrate phylogeography. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 265(1407):1707–1712.

Barber, P. H., Palumbi, S. R., Erdmann, M. V., and Moosa, M. K. (2000). Biogeography. A marine Wallace's line? *Nature*, 406(6797):692–693.

Bargelloni, L., Alarcon, J. A., Alvarez, M. C., Penzo, E., Magoulas, A., Palma, J., and Patarnello, T. (2005). The Atlantic–Mediterranean transition: Discordant genetic patterns in two seabream species, Diplodus puntazzo (Cetti) and Diplodus sargus (L.). *Molecular Phylogenetics and Evolution*, 36(3):523–535.

Bargelloni, L., Alarcon, J. A., Alvarez, M. C., Penzo, E., Magoulas, A., Reis, C., and Patarnello, T. (2003). Discord in the family Sparidae (Teleostei): Divergent phylogeographical patterns across the Atlantic–Mediterranean divide. *Journal of Evolutionary Biology*, 16(6):1149–1158.

Barton, N. and Bengtsson, B. O. (1986). The barrier to genetic exchange between hybridising populations. *Heredity*, 57(3):357–376.

Barton, N. H. and Hewitt, G. M. (1981). The genetic basis of hybrid inviability in the grasshopper Podisma pedestris. *Heredity*, 47(3):367–383.

Barton, N. H. and Hewitt, G. M. (1985). Analysis of Hybrid Zones. *Annual Review of Ecology and Systematics*, 16(1):113–148.

Bateson, W. (1909). "Heredity and Variation in Modern Lights". *Heredity and variation in modern lights*.

Bay, R. A. and Ruegg, K. (2017). Genomic islands of divergence or opportunities for introgression? *Proceedings. Biological Sciences*, 284(1850):20162414.

Bendall, E. E., Vertacnik, K. L., and Linnen, C. R. (2017). Oviposition traits generate extrinsic postzygotic isolation between two pine sawfly species. *BMC Evolutionary Biology*, 17(1):26.

Berdan, E. L., Mérot, C., Pavia, H., Johannesson, K., Wellenreuther, M., and Butlin, R. K. (2021). A large chromosomal inversion shapes gene expression in seaweed flies (Coelopa frigida). *Evolution Letters*, 5(6):607–624.

Bikard, D., Patel, D., Le Metté, C., Giorgi, V., Camilleri, C., Bennett, M. J., and Loudet, O. (2009). Divergent Evolution of Duplicate Genes Leads to Genetic Incompatibilities Within A. thaliana. *Science*, 323(5914):623–626.

Blackman, B. (2016). Speciation Genes. In *Encyclopedia of Evolutionary Biology*, pages 166–175. Elsevier.

Bolnick, D. I. and Near, T. J. (2005). Tempo of Hybrid Inviability in Centrarchid Fishes (teleostei: Centrarchidae). *Evolution*, 59(8):1754–1767.

Borsa, P., Naciri, M., Bahri-Sfar, L., Chikhi, L., Garcia De Leon, F., Kotoulas, G., and Bonhomme, F. (1997). Infraspecific zoogeography of the Mediterranean: Population genetic analysis on sixteen Atlanto-Mediterranean species (fishes and invertebrates). *Vie et Milieu*, 47:295–305.

Brüniche-Olsen, A., Kellner, K. F., and DeWoody, J. A. (2019). Island area, body size and demographic history shape genomic diversity in Darwin's finches and related tanagers. *Molecular Ecology*, 28(22):4914–4925.

Burgess, R. and Yang, Z. (2008). Estimation of hominoid ancestral population sizes under bayesian coalescent models incorporating mutation rate variation and sequencing errors. *Molecular Biology and Evolution*, 25(9):1979–1994.

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

Burri, R., Nater, A., Kawakami, T., Mugal, C. F., Olason, P. I., Smeds, L., Suh, A., Dutoit, L., Bureš, S., Garamszegi, L. Z., Hogner, S., Moreno, J., Qvarnström, A., Ružić, M., Sæther, S.-A., Sætre, G.-P., Török, J., and Ellegren, H. (2015). Linked selection and recombination rate variation drive the evolution of the genomic landscape of differentiation across the speciation continuum of Ficedula flycatchers. *Genome Research*, 25(11):1656–1665.

Calfee, E., Gates, D., Lorant, A., Perkins, M. T., Coop, G., and Ross-Ibarra, J. (2021). Selective sorting of ancestral introgression in maize and teosinte along an elevational cline. *PLOS Genetics*, 17(10):e1009810.

Carrió, E. and Güemes, J. (2014). The effectiveness of pre- and post-zygotic barriers in avoiding hybridization between two snapdragons (Antirrhinum L.: Plantaginaceae). *Botanical Journal of the Linnean Society*, 176(2):159–172.

Cayuela, H., Rougemont, Q., Laporte, M., Mérot, C., Normandeau, E., Dorant, Y., Tørresen, O. K., Hoff, S. N. K., Jentoft, S., Sirois, P., Castonguay, M., Jansen, T., Praebel, K., Clément, M., and Bernatchez, L. (2020). Shared ancestral polymorphisms and chromosomal rearrangements as potential drivers of local adaptation in a marine fish. *Molecular Ecology*, 29(13):2379–2398.

Charlesworth, B. (1998). Measures of divergence between populations and the effect of forces that reduce variability. *Molecular Biology and Evolution*, 15(5):538–543.

Charlesworth, B. (2009). Effective population size and patterns of molecular evolution and variation. *Nature Reviews Genetics*, 10(3):195–205.

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

Charrier, G., Chenel, T., Durand, J. D., Girard, M., Quiniou, L., and Laroche, J. (2006). Discrepancies in phylogeographical patterns of two European anglerfishes (Lophius budegassa and Lophius piscatorius). *Molecular Phylogenetics and Evolution*, 38(3):742–754.

Chen, J., Glémin, S., and Lascoux, M. (2017). Genetic Diversity and the Efficacy of Purifying Selection across Plant and Animal Species. *Molecular Biology and Evolution*, 34(6):1417–1428.

Christie, K. and Strauss, S. Y. (2018). Along the speciation continuum: Quantifying intrinsic and extrinsic isolating barriers across five million years of evolutionary divergence in California jewelflowers. *Evolution*, 72(5):1063–1079.

Coluzzi, M., Petrarca, V., and di Deco, M. A. (1985). Chromosomal inversion intergradation and incipient speciation in Anopheles gambiae. *Bollettino di zoologia*, 52(1-2):45–63.

Corbett-Detig, R. B., Hartl, D. L., and Sackton, T. B. (2015). Natural Selection Constrains Neutral Diversity across A Wide Range of Species. *PLOS Biology*, 13(4):e1002112.

Coughlan, J. M. and Matute, D. R. (2020). The importance of intrinsic postzygotic barriers throughout the speciation process. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 375(1806):20190533.

Coyer, J., Diekmann, O., Serrao, E., Procaccini, G., Milchakova, N., Pearson, G., Stam, W., and Olsen, J. (2004). Population genetics of dwarf eelgrass Zostera noltii throughout its geographic range. *Marine Ecology Progress Series*, 281:51–62.

Coyne, J. A. (1974). The evolutionary origin of hybrid inviability. *Evolution; International Journal of Organic Evolution*, 28(3):505–506.

Coyne, J. A. and Orr, H. A. (1989). Patterns of Speciation in Drosophila. *Evolution*, 43(2):362–381.

Coyne, J. A. and Orr, H. A. (2004). *Speciation*. Sinauer Associates, Sunderland, Mass.

Cracraft, J. (1983). Species Concepts and Speciation Analysis. In Johnston, R. F., editor, *Current Ornithology*, Current Ornithology, pages 159–187. Springer US, New York, NY.

Crapon de Caprona, M. D. and Ryan, M. J. (1990). Conspecific mate recognition in swordtails, Xiphophorus nigrensis and X. pygmaeus (Poeciliidae): Olfactory and visual cues. *Animal Behaviour*, 39(2):290–296.

Crow, J. F. and Kimura, M. (1970). *An Introduction to Population Genetics Theory*. Harper & Row.

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

Curtis, J. M. R. (2007). Validation of a method for estimating realized annual fecundity in a multiple spawner, the long-snouted seahorse (Hippocampus guttulatus), using underwater visual census. *Fishery Bulletin*, 105(3):327–337.

Cutter, A. D. (2012). The polymorphic prelude to Bateson–Dobzhansky–Muller incompatibilities. *Trends in Ecology & Evolution*, 27(4):209–218.

Cutter, A. D. (2019). *A Primer of Molecular Population Genetics*. Oxford University Press.

Dagilis, A. J., Peede, D., Coughlan, J. M., Jofre, G. I., D'Agostino, E. R. R., Mavengere, H., Tate, A. D., and Matute, D. R. (2021). 15 years of introgression studies: Quantifying gene flow across Eukaryotes.

Darwin, C. (1859). *The Origin of Species*. Collector's Library, London.

Dasmahapatra, K. K., Walters, J. R., Briscoe, A. D., Davey, J. W., Whibley, A., Nadeau, N. J., Zimin, A. V., Hughes, D. S. T., Ferguson, L. C., Martin, S. H., Salazar, C., Lewis, J. J., Adler, S., Ahn, S.-J., Baker, D. A., Baxter, S. W., Chamberlain, N. L., Chauhan, R.,

Counterman, B. A., Dalmay, T., Gilbert, L. E., Gordon, K., Heckel, D. G., Hines, H. M., Hoff, K. J., Holland, P. W. H., Jacquin-Joly, E., Jiggins, F. M., Jones, R. T., Kapan, D. D., Kersey, P., Lamas, G., Lawson, D., Mapleson, D., Maroja, L. S., Martin, A., Moxon, S., Palmer, W. J., Papa, R., Papanicolaou, A., Pauchet, Y., Ray, D. A., Rosser, N., Salzberg, S. L., Supple, M. A., Surridge, A., Tenger-Trolander, A., Vogel, H., Wilkinson, P. A., Wilson, D., Yorke, J. A., Yuan, F., Balmuth, A. L., Eland, C., Gharbi, K., Thomson, M., Gibbs, R. A., Han, Y., Jayaseelan, J. C., Kovar, C., Mathew, T., Muzny, D. M., Ongeri, F., Pu, L.-L., Qu, J., Thornton, R. L., Worley, K. C., Wu, Y.-Q., Linares, M., Blaxter, M. L., ffrench-Constant, R. H., Joron, M., Kronforst, M. R., Mullen, S. P., Reed, R. D., Scherer, S. E., Richards, S., Mallet, J., Owen McMillan, W., Jiggins, C. D., and The Heliconius Genome Consortium (2012). Butterfly genome reveals promiscuous exchange of mimicry adaptations among species. *Nature*, 487(7405):94–98.

Di Franco, A., Calò, A., Pennetta, A., De Benedetto, G., Planes, S., and Guidetti, P. (2015). Dispersal of larval and juvenile seabream: Implications for Mediterranean marine protected areas. *Biological Conservation*, 192:361–368.

Diamond, J. M. (1992). Horrible plant species. *Nature*, 360(6405):627–628.

Dobzhansky, T. (1936). Studies on Hybrid Sterility. II. Localization of Sterility Factors in Drosophila Pseudoobscura Hybrids. *Genetics*, 21(2):113–135.

Doherty, P. J., Planes, S., and Mather, P. (1995). Gene Flow and Larval Duration in Seven Species of Fish from the Great Barrier Reef. *Ecology*, 76(8):2373–2391.

Duranton, M., Allal, F., Fraïsse, C., Bierne, N., Bonhomme, F., and Gagnaire, P.-A. (2018). The origin and remolding of genomic islands of differentiation in the European sea bass. *Nature Communications*, 9(1):1–11.

Edmands, S. (2002). Does parental divergence predict reproductive compatibility? *Trends in Ecology & Evolution*, 17(11):520–527.

Edward, D. A. (2015). The description of mate choice. *Behavioral Ecology*, 26(2):301–310.

Ellegren, H. and Galtier, N. (2016). Determinants of genetic diversity. *Nature Reviews Genetics*, 17(7):422–433.

Ellegren, H., Smeds, L., Burri, R., Olason, P. I., Backström, N., Kawakami, T., Künstner, A., Mäkinen, H., Nadachowska-Brzyska, K., Qvarnström, A., Uebbing, S., and Wolf, J. B. W. (2012). The genomic landscape of species divergence in Ficedula flycatchers. *Nature*, 491(7426):756–760.

Faria, R., Chaube, P., Morales, H. E., Larsson, T., Lemmon, A. R., Lemmon, E. M., Rafajlović, M., Panova, M., Ravinet, M., Johannesson, K., Westram, A. M., and Butlin, R. K. (2019). Multiple chromosomal rearrangements in a hybrid zone between Littorina saxatilis ecotypes. *Molecular Ecology*, 28(6):1375–1393.

Faria, R. and Navarro, A. (2010). Chromosomal speciation revisited: Rearranging theory with pieces of evidence. *Trends in Ecology & Evolution*, 25(11):660–669.

Feder, J. L., Nosil, P., Wacholder, A. C., Egan, S. P., Berlocher, S. H., and Flaxman, S. M. (2014). Genome-Wide Congealing and Rapid Transitions across the Speciation Continuum during Speciation with Gene Flow. *Journal of Heredity*, 105(S1):810–820.

Feder, J. L., Opp, S. B., Wlazlo, B., Reynolds, K., Go, W., and Spisak, S. (1994). Host fidelity is an effective premating barrier between sympatric races of the apple maggot fly. *Proceedings of the National Academy of Sciences*, 91(17):7990–7994.

Felsenstein, J. (1981). Skepticism Towards Santa Rosalia, or Why Are There so Few Kinds of Animals? *Evolution*, 35(1):124–138.

Fisher, H. S. and Rosenthal, G. G. (2006). Female swordtail fish use chemical cues to select well-fed mates. *Animal Behaviour*, 72(3):721–725.

Fontaine, M. C., Pease, J. B., Steele, A., Waterhouse, R. M., Neafsey, D. E., Sharakhov, I. V., Jiang, X., Hall, A. B., Catteruccia, F., and Kakani, E. (2015). Extensive introgression in a malaria vector species complex revealed by phylogenomics. *Science*, 347(6217):1258524.

Fruciano, C., Tigano, C., and Ferrito, V. (2011). Geographical and morphological variation within and between colour phases in Coris julis (L. 1758), a protogynous marine fish. *Biological Journal of the Linnean Society*, 104(1):148–162.

Fuller, Z., Leonard, C., Young, R., Schaeffer, S., and Phadnis, N. (2017). The role of chromosomal inversions in speciation. Preprint, Evolutionary Biology.

Funk, D. J., Nosil, P., and Etges, W. J. (2006). Ecological divergence exhibits consistently positive associations with reproductive isolation across disparate taxa. *Proceedings of the National Academy of Sciences*, 103(9):3209–3213.

Gagnaire, P.-A. (2020). Comparative genomics approach to evolutionary process connectivity. *Evolutionary Applications*, 13(6):1320–1334.

Gagnaire, P.-A., Pavey, S. A., Normandeau, E., and Bernatchez, L. (2013). The Genetic Architecture of Reproductive Isolation During Speciation-with-Gene-Flow in Lake Whitefish Species Pairs Assessed by Rad Sequencing. *Evolution*, 67(9):2483–2497.

Galarza, J. A., Carreras-Carbonell, J., Macpherson, E., Pascual, M., Roques, S., Turner, G. F., and Rico, C. (2009). The influence of oceanographic fronts and early-life-history traits on connectivity among littoral fish species. *Proceedings of the National Academy of Sciences*, 106(5):1473–1478.

Galvão, T. B., Bertollo, L. A. C., and Molina, W. F. (2011). Chromosomal complements of some Atlantic Blennioidei and Gobioidei species (Perciformes). *Comparative Cytogenetics*, 5(4):259–275.

García-Souto, D., Troncoso, T., Pérez, M., and Pasantes, J. J. (2015). Molecular Cytogenetic Analysis of the European Hake Merluccius merluccius (Merlucciidae, Gadiformes): U1 and U2 snRNA Gene Clusters Map to the Same Location. *PLOS ONE*, 10(12):e0146150.

Garud, N. R., Messer, P. W., and Petrov, D. A. (2021). Detection of hard and soft selective sweeps from Drosophila melanogaster population genomic data. *PLOS Genetics*, 17(2):e1009373.

Good, J. M., Handel, M. A., and Nachman, M. W. (2008). Asymmetry and polymorphism of hybrid male sterility during the early stages of speciation in house mice. *Evolution; International Journal of Organic Evolution*, 62(1):50–65.

Gosset, C. C. and Bierne, N. (2013). Differential introgression from a sister species explains high F(ST) outlier loci within a mussel species. *Journal of Evolutionary Biology*, 26(1):14–26.

Gourbière, S. and Mallet, J. (2010). Are species real? The shape of the species boundary with exponential failure, reinforcement, and the "missing snowball". *Evolution; International Journal of Organic Evolution*, 64(1):1–24.

Grant, P. R., Grant, B. R., and Petren, K. (2005). Hybridization in the Recent Past. *The American Naturalist*, 166(1):56–67.

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

Gutenkunst, R. N., Hernandez, R. D., Williamson, S. H., and Bustamante, C. D. (2009). Inferring the Joint Demographic History of Multiple Populations from Multidimensional SNP Frequency Data. *PLOS Genetics*, 5(10):e1000695.

Haenel, Q., Laurentino, T. G., Roesti, M., and Berner, D. (2018). Meta-analysis of chromosome-scale crossover rate variation in eukaryotes and its significance to evolutionary genomics. *Molecular Ecology*, 27(11):2477–2497.

Hahn, M. W. (2018). *Molecular Population Genetics*. Oxford University Press ; Sinauer Associates, New York : Sunderland, MA.

Haldane, J. B. S. (1922). Sex ratio and unisexual sterility in hybrid animals. *Journal of Genetics*, 12(2):101–109.

Harris, K. and Nielsen, R. (2013). Inferring Demographic History from a Spectrum of Shared Haplotype Lengths. *PLOS Genetics*, 9(6):e1003521.

Harris, K. and Nielsen, R. (2016). The Genetic Cost of Neanderthal Introgression. *Genetics*, 203(2):881–891.

Harrison, R. G. (1990). Hybrid zones: Windows on evolutionary process. *Oxford Surveys in Evolutionary Biology*, 7:69–128.

Henriques, R., Potts, W. M., Santos, C. V., Sauer, W. H. H., and Shaw, P. W. (2014). Population Connectivity and Phylogeography of a Coastal Fish, Atractoscion aequidens (Sciaenidae), across the Benguela Current Region: Evidence of an Ancient Vicariant Event. *PLOS ONE*, 9(2):e87907.

Hewitt, G. (2000). The genetic legacy of the Quaternary ice ages. *Nature*, 405(6789):907–913.

Hey, J. (2001). The mind of the species problem. *Trends in Ecology & Evolution*, 16(7):326–329.

Hey, J. and Nielsen, R. (2007). Integration within the Felsenstein equation for improved Markov chain Monte Carlo methods in population genetics. *Proceedings of the National Academy of Sciences*, 104(8):2785–2790.

Hoballah, M. E., Gübitz, T., Stuurman, J., Broger, L., Barone, M., Mandel, T., Dell'Olivo, A., Arnold, M., and Kuhlemeier, C. (2007). Single Gene–Mediated Shift in Pollinator Attraction in Petunia. *The Plant Cell*, 19(3):779–790.

Hohenlohe, P. A., Bassham, S., Etter, P. D., Stiffler, N., Johnson, E. A., and Cresko, W. A. (2010). Population Genomics of Parallel Adaptation in Threespine Stickleback using Sequenced RAD Tags. *PLOS Genetics*, 6(2):e1000862.

Hudson, R. R. (1983). Testing the Constant-Rate Neutral Allele Model with Protein Sequence Data. *Evolution*, 37(1):203–217.

Huerta-Sánchez, E., Jin, X., Asan, Bianba, Z., Peter, B. M., Vinckenbosch, N., Liang, Y., Yi, X., He, M., Somel, M., Ni, P., Wang, B., Ou, X., Huasang, Luosang, J., Cuo, Z. X. P., Li, K., Gao, G., Yin, Y., Wang, W., Zhang, X., Xu, X., Yang, H., Li, Y., Wang, J., Wang, J., and Nielsen, R. (2014). Altitude adaptation in Tibetans caused by introgression of Denisovan-like DNA. *Nature*, 512(7513):194–197.

Hughes, L. C., Ortí, G., Huang, Y., Sun, Y., Baldwin, C. C., Thompson, A. W., Arcila, D., Betancur-R, R., Li, C., Becker, L., Bellora, N., Zhao, X., Li, X., Wang, M., Fang, C., Xie, B., Zhou, Z., Huang, H., Chen, S., Venkatesh, B., and Shi, Q. (2018). Comprehensive phylogeny of ray-finned fishes (Actinopterygii) based on transcriptomic and genomic data. *Proceedings of the National Academy of Sciences*, 115(24):6249–6254.

International Human Genome Sequencing Consortium (2004). Finishing the euchromatic sequence of the human genome. *Nature*, 431(7011):931–945.

Irwin, D. E. (2020). Assortative Mating in Hybrid Zones Is Remarkably Ineffective in Promoting Speciation. *The American Naturalist*, 195(6):E150–E167.

Jenkins, D. G., Brescacin, C. R., Duxbury, C. V., Elliott, J. A., Evans, J. A., Grablow, K. R., Hillegass, M., Lyon, B. N., Metzger, G. A., Olandese, M. L., Pepe, D., Silvers, G. A., Suresch, H. N., Thompson, T. N., Trexler, C. M., Williams, G. E., Williams, N. C., and Williams,

S. E. (2007). Does size matter for dispersal distance? *Global Ecology and Biogeography*, 16(4):415–425.

Jenkins, T. L., Ellis, C. D., Triantafyllidis, A., and Stevens, J. R. (2019). Single nucleotide polymorphisms reveal a genetic cline across the north-east Atlantic and enable powerful population assignment in the European lobster. *Evolutionary Applications*, 12(10):1881–1899.

Jiggins, C. D., Linares, M., Naisbit, R. E., Salazar, C., Yang, Z. H., and Mallet, J. (2001). Sex-Linked Hybrid Sterility in a Butterfly. *Evolution*, 55(8):1631–1638.

Johannesson, K., Butlin, R. K., Panova, M., and Westram, A. M. (2020a). Mechanisms of Adaptive Divergence and Speciation in Littorina saxatilis: Integrating Knowledge from Ecology and Genetics with New Data Emerging from Genomic Studies. In Oleksiak, M. F. and Rajora, O. P., editors, *Population Genomics: Marine Organisms*, Population Genomics, pages 277–301. Springer International Publishing, Cham.

Johannesson, K., Moan, A. L., Perini, S., and André, C. (2020b). A Darwinian Laboratory of Multiple Contact Zones. *Trends in Ecology & Evolution*, 35(11):1021–1036.

Jones, G. P., Milicich, M. J., Emslie, M. J., and Lunow, C. (1999). Self-recruitment in a coral reef fish population. *Nature*, 402(6763):802–804.

Joyeux, J.-C., Bouchereau, J.-L., and Tomasini, J.-A. (1991). LA REPRODUCTION DE GOBIUS NIGER (PISCES, GOBIIDAE) DANS LA LAGUNE DE MAUGUIO -FRANCE Rapports gonosomatiques, fécondités, ponte, oeufs et larves. *Vie et Milieu / Life & Environment*, pages 97–106.

Kelleher, J., Etheridge, A. M., and McVean, G. (2016). Efficient Coalescent Simulation and Genealogical Analysis for Large Sample Sizes. *PLOS Computational Biology*, 12(5):e1004842.

Kelleher, J., Wong, Y., Wohns, A. W., Fadil, C., Albers, P. K., and McVean, G. (2019). Inferring whole-genome histories in large population datasets. *Nature Genetics*, 51(9):1330–1338.

Kimura, M. (1957). Some Problems of Stochastic Processes in Genetics. *The Annals of Mathematical Statistics*, 28(4):882–901.

Kimura, M. (1983). *The Neutral Theory of Molecular Evolution*. Cambridge University Press, Cambridge.

Knowlton, N. (1993). Sibling Species in the Sea. *Annual Review of Ecology and Systematics*, 24(1):189–216.

Knowlton, N. and Weigt, L. A. (1998). New dates and new rates for divergence across the Isthmus of Panama. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 265(1412):2257–2263.

Kubo, T., Yoshimura, A., and Kurata, N. (2011). Hybrid Male Sterility in Rice Is Due to Epistatic Interactions with a Pollen Killer Locus. *Genetics*, 189(3):1083–1092.

Lambeck, K., Esat, T. M., and Potter, E.-K. (2002). Links between climate and sea levels for the past three million years. *Nature*, 419(6903):199–206.

Leitwein, M., Cayuela, H., Ferchaud, A.-L., Normandeau, É., Gagnaire, P.-A., and Bernatchez, L. (2019). The role of recombination on genome-wide patterns of local ancestry exemplified by supplemented brook charr populations. *Molecular Ecology*, 28(21):4755–4769.

Lemaire, C., Versini, J.-J., and Bonhomme, F. (2005). Maintenance of genetic differentiation across a transition zone in the sea: Discordance between nuclear and cytoplasmic markers. *Journal of Evolutionary Biology*, 18(1):70–80.

Liang, M. and Nielsen, R. (2014). The lengths of admixture tracts. *Genetics*, 197(3):953–967.

Limborg, M. T., Waples, R. K., Seeb, J. E., and Seeb, L. W. (2014). Temporally Isolated Lineages of Pink Salmon Reveal Unique Signatures of Selection on Distinct Pools of Standing Genetic Variation. *Journal of Heredity*, 105(6):835–845.

Lohse, K. (2017). Come on feel the noise – from metaphors to null models. *Journal of evolutionary biology*.

Ludlow, A. and Magurran, A. (2006). Gametic isolation in guppies (Poecilia reticulata). *Proceedings of the Royal Society B: Biological Sciences*, 273(1600):2477–2482.

Lynch, M. (2007). *The Origins of Genome Architecture*. Sinauer Associates, Sunderland, Mass.

Mackintosh, A., Laetsch, D. R., Hayward, A., Charlesworth, B., Waterfall, M., Vila, R., and Lohse, K. (2019). The determinants of genetic diversity in butterflies. *Nature Communications*, 10(1):1–9.

Macpherson, E. and Raventos, N. (2006). Relationship between pelagic larval duration and geographic distribution of Mediterranean littoral fishes. *Marine Ecology Progress Series*, 327:257–265.

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

Mallet, J. (2006). What does Drosophila genetics tell us about speciation? *Trends in Ecology & Evolution*, 21(7):386–393.

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

Marshall, D. C. and Cooley, J. R. (2000). Reproductive character displacement and speciation in periodical cicadas, with description of new species, 13-year Magicicada neotredecem. *Evolution; International Journal of Organic Evolution*, 54(4):1313–1325.

Martin, S. H., Dasmahapatra, K. K., Nadeau, N. J., Salazar, C., Walters, J. R., Simpson, F., Blaxter, M., Manica, A., Mallet, J., and Jiggins, C. D. (2013). Genome-wide evidence for speciation with gene flow in Heliconius butterflies. *Genome Research*, 23(11):1817–1828.

Martin, S. H., Davey, J. W., Salazar, C., and Jiggins, C. D. (2019). Recombination rate variation shapes barriers to introgression across butterfly genomes. *PLOS Biology*, 17(2):e2006288.

Martin, S. H. and Jiggins, C. D. (2017). Interpreting the genomic landscape of introgression. *Current Opinion in Genetics & Development*, 47:69–74.

Martínez-Rodríguez, G., Thode, G., Álvarez, M., and López, J. (1989). C-banding and Ag-NOR reveal heterogeneity among karyotypes of serranids (Perciformes). *Cytobios*, 58:53–60.

Maxwell, C. S., Sepulveda, V. E., Turissini, D. A., Goldman, W. E., and Matute, D. R. (2018). Recent admixture between species of the fungal pathogen Histoplasma. *Evolution Letters*, 2(3):210–220.

Maya-Lastra, C. A. and Eaton, D. A. R. (2021). Genetic incompatibilities do not snowball in a demographic model of speciation. Preprint, Evolutionary Biology.

Maynard Smith, J. M. and Haigh, J. (1974). The hitch-hiking effect of a favourable gene. *Genetics Research*, 23(1):23–35.

Mayr, E. (1963). *Animal Species and Evolution*.

Mazzoldi, C. and Rasotto, M. B. (2002). Alternative male mating tactics in Gobius niger. *Journal of Fish Biology*, 61(1):157–172.

McEntee, J. P., Burleigh, J. G., and Singhal, S. (2020). Dispersal Predicts Hybrid Zone Widths across Animal Diversity: Implications for Species Borders under Incomplete Reproductive Isolation. *The American Naturalist*, 196(1):9–28.

McRae, B. H. (2006). Isolation by Resistance. *Evolution*, 60(8):1551–1561.

Moran, B. M., Payne, C., Langdon, Q., Powell, D. L., Brandvain, Y., and Schumer, M. (2021). The genomic consequences of hybridization. *eLife*, 10:e69016.

Muller, H. J. (1942). Isolating mechanisms, evolution, and temperature. *Biology Symposium*, (6):71–125.

Natoli, A., Peddemors, V. M., and Rus Hoelzel, A. (2004). Population structure and speciation in the genus Tursiops based on microsatellite and mitochondrial DNA analyses. *Journal of Evolutionary Biology*, 17(2):363–375.

Nei, M. (1975). Molecular population genetics and evolution. *Frontiers of Biology*, 40:I–288.

Nei, M. and Li, W. H. (1973). Linkage disequilibrium in subdivided populations. *Genetics*, 75(1):213–219.

Noor, M. a. F. and Bennett, S. M. (2009). Islands of speciation or mirages in the desert? Examining the role of restricted recombination in maintaining species. *Heredity*, 103(6):439–444.

Nosil, P. (2012). *Ecological Speciation*. OUP Oxford.

Nosil, P., Funk, D. J., and Ortiz-Barrientos, D. (2009). Divergent selection and heterogeneous genomic divergence. *Molecular Ecology*, 18(3):375–402.

Orr, H. A. (1995). The population genetics of speciation: The evolution of hybrid incompatibilities. *Genetics*, 139(4):1805–1813.

Orr, H. A. and Turelli, M. (2001). The Evolution of Postzygotic Isolation: Accumulating Dobzhansky-Muller Incompatibilities. *Evolution*, 55(6):1085–1094.

Owens, G. L. and Rieseberg, L. H. (2014). Hybrid incompatibility is acquired faster in annual than in perennial species of sunflower and tarweed. *Evolution; International Journal of Organic Evolution*, 68(3):893–900.

Palumbi, S. R. (1996). What can molecular genetics contribute to marine biogeography? An urchin's tale. *Journal of Experimental Marine Biology and Ecology*, 203(1):75–92.

Papadopulos, A. S. T., Igea, J., Dunning, L. T., Osborne, O. G., Quan, X., Pellicer, J., Turnbull, C., Hutton, I., Baker, W. J., Butlin, R. K., and Savolainen, V. (2019). Ecological speciation in sympatric palms: 3. Genetic map reveals genomic islands underlying species divergence in Howea. *Evolution*, 73(9):1986–1995.

Pappers, S. M., van der Velde, G., and Ouborg, J. N. (2002). Host preference and larval performance suggest host race formation in Galerucella nymphaeae. *Oecologia*, 130(3):433–440.

Patarnello, T., Volckaert, F. a. M. J., and Castilho, R. (2007). Pillars of Hercules: Is the Atlantic–Mediterranean transition a phylogeographical break? *Molecular Ecology*, 16(21):4426–4444.

Peart, C. R., Tusso, S., Pophaly, S. D., Botero-Castro, F., Wu, C.-C., Aurioles-Gamboa, D., Baird, A. B., Bickham, J. W., Forcada, J., Galimberti, F., Gemmell, N. J., Hoffman, J. I., Kovacs, K. M., Kunnasranta, M., Lydersen, C., Nyman, T., de Oliveira, L. R., Orr, A. J., Sanvito, S., Valtonen, M., Shafer, A. B. A., and Wolf, J. B. W. (2020). Determinants of genetic variation across eco-evolutionary scales in pinnipeds. *Nature Ecology & Evolution*, pages 1–10.

Peñalba, J. V. and Wolf, J. B. W. (2020). From molecules to populations: Appreciating and estimating recombination rate variation. *Nature Reviews Genetics*, 21(8):476–492.

Powell, D. L., García-Olazábal, M., Keegan, M., Reilly, P., Du, K., Díaz-Loyo, A. P., Banerjee, S., Blakkan, D., Reich, D., Andolfatto, P., Rosenthal, G. G., Schartl, M., and Schumer, M. (2020). Natural hybridization reveals incompatible alleles that cause melanoma in swordtail fish. *Science*, 368(6492):731–736.

Presgraves, D. C. (2002). Patterns of postzygotic isolation in Lepidoptera. *Evolution*, 56(6):1168–1183.

Presgraves, D. C. (2010). The molecular evolutionary basis of species formation. *Nature Reviews Genetics*, 11(3):175–180.

Purcell, J. F., Cowen, R. K., Hughes, C. R., and Williams, D. A. (2006). Weak genetic structure indicates strong dispersal limits: A tale of two coral reef fish. *Proceedings of the Royal Society B: Biological Sciences*, 273(1593):1483–1490.

Racimo, F., Sankararaman, S., Nielsen, R., and Huerta-Sánchez, E. (2015). Evidence for archaic adaptive introgression in humans. *Nature Reviews Genetics*, 16(6):359–371.

Rajkov, J., Weber, A. A.-T., Salzburger, W., and Egger, B. (2018). Immigrant and extrinsic hybrid inviability contribute to reproductive isolation between lake and river cichlid ecotypes. *Evolution*, 72(11):2553–2564.

Ramon, M. L., Nelson, P. A., De Martini, E., Walsh, W. J., and Bernardi, G. (2008). Phylogeography, historical demography, and the role of post-settlement ecology in two Hawaiian damselfish species. *Marine Biology*, 153(6):1207–1217.

Rasmussen, M. D., Hubisz, M. J., Gronau, I., and Siepel, A. (2013). Genome-wide inference of ancestral recombination graphs. *arXiv:1306.5110 [q-bio]*.

Raup, D. M. (1979). Size of the permo-triassic bottleneck and its evolutionary implications. *Science (New York, N.Y.)*, 206(4415):217–218.

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

Ravinet, M., Yoshida, K., Shigenobu, S., Toyoda, A., Fujiyama, A., and Kitano, J. (2018). The genomic landscape at a late stage of stickleback speciation: High genomic divergence interspersed by small localized regions of introgression. *PLOS Genetics*, 14(5):e1007358.

Remington, C. L. (1968). Suture-Zones of Hybrid Interaction Between Recently Joined Biotas. In Dobzhansky, T., Hecht, M. K., and Steere, W. C., editors, *Evolutionary Biology: Volume 2*, pages 321–428. Springer US, Boston, MA.

Reuschel, S., Cuesta, J. A., and Schubart, C. D. (2010). Marine biogeographic boundaries and human introduction along the European coast revealed by phylogeography of the prawn Palaemon elegans. *Molecular Phylogenetics and Evolution*, 55(3):765–775.

Rice, W. R. and Salt, G. W. (1990). THE EVOLUTION OF REPRODUCTIVE ISOLATION AS A CORRELATED CHARACTER UNDER SYMPATRIC CONDITIONS: EXPERIMENTAL EVIDENCE. *Evolution; International Journal of Organic Evolution*, 44(5):1140–1152.

Ricklefs, R. E. and Miller, G. L. (1999). *Ecology*. W.H.Freeman & Co Ltd, New York, 4th edition edition.

Riechert and Hall (2001). Local population success in heterogeneous habitats: Reciprocal transplant experiments completed on a desert spider. *Journal of Evolutionary Biology*, 13:541–550.

Riesch, R., Muschick, M., Lindtke, D., Villoutreix, R., Comeault, A. A., Farkas, T. E., Lucek, K., Hellen, E., Soria-Carrasco, V., Dennis, S. R., de Carvalho, C. F., Safran, R. J., Sandoval, C. P., Feder, J., Gries, R., Crespi, B. J., Gries, G., Gompert, Z., and Nosil, P. (2017). Transitions between phases of genomic differentiation during stick-insect speciation. *Nature Ecology & Evolution*, 1(4):1–13.

Rieseberg, L. H. (2001). Chromosomal rearrangements and speciation. *Trends in Ecology & Evolution*, 16(7):351–358.

Rieseberg, L. H. and Blackman, B. K. (2010). Speciation genes in plants. *Annals of Botany*, 106(3):439–455.

Riginos, C., Douglas, K. E., Jin, Y., Shanahan, D. F., and Treml, E. A. (2011). Effects of geography and life history traits on genetic differentiation in benthic marine fishes. *Ecography*, 34(4):566–575.

Riginos, C. and Victor, B. C. (2001). Larval spatial distributions and other early life–history characteristics predict genetic differentiation in eastern Pacific blennioid fishes. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 268(1479):1931–1936.

Riquet, F., Liautard-Haag, C., Woodall, L., Bouza, C., Louisy, P., Hamer, B., Otero-Ferrer, F., Aublanc, P., Béduneau, V., Briard, O., El Ayari, T., Hochscheid, S., Belkhir, K., Arnaud-Haond, S., Gagnaire, P.-A., and Bierne, N. (2019). Parallel pattern of differentiation at a genomic island shared between clinal and mosaic hybrid zones in a complex of cryptic seahorse lineages. *Evolution*, 73(4):817–835.

Rispoli, V. F. and Wilson, A. B. (2008). Sexual size dimorphism predicts the frequency of multiple mating in the sex-role reversed pipefish Syngnathus typhle. *Journal of Evolutionary Biology*, 21(1):30–38.

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

Rogers, S. M. and Bernatchez, L. (2006). The genetic basis of intrinsic and extrinsic post-zygotic reproductive isolation jointly promoting speciation in the lake whitefish species complex (Coregonus clupeaformis). *Journal of Evolutionary Biology*, 19(6):1979–1994.

Romiguier, J., Gayral, P., Ballenghien, M., Bernard, A., Cahais, V., Chenuil, A., Chiari, Y., Dernat, R., Duret, L., Faivre, N., Loire, E., Lourenco, J. M., Nabholz, B., Roux, C., Tsagkogeorga, G., a. T. Weber, A., Weinert, L. A., Belkhir, K., Bierne, N., Glémin, S., and Galtier, N. (2014). Comparative population genomics in animals uncovers the determinants of genetic diversity. *Nature*, 515(7526):261–263.

Ronce, O. (2007). How Does It Feel to Be Like a Rolling Stone? Ten Questions About Dispersal Evolution. *Annual Review of Ecology, Evolution, and Systematics*, 38(1):231–253.

Roux, C., Fraïsse, C., Romiguier, J., Anciaux, Y., Galtier, N., and Bierne, N. (2016). Shedding Light on the Grey Zone of Speciation along a Continuum of Genomic Divergence. *PLOS Biology*, 14(12):e2000234.

Roux, C., Tsagkogeorga, G., Bierne, N., and Galtier, N. (2013). Crossing the Species Barrier: Genomic Hotspots of Introgression between Two Highly Divergent Ciona intestinalis Species. *Molecular Biology and Evolution*, 30(7):1574–1587.

Russell, S. T. (2003). Evolution of intrinsic post-zygotic reproductive isolation in fish. *Annales Zoologici Fennici*, 40(4):321–329.

Salzburger, W., Baric, S., and Sturmbauer, C. (2002). Speciation via introgressive hybridization in East African cichlids? *Molecular Ecology*, 11(3):619–625.

Sasa, M. M., Chippindale, P. T., and Johnson, N. A. (1998). Patterns of Postzygotic Isolation in Frogs. *Evolution*, 52(6):1811–1820.

Savolainen, V., Anstett, M.-C., Lexer, C., Hutton, I., Clarkson, J. J., Norup, M. V., Powell, M. P., Springate, D., Salamin, N., and Baker, W. J. (2006). Sympatric speciation in palms on an oceanic island. *Nature*, 441(7090):210–213.

Schilthuizen, M., Giesbers, M. C. W. G., and Beukeboom, L. W. (2011). Haldane's rule in the 21st century. *Heredity*, 107(2):95–102.

Scopece, G., Musacchio, A., Widmer, A., and Cozzolino, S. (2007). Patterns of Reproductive Isolation in Mediterranean Deceptive Orchids. *Evolution*, 61(11):2623–2642.

Sedghifar, A., Brandvain, Y., and Ralph, P. (2016). Beyond clines: Lineages and haplotype blocks in hybrid zones. *Molecular Ecology*, 25(11):2559–2576.

Selkoe, K. A., Gaggiotti, O. E., Laboratory, T., Bowen, B. W., and Toonen, R. J. (2014). Emergent patterns of population genetic structure for a coral reef community. *Molecular Ecology*, 23(12):3064–3079.

Servedio, M. R. (2004). The evolution of premating isolation: Local adaptation and natural and sexual selection against hybrids. *Evolution; International Journal of Organic Evolution*, 58(5):913–924.

Shanks, A. L. (2009). Pelagic Larval Duration and Dispersal Distance Revisited. *The Biological Bulletin*, 216(3):373–385.

Shaw, P. W., Arkhipkin, A. I., and Al-Khairulla, H. (2004). Genetic structuring of Patagonian toothfish populations in the Southwest Atlantic Ocean: The effect of the Antarctic Polar Front and deep-water troughs as barriers to genetic exchange. *Molecular Ecology*, 13(11):3293–3303.

Shulman, M. J. and Bermingham, E. (1995). Early Life Histories, Ocean Currents, and the Population Genetics of Caribbean Reef Fishes. *Evolution*, 49(5):897–910.

Simpson, G. G. (1961). *Principles of Animal Taxonomy*. Columbia University Press.

Slatkin, M. (1985). Gene Flow in Natural Populations. *Annual Review of Ecology and Systematics*, 16(1):393–430.

Sobel, J. M., Morris, A. E. W. F., and Kalisz, E. S. (2014). Ecogeographic Isolation and Speciation in the Genus Mimulus. *The American Naturalist*, 184(5):565–579.

Speidel, L., Forest, M., Shi, S., and Myers, S. R. (2019). A method for genome-wide genealogy estimation for thousands of samples. *Nature Genetics*, 51(9):1321–1329.

Stankowski, S. and Ravinet, M. (2021a). Defining the speciation continuum. *Evolution*, 75(6):1256–1273.

Stankowski, S. and Ravinet, M. (2021b). Quantifying the use of species concepts. *Current Biology*, 31(9):R428–R429.

Stevens, V. M., Trochet, A., Blanchet, S., Moulherat, S., Clobert, J., and Baguette, M. (2013). Dispersal syndromes and the use of life-histories to predict dispersal. *Evolutionary Applications*, 6(4):630–642.

Stobutzki, I. and Bellwood, D. (1997). Sustained swimming abilities of the late pelagic stages of coral reef fishes. *Marine Ecology Progress Series*, 149:35–41.

Sturtevant, A. H. (1921). A Case of Rearrangement of Genes in Drosophila. *Proceedings of the National Academy of Sciences*, 7(8):235–237.

Suh, A., Smeds, L., and Ellegren, H. (2015). The Dynamics of Incomplete Lineage Sorting across the Ancient Adaptive Radiation of Neoavian Birds. *PLOS Biology*, 13(8):e1002224.

Swearer, S. E., Caselle, J. E., Lea, D. W., and Warner, R. R. (1999). Larval retention and recruitment in an island population of a coral-reef fish. *Nature*, 402(6763):799–802.

Szymura, J. M. (1983). Genetic differentiation between hybridizing species Bombina bombina and Bombina variegata (Salientia, Discoglossidae) in Poland*. *Amphibia-Reptilia*, 4(2):137–145.

Tajima, F. (1989). Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*, 123(3):585–595.

Tanaka, K. M., Kamimura, Y., and Takahashi, A. (2018). Mechanical incompatibility caused by modifications of multiple male genital structures using genomic introgression in Drosophila. *Evolution*, 72(11):2406–2418.

Teeter, K. C., Payseur, B. A., Harris, L. W., Bakewell, M. A., Thibodeau, L. M., O'Brien, J. E., Krenz, J. G., Sans-Fuentes, M. A., Nachman, M. W., and Tucker, P. K. (2008). Genome-wide patterns of gene flow across a house mouse hybrid zone. *Genome Research*, 18(1):67–76.

Thompson, K. A., Peichel, C. L., Rennison, D. J., McGee, M. D., Albert, A. Y. K., Vines, T. H., Greenwood, A. K., Wark, A. R., Brandvain, Y., Schumer, M., and Schluter, D. (2022). Analysis of ancestry heterozygosity suggests that hybrid incompatibilities in threespine stickleback are environment dependent. *PLOS Biology*, 20(1):e3001469.

Thomson, F. J., Moles, A. T., Auld, T. D., and Kingsford, R. T. (2011). Seed dispersal distance is more strongly correlated with plant height than with seed mass. *Journal of Ecology*, 99(6):1299–1307.

Tine, M., Kuhl, H., Gagnaire, P.-A., Louro, B., Desmarais, E., Martins, R. S. T., Hecht, J., Knaust, F., Belkhir, K., Klages, S., Dieterich, R., Stueber, K., Piferrer, F., Guinand, B., Bierne, N., Volckaert, F. A. M., Bargelloni, L., Power, D. M., Bonhomme, F., Canario, A. V. M., and Reinhardt, R. (2014). European sea bass genome and its variation provide insights into adaptation to euryhalinity and speciation. *Nature Communications*, 5:5770.

Todesco, M., Owens, G. L., Bercovich, N., Légaré, J.-S., Soudi, S., Burge, D. O., Huang, K., Ostevik, K. L., Drummond, E. B. M., Imerovski, I., Lande, K., Pascual-Robles, M. A., Nanavati, M., Jahani, M., Cheung, W., Staton, S. E., Muños, S., Nielsen, R., Donovan, L. A., Burke, J. M., Yeaman, S., and Rieseberg, L. H. (2020). Massive haplotypes underlie ecotypic differentiation in sunflowers. *Nature*, 584(7822):602–607.

Tolimieri, N., Jeffs, A., and Montgomery, J. (2000). Ambient sound as a cue for navigation by the pelagic larvae of reef fishes. *Marine Ecology Progress Series*, 207:219–224.

Trickett, A. J. and Butlin, R. K. (1994). Recombination suppressors and the evolution of new species. *Heredity*, 73(4):339–345.

Turissini, D. A. and Matute, D. R. (2017). Fine scale mapping of genomic introgressions within the Drosophila yakuba clade. *PLOS Genetics*, 13(9):e1006971.

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

van Doorn, G. S., Edelaar, P., and Weissing, F. J. (2009). On the Origin of Species by Natural and Sexual Selection. *Science*, 326(5960):1704–1707.

Van Valen, L. (1976). Ecological Species, Multispecies, and Oaks. *TAXON*, 25(2-3):233–239.

Veller, C., Edelman, N. B., Muralidhar, P., and Nowak, M. A. (2021). Recombination and selection against introgressed DNA.

Via, S. and West, J. (2008). The genetic mosaic suggests a new role for hitchhiking in ecological speciation. *Molecular Ecology*, 17(19):4334–4345.

Vijay, N., Weissensteiner, M., Burri, R., Kawakami, T., Ellegren, H., and Wolf, J. B. W. (2017). Genomewide patterns of variation in genetic diversity are shared among populations, species and higher-order taxa. *Molecular Ecology*, 26(16):4284–4295.

Villanueva, R. (2015). Cryptic speciation in the stony octocoral Heliopora coerulea : Temporal reproductive isolation between two growth forms. *Marine Biodiversity*, 46.

Vitturi, R., Catalano, E., and Barbieri, R. (1992). Karyological and Molecular Characterization of *Mullus surmuletus* and *Mullus barbatus* (Pisces, Mullidae). *Cytologia*, 57(1):65–74.

Vitturi, R., Libertini, A., Campolmi, M., Calderazzo, F., and Mazzola, A. (1998). Conventional karyotype, nucleolar organizer regions and genome size in five Mediterranean species of Syngnathidae (Pisces, Syngnathiformes). *Journal of Fish Biology*, 52(4):677–687.

Vitturi, R., Libertini, A., Mazzola, A., Colomba, M. S., and Sara, G. (1996). Characterization of mitotic chromosomes of four species of the genus Diplodus: Karyotypes and chromosomal nucleolar organizer region phenotypes. *Journal of Fish Biology*, 49(6):1128–1137.

Wade, M. J. and Johnson, N. A. (1994). Reproductive isolation between two species of flour beetles, Tribolium castaneum and T. freemani: Variation within and among geographical populations of T. castaneum. *Heredity*, 72(2):155–162.

Walsh, J. B. (1982). Rate of Accumulation of Reproductive Isolation by Chromosome Rearrangements. *The American Naturalist*, 120(4):510–532.

Waples, R. S. (1987). A Multispecies Approach to the Analysis of Gene Flow in Marine Shore Fishes. *Evolution*, 41(2):385–400.

Waples, R. S. (2016). Life-history traits and effective population size in species with overlapping generations revisited. *Heredity*, 117(4):241–250.

Waples, R. S., Grewe, P. M., Bravington, M. W., Hillary, R., and Feutry, P. (2018). Robust estimates of a high Ne/N ratio in a top marine predator, southern bluefin tuna. *Science Advances*, 4(7):eaar7759.

Watson, J. D., Baker, T. A., Bell, S. P., Gann, A., Levine, M., and Losick, R. (2013). *Molecular Biology of the Gene*. Pearson, Boston, 7e édition edition.

Watson, J. D. and Crick, F. H. C. (1953). Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid. *Nature*, 171(4356):737–738.

White, E. P., Ernest, S. K. M., Kerkhoff, A. J., and Enquist, B. J. (2007). Relationships between body size and abundance in ecology. *Trends in Ecology & Evolution*, 22(6):323–330.

White, M. J. D. (1978). *Modes of Speciation*. A Series of Books in Biology. W. H. Freeman, San Francisco.

Whitlock, M. C. and McCauley, D. E. (1999). Indirect measures of gene flow and migration: FST not equal to 1/(4Nm + 1). *Heredity*, 82 ( Pt 2):117–125.

Wilson, A. B., Ahnesjo, I., Vincent, A. C. J., and Meyer, A. (2003). The Dynamics of Male Brooding, Mating Patterns, and Sex Roles in Pipefishes and Seahorses (family Syngnathidae). *Evolution*, 57(6):1374–1386.

Wilson, A. B. and Eigenmann Veraguth, I. (2010). The impact of Pleistocene glaciation across the range of a widespread European coastal species. *Molecular Ecology*, 19(20):4535–4553.

Wong, B. B., Fisher, H. S., and Rosenthal, G. G. (2005). Species recognition by male swordtails via chemical cues. *Behavioral Ecology*, 16(4):818–822.

Wright, S. (1931). Evolution in Mendelian Populations. *Genetics*, 16(2):97–159.

Wright, S. and Wright, S. (1984). *Variability within and among Natural Populations*. Number Sewall Wright ; Vol. 4 in Evolution and the Genetics of Populations. Univ. of Chicago Press, Chicago, Ill., paperback ed edition.

Wu, C.-I. (2001). The genic view of the process of speciation. *Journal of Evolutionary Biology*, 14(6):851–865.

Yukilevich, R. (2012). Asymmetrical Patterns of Speciation Uniquely Support Reinforcement in Drosophila. *Evolution*, 66(5):1430–1446.

Zardoya, R., Castilho, R., Grande, C., Favre-Krey, L., Caetano, S., Marcato, S., Krey, G., and Patarnello, T. (2004). Differential population structuring of two closely related fish species, the mackerel (Scomber scombrus) and the chub mackerel (Scomber japonicus), in the Mediterranean Sea. *Molecular Ecology*, 13(7):1785–1798.

# Chapitre 1

# La survie et la fécondité age spécifique déterminent la diversité génétique chez les poissons marins

**Résumé**

La diversité génétique ($\pi$) d'une espèce correspond au nombre moyen de différences entre deux séquences d'ADN tirées au hasard d'une population ou d'une espèce (Ellegren and Galtier, 2016). D'après la théorie neutre, elle est proportionnelle à la taille efficace de la population ($N_e$) et le taux de mutation ($\mu$) (Kimura, 1983). Depuis l'émergence de données moléculaires, de nombreuses études ont mis en évidence que les espèces possèdent des niveaux de diversité génétique différents (Leffler et al., 2012) : comment expliquer cette variabilité ? A travers une analyse de 76 espèces d'animaux, Romiguier et al. (2014) a montré que les espèces à investissement parental élevé (forte fécondité et faible taille de la propagule, stratégie $K$, par exemple les tortues) ont une plus faible diversité génétique que les espèces à faible investissement parental (par exemple les moules). Cependant, d'autres auteurs ont démontré des résultats différents au sein de taxa spécifiques dans la phylogénie des Métazoaires et notamment une corrélation négative avec la taille du corps chez les papillons européens (Mackintosh et al., 2019), chez les pinnipèdes (Peart et al., 2020), et chez les pinsons de Darwin (Brüniche-Olsen et al., 2019). L'abondance d'une espèce (corrélée négativement avec la taille du corps, White et al. (2007) semble être un meilleur prédicteur de la diversité génétique au sein de ces taxa. L'impact du gradient $r-K$ sur la diversité génétique à une échelle phylogénétique large pourrait cacher d'autres déterminants de la diversité génétique à des échelles plus fines. De plus, notre connaissance sur les causes biologiques qui permettent d'expliquer les relations entre traits d'histoire de vie et diversité génétique reste encore aujourd'hui parcellaire.

Dans ce premier chapitre, je me suis intéressé aux déterminants de la diversité génétique de 16 poissons marins téléostéens. Nous avons échantillonné et séquencé entre 12 et 20 individus par espèce dans 4 localités dans l'Océan Atlantique et la Mer Méditerranée (Fig. 1A). Nous avons ensuite estimé la diversité génétique de chaque individu par une analyse de la distribution de $k$-mers des lectures des séquences d'ADN (*reads*) non assemblés. Un $k$-mer est une sous-séquence d'ADN de longueur $k$. Si $k = 2$, il existe alors 16 possibles 2-mers à partir des quatre bases : AA, AT, AG, AC, TA, TG, TC, TT, CA, CT, CG, CC, GA, GG, GC, GT. A partir d'une analyse de la distribution des fréquences des 21-mers, nous avons estimé la taille du génome, la diversité génétique de chaque individu (voir les Méthodes et le Matériel Supplémentaire pour plus de détails sur la méthode `GenomeScope` Vurture et al. (2017), et la diversité génétique de chaque espèce comme la médiane des estimations individuelles. Les estimations de la diversité génétique avec `GenomeScope` sont identiques à celles inférées à partir de génomes alignés sur un génome de référence (Fig. 1B), validant la fiabilité de cette méthode.

En comparant avec 8 traits d'histoire de vie, nous avons trouvé une relation négative entre la longévité adulte (i.e. définie comme la différence entre la longévité et l'âge à maturité) et la diversité génétique (voir Figure 2) : par exemple, la baudroie (*L. budegassa*) avec une longévité adulte de 13,5 années montre une faible diversité ($\sim$0.2%) alors que la sardine (*S. pilchardus*),

avec une longévité adulte de 4 ans, une forte diversité (∼1.4%). Nous avons montré que les espèces avec des comportements de soins parentaux sur leur progéniture, comme l'hippocampe moucheté (*H. guttulatus*), par une incubation des oeufs dans la poche ventrale du mâle ou comme le crénilabre cendré (*S. cinereus*) par une garde du nid par les mâles, avaient également une diversité génétique réduite par rapport à leur longévité adulte.

En prenant en compte les tables de vie particulière des poissons marins, et notamment une forte mortalité juvénile et une augmentation de la fécondité avec l'âge (i.e. courbe de type III), nous avons montré que la réduction de la diversité génétique avec la longévité pouvait s'expliquer par une augmentation de la variance du succès reproducteur ($V_k$) chez les espèces longévives à cause de la présence de femelles très âgées et fécondes, contribuant de manière disproportionnée au pool génétique de la génération suivante (Fig. 3). Nous avons confirmé ces résultats en utilisant deux méthodes : l'une analytique (`AgeNe` , Waples et al. (2011) et l'autre simulatoire (`SLiM`, Haller and Messer (2017).

Enfin, dans une dernière partie, nous avons voulu étudier l'impact de différentes caractéristiques des tables de vie sur une corrélation éventuelle entre la longévité adulte et la diversité génétique. Nous avons simulé plusieurs courbes de mortalité par âge caractéristique de différents organismes : une forte mortalité juvénile avec une faible mortalité adulte, caractéristique des poissons marins (type III), une mortalité constante entre chaque âge, comme celles de certains oiseaux (type II) et une faible mortalité tout au cours de la vie et une forte mortalité proche de la longévité comme celles de certains mammifères et par exemple de l'espèce humaine (type I). En simulant un continuum de courbe allant du type III au type I pour diverses espèces avec des longévités différentes (Fig. 4A), nous avons trouvé qu'une corrélation entre longévité et $V_k$ n'était observée que pour les courbes de type III, caractéristiques des poissons marins (Fig. 4B).

Cette étude a permis d'affiner notre compréhension des déterminants de la diversité génétique. Notamment, nous avons proposé une explication causale pour comprendre la relation entre un trait d'histoire de vie (ici la longévité adulte) et la diversité génétique chez les poissons marins à travers la prise en compte des tables de vie. Cette explication peut également permettre de comprendre pourquoi la diversité génétique peut être corrélée à différents traits dans différents taxons.

# References

Brüniche-Olsen, A., Kellner, K. F., and DeWoody, J. A. (2019). Island area, body size and demographic history shape genomic diversity in Darwin's finches and related tanagers. *Molecular Ecology*, 28(22):4914–4925.

Ellegren, H. and Galtier, N. (2016). Determinants of genetic diversity. *Nature Reviews Genetics*, 17(7):422–433.

Haller, B. C. and Messer, P. W. (2017). SLiM 2: Flexible, Interactive Forward Genetic Simulations. *Molecular Biology and Evolution*, 34(1):230–240.

Kimura, M. (1983). *The Neutral Theory of Molecular Evolution*. Cambridge University Press, Cambridge.

Leffler, E. M., Bullaughey, K., Matute, D. R., Meyer, W. K., Ségurel, L., Venkat, A., Andolfatto,

doi:10.1002/evl3.265

# Age-specific survivorship and fecundity shape genetic diversity in marine fishes

**Pierre Barry,[1] Thomas Broquet,[2] and Pierre-Alexandre Gagnaire[1,3]**

[1]*ISEM, Univ Montpellier, CNRS, EPHE, IRD, Montpellier, France*

[2]*UMR 7144, Station Biologique de Roscoff, CNRS & Sorbonne Université, Roscoff, France*

[3]*E-mail: pierre-alexandre.gagnaire@umontpellier.fr*

Received December 18, 2020

Accepted November 9, 2021

**Genetic diversity varies among species due to a range of eco-evolutionary processes that are not fully understood. The neutral theory predicts that the amount of variation in the genome sequence between different individuals of the same species should increase with its effective population size ($N_e$). In real populations, multiple factors that modulate the variance in reproductive success among individuals cause $N_e$ to differ from the total number of individuals ($N$). Among these, age-specific mortality and fecundity rates are known to have a direct impact on the $N_e/N$ ratio. However, the extent to which vital rates account for differences in genetic diversity among species remains unknown. Here, we addressed this question by comparing genome-wide genetic diversity across 16 marine fish species with similar geographic distributions but contrasted lifespan and age-specific survivorship and fecundity curves. We sequenced the whole genome of 300 individuals to high coverage and assessed their genome-wide heterozygosity with a reference-free approach. Genetic diversity varied from 0.2% to 1.4% among species, and showed a negative correlation with adult lifespan, with a large negative effect ($slope = -0.089$ per additional year of lifespan) that was further increased when brooding species providing intense parental care were removed from the dataset ($slope = -0.129$ per additional year of lifespan). Using published vital rates for each species, we showed that the $N_e/N$ ratio resulting simply from life tables parameters can predict the observed differences in genetic diversity among species. Using simulations, we further found that the extent of reduction in $N_e/N$ with increasing adult lifespan is particularly strong under Type III survivorship curves (high juvenile and low adult mortality) and increasing fecundity with age, a typical characteristic of marine fishes. Our study highlights the importance of vital rates as key determinants of species genetic diversity levels in nature.**

**KEY WORDS:** **Adult lifespan, genetic diversity, life tables, marine fishes, variance in reproductive success.**

## *Impact Summary*

Understanding how and why genetic diversity varies across species has important implications for evolutionary and conservation biology. Although genomics has vastly improved our ability to document intraspecific DNA sequence variation at the genome level, the range and determinants of genetic diversity remain partially understood. At a broad taxonomic scale in eukaryotes, the main determinants of diversity are reproductive strategies distributed along a trade-off between the quantity and the size of offspring, which likely affect the long-term effective population size. Long-lived species also tend to show lower genetic diversity, a result that has however not been reported by comparative studies of genetic diversity at lower taxonomic scales. Here, we compared genetic diversity across 16 European marine fish species showing marked differences in longevity. Adult lifespan was the best predictor of genetic diversity, with genome-wide average heterozygosity ranging from 0.2% in the black anglerfish (*Lophius budegassa*) to 1.4% in the European pilchard (*Sardina pilchardus*). Using life tables summarizing age-specific mortality and fecundity rates for each species, we showed that the variance in lifetime reproductive success resulting from age structure, iteroparity, and overlapping generations can predict the range of observed differences in genetic diversity among marine fish species. We then used computer simulations to explore how



© 2021 The Authors. *Evolution Letters* published by Wiley Periodicals LLC on behalf of Society for the Study of Evolution (SSE) and European Society for Evolutionary Biology (ESEB).
This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

combinations of vital rates characterizing different life histories affect the relationship between adult lifespan and genetic diversity. We found that marine fishes that display high juvenile but low adult mortality, and increasing fecundity with age, are typically expected to show reduced genetic diversity with increased adult lifespan. However, the impact of adult lifespan vanished using bird and mammal-like vital rates. Our study shows that variance in lifetime reproductive success can have a major impact on species genetic diversity and explains why this effect varies widely across taxonomic groups.

Genetic diversity, the substrate for evolutionary change, is a key parameter for species adaptability and vulnerability in conservation and management strategies (Frankham 1995; Lande 1995; DeWoody et al. 2021). Understanding the determinants of species' genetic diversity has been, however, a long-standing puzzle in evolutionary biology (Lewontin 1974). Advances in DNA sequencing technologies have allowed to describe the range of genetic diversity levels across eukaryote species and identify the main evolutionary processes governing that variation (Leffler et al. 2012; Romiguier et al. 2014). Yet, the extent and reasons for which life history traits, and in particular reproductive strategies, influence genetic diversity remain to be clarified (Ellegren and Galtier 2016).

The neutral theory provides a quantitative prediction for the amount of genetic variation at neutral sites (Kimura 1983). Assuming equilibrium between the introduction of new variants by mutation occurring at rate $\mu$, and their removal by genetic drift at a rate inversely proportional to the effective population size $N_e$, the amount of genetic diversity ($\theta$) of a stable randomly mating population is equal to $4N_e\mu$ (Kimura and Crow 1964). This quantity should basically determine the mean genome-wide heterozygosity expected at neutral sites for any given individual in that population. However, because the neutral mutation-drift balance can be slow to achieve, contemporary genetic diversity often keeps the signature of past demographic fluctuations rather than being entirely determined by the current population size. Therefore, genetic diversity should be well predicted by estimates of $N_e$ that integrate the long-term effect of drift over the coalescent time. Unfortunately, such estimates are very difficult to produce using demographic data only.

Demographic variations set aside the most proximate determinant of $N_e$ is the actual number of individuals ($N$), also called the census population size. Comparative genomic studies in mammals and birds have shown that current species abundance correlates with the long-term coalescence $N_e$, despite a potential deviation from long-term population stability in several of the species studied (Díez-Del-Molino et al. 2018; Leroy et al. 2020; Peart et al. 2020). General laws in ecology, such as the negative relationship between species abundance and body size (White et al. 2007) have also been used to predict the long-term $N_e$. Higher genetic diversity in small body size species was found in butterflies and Darwin's finches (Brüniche-Olsen et al. 2019; Mackintosh et al. 2019), while in the latter genetic diversity also positively correlated with island size, another potential proxy for the long-term $N_e$ (Brüniche-Olsen et al. 2019). Surprisingly, however, genetic diversity variation across metazoans is much better explained by fecundity and propagule size than classical predictors of species abundance such as body size and geographic range (Romiguier et al. 2014). This result has been attributed to differences in extinction risk for species that have contrasted reproductive strategies. Under this hypothesis, species with low fecundity and large propagule size ($K$-strategists) would be more resilient to low population size episodes compared to species with high fecundity and small propagule size ($r$-stategists), which would go extinct if they reach such population sizes (Romiguier et al. 2014). By contrast, Mackintosh et al. (2019) found no effect of propagule size on genetic diversity within Papilionoidea, a superfamily showing little variation in reproductive strategy. Therefore, the major effect of the $r/K$ gradient on genetic diversity variation across metazoa probably hides other determinants that act within smaller branches of the tree of life. In particular, how demography and evolutionary processes influence genetic variation in different taxa remains unclear.

Other factors than fluctuations in population size are known to reduce the value of $N_e$ relative to the census population size, impacting the $N_e/N$ ratio to a different extent from one species to another. These factors include unbalanced sex ratios, variance in lifetime reproductive success among individuals, age structure, kinship-correlated survival, and some metapopulation configuration (Wright 1969; Lande and Barrowclough 1987; Falconer 1989). A potentially strong effect comes from variance in the number of offspring per parent ($V_k$), which reduces $N_e$ compared to $N$ following $N_e = \frac{4N-4}{V_k+2}$ (Crow and Kimura 1970). Variance in reproductive success can naturally emerge from particular age-specific demographic characteristics summarized in life tables that contain age-specific (or stage-specific) survival and fecundity rates (Ricklefs and Miller 1999). The impact of life tables characteristics on expected $N_e/N$ ratio has been the focus of a large body of theoretical and empirical works (Nunney 1991, 1996; Waples 2002, 2016a, 2016b; Waples et al. 2018). Accounting for iteroparity and overlapping generations, a meta-analysis of vital rates in 63 species of plants and animals revealed that half of the variance in $N_e/N$ among species can be explained by just two life history traits: adult lifespan and age at maturity (Waples et al. 2013). Interestingly, longevity was the second most important factor explaining differences in genetic diversity across metazoans (Romiguier et al. 2014). However, there is still no attempt to evaluate the extent to which lifetime variance in reproductive

success explains differences in genetic diversity between species with different life table components.

Marine fishes are good candidates to address this issue. They are expected to show a particularly high variance in reproductive success as a result of high abundance, type III survivorship curves (i.e., high juvenile mortality and low adult mortality) and increasing fecundity with age. Consequently, it has been suggested that marine fish species show a marked discrepancy between adult census size and effective population size, resulting in $N_e/N$ ratios potentially smaller than $10^{-3}$. The disproportionate contribution of a few lucky winners to the offspring of the next generation is sometimes referred as the "big old fat fecund female fish" (BOFFFF) effect, a variant of the "sweepstakes reproductive success" hypothesis (Hedgecock 1994; Hedrick 2005; Hedgecock and Pudovkin 2011) that is often put forward to explain low empirical estimates of effective population sizes from genetic data (Hauser and Carvalho 2008). However, subsequent theoretical work showed that low values of $N_e/N$ less than 0.01 can only be generated with extreme age-structure characteristics (Waples 2016b). The real impact of lifetime variance in reproductive success on genetic diversity thus remains unclear, even in species like fish in which its impact is supposed to be strong. Contrasting results have been obtained by comparative studies in marine fishes, including negative relationship between diversity and body size (Waples 1991; Pinsky and Palumbi 2014), fecundity (Martinez et al. 2018), and overfishing (Pinsky and Palumbi 2014). However, these studies relied on few nuclear markers, that could provide inaccurate or biased estimates of genetic diversity (Väli et al. 2008). They also compared species sampled from different locations, thus, likely having different demographic histories, which could blur the relationship between species characteristics and genetic diversity (Ellegren and Galtier 2016).

Here, we compared the genome-average heterozygosity to the life history traits and life table characteristics of 16 marine teleostean species sharing similar Atlantic and Mediterranean distributions. We estimated genetic diversity from unassembled whole-genome reads using `GenomeScope` (Vurture et al. 2017) and checked the validity of these estimates with those obtained using a high-standard reference-based variant calling approach. Using these data, we related species genetic diversity to eight simple quantitative and qualitative life history traits. Then, we built species life tables and determined if the lifetime variance in reproductive success induced by these tables could explain observed differences in genetic diversity using an analytical and a forward-in-time simulation approach. Finally, we generalized our findings by exploring the influence of age-specific survival and fecundity rates on the variance in reproductive success and ultimately genetic diversity via simulated lifetimes tables.

# *Materials and Methods*

## SAMPLING, DNA EXTRACTION, AND WHOLE-GENOME SEQUENCING

We sampled 16 marine teleostean fish species presenting a wide diversity of life history strategies expected to affect genetic diversity (Table 1). All these species share broadly overlapping distributions across the northeastern Atlantic and Mediterranean regions. Sampling was performed at the same four locations for all species: two in the Atlantic (the Bay of Biscay in southwestern France or northwestern Spain and the Algarve in Portugal), and two in the western Mediterranean Sea (the Costa Calida region around Mar Menor in Spain and the Gulf of Lion in France; see Fig. 1A). Individual whole-genome sequencing libraries were prepared following the Illumina TruSeq DNA PCR-Free Protocol and sequenced to an average depth of 20× on an Illumina NovaSeq 6000 platform by Genewiz, Inc. (USA). Raw reads were preprocessed with `fastp` version 0.20.0 (Chen et al. 2018) using default parameters (see Supporting Information).

## ESTIMATION OF GENETIC DIVERSITY

We used `GenomeScope` version 1.0 to estimate individual genome-wide heterozygosity (Vurture et al. 2017). Briefly, this method uses a *k*-mers-based statistical approach to infer overall genome characteristics, including total haploid genome size, percentage of repeat content, and genetic diversity from unassembled short-read sequencing data. We used `jellyfish` version 2.2.10 to compute the *k*-mer profile of each individual (Marçais and Kingsford 2011). The genetic diversity of each species was determined as the median of the individual genome-wide heterozygosity values. We chose the median instead of the mean diversity because it is less sensitive to the possible presence of individuals with nonrepresentative genetic diversity values (e.g. inbred or hybrid individuals) in our samples.

To assess the reliability of `GenomeScope` and detect potential systematic bias, we compared our results with high-standard estimates of genetic diversity obtained after read alignment against available reference genomes (see details in Supporting Information). To perform this test, we used the sea bass (*Dicentrarchus labrax*) and the European pilchard (*Sardina pilchardus*), two species that represent the lower and upper limits of the range of genetic diversity in our dataset (Table 1, Fig. 1D).

## LIFE HISTORY TRAITS DATABASE

We collected seven simple quantitative variables describing various aspects of the biology and ecology of the 16 species: body size, trophic level, fecundity, propagule size, age at maturity, lifespan, and adult lifespan (Tables 1 and S4 for detailed information on bibliographic references). We used the most representative values for each species and each trait when reported traits

**Table 1.** Life history traits and observed genetic diversity of the 16 teleostean marine species.

| Species | Vernacular name | *N* | Genetic diversity (%) | Body size (cm) | Trophic level | Fecundity | Propagule size (mm) | Maturity (years) | Lifespan (years) | Adult lifespan (years) | Parental care | Hermaph-roditism |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *Coryphoblennius galerita* | Montagu's blenny | 16 | 0.607(±0.014) | 7 | 2.28 | NA | 3.3 | 1.5 | 6 | 4.5 | NG | – |
| *Coris julis* | Rainbow wrasse | 20 | 1.172(±0.056) | 27.2 | 3.24 | 169.81 | 0.63 | 1 | 7 | 6 | – | PG |
| *Dicentrarchus labrax* | European sea bass | 20 | 0.375(±0.031) | 102.15 | 3.47 | 12,436.52 | 1.15 | 3 | 15 | 12 | – | – |
| *Diplodus puntazzo* | Sharp-snout seabream | 19 | 0.533(±0.074) | 49.69 | 3.07 | 277.87 | 0.87 | 2 | 10 | 8 | – | RUD |
| *Hippocampus guttulatus* | Long-snouted seahorse | 12 | 0.313(±0.090) | 19.8 | 3.5 | 1.21 | 12 | 0.5 | 5 | 4.5 | MP | – |
| *Lophius budegassa* | Black anglerfish | 20 | 0.225(±0.015) | 103 | 4.23 | 2304.03 | 1.88 | 7.5 | 21 | 13.5 | – | – |
| *Lithognathus mormyrus* | Striped seabream | 20 | 0.553(±0.027) | 37.85 | 3.42 | 214.09 | 0.75 | 2 | 12 | 10 | – | PA |
| *Merluccius merluccius* | European hake | 20 | 0.844(±0.025) | 88.9 | 4.43 | 2294.54 | 1.07 | 3 | 11 | 8 | – | – |
| *Mullus surmuletus* | Striped red mullet | 19 | 1.135(±0.048) | 30.18 | 3.46 | 2569.32 | 0.86 | 1.5 | 6 | 4.5 | – | – |
| *Pagellus erythrinus* | Common pandora | 19 | 1.100(±0.020) | 36 | 3.46 | 2280.46 | 0.77 | 2 | 8 | 6 | – | PG |
| *Serranus cabrilla* | Comber | 19 | 1.205(±0.055) | 30.8 | 3.68 | 37.97 | 0.91 | 2 | 6 | 4 | – | – |
| *Spondyliosoma cantharus* | Black seabream | 19 | 0.478(±0.034) | 35.7 | 3.27 | 425.62 | 2.1 | 3 | 10 | 7 | NG | PG |
| *Symphodus cinereus* | Gray wrasse | 10 | 0.660(±0.125) | 14.1 | 3.3 | 13.20 | 2.87 | 1.5 | 6 | 4.5 | NG | – |
| *Sardina pilchardus* | European pilchard | 20 | 1.415(±0.182) | 20.35 | 2.94 | 22.89 | 1.64 | 1 | 5 | 4 | – | – |
| *Syngnathus typhle* | Broadnosed pipefish | 20 | 0.859(±0.047) | 26.2 | 3.75 | 0.38 | 20 | 1 | 3 | 2 | MP | – |
| *Sarda sarda* | Atlantic bonito | 20 | 0.896(±0.208) | 68.9 | 4.34 | 15,647.73 | 1.3 | 1 | 4 | 3 | – | – |

***Note*: For each species, the number of individuals used for the estimation of genetic diversity; observed median genetic diversity among all individuals (±standard deviation); body size (in centimeters); trophic level; fecundity (in eggs/day); propagule size (in millimeters); age at first maturity (in years), lifespan (in years), adult lifespan (in years, defined as the difference between lifespan and age at maturity), parental care behavior (–, no egg protection; NG, nest guarders; MP, male brood-pouch), and hermaphroditism (–, no hermaphroditism; PG, protogynous; PA, protandrous, RUD, rudimentary). Detailed bibliographic references are provided in the Supporting Information.**

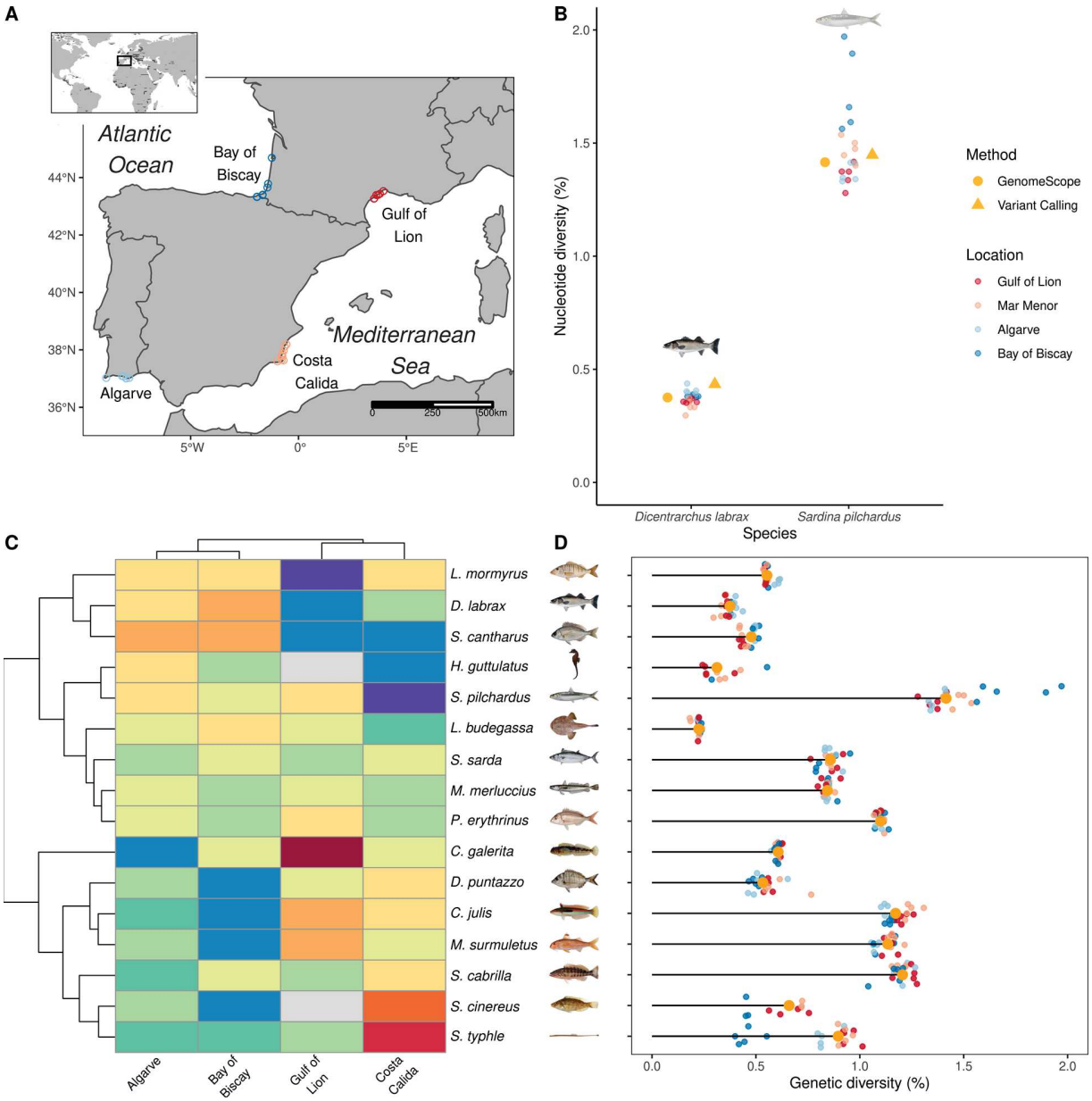



**Figure 1.** **Sampling and estimation of genetic diversity in 16 marine fish species. In panels A, B, and D, the geographical origin of samples is represented by colors. Atlantic: Bay of Biscay (dark blue), Faro region in Algarve (light blue). Mediterranean: Murcia region in Costa Calida (pink), and Gulf of Lion (red). (A) Sampling map of all individuals included in this study. Each point represents the coordinates of a sample taken from one of four locations: two in the Atlantic Ocean and two in the Mediterranean Sea. (B) Genome-wide diversity in the European pilchard (*S. pilchardus*) and European sea bass (*D. labrax*) estimated after variant calling (orange triangle) or from GenomeScope (orange dot: median; smaller dots: individual estimates) (C) Heatmap clustering showing the variance in scaled genetic diversity within species among locations. Each line represents one species, with the corresponding species name written on the right side; every column represents one location. Blue and red colors, respectively, indicate lower and higher genetic diversity within a location for a given species compared to the average species genetic diversity. (D) Individual and median genetic diversity within each species estimated with GenomeScope. Species illustrations were retrieved from Iglésias (2013) with permissions.**

varied among studies due to plasticity, selection, or methodology. In addition, we collected two qualitative variables describing the presence/absence of hermaphroditism and brooding behavior, as revealed by males carrying the eggs in a brood pouch (*Hippocampus guttulatus* and *Syngnathus typhle*) or nest-guarding (*Coryphoblennius galerita*, *Symphodus cinereus*, and *Spondyliosoma cantharus*). Detailed information on data collection is available in Supporting Information.

## CONSTRUCTION OF LIFE TABLES

Life tables summarize survival rates and fecundities at each age during lifetime (Ricklefs and Miller 1999). Thus, they provide detailed information on vital rates that influence the variance in lifetime reproductive success among individuals. This tool is well designed to describe population structure from the probability of survival to a specific age at which a specific number of offspring are produced. Ideally, age-specific survival is estimated by direct demographic measures, such as mark-recapture. Unfortunately, direct estimates of survival were not available for the 16 studied species. We thus followed Benvenuto et al. (2017) to construct species life tables. Age-specific mortality of species $sp$, $m_{sp,a}$, is a function of species body length at age $a$, $L_{sp,a}$, species asymptotic Von Bertalanffy length $L_{sp,inf}$, and species Von Bertalanffy growth coefficient, $K_{sp}$:

$$m_{sp,a} = \left[\left(\frac{L_{sp,a}}{L_{sp,inf}}\right)\right]^{-\frac{1}{5}} \times K_{sp}. \quad (1)$$

Age-specific survival rates, $s_{sp,a}$ were then estimated as

$$s_{sp,a} = e^{-m_{sp,a}}. \quad (2)$$

We collected age-specific length from empirical data and estimated $L_{inf}$ and $K$ values from age-length data as explained in the Supporting Information Appendix, setting survival probability to zero at the maximum age (Appendix S1). When differences in age-specific lengths between sexes were apparent in the literature, we estimated a different age-specific survival curve for each sex. The relationship between absolute fecundity and individual length is usually well fitted with the power-law function ($F = \alpha L^{\beta}$), although some studies also used an exponential function ($F = \alpha e^{\beta L}$) or a linear function ($F = \alpha + L\beta$). We collected empirical estimates of $\alpha$ and $\beta$ and determined age-specific fecundity from the age-specific length and the fecundity-length function reported in the literature for each species. Fecundity was set to zero before the age at first maturity.

## EFFECT OF THE VARIANCE IN REPRODUCTIVE SUCCESS ON THE $N_e/N$ RATIO

To understand how differences in life tables drive differences in genetic diversity between species, we estimated the variance in lifetime reproductive success, $V_k$ and the ensuing ratio $N_e/N$ using the analytic framework developed in `AgeNe` (Waples et al. 2011). `AgeNe` infers $V_k$ using information from life tables only. Hence, the estimated variance in reproductive success estimated is only generated by interindividual differences in fecundity and survival. `AgeNe` assumes constant population size, stable age structure, and no heritability of survival and fecundity. We used the life tables constructed as described above and set the number of new offspring to 1000 per year. This setting is an arbitrary value that has no influence on the estimation of either $V_k$ or $N_e/N$ by `AgeNe`. For all species, we set an initial sex ratio of 0.5 and equal contribution of individuals of the same age (i.e., no sweepstakes reproductive success among same-age individuals). We ran `AgeNe` and estimated $N_e/N$ for each species.

Four life tables components can generate differences in $N_e/N$ between species: age at maturity, age-specific survival rates, age-fecundity relationships and sex-related differences in these components. To determine the role that each parameter plays in shaping levels of genetic diversity among species, we built 16 alternative life tables where the effect of each component was added one after the other, while the others were kept constant across species. Thus, in our null model, age at maturity was set at 1 year for all species, fecundity and survival did not vary with age (constant survival chosen to have 0.01% of individuals remaining at maximum age, following Waples 2016b), and there were no differences between sexes. Next, the effect of each component was tested by replacing these constant values with their biological values in species' life tables. For each of the 16 life tables thus constructed, we tested whether variation in $N_e/N$ explained the variation in observed genetic diversity after scaling these two variables by their maximum value. With this scaling, the correlation between $N_e/N$ and genetic diversity should overlap with the $y = x$ function in cases where a decrease in $N_e/N$ predicts an equal decrease in genetic diversity, indicating a strong predictive power of the components included in life tables.

## FORWARD SIMULATIONS

A complementary analysis of the contribution of life table properties on genetic diversity was performed using forward simulations in `SLiM` version 3.3.1 (Haller and Messer 2017). Compared to the deterministic model implemented in `AgeNe`, the forward simulations include the stochastic variation inherent to the coalescent process and directly predict genetic diversity. Thus, they provide another approach to the problem and can lead to a more intuitive understanding of why vital rates affect $N_e$ over the long-term, and ultimately genetic diversity. We simulated populations with overlapping generations, sex-specific lifespan, and age- and sex-specific fecundity and survival. We used life tables estimated as previously, and sex-specific lifespan estimates were collected in the literature as described above. However, age at maturity was not taken into account in these forward simulations for technical reasons. Age and species-specific fecundity were determined as previously and scaled between 0 (age 0) and 100 (maximum age) within each species. In the simulations, each individual first reproduces and then either survives to the next year or dies following a probability determined by its age and the corresponding life table. We kept population size constant and estimated the mean genetic diversity (i.e., the proportion of heterozygous sites along the locus) over the last 10,000 years of the simulation after the

mutation-drift equilibrium was reached and using 50 replicates (see Supporting Information for further information).

As previously, we evaluated the contribution of each component among 8 alternative life tables by comparing scaled observed and simulated genetic diversity.

### EVALUATING THE IMPACT OF LIFE TABLES BEYOND MARINE FISH

To generalize our understanding of the influence of life tables on genetic diversity beyond the species analyzed in this study, we simulated a wide range of age-specific survival and fecundity curves and explored their effect on the relationship between adult lifespan and variance in reproductive success. To this end, we defined 16 theoretical species with age at first maturity and lifespan equal to that of our real species and then introduced variation in survival and fecundity curves. First, age-specific mortality was simulated following Pinder et al. (1978):

$$M(\text{Age}, \text{Age}+1) = 1 - \exp^{\left(\frac{\text{Age}}{b}\right)^c - \left(\frac{\text{Age}+1}{b}\right)^c}, \quad (3)$$

where $c$ defines the form of the survivorship curve, with $c > 1$, $c = 1$ and $c < 1$ defining respectively *Type I* (e.g., mammals), *Type II* (e.g., birds), and *Type III* (e.g., fish) survival curves. We took values of $c$ from 0.01 to 30 (Fig. 4A). Parameter $b$ was equal to $-\frac{\text{Lifespan}}{\log(0.01)^{1/c}}$ to scale survivorship curves in such a way that 1% of the initial population remains at maximum age.

Second, age-specific fecundity was simulated with two models: constant and exponential. In the first model, fecundity is constant for all ages since maturity. In the second model, fecundity increases or decreases exponentially with age following $F_{\text{Age}} = \exp^{f \times \text{Age}}$, as it is often observed in marine fishes (Curtis and Vincent 2006). We first set $f = 0.142$ as the median of the $f$ values for the 16 species. Second, we took values of $f$ ranging from $-1$ to 1 (Fig. 4A). We scaled maximum fecundity to 1 for all simulations.

For each combination of $c$ and $f$, and for each fecundity model, we simulated all species life tables given age at maturity and lifespan. Then, we ran `AgeNe` and estimated $N_e/N$ for each simulated species and estimated the slope of the regression between adult lifespan and $N_e/N$ across all 16 species. We explored the impact of alternative fecundity–age models on the relationship between adult lifespan and $N_e/N$ (see details in Supporting Information).

### INTRASPECIFIC VARIATION IN GENETIC DIVERSITY

We addressed the potential effects of population structure, demography, and historical contingencies on genetic diversity by examining the extent of spatial variation in genetic diversity between the four populations within each species. First, we evaluated the relative amount of intraspecific compared to interspecific variation in genetic diversity. Then, we applied a $z$-transformation of individual genetic diversity within each species to put spatial differences in within-species diversity on the same scale. To detect similar spatial patterns of genetic diversity among species, we finally performed a hierarchical clustering analysis of the matrix of $z$-transformed genetic diversity values with the `pheatmap` function available in `pheatmap v1.0.12` R package.

### STATISTICAL ANALYSES

All statistical analyses were carried out using `R-3.6.1` (R Core Team, 2018). We fitted beta regression models between genetic diversity and any covariate with the R-package `betareg` version 3.1-3 (Cribari-Neto and Zeileis 2010). We tested statistical interactions between any quantitative and qualitative covariates using likelihood tests with the `lmtest` version 0.9-37 package (Zeileis and Hothorn 2002).

## Results

### WHOLE-GENOME RESEQUENCING DATASET

We resequenced 300 individual genomes from 16 marine teleostean species, with high read quality scores (mean Q30 rate = 92.4%) and moderate duplication rates (10.8%) (Fig. S2). GC content was moderately variable among species and highly consistent among individuals of the same species, except for three individuals that showed a marked discrepancy with the overall GC content of their species (Fig. S2). These three individuals were thus removed from downstream analyses to avoid potential issues due to contamination or poor sequencing quality.

### ESTIMATION OF GENETIC DIVERSITY WITH `GENOMESCOPE`

The `GenomeScope` model successfully converged for all of the 297 individual genomes retained (Fig. S6E). The average depth of sequencing coverage per diploid genome exceeded 20× in most individuals. Estimated genome sizes were very consistent within species (Fig. S6A–C). Estimated levels of genetic diversity were also homogeneous among individuals of the same species with some few exceptions (e.g. *S. cinereus* and *S. typhle*) and most of the variability in genetic diversity was observed between species (Fig. 1D). Two individuals (one *D. puntazzo* and one *P. erythrinus*) showed a surprisingly high genetic diversity (more than twice the average level of their species), indicating possible issues in the estimation of genome-wide heterozygosity. Therefore we removed these individuals from subsequent analysis, although their estimated genome size and GC content matched their average species values (therefore excluding contamination as a cause of genetic diversity estimation failures).

Observed values of genetic diversity ranged from 0.225% for *Lophius budegassa* to 1.415% for *Sardina pilchardus*. We found

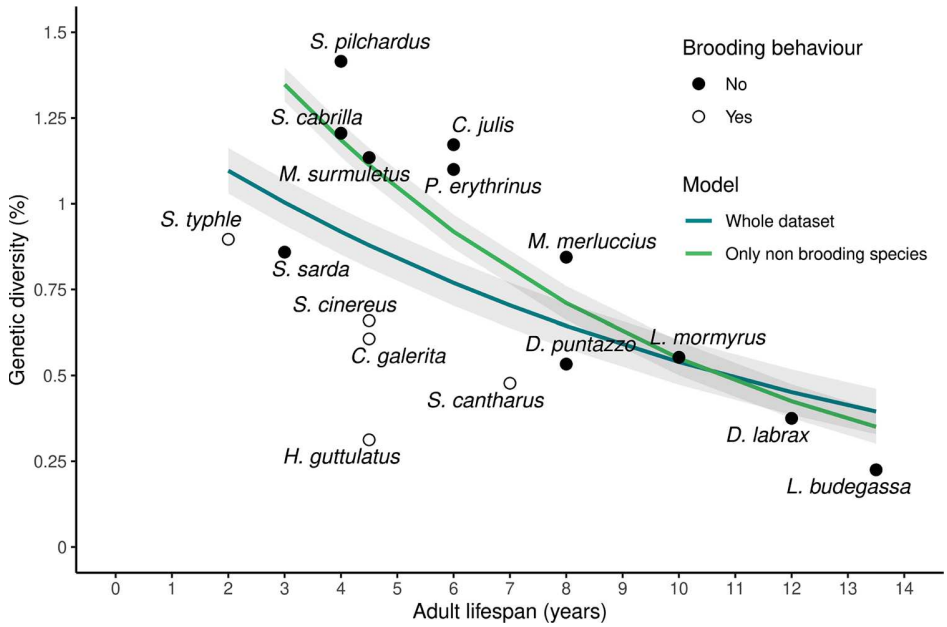



**Figure 2.** **Relationship between species median genetic diversity (%) and adult lifespan, Each point represents the median of the individual genetic diversities for a given species. Adult lifespan is defined as the difference between lifespan and age at first maturity in years. Dot points and empty circles represent nonbrooding species and brooding species, respectively. Blue and green lines represent the beta regression between adult lifespan and genetic diversity considering either the whole dataset (16 species), or the 11 nonbrooding species only, respectively.**

no correlation between species genetic diversity and genome size ($p - value = 0.983$). The estimation of genetic diversity was robust to the choice for $k$-mer lengths ranging from 21 to 25, suggesting a low sensitivity of `GenomeScope` regarding this parameter (Fig. S4). The fraction of reads mapped against reference genomes ranged between 96.72 and 98.50% for *D. labrax* and between 87.45 and 96.42 % for *S. pilchardus* (Table S2, Fig. S3). We found similar species genetic diversity estimates between `GenomeScope` and the `GATK` reference-based variant calling approach for the two control species, representing extreme values within the range of genetic diversity in our dataset (Fig. 1B).

## ADULT LIFESPAN IS THE BEST PREDICTOR OF GENETIC DIVERSITY

We evaluated the effect of several key life history traits that potentially affect species genetic diversity (Table S1).

Two widely used predictors of population size, body size and trophic level, were not significantly correlated to genetic diversity ($p\text{-}value = 0.119$ and 0.676, respectively, Fig. S8A and B). Although we detected a significant negative relationship between the logarithm of fecundity and propagule size ($p\text{-}value = 0.00131$, slope $= -0.4385 \pm 0.1076$) as in Romiguier et al. (2014), we found no significant correlation between either propagule size ($p\text{-}value = 0.561$), or the logarithm of fecundity ($p\text{-}value = 0.785$) and genetic diversity (Fig. S8C and D).

By contrast, both lifespan ($p\text{-}value = 0.011$) and adult lifespan ($p\text{-}value = 0.007$) were significantly negatively correlated with genetic diversity (Table S1, Fig. 2). The percentage of variance explained by each variable reached 43.8% and 42.9 %, respectively. Repeating the same statistical analyses with genetic diversity estimates obtained either only from Mediterranean or Atlantic individuals led to the same results, revealing no effect of within-species population structure on the relationship between genetic diversity and life history traits (Figs. 1C and S9, Table S3).

We found no significant interaction between hermaphroditism and any of the previous variables on genetic diversity. By contrast, parental care showed a significant interaction with lifespan ($p\text{-}value = 0.0011$), adult lifespan ($p\text{-}value = 0.0008$), and body size ($p\text{-}value = 0.0035$) on genetic diversity. Brooding species (nest protection by males for *C. galerita*, *S. cinereus*, and *S. cantharus* and male abdominal brood-pouch for *H. guttulatus* and *S. typhle*) had systematically lower genetic diversity than nonbrooding species with similar adult lifespan.

When considering only nonbrooding species, we found steeper negative correlations and higher percentages of between-species variance in genetic diversity explained by lifespan ($p\text{-}value = 1.017 \times 10^{-7}$, pseudo-$R^2 = 0.851$) and adult lifespan ($p\text{-}value = 1.645 \times 10^{-7}$, pseudo-$R^2 = 0.829$, Fig. 2, Table

S1). To test the relevance of considering this sub-dataset, we estimated the slope of the regression and the pseudo-$R^2$ for all combinations of 11 out of 16 species and compared the distribution of these values to the estimated slope and pseudo-$R^2$ obtained for the 11 nonbrooding species (Fig. S13). The estimated slope for nonbrooders lied outside of the 95% confidence interval of the distribution of estimated slopes ($slope = -0.129$, 95%CI $= [-0.122, -0.049]$) and the same was found for pseudo-$R^2$ (pseudo-$R^2 = 0.829$, 95%CI$= [0.073, 0.727]$). Furthermore, considering nonbrooding species only, there was still no significant correlation between genetic diversity and trophic level ($p\text{-}value = 0.259$), propagule size ($p\text{-}value = 0.170$), and fecundity ($p\text{-}value = 0.390$), but genetic diversity appeared significantly negatively correlated to body size ($p\text{-}value = 6.602 \times 10^{-5}$, pseudo-$R^2 = 0.616$). We did not detect any significant correlation between any trait variable and genetic diversity within the sub-dataset of brooding species. However, this should be taken with caution given the very low number of brooding species ($n = 5$) in our dataset.

Body size and lifespan were highly positively correlated traits in our dataset ($p\text{-}value = 0.0013$, $R^2 = 0.536$, Fig. S7). Thus, using empirical observations only, it was not possible to fully disentangle the impact of each of these traits among the possible determinants of genetic diversity in marine fishes. However, we found important differences in effect sizes for body size (slope $= -0.014$), lifespan ($-0.095$), and adult lifespan ($-0.129$), which rule out body size as a major determinant of diversity in our dataset.

## VARIANCE IN REPRODUCTIVE SUCCESS EXPLAINS LEVELS OF OBSERVED GENETIC DIVERSITY

To understand the mechanisms by which adult lifespan affects genetic diversity and test if it can alone explain our results, we built life tables for each of the 16 species by gradually incorporating age-specific fecundity and survival, age at first maturity, lifespan, and sex-specific differences in these parameters.

Nongenetic estimates of $N_e/N$ ratio obtained with `AgeNe` ranged from 0.104 in *L. budegassa* to 0.671 for *S. cinereus*. When considering the 16 species together, the $N_e/N$ ratio was not significantly correlated with genetic diversity ($p$-value $= 0.0935$). However, four out of five brooding species had low genetic diversity despite high $N_e/N$ ratios (Fig. 3A). As previously observed, removing the five brooders increased the slope and the percentage of variance in genetic diversity explained by the $N_e/N$ ratio above null expectations obtained by removing groups of five species at random ($slope = 1.849$, 95%CI $= [0.048, 1.582]$, pseudo-$R^2 = 0.55$, 95%CI $= [0.004, 0.533]$, Fig. S14). Thus, the $N_e/N$ ratio predicted by life tables was positively correlated to genetic diversity when considering nonbrooding species only (Fig. 3A, $p$-value $= 0.000966$).

Our next step was to determine the impact of each component of life tables as well as their combinations on genetic diversity (Fig. 3C–G). Starting from a null model (Fig. 3C), in which species life tables differed only in lifespan, we found that the $N_e/N$ ratio ranged from 0.558 to 0.733, a variance much lower than that of observed genetic diversities. Then, adding separately age at maturity (Fig. 3D) or age-specific survival (Fig. 3E) did not better predict the range of observed genetic diversities. However, combining age at maturity and age-specific survival (Fig. 3F) or adding only age-specific fecundity (Fig. 3G) enabled us to explain the range of observed diversity values. Finally, combining these three parameters together (age at maturity, age-specific survival, and fecundity, model 8, Fig. S10H) resulted in the best fitted slope for both nonbrooding species and the whole dataset. Adding sex-specific differences in life tables did not improve the fit, however (models 9–16, Fig. S10I–P).

Our final step was to further explore the role of the variance in reproductive success on genetic diversity by simulating genetic diversity at mutation-drift equilibrium with the age-specific vital rates of the 16 species.

We simulated a population of 2000 individuals with age-specific survival and fecundity. As expected, including age-specific vital rates decreased the equilibrium level of genetic diversity compared to expectations under the classical Wright-Fisher model ($\theta = 4N_e\mu = 0.08\%$). It was reduced to 0.070% in the species with the least effect of age-specific vital rates (*C. galerita*), and down to 0.010% in the species with the greatest effect (*L. budegassa*). Again, simulated genetic diversity was not correlated to genetic diversity considering all 16 species ($p - value = 0.297$, Fig. 3B), but significantly positively correlated within the subsample of the 11 nonbrooding species ($p - value = 0.0115$).

## LIFE TABLES DRIVE THE CORRELATION BETWEEN LIFESPAN AND THE $N_e/N$ RATIO

To determine the general effect of life table properties on the relation between adult lifespan and $N_e/N$ beyond the case of marine fish, we modeled 16 life tables with age at maturity and lifespan similar to those observed in our species but with simulated age-specific survival and fecundity (Fig. 4A).

Considering models including constant fecundity with age, we found a significant relationship between adult lifespan and $N_e/N$ for species with type III survivorship curves ($c < 1$) but not for species having an age-specific survivorship curve constant, $c$, superior to 2, including type I species (Fig. 4B). The slope between adult lifespan and $N_e/N$ was steepest for type III species, reaching $-0.053$ for $c = 0.1$. For $c < 2$, the percentage of variation in $N_e/N$ explained by adult lifespan was higher than 60%. Interestingly, it reached a maximum for $c = 1.03$ at 89% and abruptly dropped down around $c = 2$ (Fig. 4B).

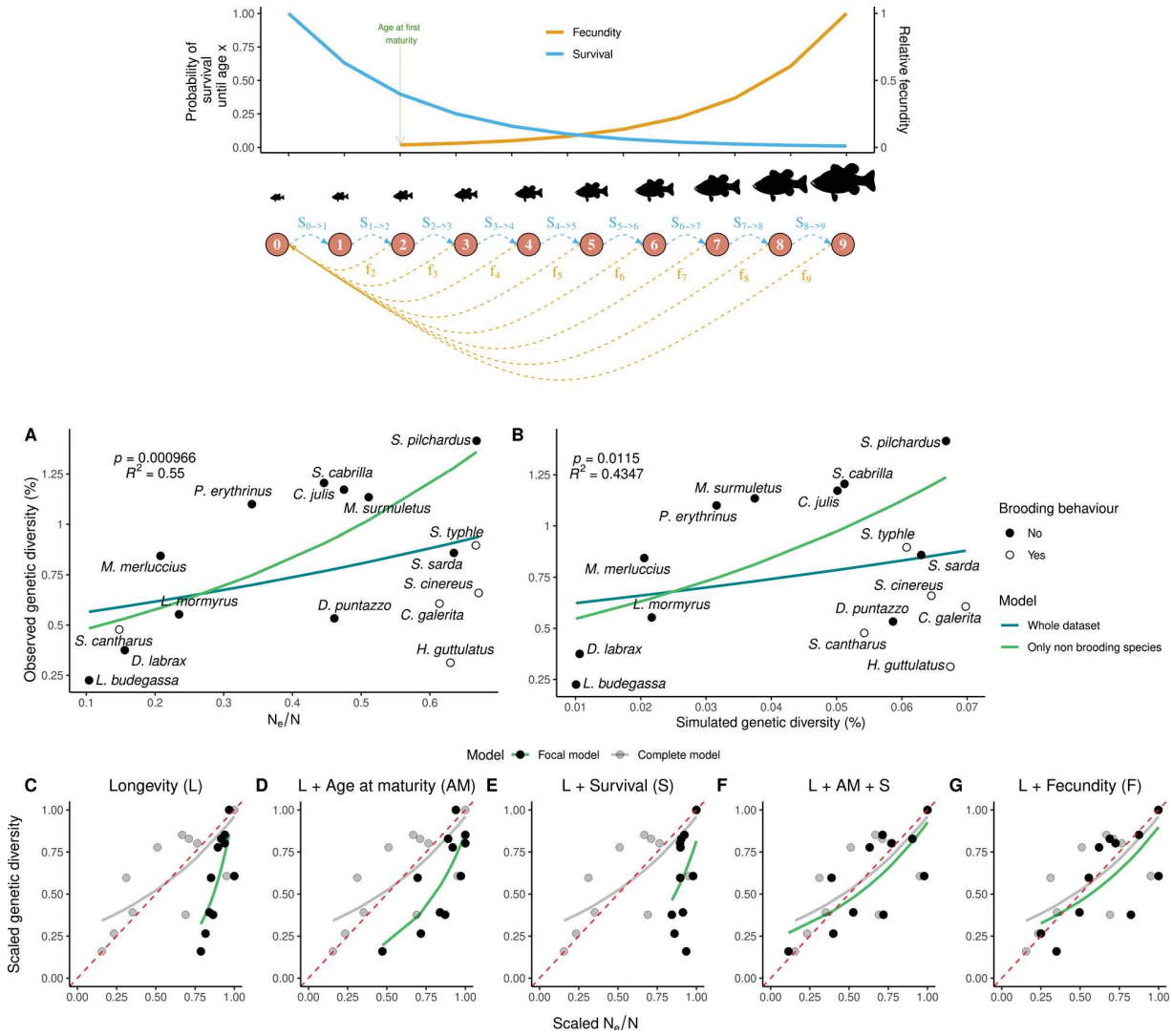



**Figure 3.** **Variance in reproductive success induced by age-specific vital rates and adult lifespan correlates with observed genetic diversity. On top, schematic illustration of age-specific fecundity ($F_{Age}$, in orange) and survival ($S_{Age->Age+1}$, blue) for a given species. (A) and (B) represent the relationship between observed genetic diversity on the *y*-axis and, respectively, $N_e/N$ estimated by `AgeNe`, and simulated genetic diversity with forward-in-time simulations in `SLiM` version 3.31 (Haller and Messer 2017), on *x*-axis. Life tables containing information on age-specific survival, fecundity and lifespan were used for the 16 species. Age at maturity was used only with `AgeNe`. Dot points represent nonbrooding species and empty circles, brooding species. Blue and green lines represent the beta regression between adult lifespan and genetic diversity considering the whole dataset (16 species), and the 11 nonbrooding species only, respectively. The *p*-value and the pseudo-$R^2$ are represented on the top left for each of the two top panels for the nonbrooders model. Panels (C)–(G) represent the relationship between scaled genetic diversity and scaled $N_e/N$ (i.e., divided by the maximum corresponding value) for the 11 nonbrooding species. In each panel, the gray points represent scaled $N_e/N$ estimated from life tables including age at maturity, age-specific fecundity and survival and sex-specific differences (as in Panel A). Black points are scaled estimates of $N_e/N$ from life tables with only: (C) longevity (L); (D) longevity (L) and age at maturity (AM); (E) longevity (L) and age-specific survival (S); (F) longevity (L), age at maturity (AM) and age-specific survival (S); and (G) longevity (L) and age-specific fecundity (F). Beta regression models (gray and green lines) that closely overlap the red dotted line indicate that a decrease in $N_e/N$ leads to a similar decrease in genetic diversity.**

Then, we added an exponential increase in fecundity with age, first taking $f = 0.142$, which is close to the mean empirical estimation across our 16 species (Fig. 4B). The slope between adult lifespan and $N_e/N$ became steeper for type I and type II species and reached $-0.074$ for extreme type III species ($c = 0.01$). When we included this exponential increase of fecundity with age, the percentage of variation explained was superior for approximately all values of $c$, and the abrupt drop of the

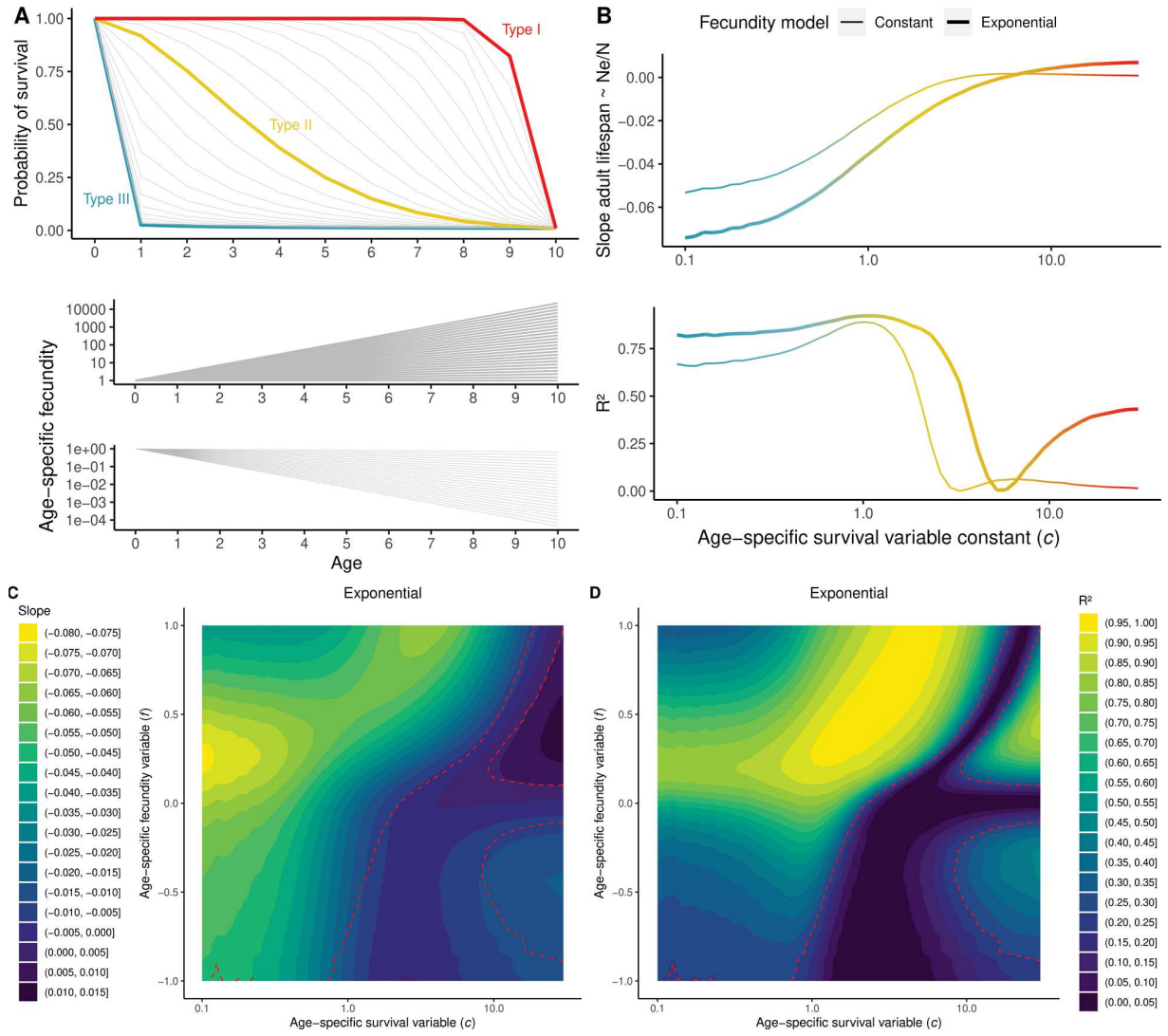



**Figure 4.** Slope of the linear model between adult lifespan and $N_e/N$ ratio estimated with `AgeNe` for different combinations of age-specific survival and fecundity. (A) On top, gradient of survivorship curves simulated, ranging from type III (blue, $c < 1$) characterized by high juvenile mortality and low adult mortality; to type II (orange, $c$ around 1), constant mortality and type I (red), low juvenile mortality and high adult mortality. At the bottom, simulated fecundity either increases or decreases exponentially with age following $F_{Age} = \exp^{f \times Age}$, with $f$ ranging from −1 to 1. Sixteen simulated life tables were constructed with the same values of age at maturity and lifespan as the 16 studied species, and all possible survivorship curves and fecundity-age relationships shown in Panels A and B). Slope and $R^2$ of the regression between adult lifespan and $N_e/N$ ratio for the 16 simulated species as a function of $c$, for constant fecundity with age (thin line) and exponential increase of fecundity with age with $f = 0.142$ (thick line). (C) Slope and (D) $R^2$ of the regression between adult lifespan and $N_e/N$ ratio for the 16 simulated species for a gradient of values of $c$ and $f$. In (C), warmer colors indicate steeper negative slopes; in (D) higher $R^2$.

percentage of variation explained shifted toward higher $c$ values, around $c = 3$. Interestingly, we found significant positive relationships associated with low slope values when $c$ became superior to 10 (type I species).

Then, we compared values of slope and $R^2$ for all $c$ values and for $f$ ranging from −1 to 1 (Fig. 4C and D). The steepest slope between adult lifespan and $N_e/N$ that we obtained reached −0.076 for extreme type III species ($c$ around 0.1), and exponential constant, $f$, between 0.18 and 0.31. For type III and type II species ($c < 1$), both the slope and the percentage of variation explained first increased with increasing exponential constant and then decreased. Significant negative relationships were found for $c < 1$ for any values of $f$, except some extreme values near −1, whereas no significant relationship was found for $c > 1$ when $f$ is negative except for values of $c$ near 1 and values of $f$ near 0. The steepest slope and the highest percentage of variation explained

were obtained for type III species with intermediate values of $f$ ($0.1 < f < 0.5$) and for type II species ($1 < c < 5$) for positive values of $f$. For type I species, as $c$ values increased, higher values of $f$ are needed to obtain a significant negative relationship between adult lifespan and the $N_e/N$ ratio. Above $c > 20$, no significant negative relationship was found for any values of $f$. Again, we found significant positive relationships and low slopes for $c > 15$ and intermediate positive values of $f$.

We found similar results considering a power-law relationship between age and fecundity, with slightly flatter slopes between $N_e/N$ and adult lifespan, and no significant correlations for extreme positive values of $f$ and extreme low values of $c$ (Fig. S16C). In contrast, we found limited or no impact of $f$ on the relationship between $N_e/N$ and adult lifespan, respectively, for the linear and the polynomial age-fecundity models (Fig. S16A and B).

## Discussion

In this study, we used whole-genome high-coverage sequencing data to estimate the genetic diversity of 16 marine teleost fish with similar geographic distribution ranges. We found that adult lifespan was the best predictor of genetic diversity, species with long reproductive lifespans generally having lower genetic diversities (Fig. 2). Longevity was already identified as one of the most important determinants of genetic diversity across metazoans and plants, in which it also correlates with the efficacy of purifying selection (Romiguier et al. 2014; Chen et al. 2017). A positive correlation between longevity and the ratio of nonsynonymous to synonymous substitutions ($dN/dS$) was also found in teleost fishes (Rolland et al. 2020), thus suggesting lower $N_e$ in long-lived species. However, the mechanisms by which lifespan impacts genetic diversity remain poorly understood and may differ among taxonomic groups. Here we showed that age-specific fecundity and survival (i.e., vital rates), summarized in life tables, naturally predict the empirical correlation between adult lifespan and genetic diversity in marine fishes.

### IMPACT OF LIFE TABLES ON GENETIC DIVERSITY

On a broad taxonomic scale including plants and animals, Waples et al. (2013) showed that almost half of the variance in $N_e/N$ estimated from life tables can be explained with only two life history traits: age at maturity and adult lifespan. Therefore, the effect of adult lifespan on genetic diversity should reflect variations in age-specific fecundity and survival across species. If the species vital rates used to derive $N_e/N$ ratios are relatively stable over time, the reduction in $N_e$ due to lifetime variance in reproductive success should not only apply to contemporary time scales but more generally throughout the coalescent time. Thus, a direct impact of life tables on genetic diversity can be expected for iteroparous species with overlapping generations.

Using both an analytical (with `AgeNe`) and a simulation-based (with `SLiM`) approach, we showed that age-specific survival and fecundity rates alone can explain a significant fraction of the variance in genetic diversity among species (Fig. 3A and B). This may appear surprising at first sight, considering that we did not account for variation in census population size among species, which vary by several orders of magnitude in marine fishes (Hauser and Carvalho 2008). Our results thus support that intrinsic vital rates are crucial demographic components of the neutral model to understand differences in levels of genetic diversity in marine fishes. But how generalizable is this finding to other taxa?

Age-specific survivorship curves are one of the main biological components of life tables. Three main types of survivorship curves are classically distinguished: type I curves are characterized by low juvenile and adult mortality combined with an abrupt decrease of survival when approaching the maximum age (e.g., mammals); in type II curves, survival is relatively constant during lifetime (e.g., birds) while type III curves are characterized by high juvenile mortality followed by low adult mortality (e.g., fishes and marine invertebrates). Type III survivorship curves favor the disproportionate contribution of a few lucky winners that survive to old age, compared to type I survivorship curves, where individuals have more equal contributions to reproduction, generating a lower variance in reproductive success. Thus, in type III species, higher lifetime variance in reproductive success is expected as the lifespan of a species increases. By simulating extreme type III survivorship curves ($c = 0.1$) for our 16 species while keeping their true adult lifespans, we found that $N_e/N$ can decrease by at most 0.05 per year of lifespan (Fig. 4B, extreme left). This can theoretically induce up to a 60% difference in genetic diversity between the species with the shortest and the longest lifespans of our dataset. In contrast, we found no correlation between adult lifespan and $N_e/N$ when simulating type I survivorship curves with the true lifespan values of the 16 species studied here (Fig. 4B, $c > 2$), meaning than lifespan and variance in reproductive success may have limited influence in other taxonomic groups, such as birds or mammals.

Another important component of life tables is age-specific fecundity. In marine fishes, fecundity is positively correlated to female ovary size, and the relationship between fecundity and age is usually well approximated with an exponential ($F = a\exp^{Ab}$) or power-law ($F = aA^b$) function. By adding an exponential increase in fecundity with age to our simulations, we found that $N_e/N$ decreases even more strongly with increasing adult lifespan ($N_e/N$ decreases by up to 0.07 per extra year of reproductive life). Using both type III survivorship and exponentially increasing fecundity with age, we could thus predict up to 84% of the

variance in genetic diversity between species with the shortest and longest lifespans.

We found that $N_e/N$ predicted from fecundity alone or age at maturity combined with age-specific survival, explained as much variation in genetic diversity as life tables with both these components (Fig. 3). This is because both these scenarios create sharp differences in fitness between young and old age classes. By contrast, variation in age at maturity alone (all other parameters being held constant across species) introduces some variation in $N_e/N$ because the onset of reproduction age varies from 1 to 7 years depending on the species, but this effect is buffered by the long subsequent period during which adults will reproduce equally. Similarly, the effect of survival alone is insufficient if individuals of all species start reproducing early enough at the age of 1 year.

Although these predicted relationships were pretty close to our empirical findings, genome-wide heterozygosity decreased by about 0.09 per additional year of lifespan in our real dataset (Fig. 2), which seems to be a stronger effect compared to theoretical predictions based on vital rates alone. It is thus likely that other correlates of adult lifespan and unaccounted factors also contribute to observed differences in genetic diversity among species.

## CORRELATED EFFECTS

When relating measures of diversity with the estimates of $N_e/N$ derived from life tables, we did not take into account differences in census size ($N$) between species. Population census sizes can be huge and are notoriously difficult to estimate in marine fishes. For that reason, abundance data remain largely unavailable for the 16 species of this study. We nevertheless expect long-lived species to have lower abundance compared to short-lived species because in marine fishes $N$ is generally negatively correlated to body size (White et al. 2007), which is itself positively correlated to adult lifespan in our dataset (Fig. S7). Hence, while we have demonstrated here that variation in vital rates has a direct effect on long-term genetic diversity, the slope between adult lifespan and genetic diversity may be inflated by uncontrolled variation in $N$. Recent genome-wide comparative studies found negative correlations between $N_e/N$ and $N$ in pinnipeds (Peart et al. 2020) as well as between genetic diversity and body size in butterflies and birds (Brüniche-Olsen et al. 2019; Mackintosh et al. 2019). Here, a highly significant negative correlation was found between genetic diversity and body size and the strength of that correlation was comparable to that found in a meta-analysis of microsatellite diversity using catch data and body size as proxies for fish abundance (Mccusker and Bentzen 2010). We note, however, that body size was not as good a predictor of genetic diversity as lifespan and adult lifespan for the 11 nonbrooding species and it was even not significant in the whole dataset of the 16 species (Table S1).

Another potentially confounding effect is the impact of $r/K$ strategies, which are the main determinant of genetic diversity across metazoans (Romiguier et al. 2014). In our dataset, fecundity and propagule size (proxies for the $r/K$ gradient) showed only little variance compared to their range of variation across metazoans, and none of them were correlated to adult lifespan. However, we found that the five brooding species of our dataset, which are typical $K$-strategists, displayed lower genetic diversities with respect to their adult lifespan (Fig. 2). Most interestingly, when these species were removed from the analysis, the effect of adult lifespan on genetic diversity was amplified, indicating a potentially confounding effect of parental care in marine fishes. Alternatively, low levels of genetic diversity in brooding species can also be explained by underestimated lifetime variance in reproductive success by `AgeNe` due to unaccounted variance in reproductive success within age classes. This may be particularly important in males as the age-fecundity relationship is empirically estimated for females only. This effect could be high for species with strong sexual selection and mate choice (Hastings 1988; Naud et al. 2009). Moreover, most of these species inhabit lagoons and coastal habitats, corresponding to smaller and more instable ecological niches compared to species with no parental care, thus potentially resulting in lower long-term abundances. The discrepancy introduced by brooders in the relationship that we observed here between adult lifespan and genetic diversity may thus involve a variety of effects that remain to be elucidated.

Temporal fluctuations of effective population size may also have impacted observed levels of genetic diversity (Nei et al. 1975). All studied species possibly went through a bottleneck during the Last Glacial Maximum (Jenkins et al. 2018), which may have simultaneously decreased their genetic diversities. As the time of return to mutation-drift equilibrium is positively correlated to generation time, which is itself directly linked to adult lifespan, we may expect long-lived species to have recovered less genetic variation than short-lived species following their latest bottleneck. Moreover, long-lived species may not have recovered their pre-bottleneck population sizes as rapidly as short-lived species. If true, the negative relationship between adult lifespan and genetic diversity may be inflated compared to the sole effect of life tables.

Variation in mutation rates between species could not be accounted for due to a lack of estimates. However, if species-specific mutation rates were correlated with adult lifespan, we would expect mutation rate variation to have a direct effect on genetic diversity. Mutation rate could be linked with species life history traits through three possible mechanisms. First, the drift-barrier hypothesis predicts a negative correlation between species effective population size and the per-generation mutation rate (Sung et al. 2012). However, this hypothesis cannot explain our results because species with the highest effective population sizes

have the highest genetic diversity. Second, species with larger genome size tend to have more germline cell divisions, hence possibly higher mutation rates. But we did not find any correlation between genome size and genetic diversity or any other qualitative and quantitative life history traits. Third, species with longer generation time, which is positively correlated to lifespan and age at maturity, may have higher per-generation mutation rate as older individuals accumulate more germinal mutations throughout their lives. Again, under this assumption, we would expect species with longer lifespan to have higher mutation rate and genetic diversity, which goes against our observations. In summary, variation in mutation rates among species due to differences in lifespan is unlikely to explain the negative lifespan-diversity relationship we observed. If anything, variation in mutation rates should theoretically oppose this relationship.

Using one of the few direct estimates of the per-generation mutation rate in fish, Feng et al. (2017) explained the surprisingly low nucleotide diversity found in the Atlantic herring *Clupea harengus* ($\pi = 0.3\%$) by a very low mutation rate of $2 \times 10^9$ per base per generation estimated from pedigree analysis. Although the herring is one of the most abundant and fecund pelagic species in the North Atlantic Ocean, its genetic diversity appears approximately 80% lower than that of the European pilchard *S. pilchardus*, another member of the *Clupeidae* family that shows the highest diversity in our study. Even if *C. harengus* has a larger body size (approximately 30 cm, compared to 20 cm for *S. pilchardus*; Froese et al. 2000), it has above all a much longer lifespan (between 12 and 25 years) and a later age at maturity (between 2 and 6.5 years) (Jennings and Beverton 1991). Considering even the lowest estimate of adult lifespan reported for the herring (10 years), the corresponding genetic diversity predicted by our model linking adult lifespan to genetic diversity would be around 0.5%, which is pretty close to the empirical estimate.

Finally, we did not take into account the erosion of neutral diversity through linked selection. Addressing that issue would need to generate local estimates of nucleotide diversity and population recombination rate along the genome of each species using resequencing data aligned to a reference assembly, which was out of the scope of this study. The predicted effect of linked selection could be, however, to remove more diversity in species with large compared to small $N_e$. It is therefore likely that linked selection would rather attenuate the negative relationship between adult lifespan and genetic diversity compared to neutral predictions.

## CONCLUSION

Here we used a simple approach to generate reference-free genome-wide estimates of diversity with *k*-mer analysis. Tested on two species with genetic diversities ranging from 0.22% to 1.42% the *k*-mer approach performed close to the level of a high-standard reference-based method in capturing fine-scale variation in diversity between evolutionary lineages and even populations of the same species. This opens the possibility to address the determinants of genetic diversity in other groups of taxa at limited costs without relying on existing genomics resources. Across metazoans, the level of genetic diversity showed no significant relationship with the species' conservation status (Romiguier et al. 2014). Studies performed at lower phylogenetical scales such as in Darwin's finches and pinnipeds, however, found reduced contemporary genetic diversity in threatened compared to nonthreatened species (Brüniche-Olsen et al. 2019; Peart et al. 2020). Our results complement and extend this literature by showing the importance of taking into account life tables in comparisons of genetic diversity between species.

## AUTHOR CONTRIBUTIONS

P.B., T.B., and P.-A.G. wrote the manuscript. P.B. and P.-A.G performed fieldwork. P.B. performed molecular experiments, and all bioinformatics and evolutionary genomics analyses with inputs from T.B. and P.-A.G. P.-A.G. conceived the project and managed financial support and genome sequencing.

## ACKNOWLEDGMENTS

The data used in this work were partly produced with the support of the GenSeq genotyping and sequencing platform, and bioinformatics data analysis benefited from the Montpellier Bioinformatics Biodiversity MBB platform, both platforms being supported by ANR program "Investissements d'avenir" (ANR-10-LABX-04-01). We would like to thank Rémy Dernat and Khalid Belkhir for their invaluable assistance in data storage, management and processing. We are grateful to the colleagues who provided us with samples as well as to those who facilitated or participated in sampling: F. Schlichta, T. Pastor, R. Castilho, R. Cunha, R. Lechuga, D. Pilo, C. Mena, J. Charton, T. Robinet, A. Darnaude, S. Vaz, M. Duranton, N. Bierne, S. Villéger, S. Blouet, as well as the fishermen and employees of fish markets and fish auctions. This work was supported by the ANR grant CoGeDiv ANR-17-CE02-0006-01. The authors declare no conflicts of interest.

## DATA ARCHIVING

Data and scripts used in this study are freely available in the GitHub repository https://github.com/pierrebarry/life_tables_genetic_diversity_marine_fishes. All sampling metadata are accessible under GEOME at the CoGeDiv Project Homepage: https://geome-db.org/workbench/project-overview?projectId=357. Sequence reads have been deposited in the GenBank Sequence Read Archive under the accession code BioProject ID PRJNA777424.

## REFERENCES

Benvenuto, C., I. Coscia, J. Chopelet, M. Sala-Bozano, and S. Mariani. 2017. Ecological and evolutionary consequences of alternative sex-change pathways in fish. Sci. Rep. 7:9084.

Brüniche-Olsen, A., K. F. Kellner, and J. A. DeWoody. 2019. Island area, body size and demographic history shape genomic diversity in Darwin's finches and related tanagers. Mol. Ecol. 28:4914–4925.

Chen, J., S. Glémin, and M. Lascoux. 2017. Genetic diversity and the efficacy of purifying selection across plant and animal species. Mol. Biol. Evol. 34:1417–1428.

Chen, S., Y. Zhou, Y. Chen, and J. Gu. 2018. Fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics 34:i884–i890.

Cribari-Neto, F., and A. Zeileis. 2010. Beta regression. R. J. Stat. Softw. 34:1–24.

Crow, J. F., and M. Kimura. 1970. An Introduction to Population Genetics Theory. Harper & Row, Manhattan, NY.

Curtis, J. M. R., and A. C. J. Vincent. 2006. Life history of an unusual marine fish: survival, growth and movement patterns of *Hippocampus guttulatus* Cuvier 1829. J. Fish Biol. 68:707–733.

DeWoody, J. A., A. M. Harder, S. Mathur, and J. R. Willoughby. 2021. The long-standing significance of genetic diversity in conservation. Mol. Ecol. 30:4147–4154.

Díez-Del-Molino, D., F. Sánchez-Barreiro, I. Barnes, M. T. P. Gilbert, and L. Dalén. 2018. Quantifying temporal genomic erosion in endangered species. Trends Ecol. Evol. 33:176–185.

Ellegren, H., and N. Galtier. 2016. Determinants of genetic diversity. Nat. Rev. Genet. 17:422–433.

Falconer, D. S. 1989. Introduction to quantitative genetics. 3rd ed. Longman Scientific & Technical, Harlow.

Feng, C., M. Pettersson, S. Lamichhaney, C.-J. Rubin, N. Rafati, M. Casini, A. Folkvord, and L. Andersson. 2017. Moderate nucleotide diversity in the Atlantic herring is associated with a low mutation rate. eLife 6:e23907.

Frankham, R. 1995. Conservation genetics. Annu. Rev. Genet. 29:305–327.

Froese, R and D. Pauly, eds. 2000. FishBase 2000: concepts, design and data sources. WorldFish, 1594 Penang, Malaysia.

Haller, B. C., and P. W. Messer. 2017. SLiM 2: flexible, interactive forward genetic simulations. Mol. Biol. Evol. 34:230–240.

Hastings, P.A. 1988. Female choice and male reproductive success in the angel blenny, *Coralliozetus angelica* (Teleostei: Chaenopsidae). Anim. Behav. 36:115–124.

Hauser, L., and G. R. Carvalho. 2008. Paradigm shifts in marine fisheries genetics: ugly hypotheses slain by beautiful facts. Fish Fish. 9:333–362.

Hedgecock, D. 1994. Does variance in reproductive success limit effective population sizes of marine organisms? Pp. 122–134 *in*, ed. Genetic and evolution of aquatic organisms. Chapman and Hall, London.

Hedgecock, D., and A. I. Pudovkin. 2011. Sweepstakes reproductive success in highly fecund marine fish and shellfish: a review and commentary. Bull. Mar. Sci. 87:971–1002.

Hedrick, P. 2005. Large variance in reproductive success and the $N_e/N$ ratio. Evolution 59:1596–1599.

Jenkins, T. L., R. Castilho, and J. R. Stevens. 2018. Meta-analysis of north-east Atlantic marine taxa shows contrasting phylogeographic patterns following post-LGM expansions. PeerJ 6:e5684.

Jennings, S., and R. J. H. Beverton. 1991. Intraspecific variation in the life history tactics of Atlantic herring (*Clupea harengus* L.) stocks. ICES J. Marine Sci. 48:117–125.

Kimura, M. 1983. The neutral theory of molecular evolution. Cambridge Univ. Press, Cambridge.

Kimura, M., and J. F. Crow. 1964. The number of alleles that can be maintained in a finite population. Genetics 49:725–738.

Lande, R. 1995. Mutation and conservation. Conserv. Biol. 9:782–791.

Lande, R., and G. F. Barrowclough. 1987. Effective population size, genetic variation, and their use in population management. Pp. 87–124 *in* M. E. Soulé, ed. Viable populations for conservation. Cambridge Univ. Press, Cambridge.

Leffler, E. M., K. Bullaughey, D. R. Matute, W. K. Meyer, L. Ségurel, A. Venkat, P. Andolfatto, and M. Przeworski. 2012. Revisiting an old riddle: what determines genetic diversity levels within species? PLoS Biol. 10:e1001388.

Leroy, T., M. Rousselle, M.-K. Tilak, A. Caizergues, C. Scornavacca & M. Recuerda et al. 2021. Island songbirds as windows into evolution in small populations. Current Biology, 31:1303–1310.

Lewontin, R. C. 1974. The genetic basis of evolutionary change. Columbia biological series, No. 25. Columbia Univ. Press, New York.

Mackintosh, A., D. R. Laetsch, A. Hayward, B. Charlesworth, M. Waterfall, R. Vila, et al. 2019. The determinants of genetic diversity in butterflies. Nat. Commun. 10:1–9.

Marçais, G., and C. Kingsford. 2011. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. Bioinformatics 27:764–770.

Martinez, A. S., J. R. Willoughby, and M. R. Christie. 2018. Genetic diversity in fishes is influenced by habitat type and life-history variation. Ecol. Evol. 8:12022–12031.

Mccusker, M. R., and P. Bentzen. 2010. Positive relationships between genetic diversity and abundance in fishes. Mol. Ecol. 19:4852–4862.

Naud, M.-J., J. M. R. Curtis, L. C. Woodall, and M. B. Gaspar. 2009. Mate choice, operational sex ratio, and social promiscuity in a wild population of the long-snouted seahorse *Hippocampus guttulatus*. Behav. Ecol. 20:160–164.

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

Nunney, L. 1991. The influence of age structure and fecundity on effective population size. Proc. Biol. Sci. 246:71–76.

———.. 1996 The influence of variation in female fecundity on effective population size. Biol. J. Linn. Soc. 59:411–425.

Peart, C. R., S. Tusso, S. D. Pophaly, F. Botero-Castro, C.-C. Wu, D. Aurioles-Gamboa, et al. 2020. Determinants of genetic variation across eco-evolutionary scales in pinnipeds. Nat. Ecol. Evol. 4:1–10.

Pinder, J. E., J. G. Wiener, and M. H. Smith. 1978. The Weibull distribution: a new method of summarizing survivorship data. Ecology 59:175–179.

Pinsky, M. L., and S. R. Palumbi. 2014. Meta-analysis reveals lower genetic diversity in overfished populations. Mol. Ecol. 23:29–39.

Ricklefs, R. E., and G. L. Miller. 1999. Ecology. 4th ed. W. H. Freeman, New York.

Rolland, J., D. Schluter, and J. Romiguier. 2020. Vulnerability to fishing and life history traits correlate with the load of deleterious mutations in teleosts. Mol. Biol. Evol. 37:2192–2196.

Romiguier, J., P. Gayral, M. Ballenghien, A. Bernard, V. Cahais, A. Chenuil, et al. 2014. Comparative population genomics in animals uncovers the determinants of genetic diversity. Nature 515:261–263.

Sung, W., M. S. Ackerman, S. F. Miller, T. G. Doak, and M. Lynch. 2012. Drift-barrier hypothesis and mutation-rate evolution. Proc. Natl. Acad. Sci. 109:18488–18492.

Väli, Ü., A. Einarsson, L. Waits, and H. Ellegren. 2008. To what extent do microsatellite markers reflect genome-wide genetic diversity in natural populations? Mol. Ecol. 17:3808–3817.

Vurture, G. W., F. J. Sedlazeck, M. Nattestad, C. J. Underwood, H. Fang, J. Gurtowski, et al. 2017. GenomeScope: fast reference-free genome profiling from short reads. Bioinformatics 33:2202–2204.

Waples, R. S. 1991. Heterozygosity and life-history variation in bony fishes: an alternative view. Evolution 45:1275–1280.

Waples, R. S. 2002. Evaluating the effect of stage-specific survivorship on the $N_e/N$ ratio. Mol. Ecol. 11:1029–1037.

Waples, R. S. 2016a. Life-history traits and effective population size in species with overlapping generations revisited. Heredity 117:241–250.

Waples, R. S. 2016b. Tiny estimates of the Ne/N ratio in marine fishes: are they real? J. Fish Biol. 89:2479–2504.

Waples, R. S., C. Do, and J. Chopelet. 2011. Calculating $N_e/N$ in age-structured populations: a hybrid Felsenstein-Hill approach. Ecology 92:1513–1522.

Waples, R. S., G. Luikart, J. R. Faulkner, and D. A. Tallmon. 2013. Simple life-history traits explain key effective population size ratios across diverse taxa. Proc. R. Soc. B 280:20131339.

Waples, R. S., S. Mariani, and C. Benvenuto. 2018. Consequences of sex change for effective population size. Proc. R. Soc. B 285:20181702.

White, E. P., S. K. M. Ernest, A. J. Kerkhoff, and B. J. Enquist. 2007. Relationships between body size and abundance in ecology. Trends Ecol. Evol. 22:323–330.

Wright, S. 1969. The theory of gene requencies. Evolution and the genetics of populations: A Treatise in Three Volumes, Vol. 2. Univ. of Chicago Press, Chicago, IL.

Zeileis, A., and T. Hothorn. 2002. Diagnostic checking in regression relationships. R News 2:7–10.

## *Supporting Information*

Additional supporting information may be found online in the Supporting Information section at the end of the article.

**Figure S1**: *k-mer* frequency-coverage relationship and estimation by `GenomeScope` v.1.0 (Vurture et al., 2017) for two species, *L. budegassa* and *M. surmuletus*.

**Table S1**: Statistical relationships between species genetic diversity and life history traits.

**Table S2**: Mapping and variant calling statistics for *D. labrax* and *S. pilchardus* individuals.

**Table S3**: Statistical relationships between different estimations of species genetic diversity and life history traits.

**Figure S2**: Number of reads ($10^9$ bp), percentage of reads with quality superior to Q30, duplication rate and GC content after filtering, correcting and trimming steps carried out with `fastp` v.0.20.0 (Chen et al. 2018).

**Figure S3**: Mapping statistics.

**Figure S4**: Effect of $k-mer$ length on genetic diversity estimation with `GenomeScope` v.1.0 (Vurture et al. 2017).

**Figure S5**: Individual whole-genome sequences features estimated with `GenomeScope` v.1.0 (Vurture et al. 2017).

**Figure S6**: Relationship between individual mean genome-wide heterozygosity estimated with the k-mer based reference-free approach in GenomeScope (x-axis), and the high standard reference-based approach in GATK (y-axis), for european sea bass (*D. labrax*, dots) and european pilchard (*S. pilchardus*, triangle).

**Figure S7**: Correlation matrix between genetic diversity and all quantitative life history traits.

**Figure S8**: Relationship between species median genetic diversity (%) and 5 covariables.

**Figure S9**: Effect of population structure on genetic diversity estimates.

**Figure S10**: Relationship between relative species genetic diversity and relative $N_e/N$ estimated by `AgeNe` for 16 sets of life tables.

**Figure S11**: Relationship between relative species genetic diversity and simulated genetic diversity with forward-in-time simulations for 16 sets of life tables.

**Figure S12**: Residuals of the linear model between genetic diversity and variance in reproductive success estimated from various combinations of life tables from a model with slope equals 1 and intercept 0.

**Figure S13**: Distribution of slope and pseudo $R^2$ of the beta regression between adult lifespan and genetic diversity for random subsets of 11 species.

**Figure S14**: Distribution of slope and pseudo $R^2$ of the beta regression between $N_e/N$ and genetic diversity for random subsets of 11 species.

**Figure S15**: Distribution of slope and pseudo $R^2$ of the beta regression between simulated, with `SLiM` v.3.3.1, and observed genetic diversity for random subsets of 11 species.

**Figure S16**: Slope of and proportion of variance explained by linear models between adult lifespan and $N_e/N$ estimated with `AgeNe` for different combinations of age-specific survival and fecundity for three fecundity-age models: linear, polynomial and power-law.

**Figure S17**: Population size count for the 50 iterations of the 16 species for set 1 of life tables (age at first maturity at 1 year old, constant age-specific survival rate, constant age-specific fecundity and no differences between sex-specific life tables).

**Figure S18**: Population size count for the 50 iterations of the 16 species for set 2 of life tables (age at first maturity at 1 year old, increasing age-specific survival rate, constant age-specific fecundity and no differences between sex-specific life tables).

**Figure S19**: Population size count for the 50 iterations of the 16 species for set 3 of life tables (age at first maturity at 1 year old, constant age-specific survival rate, increasing age-specific fecundity and no differences between sex-specific life tables).

**Figure S20**: Population size count for the 50 iterations of the 16 species for set 4 of life tables (age at first maturity at 1 year old, increasing age-specific survival rate, increasing age-specific fecundity and no differences between sex-specific life tables).

**Figure S21**: Population size count for the 50 iterations of the 16 species for set 5 of life tables (age at first maturity at 1 year old, constant age-specific survival rate, constant age-specific fecundity and sex-specific differences in life tables).

**Figure S22**: Population size count for the 50 iterations of the 16 species for set 6 of life tables (age at first maturity at 1 year old, increasing age-specific survival rate, constant age-specific fecundity and sex-specific differences in life tables).

**Figure S23**: Population size count for the 50 iterations of the 16 species for set 7 of life tables (age at first maturity at 1 year old, constant age-specific survival rate, increasing age-specific fecundity and sex-specific differences in life tables).

**Figure S24**: Population size count for the 50 iterations of the 16 species for set 8 of life tables (age at first maturity at 1 year old, increasing age-specific survival rate, increasing age-specific fecundity and sex-specific differences in life tables).

**Figure S25**: Genetic diversity simulated for each of the 16 species for set 1 of life tables (age at first maturity at 1 year old, constant age-specific survival rate, constant age-specific fecundity and no differences between sex-specific life tables).

**Figure S26**: Genetic diversity simulated for each of the 16 species for set 2 of life tables (age at first maturity at 1 year old, increasing age-specific survival rate, constant age-specific fecundity and no differences between sex-specific life tables).

**Figure S27**: Genetic diversity simulated for each of the 16 species for set 3 of life tables (age at first maturity at 1 year old, constant age-specific survival rate, increasing age-specific fecundity and no differences between sex-specific life tables).
**Figure S28**: Genetic diversity simulated for each of the 16 species for set 4 of life tables (age at first maturity at 1 year old, increasing age-specific survival rate, increasing age-specific fecundity and no differences between sex-specific life tables).
**Figure S29**: Genetic diversity simulated for each of the 16 species for set 5 of life tables (age at first maturity at 1 year old, constant age-specific survival rate, constant age-specific fecundity and sex-specific differences in life tables).
**Figure S30**: Genetic diversity simulated for each of the 16 species for set 6 of life tables (age at first maturity at 1 year old, increasing age-specific survival rate, constant age-specific fecundity and sex-specific differences in life tables).
**Figure S31**: Genetic diversity simulated for each of the 16 species for set 7 of life tables (age at first maturity at 1 year old, constant age-specific survival rate, increasing age-specific fecundity and sex-specific differences in life tables).
**Figure S32**: Genetic diversity simulated for each of the 16 species for set 8 of life tables (age at first maturity at 1 year old, increasing age-specific survival rate, increasing age-specific fecundity and sex-specific differences in life tables).
Data S1
Data S2

# Chapitre 2

# Impact des traits d'histoire de vie sur la spéciation des poissons marins: génomique comparative de la zone de suture Atlantique - Mediterranée

**Résumé**

La spéciation est un processus durant lequel une espèce diverge en deux lignées qui accumulent des barrières au flux génique jusqu'à l'isolement reproductif total (Coyne and Orr, 2004). Ce processus est complexe et implique de nombreuses forces évolutives qui peuvent accélérer, maintenir ou éroder les barrières à l'isolement reproductif (Coyne and Orr, 2004). Comprendre les facteurs qui modèlent la dynamique de mise en place des barrières est nécessaire pour comprendre les déterminants de la spéciation. Avec l'avènement des techniques de séquençage à haut débit, de nombreuses études ont caractérisé le paysage génomique de différentiation et de divergence de lignées en cours de spéciation dans de nombreuses taxa différents (Martin et al., 2013; Burri et al., 2015). Ces études ont aussi permis de mettre en évidence l'importance de l'histoire démographique et de l'architecture génomique sur la mise en place des barrières. Toutefois, on connaît très peu aujourd'hui l'impact des traits d'histoire de vie sur la spéciation. On peut par exemple supposer que l'érosion de barrières au flux génique soit plus importantes chez des espèces à forte capacité de dispersion (Barton and Bengtsson, 1986) ; ou que la rapidité de mise en place des barrières soit plus importante chez les espèces à faible temps de génération (Orr, 1995).

Dans ce deuxième chapitre, j'ai comparé l'histoire évolutive de 19 espèces de poissons marins téléostéens subdivisées en deux lignées Atlantiques (Atl) et Méditerranéennes (Med) avec des niveaux variés de divergence. Nous avons échantillonné 20 individus par espèce dans 4 localités en Atlantique et en Méditerranée. Par une analyse du polymorphisme du génome complet, nous avons comparé les multiples facettes de la spéciation de chaque espèce et notamment : i) le niveau de différentiation génétique Atl-Med mesuré par le $F_{ST}$ (Fig. 1B), ii) la structure de la population (Fig. 1D), iii) le niveau de divergence absolue ($d_{XY}$) et iv) de divergence nette ($d_a$) (Fig. 1B), v) la présence d'introgression dans la zone de suture (Fig. 2), vi) la distribution des temps de coalescence (Fig. 3) et vii) le niveau de semi-perméabilité des génomes (Fig. 4). Enfin, nous avons comparé ces différentes caractéristiques génétiques à 9 traits d'histoire de vie qui sont supposés impactés la taille efficace des populations, le taux de migration et le temps de génération (Fig. 6-7).

Les 19 espèces étudiées présentent un gradient de différentiation génétique de 0 pour la bonite (*S. sarda*) à 0.64 pour le marbré (*L. mormyrus*); ce gradient de différentiation est également corrélée à la divergence nette, c'est à dire aux différences accumulées depuis la subdivision dans les séquences des populations Atl. et Med allant de 0 à 1.05% pour les deux espèces précédentes : les espèces les plus différenciées sont également celles les plus divergentes.

Nous avons aussi montré une grande diversité de structure de populations : certaines espèces présentent une différentiation géographique faible comme la sardine *S. pilchardus*; d'autres une

subdivision génétique et un mélange des génomes des deux lignées en Atlantique pour le bar commun (*D. labrax*) ou en Méditerranée pour le rouget (*M. surmuletus*); enfin certaines sont divisées en deux lignées très divergentes comme la blennie coiffée (*C. galerita*).

Malgré ces différences, l'introgression dans la zone de transition est présente chez plusieurs espèces tout le long du gradient de différentiation. Toutefois, les espèces les plus différenciées sont celles également chez lesquelles on ne détecte plus de flux de gène contemporain entre les populations à l'extérieur de la zone de suture. L'isolement reproducteur pourrait être assez fort pour que les fragments d'ADN introgressés soit rapidement éliminés par la sélection au sein de la zone de suture.

Enfin, nous avons montré que les relations entre traits d'histoire de vie et spéciation semblent complexes. Cependant, nous avons identifié 3 traits d'histoire de vie qui permettent d'expliquer chacun une des facettes de la spéciation. Premièrement, la différentiation génétique Atlantique-Méditerranée est négativement corrélée à la durée de vie larvaire (PLD), supposée être une bonne indicatrice des capacités de dispersion des poissons marins (Selkoe and Toonen, 2011), ce qui est conforme à un modèle d'isolement avec migration ou une ré-homogénéisation de lignées après une période d'allopatrie (Wright, 1931; Sedghifar et al., 2016). Deuxièmement, la divergence ($d_{XY}$) est négativement corrélée à la taille du corps, ce qui pourrait montrer un effet de la taille efficace de la population sur la diversité ancestrale et la divergence. Enfin, la longévité est également négativement corrélée au temps de séparation ancestral des deux lignées (en génération), ce résultat étant conforme à une barrière biogéographique identique à toutes les espèces et à une évolution plus lente chez les espèces à temps de génération élevé.

A travers cette étude, nous avons démontré une grande diversité d'histoire évolutive chez 19 espèces de poissons marins dans une même zone de transition. L'histoire de divergence des lignées Atlantiques - Mediterranée pourrait être plus ancienne comme suggérée par la distribution des temps de coalescence (Fig. 3). De futures analyses sur l'effet de la sélection en liaison et de l'impact de la recombinaison sur la divergence pourrait éclairer l'histoire évolutive de chaque paire d'espèce et les relations entre traits d'histoire de vie et spéciation.

# References

Barton, N. and Bengtsson, B. O. (1986). The barrier to genetic exchange between hybridising populations. *Heredity*, 57(3):357–376.

Burri, R., Nater, A., Kawakami, T., Mugal, C. F., Olason, P. I., Smeds, L., Suh, A., Dutoit, L., Bureš, S., Garamszegi, L. Z., Hogner, S., Moreno, J., Qvarnström, A., Ružić, M., Sæther, S.-A., Sætre, G.-P., Török, J., and Ellegren, H. (2015). Linked selection and recombination rate variation drive the evolution of the genomic landscape of differentiation across the speciation continuum of Ficedula flycatchers. *Genome Research*, 25(11):1656–1665.

Coyne, J. A. and Orr, H. A. (2004). *Speciation*. Sinauer Associates, Sunderland, Mass.

Martin, S. H., Dasmahapatra, K. K., Nadeau, N. J., Salazar, C., Walters, J. R., Simpson, F., Blaxter, M., Manica, A., Mallet, J., and Jiggins, C. D. (2013). Genome-wide evidence for speciation with gene flow in Heliconius butterflies. *Genome Research*, 23(11):1817–1828.

Orr, H. A. (1995). The population genetics of speciation: The evolution of hybrid incompatibilities. *Genetics*, 139(4):1805–1813.

Sedghifar, A., Brandvain, Y., and Ralph, P. (2016). Beyond clines: Lineages and haplotype blocks in hybrid zones. *Molecular Ecology*, 25(11):2559–2576.

Selkoe, K. and Toonen, R. (2011). Marine connectivity: A new look at pelagic larval duration and genetic metrics of dispersal.

Wright, S. (1931). Evolution in Mendelian Populations. *Genetics*, 16(2):97–159.

# Comparative population genomics analysis of the Atlantic/Mediterranean suture zone to assess the life-history determinants of speciation in marine fishes

Pierre Barry[1], Christine Arbiol[1], Khalid Belkhir[1], Rémy Dernat[1], *technical MGX staff*[3], *bioinformatics MGX staff*[3], Thomas Broquet[2], Pierre-Alexandre Gagnaire[1]

[1]ISEM, Univ Montpellier, CNRS, EPHE, IRD, Montpellier, France.

[2]CNRS & Sorbonne Université, UMR 7144, Station Biologique de Roscoff, 29680 Roscoff, France.

[3]MGX-Montpellier GenomiX, Univ. Montpellier, CNRS, INSERM, Montpellier, France

# Abstract

The formation of new species involves a wide variety of evolutionary mechanisms ranging from population-level microevolutionary changes to long-term molecular evolution. Understanding how these different mechanisms affect the rise of reproductive isolation barriers and the consecutive build-up of genome-wide divergence is a central objective in speciation research. Besides the well-documented roles of biogeographical history, ecological adaptation and genomic architecture on the unfolding of speciation, the effect of species' life-history traits remains unclear. For example, the rate at which reproductive isolation barriers accumulate and how much they resist gene flow may depend on traits affecting population sizes, dispersal capacities or generation time. A comparative approach of several species pairs with differences in life-history traits but sharing the same biogeographic context and a similar genetic architecture is therefore needed to assess the role of life-history traits on speciation. Through the sequencing of 400 whole genomes, we compared the genomic landscapes of differentiation of 19 marine teleost fish species showing genetic subdivision in the transition zone between the Atlantic Ocean and the Mediterranean Sea. We found a surprisingly high variability of evolutionary histories, associated with a gradient of genetic divergence ranging from near panmixia to divergent geographic lineages that have almost completed speciation. Introgression between Atlantic and Mediterranean populations was found to occur across the whole gradient, although the boundaries between the most divergent lineages remain impermeable to gene flow beyond the region of the suture zone. We found life-history traits to impact all critical components of the evolution of divergence: the amount of ancestral genetic variation, the intensity of gene flow between diverging lineages and even the time during which divergence can build up. Although these effects remain strongly intertwined and difficult to untangle, our findings illustrate the multifarious nature of speciation, and the diverse ways by which life-history traits influence organismal diversification at the molecular level.

**Key words:** speciation, marine fishes, genome divergence, life-history traits, Atlantic-Mediterranean suture zone

# Introduction

Understanding how species arise and persist is one of the key questions in evolutionary biology. Speciation studies have identified a variety of reproductive isolation mechanisms acting throughout the life cycle (Coyne and Orr, 2004). Detailed analyses of the effects of reproductive isolation barriers have revealed moderate to drastic impacts on the reduction of gene flow between incipient species (Sobel and Chen, 2014; Coughlan and Matute, 2020). In parallel, QTL and admixture mapping studies have started to identify barrier loci (Hermann et al., 2013; Arnegard et al., 2014; Powell et al., 2020), and advances in sequencing technologies have allowed an in-depth description of the genomic landscape of divergence between nascent species in various organisms (Turner et al., 2005; Harr, 2006; Hohenlohe et al., 2010; Martin et al., 2013; Renaut et al., 2013). Together, these studies have highlighted the semi-permeable nature of the genome until reproductive isolation is complete (Feder et al., 2012; Harrison and Larson, 2016), with a growing number of barrier loci reducing the effective migration rate of linked genomic regions with increasing strength along the speciation continuum (Barton and Bengtsson, 1986; Stankowski and Ravinet, 2021). Finally, in order to understand which contexts are most propitious to speciation, the ecological conditions, demographic histories and the genetic architectures that have favored the emergence and maintenance of reproductive isolation barriers have been identified in a large number of species (Renaut et al., 2013; Samuk et al., 2017; Van Belleghem et al., 2018; Westram et al., 2018). Although the convergence of these different approaches has provided an increasingly detailed picture of how speciation works, the impact of other determinants of speciation that are more directly linked with species biology and ecology remains elusive.

One aspect of speciation that remains particularly unclear is the role played by species-specific life-history traits in the formation, maintenance and accumulation of reproductive isolation barriers. Many different traits related to abundance, dispersal, fecundity or behavior should have an impact on how evolutionary forces (i.e. mutation, drift, migration and selection) interact in the different mechanisms involved in speciation. For example, propagule size, fecundity and lifespan determine the long-term effective population size ($N_e$) in animals, explaining different levels of genetic diversity and efficiency of purifying selection among species (Romiguier et al., 2014; Chen et al., 2017; Barry et al., 2022). Likewise, traits determining dispersal capabilities such as wingspan in butterflies or pelagic larval duration (PLD) in marine organisms affect migration rate ($m$) and thus shape the extent of spatial genetic differentiation (Selkoe and Toonen, 2011; Dapporto et al., 2019). However, how life-history traits impact speciation will ultimately depend on their effect on the accumulation of genetic barriers to gene flow, with a series of non-trivial predictions of possibly conflicting effects. For example, if speciation occurs mainly through the accumulation of genetic incompatibilities in allopatry (e.g. Bateson-Dohbzansky-Muller incompatibilities), the strength of reproductive isolation built up in a given amount of time (in years) might be higher for low $N_e$ species (Nei, 1975) because of the faster fixation of incompatibilities, as well as for short-lived species that produce more generations per time unit (Orr, 1995; Orr and Turelli, 2001). On the contrary, if speciation occurs through the differential fixation of beneficial mutations by divergent selection (Nosil, 2012; Via and West, 2008), barriers may accumulate faster for high $N_e$ species, which have a higher probability of fixing advantageous mutations for a given intensity of selection (Charlesworth, 2009). Predictions for the effect of traits related to migration rate may also appear contradictory. Increased migration is generally a limiting factor that prevents the establishment of genetic barriers in models of speciation-with-gene-flow (Barton and Bengtsson, 1986; Feder and Nosil, 2010; Yeaman et al., 2016). On the other hand, hybridization can also promote speciation (Mallet, 2007; Schumer et al., 2014; Marques et al., 2019). Finally, behavioral traits such

as mate discrimination or parental care might promote the emergence of prezygotic barriers to gene flow through diverse mechanisms (Coyne, 1974; Servedio, 2004; Verzijden et al., 2012).

Empirical studies aimed at identifying the role of life-history traits in speciation remain scarce. Comparing different divergent lineages of plants, Owens and Rieseberg (2014) found that annual species accumulate post-zygotic barriers more rapidly than perennial species. Considering a higher prevalence of chromosomal rearrangements in annual species, they argued that lower $N_e$ combined with shorter generation time might contribute to increased the probability of fixation of rearrangements that might subsequently act as reproductive isolation barriers. In a meta-analysis in animals, McEntee et al. (2020) found that the width of clines in tension zones is positively correlated with species dispersal capacities, particularly in insects, birds and mammals, reflecting the antagonism between migration and selection predicted by hybrid zone theory (Barton and Hewitt, 1985). Thus, both of these comparative studies emphasized the role of life-history traits that affect $N_e$ and $m$ on the accumulation of reproductive isolation barriers. They were, however, limited by only considering postzygotic barriers through experimental procedures for the former, and not controlling for confounding factors affecting cline width for the latter (such as non-equilibrium demographic histories or differences in biogeographic barriers to gene flow). A more integrative approach capturing the genome-wide effect of barriers to gene flow within an ABC framework accounting for both demography and indirect selection was developed in Roux et al. (2016). However, this comparative analysis, based on phylogenetically distant species from different biogeographic areas, did not detect any significant effect of life-history traits on the probability of ongoing gene flow along a continuum of divergence. Because we predict life-history traits to have complex and potentially conflicting effects, it may be necessary to run a comparative analysis not only across multiple species but also looking simultaneously at multiple life-history traits and their respective effects on the different components of speciation (e.g. genetic diversity, gene flow, semi-permeability). In addition, the greatest power should be achieved by comparing several independent species that are currently subdivided into two (or more) divergent lineages and that share a similar genomic architecture and a common biogeographical context.

The Atlantic Ocean – Mediterranean Sea suture zone is an ideal location for such a comparative analysis. The region surrounding the Strait of Gibraltar and the Alboran Sea has been shown to be a hotspot of genetic differentiation and phylogeographic breaks for various species of marine algae, vertebrates, invertebrates and plants (reviewed in Patarnello et al. (2007). Two explanations have been proposed for this pattern. The first one attributed the multispecies genetic differentiation hotspot to the Almeria-Oran oceanographic Front, which may act as a physical barrier to dispersal that locally restricts larval drift between the two basins (Schunter et al., 2011; Pascual et al., 2017; Carreras et al., 2020). The second hypothesis considers the Atlantic-Mediterranean suture zone as a region where multiple pairs of semi-isolated lineages have come into postglacial secondary contact following allopatric divergence (Lemaire et al., 2005; Bierne et al., 2011). This latter view has been supported by genome-wide studies of divergence between hybridizing lineages showing partial reproductive isolation and semi-permeable barriers to gene flow (Tine et al., 2014; Duranton et al., 2018). As for other multispecies contact zones (Hewitt, 1988, 2000; Johannesson et al., 2020), the Atlantic-Mediterranean suture zone thus offers an interesting study area for comparing histories of divergence and gene flow and the resulting levels of reproductive isolation between species that have evolved in a shared biogeographic context. Here, we performed a comparative population genomics analysis of genome-wide differentiation in 19 marine teleostean fish species with broadly overlapping geographic distributions in the northeastern Atlantic and the Mediterranean Sea. We chose to focus on fish for three reasons. (i) Despite 400 Mya of evolution, the architecture of fish genomes has been remarkably conserved, with similar chromosome numbers (Mank and Avise, 2006), highly conserved synteny (Schartl et al., 2013) and similar broad-scale patterns of recombination rate

variation within chromosomes (Roesti et al., 2013; Tine et al., 2014; Haenel et al., 2018). (ii) Several fish species that were selected for this study are known to be genetically subdivided into an Atlantic and a Mediterranean lineage, with various levels of molecular divergence between lineage pairs across species. (iii) These species display a large variability in life-history traits, including body size, lifespan, fecundity, parental care behavior, or hermaphroditism, which may impact speciation dynamics in different ways. We generated 17 new reference genomes and re-sequenced the whole-genome of 20 individuals per species to disentangle the divergence, gene flow and semi-permeability components of genome-wide differentiation in each species. Finally, we evaluated the links between the inferred evolutionary parameters and the life-history traits thought to have an impact on the course of speciation.

# Material and Methods

## Species selection, sampling and DNA extraction

We selected 19 teleostean marine fish species based on three criteria. Firstly, all species share broadly overlapping distributions across the North-eastern Atlantic and Mediterranean regions. Secondly, previous studies have described some degree of molecular differentiation between Atlantic and Mediterranean populations within each species (sometimes associated with morphological differences) ranging from weak population structure to strong divergence between geographical lineages or ecotypes. Divergent mitochondrial lineages were found to segregate in most species, with varied extent of spatial structure ranging from spatially homogeneously distributed lineages to marked phylogeographical breaks at the Atlantic-Mediterranean suture zone. Finally, the 20 selected species present a wide diversity of life-history strategies (Fig. S4) as revealed by extensive variation in life-history traits including body size, trophic level, lifespan, fecundity, parental care, or pelagic larval duration (PLD). This diversity of life histories is expected to affect demographic and evolutionary parameters such as effective population sizes, migration rates, and the efficacy of selection.

Sampling was performed at the same four locations for all species to standardize geographical distances (Fig. 1A). Two inner sites were sampled within the suture zone (one in the Atlantic Ocean: the Algarve region in Portugal, hereafter referred to as “Atl-in”; one in the Mediterranean Sea: the Costa Calida region around Mar Menor in Spain, “Med-in”), and two outer sites were sampled away from the suture zone (Atlantic: the Bay of Biscay in South-western France or North-western Spain, “Atl-out”; Mediterranean Sea: Gulf of Lion in South of France,“Med-out”). This sampling design specifically aims at distinguishing introgression from incomplete lineage sorting by taking advantage of the higher rates of introgression in the suture zone compared to the remote outer sites. We sampled a fixed number of 5 individuals per site and per species, amounting to 20 individuals for almost all species (Table S3).

Fin clips were collected from live fishes or commercial landings and stored at -20 °C in 90% ethanol until genomic DNA extraction following the Macherey-Nagel NucleoSpin standard protocol with RNase treatment. For 17 species with no reference genome available (see below), fresh gill tissue (or a piece of the body for small species) was collected from an additional individual and placed in a TNES-Urea solution (10 mM Tris-HCl, 120 mM NaCl, 10 mM EDTA Ph 8.0, 05% SDS, 4M urea). After 4 weeks at ambient temperature, the lysis solution was treated with proteinase K (150 g/ml) followed by two phenol-chloroform and two chloroform extractions. High molecular weight genomic DNA (HWM gDNA) was finally precipitated with 2 volumes of ethanol, rinsed in 80% ethanol, dried and resuspended in TE before being treated with the NEBNext FFPE DNA Repair Mix to repair basic DNA damages such as single-strand nicks (see Supplementary Material for further details).

## Sequencing and de-novo assembly of reference genomes

A reference genome was available for 3 out of the 20 species used in this study (*S. pilchardus*: Louro et al. (2019), *D. labrax*: Tine et al. (2014), *A. alosa*: GenBank accession: GCA_017589495.1). We produced a new reference genome for the 17 remaining species using a similar linked-read sequencing strategy (Zheng et al., 2012). Cleaned HMW gDNA were submitted to size selection on the pippin XT (Sage Science) to remove fragments under 40 kb before 10X Genomics library preparation following the Chromium Genome Reagent Kit v2 Protocol. Pooled genome libraries were sequenced to at least 60X per species on one S1 and one S4 lane of an Illumina NovaSeq6000 in 150bp paired-end mode by Genewiz Inc (USA) and the MGX platform (CNRS, France).

The raw demultiplexed reads of each species were deduplicated and processed to retain only reads having a shared barcode (Table S1) with a number of other reads compatible with the experimental design, thus eliminating reads carrying rare barcodes (i.e. suggesting sequencing errors) or over-represented barcodes. De novo genome assembly was finally performed with filtered reads using the `Supernova-2.1.1` software package (Weisenfeld et al., 2017) to generate consensus scaffolds.

Newly assembled reference genomes were evaluated using classical assembly quality metrics and BUSCO analysis to estimate the percentage of genes present, fragmented and missing using 3640 genes from the actinopterygian database (Simão et al., 2015).

## Whole-genome resequencing, mapping and variant calling

Individual whole-genome sequencing (WGS) libraries were prepared following the Illumina TruSeq DNA PCR-Free Protocol and sequenced to an average depth of 20X on an Illumina NovaSeq 6000 platform by Genewiz Inc (USA) and the MGX platform (CNRS, France). Raw reads were preprocessed with fastp v.0.20.0 (Chen et al., 2018) using default parameters (see Barry et al. (2022).

WGS sequence data from the 20 re-sequenced individuals per species were aligned with `bwa-mem` v.0.7.17 (Li and Durbin, 2009) against the corresponding species' reference genome produced either in this (17 species) or previous studies (*S. pilchardus*: Louro et al. (2019), *D. labrax*: (Tine et al., 2014), *A. alosa*: GenBank accession: GCA_017589495.1). We evaluated mapping statistics for each individual with `samtools stats` and `samtools flagstat` v1.10 (Danecek et al., 2021). We then removed PCR duplicates with the `Picard` tools `MarkDuplicates` v.2.23.2 (`https://broadinstitute.github.io/picard/`).

Variant calling was performed separately for each species following the best-practice pipeline in `GATK` v.4.1.6.0 (Poplin et al., 2018). We first ran `HaplotypeCaller` with default options to generate individual GVCF files. Then, we combined the GVCF data of each species into a database with `GenomicsDBImport` and computed one VCF file per scaffold in which all samples have been jointly genotyped with `GenotypeGVCFs`. Finally, we merged all scaffolds' VCF files with `vcf-concat` from `vcftools` v.0.1.17 (Danecek et al., 2011). We followed Fuller et al. (2020) and removed variants located within 5bp around indels with `bcftools filter` v.1.10.2 (Danecek et al., 2021) and then removed indels with `vcftools` v.0.1.17. We evaluated variant calling performance with `RTG Tools` v3.12 ((Cleary et al., 2015), `bcftools stats` v.1.10.2, `vcftools` v.0.1.17 and `scikit-allel` v1.2.1 (Miles et al., 2020).

## Population differentiation and divergence statistics

We estimated pairwise genome-wide $F_{ST}$ between each of the four sampling sites by computing Hudson's (Hudson et al., 1992; Bhatia et al., 2013) and Weir and Cockerham's (Weir and Cockerham, 1984) estimators and their respective standard errors, using the block-jackknife method in `scikit-allel` v1.2.1 (Miles et al., 2020).

We estimated the per-site nucleotide diversity ($\pi$) following Nei and Li (1979) equation:

$$\pi = \frac{\sum_{i<j} k_{ij}}{\binom{n}{2}} \tag{1}$$

where $k_{ij}$ equals 1 in the presence of a mismatch or 0 otherwise, and $n$ is the total number of sampled chromosomes.

Likewise, we estimated pairwise absolute divergence ($d_{XY}$) using Nei and Li (1979):

$$d_{XY} = \frac{1}{n_X}\frac{1}{n_Y}\sum_{i=1}^{n_X}\sum_{j=1}^{n_Y} k_{ij} \tag{2}$$

where $n_X$ and $n_Y$ represent the total number of sampled chromosomes in populations X and Y, respectively. Net divergence $d_a$ was estimated following:

$$d_a = d_{XY} - \frac{\pi_X + \pi_Y}{2} \tag{3}$$

For these three statistics, we used a custom-script and `scikit-allel` v1.2.1 (Miles et al., 2020) with VCF files containing both variable and invariant sites to avoid underestimation due to the presence of missing sites in the reference genomes or missing individual genotypes (Korunes and Samuk (2021) see Supplementary Material for further information).

In addition to genome-wide estimates, non-overlapping sliding window averages were calculated over variable and invariant sites for each statistic using a 50kbp window size.

## Multivariate analysis of population structure and divergence

We conducted principal component analysis (PCA) on the variant dataset of each species with `SNPRelate` v.1.20.1 (Zheng et al., 2012). To evaluate the effect of linked variants on population structure, we compared PCA performed with either the whole SNP dataset or a pruned subdataset obtained by discarding variants with linkage disequilibrium, estimated as a composite measure, above 0.2 with the `snpgdsLDpruning` function. Each PCA was also repeated with filters controlling for varied minor allele frequency thresholds, ranging from 0 to 0.5 by steps of 0.05.

The diversity of spatial patterns and extent of genetic structure among species was illustrated with a multidimensional scaling (MDS) analysis of species population trees reconstructed from pairwise $F_{ST}$ matrices, using the `R` package `treespace` (Jombart et al., 2017) with the Kuhner-Felsenstein (KF) metric accounting for branch distances.

## Detection of admixture and introgression

We searched for genome-wide signals of admixture and introgression using several tests that distinguish gene flow from incomplete lineage sorting (ILS). The $f_3$ statistics measures the correlation of allele frequency differences between a focal population and two reference populations (Reich et al., 2009; Patterson et al., 2012). Specifically, a negative $f_3$ reveals that the focal population is the result of admixture between the two reference populations. We estimated genome-wide $f_3$ with `scikit-allel` v1.2.1 (Miles et al., 2020) for each of the

12 possible topologies that can be built from our set of four populations (see Supplementary Material).

Then, we used the $D$ statistics (Green et al., 2010; Durand et al., 2011) to detect genowe-wide signals of introgression between each population and the two populations from the opposite side of the suture zone. (e.g., differential spatial introgression from $P_2$=Atl-in (Algarve) to $P_3$=Med-out (Gulf of Lion) vs $P_2$ to $P_4$=Med-in (Costa Calida) , using $P_1$=Atl-out (Bay of Biscay) as a reference). Similarly, we inferred $D$ statistics for all possible topologies with `D-suite` (Malinsky et al., 2021) for 10 species with phased VCF (see below).

For both $f_3$ and $D$ statistics, we estimated standard error and $Z$ score using the block-jackknife method with a block size of 10 000bp, to assess deviation from zero when $|Z| > 3$ (Patterson et al., 2012). The $D$ statistic is less sensitive to divergence time than $f_3$, because no matter how much differences are fixed between ($P_1$, $P_2$) and ($P_3$, $P_4$), only the sites that show ABBA and BABA genealogical patterns are considered to distinguish ILS from differential introgression. In species with a high divergence between Atlantic and Mediterranean lineages, $D$ is thus expected to be more sensitive to introgression compared to $f_3$.

## Estimation of divergence times from inferred gene genealogies

We inferred genome-wide genealogies of non-recombining genome segments using tsinfer (Kelleher et al., 2019) and estimated divergence times with tsdate (Wohns et al., 2021). Since this approach requires the reconstruction of phased haplotypes and the identification of ancestral allelic states, it was only applied to the 10 species (Table S2) for which we had good-quality reference genomes (because genome fragmentation negatively impacts phasing performance), and available reference assemblies for three outgroup species (needed for variant orientation).

For each of these ten species, variable sites mapping to the same scaffold were phased to reconstruct maternal and paternal haplotypes following a two-step approach. First, we used a read-aware physical phasing method implemented in WhatsHap (Martin et al., 2016) for prephasing neighboring heterozygous positions within individuals based on the information contained in paired-end reads. We evaluated the number of prephased SNPs and the length distribution of phased blocks, retaining only scaffolds longer than 10kb, which represented from 67% (*S. pilchardus*) to 97% (*H. guttulatus*) of the total reference genome size. Then, we used population-level linkage disequilibrium information at the species scale to complete the phasing of prephased blocks scaffold-wide using `SHAPEit4` (Delaneau et al., 2019).

We determined the most likely ancestral state at each variable position using the likelihood method implemented in `est-sfs` (Keightley and Jackson, 2018), which uses phylogenetic information from up to three outgroup species in addition to allele frequencies in the ingroup. For each species, we searched databases for three high-quality reference genomes from fish species phylogenetically close to the ingroup (Table S4). We identified orthologous genomic regions between ingroup and outgroup species around each variable site (i.e. 200 bp centered on each SNP present in the ingroup VCF) using blast search. Allele counts within the ingroup were directly retrieved from the VCF, combined to the outgroups allelic states and passed to `est-sfs`.

We inferred gene genealogies from phased and oriented VCF with `tsinfer` (Kelleher et al., 2019), with a recombination rate constant of $1e^{-8}$ per base pair per generation and a mismatch ratio of 1. Then, we estimated branch lengths and node ages of inferred genealogies with tsdate (Wohns et al., 2021), using a mutation rate constant equal to $1e^{-8}$ per base pair per generation and the species' effective population size estimated from the data ($N_e = \pi/(4\mu)$). Inferred and dated tree sequences from our 10 empirical datasets were finally used to generate within- and between-basin distributions of TMRCA using `tskit` (`https://tskit.dev/tskit`). For each species, inferred coalescence times in units of generations were converted into years using the

species-specific generation times calculated from life tables in (Barry et al., 2022). We compared each inferred distribution to the expected exponential distribution of coalescence times under the standard (i.e. single population) coalescent model following (Brandt et al., 2022), using the estimated species' effective population size as a scaling factor.

Coalescent simulations were performed with `msprime` (Kelleher et al., 2016) to evaluate whether coalescence times contained in tree sequences (i.e. TMRCA between all pairs of haplotypes within genealogies) could be used for dating divergence under varied demographic scenarios. We used `stdpopsim` (Adrion et al., 2020) to set up five generic demographic divergence models (`https://github.com/PA-GAGNAIRE/stdpopsim/blob/main/stdpopsim/models.py`): Strict Isolation (SI), Isolation-with-Migration (IM), Secondary Contact (SC), Migration Pulse (MP) and Ancient Introgression (AI). All effective population sizes were set to 10 000, migration rates equaled 0.0001 in both directions in IM and SC models, the fraction of population 2 coming from population 1 or the ancient donor lineage was set to 0.05 in the MP and AI models, respectively. Secondary gene flow or migration pulse occurred 1 000 generations before present in SC, MP and AI models, and divergence time with the donor lineage occurred 200 000 generations ago in the AI model. In order to explore the influence of divergence time between population 1 and 2 on the distribution of inferred TMRCA, splitting times ranging from $0.5N_e$ to $10N_e$ generations were used to run different simulations under each model, assuming one chromosome of size 20Mb and uniform mutation and recombination rates of $10e^{-8}$ per site per generation. After the completion of simulations, 20 chromosomes were sampled from each population and their tree sequence was outputted. We used `tskit` to examine the distribution of TMRCA between haplotype pairs both within populations 1 and 2, as well as between populations.

To evaluate the performance of `tsinfer` and `tsdate` in inferring coalescence times, we simulated 1Mb scaffolds under a SC model whereby an ancestral population of size 20 000 split 100 000 generations ago in two populations of size 10 000, which diverged in complete isolation before coming back into contact 1 000 generations ago, with a migration rate of $1e^{-3}$ individuals per generation. Recombination and migration rates were set to $1e^{-8}$ per base pair per generation. We created sample polymorphism data (stored in VCF files) from simulated individuals and used it as an input to `tsinfer` and `tsdate` before inferring TMRCA distributions within and between populations from inferred genealogies and comparing results with those obtained from simulated tree sequences in `msprime`. We ran simulations using 3 numbers of sampled haplotypes per population (20, 200, 500) and performed inferences with 3 different mismatch ratios (0.1, 1, 10) and two different prior distributions to approximate the coalescent distributions times (gamma and lognormal distributions).

## ABC inference of demographic history

We used `DILS` (Demographic inference with linked selection) to infer the most likely demographic history of isolation/contact between Atlantic and Mediterranean populations of each species, accounting for temporal variation in migration rates and heterogeneous effective population size and migration rates among genes (Fraïsse et al., 2020). In order to run these analyses using a dataset that is comparable across all species, we extracted individual FASTA-formated coding sequences using genome annotations from the `BUSCO` actinopterygian database. As `DILS` does not rely on any population genetic statistics requiring phase information, individual haplotypes were simply generated by randomly assigning alleles at heterozygous genotypes to any of the two haplotypes, as recommended. `DILS` is designed to infer the history of a pair of populations (2-population model).

For each species, we used two sub-datasets to infer two demographic histories: one between

the two outer "reference" populations (Gulf of Lion for Mediterranean Sea and Bay of Biscay for the Atlantic Ocean) and one between the two inner populations from the admixture zone (Costa Calida for the Mediterranean Sea and Algarve for the Atlantic Ocean). In this way, we could assess differential introgression between the two population pairs (Atl-out ⇔ Med-out versus Atl-in ⇔ Med-in) while avoiding mixing samples from distinct locations in the same population (i.e. we refrained to mix in and out populations within each basin to run a unique Atl ⇔ Med analysis). For each inference, we excluded individual gene sequences that contained more than 10% of missing data and genes having less than 30 nucleotides. We then excluded genes that contained less than 5 haplotype sequences per population (out of five individuals sampled per population). We set the mutation rate ($\mu$) to $3e^{-9}$ per base pair per generation and the within-gene recombination rate ($r$) to $3e^{-10}$ as default parameters in `DILS`. We did not use precise biological estimates for mutation and recombination rates here since these values are not available for most of the studied species, mutation rate only acts as a scalar, and no haplotype-based statistics are used in the inference. We set uniform priors for the time of split (between 100 and 1 000 000 generations), effective population size (100 to 1 000 000 individuals), and migration rate parameters (0.4 to 20 migrants per generation).

For each species and each pair of populations, `DILS` was used to determine the probability of ongoing gene flow and estimate the following model parameters: effective population size of the Mediterranean Sea population ($N_1$), Atlantic Ocean population ($N_2$) and ancestral population ($N_a$); the time of split between the two populations ($T_{split}$), the time of secondary contact when SC was the best model ($T_{sc}$), the time of ancient migration when AM was the best model ($T_{am}$); the number of migrants $M_{ij}$ (expressed in units of $4N_i m_{ij}$ where $m_{ij}$ is immigration rate into population $j$ from population $i$) from the Mediterranean Sea to the Atlantic Ocean ($M_{12}$) and from the Atlantic Ocean to Mediterranean Sea ($M_{21}$). If heterogeneous effective population size and/or heterogeneous effective migration rates among loci were included in the best model, we estimated the two parameters of the beta distribution controlling for variation of either $N_e$ and/or $M_{12}$ and $M_{21}$ between different genes. For each parameter, we set the estimate and the standard deviation as the mean and the 95% credible interval of the posterior distribution. From these estimates, we measured the ratio between contemporary ($N_1$ and $N_2$) and ancestral population sizes ($N_a$), and the ratio between either $T_{am}$ or $T_{sc}$ and $T_{split}$, which do not depend on the mutation rate parameter. As defined in `DILS`, we estimated $m_{12} = M_{12}/(4N_2)$, with $m_{12}$ defined as the proportion of migrants from $N_1$ that contributes to $N_2$ per generation.

## Evaluation of semi-permeability and reproductive isolation

Heterogeneous gene flow across the genome is an important hallmark of ongoing (or incomplete) speciation, which can be used to characterize the extent of reproductive isolation between two taxa. In addition to the distribution of $F_{ST}$ estimated from 50kb genomics windows and ABC inference, which will both provide some insight on semi-permeability, we used two other empirical approaches to evaluate the degree of semi-permeability between Atlantic and Mediterranean populations in each species.

Our first approach was based on the rationale that in the presence of genetic barriers to gene flow, the level of genetic differentiation measured by $F_{ST}$ in a given genomic region reflects a balance between the homogenizing effect of gene flow, the direct or indirect effect (i.e. through linkage with a nearby selected locus) of selection against introgressed ancestry, and genetic drift. Assuming neutrality and a similar drift intensity across the four sampled populations, the difference between $F_{ST,out}$ (i.e. $F_{ST}$ between Gulf of Lion and Bay of Biscay) and $F_{ST,in}$ (i.e. $F_{ST}$ between Costa Calida and Algarve) for a given genomic region should mostly reflect a higher rate of gene flow in the contact zone as compared to between remote reference popu-

lations. Therefore, increased gene flow between inner populations vs outer populations should cause a relative decay in genetic differentiation shared by all loci. By contrast, a genomic region influenced by a barrier to gene flow should exhibit a smaller relative decay in genetic differentiation, because selection against introgressed ancestry maintains genetic differentiation at more similar levels between outer and inner pairs (i.e. such a locus my resist introgression and thus have an $F_{ST}$ in the suture zone that is not too different from the $F_{ST}$ between outer populations). Semi-permeability can thus be revealed by a strong variance in the relative decay of genetic differentiation among loci (i.e. differential introgression). At one end of the speciation continuum, neutral genetic differentiation between "simple" populations (no reproductive isolation at all) would be equally eroded by increased gene flow in all genomic regions (i.e. no differential introgression would be visible). At the other end of the continuum, strong reproductive isolation would be characterized by a large fraction of the genome maintaining equally strong genetic divergence in both the outer and inner pairs (again, no differential introgression). The degree of semi-permeability was thus captured by the genome-wide distribution of the following ratio:

$$\Delta_{Fst} = \frac{F_{ST,out} - F_{ST,in}}{F_{ST,out}} \quad (4)$$

measured in non-overlapping 50kb windows. Regions located within 1kb of scaffold extremities were removed to avoid possible artifacts due to reference genome fragmentation. This ratio aims to standardize the decay in $F_{ST}$ between out and in populations in an effort to capture the variance in decay across loci, that is, semi-permeability.

Our second approach relied on quantifying variation in introgression rates across genomic windows for each species using $f_d$ (Martin et al., 2015). Our sampling design enabled us to capture the magnitude of the difference in introgression rates between two populations located in the same basin (e.g. excess in Mediterranean introgression in Atl-in with reference to Atl-out). Therefore, these relative estimations should be considered as minimum values, which may appear underestimated (see results), especially for species with high gene flow between populations within basins. Variation in $f_d$ among 10 000 SNPs genomic regions were used to characterize heterogeneous gene flow among loci, a measure reflecting the degree of semi-permeability and reproductive isolation. Since the calculation of $f_d$ requires oriented SNP data, it was only applied to the 10 species used for inferring genome-wide gene genealogies.

## Genomic architecture and phylogenetic controls

Before exploring biological and ecological drivers of divergence and reproductive isolation, we tested the influence of parameters pertaining to intrinsic evolutionary constraints imposed by the genome architecture and phylogeny of the studied species.

We used `BUSCO` genes from the actinopterygian database to build a high-confidence ortholog genes dataset to make phylogenetic reconstruction. We inferred the phylogenetic relationships among the 19 studied species using `RAxML` v8.2.12 (Stamatakis, 2014) with a subset of 87 single-copy orthologs, including *Lepisosteus oculatus* as an outgroup species (Braasch et al., 2016). Ortholog genes were first individually aligned with `mafft` (Katoh et al., 2019) and then concatenated to generate a multi-species alignment. Phylogenetic tree inference was made with `iqtree` v2.1.2 (Minh et al., 2020) under the best-fit protein evolution model (JTT+F+R4), using 1000 bootstrap iterations. We used `ggtree` v2.0.4 (Yu, 2020) utilities for visualizing the inferred phylogenetic tree.

Then, for each possible species pair among the 19 species, we built a larger high-confidence ortholog genes dataset to assess the between-species correlation in genetic differentiation ($F_{ST}$),

and absolute divergence ($d_{XY}$) across genes. Our rationale was to use high confidence orthologs as markers in the genome to assess the strength of correlation between the genomic landscapes (of $F_{ST}$ and $d_{XY}$) of species with varying degrees of divergence, without the need to perform whole-genome alignments. For each pair of species and each statistic, we evaluated the correlation between landscapes using linear regression. The coefficient of determination ($R^2$) in each species pair was finally represented as a function of the phylogenetic distance between species to evaluate the temporal decay in correlated landscapes of genetic differentiation and gene evolutionary constraints across the phylogeny.

In addition, for each species, we measured $\pi$, $F_{ST}$, $d_{XY}$ and $f_3$ statistics for each `BUSCO` gene and compared these values to those of 10kb genomic windows that do not contain `BUSCO` genes. With this comparison, we tested whether diversity, differentiation, divergence and introgression take different values between a subset of highly constrained coding gene sequences and the remainder of the genome. Distributions were compared with Student's t-Test and the difference between the means of the two distributions was calculated.

## Identification of life-history correlates of divergence and speciation

We collected eight simple quantitative traits (body size, trophic level, fecundity, propagule size, age at maturity, lifespan, adult lifespan, and pelagic larval duration) and two qualitative variables (presence/absence of hermaphroditism and brooding behavior) describing various aspects of the biology and ecology of the 19 species (Fig. S4). These life-history traits are supposed to affect divergence and speciation dynamics through their impact on population evolutionary parameters including drift intensity, migration rates, and the efficiency of selection. Detailed information on data collection is available in Barry et al. (2022) for all traits except for pelagic larval duration (PLD). We used PLD estimates from Macpherson and Raventos (2006) for 10 species, and species-specific references for the other 10 species (see Supplementary Material for further details). The correlation between life-history traits is shown in chapter 1 (supplementary figure S7).

We tested for phylogenetic signals in life-history traits and all genetic statistics using Pagel's $\lambda$ (Pagel, 1999)) and Blomberg 's K (Blomberg et al., 2003) in `phytools` v.0.7.90 (Revell, 2012). Pagel's $\lambda$ tests whether a variable can be explained only by branch length of a given phylogeny (corresponding to the null hypothesis: $\lambda = 1.0$). Similarly, Blomberg's K evaluates variation of the trait variable between and within clades: $K$ values less than or greater than 1 means that neighboring tips are less or more similar to each other, respectively, than expected if branch length explains the variation of the trait. In summary, to exclude phylogenetic signals that could explain the correlation between any life-history traits and genetic statistics, we checked that there was no covariation between phylogeny and any of the trait and genetic covariables of interest (hence $\lambda < 1.0$ and $K =< 1$).

Statistical associations between life-history traits and genetic data were first explored with a redundancy analysis (RDA) using the `R` package `vegan` (Dixon, 2003) to summarize multiple linear regression between response variables (genetic data) and explanatory variables (life-history traits). We used empirical genetic summary statistics including genetic diversity ($\pi$), divergence ($F_{ST}$, $d_{XY}$, $d_a$) and gene flow ($f_3$) obtained for each species to evaluate the extent to which life-history trait variables influence species divergence patterns. We coded each qualitative variable (hermaphroditism, parental care and presence of genetic admixture) as a binary numeric variable (1 = Yes, 0 = No) and processed a $z$-transform on all explanatory variables. The significance of all constraints taken simultaneously and the marginal effect of each constraint was tested with an ANOVA using 10 000 permutations. Finally, significant terms were individually submitted to a conditional RDA by partialling out the other constraining variables

from the analysis.

## Statistical analyses and graphical illustration

All statistical analyses and figure visualization were carried out using `R`-3.6.1 (R Core Team, 2018) and `python`-3.7.6 (Rossum and Drake, 2009). We followed Muff et al. (2021) and addressed an equivalent level of evidence for each statistical test using $p - values$. All plots and graphical illustrations were created using `tidyverse` v.1.3.0 (Wickham et al., 2019) and `ggplot2` v.3.3.2 `R` packages (Wickham, 2016). All scripts and interactive notebooks representing data and intermediate results are freely available at `https://github.com/pierrebarry/speciation_life_history_traits_comparative_genomics_fish`.

# Results

## De novo genome assembly and whole-genome re-sequencing in 19 species

We sequenced and de-novo assembled new reference genomes for 17 marine fish species (Table S2, Fig. S5). Megabase-scale scaffolds were reconstructed in a majority of species, with an across-species mean scaffold N50 of 2Mb. Although assembly contiguity was variable among species, the fraction of total genome size spanned by assembled scaffolds longer than 10kb averaged 85.4% in our 15 best assemblies. More fragmented and partial genome assemblies were obtained for one species (*G. niger*). Overall, the 19 reference genomes used in this study (i.e. including three taken from genome databases) showed relatively good BUSCO quality scores (Fig. S6), with a mean level of ortholog completeness of 81.4% (94.9% of which being single-copy), a mean fragmentation level of 5.4%, and mean missingness of 13.2%. Therefore, these reference genomes were suitable to document genome-wide distributions of genetic diversity and divergence in the 20 studied species, as confirmed by the mapping statistics of whole-genome re-sequencing data (Fig. S7). We re-sequenced the whole genome of 400 individuals equally distributed among the 19 species. Following read quality filtering, we obtained a mean amount of 19.3 Gigabases of sequences per individual (sd=3.7Gb), with a mean Q30 rate of 92.3% (sd=1.1%) and a moderate average duplication rate of 13% (sd=5.5%). The fraction of individual samples' reads mapped to their corresponding reference genome ranged between 82% and 99% (overall mean=95.9%, sd=5.4%), reflecting the relatively high completeness of the reference genomes. No effect of species' mean scaffold length was found on the percentage of mapped reads ($t = 0.729$, $p - value = 0.476$), as for the total number of scaffolds in assemblies ($t = -2.013$, $p - value = 0.0602$). The mean individual coverage depth per species was generally close to 20X or larger (up to 50X in *S. typhle*).

After variant calling, we generated 20 species-specific genome-wide SNP datasets stored in VCF format (Fig. S2), which contained between $2.56e^6$ (*H. guttulatus*) and $47.52e^6$ SNPs (*A. boyeri*) passing filters (mean=$20.13e^6$ , sd=$11.72e^6$). Again, the number of retained SNPs was not correlated to the mean scaffold length ($t = -1.016; p - value = 0.325$). Ten species' SNP datasets were fully oriented using outgroup information and phased at the scaffold level.

## Gradient of genetic diversity, differentiation and divergence

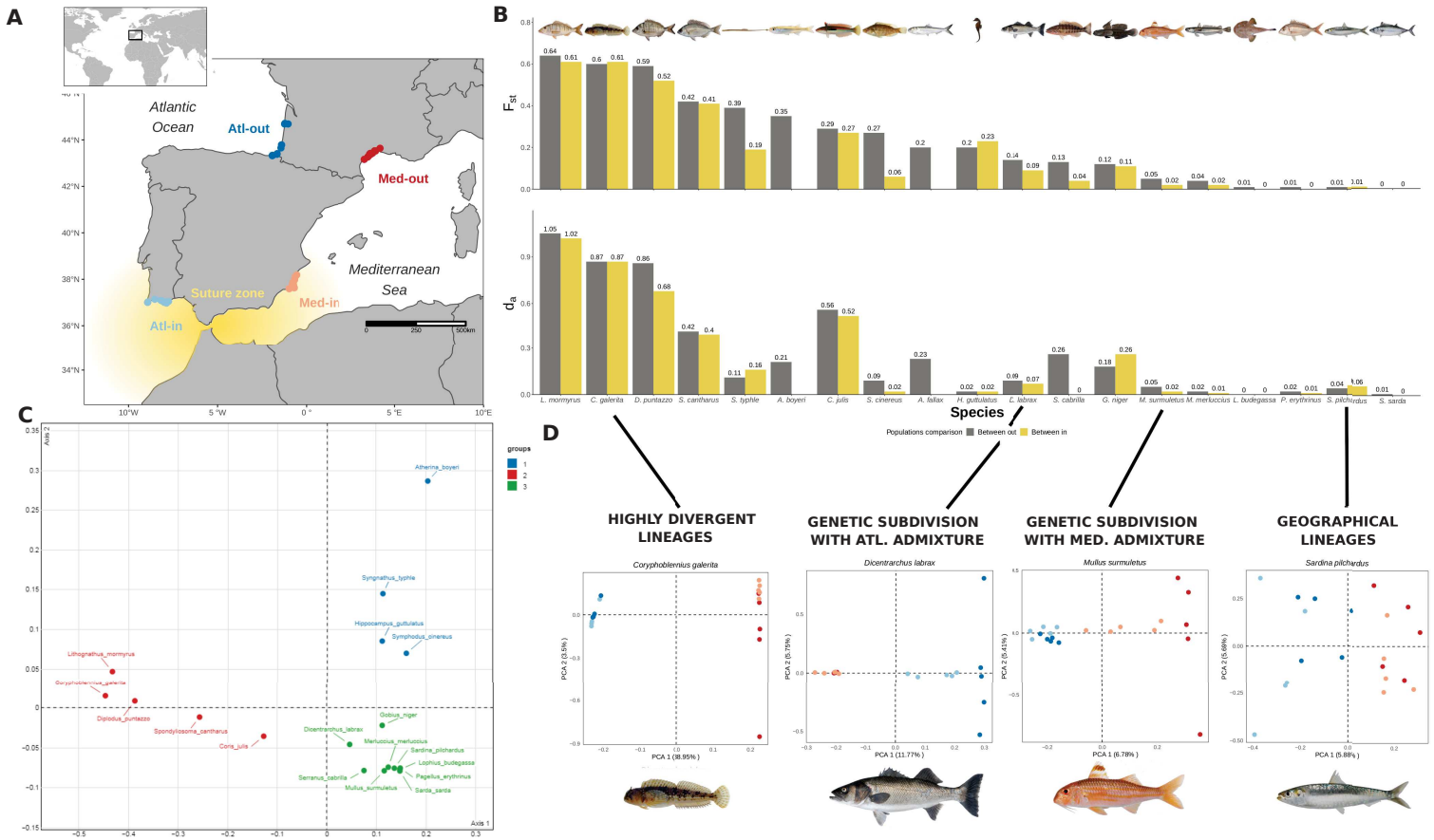



Figure 1: **Gradient of genetic differentiation and patterns of spatial population structure in 19 teleostean marine fish species across the Atlantic Ocean - Mediterranean Sea suture zone**. Top left: Sampling map of the 20 individuals for each of 19 fish species equally sampled from four different localities outside the suture zone (Atl-out, dark blue and Med-out, dark red) and within the suture zone (Atl-in, light blue and Med-in, light red). Top right: Gradient of genetic differentiation ($F_{ST}$) and net divergence ($d_a$) between outer and inner populations (shown as gray and orange bars, respectively), ordered from higher to lower genetic differentiation in outer pairs from left to right. Bottom left: MDS analysis of pairwise population differentiation across 19 species, showing a gradient in the level of genetic differentiation between basins on the first axis. Hierarchical clustering separates five strongly differentiated species (in red) from ten species showing weaker differentiation between Atlantic and Mediterranean locations, all of which are weakly or not genetically differentiated within basins. MDS axis 2 captured a third group of four species (in blue) showing more complex population structures that include an ecotypic differentiation component. Bottom right: Four examples of within-species population structure assessed with a PCA analysis of whole-genome polymorphism data. Each point represents an individual with its color indicating sampling location on the map. From left to right: strong differentiation between phylogeographical lineages of Montagu's blenny (C galerita), population subdivision with admixture in the Atlantic Ocean or the Mediterranean Sea in the European sea bass (D. labrax) and the striped red mullet (M. surmuletus), and spatial genetic differentiation in the European pilchard (S. pilchardus).

Our comparative analysis of genomic variation revealed a 6 fold variation in mean within-population genetic diversity across species, with within-populations nucleotide diversity ($\pi_{pop}$) ranging from 0.24% for *L. budegassa* to 1.54% for *A. boyeri*, consistent with previous reference-free estimates (Barry et al., 2022). A 10 fold variation was found in total species genetic diversity ($\pi_{tot}$) ranging from 0.24% for L. budegassa to 2.41% for *A. boyeri*, reflecting variable contributions of between-population diversity to $\pi_{pop}$ across species. Accordingly, the genome-wide average differentiation calculated as $F_{ST}$ between Atlantic and Mediterranean outer populations revealed a wide gradient (Fig 1B, top) extending from nearly zero for 4 species (*L. budegassa*, *S. pilchardus*, *S. sarda* and *P. erythrinus*), to weak (e.g., 0.037 for *M. merluccius* and 0.051 for *M. surmuletus*), moderate (e.g., 0.144 for *D. labrax* and 0.204 for *H. guttulatus*) and strong genetic differentiation (e.g., 0.601 for *C. galerita* and 0.640 for *L. mormyrus*). These marked differences in within-species subdivision were not related to phylogenetic distance among species pairs (Pagel's $\lambda = 6.61e-5$ ; $p-value = 1$ - Blomberg's $K = 0.519$ ; $p-value = 0.1361$). Similarly, we observed a broad gradient in both absolute ($d_{XY}$, Fig S9A) and net divergence ($d_a$, Fig. 1B, bottom) across species, with the strongest levels of molecular divergence between Atlantic and Mediterranean populations exceeding one percent (1.04 for *L. mormyrus* and 1.01 for *D. puntazzo*). No correlation was found between $F_{ST}$ and $d_{XY}$ ($t = 1.677$ ; $p-value = 0.112$, Fig. S9A) nor between $d_a$ and $d_{XY}$ ($t = 2.063, p-value = 0.0547$, Fig. S9B), but $F_{ST}$ was very strongly positively correlated to $d_a$ ($t = 7.567$ ; $p-value = 7.72e-7$, Fig. S9C). Therefore, the most strongly genetically differentiated species between Atlantic Ocean and Mediterranean Sea populations also displayed the strongest levels of net molecular divergence between basins. Again, we detected no phylogenetic signal for between-species variation in $d_{XY}$ (Pagel's $\lambda = 6.61e-5$ ; $p-value = 1$ - Blomberg's $K = 0.390$ ; $p-value = 0.3723$) and $d_a$ (Pagel's $\lambda = 6.61e-5$ ; $p-value = 1$ - Blomberg's $K = 0.380$ ; $p-value = 0.4077$).

The comparison of within-species population trees based on pairwise $F_{ST}$ between sampling locations revealed two main axes of variation in spatial population structure across the 19 species. The principal source of variance was a broad gradient in the strength of differentiation between Atlantic and Mediterranean populations along MDS axis 1 (Fig. 1C). Fifteen species were distributed along this gradient, defining two groups of species characterized by strong and weak inter-basin differentiation, respectively, but always showing weaker or non-significant genetic structure within basins. A third group of 4 species is characterized by spatial structure at both the between- and within-basin scales was captured by MDS axis 2 (Fig. 1C). Interestingly, these four species occupy both coastal marine and lagoonal/estuarine habitats and showed genetic differentiation between groups of individuals sampled in different habitats, sometimes even within the same location. Accordingly, species PCA based on individual genomes revealed varied contributions of geography and habitat type to genetic variation in this group (see below and Fig. S3), reflecting previous findings of morphological and/or genetic differentiation between marine and lagoonal ecotypes in some of these species (Riquet et al., 2019).

This coarse classification into three clusters of species distributed along two gradients of spatial structure was refined by the species' PCAs based on individual genomes, which suggested the presence of admixed individuals or populations in some species, either in the Mediterranean Sea or the Atlantic Ocean (Fig 1D, Fig. S3, Fig. S8). For all species, except for *A. boyeri*, the first component of the PCA analysis separated Atlantic and Mediterranean samples but with various proportions of variance explained, ranging from 6.25% for *L. budegassa* to 50.84% for *L. mormyrus*. PCA plots showed 5 broad different types of population structures: (i) Nearly absent (i.e. single species without almost null differentiation between populations, e.g. *S. sarda*), or weak geographic structure that might correspond to isolation-by-distance (e.g. for *S. pilchardus* and *P. erythrinus*). (ii) Genetic subdivision with admixture within the suture zone reflected by inter-individuals variation in foreign ancestry (either in the Mediterranean Sea: *M.*

*surmuletus*, or in the Atlantic Ocean: *D. labrax*). (iii) Genetically differentiated populations with the presence of migrant individuals of foreign ancestry in one population (*S. cabrilla* and *L. budegassa*) or hybrids (*M. merluccius*) suggesting ongoing contemporary migration and hybridization. (iv) Complex population structure with marine/lagoon and strong within-basin components: *A. boyeri*, *H. guttulatus*, *S. cinereus*, *S. typhle*. (v) Divergent phylogeographical lineages showing strong between but no within-basin differentiation and no or little evidence for introgression (*C. galerita*, *L. mormyrus*, *D. puntazzo*, *S. cantharus*, *C. julis*).

## ABC Inference of the demographic divergence history

We inferred the demographic history of divergence in each species using an ABC method applied both between the two outer reference populations (Gulf of Lion vs Bay of Biscay) and the two inner populations within the suture zone (Costa Calida vs Algarve). These analyses revealed a wide diversity of gene flow histories across species, although they generally lacked the power to clearly distinguish between Isolation-with-Migration and Secondary Contact scenarios (Fig. S24). Most population pairs showed evidence for ongoing gene flow between Atlantic and Mediterranean populations, especially within the suture zone (Fig. 2C). Inferred gene flow ($4N_em$, in absolute number of migrants) in these pairs ranged from 0.42 in *A. boyeri* from Gulf of Lion to Bay of Biscay, to 26.72 for *H. guttulatus* from Algarve to Costa Calida (Fig. S24). Whether there was or not contemporary gene flow, the time of split estimated between Atlantic and Mediterranean populations ranged from 40 860 generations for *L. budegassa* to 1,569,728 generations for *A. boyeri*, assuming an equal mutation rate of $10e^{-8}$ per site per generation for all species. Using an estimate of the generation time based on the life table characteristics of each species (Table S3), the time of split estimated from ABC ranged from 672 703 for *S. pilchardus* years to 9 013 196 years for *L. mormyrus*.

Heterogeneous effective population sizes were inferred among loci for all species and all population comparisons, except for *H. guttulatus*, suggesting that linked selection significantly impacts variation in the rate of polymorphism and lineage sorting across species' genomes. Likewise, heterogeneous effective migrations rates indicating semi-permeability to gene flow were inferred, but only for 11 species. Since these ABC inferences do not provide accurate estimates of demographic and selective model parameters, the analyses in the following sections present more dedicated approaches to measuring the impact of gene flow, divergence times, degrees of semi-permeability and genomic constraints on genomic differentiation landscapes between population pairs in each species.

## Widespread admixture along the divergence gradient

We found evidence for almost ubiquitous gene flow between Atlantic and Mediterranean basins across the divergence continuum. This was captured by leveraging the effect of an increased potential for gene flow between populations located within as compared to outside of the transition zone. The presence of introgression between populations of the two basins should be somehow related to the level of genome-wide genetic divergence because gene flow might be impeded between highly divergent genomes if they have accumulated barriers to reproductive isolation. We aimed to test this prediction with various approaches.

First, increased introgression within the transition zone might be reflected by lower genetic differentiation and divergence within the suture zone (Atl-in / Med-in) as compared to outside of the transition zone (Atl-out / Med-out) (Grant et al., 2005; Martin et al., 2013). Among the 17 species in which such a comparison could be made (i.e. removing *A. fallax* with no samples in

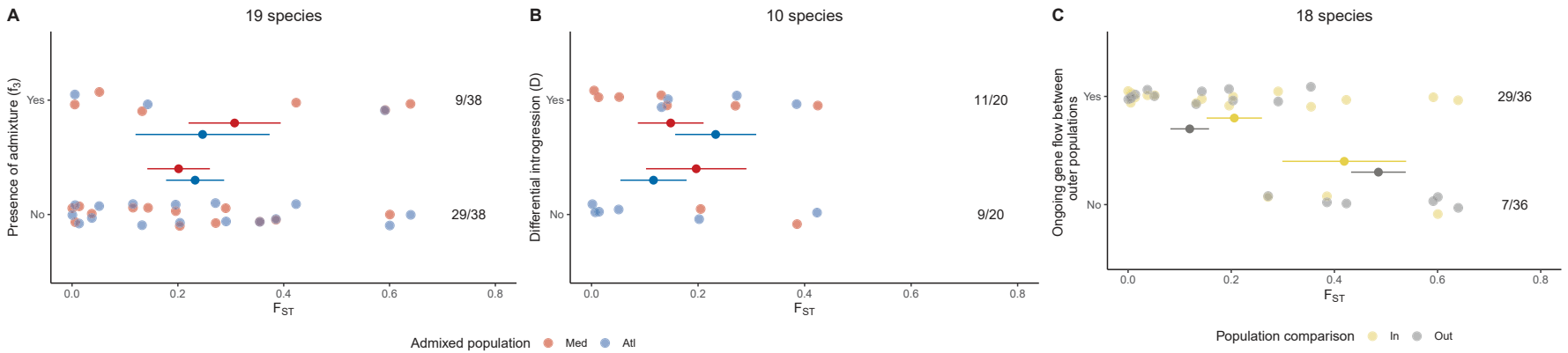



Figure 2: **Admixture, introgression and ongoing gene flow along the gradient of differentiation**. Panel A: $f_3$ test showing that several inner populations (either Med-in in violet or Atl-in in blue) result from genetic admixture between the outer populations of the two basins. Panel B: Results of the $D$ statistics showing differential introgression between populations of the same basin (either Med-in violet or Atl-in blue), originating from a population of the other basin. Panel C: Results of ABC inference of contemporary gene flow between either inner (yellow) or outer (grey) populations. In each panel, gene flow results are plotted against the gradient of differentiation measured by $F_{ST}$ between the two outer populations. The mean and standard deviation of Fst across species is reported for each gene flow outcome, as well as the number of observations for each outcome over the total number of tests performed. Gene flow between outer populations is only tested in panel C (grey points), showing a limit to gene flow beyond a certain level of differentiation.

Med-in) and showing significant Atl-out / Med-out differentiation (i.e. removing *S. sarda*), 12 species showed reduced genetic differentiation in the suture zone (*L. mormyrus*, *D. puntazzo*, *S. cantharus*, *S.typhle*, *C.julis*, *S. cinereus*, *D. labrax*, S. cabrilla, *M. surmuletus*, *M. merluccius*, *P. erythrinus*, *L. budegassa*), 4 species showed no significant difference in differentiation between inner versus outer population pairs (*C. galerita*, *H. guttulatus*, *G. niger*, *S. pilchardus*) and 1 showed higher differentiation (*A. boyeri*) within the suture zone (Fig. 1B, top). For this last species, this result was likely explained by a confounding effect of the marine/lagoon ecotype structure overlapping geography. Indeed, the Med-in samples were taken from a lagoon habitat while the Atl-in population was from a marine environment. After removing this species, we found little evidence of higher occurrence of species showing lower genetic admixture within the suture zone (exact binomial test, 12 versus 4 species, $H_0$: p = 0.5, $p-value = 0.08$), although we note that sample size limited the power of this test.

Secondly, we tested for the presence of genetic admixture within the transition zone with the $f_3$ statistics (Fig 2A, Fig. S12). Among 18 species (removing again *A. fallax*), we found evidence of genetic admixture (i.e. $f_3 < 0$) within the suture zone for 7 species, with 4 in the Mediterranean Sea only (*C. julis*, *L. mormyrus*, *S. cantharus*, *S. cabrilla*), 1 in the Atlantic Ocean only (*D. labrax*), and 2 in both sides of the suture zone (*D. puntazzo* and *L. budegassa*). We found no evidence of genetic admixture in the populations outside the transition zone. We found no difference in $F_{ST}$ between species showing genetic admixture or not in the Mediterranean Sea (Kruskal Wallis test, $\chi^2 = 0.49$, $p-value = 0.48$) and the Atlantic Ocean (Kruskal Wallis test, $\chi^2 = 0.05$, $p-value = 0.82$), or admixture within one the two basins (Kruskal Wallis test, $\chi^2 = 0.35$, $p-value = 0.55$).

Thirdly, as $f_3$ has low power to detect small levels of introgression especially when divergence is strong, we tested for unbalanced introgression between two populations from the same basin, originating from one population in the alternate basin using the $D$ statistic (Fig 2B, Fig. S13). We detected significant signals of differential introgression ($D = / = 0$) for 8 out of 10 species for which variant orientation and outgroup species were available (*S. cantharus*, *S. typhle*, *S. cinereus*, *D. labrax*, *S. cabrilla*, *M. surmuletus*, *S. pilchardus*, *L. budegassa*). Again, we found no difference in $F_{ST}$ between species showing differential introgression or not (Mediterranean Sea: Kruskal Wallis test, $\chi^2 = 0.01$, $p-value = 0.91$; Atlantic Ocean: $\chi^2 = 1.64$, $p-value = 0.20$). These two results support that genetic admixture occurs all along the differentiation continuum.

Finally, we inferred the probability of contemporary gene flow between basins both between inner and outer populations using the ABC framework implemented in `DILS` (Fig. 2C, Fig. S24, Fig. S11). Among 18 tested species, we found 3 showing no ongoing gene flow between populations within the suture zone (*C. galerita*, *S. cinereus* and *S. typhle*). By contrast, 6 species showed no ongoing gene flow between the two outer populations (corresponding to the 3 previous species, plus *L. mormyrus*, *D. puntazzo* and *S. cantharus*). Species showing ongoing gene flow within the suture zone were not limited to the lower end of the differentiation continuum, and consistently with previous admixture tests, they did not have reduced differentiation (as measured by $F_{ST,out}$) than species without ongoing gene flow in the suture zone (Kruskal Wallis test, $\chi^2 = 2.56$, $p-value = 0.11$). However, species showing no contemporary gene flow between outer populations had significantly higher differentiation on average (Kruskal Wallis test, $\chi^2 = 10.14$, $p-value = 0.0015$), suggesting a limit to the spread of gene flow beyond the contact zone in the most divergent species.

## Distribution of coalescence times

The inferred distributions of coalescence times (i.e. TMRCA between non-recombining blocks of sampled chromosomes) obtained with `tsinfer` and `tsdate` within and between basins

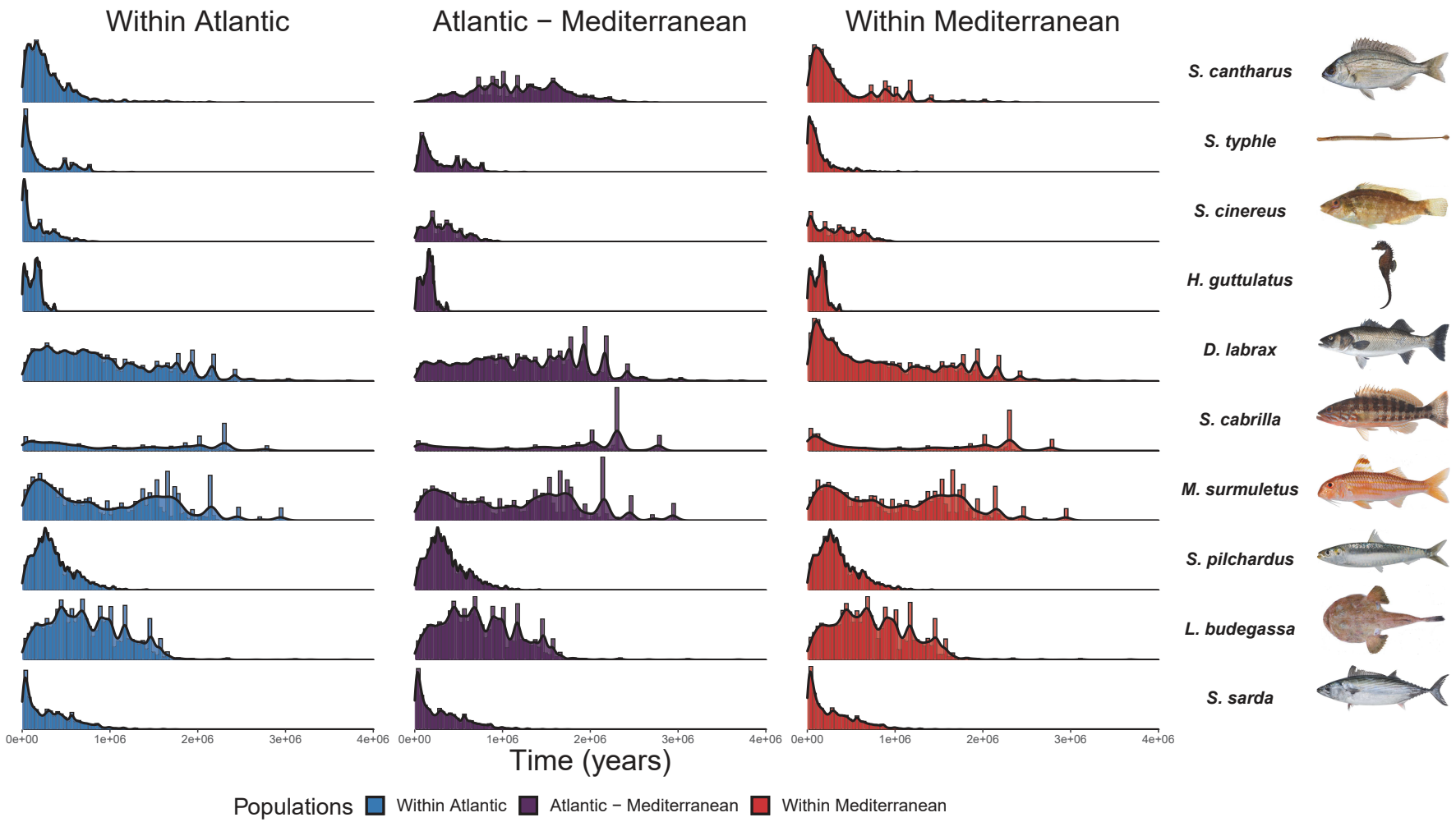



Figure 3: **Distribution of coalescent times within and between basins**. The genome-wide tree sequence of non-recombining genomic segments was inferred with `tsinfer` and branch lengths were estimated with `tsdate` using individual haplotypes within contiguous scaffolds longer than 10kb. Each panel represents the genome-wide distribution of time to the most recent common ancestor (TMRCA) converted to years, using all haplotype pairs taken either within the Atlantic (left, blue), the Mediterranean (right, red) or between Atlantic and Mediterranean samples (middle, purple), for ten species in rows. Branch lengths were first estimated in units of number of generations and subsequently rescaled with the mean generation time estimated with AgeNe, taking into account lifespan, age at maturity and age-specific fecundity and survival summarized in life tables. See Fig. S23, for unscaled branch lengths

showed a broad variation among species (Fig 3). For some of them, within-basin distributions contained ancient coalescent times that clearly exceeded the predicted distribution for a single population under the standard coalescent model (e.g. *S. cantharus*, *D. labrax*, *S. cabrilla*, *M. surmuletus*). These old coalescence times were enriched in the between-basin TMRCA distributions to different extents (i.e. most strongly for *S. cantharus*), reflecting past divergence events followed by more or less pronounced mixing of lineages due to gene flow between basins. By contrast, some species and in particular those inhabiting both coastal marine and lagoon habitats showed coalescence time distributions that were limited to the relatively recent past (*S. typhle*, *S. cinereus*, *H. guttulatus*). Overall, there was no clear association between the genome-wide average genetic differentiation ($F_{ST}$) and the shape of the TMRCA distributions, suggesting a wide variety of scenarios of divergence accumulation and erosion by gene flow across species.

Our coalescent simulations provided a reference framework for interpreting the obtained empirical TMRCA distributions (Fig. S18-S20). The excess of old within-population coalescence times beyond predictions of the single population model was not observed under the strict isolation model. However, all models of divergence with gene flow, migration pulse, secondary contact and ancient introgression from a divergent lineage generated excesses of ancient coalescent times, that keep track of the divergence time between admixed lineages. Although recording this information in the real tree sequence of non-recombining genealogies may seem trivial, our simulations integrating the tree sequence inference and dating steps by tsinfer and tsdate showed that the signature of divergence and gene flow remained detectable despite estimation uncertainties. For instance, the expected bimodality of TMRCAs under secondary contact was detected by `tsdate` with little influence of the parameters (i.e. prior distribution of coalescence times and mismatch ratio) on the inference accuracy. However, a small sample size similar to the one used here (i.e. 20 haplotypes per population) probably led us to underestimate the mean age of ancient blocks in the right end of the TMRCA distribution represented by divergent lineages produced during the allopatric phase (Fig. S21-S22). Therefore, our simulation work shows that the extended TMRCA distributions obtained for some species are likely to result from complex histories of divergence and gene flow. The bi- or multi-modal distributions such as those detected in *D. labrax*, *S. cabrilla* or *M. surmuletus* most probably indicate past episodes of divergence and secondary contact, or possibly ancient admixture with a sister lineage, with uncertainties probably leading to underestimated datings as we move further into the past. Interestingly, the splitting times estimated by ABC inference also pointed to divergence events going back several hundred thousand to more than a million generations in a majority of species (Fig. S24). Our results thus provide indications from different aspects of the data, that anciently diverged alleles are involved in the different genetic make-up between Atlantic and Mediterranean populations in several species.

## Decreasing permeability to gene flow with increasing genome-wide divergence

A first indication of the degree of semi-permeability to gene flow across species was obtained from the comparison of the genomic distributions of windowed $F_{ST}$ values calculated at 50kb resolution between outer populations (Fig. 4A). For all species with low to moderate genome-wide differentiation (e.g. eight first species at the bottom of Fig. 4A), the mode of the distribution was close to zero, but the right tail of most of these distributions showed the existence of more strongly differentiated genomic regions. The spread of these distributions

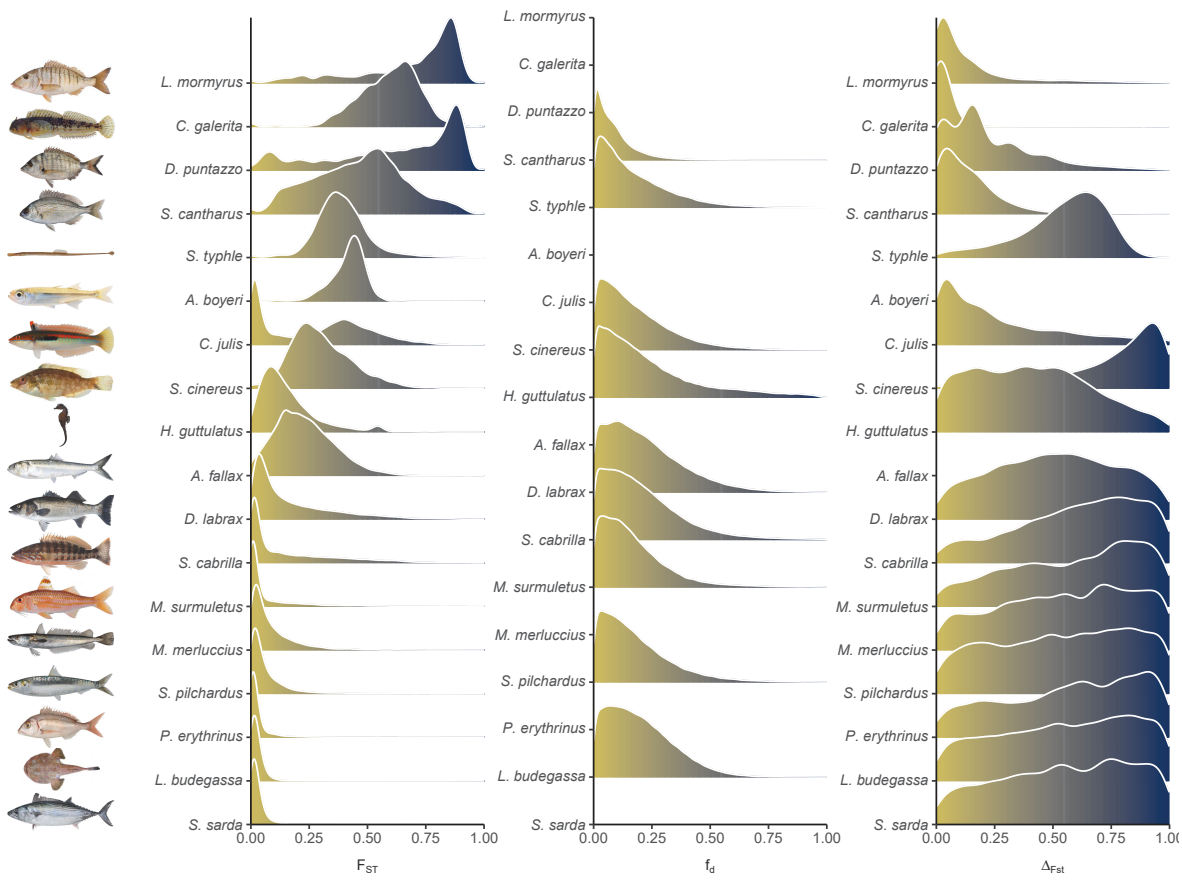



Figure 4: **Variation in semi-permeability along the divergence continuum**. The degree of semi-permeability to gene flow across species genomes was described by comparing the distributions of three statistics calculated in non-overlapping 50kb windows. Left: Genome-wide distribution of $F_{ST}$ assessed between outer populations for 18 species and ordered from lower (bottom) to higher (top) average differentiation. Middle: Genome-wide distribution of $f_d$ for 10 species with outgroup information. Right: Genome-wide distribution of relative decay in differentiation measured by the $\Delta_{Fst} = ()F_{ST,out} - F_{ST,in})/F_{ST,out}$ ratio for 16 species with complete sampling in all four populations. Genomic windows in these distributions are expected to range between 0 (i.e. strong resistance to gene flow due to reproductive isolation barriers) and 1 (i.e. complete erosion of differentiation in the absence of barrier loci). In each panel, the spread of each distribution is illustrated with a color gradient from yellow to dark blue.

towards increasingly extreme Fst values generally increased with the average genome-wide differentiation. This might reflect an increase in the total number and intensity of barriers to gene flow with increasing divergence.

For species with intermediate to high levels of differentiation (i.e. $F_{ST} > 0.2$), we found markedly different situations. Some species showed a relatively homogeneous distribution of $F_{ST}$ centered around the mean genome-wide value with limited variance across windows (e.g. *S. typhle* and *A. boyeri*). We note, however, that this concerned species has both geographical and ecotype components of population structure. On the contrary, four species showed huge variation in windowed $F_{ST}$, indicating a possibly very large variance in permeability across their genome. The most extreme case was found in *C. julis*, which showed a strongly bi-modal distribution with window values either close to $F_{ST} = 0$, or 0.5. This suggests that genome permeability to Atl-Med gene flow is highly bi-modal in this species, with genomic regions experiencing unrestricted gene flow while others being strongly resistant to gene flow. The sharp-snout seabream (*D. puntazzo*) displayed a somewhat similar but more extended bi-modality, ranging from almost null ($F_{ST}$ near 0) to complete ($F_{ST} \approx 1$) genetic differentiation across windows. By contrast, *S. cantharus* showed a flattened, mostly unimodal distribution ranging nearly from 0 to 1 and centered around $F_{ST} = 0.5$, while *L. mormyrus* showed a sharp peak of $F_{ST}$ at around 0.8, with a long left-tailed distribution extending to low differentiation windows, likely indicating a small proportion of regions permeable to introgression in this species.

Heterogeneity in the degree of semi-permeability across species was also reflected by variation in introgression rates estimated across the whole-genomes using $f_d$ (Fig. S14). For the species' triplet topologies that previously showed evidence for introgression using the $D$ statistic, we inferred average percentages of introgression with $f_d$ ranging from 0 for *S. sarda* to 0.125 for *S. cabrilla*. We note, however, that $f_d$ was not perfectly adapted to quantify absolute levels of introgression given our sampling scheme, especially in high gene flow species with no spatial pattern of population structure. Thus, introgression was likely underestimated in the least differentiated species, and accordingly, we found a significant inverse polynomial relationship between mean $f_d$ and $F_{ST,ou}$ ($LL$ linear model = 16.540 ; $LL$ polynomial model = 20.223, $\chi^2 = 7.36$, $p - value = 0.0066$), indicating that species with high genetic differentiation tended to show reduced $f_d$ compared to species with intermediate $F_{ST}$ values (Fig. S14). Beyond this genome-scale trend, the comparison of $f_d$ distributions along the differentiation gradient (Fig. 4B) did not provide a detailed picture of the variance in semi-permeability across species.

The relative decay in genetic differentiation in inner compared to outer pairs measured by the $\Delta_{Fst} = ()F_{ST,out} - F_{ST,in})/F_{ST,out}$ ratio was a better suited statistic for measuring introgression in four-population tree. The median value of species' genomic distributions showed a negative correlation with the mean genome-wide $F_{ST,out}$ ($R^2 = 0.58$, $p - value = 5.7e^{-4}$). This result indicated that the most divergent population pairs were generally more resistant to gene flow, with genetic differentiation in the inner pair maintained at more similar levels to the outer pair due to reduced introgression in the suture zone. This measure of resistance to gene flow was even more informative when examining variation in the relative decay of differentiation across 50 kb windows of the genome (Fig 4C). At the upper end of the divergence gradient, only a minor fraction of the *C. galerita* genome (and to a lower extent *L. mormyrus*) was found to introgress within the transition zone. Species like *D. puntazzo*, *S. cantharus* or *C. julis* showed broader distributions with higher proportions of genomic regions almost losing entirely their divergence between populations from the suture zone. Intermediate degrees of semi-permeability were observed in *H. guttulatus*, *D. labrax*, *S. cabrilla* and *M. surmuletus*, which displayed increasing genome fractions with elevated erosion of divergence in the transition zone. By contrast, the relative decay in genetic differentiation was more homogeneous in *S. typhle*, reflecting the low variance in divergence observed across its genome. The other species had most of their genome

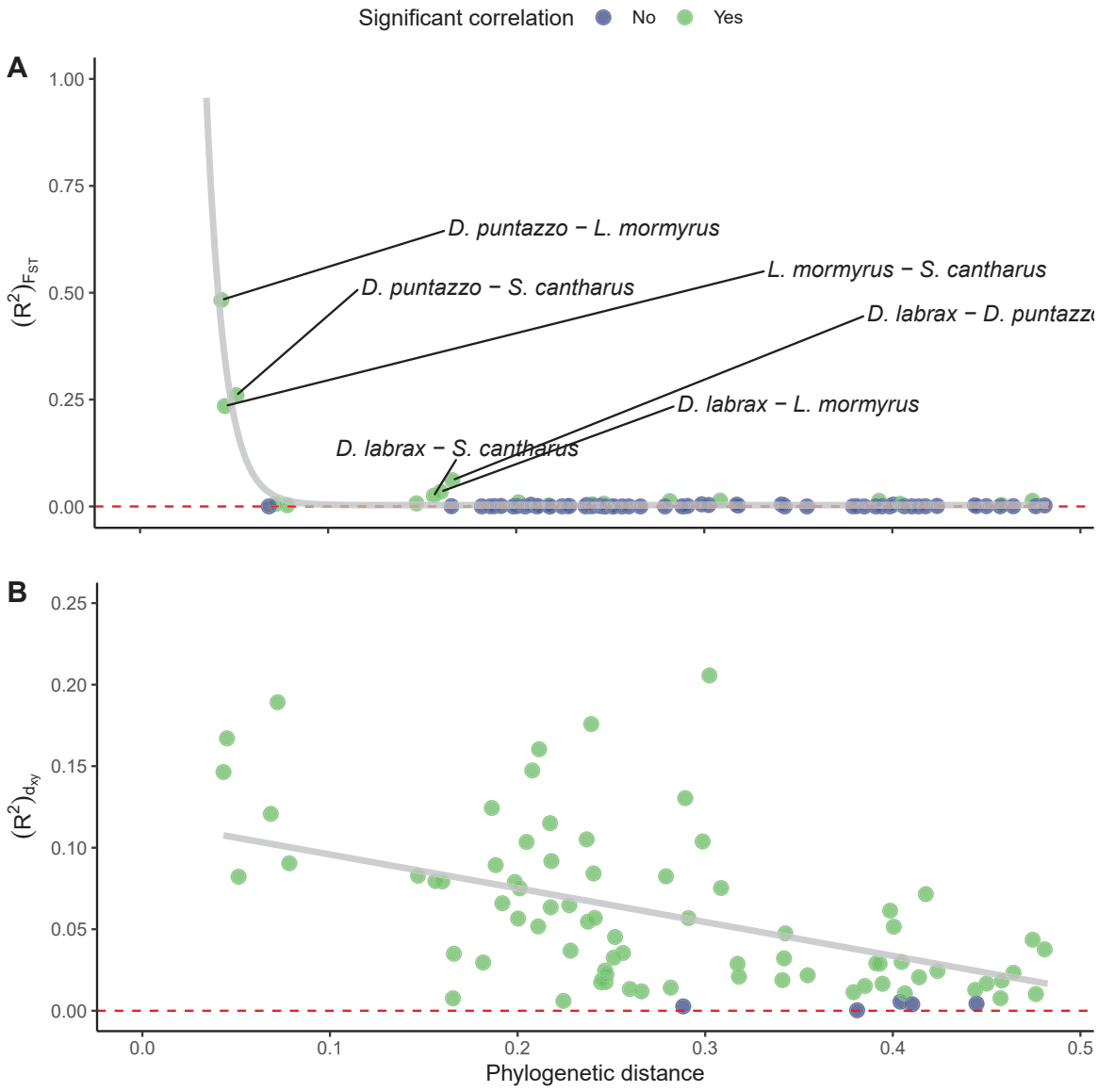



Figure 5: **Correlation of genomic landscapes of differentiation and absolute divergence between different species**. Genetic differentiation ($F_{ST}$) and absolute divergence ($d_{XY}$) were compared at highly conserved Actinoptegyian ortholog genes across all pairwise comparisons among 14 species. The percentage of variation explained by a linear model between gene-specific $F_{ST}$ across species (panel A) and the same for $d_{XY}$ (B) is represented against phylogenetic distance for each pair of species. Green and blue points represent significant and non-significant linear regression, respectively. Grey lines show the best fit using either an exponential model (panel A, $y = a \times exp^{-S \times x} + K$) or a simple linear regression (B). In panel A, the six species pairs with the strongest $R^2_{FST}$ are labeled: these pairs involved the three Sparidae species of the dataset (*D. puntazzo*, *L. mormyrus* and *S. cantharus*) in addition to the European sea bass (*D. labrax*).

losing more than 50% of the $F_{ST,out}$ value in the inner pair, suggesting high permeability to gene flow.

## Limited effect of genome architecture on the similarity of differentiation landscapes between species

Teleost fish species share relatively conserved genomic architectures in terms of both chromosome number and shared synteny (Mank and Avise, 2006; Schartl et al., 2013), with most structural rearrangements occurring within rather than between chromosomes. This stability of genome architecture could constrain the evolution of homologous genome regions in a similar way between distantly related species due to correlated distributions of recombination rates, and correlated exposure to the effects of linked selection along the genome. It was, therefore, important to verify whether the species studied here provide independent observations regarding their genomic differentiation landscapes, or whether on the contrary the similarity of differentiation landscapes is constrained by the genomic architecture and the phylogenetic distance between species (Fig. S26). For this purpose, we regressed the values of differentiation and divergence for each pair of species at highly conserved orthologous loci contained in the actinopterygian `BUSCO` database (i.e. $F_{ST,out}$ sp1 vs $F_{ST,out}$ sp2, and $d_{XY,out}$ sp1 vs $d_{XY,out}$ sp2). The strength of these correlations (measured here by the percentage of variation $R^2_{FST}$ and $R^2_{dxy}$ explained by linear regression) was used to quantify the level of intrinsic constraint shared by each pair of species.

We found an inverse exponential decay of $R^2_{FST}$ as a function of increasing phylogenetic distance between species ($t = 9.338$, $p - value = 8.28e^{-15}$, Fig. 5A). Non-negligible $R^2_{FST}$ ($> 0.2$) were only found between the three closest species within the Sparidae family, which also show substantial Atlantic - Mediterranean differentiation (*D. puntazzo*, *L. mormyrus* and *S. cantharus*), while the strength of the correlation decreased sharply as the phylogenetic distance between species increased. The correlation was also statistically significant in a few other pairs, notably involving *D. labrax* and the three species mentioned above, but with qualitatively weaker strengths as illustrated by their low $R^2_{FST}$ values. Therefore, conserved genome architecture across the phylogeny of teleost fishes was unlikely to generate correlated genetic differentiation landscapes beyond the most closely related sparid species.

In sharp contrast, $R^2_{dxy}$ displayed a highly significant linear decay with phylogenetic distance ($t = -6.863$, $p - value = 8.7e^{-10}$, Fig. 5B), with absolute divergence being significantly correlated even between the most distant species studied. This correlation, however, rather reflected similar rankings of genes' evolutionary constraints between species, since $d_{XY}$ also captures the direct effect of selection acting on constrained gene sequence.

We finally tested whether the subset of highly conserved BUSCO orthologs had different distributions of $F_{ST}$ and introgression rates between basins as compared to the remainder of the genome. Our rationale was that stronger purifying selection acting on BUSCO genes would tend to increase $F_{ST}$ above its background value more markedly between isolated populations than between populations connected by high migration rates (Cruickshank and Hahn, 2014). We found a significant difference in mean $F_{ST}$ between BUSCO and non-BUSCO genome regions for nearly all species (except for *S. typhle*), but surprisingly the sign of the difference varied among species (Fig. S25A). However, consistent with the prediction, a marginal correlation was found indicating higher differentiation at BUSCO compared to non-BUSCO sequences with increasing mean genome-wide $F_{ST}$ ($p - value = 0.038$). In addition, we found very strong evidence for differential introgression between coding and non-coding sequences for 12 species (*D. labrax*, *L. budegassa*, *M. merluccius*, *S. typhle*, *D. puntazzo*, *S. sarda*, *P. erythrinus*, *C. julis*, *S. cabrilla*, *H. guttulatus*, *M. surmuletus* and *S. cantharus*), and all exhibiting reduced introgression rates

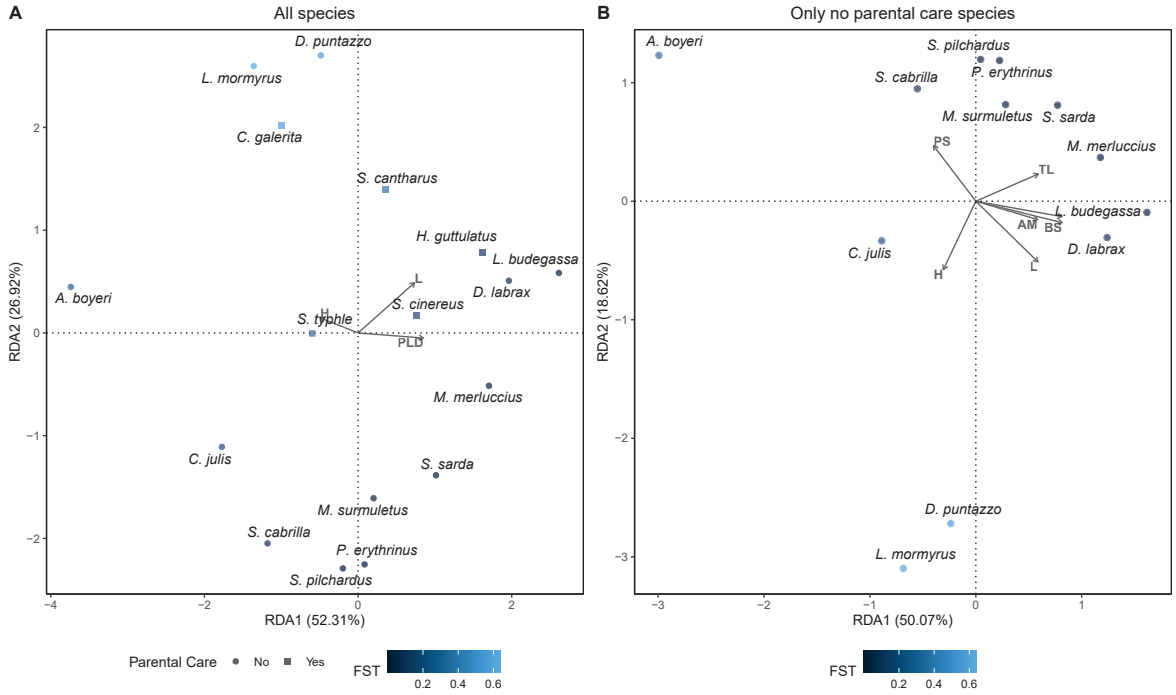



Figure 6: **RDA of 17 species of genetic differentiation, divergence, diversity and introgression explained by several life-history traits**. A) Complete dataset: genetic characteristics are explained by 3 life-history traits: hermaphroditism (H), lifespan (L) and pelagic larval duration (PLD), B) Only species with no parental care behavior: genetic characteristics are explained by 7 life-history traits: hermaphroditism, lifespan, pelagic larval duration, propagule size (PS), trophic level (TL), body size (BS) and age at first maturity (AM). Darker to lighter colors indicate lower and higher $F_{ST}$ respectively.

in conserved BUSCO genes compared to the remainder of the genome (Table S6). Thus, there was a general tendency for increased resistance to gene flow within genome regions containing conserved genes.

## Effect of life-history traits on speciation

Our overarching goal was to evaluate the role of life history traits on the evolutionary mechanisms involved in speciation (Fig. 6-7, Fig. S15-S17). The simultaneous analysis of genetic summary statistics combined with 10 life history traits used as constraining factors in a RDA showed a non-significant multivariate relationship between these traits and our measures of genetic diversity, differentiation, divergence and introgression (Permutation test, $F = 1.67$, $p - value = 0.18$). Following stepwise reduction to eliminate collinear explanatory variables, we found that three life-history traits (lifespan, pelagic larval duration and hermaphroditism) collectively explained 40.48% of the total variance in genetic summary statistics (Permutation test, $F = 2.95$, $p-value = 0.00471$) (Fig. 6A). Each of these traits had significant or marginally significant effects (PLD, $F = 2.59$, $p-value = 0.046$; Hermaphroditism, $F = 2.59$, $p-value = 0.040$ and Lifespan, $F = 2.41$, $p - value = 0.060$). When removing the 5 species with parental care behavior (Fig. 6B), 7 life-history traits (body size, trophic level, propagule size, age at first maturity, pelagic larval duration and hermaphroditism) collectively explained 88% of the total variance (Permutation test, $F = 4.22$, $p - value = 0.00256$), with marginally significant effect of propagule size ($F = 2.99$, $p - value = 0.075$) and lifespan ($F = 3.11$, $p - value = 0.067$).

Looking at specific relationships between life history traits and genetic descriptors of differentiation and divergence provided more complex and complementary insights. In agreement with the RDA results, $F_{ST}$ appeared to be correlated with pelagic larval duration (beta regression model, $z$-value $= -2.432$, $\rho = -0.03751$, $p$-value $= 0.015$, Fig. 7A) following a beta negative correlation rather a linear relationship (likelihood ratio test, log-likelihood for linear model $= 4.59$, log-likelihood $= 13.93$, $\chi^2 = 18.69$, $p$-value $< 2.2e - 16$). This relationship was also observed when focusing only on populations from the suture zone ($z$-value $= -2.086$, $\rho = -0.03325$, $p$-value $= 0.037$) and was stronger considering only species with ongoing gene flow between outer populations (i.e. excluding six species shown in blue in Fig. 7, between outer populations $z$-value $= 0.01627$, $\rho = -0.04945$, $p$-value $= 0.0023$; and inner populations $z$-value $= -3.818$, $\rho = -0.06554$, $p$-value $= 0.000135$).

We found negative correlation between $d_{XY}$ between outer populations and body size ($t$-value $= -3.289$, $p$-value $= 0.00433$, Fig. 7B), trophic level ($t$-value $= -2.402$, $p$-value $= 0.02803$), lifespan ($t$-value $= -2.231$, $p$-value $= 0.0394$), and PLD ($t$-value $= -2.335$, $p$-value $= 0.032$), all traits that are strongly correlated among each other (see Fig. S5 in chapter 1 supplementary material). These four traits explain 40.21% of total variance of genetic divergence. Weaker correlations was only observed with body size ($t = -2.528$, $p$-value $= 0.0232$) and PLD ($t = -2.202$, $p$-value $= 0.044$). Absolute divergence ($d_a$) was higher for hermaphroditic species (Kruskall Wallis, $\chi^2 = 4.4308$, $p$-value $= 0.0353$) and negatively correlated to trophic level ($t = -2.655$, $p$-value $= 0.01668$).

Species with long adult lifespan also had a higher probability of admixture as detected by $f_3$ (Kruskall Wallis, $\chi^2 = 6.6455$, $p$-value $= 0.00994$) and species with parental care behavior have less probability of ongoing gene flow (Kruskal Wallis, $\chi^2 = 6.4077$, $p$-value $= 0.01136$). We found a negative correlation between time of ancestral split and lifespan ($t = -3.941$, $p$-value $= 0.00117$, $R^2 = 0.49$, Fig. 7C). We found no phylogenetic signal for any life-history traits except for propagule size (Pagel's $\lambda = 0.99934$ ; $p - value = 0.0054$ and Bloomberg's $K = 1.12688$ ; $p - value = 0.0544$, Table S5).

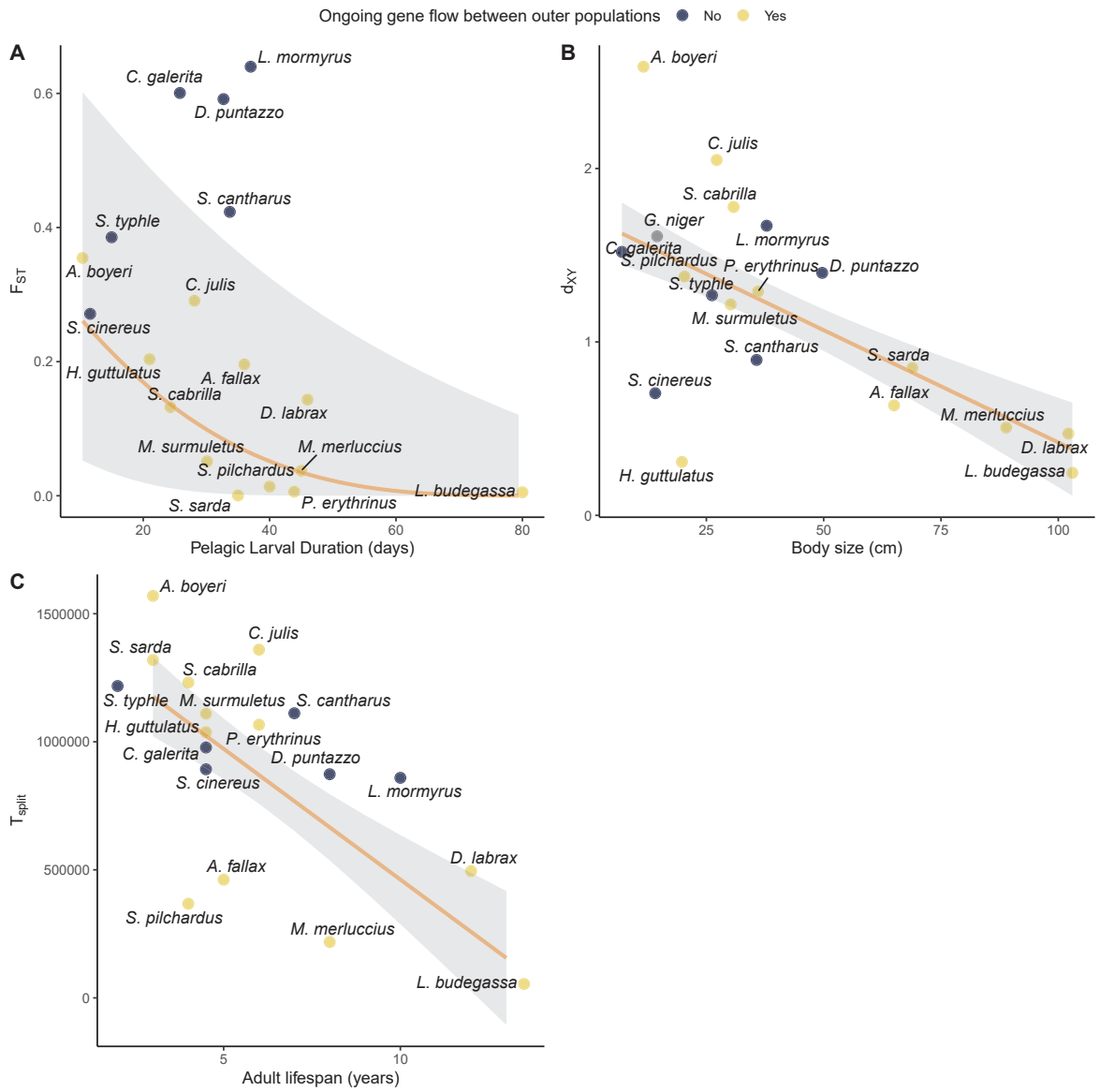



Figure 7: **Relationships between life-history traits and genetic characteristics**. A) shows a negative beta regression between $F_{ST}$ outer populations and pelagic larval duration, representative of dispersal capacities; B) negative relationship between absolute divergence ($d_{XY}$) and individual body size, representative of abundance and effective population size, C) negative relationship between split time between Atlantic and Mediterranean outer populations and adult lifespan, representative of generation time. Yellow and violet points represent species showing and not showing contemporary gene flow between outer populations.

# Discussion

Our multispecies comparative population genomics study based on whole-genome polymorphism data revealed a surprising diversity of evolutionary histories in 19 marine fishes with broadly similar large-scale distributions. Although they have evolved in the same biogeographical context, these species show a wide spectrum of contemporary genetic structures ranging from near panmixia to strong subdivision into divergent reproductively isolated lineages (Fig. 1). Moreover, while most species show some degree of geographical subdivision between Atlantic and Mediterranean basins, four species from the coastal zone have an additional ecological component of differentiation, which might involve ecotype variation associated with the use of marine and lagoon habitats. The diversity of evolutionary trajectories associated with these patterns was captured at all the levels that were explored by our analyses. (i) First, species differ from each other in their amount of ancestral genetic variation. (ii) Secondly, genetic exchanges between basins are highly variable from one species to another, revealing a wide diversity of histories, directionalities, and current intensities in gene flow (Fig. 2). (iii) Thirdly, some species carry in their genome the legacies of past divergence events, which contribute to different extents to the present genetic structure (Fig. 3). (iv) Finally, the genomic heterogeneity of permeability to gene flow differed between species, reflecting a wide range of reproductive isolation levels along a speciation continuum (Fig. 4). These four levels represent key components in the process of genomic divergence that can lead to speciation, ranging from the sorting of ancestral polymorphism to the accumulation of barriers to gene flow. Distinguishing and measuring these different compounds is therefore of prime importance to understanding the influence of factors such as life-history traits on organismal diversification.

## Gene flow variation along a speciation continuum

As reproductive isolation increases due to the accumulation of barriers over time, the intensity of gene flow effectively leading to introgression between diverging lineages is expected to decrease (Coyne and Orr, 1989; Matute and Cooper, 2021). Empirically, this effect has been detected through negative relationships between measures of introgression and molecular divergence - though, few comparative studies have addressed this relationship with genome-scale data. Using ABC inference in 61 pairs of animal taxa ranging from populations to real biological species, Roux et al. (2016) found that the probability of contemporary gene flow becomes very small above 2% of net synonymous divergence. In a meta-analysis of 123 studies, Dagilis et al. (2021) found a weak negative relationship between genetic distance and the strength of introgression (quantified with Patterson's D), which was furthermore not supported in all taxonomic groups. Hamlin et al. (2020) also found a very weak negative relationship between genetic distance and the fraction of introgression (measured by $D_p$) among several *Solanum* tomato species. All these studies suggest that there is no simple, strong relationship between introgression and divergence.

Here, we found evidence for gene flow across the entire divergence gradient covered by our study, which extends to 1.04% of net genome-wide divergence. However, effective gene flow was not always detected between populations that are not directly located within the suture zone, including for species with a net divergence well below 1% (e.g. 0.44% in *S. cantharus*). Since this result could not be explained by the lack of opportunity for gene flow in the suture zone, it most likely reflects the multigenerational effect of selection against introgressed genome segments, which restricts the spread of foreign ancestry within recipient lineages and eventually eliminates introgression at some distance from the contact zone (Sedghifar et al., 2016). Our

results thus illustrate the importance of the spatial sampling design to improve the detection of gene flow and selection against introgressed regions in the presence of incomplete lineage sorting.

A limitation in comparing introgression to genetic divergence, however, is that measures of genetic divergence between closely related taxa do not accurately (and even sometimes do very badly) reflect divergence time for several reasons. Divergence depends essentially on demographic time over a potentially long time ($\sim$10 $N_e$ generations), but also on gene flow and selection. This means that introgression and molecular divergence are not independent variables, because higher introgression will reduce genetic divergence. Gene flow can therefore potentially artificially amplify the negative correlation between the strength of introgression and genetic divergence. To circumvent this problem, it may be preferable to estimate coalescence times directly between lineages, in a way that is not influenced by gene flow and recombination. Inference of the ancestral recombination graph under a demographic divergence model including gene flow has proven a powerful approach to map Neanderthal and Denisovan ancestry across the human genome, even with small sample sizes (Hubisz et al., 2020). Here, we took a simpler but similar approach by analyzing the distributions of Time to the Most Recent Common Ancestor (TMRCA) inferred from genome-wide genealogies on genomic intervals bounded by historical recombination events. Although such estimates are subject to some uncertainties (Brandt et al., 2022), these are unlikely to explain the large differences in TMRCA among species that we observed. Beyond the fact that our results suggest little synchrony between the divergence times of species that evolved in a shared biogeographic context, they also provide evidence against a systematic increase in reproductive isolation with divergence time. Three particularly striking examples are provided by the European sea bass (*D. labrax*), the sea comber (*S. cabrilla*) and the striped red mullet (*M. surmuletus*), which display low average genetic differentiation ($F_{ST}$ <0.14) and low net divergence ($d_a$ <0.26), but relatively large fractions of anciently diverged alleles (>2 Myrs) in their genomes. In contrast, the black seabream (*S. cantharus*) shows much stronger $F_{ST}$ and $d_a$ and even no contemporary gene flow between populations outside of the transition zone, seemingly pointing to higher reproductive isolation. Yet, the oldest alleles in this species are younger than those found in the 3 previous species. In *D. labrax*, the most divergent alleles now involved in reproductive isolation between Atlantic and Mediterranean lineages have been found to originate from an ancient episode of admixture with a closely related species, the spotted sea bass *D. punctatus* (Duranton et al., 2020). Therefore, a similar scenario could possibly explain the presence of anciently diverged alleles in the two other species cited above (*S. cabrilla* and *M. surmuletus*), especially because they are both found in sympatry with a closely related species: the red mullet (*M. barbatus*) and the blacktail comber (*S. atricauda*), with which they were possibly admixed in the past. Whether the old alleles found in *S. cabrilla* and *M. surmuletus* are also associated with significant reductions of gene flow remains, however, to be elucidated. Ancestry inference based on genome-wide genealogies is therefore a useful complement to divergence measures that may overlook the signatures of old evolutionary processes due to the rejuvenating effect of gene flow and recombination.

## Complex relationships between life-history traits and speciation

Overall, we found little evidence for simple, direct effects of unique life-history traits on the genetic components of speciation that were analyzed here. This lack of a major-effect trait probably reflects the multifarious and complex influence of these traits on the emergence and erosion of divergence and reproductive barriers. In agreement with this view, redundancy analyses suggested that the combined effects of several traits taken together could explain a

very significant fraction of the variance in genetic diversity, differentiation, divergence and gene flow among species. Looking into more details, we found, however, several interesting patterns that enlighten the impact of specific traits.

We knew from the analyses presented in chapter 1 that lifespan and parental care are the main determinants of ancestral genetic diversity in marine fishes (Barry et al., 2022). Species with long adult lifespan and with parental care behavior should thus have less ancestral genetic diversity on which reproductive isolation may build up. However, it is still not clear to what extent the number and effect of reproductive isolation mutations initially present as standing variants, correlate with the size of the reservoir of neutral ancestral variation before split. Here, we also found that lifespan is negatively correlated to the time (in generations) of ancestral split between lineages (Fig. 7C). This pattern might indicate that the time of split (measured in years) estimated from ABC inference is relatively similar across species, since species that have longer lifespans also have longer generation times. Such finding would not necessarily be inconsistent with the high heterogeneity of TMRCA distribution between species. Indeed, the age of inferred population splits, possibly imposed by more recent biogeographic constraints, does not necessarily capture older admixture events that possibly led to the old alleles detected in some species.

We also found a negative relationship between genetic differentiation and pelagic larval duration among species that continue to exchange genes between populations located outside of the transition zone. Pelagic larval duration is an approximation of the dispersal capacities of marine organisms, assuming passive larval drift (Macpherson and Raventos, 2006; Selkoe et al., 2014). Such an inverse relationship is expected under migration-drift equilibrium in the simplest situation of the infinite-island model (Wright, 1931), where $F_{ST} = \frac{1}{4N_e m+1}$. However, it is also expected in the presence of genetic barriers to gene flow due to migration-selection antagonism in semipermeable genomes (Barton and Bengtsson, 1986; Charlesworth et al., 1997; Feder and Nosil, 2010). It thus seems logical that the relationship between PLD and genetic differentiation holds only for species that are still exchanging genes between populations outside of the transition zone. Once reproductive barriers become strong (i.e. genome hitchhiking stage, sensu Feder et al. (2012), dispersal becomes irrelevant because no matter how frequently immigrant individuals interact with residents, effective gene flow will be strongly reduced by hybrid breakdown and/or selection against introgressed ancestry.

Finally, absolute genetic divergence ($d_{XY}$) was negatively correlated with body size, trophic level, lifespan and pelagic larval duration, four traits that are highly correlated with each other and all more or less direct proxies for $N_e$. As absolute genetic divergence reflects the sum of ancestral genetic diversity and net divergence, the observed relationship can be explained in part by the previously mentioned effect of adult lifespan on long-term genetic diversity. When focusing on net divergence (which only depends on mutation rate and time), only the effect of trophic level remained (weakly) significant. Hence, differences in $N_e$ between species do not seem to explain the amount of molecular divergence accumulated since ancestral split, as expected. More surprisingly, lifespan was not correlated to $d_a$, although this measure of divergence reflects the accumulation of mutations across generations. This can be explained by the previous observation that introgression is widespread along the differentiation continuum, which can lead to underestimate $d_a$ under high gene flow and blur its potential relationship with lifespan.

The comparative approach developed here also has some limitations. First, life-history traits are not perfect predictors of demographic parameters. In a meta-analysis of 44 marine invertebrate and vertebrate species, Shanks (2009) found no correlation between pelagic larval duration and dispersal distances for species that disperse beyond 20km, which is probably the case for nearly all of the 19 species studied here. This study criticized the use of PLD as a proxy for dispersal capacities because of considerable variability in dispersal distances for

species having PLD superior to 10 hours (like most marine fishes). This variability can be explained by the fact that larvae do not drift passively with currents. The larvae of some species are active swimmers that respond to environmental cues involving for instance chemical signals or sounds and therefore do not drift passively in the water column. This active behavior can result in reduced dispersal movement (e.g. larval retention, self-recruitment, or homing), leading to dispersal patterns that do not follow oceanographic distances mediated by currents (Jones et al., 1999; Swearer et al., 1999). Thus, results related to pelagic larval duration should be considered with that limitation in mind, and consolidated with other behavioral life-history traits, that are, unfortunately, difficult to document (especially for the species studied here). Another limitation comes from the fact that correlative analysis between life-history traits and any genetic characteristics does not prove direct causality. Finally, these general correlations were based on a set of 19 species, which may also have limited somewhat the statistical power of our analyses.

## Cryptic diversity and implications for conservation biology

The net absolute divergence ($d_a$) showed a 500-fold variation among species, ranging from 0.002 in the blackbellied monkfish *L. budegassa* to 1.036% in the striped seabream *L. mormyrus*. Consistent with the definition of the "speciation grey zone" (De Queiroz, 2007; Roux et al., 2016), the net divergence interval between $d_a$ values of 0.5% and 2% comprised both partially and completely reproductively isolated pairs of fish lineages in our study. No species had Atlantic-Mediterranean net divergence higher than Roux et al's upper limit, but 11 species fell within the so-called grey zone. Species showing clear evidence of partial reproductive isolation as revealed by their semipermeable genomes had $d_a$ values as low as 0.05% (*M. surmuletus*). This result shows that pairs of marine fish taxa classically considered as simple populations of the same species can be engaged in speciation pathways even at very low levels of molecular divergence. In his comparative analysis of reproductive isolation in fish, Russell (2003) found that postzygotic barriers causing hybrid breakdown start to appear after 7% of mitochondrial divergence (based on cytochrome b) and always result in complete reproductive isolation beyond 20% divergence. Here, the most divergent species pair showing complete reproductive isolation (*C. galerita*) had 7.2% divergence on cytochrome-b. This suggests that the upper range of our reproductive isolation continuum coincides well with the onset of post-zygotic barriers with sufficiently strong effects to be detectable in experimental crosses. Despite high genetic divergence and reproductive isolation, most pairs of marine fish lineages showed no or little morphological differentiation. Some authors found differences in male coloration between Mediterranean and Atlantic individuals in the rainbow wrasse (*C. julis*) and black seabream (*S. cantharus*), and differences in crest size and width in Montagu's blenny (*C. galerita*) (Domingues et al., 2007). Yet, these differences were so slight that they were usually considered as intraspecific variation rather than evidence of reproductive isolation (but see Ramírez-Amaro et al. (2021). Furthermore, for two of our species showing the strongest divergence between basins, the sharp-snout seabream (D. puntazzo) and the striped seabream (*L. mormyrus*), no authors have found any noticeable morphological differentiation between the two lineages. This raises the importance of genetic and genomic approaches to characterize hidden layers of cryptic species diversity below the officially recognized taxonomic units. Some of these species are furthermore targeted by commercial and recreational fisheries. Stock delimitations should thus take into account existing genetic subdivisions to avoid future erosion of cryptic diversity in one or the other lineage.

Phylogeographical patterns of genetic differentiation between Atlantic and Mediterranean populations have been classically attributed to particular water flows in the Alboran Sea to

the East of the Strait of Gibraltar, which create a physical barrier to larval dispersal between the two basins (i.e. the Almerian-Oran Front) (Patarnello et al., 2007; Pascual et al., 2017). An alternative hypothesis involves, in addition, the existence of genetic barriers (e.g., genetic incompatibilities and/or local adaptations), which appear necessary to effectively reduce gene flow in the presence of frequent migrants (Bierne et al., 2011). These genetic barriers were likely built-up in allopatry during glacial periods of the Pleistocene, alternating with secondary contact episodes during which genetic homogenization has occurred in the permeable fraction of the genome (Tine et al., 2014; Duranton et al., 2018). Our results strongly support this latter view, although pushing back some divergence events before the Mid-Pleistocene Transition, as suggested by both ABC and gene genealogies inferences (supposing that the mutation rate is correct). Some highly divergent species like the rainbow wrasse (*C. julis*) showed a bimodal distribution of genetic differentiation along their genome, a pattern inconsistent with a simple migration-drift balance that would affect genetic differentiation homogeneously across the genome. One could argue that the least divergent species in our dataset have truly neutrally differentiated populations. However, the example of hake (*M. merluccius*), whose average genomic differentiation of $F_{ST} = 0.04$ does not seem compatible with the finding of one Atlantic-Mediterranean hybrid genotype among 20 sampled individuals, already shows signs of semi-permeability in its genome. Therefore, oceanographic current features at the Atlantic-Mediterranean suture zone explain the location rather than the origin of phylogeographical breaks (Bierne et al., 2011), and show as main effect a migratory asymmetry from the Atlantic to the Mediterranean shared by a majority of species in our dataset.

The approach developed here shows the strength of comparative phylogeography applied with whole-genome sequence data (Edwards et al., 2022), and the potential of recombination-aware analysis of genome-wide genealogies to reveal different facets of lineage divergence. The comparison of multiple species in a common biogeographic context will certainly help to further understand how reproductive isolation barriers accumulate and shape organismal diversification in interaction with species life-history traits. The analysis of other suture zones (e.g. Johannesson et al. (2020) will ultimately help to understand how different eco-biogeographical contexts influence speciation.

# Conclusion

Using a comparative genomic analysis of 19 species, we found that marine fish populations across the Atlantic Ocean - Mediterranean Sea suture zone have gone through varied evolutionary histories engaging them to different degrees on the way to speciation. This variability is reflected in a gradient of differentiation and divergence from near panmixia to highly reproductively isolated divergent lineages. By disentangling the different stages that influence both the progression and the regression of the speciation process, we related this diversity to variable contributions of deep divergence episodes, genetic admixture and selection against introgressed ancestry. Because of the multimechanistic nature of speciation, different life-history traits affecting different stages of species diversification were identified, including key determinants of long-term abundance, dispersal capacity, and generation time.

# Acknowledgments

The sequence data used in this work were produced by the Montpellier MGX platform and the Genewiz company, with the support of the GenSeq genotyping and sequencing platform

and the Montpellier Bioinformatics Biodiversity MBB platform, both being supported by ANR program "Investissements d'avenir" (ANR-10-LABX-04-01). We thank Marion Talbi for the development of the pipeline to infer phased VCFs. We thank colleagues who provided us with samples as well as those who facilitated or participated in sampling: F. Schlichta, T. Pastor, R. Castilho, R. Cunha, R. Lechuga, D. Pilo, C. Mena, J. Charton, T. Robinet, A. Darnaude, S. Vaz, M. Duranton, N. Bierne, S. Villéger, S. Blouet, as well as the fishermen and employees of fish markets and fish auctions. This work was supported by the ANR grant CoGeDiv ANR-17-CE02-0006-01.

# References


Adrion, J. R., Cole, C. B., Dukler, N., Galloway, J. G., Gladstein, A. L., Gower, G., Kyriazis, C. C., Ragsdale, A. P., Tsambos, G., Baumdicker, F., Carlson, J., Cartwright, R. A., Durvasula, A., Gronau, I., Kim, B. Y., McKenzie, P., Messer, P. W., Noskova, E., Ortega-Del Vecchyo, D., Racimo, F., Struck, T. J., Gravel, S., Gutenkunst, R. N., Lohmueller, K. E., Ralph, P. L., Schrider, D. R., Siepel, A., Kelleher, J., and Kern, A. D. (2020). A community-maintained standard library of population genetic models. *eLife*, 9:e54967.

Arnegard, M. E., McGee, M. D., Matthews, B., Marchinko, K. B., Conte, G. L., Kabir, S., Bedford, N., Bergek, S., Chan, Y. F., Jones, F. C., Kingsley, D. M., Peichel, C. L., and Schluter, D. (2014). Genetics of ecological divergence during speciation. *Nature*, 511(7509):307–311.

Barry, P., Broquet, T., and Gagnaire, P.-A. (2022). Age-specific survivorship and fecundity shape genetic diversity in marine fishes. *Evolution Letters*, 6(1):46–62.

Barton, N. and Bengtsson, B. O. (1986). The barrier to genetic exchange between hybridising populations. *Heredity*, 57(3):357–376.

Barton, N. H. and Hewitt, G. M. (1985). Analysis of Hybrid Zones. *Annual Review of Ecology and Systematics*, 16(1):113–148.

Bhatia, G., Patterson, N., Sankararaman, S., and Price, A. L. (2013). Estimating and interpreting FST: The impact of rare variants. *Genome Research*, 23(9):1514–1521.

Bierne, N., Welch, J., Loire, E., Bonhomme, F., and David, P. (2011). The coupling hypothesis: Why genome scans may fail to map local adaptation genes. *Molecular Ecology*, 20(10):2044–2072.

Blomberg, S. P., Garland, T., and Ives, A. R. (2003). Testing for phylogenetic signal in comparative data: Behavioral traits are more labile. *Evolution; International Journal of Organic Evolution*, 57(4):717–745.

Braasch, I., Gehrke, A. R., Smith, J. J., Kawasaki, K., Manousaki, T., Pasquier, J., Amores, A., Desvignes, T., Batzel, P., Catchen, J., Berlin, A. M., Campbell, M. S., Barrell, D., Martin, K. J., Mulley, J. F., Ravi, V., Lee, A. P., Nakamura, T., Chalopin, D., Fan, S., Wcisel, D., Cañestro, C., Sydes, J., Beaudry, F. E. G., Sun, Y., Hertel, J., Beam, M. J., Fasold, M., Ishiyama, M., Johnson, J., Kehr, S., Lara, M., Letaw, J. H., Litman, G. W., Litman, R. T., Mikami, M., Ota, T., Saha, N. R., Williams, L., Stadler, P. F., Wang, H., Taylor, J. S., Fontenot, Q., Ferrara, A., Searle, S. M. J., Aken, B., Yandell, M., Schneider, I., Yoder, J. A., Volff, J.-N., Meyer, A., Amemiya, C. T., Venkatesh, B., Holland, P. W. H., Guiguen, Y., Bobe, J., Shubin, N. H., Di Palma, F., Alföldi, J., Lindblad-Toh, K., and Postlethwait, J. H. (2016). The spotted gar genome illuminates vertebrate evolution and facilitates human-teleost comparisons. *Nature Genetics*, 48(4):427–437.

Brandt, D. Y. C., Wei, X., Deng, Y., Vaughn, A. H., and Nielsen, R. (2022). Evaluation of methods for estimating coalescence times using ancestral recombination graphs.

Carreras, C., García-Cisneros, A., Wangensteen, O. S., Ordóñez, V., Palacín, C., Pascual, M., and Turon, X. (2020). East is East and West is West: Population genomics and hierarchical analyses reveal genetic structure and adaptation footprints in the keystone species Paracentrotus lividus (Echinoidea). *Diversity and Distributions*, 26(3):382–398.

Charlesworth, B. (2009). Effective population size and patterns of molecular evolution and variation. *Nature Reviews Genetics*, 10(3):195–205.

Charlesworth, B., Nordborg, M., and Charlesworth, D. (1997). The effects of local selection, balanced polymorphism and background selection on equilibrium patterns of genetic diversity in subdivided populations. *Genetics Research*, 70(2):155–174.

Chen, J., Glémin, S., and Lascoux, M. (2017). Genetic Diversity and the Efficacy of Purifying Selection across Plant and Animal Species. *Molecular Biology and Evolution*, 34(6):1417–1428.

Chen, S., Zhou, Y., Chen, Y., and Gu, J. (2018). Fastp: An ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*, 34(17):i884–i890.

Cleary, J. G., Braithwaite, R., Gaastra, K., Hilbush, B. S., Inglis, S., Irvine, S. A., Jackson, A., Littin, R., Rathod, M., Ware, D., Zook, J. M., Trigg, L., and De La Vega, F. M. (2015). Comparing Variant Call Files for Performance Benchmarking of Next-Generation Sequencing Variant Calling Pipelines. Preprint, Bioinformatics.

Coughlan, J. M. and Matute, D. R. (2020). The importance of intrinsic postzygotic barriers throughout the speciation process. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 375(1806):20190533.

Coyne, J. A. (1974). The evolutionary origin of hybrid inviability. *Evolution; International Journal of Organic Evolution*, 28(3):505–506.

Coyne, J. A. and Orr, H. A. (1989). Patterns of Speciation in Drosophila. *Evolution*, 43(2):362–381.

Coyne, J. A. and Orr, H. A. (2004). *Speciation*. Sinauer Associates, Sunderland, Mass.

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

Dagilis, A. J., Peede, D., Coughlan, J. M., Jofre, G. I., D'Agostino, E. R. R., Mavengere, H., Tate, A. D., and Matute, D. R. (2021). 15 years of introgression studies: Quantifying gene flow across Eukaryotes.

Dai, H. and Guan, Y. (2020). The Nubeam reference-free approach to analyze metagenomic sequencing reads. *Genome Research*, 30(9):1364–1375.

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., and Durbin, R. (2011). The variant call format and VCFtools. *Bioinformatics*, 27(15):2156–2158.

Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., Whitwham, A., Keane, T., McCarthy, S. A., Davies, R. M., and Li, H. (2021). Twelve years of SAMtools and BCFtools. *GigaScience*, 10(giab008).

Dapporto, L., Cini, A., Vodă, R., Dincă, V., Wiemers, M., Menchetti, M., Magini, G., Talavera, G., Shreeve, T., Bonelli, S., Casacci, L. P., Balletto, E., Scalercio, S., and Vila, R. (2019). Integrating three comprehensive data sets shows that mitochondrial DNA variation is linked to species traits and paleogeographic events in European butterflies. *Molecular Ecology Resources*, 19(6):1623–1636.

De Queiroz, K. (2007). Species Concepts and Species Delimitation. *Systematic Biology*, 56(6):879–886.

Delaneau, O., Zagury, J.-F., Robinson, M. R., Marchini, J. L., and Dermitzakis, E. T. (2019). Accurate, scalable and integrative haplotype estimation. *Nature Communications*, 10(1):5436.

Dixon, P. (2003). VEGAN, a package of R functions for community ecology. *Journal of Vegetation Science*, 14(6):927–930.

Domingues, V. S., Faria, C., Stefanni, S., Santos, R. S., Brito, A., and Almada, V. C. (2007). Genetic divergence in the Atlantic–Mediterranean Montagu's blenny, Coryphoblennius galerita (Linnaeus 1758) revealed by molecular and morphological characters. *Molecular Ecology*, 16(17):3592–3605.

Durand, E. Y., Patterson, N., Reich, D., and Slatkin, M. (2011). Testing for Ancient Admixture between Closely Related Populations. *Molecular Biology and Evolution*, 28(8):2239–2252.

Duranton, M., Allal, F., Fraïsse, C., Bierne, N., Bonhomme, F., and Gagnaire, P.-A. (2018). The origin and remolding of genomic islands of differentiation in the European sea bass. *Nature Communications*, 9(1):1–11.

Duranton, M., Allal, F., Valière, S., Bouchez, O., Bonhomme, F., and Gagnaire, P.-A. (2020). The contribution of ancient admixture to reproductive isolation between European sea bass lineages. *Evolution letters*, 4(3):226–242.

Edwards, S. V., Robin, V. V., Ferrand, N., and Moritz, C. (2022). The Evolution of Comparative Phylogeography: Putting the Geography (and More) into Comparative Population Genomics. *Genome Biology and Evolution*, 14(1):evab176.

Emms, D. M. and Kelly, S. (2019). OrthoFinder: Phylogenetic orthology inference for comparative genomics. *Genome Biology*, 20(1):238.

Feder, J. L., Egan, S. P., and Nosil, P. (2012). The genomics of speciation-with-gene-flow. *Trends in Genetics*, 28(7):342–350.

Feder, J. L. and Nosil, P. (2010). The Efficacy of Divergence Hitchhiking in Generating Genomic Islands During Ecological Speciation. *Evolution*, 64(6):1729–1747.

Fraïsse, C., Popovic, I., Mazoyer, C., Romiguier, J., Loire, É., Simon, A., Galtier, N., Duret, L., Bierne, N., Vekemans, X., and Roux, C. (2020). DILS : Demographic Inferences with Linked Selection by using ABC. *bioRxiv*, page 2020.06.15.151597.

Fuller, Z. L., Mocellin, V. J. L., Morris, L. A., Cantin, N., Shepherd, J., Sarre, L., Peng, J., Liao, Y., Pickrell, J., Andolfatto, P., Matz, M., Bay, L. K., and Przeworski, M. (2020). Population genetics of the coral Acropora millepora: Toward genomic prediction of bleaching. *Science*, 369(6501).

Grant, P. R., Grant, B. R., and Petren, K. (2005). Hybridization in the Recent Past. *The American Naturalist*, 166(1):56–67.

Green, R. E., Krause, J., Briggs, A. W., Maricic, T., Stenzel, U., Kircher, M., Patterson, N., Li, H., Zhai, W., Fritz, M. H.-Y., Hansen, N. F., Durand, E. Y., Malaspinas, A.-S., Jensen, J. D., Marques-Bonet, T., Alkan, C., Prüfer, K., Meyer, M., Burbano, H. A., Good, J. M., Schultz, R., Aximu-Petri, A., Butthof, A., Höber, B., Höffner, B., Siegemund, M., Weihmann, A., Nusbaum, C., Lander, E. S., Russ, C., Novod, N., Affourtit, J., Egholm, M., Verna, C., Rudan, P., Brajkovic, D., Kucan, Ž., Gušic, I., Doronichev, V. B., Golovanova,

L. V., Lalueza-Fox, C., de la Rasilla, M., Fortea, J., Rosas, A., Schmitz, R. W., Johnson, P. L. F., Eichler, E. E., Falush, D., Birney, E., Mullikin, J. C., Slatkin, M., Nielsen, R., Kelso, J., Lachmann, M., Reich, D., and Pääbo, S. (2010). A Draft Sequence of the Neandertal Genome. *Science*, 328(5979):710–722.

Haenel, Q., Laurentino, T. G., Roesti, M., and Berner, D. (2018). Meta-analysis of chromosome-scale crossover rate variation in eukaryotes and its significance to evolutionary genomics. *Molecular Ecology*, 27(11):2477–2497.

Hamlin, J. A. P., Hibbins, M. S., and Moyle, L. C. (2020). Assessing biological factors affecting postspeciation introgression. *Evolution Letters*, 4(2):137–154.

Harr, B. (2006). Genomic islands of differentiation between house mouse subspecies. *Genome Research*, 16(6):730–737.

Harrison, R. G. and Larson, E. L. (2016). Heterogeneous genome divergence, differential introgression, and the origin and structure of hybrid zones. *Molecular ecology*, 25(11):2454–2466.

Hermann, K., Klahre, U., Moser, M., Sheehan, H., Mandel, T., and Kuhlemeier, C. (2013). Tight Genetic Linkage of Prezygotic Barrier Loci Creates a Multifunctional Speciation Island in Petunia. *Current Biology*, 23(10):873–877.

Hewitt, G. (2000). The genetic legacy of the Quaternary ice ages. *Nature*, 405(6789):907–913.

Hewitt, G. M. (1988). Hybrid zones-natural laboratories for evolutionary studies. *Trends in Ecology & Evolution*, 3(7):158–167.

Hohenlohe, P. A., Bassham, S., Etter, P. D., Stiffler, N., Johnson, E. A., and Cresko, W. A. (2010). Population Genomics of Parallel Adaptation in Threespine Stickleback using Sequenced RAD Tags. *PLOS Genetics*, 6(2):e1000862.

Hubisz, M. J., Williams, A. L., and Siepel, A. (2020). Mapping gene flow between ancient hominins through demography-aware inference of the ancestral recombination graph. *PLOS Genetics*, 16(8):e1008895.

Hudson, R. R., Slatkin, M., and Maddison, W. P. (1992). Estimation of levels of gene flow from DNA sequence data. *Genetics*, 132(2):583–589.

Johannesson, K., Moan, A. L., Perini, S., and André, C. (2020). A Darwinian Laboratory of Multiple Contact Zones. *Trends in Ecology & Evolution*, 35(11):1021–1036.

Jombart, T., Kendall, M., Almagro-Garcia, J., and Colijn, C. (2017). Treespace: Statistical exploration of landscapes of phylogenetic trees. *Molecular Ecology Resources*, 17(6):1385–1392.

Jones, G. P., Milicich, M. J., Emslie, M. J., and Lunow, C. (1999). Self-recruitment in a coral reef fish population. *Nature*, 402(6763):802–804.

Katoh, K., Rozewicki, J., and Yamada, K. D. (2019). MAFFT online service: Multiple sequence alignment, interactive sequence choice and visualization. *Briefings in Bioinformatics*, 20(4):1160–1166.

Keightley, P. D. and Jackson, B. C. (2018). Inferring the Probability of the Derived vs. the Ancestral Allelic State at a Polymorphic Site. *Genetics*, 209(3):897–906.

Kelleher, J., Etheridge, A. M., and McVean, G. (2016). Efficient Coalescent Simulation and Genealogical Analysis for Large Sample Sizes. *PLOS Computational Biology*, 12(5):e1004842.

Kelleher, J., Wong, Y., Wohns, A. W., Fadil, C., Albers, P. K., and McVean, G. (2019). Inferring whole-genome histories in large population datasets. *Nature Genetics*, 51(9):1330–1338.

Korunes, K. L. and Samuk, K. (2021). Pixy: Unbiased estimation of nucleotide diversity and divergence in the presence of missing data. *Molecular Ecology Resources*, 21(4):1359–1368.

Lemaire, C., Versini, J.-J., and Bonhomme, F. (2005). Maintenance of genetic differentiation across a transition zone in the sea: Discordance between nuclear and cytoplasmic markers. *Journal of Evolutionary Biology*, 18(1):70–80.

Li, H. and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics (Oxford, England)*, 25(14):1754–1760.

Louro, B., De Moro, G., Garcia, C., Cox, C. J., Veríssimo, A., Sabatino, S. J., Santos, A. M., and Canário, A. V. M. (2019). A haplotype-resolved draft genome of the European sardine (Sardina pilchardus). *GigaScience*, 8(5).

Macpherson, E. and Raventos, N. (2006). Relationship between pelagic larval duration and geographic distribution of Mediterranean littoral fishes. *Marine Ecology Progress Series*, 327:257–265.

Malinsky, M., Matschiner, M., and Svardal, H. (2021). Dsuite - Fast D-statistics and related admixture evidence from VCF files. *Molecular Ecology Resources*, 21(2):584–595.

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

Mank, J. E. and Avise, J. C. (2006). Phylogenetic conservation of chromosome numbers in Actinopterygiian fishes. *Genetica*, 127(1):321–327.

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

Martin, M., Patterson, M., Garg, S., Fischer, S. O., Pisanti, N., Klau, G. W., Schöenhuth, A., and Marschall, T. (2016). WhatsHap: Fast and accurate read-based phasing.

Martin, S. H., Dasmahapatra, K. K., Nadeau, N. J., Salazar, C., Walters, J. R., Simpson, F., Blaxter, M., Manica, A., Mallet, J., and Jiggins, C. D. (2013). Genome-wide evidence for speciation with gene flow in Heliconius butterflies. *Genome Research*, 23(11):1817–1828.

Martin, S. H., Davey, J. W., and Jiggins, C. D. (2015). Evaluating the Use of ABBA–BABA Statistics to Locate Introgressed Loci. *Molecular Biology and Evolution*, 32(1):244–257.

Matute, D. R. and Cooper, B. S. (2021). Comparative studies on speciation: 30 years since Coyne and Orr. *Evolution; International Journal of Organic Evolution*, 75(4):764–778.

McEntee, J. P., Burleigh, J. G., and Singhal, S. (2020). Dispersal Predicts Hybrid Zone Widths across Animal Diversity: Implications for Species Borders under Incomplete Reproductive Isolation. *The American Naturalist*, 196(1):9–28.

Miles, A., Bot, P., Murillo, R., Ralph, P., Harding, N., Pisupati, R., Rae, S., and Millar, T. (2020). Cggh/scikit-allel: V1.3.2. Zenodo.

Minh, B. Q., Schmidt, H. A., Chernomor, O., Schrempf, D., Woodhams, M. D., von Haeseler, A., and Lanfear, R. (2020). IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Molecular Biology and Evolution*, 37(5):1530–1534.

Muff, S., Nilsen, E. B., O'Hara, R. B., and Nater, C. R. (2021). Rewriting results sections in the language of evidence. *Trends in Ecology & Evolution*, 0(0).

Nakayama, I., Foresti, F., Tewari, R., Schartl, M., and Chourrout, D. (1994). Sex chromosome polymorphism and heterogametic males revealed by two cloned DNA probes in the ZW/ZZ fish Leporinus elongatus. *Chromosoma*, 103(1):31–39.

Nei, M. (1975). Molecular population genetics and evolution. *Frontiers of Biology*, 40:I–288.

Nei, M. and Li, W. H. (1979). Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proceedings of the National Academy of Sciences of the United States of America*, 76(10):5269–5273.

Nosil, P. (2012). *Ecological Speciation*. OUP Oxford.

Orr, H. and Turelli, M. (2001). Orr HA, Turelli M. The evolution of postzygotic isolation: Accumulating Dobzhansky-Muller incompatibilities. Evolution Int J Org Evolution 55: 1085-1094. *Evolution; international journal of organic evolution*, 55:1085–94.

Orr, H. A. (1995). The population genetics of speciation: The evolution of hybrid incompatibilities. *Genetics*, 139(4):1805–1813.

Owens, G. L. and Rieseberg, L. H. (2014). Hybrid incompatibility is acquired faster in annual than in perennial species of sunflower and tarweed. *Evolution; International Journal of Organic Evolution*, 68(3):893–900.

Pagel, M. (1999). Inferring the historical patterns of biological evolution. *Nature*, 401(6756):877–884.

Pascual, M., Rives, B., Schunter, C., and Macpherson, E. (2017). Impact of life history traits on gene flow: A multispecies systematic review across oceanographic barriers in the Mediterranean Sea. *PLOS ONE*, 12(5):e0176419.

Patarnello, T., Volckaert, F. a. M. J., and Castilho, R. (2007). Pillars of Hercules: Is the Atlantic–Mediterranean transition a phylogeographical break? *Molecular Ecology*, 16(21):4426–4444.

Patterson, N., Moorjani, P., Luo, Y., Mallick, S., Rohland, N., Zhan, Y., Genschoreck, T., Webster, T., and Reich, D. (2012). Ancient Admixture in Human History. *Genetics*, 192(3):1065–1093.

Poplin, R., Ruano-Rubio, V., DePristo, M. A., Fennell, T. J., Carneiro, M. O., der Auwera, G. A. V., Kling, D. E., Gauthier, L. D., Levy-Moonshine, A., Roazen, D., Shakir, K., Thibault, J., Chandran, S., Whelan, C., Lek, M., Gabriel, S., Daly, M. J., Neale, B., MacArthur, D. G., and Banks, E. (2018). Scaling accurate genetic variant discovery to tens of thousands of samples. *bioRxiv*, page 201178.

Powell, D. L., García-Olazábal, M., Keegan, M., Reilly, P., Du, K., Díaz-Loyo, A. P., Banerjee, S., Blakkan, D., Reich, D., Andolfatto, P., Rosenthal, G. G., Schartl, M., and Schumer, M. (2020). Natural hybridization reveals incompatible alleles that cause melanoma in swordtail fish. *Science*, 368(6492):731–736.

Ramírez-Amaro, S., Ordines, F., Fricke, R., Ruiz-Jarabo, I., Bolado, I., and Massutí, E. (2021). Genetic and Morphological Evidence to Split the Coris julis Species Complex (Teleostei: Labridae) Into Two Sibling Species: Resurrection of Coris melanura (Lowe, 1839) Redescription of Coris julis (Linnaeus, 1758). *Frontiers in Marine Science*, 8:744639.

Reich, D., Thangaraj, K., Patterson, N., Price, A. L., and Singh, L. (2009). Reconstructing Indian population history. *Nature*, 461(7263):489–494.

Renaut, S., Grassa, C. J., Yeaman, S., Moyers, B. T., Lai, Z., Kane, N. C., Bowers, J. E., Burke, J. M., and Rieseberg, L. H. (2013). Genomic islands of divergence are not affected by geography of speciation in sunflowers. *Nature Communications*, 4:1827.

Revell, L. J. (2012). Phytools: An R package for phylogenetic comparative biology (and other things). *Methods in Ecology and Evolution*, 3(2):217–223.

Riquet, F., Liautard-Haag, C., Woodall, L., Bouza, C., Louisy, P., Hamer, B., Otero-Ferrer, F., Aublanc, P., Béduneau, V., Briard, O., El Ayari, T., Hochscheid, S., Belkhir, K., Arnaud-Haond, S., Gagnaire, P.-A., and Bierne, N. (2019). Parallel pattern of differentiation at a genomic island shared between clinal and mosaic hybrid zones in a complex of cryptic seahorse lineages. *Evolution*, 73(4):817–835.

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

Romiguier, J., Gayral, P., Ballenghien, M., Bernard, A., Cahais, V., Chenuil, A., Chiari, Y., Dernat, R., Duret, L., Faivre, N., Loire, E., Lourenco, J. M., Nabholz, B., Roux, C., Tsagkogeorga, G., a. T. Weber, A., Weinert, L. A., Belkhir, K., Bierne, N., Glémin, S., and Galtier, N. (2014). Comparative population genomics in animals uncovers the determinants of genetic diversity. *Nature*, 515(7526):261–263.

Rossum, G. V. and Drake, F. L. (2009). *Python 3 Reference Manual:*. CreateSpace Independent Publishing Platform, Hampton, NH.

Roux, C., Fraïsse, C., Romiguier, J., Anciaux, Y., Galtier, N., and Bierne, N. (2016). Shedding Light on the Grey Zone of Speciation along a Continuum of Genomic Divergence. *PLOS Biology*, 14(12):e2000234.

Russell, S. T. (2003). Evolution of intrinsic post-zygotic reproductive isolation in fish. *Annales Zoologici Fennici*, 40(4):321–329.

Samuk, K., Owens, G. L., Delmore, K. E., Miller, S. E., Rennison, D. J., and Schluter, D. (2017). Gene flow and selection interact to promote adaptive divergence in regions of low recombination. *Molecular Ecology*, 26(17):4378–4390.

Schartl, M., Walter, R. B., Shen, Y., Garcia, T., Catchen, J., Amores, A., Braasch, I., Chalopin, D., Volff, J.-N., Lesch, K.-P., Bisazza, A., Minx, P., Hillier, L., Wilson, R. K., Fuerstenberg, S., Boore, J., Searle, S., Postlethwait, J. H., and Warren, W. C. (2013). The genome of the platyfish, Xiphophorus maculatus, provides insights into evolutionary adaptation and several complex traits. *Nature Genetics*, 45(5):567–572.

Schumer, M., Rosenthal, G. G., and Andolfatto, P. (2014). How Common Is Homoploid Hybrid Speciation? *Evolution*, 68(6):1553–1560.

Schunter, C., Carreras-Carbonell, J., Macpherson, E., Tintoré, J., Vidal-Vijande, E., Pascual, A., Guidetti, P., and Pascual, M. (2011). Matching genetics with oceanography: Directional gene flow in a Mediterranean fish species. *Molecular Ecology*, 20(24):5167–5181.

Sedghifar, A., Brandvain, Y., and Ralph, P. (2016). Beyond clines: Lineages and haplotype blocks in hybrid zones. *Molecular Ecology*, 25(11):2559–2576.

Selkoe, K. and Toonen, R. (2011). Marine connectivity: A new look at pelagic larval duration and genetic metrics of dispersal.

Selkoe, K. A., Gaggiotti, O. E., Laboratory, T., Bowen, B. W., and Toonen, R. J. (2014). Emergent patterns of population genetic structure for a coral reef community. *Molecular Ecology*, 23(12):3064–3079.

Servedio, M. R. (2004). The evolution of premating isolation: Local adaptation and natural and sexual selection against hybrids. *Evolution; International Journal of Organic Evolution*, 58(5):913–924.

Shanks, A. L. (2009). Pelagic Larval Duration and Dispersal Distance Revisited. *The Biological Bulletin*, 216(3):373–385.

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

Sobel, J. M. and Chen, G. F. (2014). Unification of Methods for Estimating the Strength of Reproductive Isolation. *Evolution*, 68(5):1511–1522.

Stamatakis, A. (2014). RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics (Oxford, England)*, 30(9):1312–1313.

Stankowski, S. and Ravinet, M. (2021). Defining the speciation continuum. *Evolution*, 75(6):1256–1273.

Swearer, S. E., Caselle, J. E., Lea, D. W., and Warner, R. R. (1999). Larval retention and recruitment in an island population of a coral-reef fish. *Nature*, 402(6763):799–802.

Tine, M., Kuhl, H., Gagnaire, P.-A., Louro, B., Desmarais, E., Martins, R. S. T., Hecht, J., Knaust, F., Belkhir, K., Klages, S., Dieterich, R., Stueber, K., Piferrer, F., Guinand, B., Bierne, N., Volckaert, F. A. M., Bargelloni, L., Power, D. M., Bonhomme, F., Canario, A. V. M., and Reinhardt, R. (2014). European sea bass genome and its variation provide insights into adaptation to euryhalinity and speciation. *Nature Communications*, 5:5770.

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

Van Belleghem, S. M., Vangestel, C., De Wolf, K., De Corte, Z., Möst, M., Rastas, P., De Meester, L., and Hendrickx, F. (2018). Evolution at two time frames: Polymorphisms from an ancient singular divergence event fuel contemporary parallel evolution. *PLoS genetics*, 14(11):e1007796.

Verzijden, M. N., ten Cate, C., Servedio, M. R., Kozak, G. M., Boughman, J. W., and Svensson, E. I. (2012). The impact of learning on sexual selection and speciation. *Trends in Ecology & Evolution*, 27(9):511–519.

Via, S. and West, J. (2008). The genetic mosaic suggests a new role for hitchhiking in ecological speciation. *Molecular Ecology*, 17(19):4334–4345.

Weir, B. S. and Cockerham, C. C. (1984). Estimating F-Statistics for the Analysis of Population Structure. *Evolution*, 38(6):1358–1370.

Weisenfeld, N. I., Kumar, V., Shah, P., Church, D. M., and Jaffe, D. B. (2017). Direct determination of diploid genome sequences. *Genome Research*, 27(5):757–767.

Westram, A. M., Rafajlović, M., Chaube, P., Faria, R., Larsson, T., Panova, M., Ravinet, M., Blomberg, A., Mehlig, B., Johannesson, K., and Butlin, R. (2018). Clines on the seashore: The genomic architecture underlying rapid divergence in the face of gene flow. *Evolution Letters*, 2(4):297–309.

Wickham, H. (2016). *Ggplot2: Elegant Graphics for Data Analysis*. Use R! Springer International Publishing : Imprint: Springer, Cham, 2nd ed. 2016 edition.

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T., Miller, E., Bache, S., Müller, K., Ooms, J., Robinson, D., Seidel, D., Spinu, V., Takahashi, K., Vaughan, D., Wilke, C., Woo, K., and Yutani, H. (2019). Welcome to the Tidyverse. *Journal of Open Source Software*, 4(43):1686.

Wohns, A. W., Wong, Y., Jeffery, B., Akbari, A., Mallick, S., Pinhasi, R., Patterson, N., Reich, D., Kelleher, J., and McVean, G. (2021). A unified genealogy of modern and ancient genomes.

Wright, S. (1931). Evolution in Mendelian Populations. *Genetics*, 16(2):97–159.

Yeaman, S., Aeschbacher, S., and Bürger, R. (2016). The evolution of genomic islands by increased establishment probability of linked alleles. *Molecular ecology*, 25(11):2542–2558.

Yu, G. (2020). Using ggtree to Visualize Data on Tree-Like Structures. *Current Protocols in Bioinformatics*, 69(1).

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

# Chapitre 3

# Corrélation des paysages de divergence mitochondriaux chez les poissons marins

**Résumé**

Comprendre comment les forces évolutives modèlent les génomes des espèces durant la spéciation est nécessaire pour isoler les causes qui permettent l'isolement reproductif entre des lignées divergentes. Chaque cellule eucaryote possède plusieurs mitochondries qui sont le résultat d'une fusion ancestrale entre un eucaryote et une bactérie. Les mitochondries possèdent un court génome circulaire de 16 000 paires de bases qui codent pour 13 protéines, 2 ARN ribosomaux (ARN-r) et 22 ARN transfert (ARN-t). A cause de son haploïdie et de sa transmission uniquement maternelle, le génome mitochondrial est supposé être soumis à une intensité de la dérive quatre fois plus forte que le génome nucléaire (Ballard and Whitlock, 2004). Ainsi, la diversité est supposée être réduite d'un quart chez le génome mitochondrial par rapport au génome nucléaire ; de manière similaire, toutes les forces évolutives impactées par la taille efficace de la population (tri de lignées, sélection en liaison) sont censées agir de manière différente sur le génome mitochondriale.

Les marqueurs génétiques mitochondriaux ont été très utilisés dans les études phylogéographiques et génétiques pour inférer l'histoire évolutive des espèces. La facilité d'amplification combinée à sa sensibilité aux évènements récents dû à son taux d'évolution rapide (Allio et al., 2017) les rendent particulièrement attrayants pour retracer la démographie récente des populations et des espèces. Par exemple, les périodes d'allopatrie dans des refuges glaciaires du Quaternaire suivie d'un contact secondaire dans une grande variabilité d'espèces terrestres et marines ont pu être inférées à partir d'une observation d'une divergence de deux ou plusieurs haplogroupes mitochondriaux suivies (ou non) d'un re-mélange entre différents populations refuges (Avise et al., 1987). De plus, les génomes mitochondriaux sont supposées être également directement impliquées dans les barrières à l'isolement reproductif par la mise en place d'incompatibilités mito-nucléaires (Burton and Barreto, 2012). Les protéines codées par le génome mitochondrial sont impliquées dans le métabolisme de phosphorylation oxydative (OXPHOS) en interaction avec des protéines codées par le génome nucléaire. Une coévolution différentielle des gènes des deux génomes entre plusieurs lignées évolutives pourrait entraîner une incompatibilité des génomes mitochondriaux et nucléaires chez les hybrides, diminuer le flux de gène efficace entre lignées et favoriser la mise en place de barrières évolutives. La compréhension de la prévalence de chacune des différents mécanismes évolutifs agissant sur les génomes mitochondriaux restent cependant encore à élucider.

Dans ce dernier chapitre, nous avons assemblé et annoté *de novo* 380 génomes mitochondriaux complets de 20 espèces subdivisées en deux lignées Atlantique et Méditerrannéene. Puis, nous avons analysé les patrons de polymorphisme, de différentiation et de divergence intraspécifique sur le génome mitochondrial et sur chaque gène séparément. Nous avons montré une grande variabilité de structure mitochondrial allant de la monophilie réciproque entre populations Atlantique et Méditerranée, à forte divergence pour la girelle (*C. julis*), ou à faible

divergence pour le gobie noir (*G. niger*), une monophilie réciproque incomplète avec la présence d'un ou deux individus possédant une mitochondrie caractéristique de l'autre lignée comme pour le marbré (*L. mormyrus*), un remélange de deux haplogroupes mitochondriaux divergents pour le serran chevrette (*S. cabrilla*), une absence de structure génétique comme pour le merlu (*M. merluccius*) ou des structures plus complexes comme pour l'hippocampe moucheté (*H. guttulatus*). Par une comparaison avec les données de polymorphisme nucléaire, nous avons montré une diversité mitochondriale plus faible n'atteignant pas, toutefois, la réduction de 3/4 attendue d'après les différences du génome mitochondrial (haploïdie et une transmission uniquement maternelle). Nous avons aussi observé des niveaux de divergence et de différentiation plus élevés que sur le génome nucléaire pour les espèces très structurées. En comparant les variations intragénomiques de polymorphisme au sein de chaque espèce, nous avons trouvé que la diversité et la divergence des gènes mitochondriaux étaient corrélées négativement au ratio de sites non-synonymes (ceux dont le changement de bases entraîne un changement de l'acide aminé) par rapport au nombre de sites synonymes (ceux dont un changement de base ne change pas la structure de la protéine). Ceci est conforme à la grande contrainte des gènes mitochondriaux : un changement de la structure protéique entraînerait une baisse de l'efficacité du métabolisme OXPHOS dans les membranes mitochondriales et ainsi une réduction de la valeur sélective individuelle. Les mutations non-synonymes seraient ainsi rapidement supprimées et la diversité et la divergence ne se baserait principalement que sur les sites synonymes.

Au contraire, les niveaux de différentiation génétique ne semblent pas être corrélées à la fréquence de sites synonymes. Les mécanismes évolutifs qui permettent d'empêcher le remélange d'haplogroupes mitochondriaux divergents restent donc encore à élucider. L'analyse de la divergence et la résistance à l'introgression des gènes nucléaires impliqués dans le métabolisme OXPHOS permettrait de tester si les incompatibilités mito-nucléaires permettent d'expliquer le maintien de la différentiation mitochondriale chez certaines espèces.

# References

Allio, R., Donega, S., Galtier, N., and Nabholz, B. (2017). Large Variation in the Ratio of Mitochondrial to Nuclear Mutation Rate across Animals: Implications for Genetic Diversity and the Use of Mitochondrial DNA as a Molecular Marker. *Molecular Biology and Evolution*, 34(11):2762–2772.

Avise, J. C., Arnold, J., Ball, R. M., Bermingham, E., Lamb, T., Neigel, J. E., Reeb, C. A., and Saunders, N. C. (1987). Intraspecific Phylogeography: The Mitochondrial DNA Bridge Between Population Genetics and Systematics. *Annual Review of Ecology and Systematics*, 18:489–522.

Ballard, J. W. O. and Whitlock, M. C. (2004). The incomplete natural history of mitochondria. *Molecular Ecology*, 13(4):729–744.

Burton, R. S. and Barreto, F. S. (2012). A disproportionate role for mtDNA in Dobzhansky–Muller incompatibilities? *Molecular Ecology*, 21(20):4942–4957.

# Correlated landscapes of mitochondrial and divergence in marine fishes

Pierre Barry[1], Thomas Broquet[2], Pierre-Alexandre Gagnaire[1]

[1]ISEM, Univ Montpellier, CNRS, EPHE, IRD, Montpellier, France.

[2]UMR 7144, Station Biologique de Roscoff, CNRS & Sorbonne Université, Roscoff, France

## Abstract

Analyzing how evolutionary forces affect the genomes during speciation is mandatory to understand how species arise. Due to its haploidy and only maternal inheritance, the mitochondrial genome is supposed to experience a stronger intensity of drift that can drive the differentiation, divergence and accumulation of barriers to gene flow notably through mito-nuclear incompatibilities. Comparing the patterns and drivers of genetic polymorphism in mitochondrial genomes of different species might evaluate the relative importance of each of these forces to shape the species differentiation and divergence landscapes. Here, we *de novo* assembled and annotated the mitochondrial genome of 380 individuals of 20 marine teleostan fish species subdivided in two Atlantic and Mediterranean lineages and evaluated the levels of diversity, differentiation and divergence at the whole-mitochondrial genome and within each gene. We found a considerable variation in mitochondrial population structure between Atlantic and Mediterranean populations from complete reciprocal monophyly to the absence of genetic differentiation. Despite this variation, within-species in diversity and divergence between genes is strongly explained by its frequency of synonymous sites suggesting the importante of genetic constraints in shaping similar divergence landscapes between independent species. In contrast, the variation of genetic differentiation between genes is not explained by this frequency. The potential presence of putative mito-nuclear incompatibilities in these species remains to be assessed to better explain the establishment and the maintenance of reproductive barriers during speciation.


**Key words**: speciation, mitochondrial genome, Atlantic - Mediterranean suture zone, divergence, $\pi_n/\pi_s$

# Introduction

Understanding the biogeographic history that affects the formation of evolutionary lineages is an important step to understand how species arise. Examining the evolutionary forces that shaped lineages' divergence through time and on genomes with different characteristics is a necessary prerequisite to approach (Avise et al., 1987; Godinho et al., 2008). Eukaryotes species have a mitochondrial genome that originates from an ancient fusion with a bacteria. Each eukaryote cells generally have a unique circular mitochondrial genome of approximately 16 500 base pairs that is essentially composed of coding sequences with 13 protein-coding genes, 2 ribosomal r-RNAs and 22 transfer RNAs (t-RNAs). The vertical transmission of the mitochondrial genome is most commonly insured through maternal inheritance (but see double-parent inheritance in bivalves and *Drosophila* flies, Zouros et al. (1994) and Kondo et al. (1990)). Eukaryote cells have a haploid mitochondrial genome, which implies a four-fold reduction in the mitochondrial effective population size as compared to nuclear genes in diploid species. Therefore, the mitochondrial genome does not recombine and experiences stronger genetic drift as compared to the nuclear genome (Ballard and Whitlock, 2004). This particular mode of evolution makes the mitochondrial genome more sensitive to the depletion of genetic diversity due to linked selection because, in the absence of recombination, selective sweeps and background selection provoke hitchhiking of the whole-mitochondrial genome. Finally, mitochondrial genomes have higher mutations rates compared to nuclear genes that can reach up to a 50 fold difference (Allio et al., 2017). Thus, any genetic characteristics that depend on $N_e$, such as genetic diversity ($\pi$), differentiation ($F_{ST}$), divergence ($d_{XY}$) and genetic load ($\pi_n/\pi_s$ ratio) should show differences compared to nuclear genome values.

Mitochondrial markers have been widely used in phylogeographic studies to assess species' evolutionary history because of high evolutionary rates that help record recent evolutionary changes. Patterns of genetic polymorphism, differentiation and divergence of the mitochondrial genome between populations can help to understand how contemporary patterns of the population genetic structure reflect the footprint of past evolutionary events such as those imposed by quaternary climate oscillations (Hewitt, 2000). For example, historical population separation can be detected by the presence of divergent haplotypes that fixed different mutations in allopatry. Upon secondary contact, if gene flow occurs, the two non-recombining haplotypes should display a discordant phylogeographic signal with respect to geography.

However, the mitochondrial genome can also be directly involved in the speciation process because, for example, of mitochondrial-nuclear genes incompatibilities. Both mitochondrial and nuclear genes code for proteins that interact in the mitochondrial oxidative phosphorylation (OXPHOS) metabolism that is the main source of energy for eukaryote cells. As the mitochondrial genome is more prone to fixation slightly deleterious mutations fixation because of its lower effective size and absence of recombination, this can elicit the fixation of compensatory mutations in the nuclear genes coding for proteins that interact with mitochondrial proteins (Rand et al., 2004; Oliveira et al., 2008). If different populations evolved different sets of co-adapted mitochondrial and nuclear alleles, hybrids might experience higher reduction in fitness and drive the evolution of post-zygotic reproductive isolation barriers. Mito-nuclear incompatibilities has been discovered in *Nasonia* wasps (Niehuis et al., 2008), Atlantic eels (Gagnaire et al., 2012) and *Tigriopus* copepods (Burton and Barreto, 2012).

Here, we sequenced, assembled and annotated population-scale whole-mitochondrial genomes in 20 marine teleostan fish species showing more or less pronounced subdivisions in two phylogeographical lineages associated with the Atlantic Ocean and the Mediterranean Sea. We analyzed patterns of genetic diversity, divergence and differentiation both at the whole-mitochondrial

genome scale and within each coding gene and compared these estimates with those previously obtained from whole-nuclear genome sequences. Our objectives were to i) describe the distribution of the genetic variation on mitochondrial genomes with regards to the Atlantic - Mediterranean transition zone, ii) assess whether mitochondrial genetic diversity is conformed to neutral expectations based solely on the difference in effective population size of the mitochondrial and nuclear genomes, iii) explore the potential role of differential selection in shaping mitochondrial genetic variation, and iv) investigate the tendency of the mitochondrial genome to be frequently involved in barriers to gene flow between partially reproductively isolated populations.

# Material and Methods

## Sampling and sequencing

We sampled 397 individuals of 20 marine teleostan fish species with similar geographic distributions across the northeastern Atlantic Ocean and the Mediterranean Sea. All these species show evidence from the published literature of genetic variation between Atlantic and Mediterranean populations, with various degrees of molecular differentiation. We sampled 5 individuals per species in the same four different locations: two within the Atlantic - Mediterranean transition zone (Algarve region in Portugal, Atl-in; Costa Calida region around Mar Menor, Med-in for the Atlantic Ocean and the Mediterranean Sea, respectively), and two outside this zone (the Bay of Biscay, Atl-out; the Gulf of Lion, Med-out for the Atlantic Ocean and the Mediterranean Sea respectively). DNA extraction and whole-genome sequencing were performed with the same protocol for all individuals (see Chapter 2 for further details).

## Whole-mitochondrial genome assembly

We used the `MitoZ` pipeline to assemble and annotate the 400 individual whole-mitochondrial genomes from unassembled resequencing paired-end reads (Meng et al., 2019). We first used `Megahit`, usually used in metagenomics for its suitability to treat low-depth sequencing data (Li et al., 2015), to assemble both nuclear and mitochondrial reads *de novo* with the `assemble` module. Then, among all assembled contigs, we ran the `findmitoscaf` module to identify individual mitochondrial scaffolds. As `MitoZ` is able to assemble small fragments of mitochondrial genomes from low-depth reads originating from other species or other individuals due to slight contaminations, we carefully checked and discarded these fragments. Finally, we ran the `annotate` module to annotate the corresponding mitochondrial scaffold, searching for all Protein Coding Genes (PCGs), tRNA and rRNA genes.

Then, for each species, we used all individual mitochondrial scaffolds to generate a species multiple sequence alignment. We chose the best-assembled (i.e. fully annotated assembly) individual mitochondrial scaffold as a reference assembly for each species. Then, we ran the `Mitogenome_reorder.py` script retrieved from `MitoZ` Github (`https://github.com/linzhi2013/MitoZ`) to reorder each mitochondrial scaffold using the corresponding species reference. We concatenated each individual mitochondrial scaffolds into a single fasta file and aligned the 20 genomes of each species using `mafft` (Katoh et al., 2019). As `MitoZ` did not fully annotate every individual mitochondrial scaffold, we blasted each gene annotated by `MitoZ` from the reference individual to all individuals whole-mitochondrial scaffolds to get all possible individual gene sequences.

We compiled assembly statistics for each species and each gene-species mitochondrial alignment, including the length of the species' mitochondrial genome assembly, the number of biallelic and polyallelic sites, gaps in the alignment, and GC content.

## Population genetic statistics and phylogenetic tree

For each species, we estimated global ($\pi_{tot}$) and within-population ($\pi_{pop}$) genetic diversity, pairwise absolute divergence between populations ($d_{XY}$), and net divergence ($d_a$) (Hahn, 2018) with a custom python script using `Biopython` tools (Cock et al., 2009). We converted species fasta alignment file in a VCF and estimated pairwise Weir and Cockerham's $F_{ST}$ between populations (Weir and Cockerham, 1984) with `scikit-allel` (Miles et al., 2020). We estimated all these statistics for the whole-mitochondrial genome and for each gene taken separately. We also retrieved nuclear whole-genome estimations for each of these statistics from a previous study (see Chapter 2).

We trimmed species alignment using `trimAl` (Capella-Gutiérrez et al., 2009) and inferred mitochondrial phylogenetic tree with the `iqtree` (Minh et al., 2020) with `TN+F+I` evolution model. As previously, for each species, we inferred both the phylogenetic tree for the whole-mitochondrial alignment and for each gene separately.

## $\pi_n/\pi_s$ estimation

We estimated the number of synonymous and non-synonymous sites ($sites_s$ and $sites_n$) and nucleotide diversity at these sites ($\pi_n$ and $\pi_s$) using `DNAsp` (Rozas et al., 2017) following the Nei-Gojobori method (Nei and Gojobori, 1986). We inferred the $sites_n/sites_s$ and $\pi_n/\pi_s$ ratio for each gene and each species by discarding every individual in each gene assembly that contained too many gaps.

## Comparative landscapes of divergence and differentiation

To evaluate the potential drivers of landscapes of diversity, divergence and differentiation along the mitochondrial genome between species, we fitted the following mixed model:

$$S_{sp,g} = \alpha + \beta_1 L_{s,g} + \beta_2 PI_{s,g} + \beta_3 RS_{s,g} + species_{sp} + gaps_{sp,g} \quad (1)$$

with $S$ corresponding either to $F_{ST}$, $d_a$, $d_{XY}$ or $\pi_{pop}$, corresponding to species $sp$ and gene $g$, explained in a multiple linear regression including three explanatory genetic parameters, gene length ($L$), $\pi_n/\pi_s$ ratio ($PI$) and the ratio $sites_n/sites_s$ ($RS$) , taking into account species identity ($species$) and gaps in the alignment ($gaps$) as random factors.

## Statistical analyses and graphical illustration

All statistical analyses and figure visualization were carried out using `R-3.6.1` (R Core Team, 2018) and `python-3.7.6` (Rossum and Drake, 2009). We followed Muff et al. (2021) and addressed an equivalent level of evidence for each statistical test using $p - values$. All plots and graphical illustrations were created using `tidyverse v.1.3.0` (Wickham et al., 2019) and `ggplot2 v.3.3.2` R packages (Wickham, 2016). All scripts and interactive notebooks representing data and intermediate results are freely available at `https://github.com/pierrebarry/comparative_mitochondrial_fish`.

# Results

## Species whole-mitochondrial genome alignment

We assembled *de novo* 380 mitochondrial genomes from 20 marine teleostean fish species to generate whole-mitochondrial genome alignment for each species (Table S1) using `MitoZ`

(Meng et al., 2019), discarding only 17 individuals in which the mitochondrial assembly was not possible. Overall, the median assembly size was 16 910 bp, concordant with known vertebrate mitochondrial genome size (Satoh et al., 2016), with few gaps in the alignment ranging from 4 to 2 127 (median = 93 ± 592). Using previously estimated genomic statistics for the nuclear genome (Barry et al., 2022), we found no evidence of a correlation between mitochondrial GC content and nuclear GC content ($t$-test, $t = -0.557$, $p - value = 0.585$, Fig. S2).

We fully annotated each mitochondrial assembly, i.e., 13 protein-coding genes, two ribomosal RNAs (rRNAs) and 22 transfer RNAs (tRNAs) for each species except for 5 protein-coding genes for *A. boyeri*, *C. galerita*, *D. puntazzo*, *S. cinereus* (ND3,ND4,ND4L,ND5,ND6) and 1 protein-coding genes for *L. mormyrus* and *S. pilchardus* (ND6).

## Mitochondrial diversity, differentiation and divergence

Analyzing the whole-mitochondrial phylogenetic tree of each species, we empirically classified species into 6 groups characterised by particular genealogical patterns (Fig S1). First, we found 5 species with no population structure or any sign of haplogroup divergence discordant with geography (*D. puntazzo*, *M. merluccius*, *L. budegassa*, *S. cinereus*, *S. pilchardus*). Then, 4 species showed two divergent haplogroups discordant with geography (*S. cabrilla*, *M. surmuletus*, *P. erythrinus* and *S. sarda*). On the contrary, other species showed different groups of haplotypes concordant with geography, either with low divergent haplotypes and reciprocal monophyly (*G. niger*) or highly divergent haplotypes with uncomplete reciprocal monophyly - where nearly all individuals cluster within basins (*L. mormyrus* and *D. labrax*), or reciprocal monophyly (*C. julis*, *C. galerita* and *S. cantharus*). Finally, we also detected more complex population structure in 5 species (*H. guttulatus*, *S. typhle*, *A. boyeri*, *A. fallax* and *E. encrasicolus*). The distribution of mitochondrial genetic diversity among individuals and populations thus appeared to be extremely variable between species.

Total genetic diversity (Table S2) ranged from 0.16% for *H. guttulatus* to 8.05% for *A. boyeri* ($\pi_{total}$, median = 0.86% ± 1.83) and mean within-populations genetic diversity ranged from 0.065% for *C. julis* to 4.35% for *A. boyeri* ($\pi_{within}$, median = 0.52 ± 0.95).

Absolute genetic divergence between populations located outside the transition zone (outer populations, $d_{XY,out}$) ranged from 0.17% for *H. guttulatus* to 8.27% for *A. boyeri* ($d_{XY,out}$, median = 0.74% ± 2.48) and was very strongly correlated to $\pi_{total}$ ($t$-test, $t = 64.764$, $p - value < 2e^{-16}$, Table S4, Fig 2B, grey bars). Genetic divergence between populations within the suture zone (inner populations, $d_{XY,in}$) was strongly correlated to $d_{XY,out}$ ($t$-test, $t = 40.118$, $p - value < 2e^{-16}$, Fig 2B, yellow bars). Net genetic divergence between outer populations ranged from 0 for 6 species to 6.74% for *C. galerita* ($d_{a,out}$, median = 0.14% ± 2.22, Table S5, Fig 2C, gray bars) and was also strongly correlated to net divergence between inner populations ($t$-test, $t = 7.953$, $p - value = 3.96e^{-7}$).

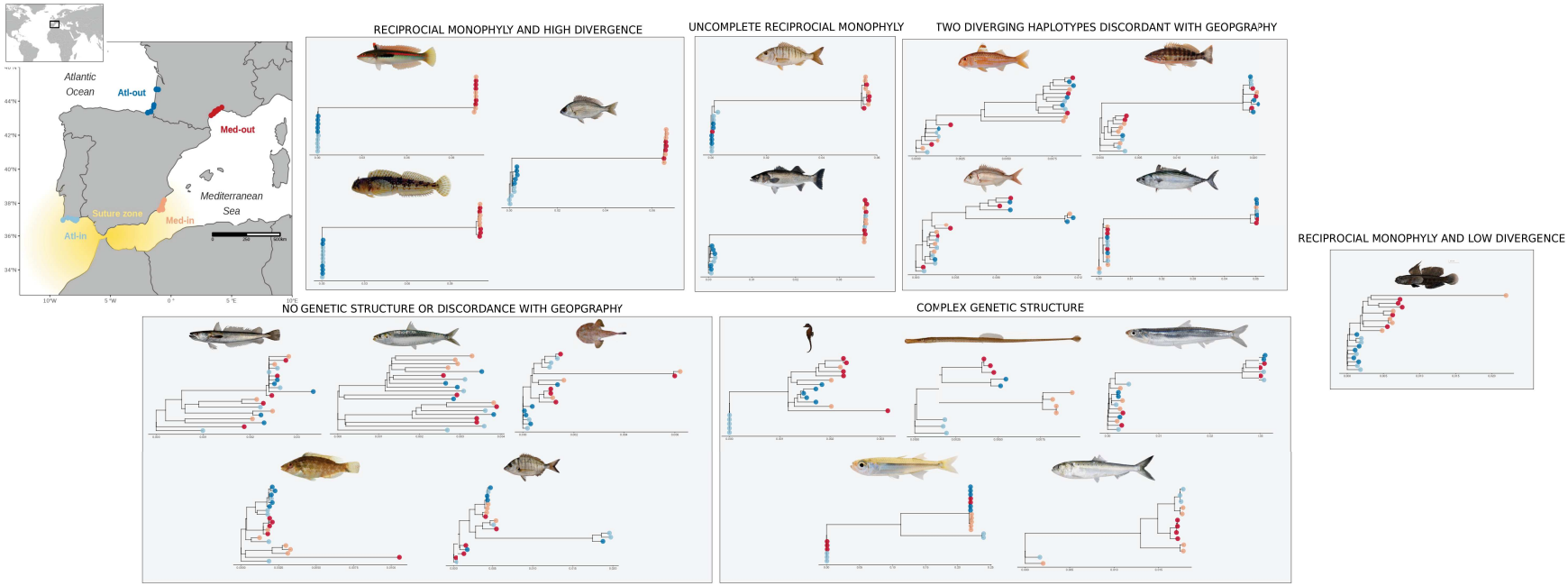



Figure 1: **Unrooted phylogenetic trees inferred from whole-mitochondrial genomes in 20 marine teleostan fish species**. 20 individuals per species were sampled in four locations in Atlantic Ocean (blue), and Mediterranean Sea (red), two within and two outside the Atlantic Ocean - Mediterranean Sea suture zone (area shown in yellow). We classified the 20 species in 6 groups based on empirical descriptions of phylogenetic trees: i) reciprocal monophyly with high haplogroup divergence for *C. julis*, *S. cantharus*, *C. galerita*, ii) incomplete reciprocal monophyly (*L. mormyrus*, *D. labrax*), iii) two divergent haplogroups discordant with geography (*M. surmuletus*, *S. cabrilla*, *P. erythrinus* and *S. sarda*), iv) no genetic structure (*M. merluccius, S. pilchardus, L. budegassa, S. cinereus, D. puntazzo*), v) reciprocal monophyly with low divergence (*G. niger*) and vi) more complex genetic structure (*H. guttulatus, S. typhle, E. encrasicolus, A. boyeri, A. fallax*).

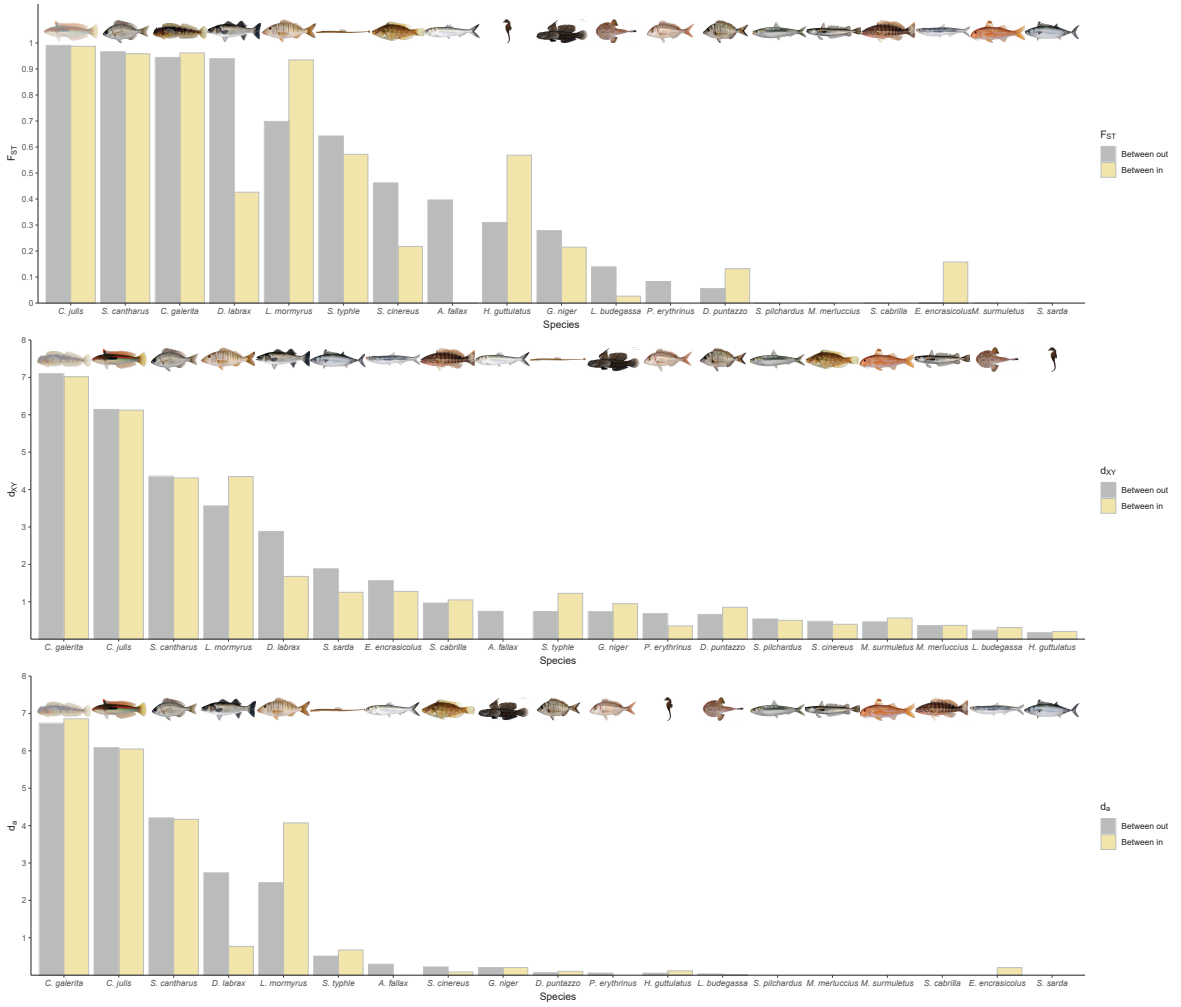



Figure 2: **Gradient of genetic differentiation ($F_{ST}$, top), absolute genetic divergence ($d_{XY}$, middle), and net genetic divergence ($d_a$, bottom) between populations located within (yellow) or outside (gray) of the suture zone.** Species are ordered to the high (left) to the low (right) genetic differentiation based on $F_{ST}$ between outer populations.

We found no genetic differentiation between outer populations for 6 species (*S. pilchardus*, *M. merluccius*, *S. cabrilla*, *E. encrasicolus*, *M. surmuletus*, *S. sarda*) (Table S3, Fig 2A, gray bars). Among the species that showed non-zero mitochondrial differentiation, $F_{ST}$ ranged from 0.056 for *D. puntazzo* to 0.99 for *C. julis* ($F_{ST,out}$,median $= 0.48 \pm 0.34$). Five out of the previous species and *P. erythrinus* showed also no genetic differentiation between their inner populations. Genetic differentiation was strongly correlated to absolute ($t$-test, $t = 10.349$ , $p - value < 2e^{-16}$) and net genetic divergence ($t$-test, $t = 17.811$, $p - value < 2e^{-16}$).

We retrieved genetic diversity, differentiation and divergence from the nuclear genome data of the studied species and compared it to mitochondrial estimates (Fig 3). Nuclear and mitochondrial genetic diversity were not correlated ($\pi$, $t$-test $= -0.12$, $R^2 = 0.000927$, $p - value = 0.91$). Most species showed lower mitochondrial compared to neutral genetic diversity but not as much as expected under a 1/4 fold reduction in effective population size for the mitochondrial genome. We found significant correlations between nuclear and mitochondrial genetic differentiation ($F_{ST}$, $t$-test $= 3.00$, $R^2 = 0.374$, $p - value = 0.009$), absolute ($d_{XY}$, $t$-test $= 3.19$, $R^2 = 0.404$, $p - value = 0.0061$) and net genetic divergence ($d_a$, $t$-test $= 2.68$, $R^2 = 0.323$, $p - value = 0.017$).

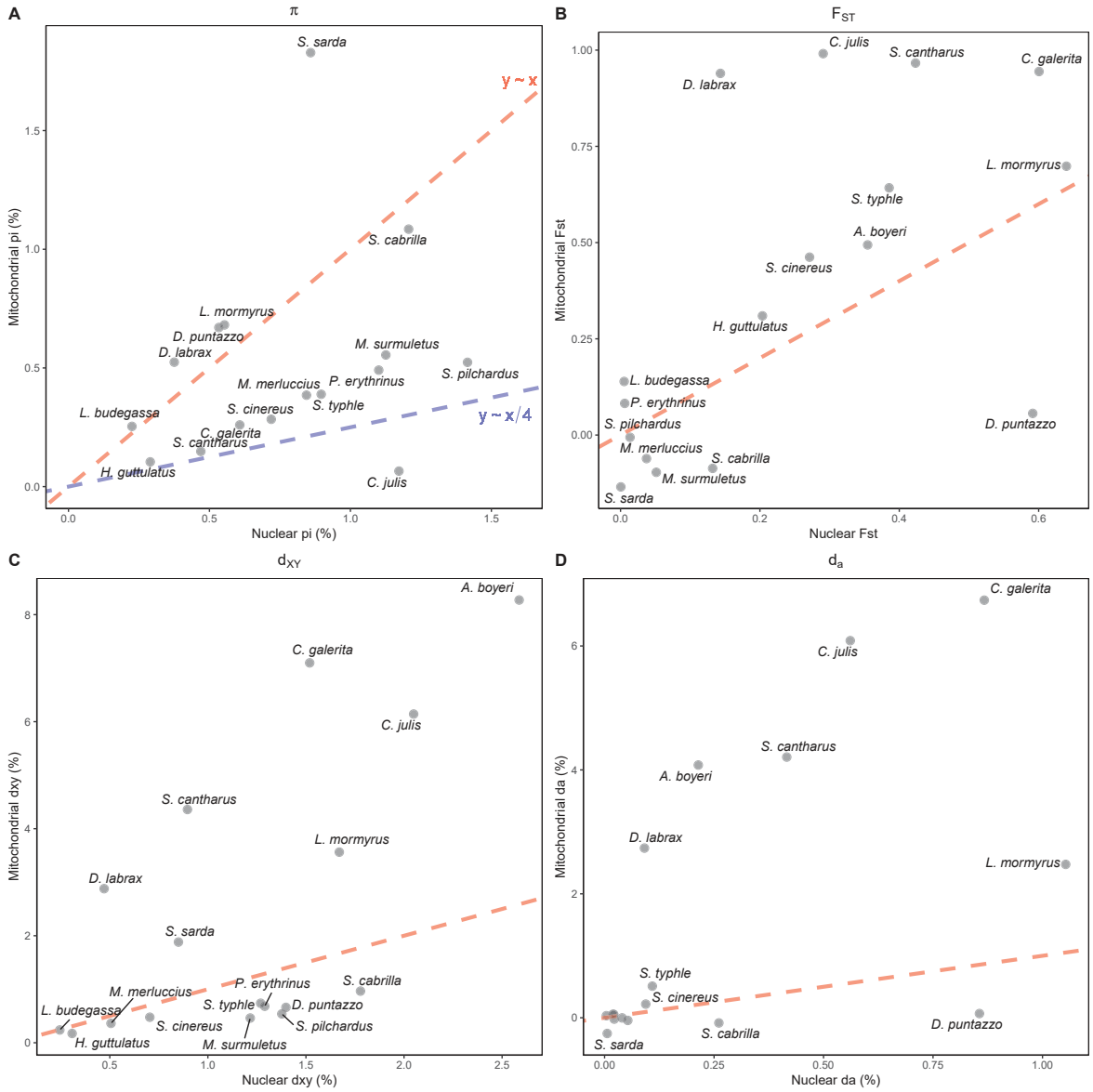



Figure 3: **Relationships between nuclear and mitochondrial total genetic diversity ($\pi_{pop}$, A), genetic differentiation ($F_{ST,out}$, B), absolute ($d_{XY,out}$, C) and net divergence ($d_{a,out}$, D).** Red lines show the $y = x$ model; in panel A, the relation $y = x/4$ that corresponds to a four-fold reduction of mitochondrial to nuclear genetic diversity is shown in blue.

## Comparative landscapes of divergence

We compared the levels of total genetic diversity ($\pi_{total}$), divergence and differentiation between outer populations at protein coding genes (PCGs) versus RNA genes. For 13 species, we found strong evidence for higher absolute genetic divergence at PCGs compared to RNA genes ($t$-test, $p - value$ ranging from $8.5e^{-3}$ to $9.65e^{-12}$), 4 species showed moderate evidence ($t$-test, $p - value$ ranging from 0.013 to 0.036) and 3 species showed no difference ($t$-test, all $p - value > 0.15$). Results were nearly identical for absolute genetic divergence. Considering genetic differentiation, we found strong evidence for higher differentiation at PCGs in 12 species ($t$-test, $p - value$ ranges from 0.0096 to $1.66e^{-9}$), 1 species showed moderate evidence ($t$-test, $p - value = 0.012$), while 7 species showed no evidence ($p - value > 0.088$).

We compared intragenomic variation in total genetic diversity, absolute and net divergence and differentiation by estimating the correlation between each of these statistics across all genes in each species pair (Table S6-S9). We found strong correlations for genetic diversity and both and absolute divergence except for *H. guttulatus*, *M. merluccius*, *S. cinereus* and *S. typhle*, but lower correlations for net genetic divergence and genetic differentiation acros all species.

To test whether gene length explained the correlated patterns of diversity and divergence across genes, we tested for an effect of gene length on all previous genetic statistics. We found strong evidence for a significant positive correlation between gene length and genetic diversity, net and absolute divergence and differentiation (all $p-values < 0.0001$), taking into account variation between species and missing data. However, this might be due to both differences in gene length and different levels of constraints between PCGS and t-RNAs, with t-RNA genes having on average stronger functional constraints. In agreement with this, we found no evidence of correlation for PCGs only ($\pi_{tot}$: $p-value = 0.4191$, $d_{XY}$: $p-value = 0.1686$, $d_a$: $p-value = 0.8938$ and $F_{ST}$: $p-value = 0.5482$).

## $\pi_n/\pi_s$ and correlated genomic landscapes

We found broad overall differences between species $\pi_n/\pi_s$ ratios (One-way ANOVA, F = 2.19, p-value = 0.00466) and between genes (One-way ANOVA, F = 3.9874, p-value = 1.549e-5).

Table 1: **Mixed models relating genetic summary statistics and gene covariables**. Mean genetic diversity within-populations ($\pi_{pop}$), absolute ($d_{XY}$) and net divergence ($d_a$) and genetic differentiation ($F_{ST}$) were fitted by a mixed model including the effect of gene length, $\pi_n/\pi_s$ and $sites_n/sites_s$ with corresponding slope ($\rho$) and $p-value$

| Statistic | Covariable | $\rho$ | $p-value$ |
|---|---|---|---|
| $\pi_{pop}$ | Length | - | 0.4486 |
| | $\pi_n/\pi_s$ | 0.031 | 0.0059 |
| | $sites_n/sites_s$ | -0.014 | 0.0257 |
| $d_{XY}$ | Length | - | 0.5756 |
| | $\pi_n/\pi_s$ | 0.0288 | 0.0351 |
| | $sites_n/sites_s$ | -0.025 | 0.0014 |
| $F_{ST}$ | Length | - | 0.4856 |
| | $\pi_n/\pi_s$ | - | 0.8846 |
| | $sites_n/sites_s$ | - | 0.1714 |
| $d_a$ | Length | - | 0.1901 |
| | $\pi_n/\pi_s$ | - | 0.8964 |
| | $sites_n/sites_s$ | -0.0127 | 0.0031 |

Taking into account variability between species and gaps in the alignment between genes, genetic diversity ($\pi_{tot}$) and absolute ($d_{XY}$) genetic divergence were positively and negatively correlated to $\pi_n/\pi_s$ and $sites_n/sites_s$, respectively, while net genetic divergence ($d_a$) was only negatively correlated to $sites_n/sites_s$. As the error variance associated with $\pi_n/\pi_s$ ratio estimates may be large because of low numbers of either non-synonymous and/or synonymous mutations, we did the same tests after removing genes with less than 150 non-synonymous sites: we found similar results except for $d_{XY}$, which became more strongly correlated to $\pi_n/\pi_s$ ($\rho = 0.083$, $p-value = 0.0001$) than $sites_n/sites_s$ ($\rho = -0.0242$, $p-value = 0.0042$). (Table 1). $F_{ST}$ between outer populations was not correlated to neither length, $\pi_n/\pi_s$ or $sites_n/sites_s$. However, as some species show complete homogeneization of haplogroups between outer populations, we did the same test but keeping only the 6 species having complete or incomplete reciprocal

monophyly between outer populations (*G. niger*,*D. labrax*,*C. julis*,*C. galerita*,*L. mormyrus*,*S. cantharus*): we found little evidence for a negative correlation between $F_{ST}$ and $\pi_n/\pi_s$ ratio ($\rho = -0.358$, $p - value = 0.0259$).

# Discussion

Here, we sequenced and assembled 380 whole-mitochondrial genomes from 20 marine teleostan fish species to document mitochondrial genome variation and molecular evolution patterns associated with haplogroup subdivisions between the Atlantic Ocean and the Mediterranean sea. On one hand, we described a great range of mitochondrial population structure, differentiation, diversity and divergence among the 20 species; on the other hand, we found that all mitochondrial diversity and divergence landscapes at protein-coding genes are all shaped by the number of synonymous sites. It is therefore important to disentangle the factors that similarly affect or not mitochondrial evolution patterns in independent species.

## Mitochondrial population structure and mito-nuclear discordance

Because of their different mutation rates, effective population sizes and exposure to linked selection, nuclear and mitochondrial genomes are expected to display different rates of evolution that result in contrasted genetic diversity and divergence patterns. Here, we showed that mitochondrial genomes can display higher differentiation and sometimes much higher molecular divergence as compared to genome-wide nuclear estimates for some species.

However, this finding mostly concerned species that are also highly differentiated on the nuclear genome (*C. galerita*, *L. mormyrus*, *A. boyeri*, *S. cantharus* and *C. julis*). Conversely, weak genetic structure on the nuclear genome was often reflected by low genetic structure on the mitochondrial genome. However, two species showed contrasting patterns: the European sea bass *D. labrax* displays high $F_{ST}$ and $d_{XY}$ on the mitochondrial genome despite moderate nuclear differentiation ($F_{ST} = 0.14$); on the contrary, we found that the sharp-snout sea bream, (*D. puntazzo*) shows no population structure on the mitochondrial genome but very high nuclear differentiation $F_{ST} = 0.58$. These mito-nuclear discordances could reflect the semipermeable nature of species boundaries before reproductive isolation is complete. In species like the European sea bass (*D. labrax*), the mitochondrial genome behaves like the minor fraction of the nuclear genome that resists gene flow, whereas, in the sea bream, it seems to be freely exchanged between Atlantic and Mediterranean lineages despite substantial reproductive isolation in that species. Although the mutation rate is generally higher in the mitochondrial genome, the large excess of divergence in comparison to the nuclear genome observed in some species could be also explained by the rejuvenation of nuclear divergence due to gene flow and recombination. By contrast, the non-recombining mitochondrial genome keeps track of the divergence history (although imprecisely due to coalescent stochasticity) as long as the divergent lineages are not lost by drift or selection. This seems to be the case in several species of our dataset, the most extreme example being in *S. sarda*, which has two divergent mitochondrial lineages segregating at similar frequencies despite a quasi absence of genetic differentiation in the nuclear genome. The frequent mito-nuclear discordances observed in our study thus reflect recent similar findings in European butterflies (Ebdon et al., 2021), showing that phylogeographical patterns and levels of mitochondrial divergence often poorly reflect nuclear divergence.

## Reduction of mitochondrial compared to nuclear genetic diversity

We found that the reduction in mitochondrial genetic diversity compared to the nuclear genome diversity is lower than the 4 fold reduction predicted by its haploidy and maternal inheritance.

This is also not coherent with the prediction of depletion of genetic diversity expected from the strongest linked selection due to reduced recombination. This observation was obviously explained by mitochondrial introgression for species with divergent mitochondrial haplotypes shared by individuals of the same population (which strongly increase mean within-population genetic diversity). However, for species with no genetic structure, other factors might explain the slight excess of mitochondrial diversity. First, genetic diversity is directly proportional to both effective population size $N_e$ and mutation rate $\mu$. Because mutation rate is higher in the mitochondrial than in the nuclear genome (Allio et al., 2017), this should increase mitochondrial genetic diversity even if $N_e$ of the mitochondrial genome is four times lower than the 4 fold reduction of nuclear $N_e$. Another possible confounding factor is differential variance in reproductive success between males and females. Variance in reproductive success ($V_k$) quantifies the unequal production of offspring among individuals of a population throughout their lifetime. If $V_k$ is similar between males and females, the reduction in $N_e$ due to $V_k$ will affect similarly both nuclear and mitochondrial genomes. However, if $V_k$ is higher for males than for females, the reduction of $N_e$ will more strongly affect the nuclear than the mitochondrial genome because of the strictly maternal inheritance of the latter. This hypothesis might be true for species with a strong sexual selection involving female mate choice, such as in nest-guarding species like the black seabream (*S. cantharus*), black goby (*G. niger*) or the gray wrasse (*S. cinereus*). On the contrary, higher $V_k$ in females could lower mitochondrial genetic diversity below the 4 fold theoretically expected reduction. Considering the diversity of reproductive strategies in our dataset, a more dedicated analysis would be required to specifically quantify the contribution of variance in reproductive success to observed differences in mitochondrial diversity among species. Evaluating the role of mutation rate would on the other hand require measures of divergence with an outgroup for each species, as well as a time-calibrated phylogeny. Thus, the factors that shape mitochondrial diversity between marine fish species remain to be further explored.

## Comparative landscapes of diversity and divergence

The variance in genetic diversity and genetic divergence among the protein-coding genes within each species' mitochondrial genome is mainly explained by the ratio of non-synonymous to synonymous sites as well as the $\pi_n/\pi_s$ ratio within each gene. This might be explained by high genetic constraints on mitochondrial genes because the corresponding coded proteins interact with nuclear-encoded proteins in the oxidative phosphorylation (OXPHOS) metabolic pathway. Non-synonymous mutations should lead to a change in the protein structure that might reduce the efficacy of the OXPHOS chain and induce fitness reduction. Therefore, these mutations might be quickly removed by purifying selection, even with reduced $N_e$ (Cooper et al., 2015). As mtDNA genome and function is strongly conserved among fish species (Satoh et al., 2016), similar selection might affect all the studied species here. On the contrary, the ratio of non-synonymous to synonymous diversity did not predict the levels of genetic differentiation ($F_{ST}$) across species. We would either mean that genetic differentiation is irrelevant of the mitochondrial genomic architecture and might represent other factors such as barriers on the nuclear genome. Many species show a narrow distributions of $F_{ST}$ among PCGs especially for highly divergent species (e.g. *C. julis*, gene $F_{ST}$ ranges from 0.98 to 1.00). Mitochondrial differentiation should therefore represent the permeability of the nuclear genome, as explained above.

# Conclusion

Here, we described the pattern of differentiation, diversity and divergence of the mitochondrial genomes of 20 marine fish species. The observed variability of these genetic characteristics seems rather large considering that we focused on species in a same biogeographical context that should display a similar population mitochondrial genome structure. However, it mirrors to some extent the results from a previous study on the species that shows huge variability on the neutral population genome structure, apart from a few exceptions of nuclear-mitochondrial discordance (see Chapter 2). This variability cannot be only explained by similar selection pressures: other factors have to be taken into account. Are mito-nuclear incompatibilities important to promote differentiation in our studied species (Burton and Barreto, 2012)? To address these questions, it is mandatory to compare the levels of genetic divergence, resistance to introgression of nuclear genes that encoded proteins in mitochondrial OXPHOS metabolism. If these nuclear genes display barriers to reproductive isolation in species with strong reciprocal monophyly, it might represent the footprints of mito-nuclear incompatibilities. This analysis will also give an estimation of its frequency in the 20 independent diverging species. Secondly, species life-history traits might explain the observed pattern of differentiation. Nabholz et al. (2008) found a strong negative correlation between mtDNA substitution rates and longevity in birds supporting a lower mutation rate in long-lived species. Whether this macro-evolutionary pattern is also seen during divergence remains to be elucidated.

# References

Allio, R., Donega, S., Galtier, N., and Nabholz, B. (2017). Large Variation in the Ratio of Mitochondrial to Nuclear Mutation Rate across Animals: Implications for Genetic Diversity and the Use of Mitochondrial DNA as a Molecular Marker. *Molecular Biology and Evolution*, 34(11):2762–2772.

Avise, J. C., Arnold, J., Ball, R. M., Bermingham, E., Lamb, T., Neigel, J. E., Reeb, C. A., and Saunders, N. C. (1987). Intraspecific Phylogeography: The Mitochondrial DNA Bridge Between Population Genetics and Systematics. *Annual Review of Ecology and Systematics*, 18:489–522.

Ballard, J. W. O. and Whitlock, M. C. (2004). The incomplete natural history of mitochondria. *Molecular Ecology*, 13(4):729–744.

Barry, P., Broquet, T., and Gagnaire, P.-A. (2022). Age-specific survivorship and fecundity shape genetic diversity in marine fishes. *Evolution Letters*, 6(1):46–62.

Burton, R. S. and Barreto, F. S. (2012). A disproportionate role for mtDNA in Dobzhansky–Muller incompatibilities? *Molecular Ecology*, 21(20):4942–4957.

Capella-Gutiérrez, S., Silla-Martínez, J. M., and Gabaldón, T. (2009). trimAl: A tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics*, 25(15):1972–1973.

Cock, P. J. A., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B., and de Hoon, M. J. L. (2009). Biopython: Freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, 25(11):1422–1423.

Cooper, B. S., Burrus, C. R., Ji, C., Hahn, M. W., and Montooth, K. L. (2015). Similar Efficacies of Selection Shape Mitochondrial and Nuclear Genes in Both Drosophila melanogaster and Homo sapiens. *G3 Genes—Genomes—Genetics*, 5(10):2165–2176.

Ebdon, S., Laetsch, D. R., Dapporto, L., Hayward, A., Ritchie, M. G., Dinca, V., Vila, R., and Lohse, K. (2021). The Pleistocene species pump past its prime: Evidence from European butterfly sister species. *Molecular Ecology*, 30(14):3575–3589.

Gagnaire, P.-A., Normandeau, E., and Bernatchez, L. (2012). Comparative Genomics Reveals Adaptive Protein Evolution and a Possible Cytonuclear Incompatibility between European and American Eels. *Molecular Biology and Evolution*, 29(10):2909–2919.

Godinho, R., Crespo, E., and Ferrand, N. (2008). The limits of mtDNA phylogeography: Complex patterns of population history in a highly structured Iberian lizard are only revealed by the use of nuclear markers. *Molecular Ecology*, 17(21):4670–4683.

Hahn, M. W. (2018). *Molecular Population Genetics*. Oxford University Press ; Sinauer Associates, New York : Sunderland, MA.

Hewitt, G. (2000). The genetic legacy of the Quaternary ice ages. *Nature*, 405(6789):907–913.

Katoh, K., Rozewicki, J., and Yamada, K. D. (2019). MAFFT online service: Multiple sequence alignment, interactive sequence choice and visualization. *Briefings in Bioinformatics*, 20(4):1160–1166.

Kondo, R., Satta, Y., Matsuura, E. T., Ishiwa, H., Takahata, N., and Chigusa, S. I. (1990). Incomplete Maternal Transmission of Mitochondrial DNA in Drosophila. *Genetics*, 126(3):657–663.

Li, D., Liu, C.-M., Luo, R., Sadakane, K., and Lam, T.-W. (2015). MEGAHIT: An ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics*, 31(10):1674–1676.

Meng, G., Li, Y., Yang, C., and Liu, S. (2019). MitoZ: A toolkit for animal mitochondrial genome assembly, annotation and visualization. *Nucleic Acids Research*, 47(11):e63–e63.

Miles, A., Bot, P., Murillo, R., Ralph, P., Harding, N., Pisupati, R., Rae, S., and Millar, T. (2020). Cggh/scikit-allel: V1.3.2. Zenodo.

Minh, B. Q., Schmidt, H. A., Chernomor, O., Schrempf, D., Woodhams, M. D., von Haeseler, A., and Lanfear, R. (2020). IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Molecular Biology and Evolution*, 37(5):1530–1534.

Muff, S., Nilsen, E. B., O'Hara, R. B., and Nater, C. R. (2021). Rewriting results sections in the language of evidence. *Trends in Ecology & Evolution*, 0(0).

Nabholz, B., Glémin, S., and Galtier, N. (2008). Strong variations of mitochondrial mutation rate across mammals–the longevity hypothesis. *Molecular Biology and Evolution*, 25(1):120–130.

Nei, M. and Gojobori, T. (1986). Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Molecular Biology and Evolution*, 3(5):418–426.

Niehuis, O., Judson, A. K., and Gadau, J. (2008). Cytonuclear Genic Incompatibilities Cause Increased Mortality in Male F2 Hybrids of Nasonia giraulti and N. vitripennis. *Genetics*, 178(1):413–426.

Oliveira, D. C. S. G., Raychoudhury, R., Lavrov, D. V., and Werren, J. H. (2008). Rapidly Evolving Mitochondrial Genome and Directional Selection in Mitochondrial Genes in the Parasitic Wasp Nasonia (Hymenoptera: Pteromalidae). *Molecular Biology and Evolution*, 25(10):2167–2180.

Rand, D. M., Haney, R. A., and Fry, A. J. (2004). Cytonuclear coevolution: The genomics of cooperation. *Trends in Ecology & Evolution*, 19(12):645–653.

Rossum, G. V. and Drake, F. L. (2009). *Python 3 Reference Manual:*. CreateSpace Independent Publishing Platform, Hampton, NH.

Rozas, J., Ferrer-Mata, A., Sánchez-DelBarrio, J. C., Guirao-Rico, S., Librado, P., Ramos-Onsins, S. E., and Sánchez-Gracia, A. (2017). DnaSP 6: DNA Sequence Polymorphism Analysis of Large Data Sets. *Molecular Biology and Evolution*, 34(12):3299–3302.

Satoh, T. P., Miya, M., Mabuchi, K., and Nishida, M. (2016). Structure and variation of the mitochondrial genome of fishes. *BMC Genomics*, 17(1):719.

Weir, B. S. and Cockerham, C. C. (1984). Estimating F-Statistics for the Analysis of Population Structure. *Evolution*, 38(6):1358–1370.

Wickham, H. (2016). *Ggplot2: Elegant Graphics for Data Analysis*. Use R! Springer International Publishing : Imprint: Springer, Cham, 2nd ed. 2016 edition.

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T., Miller, E., Bache, S., Müller, K., Ooms, J., Robinson, D., Seidel, D., Spinu, V., Takahashi, K., Vaughan, D., Wilke, C., Woo, K., and Yutani, H. (2019). Welcome to the Tidyverse. *Journal of Open Source Software*, 4(43):1686.

Zouros, E., Oberhauser Ball, A., Saavedra, C., and Freeman, K. R. (1994). An unusual type of mitochondrial DNA inheritance in the blue mussel Mytilus. *Proceedings of the National Academy of Sciences of the United States of America*, 91(16):7463–7467.

# DISCUSSION

# Discussion contents

**1 Life-history traits and speciation: what have we learned so far ? 161**

1.1 Adult lifespan and life tables impact the reservoir of ancestral genetic diversity . 161

1.2 Complex effects of life-history traits on differentiation and divergence . . . . . 162

**2 Expanding the genetic differentiation continuum... 163**

2.1 ... to the left: why some species *do not* have genetic barriers? . . . . . 163

2.2 ... and to the right: are some species completely isolated? . . . . . 164

**3 Contribution of ancient introgression to reproductive isolation 165**

**4 Another layer of structure: differentiation between marine and lagoon populations 166**

**5 How does linked selection affect species with different life-history traits? 169**

**6 Comparative genomics beyond vertebrates 171**

**7 Gene trees and ABC 171**

**8 Genetic differentiation reveals cryptic layers of biodiversity for conservation issues 174**

**9 Bridging the gap between phylogeography and comparative genomics 175**

**10 Conclusion 175**

# Discussion

## 1 Life-history traits and speciation: what have we learned so far ?

The main objective of the thesis was to broaden our knowledge on the *causes* of speciation and specifically address the role of species *life-history traits* on the accumulation and erosion of reproductive isolation barriers. We compared various genetic characteristics such as differentiation, divergence, introgression and demographic history of 20 species subdivided into the Atlantic Ocean and Mediterranean Sea populations or geographical lineages. These species supposedly followed a similar biogeographic history of separation during the Pleistocene followed by secondary contact after the Last Glacial Maximum. They also share a similar genome architecture but display a wide array of life-history characteristics.

### 1.1 Adult lifespan and life tables impact the reservoir of ancestral genetic diversity

We found that a species' genetic diversity is mainly driven by adult lifespan and parental care: long-lived species with higher parental progeny investment have lower genetic diversity, in agreement with previous studies conducted at broader phylogenetic scales (Romiguier et al., 2014; Chen et al., 2017). We specifically addressed the cause of this relationship with an analytic and a simulation approach: we found that particular age-survival and age-fecundity curves characterized by high juvenile and low adult mortality (i.e., type III survivorship curves) combined with increasing fecundity with age, specific to many marine fishes can explain the observed effect of adult lifespan on diversity. Indeed, a long adult lifespan reduces the ratio of effective population size over the census size as $N_e/N$ because it increases the variance in reproductive success between individuals, an effect that is particularly due to the disproportionate contribution of rare but highly fecund old individuals. The lower genetic diversity in species with parental care behavior is in agreement with a lower probability of extinction of species with high parental investment at low $N_e$ (Romiguier et al., 2014).

These effects were discovered by looking at the current genetic diversity of species, but they are due to vital rates effects and ecological strategies on long-term effective population size, and there is little doubt that these processes also determine the ancestral reservoir of genetic diversity of any species currently experiencing genetic divergence into several lineages. Thus, the number of genetic incompatibilities that arise in the ancestral diversity reservoir (Cutter, 2012) and of long-term balanced polymorphism that might be sieved after ancestral split (Guerrero

and Hahn, 2017) is supposed to be greater in species with short adult lifespan and no parental care. In fact, any ancestral polymorphism, whether neutral or not, will potentially contribute to divergence after the split of the ancestral population. However, the extent of reproductive isolation due to the differential sorting of ancestral polymorphism will also depend on the effect of $N_e$ and on the number of such mutations in the ancestral variation, since this number is not necessarily proportional to the amount of neutral genetic diversity. Because present genetic diversity is the product of the long-term $N_e$, it gives also a clue about the intensity of genetic drift that each subpopulation will experience after split during divergence.

## 1.2 Complex effects of life-history traits on differentiation and divergence

Three main relationships between life-history traits and genetic characteristics have been pinpointed in this thesis: a negative relationship between i) pelagic larval duration and genetic differentiation, ii) body size and absolute genetic divergence, and iii) lifespan and time of ancestral split. These relationships can be explained by the differential impact of gene flow, $N_e$ and generation time respectively and are discussed elsewhere (see discussion Chapter 2). However, we found that the relationship between life-history traits and genetic characteristics is overall complex and that no simple trait or combination of traits can fully explain on its own the different evolutionary histories experienced by the studied species (Fig. 1). Specifically, no trait has been found to explain why some species accumulate more barriers to gene flow than others. This can reflect either a true lack of simple relationships between life-history traits and genetic characteristics; or it could be that other components not directly related to speciation have to be taken into account to disentangle the putative multifarious effects of traits during divergence. The latter hypothesis seems more likely: for example, the relationship between pelagic larval duration and $F_{ST}$ would have been weak if we had considered all species. But we found a stronger negative correlation by removing the species that do not show contemporary gene flow: had we never known this parameter (which we inferred from independent ABC analyses), we would not have concluded the impact of dispersal capacities on $F_{ST}$. We think that our current lack of knowledge of some parts of the evolutionary history of the studied species here limits our ability to assess the impact of life-history traits on speciation.

Here, we will discuss some potential aspects of the study that could help better understand the impact of life-history traits: we will consider i) broaden the continuum of studied species to the 'left' (weak or no genetic differentiation) and to the 'right' (highly divergent species close to complete reproductive isolation), ii) take into account the deep divergence history possibly explain by ancient episodes of introgression from closely related species, iii) consider a marine/lagoon population structure for some studied species, iv) explore the role of a differential impact of recombination and linked selection, and v) broaden the diversity studied species

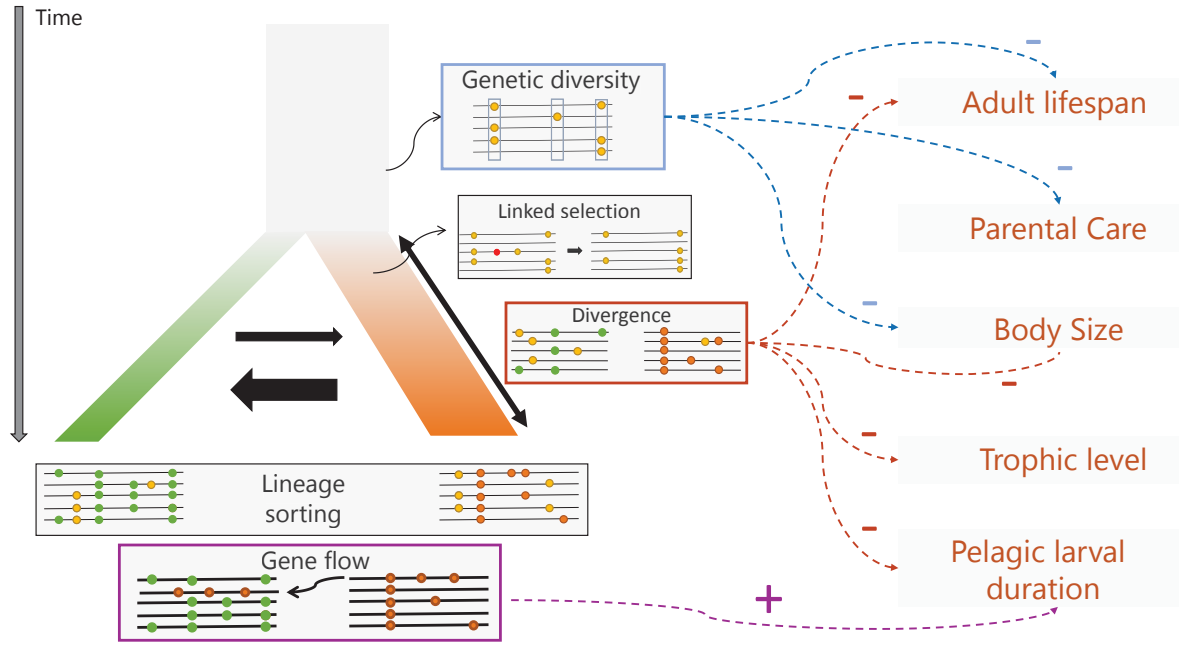



Figure 1: **Summary of the relationships between five life-history traits and the evolutionary processes that affect speciation**. The determinants of ancestral genetic diversity (blue), divergence (red), and gene flow (purple) are shown with corresponding colored arrows to each specific trait: minus (-) or plus (+) sign correspond to a negative or positive effect respectively.

beyond vertebrate taxa. Then, in the last three sections we will discuss about i) the potential of whole-genome sequences and gene genealogies to understand the species' demographic history, ii) the implications of our results on conservation biology, and finally, iii) the link between phylogeography and comparative genomics.

# 2 Expanding the genetic differentiation continuum...

## 2.1 ... to the left: why some species *do not* have genetic barriers?

Because we were interested in the determinants of the accumulation of reproductive isolation barriers, we mainly focused on species in the *gray zone of speciation*, where speciation is still incomplete (Roux et al., 2016). However, previous studies showed that some marine fish species are not genetically separated at all in Atlantic and Mediterranean lineages, such as, for the example, the chub mackerel (Zardoya et al., 2004). Some are even closely related to species with similar life-history traits that we included in our study such as the white seabream (*Diplodus sargus*) related to the sharp-snout seabream (*D. puntazzo*) (Bargelloni et al., 2005), the red mullet (*Mullus barbatus*) related to the stripe red mullet (*M. surmuletus*) (Félix-Hackradt et al., 2013) or the European angler (*L. budegassa*) related to the black-bellied angler fish (*L.*

*budegassa*) (Charrier et al., 2006). How can we explain these differences?

First, during allopatry (i.e., supposing that some periods of allopatric isolation have existed for most species), some species may have accumulated genetic differentiation and divergence that was not subsequently involved in reproductive isolation, so that divergence was later eroded by gene flow after secondary contact. Secondly, some species may have not started to diverge at all: the separation between the Atlantic Ocean and the Mediterranean Sea was not complete during the Pleistocene periods (the minimum depth in the transition zone never went below about 80 meters in the shallowest zone located on the Atlantic side of the Gibraltar Strait), and some species should have sustained enough gene flow to counterbalance any potential initial genetic differentiation.

How many life-history traits be related to these patterns? All previously exposed hypotheses between life-history traits and the accumulation of reproductive barriers can be extended to these questions. For instance, a long pelagic larval duration might confer such a strong dispersal ability to some species that they maintained intensive gene flow during the Pleistocene and did not initiate the evolution of any barrier to reproduction. Similarly, genetic drift could have been simply too weak in large $N_e$ species to let the time for new mutations to fixed and therefore most of the ancestral polymorphism would still be retained in the two populations.

To answer these questions, the current data set should be updated with species with closely related species but showing contrasted values of life-history traits. The presence of blocks of higher coalescent times compared to that expected at migration-drift-equilibrium might be interpreted as the presence of old divergence that has more recently been eroded by gene flow. Then, we could test if specific life-history traits can explain why some species have accumulated barriers (species with contemporary barriers or those with ancient barriers that were eroded) and others have not; and secondly if they can explain why barriers have been totally eroded in some species.

## 2.2 ... and to the right: are some species completely isolated?

Looking at the other end of the continuum, have some species accumulated more barriers to reproduction than those studied here? We did not find genetic studies on other fish species that showed higher levels of genetic differentiation than our most strongly divergent species (e.g. *L. mormyrus*, *C. galerita* and *D. puntazzo*). However, there are some lineages that have been characterized as distinct species and have disjunct distributions between the North-Eastern Atlantic and the Mediterranean Sea.

For example, the capelan *Trisopterus capelanus* was recently recognized as distinct from

the closely related pout *T. luscus* (Delling et al., 2011). The two species have disjunct distributions with a small overlap: the former is only found in the Mediterranean Sea and on the Atlantic coast of Morocco while the latter is found in the North-Eastern Atlantic and the Western Mediterranean Sea (Gonzalez et al., 2012). Delling et al. (2011) found a 4.5% dissimilarity in mitochondrial gene *cytochrome b* between the two species, that is, half of that of the most divergent species of our study, the Montagu's blenny (*Coryphoblennius galerita*, 7% dissimilarity between Atlantic and Mediterranean populations). To our knowledge, no study looked at the genetic differentiation, the demographic history and the semi-permeability between these two species: they may represent an advanced stage of reproductive isolation between Atlantic and Mediterranean lineages and a similar biogeographic history, but were simply not covered by the gradient of divergence defined by our set of 20 species.

Similar patterns of distribution disjunctions between closely related marine fishes species are found between the ribbonfish *Trachipeterus trachypterus* and *T. arcticus*; the Spanish *Molva macrophtalma* and the common *M. molva* lings; the *Lipophrys trigloides* and *L. pholis* combtooth blennies or the *Labrus bergylta* and *L. viridis* wrasse (Louisy, 2015). Recently, Aguirre-Sarabia et al. (2021) found frequent natural hybridization between *L. budegassa* and *L. piscatorius* with some $F_1$ hybrids and very few backcrosses: although the two species are found in sympatry within the Atlantic Ocean and the Mediterranean Sea, they might also represent a late stage of the speciation continuum with hybridization that leads to a high reduction in fitness and quick removal of introgressed foreign alleles from the gene pool. Thus, analyzing patterns of genetic differentiation between these species pairs could allow us to decipher whether they represent highly reproductively isolated lineages that possibly diverged across the Atlantic Ocean - Mediterranean Sea suture zone and help to assess which life-history traits can explain a strong accumulation of reproductive barriers.

# 3 Contribution of ancient introgression to reproductive isolation

The trans-specific origin of reproductive isolation barriers is of particular interest for the study of speciation. For example, Duranton et al. (2020) showed that alleles that introgressed about 80 000 years ago from the spotted sea bass *Dicentarchus punctatus* to the *D. labrax* Atlantic populations correspond to the regions that contribute to barriers to gene flow with *D. labrax* Mediterranean populations. Two observations suggest that such complex effects involving outsider species might have happened in some of the other systems that we looked at. First, Time to Most Recent Common Ancestor (TMRCA) distributions inferred from genome-wide gene genealogies showed the presence of very old blocks, older than 1 million years, that we sometimes not found within basins (e.g. *S. cantharus*). Although our estimations might be not precise because we made assumptions on the mutation rate, effective population

size and generation time, older tracks might represent past introgression from other species. Second, many studied species have a closely related congener species one in the Atlantic Ocean or Mediterranean Sea that might be good candidates as past donor species: the sea comber *Serranus cabrilla* is closely related to the blacktail comber *S. atricauda*, present around the Iberian Peninsula and the Coast of Morocco (Vella et al., 2021); *Alosa fallax* to *A. alosa* (Faria et al., 2006); *Mullus surmuletus* to *M. barbatus* (Turan, 2006); *Merluccius merluccius* to *M. senegalensis* (Campo et al., 2007).

To test the role of past admixture, two steps have to be fulfilled: it requires to i) detect the regions of the genomes that show a signal of differential introgression between one donor species and one population of either the Atlantic or the Mediterranean basins, ii) test whether these regions contributed to Atlantic/Mediterranean reproductive isolation by comparing the level of genetic differentiation, divergence and TMRCA in these old introgressed regions compared with the rest of the genome. These analyses need "only" the re-sequencing of one individual from the candidate donor species and the phasing of genotypes with known ancestral states at variant sites for the focal species.

As it turns out, we actually have these data to test the potential past introgression from the common angler-fish, *Lophius piscatorius* towards *L. budegassa*. These two species display about 9-10 % of genetic divergence on their mitochondrial genome. We have sampled two *L. piscatorius* samples in Med-in (Costa Calida region) and inferred *D* statistic with three different topologies using *L. piscatorius* as the P3 population - i.e., test for differential introgression between *L. piscatorius* and either P1 or P2 (Atl. or Med. populations in *L. budegassa*) (Green et al., 2010). We found very strong evidence for differential introgression of *L. piscatorius* genes between *L. budegassa* populations: especially Atlantic populations appeared to be more introgressed by foreign alleles of *L. piscatorius* origin than Mediterranean ones, consistent with a higher rate of contemporary hybridization found in the Atlantic (Aguirre-Sarabia et al., 2021). If genome regions of high introgression between the two species are also those with old TMRCA between the two basins within *L. budegassa*, then it might mean that past and/or contemporary admixture in the *L. budegassa* genome is contributing to weak, but existing reproductive isolation. This example shows the potential research path to follow to increase our knowledge on the origin of reproductive barriers.

# 4 Another layer of structure: differentiation between marine and lagoon populations

Among the 19 studied species, 4 show a more complex population structure than a simple differentiation between Atlantic and Mediterranean lineages. These four species - the gray wrasse, *S. cinereus*, long-snouted seahorse, *H. guttulatus*, the broadnosed pipefish, *S. typhle*

Table 1: **Differential introgression between *L. piscatorius* and pairwise topologies among the four populations of *L. budegassa* studied here and measured by the $D$ statistics.**

| P1 | P2 | P3 | $D$ | $Z$ | $p-value$ | $f_d$ |
|---|---|---|---|---|---|---|
| Atl-in | Atl-out | *L. piscatorius* | 0.01873 | 3.6781 | 0.00024 | 0.00068 |
| Med-out | Atl-in | *L. piscatorius* | 0.01597 | 8.0921 | $< 0.000001$ | 0.00055 |
| Med-in | Atl-in | *L. piscatorius* | 0.00504 | 1.5981 | 0.11 | 0.00018 |
| Med-out | Atl-out | *L. piscatorius* | 0.03423 | 7.3406 | $< 0.000001$ | 0.00124 |
| Med-in | Atl-out | *L. piscatorius* | 0.02376 | 3.9817 | 0.000068 | 0.00087 |
| Med-out | Med-in | *L. piscatorius* | 0.01043 | 3.3255 | 0.00083 | 0.00035 |

and the big-scale sand melt *A. boyeri* - are suspected to be genetically divided into lagoon and marine lineages (Hanel et al., 2003; Riquet et al., 2019; Louisy, 2015; Boudinar et al., 2016). Interestingly, they all display lower net genetic divergence between the Atlantic and Mediterranean basins than expected solely from the genome-wide level of genetic differentiation: this pattern might be caused by recurrent bottlenecks that drastically reduced $N_e$ causing higher lineage sorting and differential fixation of haplotypes in each location with small differences in nucleotide compositions. Recurrent bottlenecks might be more prevalent in lagoon habitats because of lower habitat stability (more fluctuations in environmental conditions). Moreover, all these species should be able to sustain drastic reductions of $N_e$ without going to extinction because of higher parental investment in the progeny (Romiguier et al., 2014).

For *A. boyeri*, Milana et al. (2008) found a 100 bp intergenic spacer in the mitochondrial genome of lagoon individuals that is not present in marine individuals. In our study, the main axis of differentiation in a PCA analysis for *A. boyeri* separated the marine habitat and lagoons in contrast with a weak differentiation within each basin: Med-in samples with only lagoon individuals and Atl-in with only marine individuals displayed the highest between-populations net and absolute divergence among all species with 3.77% and 3.22% respectively. For *S. cinereus*, we sampled and sequenced 2 individuals from Thau Lagoon and 3 individuals from Agde (marine location) in the Med-out sample (Gulf of Lion): however, contrary to *A. boyeri*, they displayed lower genetic differentiation between them, and they were more related to each other than with individuals from both Atlantic populations (all sampled in lagoons). *S. typhle* showed a very high population structure between the four populations while *H. guttulatus* showed a first axis of differentiation between both Mediterranean and Atl-out (Bay of Biscay) populations and a second axis of differentiation separating Atl-in (Algarve). However, the lack of marine individuals in these two species made the analysis of the marine/lagoon differentiation impossible. We note, however, that *H. guttulatus* and *S. typhle* both show large inversions within their genomes that are almost differentially fixed between populations that were sampled in different habitat types. The occurrence of inversions within these species is also concordant with the hypothesis of severe reductions in $N_e$ that may help to fix underdominant chromosomal rearrangements (Walsh, 1982). The emergence of these inversions and the role that they play

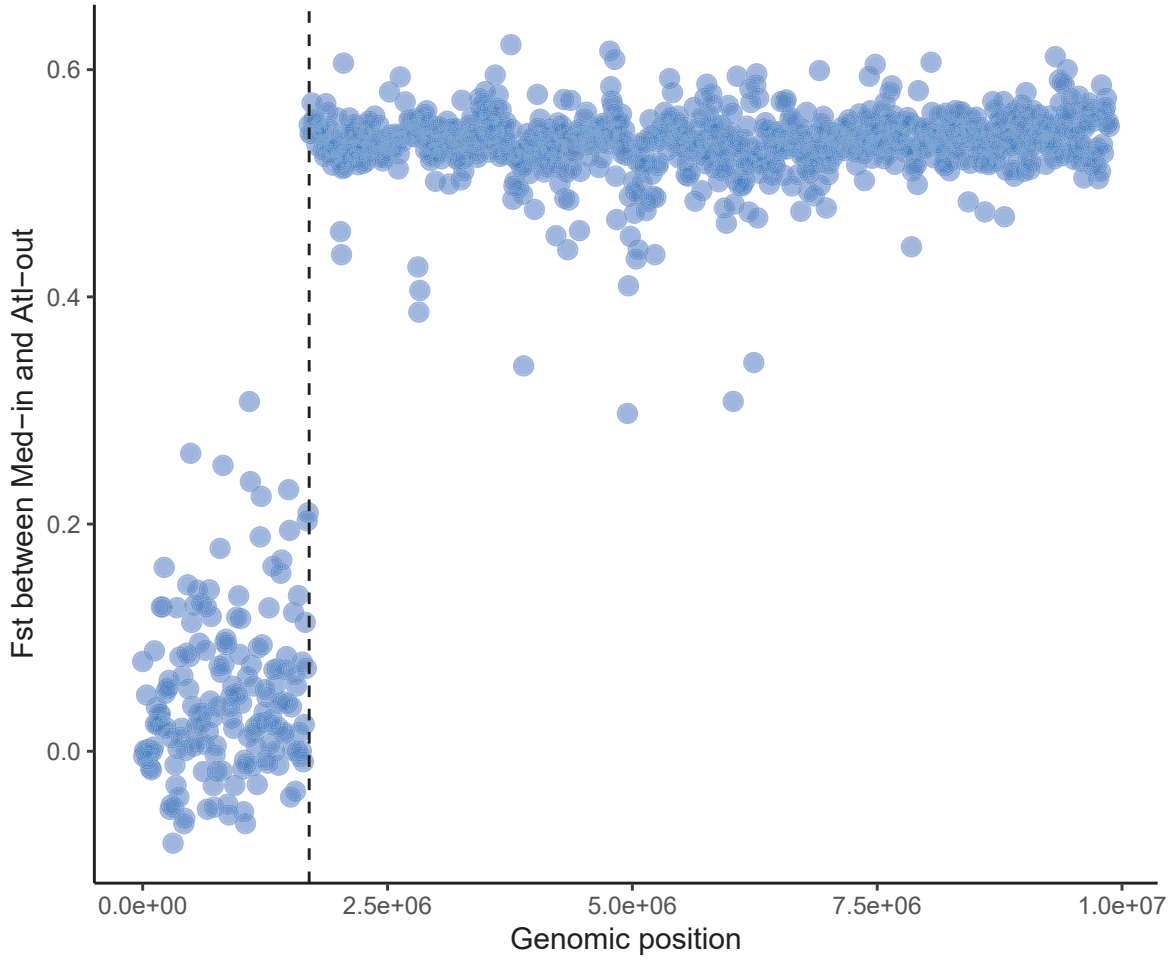



Figure 2: **Effect of chromosomal inversions in *H. guttulatus***. Genetic differentiation $F_{ST}$ between Med-in (Costa Calida) and Atl-out (Bay of Biscay) populations measured in 10kb windows.

in the differentiation between marine and lagoon populations will have to be clarified.

Should marine and lagoon lineages be considered as different *ecotypes*? To be so, it has to be proven that the two lineages are genetically adapted to the different habitats (Johannesson et al., 2010). Identification of selective sweeps at loci that might have a role in lagoon or marine adaptation, fitness measures in reciprocal transplants experiments are mandatory to assess the role of selection in the formation of marine/lagoon lineages.

# 5 How does linked selection affect species with different life-history traits?

Recessive deleterious mutations should be more prevalent in low $N_e$ species because purifying selection becomes less efficient as effective population size decreases. As a consequence, the introgression of foreign blocks in the genome could mask the effect of these mutations because it is unlikely that two diverging populations have fixed deleterious mutations on the same loci (Harris and Nielsen, 2016). Thus, recombination rate should be negatively correlated with introgression as recombination will break up the blocks of ancestry and unveil the previously masked recessive deleterious mutations. On the other hand, if reproductive barriers are polygenic, selection against foreign ancestry should be stronger in long introgressed blocks: thus, introgression should be greater and genetic differentiation lower in regions of high recombination (Martin et al., 2019). This pattern could be further enhanced by faster lineage sorting of genetic incompatibilities due to higher linked selection in high $N_e$ species.

The complex effect between $N_e$ and variation in introgression along the genome modulated by variation in recombination rate can be assessed by comparing the levels of intragenomic variation in $F_{ST}$, $d_{XY}$, $f_d$ and $\rho$, the population recombination rate ($\rho = 4N_e r$). This last parameter can be inferred using population genomics methods relying on genome-wide polymorphism data with phased genotypes: so far, we performed this inference for 10 of our species. Using 10kb windows in the European sea bass (*D. labrax*), we found a negative relationship between $F_{ST}$ and $\rho$, and a positive relationship between $d_{XY}$ and $\rho$, which is coherent with previous results on this species (see Fig. 5a and c in Duranton et al. (2018). Notably, this is coherent with an important impact of background selection that increases differentiation without increasing divergence in low-recombining regions. We found also some regions that display high $d_{XY}$ in low-$\rho$ regions, which could represent barriers and contribute to reduced gene flow as shown by Duranton et al. (2018).

A possible limitation here was that local recombination rates were inferred from polymorphism data with `LDHelmet` (Chan et al., 2012) with possible low accuracy due to small sample size and confounding factors such as genetic population structure. However, with confidence in these estimations, it is possible to develop this approach for all 19 species, to compare the differential effect of linked selection on genetic polymorphism and ask the following questions, how does linked selection affect patterns of differentiation and divergence and can some life-history traits, especially those related to $N_e$ (body size, adult lifespan, parental care) explain these differences?

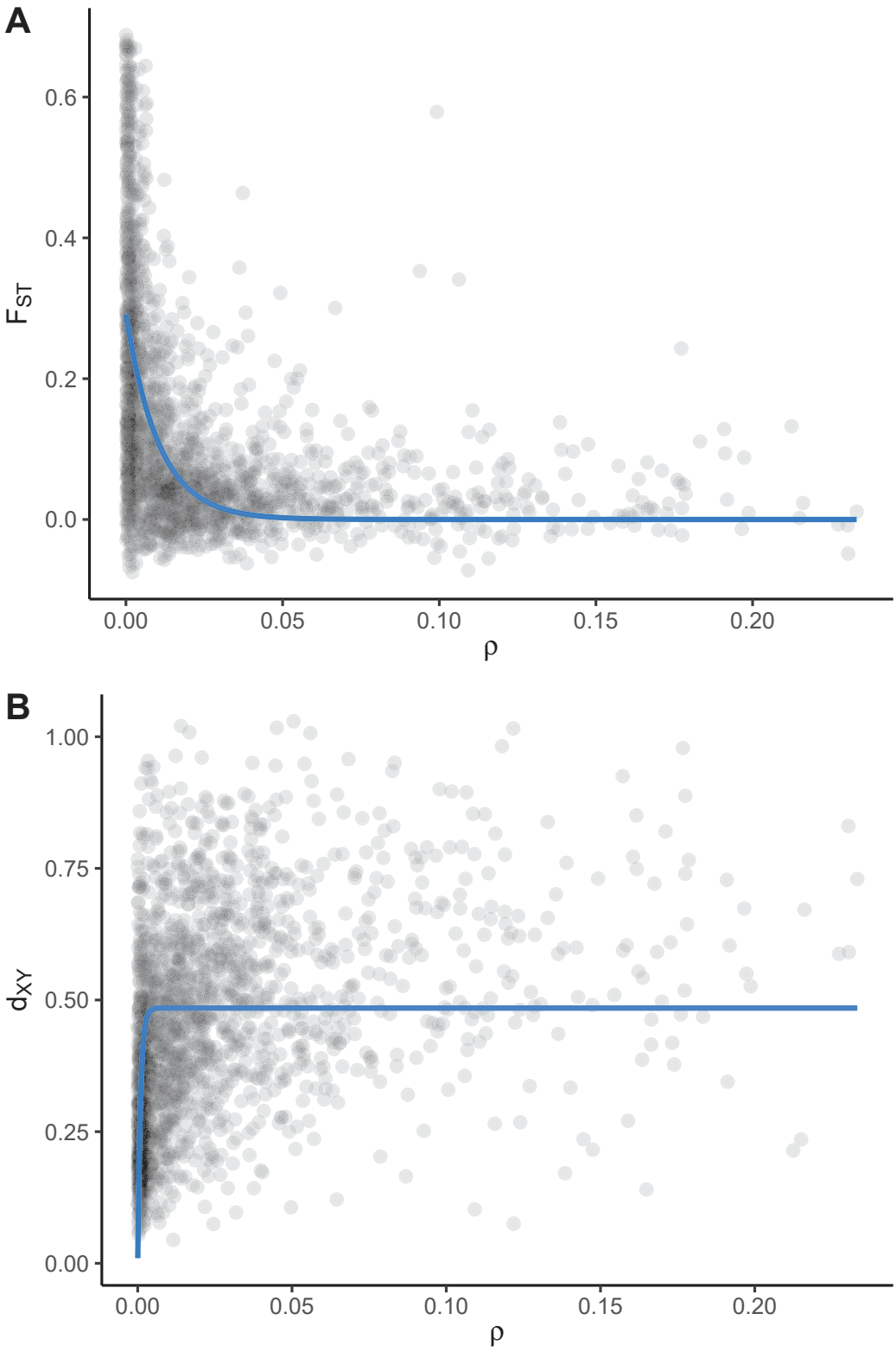



Figure 3: **Correlation between genetic differentiation $F_{ST}$ (A, top) and divergence $d_{XY}$ (bottom, B) with the population recombination rate, $\rho$, in 10kb windows.** $F_{ST}$ and $d_{XY}$ follow a inverse-exponential relationship with $\rho$ following $F_{ST} = a * (exp[-b * \rho])$ and $d_{XY} = a * (1 - exp[-b * \rho])$

# 6 Comparative genomics beyond vertebrates

The Atlantic Ocean - Mediterranean Sea phylogeographical break is not only a suture zone for marine vertebrates: various marine invertebrates, algae and plants show divergent lineages (reviewed in Patarnello et al. (2007). Would a comparative genomics approach with both vertebrates and invertebrates help to address the role of life-history traits on speciation? The answer is not clear-cut.

On one hand, it can broaden the variability of life-history traits: invertebrates usually display higher effective population size (White et al., 2007), variability in fecundity and longer lifespan - such as 42 years for the European lobsters *Hommarus gammarus* (Sheehy et al., 2011) - that can affect $N_e$ - and higher variability of pelagic larval duration, specifically very short PLD, less than one day, and the presence or not of a pelagic larval phase during the life cycle (Shanks, 2009), which can affect migration rate. A lack of relationship between any genetic statistics and life-history traits can be caused by a lack of variability. Thus, broadening the spectrum of possible traits with invertebrates could help resolve this issue.

On the other hand, the assembly of invertebrate genomes is much more difficult because of larger genome sizes and higher prevalence of repeated elements (Simakov et al., 2022). This could lead to low-quality reference genomes, low contiguity of scaffolds, which would reduce the power of whole-genome analyzes to infer gene genealogies. Secondly, higher genome size means higher sequencing effort to reach 20X individual sequencing depth: for a given limited sequencing effort, this would mean fewer individuals per species and/or fewer species sequenced. Moreover, genome architecture differs between vertebrate and invertebrate species (Simakov et al., 2022) which could blur the signals of life-history traits and any genetic characteristics related to genome architecture that can affect speciation because of its effects on the distribution of recombination rate.

# 7 Gene trees and ABC

An important aspect of speciation is to understand how introgressed alleles of foreign ancestry recombine in their new genetic background. Recombination introduces variation in gene genealogies that might represent different aspects of the species' evolutionary history. New tools have been recently developed to infer these topologies from phased whole-population genomics data (Kelleher et al., 2019; Speidel et al., 2019; Wohns et al., 2021). The decrease in the cost of sequencing has given us the ability to sequence the whole-genome of several individuals from several populations such as in classic population genetics. In the second chapter, we proposed that these topologies can trace footprints of loci that coalesced at an older time than expected under standard coalescent expectations. The comparison of the distribution of times to the

most recent common ancestor between and within populations can show how these genealogies are distributed between the populations. We think that these tools might be promising to answer questions about speciation processes.

A second approach is to use Approximate Bayesian Computations (ABC) to infer the demographic and evolutionary history (Beaumont et al., 2002). ABC relies on a comparison between empirical and simulated genetic statistics under various evolutionary models - e.g., secondary contact or strict isolation - and different values of evolutionary parameters - e.g., time of ancestral split, effective population size. These methods have been developed to take into account the heterogeneity in $N_e$ due to linked selection and variation in migration rates due to the presence of reproductive isolation barriers (Roux et al., 2016; Fraïsse et al., 2020). However, the choice of genetic summary statistics is very important because some model parameters can only be precisely inferred if the most relevant statistics are included in the analysis. This also means that the power of ABC methods to discriminate between models and to infer parameters is limited by the available statistics. Whole-genome sequences offer the opportunity to infer new statistics in an ABC framework based on blocks of local ancestry. Introgressed blocks of ancestry are eroded in the foreign population proportional to time and local recombination rate (Liang and Nielsen, 2014): high prevalence of longer blocks are either the sign of recent divergence time, small recombination rate or high migration rate. Thus, the length distribution of blocks of ancestry, conditioned on known local recombination rate, should give clues about the time of split and migration rate.

We started to develop this approach with Camille Roux from the University of Lille (`https://eep.univ-lille.fr/user/camille.roux/`). We modeled a two-population split scenario with secondary contact: we simulated a 1Mb locus with various parameters including $N_e$, time of split, time of secondary contact and migration rates. Then, we inferred these parameters using either classic ABC summary statistics or statistics derived from the blocks of ancestry: for this, we measured the number of blocks in intervals of 1000kb of size between 0 and 0.1Mb. We found an accurate estimation of the time of split and the time of secondary contact using only "block of ancestry" statistics, completely outperforming classical ABC summary statistics (Fig 4). However, there was no increase in the ability of blocks of ancestry statistics to predict gene flow intensity. The integration of new statistics specific to whole-genome sequences can be beneficial for ABC inferences; however future works are needed to specifically address in which way these statistics will be useful to infer parameters such as migration rates, which are currently very difficult to infer precisely with classic ABC statistics.

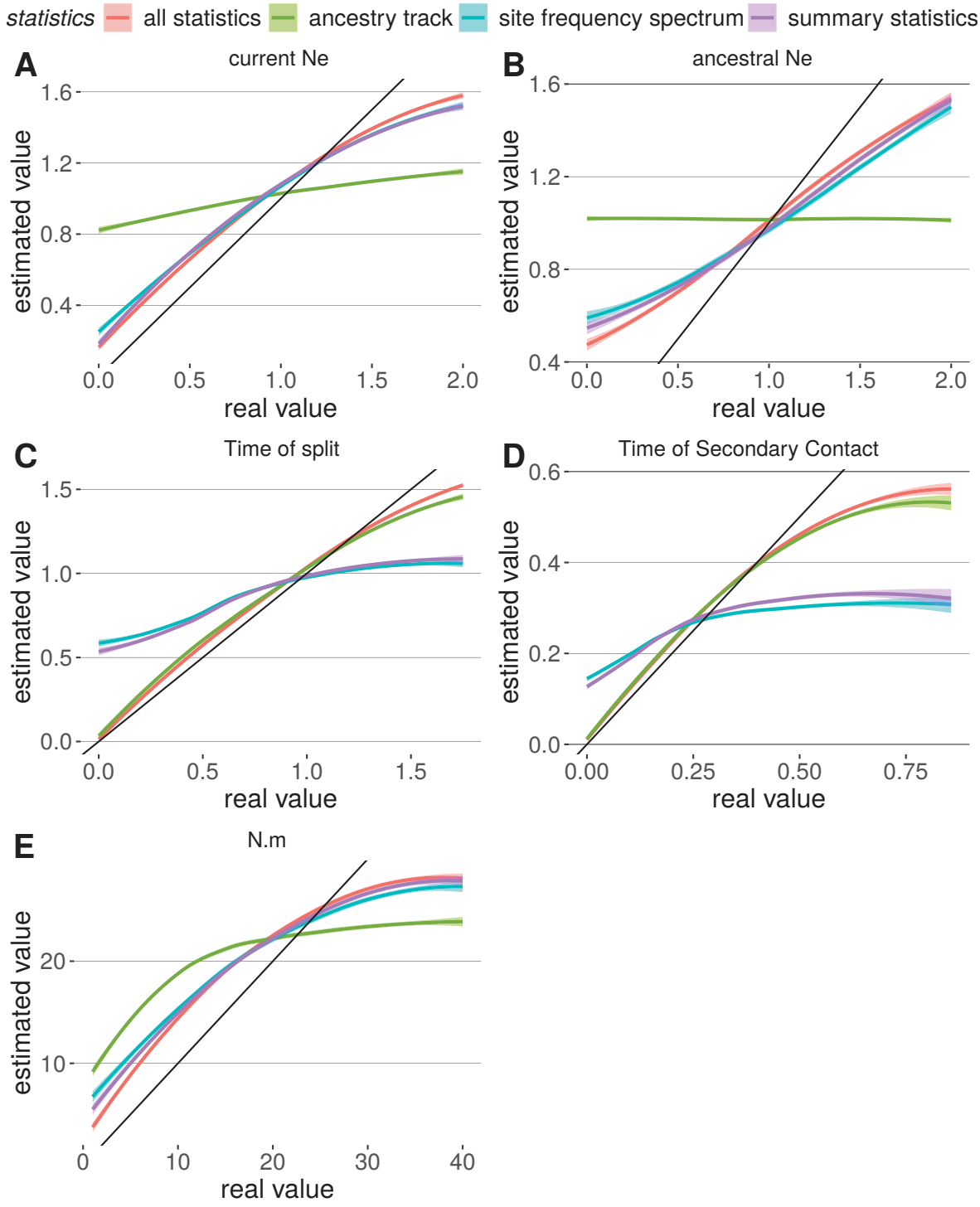



Figure 4: **Performance of ABC with or without "blocks of ancestry" statistics**. Each panel represents the real parameters on $x$-axis and the estimated parameters on $y$-axis either based on the site frequency spectrum (blue), classic summary statistics (violet), ancestry track statistics (green) or all statistics (red). Black line represents the $y = x$ line. A) current $N_e$, B) ancestral $N_e$, C) Time of split, D) Time of secondary contact, E) Gene flow ($N_e m$). Graphical illustrations made by Camille Roux.

# 8 Genetic differentiation reveals cryptic layers of biodiversity for conservation issues

Our work revealed a deep divergence between some pairs of lineages between the Atlantic Ocean and the Mediterranean Sea, including for instance the sharp-snout sea bream (*D. puntazzo*), the Montagu's blenny (*C. galerita*) and the sand seabream (*L. mormyrus*). Notably, these 3 species share a high level of genetic differentiation between Atlantic and Mediterranean populations ($F_{ST} > 0.6$) and no or low differentiation within each of the two basins. This last result indicates that high differentiation between the two populations within the suture zone could not be explained by limited dispersal due to long geographical distances. Rather, they represent strong barriers to gene flow that impede foreign alleles to persist and spread within each basin, so that we detect no contemporary gene flow between outer populations for *D. puntazzo* and *L. mormyrus* and even between inner populations in the case of *C. galerita*. Strikingly, this strong genetic differentiation does not translate into strong phenotypic differentiation. Palma and Andrade (2002) found some differences between West-Mediterranean and North-Eastern Atlantic populations for both *D. puntazzo* and *L. mormyrus* based on morphometric analysis of multiple landmarks. Domingues et al. (2007) found that crest size and width might differentiate Atlantic and Mediterranean *C. galerita* individuals. However, these phenotypic differences are so slight that it would have seemed irrelevant that they can be accompanied by such a strong genetic differentiation.

This raises the importance of genetic studies to unveil *cryptic diversity* - i.e. the presence of diverging lineages within a species with no morphological differentiation (Espíndola et al., 2016), that has consequences in conservation biology. First, *D. puntazzo* and *L. mormyrus* are of prime interest for semi-industrial, small-scale and recreational fisheries. Our work shows that the management of stocks of these two species might be improved if the divergence between the two lineages is taken into account. Secondly, *D. puntazzo* has been farmed in aquaculture for more than 25 years because of its fast growth and omnivorous diet that reduce production cost (Oikonomou et al., 2021; Karapanagiotidis et al., 2021). *L. mormyrus* could also be selected as a good candidate for aquaculture production (Divanach and Kentouri, 1983). As the presence of two lineages was never acknowledged for these two species, the development of domesticated strains in these two species can lead to human-induced hybridization between Atlantic and Mediterranean medicated by aquaculture escapees. With our current knowledge, the consequences of such hybridization are unpredictable: this can be advantageous because of transfer of beneficial alleles through *adaptive introgression* (Whitney et al., 2015) or negative because of the disruption of blocks of locally coadapted alleles, loss of genetic diversity, introgression of maladapted alleles and ecological competition or transfer of pathogens that can reduce wild population viability (Randi, 2008). In any case, genetic divergence between the two lineages in both species should be taken in account into aquaculture monitoring to assess the

potential impact of aquaculture escapees on wild eco-evolutionary dynamics (Todesco et al., 2016; Bradbury et al., 2020).

# 9 Bridging the gap between phylogeography and comparative genomics

Our approach lies at the crossroad of two major fields: comparative *phylogeography* and population *genomics*. Phylogeography has been developed by Avise et al. (1987) and coworkers to understand what biogeographical and ecological factors shape the temporal and spatial patterns of genetic polymorphism and variability of a given species (Gutiérrez-García and Vázquez-Domínguez, 2011). *Comparative phylogeograhy* has naturally emerged to identify common evolutionary histories of several independent species caused by a common biogeographic context. In parallel, *comparative genomics* aims to assess how and why evolutionary forces - mutation, drift, selection and recombination - shape the patterns of polymorphism along the genome and between taxa at various evolutionary scales (McGrath, 2022).

The approach developed in this thesis is located at the crossroad on these two research fields by taking the advantages of both (Edwards et al., 2022): we followed a common comparative phylogeographical methodology by testing the impact of the Atlantic Ocean - Mediterranean Sea suture zone on 20 independent codistributed marine fish species and we used whole-genome sequences to more precisely infer the demographic history and the semi-permeability of the genomes by analyzing the multiple gene genealogies along the genome. We think that the use of whole-genome sequences in comparative phylogeography and a better consideration of recombination will offer good opportunities to assess the impact of biogeographic history on spatial genetic variation patterns and particularly how they might differentially affect species with different life-history traits.

# 10 Conclusion

Despite several decades of research, the ecological and evolutionary factors that explain how species arise and persist remain still not fully understood. The emergence of *speciation genomics* 20 years ago has allowed us to empirically address previous assumptions based on theoretical models (Barton and Bengtsson, 1986), very reduced genome representations based on a small number of loci (Szymura, 1983) and, at the root of this research field, naturalist observation (Wallace, 1865). Advances in genome technologies have allowed us to precisely describe the patterns of genetic differentiation of diverging lineages and understand what evolutionary forces shape their genomic differentiation landscapes. This approach has been developed for many taxa across all the Tree of Life. Today, the decrease of the sequencing costs enables researchers

to study the diverging genome of multiple species pairs (Roux et al., 2016) and implemented similar comparative methods than those developed to study the evolution of post- and pre-zygotic barriers (Coyne and Orr, 1989). We believe that further similar comparative studies of multiple species pairs will broaden our knowledge of the determinants of the accumulation of reproductive isolation barriers that affect the tempo and mode of speciation.

# References

Aguirre-Sarabia, I., Díaz-Arce, N., Pereda-Agirre, I., Mendibil, I., Urtizberea, A., Gerritsen, H. D., Burns, F., Holmes, I., Landa, J., Coscia, I., Quincoces, I., Santurtún, M., Zanzi, A., Martinsohn, J. T., and Rodríguez-Ezpeleta, N. (2021). Evidence of stock connectivity, hybridization, and misidentification in white anglerfish supports the need of a genetics-informed fisheries management framework. *Evolutionary Applications*, 14(9):2221–2230.

Avise, J. C., Arnold, J., Ball, R. M., Bermingham, E., Lamb, T., Neigel, J. E., Reeb, C. A., and Saunders, N. C. (1987). Intraspecific Phylogeography: The Mitochondrial DNA Bridge Between Population Genetics and Systematics. *Annual Review of Ecology and Systematics*, 18:489–522.

Bargelloni, L., Alarcon, J. A., Alvarez, M. C., Penzo, E., Magoulas, A., Palma, J., and Patarnello, T. (2005). The Atlantic–Mediterranean transition: Discordant genetic patterns in two seabream species, Diplodus puntazzo (Cetti) and Diplodus sargus (L.). *Molecular Phylogenetics and Evolution*, 36(3):523–535.

Barton, N. and Bengtsson, B. O. (1986). The barrier to genetic exchange between hybridising populations. *Heredity*, 57(3):357–376.

Beaumont, M. A., Zhang, W., and Balding, D. J. (2002). Approximate Bayesian computation in population genetics. *Genetics*, 162(4):2025–2035.

Boudinar, A. S., Chaoui, L., Quignard, J. P., Aurelle, D., and Kara, M. H. (2016). Otolith shape analysis and mitochondrial DNA markers distinguish three sand smelt species in the Atherina boyeri species complex in western Mediterranean. *Estuarine, Coastal and Shelf Science*, 182:202–210.

Bradbury, I., Burgetz, I., Coulson, M., Verspoor, E., Gilbey, J., Lehnert, S., Kess, T., Cross, T., Vasemägi, A., Solberg, M., Fleming, I., and McGinnity, P. (2020). Beyond hybridization: The genetic impacts of nonreproductive ecological interactions of salmon aquaculture on wild populations. *Aquaculture Environment Interactions*, 12:429–445.

Campo, D., Machado-Schiaffino, G., Perez, J., and Garcia-Vazquez, E. (2007). Phylogeny of the genus Merluccius based on mitochondrial and nuclear genes. *Gene*, 406(1):171–179.

Chan, A. H., Jenkins, P. A., and Song, Y. S. (2012). Genome-Wide Fine-Scale Recombination Rate Variation in Drosophila melanogaster. *PLOS Genetics*, 8(12):e1003090.

Charrier, G., Chenel, T., Durand, J. D., Girard, M., Quiniou, L., and Laroche, J. (2006). Discrepancies in phylogeographical patterns of two European anglerfishes (Lophius budegassa and Lophius piscatorius). *Molecular Phylogenetics and Evolution*, 38(3):742–754.

Chen, J., Glémin, S., and Lascoux, M. (2017). Genetic Diversity and the Efficacy of Purifying Selection across Plant and Animal Species. *Molecular Biology and Evolution*, 34(6):1417–1428.

Coyne, J. A. and Orr, H. A. (1989). Patterns of Speciation in Drosophila. *Evolution*, 43(2):362–381.

Cutter, A. D. (2012). The polymorphic prelude to Bateson–Dobzhansky–Muller incompatibilities. *Trends in Ecology & Evolution*, 27(4):209–218.

Delling, B., Noren, M., Kullander, S. O., and González, J. A. (2011). Taxonomic review of the genus Trisopterus(Teleostei: Gadidae) with recognition of the capelan Trisopterus capelanus as a valid species. *Journal of Fish Biology*, 79(5):1236–1260.

Divanach, P. and Kentouri, M. (1983). Donnees preliminaires sur la technique de production, la croissance et la survie des larves de marbre Lithognathus mormyrus. *Aquaculture*, 31(2):245–256.

Domingues, V. S., Faria, C., Stefanni, S., Santos, R. S., Brito, A., and Almada, V. C. (2007). Genetic divergence in the Atlantic–Mediterranean Montagu's blenny, Coryphoblennius galerita (Linnaeus 1758) revealed by molecular and morphological characters. *Molecular Ecology*, 16(17):3592–3605.

Duranton, M., Allal, F., Fraïsse, C., Bierne, N., Bonhomme, F., and Gagnaire, P.-A. (2018). The origin and remolding of genomic islands of differentiation in the European sea bass. *Nature Communications*, 9(1):1–11.

Duranton, M., Allal, F., Valière, S., Bouchez, O., Bonhomme, F., and Gagnaire, P.-A. (2020). The contribution of ancient admixture to reproductive isolation between European sea bass lineages. *Evolution letters*, 4(3):226–242.

Edwards, S. V., Robin, V. V., Ferrand, N., and Moritz, C. (2022). The Evolution of Comparative Phylogeography: Putting the Geography (and More) into Comparative Population Genomics. *Genome Biology and Evolution*, 14(1):evab176.

Espíndola, A., Ruffley, M., Smith, M. L., Carstens, B. C., Tank, D. C., and Sullivan, J. (2016). Identifying cryptic diversity with predictive phylogeography. *Proceedings of the Royal Society B: Biological Sciences*, 283(1841):20161529.

Faria, R., Weiss, S., and Alexandrino, P. (2006). A molecular phylogenetic perspective on the evolutionary history of Alosa spp. (Clupeidae). *Molecular Phylogenetics and Evolution*, 40(1):298–304.

Félix-Hackradt, F. C., Hackradt, C. W., Pérez-Ruzafa, Á., and García-Charton, J. A. (2013). Discordant patterns of genetic connectivity between two sympatric species, Mullus barbatus

(Linnaeus, 1758) and Mullus surmuletus (Linnaeus, 1758), in south-western Mediterranean Sea. *Marine Environmental Research*, 92:23–34.

Fraïsse, C., Popovic, I., Mazoyer, C., Romiguier, J., Loire, É., Simon, A., Galtier, N., Duret, L., Bierne, N., Vekemans, X., and Roux, C. (2020). DILS : Demographic Inferences with Linked Selection by using ABC. *bioRxiv*, page 2020.06.15.151597.

Gonzalez, E. G., Cunha, R. L., Sevilla, R. G., Ghanavi, H. R., Krey, G., and Bautista, J. M. (2012). Evolutionary history of the genus Trisopterus. *Molecular Phylogenetics and Evolution*, 62(3):1013–1018.

Green, R. E., Krause, J., Briggs, A. W., Maricic, T., Stenzel, U., Kircher, M., Patterson, N., Li, H., Zhai, W., Fritz, M. H.-Y., Hansen, N. F., Durand, E. Y., Malaspinas, A.-S., Jensen, J. D., Marques-Bonet, T., Alkan, C., Prüfer, K., Meyer, M., Burbano, H. A., Good, J. M., Schultz, R., Aximu-Petri, A., Butthof, A., Höber, B., Höffner, B., Siegemund, M., Weihmann, A., Nusbaum, C., Lander, E. S., Russ, C., Novod, N., Affourtit, J., Egholm, M., Verna, C., Rudan, P., Brajkovic, D., Kucan, Ž., Gušic, I., Doronichev, V. B., Golovanova, L. V., Lalueza-Fox, C., de la Rasilla, M., Fortea, J., Rosas, A., Schmitz, R. W., Johnson, P. L. F., Eichler, E. E., Falush, D., Birney, E., Mullikin, J. C., Slatkin, M., Nielsen, R., Kelso, J., Lachmann, M., Reich, D., and Pääbo, S. (2010). A Draft Sequence of the Neandertal Genome. *Science*, 328(5979):710–722.

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

Gutiérrez-García, T. A. and Vázquez-Domínguez, E. (2011). Comparative Phylogeography: Designing Studies while Surviving the Process. *BioScience*, 61(11):857–868.

Hanel, R., Westneat, M., and Sturmbauer, C. (2003). Phylogenetic Relationships, Evolution of Broodcare Behavior, and Geographic Speciation in the Wrasse Tribe Labrini. *Journal of molecular evolution*, 55:776–89.

Harris, K. and Nielsen, R. (2016). The Genetic Cost of Neanderthal Introgression. *Genetics*, 203(2):881–891.

Johannesson, K., Panova, M., Kemppainen, P., André, C., Rolán-Alvarez, E., and Butlin, R. K. (2010). Repeated evolution of reproductive isolation in a marine snail: Unveiling mechanisms of speciation. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 365(1547):1735–1747.

Karapanagiotidis, I. T., Kyritsi, S., Dretaki-Stamou, G., Psofakis, P., Neofytou, M. C., Mente, E., Vlahos, N., and Karalazos, V. (2021). The effect of different dietary protein levels on growth performance and nutrient utilization of snarpsnout sea bream (Diplodus puntazzo). *Aquaculture Research*, n/a(n/a).

Kelleher, J., Wong, Y., Wohns, A. W., Fadil, C., Albers, P. K., and McVean, G. (2019). Inferring whole-genome histories in large population datasets. *Nature Genetics*, 51(9):1330–1338.

Liang, M. and Nielsen, R. (2014). The lengths of admixture tracts. *Genetics*, 197(3):953–967.

Louisy, P. (2015). *Guide d'identification des poissons marins: Europe et Méditerranée 860 espèces, 1450 photos, 1400 dessins*. Ulmer, Paris, 3e éd. entièrement revue, complétée et mise à jour edition.

Martin, S. H., Davey, J. W., Salazar, C., and Jiggins, C. D. (2019). Recombination rate variation shapes barriers to introgression across butterfly genomes. *PLOS Biology*, 17(2):e2006288.

McGrath, C. (2022). Highlight: Comparative Population Genomics—Answering Old Questions with New Data. *Genome Biology and Evolution*, 14(1):evab278.

Milana, V., Sola, L., Congiu, L., and Rossi, A. R. (2008). Mitochondrial DNA in Atherina (Teleostei, Atheriniformes): Differential distribution of an intergenic spacer in lagoon and marine forms of Atherina boyeri. *Journal of Fish Biology*, 73(5):1216–1227.

Oikonomou, S., Tsakogiannis, A., Kriaridou, C., Danis, T., Manousaki, T., Chatziplis, D., Papandroulakis, N., Mylonas, C. C., Triantafyllidis, A., and Tsigenopoulos, C. S. (2021). First linkage maps and a pilot QTL analysis for early growth performance in common dentex (Dentex dentex) and sharpsnout seabream (Diplodus puntazzo). *Aquaculture Reports*, 21:100855.

Palma, J. and Andrade, J. P. (2002). Morphological study of Diplodus sargus, Diplodus puntazzo, and Lithognathus mormyrus (Sparidae) in the Eastern Atlantic and Mediterranean Sea. *Fisheries Research*, 57(1):1–8.

Patarnello, T., Volckaert, F. a. M. J., and Castilho, R. (2007). Pillars of Hercules: Is the Atlantic–Mediterranean transition a phylogeographical break? *Molecular Ecology*, 16(21):4426–4444.

Randi, E. (2008). Detecting hybridization between wild species and their domesticated relatives. *Molecular Ecology*, 17(1):285–293.

Riquet, F., Liautard-Haag, C., Woodall, L., Bouza, C., Louisy, P., Hamer, B., Otero-Ferrer, F., Aublanc, P., Béduneau, V., Briard, O., El Ayari, T., Hochscheid, S., Belkhir, K., Arnaud-Haond, S., Gagnaire, P.-A., and Bierne, N. (2019). Parallel pattern of differentiation at a genomic island shared between clinal and mosaic hybrid zones in a complex of cryptic seahorse lineages. *Evolution*, 73(4):817–835.

Romiguier, J., Gayral, P., Ballenghien, M., Bernard, A., Cahais, V., Chenuil, A., Chiari, Y., Dernat, R., Duret, L., Faivre, N., Loire, E., Lourenco, J. M., Nabholz, B., Roux, C., Tsagkogeorga, G., a. T. Weber, A., Weinert, L. A., Belkhir, K., Bierne, N., Glémin, S., and

Galtier, N. (2014). Comparative population genomics in animals uncovers the determinants of genetic diversity. *Nature*, 515(7526):261–263.

Roux, C., Fraïsse, C., Romiguier, J., Anciaux, Y., Galtier, N., and Bierne, N. (2016). Shedding Light on the Grey Zone of Speciation along a Continuum of Genomic Divergence. *PLOS Biology*, 14(12):e2000234.

Shanks, A. L. (2009). Pelagic Larval Duration and Dispersal Distance Revisited. *The Biological Bulletin*, 216(3):373–385.

Sheehy, M., Bannister, R., Wickins, J., and Shelton, P. (2011). New perspectives on the growth and longevity of the European lobster (Homarus gammarus). *Canadian Journal of Fisheries and Aquatic Sciences*, 56:1904–1915.

Simakov, O., Bredeson, J., Berkoff, K., Marletaz, F., Mitros, T., Schultz, D. T., O'Connell, B. L., Dear, P., Martinez, D. E., Steele, R. E., Green, R. E., David, C. N., and Rokhsar, D. S. (2022). Deeply conserved synteny and the evolution of metazoan chromosomes. *Science Advances*, 8(5):eabi5884.

Speidel, L., Forest, M., Shi, S., and Myers, S. R. (2019). A method for genome-wide genealogy estimation for thousands of samples. *Nature Genetics*, 51(9):1321–1329.

Szymura, J. M. (1983). Genetic differentiation between hybridizing species Bombina bombina and Bombina variegata (Salientia, Discoglossidae) in Poland*. *Amphibia-Reptilia*, 4(2):137–145.

Todesco, M., Pascual, M. A., Owens, G. L., Ostevik, K. L., Moyers, B. T., Hübner, S., Heredia, S. M., Hahn, M. A., Caseys, C., Bock, D. G., and Rieseberg, L. H. (2016). Hybridization and extinction. *Evolutionary Applications*, 9(7):892–908.

Turan, C. (2006). Phylogenetic relationships of Mediterranean Mullidae species (Perciformes) inferred from genetic and morphologic data. *Scientia Marina*, 70(2):311–318.

Vella, A., Vella, N., and Acosta-Díaz, C. (2021). Resurrection of the Butterfly-winged Comber, Serranus papilionaceus Valenciennes, 1832 (Teleostei, Serranidae) and its phylogenetic position within genus Serranus. *ZooKeys*, 1017:111–126.

Wallace, A. R. (1865). *On the Phenomena of Variation and Geographical Distribution as Illustrated by the Papilionidae of the Malayan Region. Read March 17, 1864.* London,.

Walsh, J. B. (1982). Rate of Accumulation of Reproductive Isolation by Chromosome Rearrangements. *The American Naturalist*, 120(4):510–532.

White, E. P., Ernest, S. K. M., Kerkhoff, A. J., and Enquist, B. J. (2007). Relationships between body size and abundance in ecology. *Trends in Ecology & Evolution*, 22(6):323–330.

Whitney, K. D., Broman, K. W., Kane, N. C., Hovick, S. M., Randell, R. A., and Rieseberg, L. H. (2015). QTL mapping identifies candidate alleles involved in adaptive introgression and range expansion in a wild sunflower. *Molecular ecology*, 24(9):2194–2211.

Wohns, A. W., Wong, Y., Jeffery, B., Akbari, A., Mallick, S., Pinhasi, R., Patterson, N., Reich, D., Kelleher, J., and McVean, G. (2021). A unified genealogy of modern and ancient genomes.

Zardoya, R., Castilho, R., Grande, C., Favre-Krey, L., Caetano, S., Marcato, S., Krey, G., and Patarnello, T. (2004). Differential population structuring of two closely related fish species, the mackerel (Scomber scombrus) and the chub mackerel (Scomber japonicus), in the Mediterranean Sea. *Molecular Ecology*, 13(7):1785–1798.

# ANNEX OF CHAPTER 1

# Supplementary Material to Barry and al. "Age-specific survivorship and fecundity shape genetic diversity in marine fishes"

## Contents

**Additional methods** **2**
Sampling and sequencing . . . 2
Collection of life history traits database . . . 2
Estimation of genetic diversity with GenomeScope . . . 2
Forward simulations . . . 4
Evaluating the impact of life tables beyond marine fish . . . 4

**Additional results** **4**
Whole-genome resequencing data set . . . 4
Intraspecific variation in genetic diversity . . . 5

**Additional tables** **6**

**Additional figures** **7**

# Additional methods

## Sampling, DNA extraction, whole-genome sequencing and reads quality control

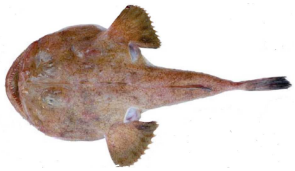
We sampled 16 marine teleostean fish species presenting a wide diversity of life history strategies expected to affect genetic diversity. For 12 of these species, 20 individuals were sampled (5 per location). For the 4 other species, the total number of samples ranged from 10 to 19 (Table 1). Individuals were either sampled from landings in local fish markets, captured in the field (using hand nets, lure fishing, spearfishing or beach seines) or provided by collaborators. The majority of the sampling was done in 2018 and 2019. Whole-genomic DNA was extracted from fin or tissue clips stored in 95% ethanol using the NucleoSpin Tissue Kit (Macherey-Nagel) and treated with RNAse A to remove residual RNA. Double-stranded nucleic acid concentration was quantified using Qubit2.0 and standardized to 20ng per $\mu$l. Individual whole-genome sequencing libraries were prepared following the Illumina TruSeq DNA PCR-Free Protocol and sequenced by Genewiz Inc (USA). Libraries were quantified and multiplexed by groups of 40 individuals and sequenced on two S4 flow cells on a NovaSeq6000 instrument (Illumina) to generate 150 bp paired-end reads, targetting an average read depth of 20X per individual. Raw reads were preprocessed with `fastp` v.0.20.0 (Chen et al., 2018) using default parameters, allowing quality control, filtering by quality, length and complexity, and adapter trimming to be performed in a single step. Base correction was performed using a quality comparison between overlapping bases of paired-end reads, and polyG tail trimming was enabled to correct for artefactual G repetitions occurring in Novaseq read tails.

## Collection of life history traits database

As growth is indeterminate in fish, we defined adult body size as the infinite length, $L_{inf}$ determined by the Von Bertalanffy equation ($L_t = L_{inf}[1 - exp^{-K(t-t0)}]$), that links individual body size $L_t$ to age $t$, with $K$ a parameter defining the shape of this relationship (Pauly et al., 1987). We estimated adult body size as the median of all $L_{inf}$ values reported for each species in the online database Fishbase (Froese et al., 2000). As $L_{inf}$ was not documented in Fishbase for *D. puntazzo* and *C. galerita*, we took the median of the values reported in (Kraljević et al., 2007) and (Domínguez-Seoane et al., 2006) for *D. puntazzo*, and the maximum length observed in (Milton, 1983) for *C. galerita*. Trophic level was retrieved from Fishbase. Fecundity was defined as the absolute fecundity, i.e. the mean number of eggs in an ovary of a female in a single spawning event. Females may spawn several times during one reproductive season (Ganias et al., 2003; Murua and Motos, 2006), so absolute fecundity is not the value most directly relevant to global genetic diversity. However, it is the most commonly reported in the literature as the number of spawnings events per reproductive season is difficult to measure. Because fecundity is proportional to individual body size, we computed fecundity at infinite length, $L_{inf}$. Propagule size was determined following Romiguier et al. (2014), as the size of the dispersal stage that becomes independent of the parents. For all species of this study, this corresponded to egg diameter, except for brooders, for which we used hatching size. All propagule size data were retrieved from species-specific references. Age at maturity was defined as the age at which 50% of the population is mature. Values for age at maturity were taken from Tsikliras and Stergiou (2015) for seven species while other values were retrieved from species-specific references. Likewise, lifespan values were taken from (Tsikliras and Stergiou, 2015) for six species and completed with specific references. Finally, adult lifespan was defined as $Lifespan - Age\ at\ Maturity$ (Waples et al., 2013).

## Estimation of genetic diversity with GenomeScope

Provided a sufficient average coverage depth (e.g. 20X), `GenomeScope` evaluates the fraction of heterozygous sites from the ratio of the height of the heterozygous to the homozygous $k$-mer peak, occurring at 50% (i.e. 10X) and 100% (i.e. 20X) of the average coverage depth, respectively. The number of different possible $k$-mers (and thus the precision of the method) increases with $k$, but so does the runtime and the probability of "wrong" $k$-mers due to sequencing errors. We set $k = 21$ as recommended by `GenomeScope` and performed a sensitivity analysis by estimating genetic diversity and genome size for one individual of *D. labrax* using $k$ from 17 to 25 (Fig S4).



*Lophius budegassa*:
high genome size and
low diversity (0.225 %)



*Mullus surmuletus*:
low genome size and
high diversity (1.135 %)

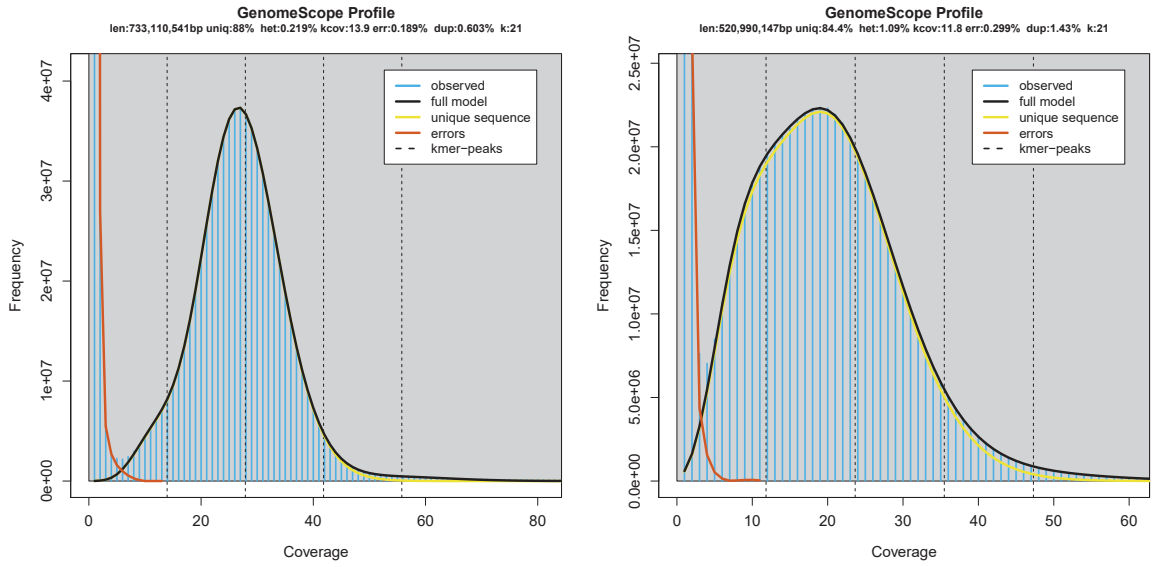



Figure S1: $k - mer$ **frequency-coverage relationship and estimation by `GenomeScope` v.1.0 (Vurture et al., 2017) for two species, *L. budegassa* and *M.surmuletus***

In order to assess the reliability of `GenomeScope`, the 20 resequenced genomes for two species (*S. pilchardus* and *D. labrax*) were aligned with `bwa-mem` v.0.7.17 (Li and Durbin, 2009) to the reference genomes retrieved from Louro et al. (2019) and Tine et al. (2014) for *S. pilchardus* and *D. labrax*, respectively. We then removed PCR duplicates with the Picard tools `MarkDuplicates` v.2.23.2. We followed the best-practice pipeline in `GATK` v.4.1.6.0 for variant calling (Poplin et al., 2018): we ran `HaplotypeCaller` with default options to generate individual GVCFs files, stored them in a database with `GenomicsDBImport` and finally computed VCF files with `GenotypeGVCFs`. We didn't apply post variant calling filtering steps, such as hard filters on genotype quality scores or Hardy-Weinberg Equilibrium criterion, in order to

avoid potential bias in the comparison of genetic diversity between species with very different rates of heterozygosity. However, we assume that possible bias due to the absence of variant filtering should not impact differences among individuals within each species. However, we generated VCF files with all sites, including non-variant ones, to avoid underestimation of genetic diversity due to the assumption that missing sites are homozygous for the reference allele. We estimated species genetic diversity as:

$$\pi = \frac{\sum^{L} \sum_{i<j} k_{ij}}{\sum^{L} \binom{n}{2}} \tag{S1}$$

where $\sum_{i<j} k_{ij}$ is the number of pairwise nucleotide differences between all haplotypes at a given site, $\binom{n}{2}$ is the total number of pairwise nucleotide comparison between all haplotypes at a given site, and $L$, the number of sites. We thus exclude missing data from the analysis. However, we include multiallelic sites as removing these sites may underestimate genetic diversity, especially for species with high genetic diversity.

## Forward simulations

We used `SLiM` v.3.3.1 (Haller and Messer, 2017) to perform forward simulations to estimate genetic diversity at mutation-drift equilibrium incorporating age-specific survival and fecundity. For each individual, the number of offspring produced per year was determined by a Poisson distribution with mean $\lambda_{S,a}$ specific to each species and each age. Age at first maturity was set to 1 for all simulations. To keep population size constant in these non Wright-Fisher forwards simulations, we introduced a carrying capacity parameter, allowing population size to fluctuate around this capacity (Fig S17 - S24). We arbitrarily set this parameter to $N = 2000$, and simulated non-recombining 1Mb loci with a mutation rate of $\mu = 1e^{-7}$. Each simulation was run for 25000 years, which was long enough for genetic diversity to reach mutation-drift equilibrium (Fig S25 - S32). For each simulation, we estimated the mean genetic diversity (i.e., the proportion of heterozygous sites along the 1Mb locus) over the last 10000 years after checking that an equilibrium has been reached. For each species, we ran 50 replicates and defined the genetic diversity predicted by a given simulation scenario as the mean genetic diversity at equilibrium averaged over the 50 replicates.

## Evaluating the impact of life tables beyond marine fish

We explored the impact of alternative fecundity-age models on the relationship between adult lifespan and $\frac{N_e}{N}$ using three additional biologically realistic models: linear ($F_{Age} = a \times Age + b$), polynomial ($F_{Age} = [Age - AgeMat][(AgeMat + Lifespan - Age)^2]$, common in mammal) (Gage, 2001) and power-law ($F_{Age} = Age^f$). For the linear and the polynomial model, $f$ describes the maximum fecundity at lifespan and age with the highest fecundity, respectively (i.e. higher absolute values of $f$ correspond to higher differences in fecundity between low and high fecund ages for both models). $f$ was between -1 and 1 for the linear and the polynomial model. For the power-law model, we took values of $f$ from $-5$ to 5.

# Additional results

## Whole-genome resequencing data set

We resequenced 300 individual genomes from 16 marine teleostean species, generating from $59.86 \times 10^6$ to $200.92 \times 10^6$ reads per individual (mean = $129 \times 10^6$, sd = $20 \times 10^6$, Fig S2). The

925 read quality score (Q30 rate) ranged between 88% and 94% (mean = 92.4%, sd = 1.1) and the
926 duplication rate lied between 5 and 15% (mean = 10.8%, sd = 2.6) (Fig S2). GC content was
927 moderately variable among species and highly consistent among individuals of the same species,
928 except for one individual of *S. cabrilla*, *D. puntazzo* and *M. surmuletus* that showed a marked
929 discrepancy with the overall GC content of their species (Fig S2). These three individuals
930 were thus removed from downstream analyses to avoid potential issues due to contamination
931 or poor sequencing quality (see discussion).

## 932 Intraspecific variation in genetic diversity

933 Although, we detected slight differences in within-species genetic diversity between individuals
934 of different basins, this does not affect the results linking species' genetic diversity to life-
935 history traits. However, regarding variation in within-species genetic diveristy, we distinguished
936 two clusters: the first one included 9 species with generally lower genetic diversity in the
937 Mediterranean than Atlantic localities, while the opposite was observed in the second cluster
938 (7 species) (Fig 1C). The species of the second cluster are often found in coastal habitats,
939 lagoons, estuaries whereas species of the first cluster are rather pelagic, epi-pelagic or benthic
940 species. The only exception was the presence of *H. guttulatus* in the Atlantic cluster.

# Additional tables

| Dataset | Predictor | *p-value* | Pseudo $R^2$ | Slope estimate (± 95% interval) |
|---|---|---|---|---|
| Whole data set | Body size | 0.119 | 0.192 | $-0.006(-0.014; 0.002)$ |
| | Trophic level | 0.676 | 0.012 | $-0.091(-0.524; 0.343)$ |
| | Propagule size | 0.562 | 0.015 | $-0.014(-0.062; 0.034)$ |
| | Fecundity | 0.653 | 0.013 | $-1.22e^{-5}(-6.63e^{-5}; 4.20e^{-5})$ |
| | Lifespan | **0.0107** | 0.438 | $-0.062(-0.111; -0.013)$ |
| | Adult lifespan | **0.0070** | 0.429 | $-0.089(-0.156; -0.023)$ |
| | Hermaphroditism | 0.434 | 0.034 | $0.1779(-0.278; 0.633)$ |
| | Parental Care | 0.274 | 0.075 | $-0.273(-0.772; 0.226)$ |
| No parental care | Body size | $\mathbf{6.60e^{-5}}$ | 0.616 | $-0.014(-0.021; -0.007)$ |
| | Trophic level | 0.256 | 0.093 | $-0.326(-0.902; 0.251)$ |
| | Propagule size | 0.170 | 0.175 | $-0.518(-1.273; 0.237)$ |
| | Fecundity | 0.390 | 0.056 | $-2.51e^{-5}(-8.35e^{-5}; 3.33e^{-5})$ |
| | Lifespan | $\mathbf{1.017e^{-7}}$ | 0.851 | $-0.095(-0.131; -0.060)$ |
| | Adult lifespan | $\mathbf{1.65e^{-7}}$ | 0.829 | $-0.129(-0.179; -0.080)$ |
| | Hermaphroditism | 0.454 | 0.044 | $0.206(-0.345; 0.757)$ |

Table S1: **Statistical relationships between species genetic diversity and life history traits** - Genetic diversity was fitted to 6 quantitative (body size, trophic level, propagule size, fecundity, lifespan and adult lifespan) and two qualitative predictors (hermaphroditism and parental care) with a beta regression model using the `betareg` R package (Zeileis and Hothorn, 2002). In the upper part of the table, regressions were performed with the whole dataset, while in the lower part only the 11 non-brooding species were considered.

Table S2: **Mapping and variant calling statistics for *D. labrax* and *S. pilchardus* individuals** - For each, individual, number and percentage of reads mapped with `bwa-mem` v.0.7.17 (Li and Durbin, 2009). Individual GCVF files were created from bam files with `HaplotypeCaller` from `GATK` v.4.1.6.0 (Poplin et al., 2018), then stored in a GCVF database with `GenomicsDBImport`, and VCF files were finally generated with `GenotypeGVCFs`. Individual heterozygosity was estimated with a custom script. $O_{het}$ is the number of observed heterozygous positions at $N_{sites}$ number of variable sites. $Het_{bwa+GATK}$ and $Het_{GenomeScope}$ correspond to genome-wide average heterozygosity estimated with the variant calling and `GenomeScope` approaches respectively. Sample names indicate geographical origin (Li = Gulf of Lion, Mu = Costa Calida, Fa = Algarve, Ga = Bay of Biscay).

| Species | Sample | Reads mapped | % reads mapped | $O_{het}$ | $N_{sites}$ | $Het_{bwa+GATK}$ | $Het_{GenomeScope}$ |
|---|---|---|---|---|---|---|---|
| *D. labrax* | DlabrFa1 | 113 750 879 | 98.50 | 2 147 840 | 503 641 870 | 0.4264618 | 0.4023185 |
| | DlabrFa3 | 113 675 336 | 98.47 | 2 208 200 | 503 847 656 | 0.4382674 | 0.4056315 |
| | DlabrFa4 | 119 760 663 | 98.26 | 2 168 714 | 503 981 403 | 0.4303163 | 0.3963415 |
| | DlabrFa5 | 111 632 052 | 98.40 | 2 280 744 | 503 764 199 | 0.4527404 | 0.4370325 |
| | DlabrFa6 | 118 251 538 | 98.39 | 2 153 188 | 503 902 634 | 0.4273024 | 0.391308 |
| | DlabrMu1 | 111 356 113 | 98.20 | 1 719 388 | 502 790 222 | 0.3419693 | 0.2962895 |
| | DlabrMu2 | 78 923 334 | 96.72 | 1 930 312 | 498 084 060 | 0.3875474 | 0.3733955 |
| | DlabrMu3 | 100 325 589 | 98.00 | 1 874 416 | 502 625 989 | 0.3729246 | 0.332756 |
| | DlabrMu4 | 113 701 410 | 98.29 | 1 955 294 | 503 912 767 | 0.3880223 | 0.355843 |
| | DlabrMu6 | 117 131 984 | 98.11 | 1 923 386 | 503 422 444 | 0.3820620 | 0.3329545 |
| | DlabrLi1 | 100 519 118 | 98.16 | 1 919 555 | 503 396 155 | 0.3813209 | 0.3627435 |
| | DlabrLi2 | 103 717 839 | 98.18 | 1 926 045 | 503 259 727 | 0.3827139 | 0.3566875 |
| | DlabrLi3 | 101 815 078 | 98.18 | 1 917 988 | 503 399 019 | 0.3810075 | 0.3713805 |
| | DlabrLi4 | 94 399 356 | 98.21 | 1 920 186 | 502 848 166 | 0.3818620 | 0.3520245 |
| | DlabrLi5 | 115 515 514 | 98.06 | 1 918 720 | 503 724 011 | 0.3809070 | 0.354731 |
| | DlabrGa2 | 129 379 987 | 98.45 | 2 114 798 | 503 898 866 | 0.4196870 | 0.3770105 |
| | DlabrGa3 | 118 863 303 | 98.02 | 2 082 274 | 503 524 710 | 0.4135396 | 0.388246 |
| | DlabrGa4 | 125 269 167 | 98.02 | 2 085 977 | 503 675 656 | 0.4141508 | 0.380342 |
| | DlabrGa5 | 121 308 016 | 98.05 | 2 093 356 | 503 693 457 | 0.4156012 | 0.3807775 |
| | DlabrGa6 | 113 020 888 | 98.04 | 2 150 834 | 503 563 171 | 0.4271230 | 0.379656 |
| *S. pilchardus* | SpilcFa1 | 98 789 082 | 96.42 | 7 128 735 | 604 856 771 | 1.178582 | 1.331 |
| | SpilcFa3 | 109 782 816 | 96.24 | 7 452 162 | 613 645 048 | 1.214409 | 1.340575 |
| | SpilcFa4 | 97 444 013 | 95.07 | 7 195 048 | 609 516 704 | 1.180451 | 1.347235 |
| | SpilcFa5 | 110 357 654 | 95.67 | 7 545 357 | 615 252 143 | 1.226385 | 1.40959 |
| | SpilcFa6 | 104 223 183 | 95.90 | 7 500 848 | 615 157 704 | 1.219337 | 1.41341 |
| | SpilcMu1 | 124 885 493 | 96.20 | 7 471 544 | 617 558 276 | 1.209852 | 1.500415 |
| | SpilcMu2 | 109 948 539 | 95.70 | 7 272 037 | 615 199 861 | 1.182061 | 1.474785 |
| | SpilcMu3 | 108 564 997 | 92.46 | 7 272 929 | 614 357 522 | 1.183827 | 1.537 |
| | SpilcMu4 | 122 541 302 | 95.47 | 7 283 499 | 616 999 758 | 1.180470 | 1.446285 |
| | SpilcMu6 | 105 613 487 | 96.25 | 7 172 809 | 614 547 029 | 1.167170 | 1.40057 |
| | SpilcLi2 | 121 934 021 | 95.06 | 7 123 340 | 614 907 641 | 1.158441 | 1.373815 |
| | SpilcLi3 | 126 015 301 | 95.03 | 7 581 636 | 617 017 545 | 1.228755 | 1.417255 |
| | SpilcLi4 | 116 445 615 | 95.68 | 7 350 852 | 615 226 870 | 1.194820 | 1.37415 |
| | SpilcLi5 | 119 470 404 | 95.48 | 7 097 138 | 612 517 142 | 1.158684 | 1.336275 |
| | SpilcLi6 | 112 805 605 | 96.10 | 6 835 691 | 611 663 555 | 1.117557 | 1.279095 |
| | SpilcGa1 | 99 664 874 | 93.16 | 7 195 528 | 612 305 296 | 1.175154 | 1.5632 |
| | SpilcGa3 | 106 538 562 | 95.80 | 7 353 467 | 615 590 542 | 1.194539 | 1.65899 |
| | SpilcGa4 | 97 977 975 | 90.70 | 6 894 734 | 609 669 475 | 1.130897 | 1.895515 |
| | SpilcGa5 | 102 032 394 | 95.10 | 7 137 835 | 612 596 676 | 1.165177 | 1.592515 |
| | SpilcGa6 | 97 318 612 | 87.45 | 7 129 996 | 611 349 731 | 1.166271 | 1.97076 |

| Dataset | Predictor | *p-value* | | | Pseudo $R^2$ | | | Slope estimate (± 95% interval) | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | All | Med. | Atl. | All | Med. | Atl. | All. | Med. | Atl. |
| Whole data set | Body size | 0.119 | 0.105 | 0.159 | 0.192 | 0.206 | 0.169 | $-0.006$ $(-0.014; 0.002)$ | $-0.006$ $(-0.014; 0.001)$ | $-0.005$ $(-0.013; 0.002)$ |
| | Trophic level | 0.676 | 0.671 | 0.672 | 0.012 | 0.012 | 0.012 | $-0.091$ $(-0.524; 0.343)$ | $-0.094$ $(-0.536; 0.349)$ | $-0.090$ $(-0.516; 0.336)$ |
| | Propagule size | 0.562 | 0.608 | 0.450 | 0.015 | 0.012 | 0.032 | $-0.014$ $(-0.062; 0.034)$ | $-0.013$ $(-0.061; 0.036)$ | $-0.018$ $(-0.066; 0.030)$ |
| | Fecundity | 0.653 | 0.605 | 0.723 | 0.013 | 0.017 | 0.008 | $-1.22e^{-5}$ $(-6.63e^{-5}; 4.20e^{-5})$ | $-1.44e^{-5}$ $(-7.00e^{-5}; 4.13e^{-5})$ | $-9.2e^{-6}$ $(-6.18e^{-5}; 4.35e^{-5})$ |
| | Lifespan | **0.0107** | **0.0085** | **0.0219** | 0.438 | 0.453 | 0.404 | $-0.062$ $(-0.111; -0.013)$ | $-0.0652$ $(-0.115; -0.016)$ | $-0.056$ $(-0.104; -0.007)$ |
| | Adult lifespan | **0.0070** | **0.0055** | **0.0216** | 0.429 | 0.444 | 0.374 | $-0.089$ $(-0.156; -0.023)$ | $-0.093$ $(-0.161; -0.026)$ | $-0.077$ $(-0.144; -0.0010)$ |
| | Hermaphroditism | 0.434 | 0.465 | 0.350 | 0.034 | 0.03 | 0.050 | 0.1779 $(-0.278; 0.633)$ | 0.1701 $(-0.296; 0.636)$ | 0.2069 $(-0.2359; 0.6497)$ |
| | Parental Care | 0.274 | 0.287 | 0.168 | 0.075 | 0.071 | 0.121 | $-0.273$ $(-0.772; 0.226)$ | $-0.271$ $(-0.781; 0.239)$ | $-0.335$ $(-0.822; 0.152)$ |
| No parental care | Body size | $\mathbf{6.60e^{-5}}$ | $\mathbf{3.85e^{-5}}$ | $\mathbf{8.17e^{-5}}$ | 0.616 | 0.636 | 0.607 | $-0.014$ $(-0.021; -0.007)$ | $-0.014$ $(-0.021; -0.007)$ | $-0.014$ $(-0.021; -0.007)$ |
| | Trophic level | 0.256 | 0.245 | 0.254 | 0.093 | 0.097 | 0.095 | $-0.326$ $(-0.902; 0.251)$ | $-0.340$ $(-0.925; 0.244)$ | $-0.328$ $(-0.903; 0.247)$ |
| | Propagule size | 0.170 | 0.155 | 0.200 | 0.175 | 0.184 | 0.157 | $-0.518$ $(-1.273; 0.237)$ | $-0.544$ $(-1.310; 0.221)$ | $-0.484$ $(-1.240; 0.272)$ |
| | Fecundity | 0.390 | 0.353 | 0.401 | 0.056 | 0.065 | 0.054 | $-2.51e^{-5}$ $(-8.35e^{-5}; 3.33e^{-5})$ | $-2.76e^{-5}$ $(-8.71e^{-5}; 3.19e^{-5})$ | $-2.44e^{-5}$ $(-8.26e^{-5}; 3.37e^{-5})$ |
| | Lifespan | $\mathbf{1.017e^{-7}}$ | $\mathbf{7.28e^{-8}}$ | $\mathbf{5.451e^{-7}}$ | 0.851 | 0.856 | 0.832 | $-0.095$ $(-0.131; -0.060)$ | $-0.097$ $(-0.134; -0.061)$ | $-0.094$ $(-0.131; -0.056)$ |
| | Adult lifespan | $\mathbf{1.65e^{-7}}$ | $\mathbf{1.36e^{-7}}$ | $\mathbf{8.778e^{-7}}$ | 0.829 | 0.837 | 0.805 | $-0.129$ $(-0.179; -0.080)$ | $-0.131$ $(-0.181; -0.081)$ | $-0.127$ $(-0.178; -0.075)$ |
| | Hermaphroditism | 0.454 | 0.407 | 0.489 | 0.044 | 0.053 | 0.038 | 0.206 $(-0.345; 0.757)$ | 0.231 $(-0.326; 0.787)$ | 0.191 $(-0.360; 0.741)$ |

Table S3: **Statistical relationships between different estimations of species genetic diversity and life history traits** - Genetic diversity, either estimated from all individuals (all), individuals from Mediterranean Sea (Med.), Atlantic Ocean (Atl.) was fitted to 6 quantitative (body size, trophic level, propagule size, fecundity, lifespan and adult lifespan) and two qualitative predictors (hermaphroditism and parental care) with a beta regression model using the `betareg` R package (Zeileis and Hothorn, 2002). In the upper part of the table, regressions were performed with the whole dataset, while in the lower part only the 11 non-brooding species were considered.

# Additional figures

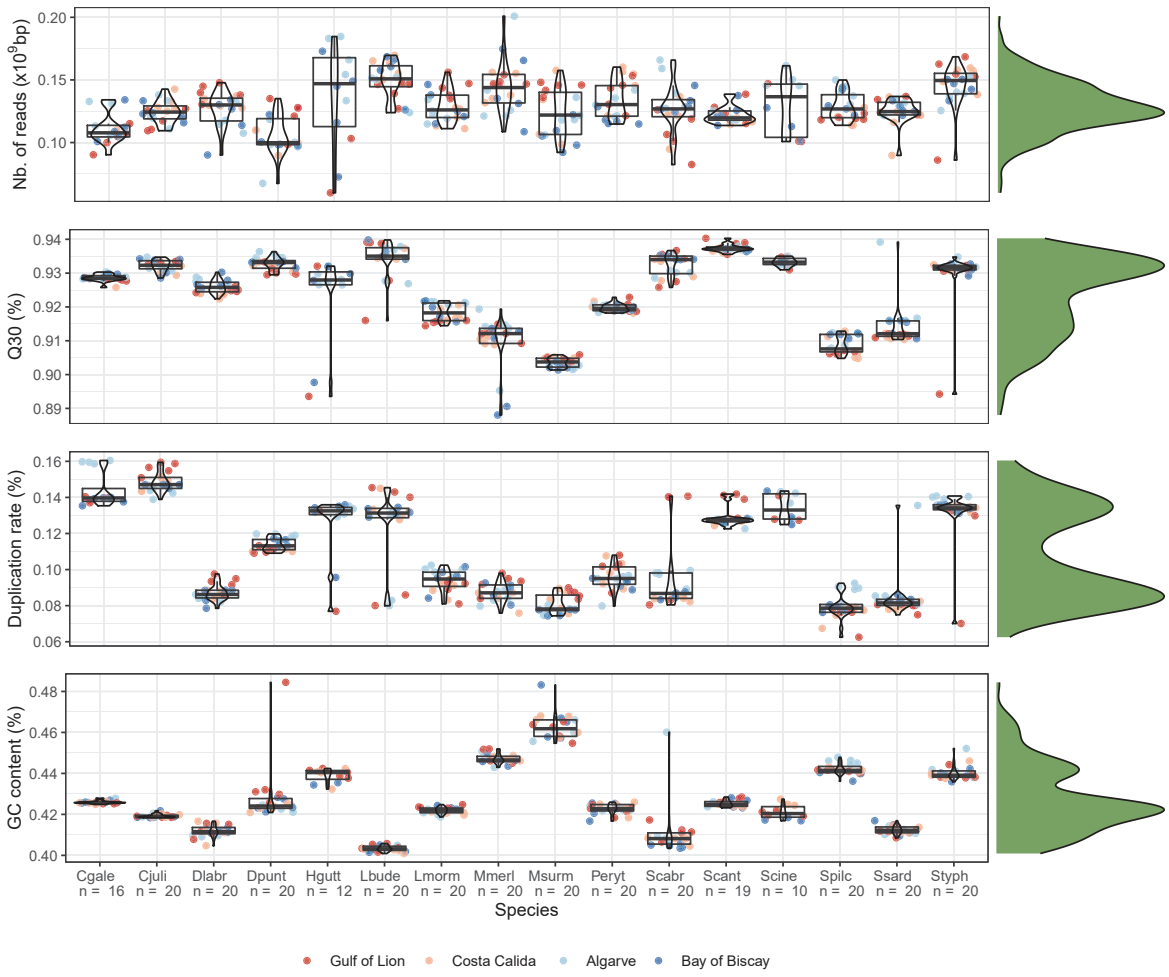



Figure S2: **Number of reads ($10^9$ bp), percentage of reads with quality superior to Q30, duplication rate and GC content after filtering, correcting and trimming steps carried out with `fastp` v.0.20.0 (Chen et al., 2018)**. Each point represents an individual unassembled genome: species are represented on the *y*-axis. Colors represent sampling locations: mediterranean locations in warm colors (Gulf of Lion in dark red, Murcia in pale red), atlantic locations in cold colors (Faro in pale blue, Bay of Biscay in dark blue). Overall distributions of each parameters are represented on the right side of each panel. Cgale = *Coryphoblennius galerita*, Cjuli = *Coris julis*, Dlabr = *Dicentrarchus labrax*, Dpunt = *Diplodus puntazzo*, Hgutt = *Hippocampus guttulatus*, Lbude = *Lophius budegassa*, Lmorm = *Lithognathus mormyrus*, Mmerl = *Merluccius merluccius*, Msurm = *Mullus surmuletus*, Peryt = *Pagellus erythrinus*, Scabr = *Serranus cabrilla*, Scant = *Spondyliosoma cantharus*, Scine = *Symphodus cinereus*, Spilc = *Sardina pilchardus*, Ssard = *Sarda sarda*, Styph = *Syngnathus typhle*.

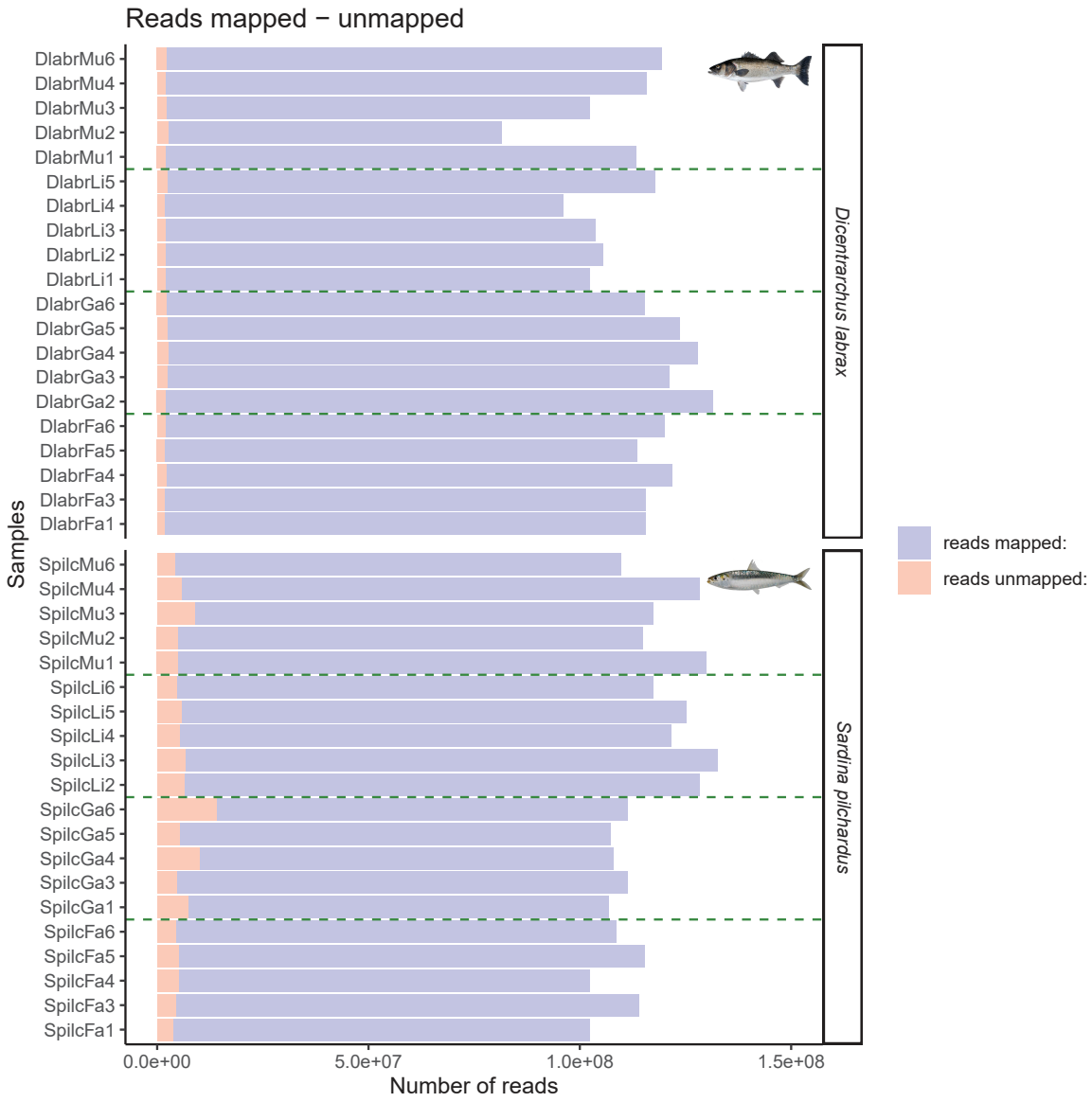



Figure S3: **Mapping statistics** - Number of reads mapped in red and unmapped blue with `bwa` v.0.7.17 for the 20 indiviuals of *D. labrax* and *S. pilchardus*.

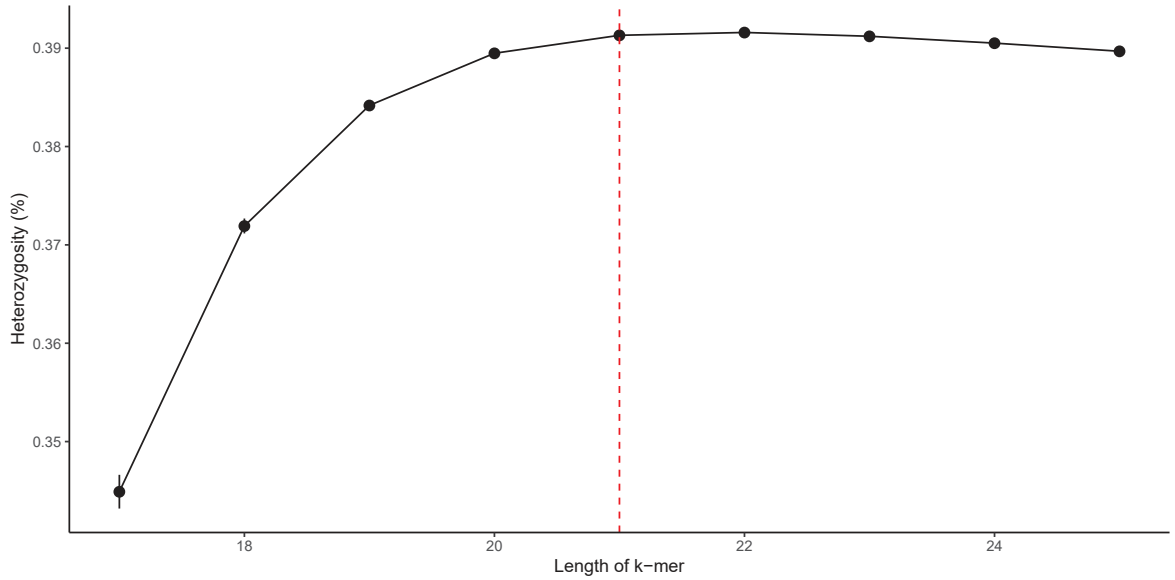



Figure S4: **Effect of $k-mer$ length on genetic diversity estimation with `GenomeScope` v.1.0 (Vurture et al., 2017)** - Genetic diversity was estimated with `GenomeScope` with different values of $k-mer$ length (17,19,21,23 and 25) on one European sea bass individual (*D. labrax*), all other parameters being equal (as detailed in the main text). The red vertical dashed line represents the value used in this study (21).

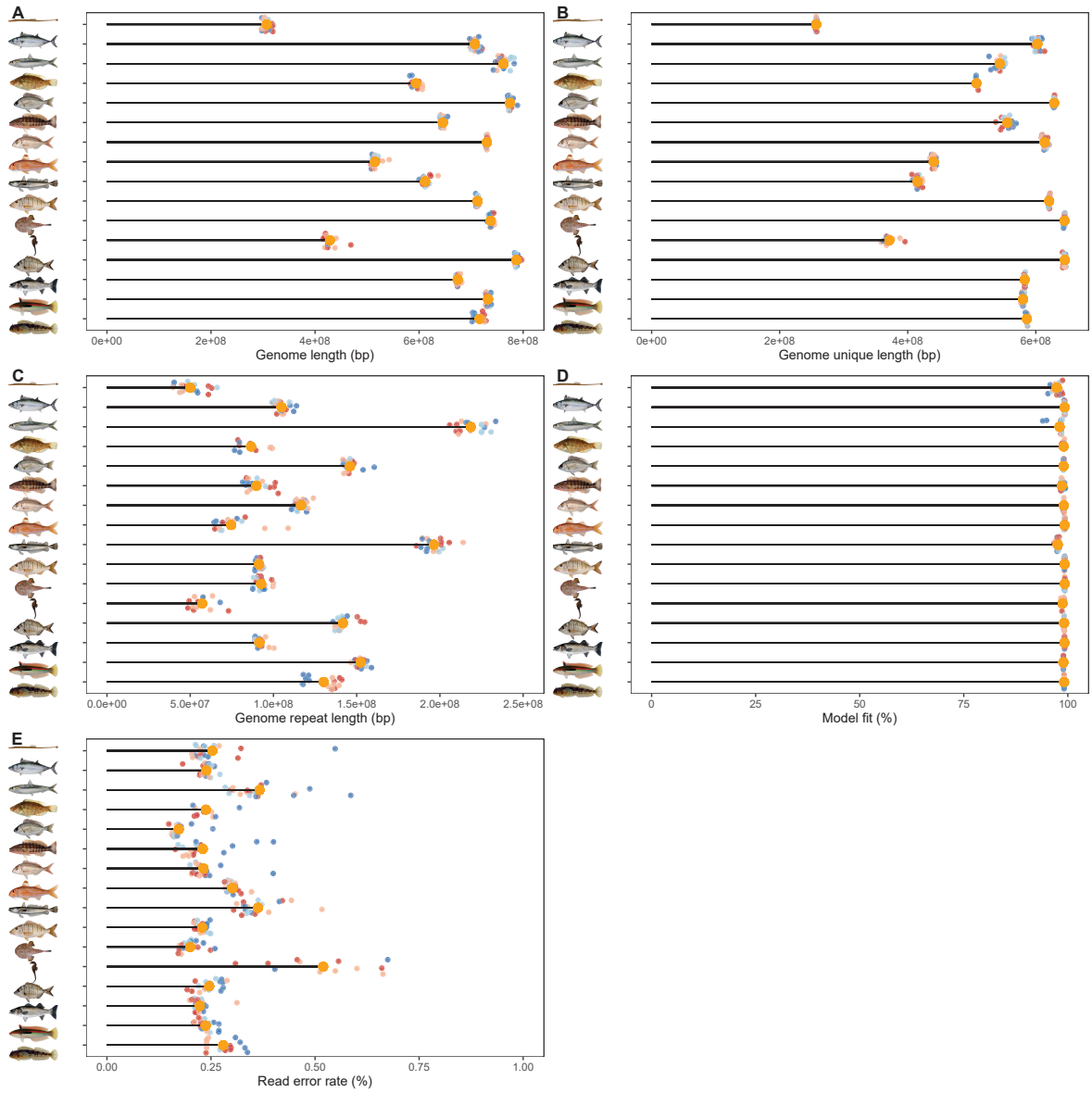



Figure S5: **Individual whole-genome sequences features estimated with `GenomeScope` v.1.0 (Vurture et al., 2017)**. A) Genome length (number of base pairs), B) genome unique length (number of base pairs), C) genome repeat length (number of base pairs), D) read error rate (%) and E) model fit (%). Individual feature and species median feature is represented in yellow in each panel. Species illustrations were retrieved from Iglésias (2013) with permissions.

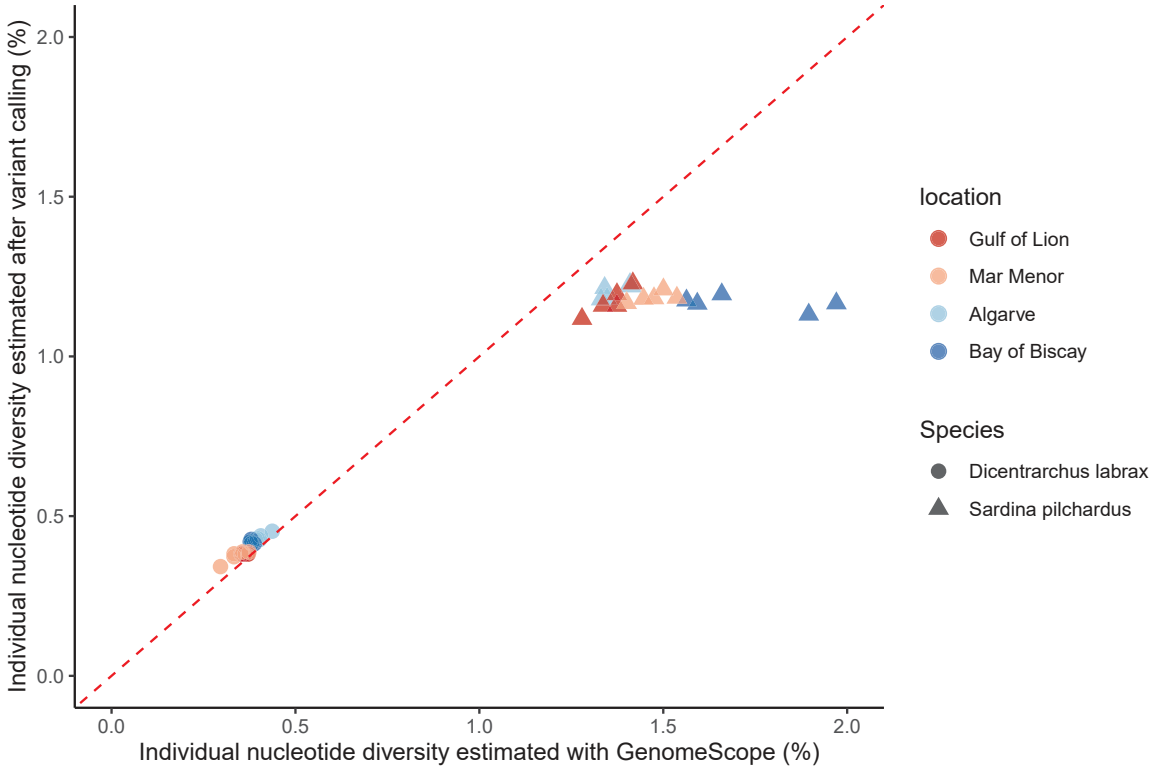



Figure S6: Relationship between individual mean genome-wide heterozygosity estimated with the k-mer based reference-free approach in GenomeScope ($x$-axis), and the high standard reference-based approach in GATK ($y$-axis), for european sea bass (*D. labrax*, dots) and european pilchard (*S. pilchardus*, triangle).

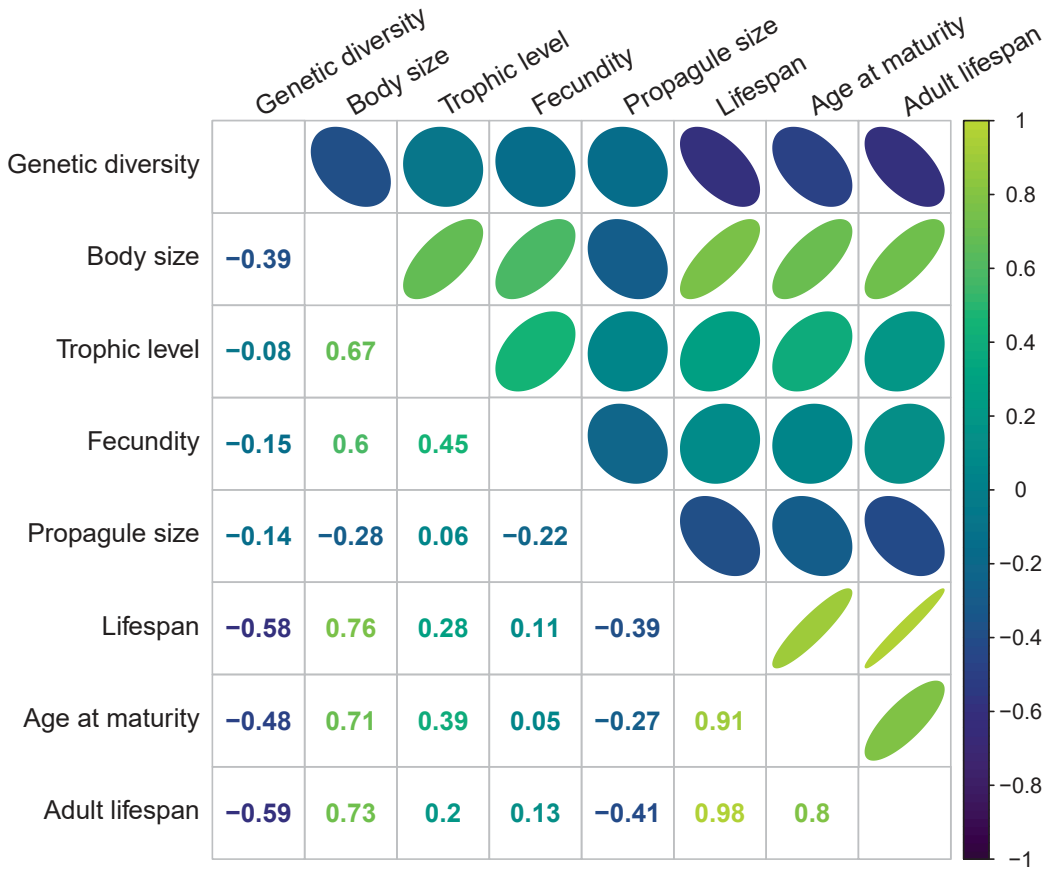



Figure S7: **Correlation matrix between genetic diversity and all quantitative life history traits** - Upper triangle represents the strength and the direction of the correlation, the lower triangle the coefficient of correlation.

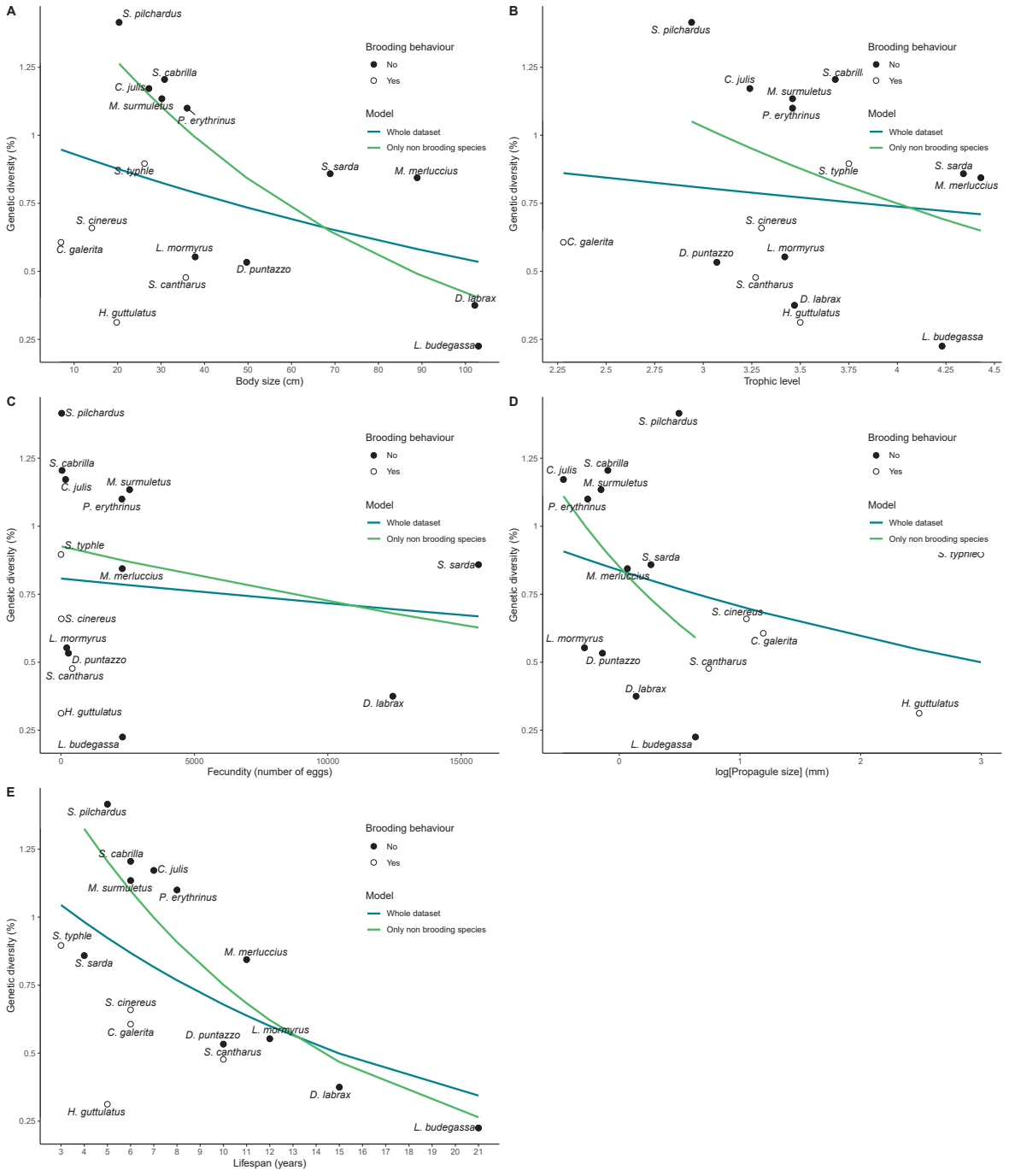



Figure S8: **Relationship between species median genetic diversity (%) and 5 covariables** - Each point represents the median of observed genetic diversity among individuals within each species. Full points represents non-brooding species, empty circles, brooding species. Blue and green line represent the beta regression between each predictive variable and genetic diversity considering either the whole dataset (16 species), or the 11 non-brooding species only, respectively. **A)** adult body size, **B)** trophic level, **C)** fecundity, **D)**, propagule size and **E)**, lifespan.

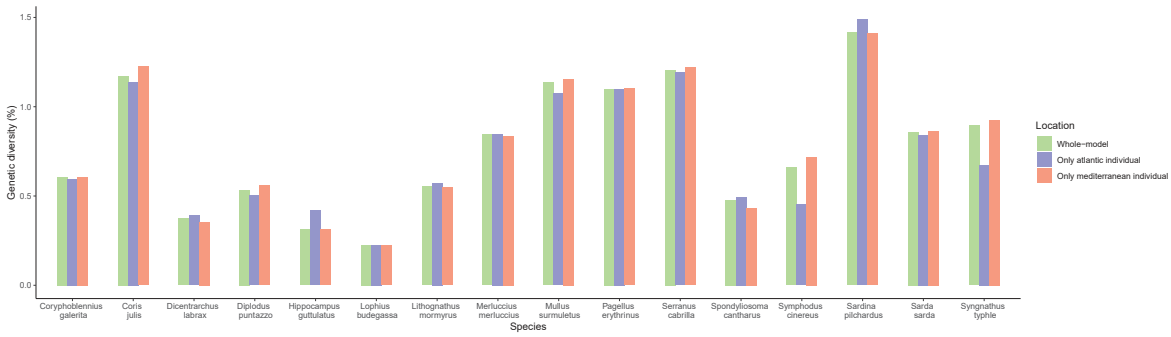



Figure S9: **Effect of population structure on genetic diversity estimates**. For each species on the $x$-axis, genetic diversity is estimate from the median of individual genome-wide heterozygosities for all the individuals (in green), from the individuals from the Mediterranean Sea (in red) or from the individuals from the Atlantic Ocean (in blue).

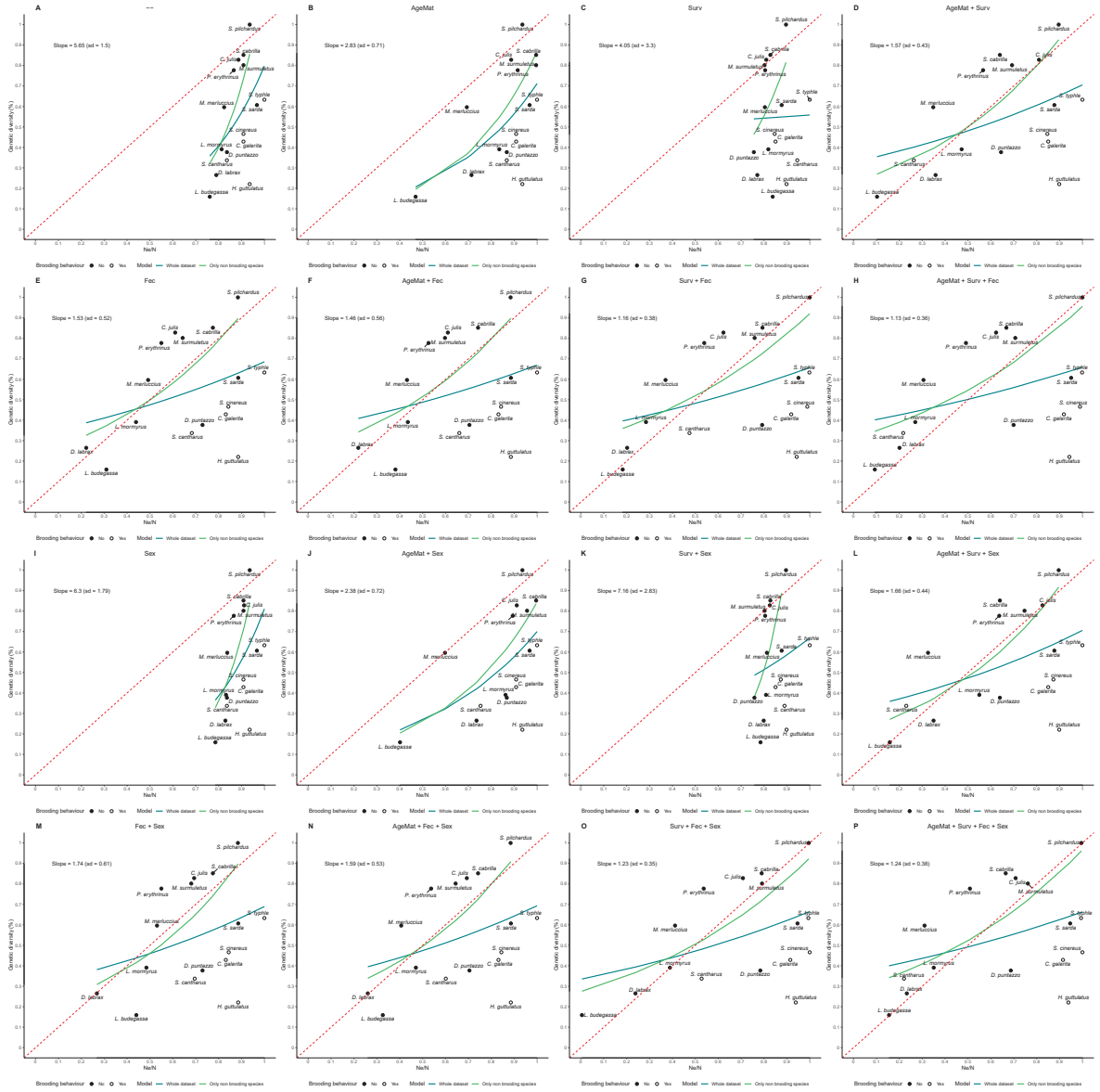


Figure S10: **Relationship between relative species genetic diversity and relative $\frac{N_e}{N}$ estimated by AgeNe for 16 sets of life tables**. Estimated species genetic diversity divided by the highest empirical estimates (i.e. *Sardina pilchardus*) on $y$-axis, and $\frac{N_e}{N}$ estimated by **AgeNe** divided by the highest estimates in each of the 16 sets of life tables. Each of the 16 panels represents the impact of species life tables considering some characteristics: AgeMat (delayed first age at maturity), Surv (age-specific survival estimated from empirical species age-length distributions), Fec (empirical estimates of increasing fecundity with age), Sex (sex-specific differences in either, age at first maturity, age-specific survival or/and fecundity). A) Age at first maturity at 1 year old; constant age-specific survival; constant age-specific fecundity and no differences between sex life tables, B) Species-specific age at first maturity; constant age-specific survival; constant age-specific fecundity and no differences between sex life tables, C) Age at first maturity at 1 year old; increasing age-specific survival rate; D) Species-specific age at first maturity; increasing age-specific survival rate; constant age-specific fecundity and no differences between sex life tables, constant age-specific fecundity and no differences between sex life tables, E) Age at first maturity at 1 year old; constant age-specific survival; increasing fecundity with age and no differences between sex life tables, F) Species-specific age at first maturity; constant age-specific survival; increasing fecundity with age and no differences between sex life tables, G) Age at first maturity at 1 year old; increasing age-specific survival rate; increasing fecundity with age and no differences between sex life tables, H) Species-specific age at first maturity; increasing age-specific survival rate; increasing fecundity with age and no differences between sex life tables, I) Age at first maturity at 1 year old; constant age-specific survival; constant age-specific fecundity and sex-specific differences between life tables, J) Species-specific age at first maturity; constant age-specific survival; constant age-specific fecundity and sex-specific differences between life tables, K) Age at first maturity at 1 year old; increasing age-specific survival rate; constant age-specific fecundity and sex-specific differences between life tables, L) Species-specific age at first maturity; increasing age-specific survival rate; constant age-specific fecundity and sex-specific differences between life tables, M) Age at first maturity at 1 year old; constant age-specific survival; increasing fecundity with age and sex-specific differences between life tables, N) Species-specific age at first maturity; constant age-specific survival; increasing fecundity with age and sex-specific differences between life tables, O) Age at first maturity at 1 year old; increasing age-specific survival rate; increasing fecundity with age and sex-specific differences between life tables, P) Species-specific age at first maturity; increasing age-specific survival rate; increasing fecundity with age and sex-specific differences between life tables. Full points represents non-brooding species, empty points, brooding species. Dashed blue line and solid green represent the beta regression between relative genetic diversity and relative estimated $\frac{N_e}{N}$. Dotted red line represents the identity curve, with slope equals 1 and intercept 0

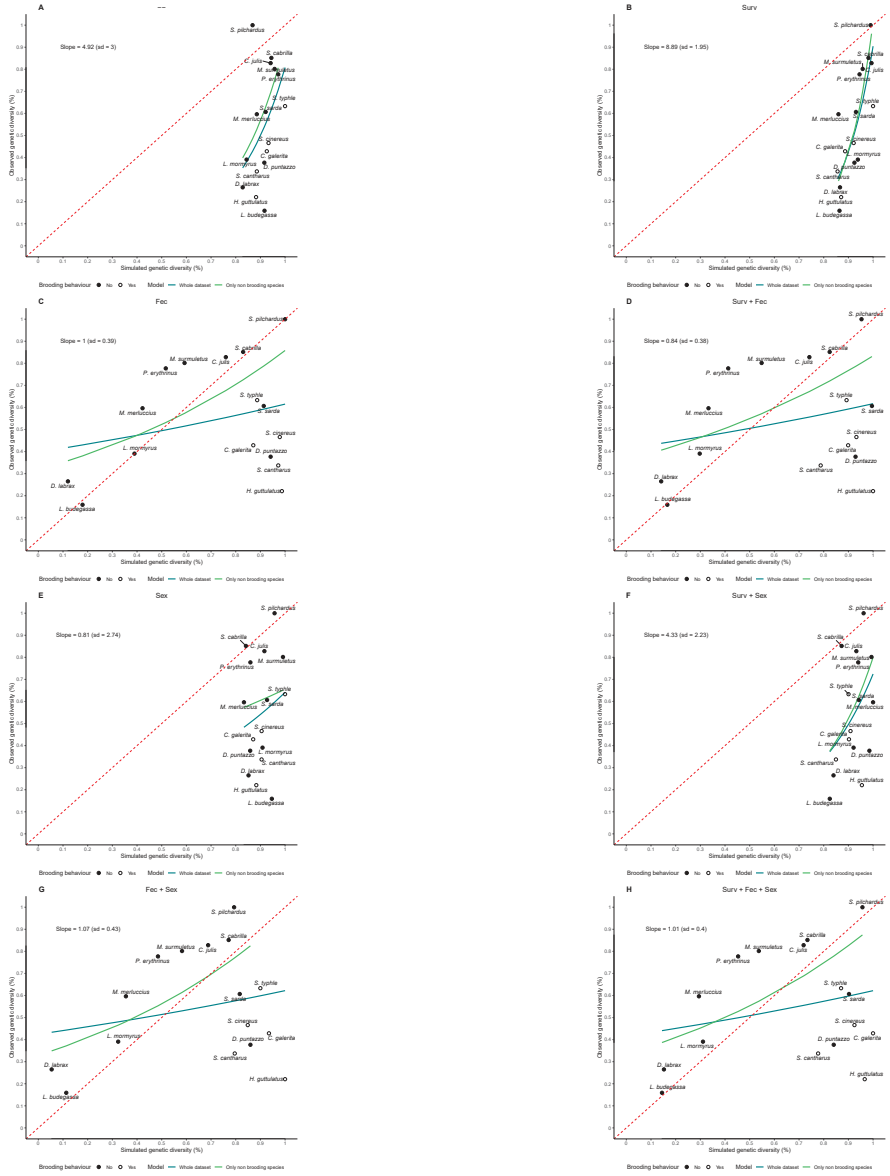



Figure S11: **Relationship between relative species genetic diversity and simulated genetic diversity with forward-in-time simulations for 16 sets of life tables**. Estimated species genetic diversity divided by the highest empirical estimates (i.e. *Sardina pilchardus*) on *y*-axis, and simulated genetic diversity estimated by `SLiM` v.3.3.1 (Haller and Messer, 2017) divided by the highest estimates in each of the 16 set of life tables. Each of the 16 panels represents species life tables considering some characteristics: AgeMat (delayed first age of maturity), Surv (age-specific survival estimated from empirical species age-length distributions), Fec (empirical estimates of increasing fecundity with age), Sex (sex-specific differences in either, age at first maturity, age-specific survival or/and fecundity). A) Age at first maturity at 1 year old; constant age-specific survival; constant age-specific fecundity and no differences between sex life tables, B) Age at first maturity at 1 year old; increasing age-specific survival rate; C) Age at first maturity at 1 year old; constant age-specific survival; increasing fecundity with age and no differences between sex life tables, D) Age at first maturity at 1 year old; increasing age-specific survival ratel; increasing fecundity with age and no differences between sex life tables, E) Age at first maturity at 1 year old; constant age-specific survival; constant age-specific fecundity and sex-specific differences between life tables, F) Age at first maturity at 1 year old; increasing age-specific survival rate; constant age-specific fecundity and sex-specific differences between life tables, G) Age at first maturity at 1 year old; constant age-specific survival; increasing fecundity with age and sex-specific differences between life tables, H) Age at first maturity at 1 year old; increasing age-specific survival rate; increasing fecundity with age and sex-specific differences between life tables, Full points represents non-brooding species, empty points, brooding species. Dashed blue line and solid green represent the beta regression between relative genetic diversity and relative estimated $\frac{N_e}{N}$. Dotted red line represents the identity curve, with slope equals 1 and intercept 0

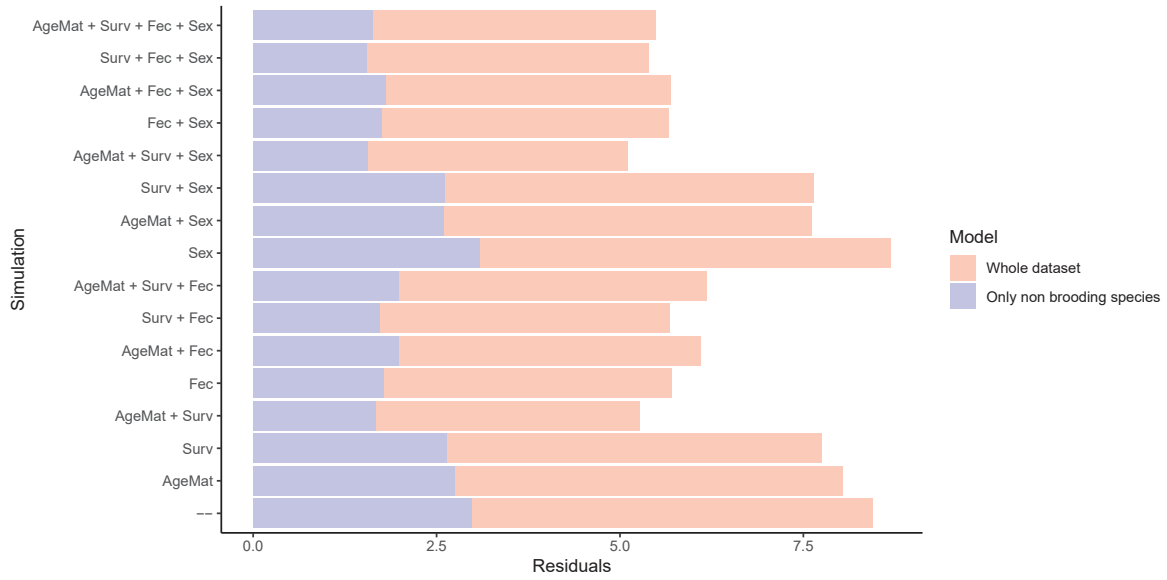



Figure S12: **Residuals of the linear model between genetic diversity and variance in reproductive success estimated from various combinations of life tables from a model with slope equals 1 and intercept 0**. For each set of life tables, we estimated $\frac{N_e}{N}$ with `AgeNe` and calculated the sum of squared deviation between these estimate and a model with slope equals 1 and intercept 0, as residuals. Higher residual number means variation in $\frac{N_e}{N}$ explain little the variation of observed genetic diversity.

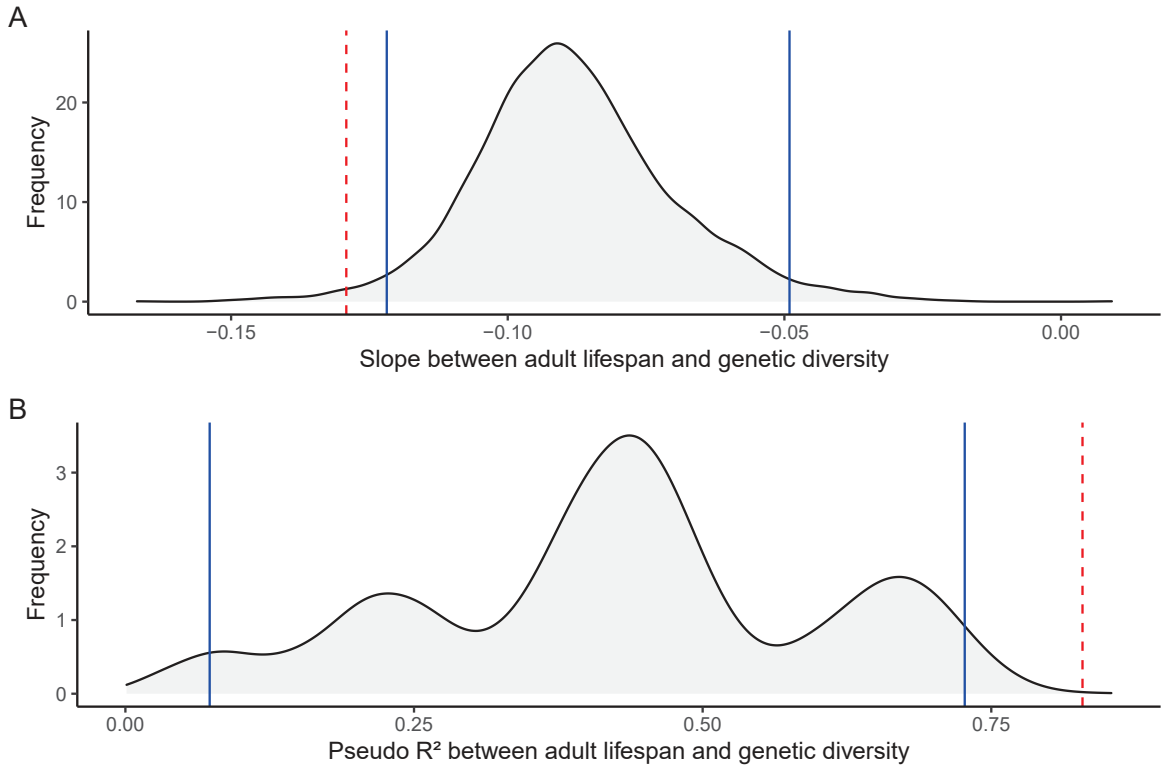



Figure S13: **Distribution of slope and pseudo $R^2$ of the beta regression between adult lifespan and genetic diversity for random subsets of 11 species** - We fit a beta regression between adult lifespan and genetic diversity for all sub-samples of 11 species and calculated the width of the 95% interval of the slope (A) and pseudo $R^2$ (B), represented with solid blue lines. Dashed red lines represent the estimated slope and pseudo $R^2$ for the sub-sample of the 11 species with no parental care behaviour. In (A) the 95% width interval ranged from -0.122 to -0.049, and the estimated slope for the non-brooders only was equal to -0.129. In (B), the 95% width interval ranged from 0.073 to 0.727, and the estimated pseudo $R^2$ for the non-brooders only was equal to 0.829.

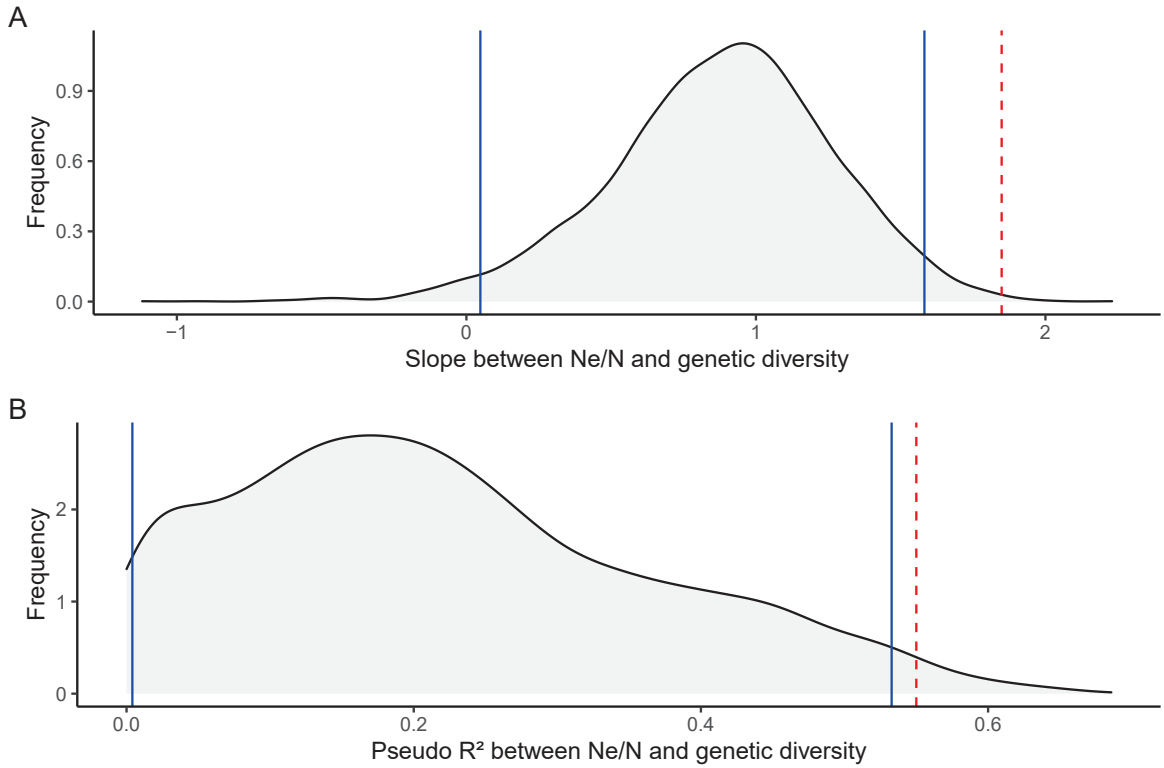



Figure S14: **Distribution of slope and pseudo $R^2$ of the beta regression between $\frac{N_e}{N}$ and genetic diversity for random subsets of 11 species** - We fit a beta regression between observed genetic diversity and $\frac{N_e}{N}$ for all sub-samples of 11 species and calculated the width of the 95% interval of the slope (A) and pseudo $R^2$ (B), represented with solid blue lines. Dashed red lines represent the estimated slope and pseudo $R^2$ for the sub-sample of the 11 species with no parental care behaviour. In (A) the 95% width interval ranged from 0.048 to 1.582, and the estimated slope for the non-brooders only was equal to 1.849. In (B), the 95% width interval ranged from 0.004 to 0.533, and the estimated pseudo $R^2$ from the non-brooders only was equal to 0.55.

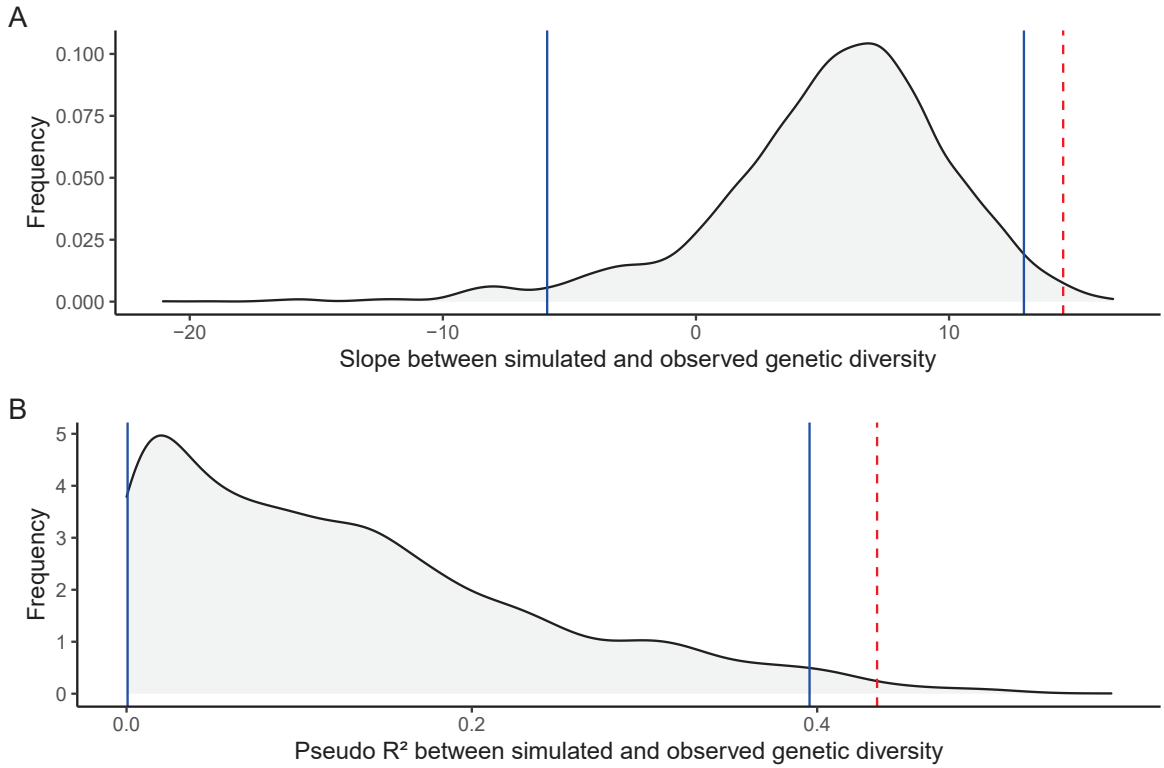



Figure S15: **Distribution of slope and pseudo $R^2$ of the beta regression between simulated, with `SLiM` v.3.3.1, and observed genetic diversity for random subset of 11 species** - We fit a beta regression between simulated genetic diversity and observed genetic diversity for all sub-samples of 11 species and calculated the width of the 95% interval of the slope (A) and pseudo $R^2$ (B), represented with solid blue lines. Dashed red lines represent the estimated slope and pseudo $R^2$ for the sub-sample of the 11 species with no parental care behaviour. In (A) the 95% width interval ranged from -5.879 to 12.957, and the estimated slope for the non-brooders only was equal to 14.509. In (B), the 95% width interval ranged from 0.001 to 0.396, and the estimated pseudo $R^2$ from the non-brooders only was equal to 0.435.

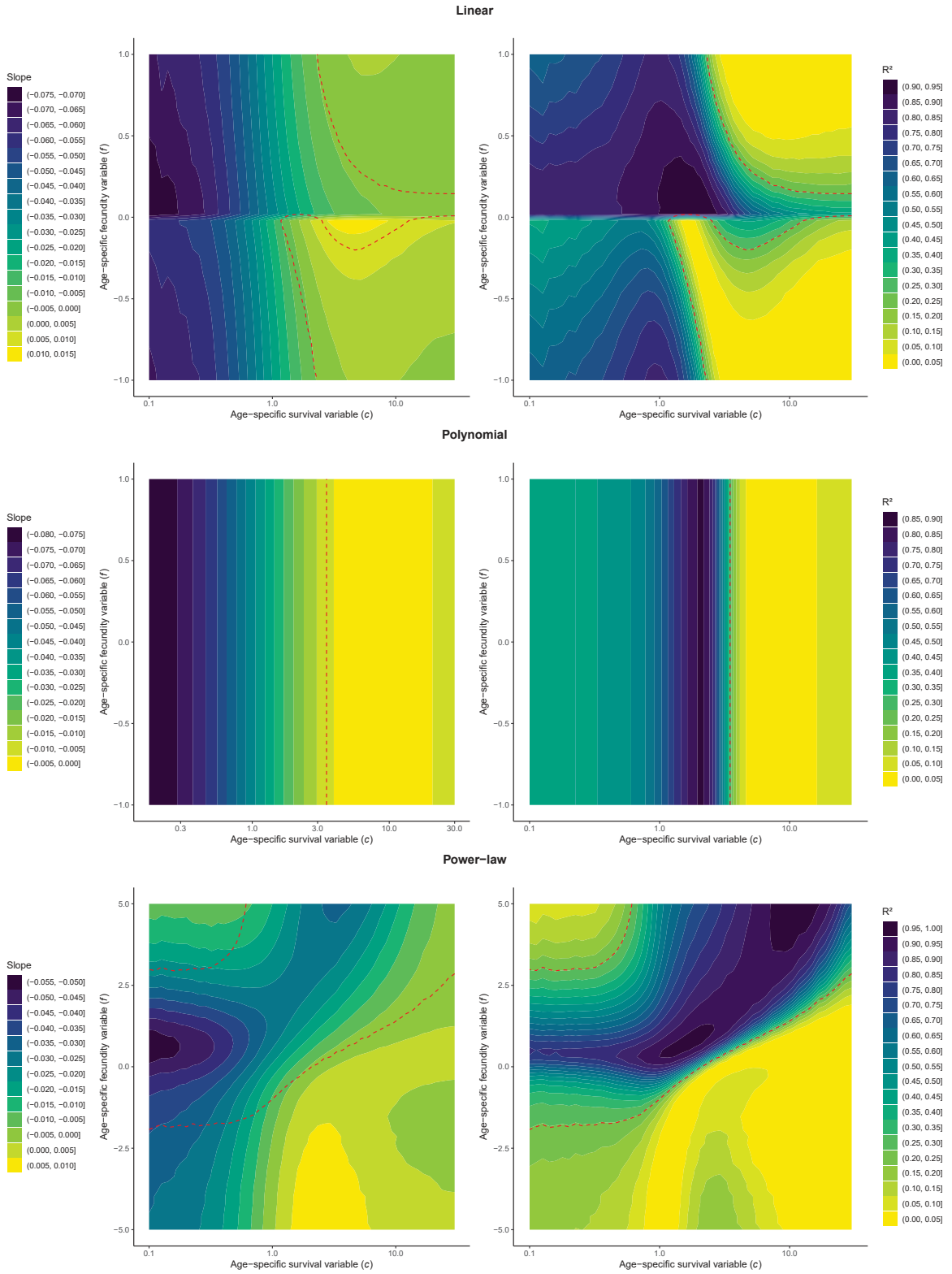



Figure S16: **Slope of and proportion of variance explained by linear models between adult lifespan and $\frac{N_e}{N}$ estimated with `AgeNe` for different combinations of age-specific survival and fecundity for three fecundity-age models: linear, polynomial and power-law**. Colder colors indicate steeper slope (left panels) or higher $R^2$ (right panels) of the regression between adult lifespan and $\frac{N_e}{N}$. On top, linear fecundity-age model ($F = \alpha \times L + \beta$), middle, polynomial fecundity-age model and on the bottom, power-law model, ($F = \alpha L^{\beta}$)

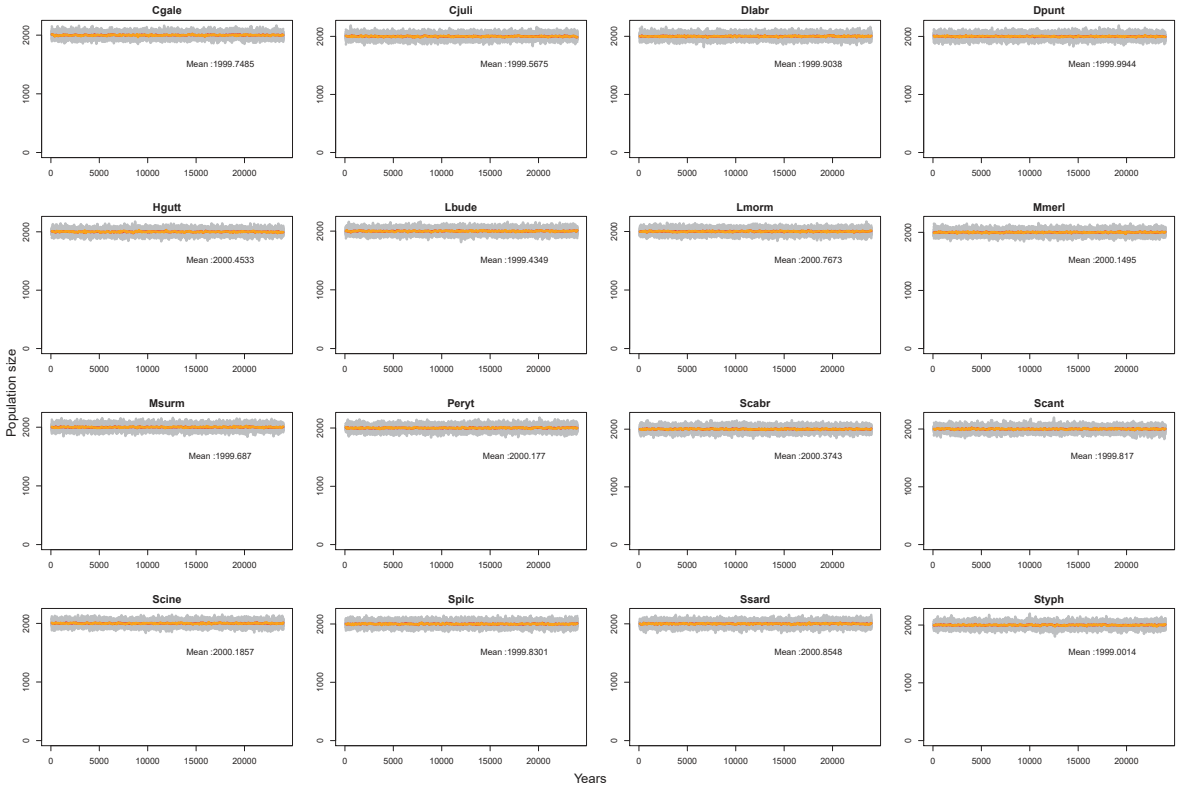



Figure S17: **Population size count for the 50 iterations of the 16 species for set 1 of life tables (age at first maturity at 1 year old, constant age-specific survival rate, constant age-specific fecundity and no differences between sex-specific life tables)**. Adult population size was count for each iteration every 100 years during the 25000 years of the simulation. Each grey line represents population size fluctuations for one iteration. For each species, orange and red line, respectively, represents the median and the mean for all the 50 iterations. Mean over all 50 iterations is written on the plot for each species. Cgale = *Coryphoblennius galerita*, Cjuli = *Coris julis*, Dlabr = *Dicentrarchus labrax*, Dpunt = *Diplodus puntazzo*, Hgutt = *Hippocampus guttulatus*, Lbude = *Lophius budegassa*, Lmorm = *Lithognathus mormyrus*, Mmerl = *Merluccius merluccius*, Msurm = *Mullus surmuletus*, Peryt = *Pagellus erythrinus*, Scabr = *Serranus cabrilla*, Scant = *Spondyliosoma cantharus*, Scine = *Symphodus cinereus*, Spilc = *Sardina pilchardus*, Ssard = *Sarda sarda*, Styph = *Syngnathus typhle*

.

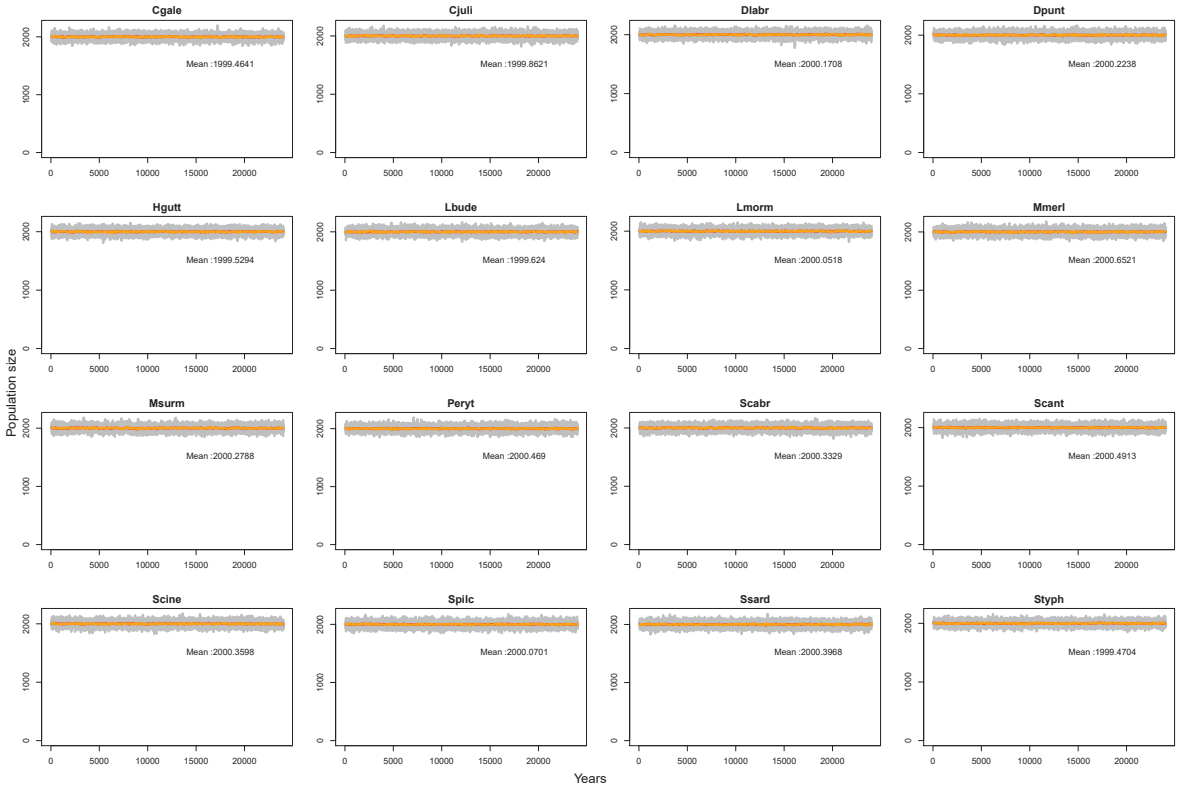



Figure S18: **Population size count for the 50 iterations of the 16 species for set 2 of life tables (age at first maturity at 1 year old, increasing age-specific survival rate, constant age-specific fecundity and no differences between sex-specific life tables)**. Adult population size was count for each iteration every 100 years during the 25000 years of the simulation. Each grey line represents population size fluctuations for one iteration. For each species, orange and red line, respectively, represents the median and the mean for all the 50 iterations. Mean over all 50 iterations is written on the plot for each species. Cgale = *Coryphoblennius galerita*, Cjuli = *Coris julis*, Dlabr = *Dicentrarchus labrax*, Dpunt = *Diplodus puntazzo*, Hgutt = *Hippocampus guttulatus*, Lbude = *Lophius budegassa*, Lmorm = *Lithognathus mormyrus*, Mmerl = *Merluccius merluccius*, Msurm = *Mullus surmuletus*, Peryt = *Pagellus erythrinus*, Scabr = *Serranus cabrilla*, Scant = *Spondyliosoma cantharus*, Scine = *Symphodus cinereus*, Spilc = *Sardina pilchardus*, Ssard = *Sarda sarda*, Styph = *Syngnathus typhle*

.




Figure S19: **Population size count for the 50 iterations of the 16 species for set 3 of life tables (age at first maturity at 1 year old, constant age-specific survival rate, increasing age-specific fecundity and no differences between sex-specific life tables)**. Adult population size was count for each iteration every 100 years during the 25000 years of the simulation. Each grey line represents population size fluctuations for one iteration. For each species, orange and red line, respectively, represents the median and the mean for all the 50 iterations. Mean over all 50 iterations is written on the plot for each species. Cgale = *Coryphoblennius galerita*, Cjuli = *Coris julis*, Dlabr = *Dicentrarchus labrax*, Dpunt = *Diplodus puntazzo*, Hgutt = *Hippocampus guttulatus*, Lbude = *Lophius budegassa*, Lmorm = *Lithognathus mormyrus*, Mmerl = *Merluccius merluccius*, Msurm = *Mullus surmuletus*, Peryt = *Pagellus erythrinus*, Scabr = *Serranus cabrilla*, Scant = *Spondyliosoma cantharus*, Scine = *Symphodus cinereus*, Spilc = *Sardina pilchardus*, Ssard = *Sarda sarda*, Styph = *Syngnathus typhle*

.




Figure S20: **Population size count for the 50 iterations of the 16 species for set 4 of life tables (age at first maturity at 1 year old, increasing age-specific survival rate, increasing age-specific fecundity and no differences between sex-specific life tables)**. Adult population size was count for each iteration every 100 years during the 25000 years of the simulation. Each grey line represents population size fluctuations for one iteration. For each species, orange and red line, respectively, represents the median and the mean for all the 50 iterations. Mean over all 50 iterations is written on the plot for each species. Cgale = *Coryphoblennius galerita*, Cjuli = *Coris julis*, Dlabr = *Dicentrarchus labrax*, Dpunt = *Diplodus puntazzo*, Hgutt = *Hippocampus guttulatus*, Lbude = *Lophius budegassa*, Lmorm = *Lithognathus mormyrus*, Mmerl = *Merluccius merluccius*, Msurm = *Mullus surmuletus*, Peryt = *Pagellus erythrinus*, Scabr = *Serranus cabrilla*, Scant = *Spondyliosoma cantharus*, Scine = *Symphodus cinereus*, Spilc = *Sardina pilchardus*, Ssard = *Sarda sarda*, Styph = *Syngnathus typhle*

.

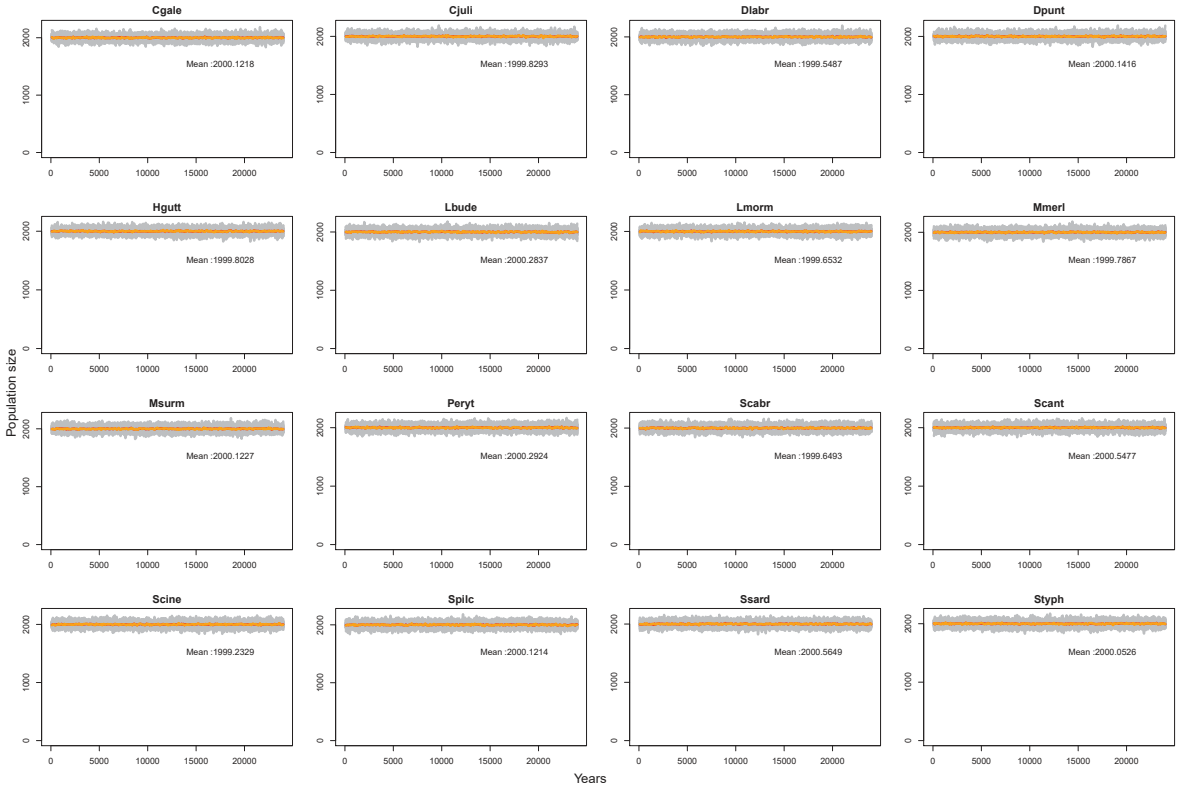



Figure S21: **Population size count for the 50 iterations of the 16 species for set 5 of life tables (age at first maturity at 1 year old, constant age-specific survival rate, constant age-specific fecundity and sex-specific differences in life tables)**. Adult population size was count for each iteration every 100 years during the 25000 years of the simulation. Each grey line represents population size fluctuations for one iteration. For each species, orange and red line, respectively, represents the median and the mean for all the 50 iterations. Mean over all 50 iterations is written on the plot for each species. Cgale = *Coryphoblennius galerita*, Cjuli = *Coris julis*, Dlabr = *Dicentrarchus labrax*, Dpunt = *Diplodus puntazzo*, Hgutt = *Hippocampus guttulatus*, Lbude = *Lophius budegassa*, Lmorm = *Lithognathus mormyrus*, Mmerl = *Merluccius merluccius*, Msurm = *Mullus surmuletus*, Peryt = *Pagellus erythrinus*, Scabr = *Serranus cabrilla*, Scant = *Spondyliosoma cantharus*, Scine = *Symphodus cinereus*, Spilc = *Sardina pilchardus*, Ssard = *Sarda sarda*, Styph = *Syngnathus typhle*

.

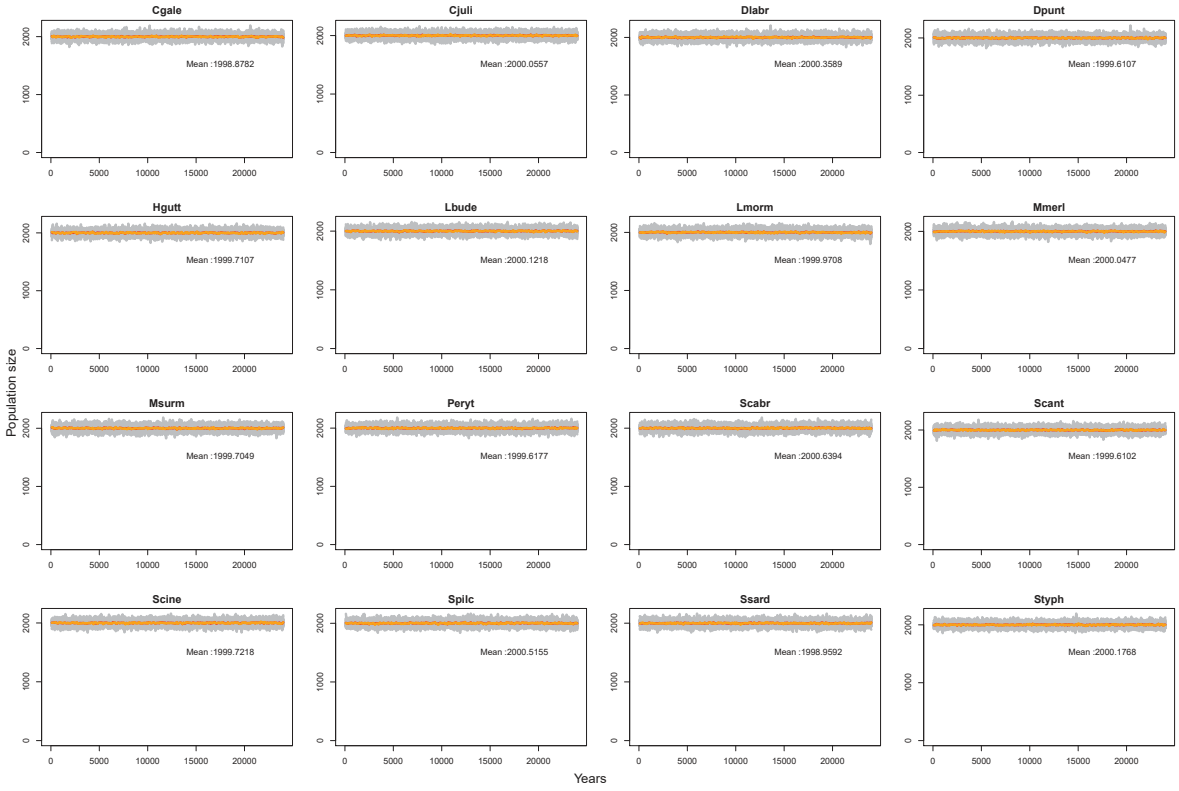



Figure S22: **Population size count for the 50 iterations of the 16 species for set 6 of life tables (age at first maturity at 1 year old, increasing age-specific survival rate, constant age-specific fecundity and sex-specific differences in life tables)**. Adult population size was count for each iteration every 100 years during the 25000 years of the simulation. Each grey line represents population size fluctuations for one iteration. For each species, orange and red line, respectively, represents the median and the mean for all the 50 iterations. Mean over all 50 iterations is written on the plot for each species. Cgale = *Coryphoblennius galerita*, Cjuli = *Coris julis*, Dlabr = *Dicentrarchus labrax*, Dpunt = *Diplodus puntazzo*, Hgutt = *Hippocampus guttulatus*, Lbude = *Lophius budegassa*, Lmorm = *Lithognathus mormyrus*, Mmerl = *Merluccius merluccius*, Msurm = *Mullus surmuletus*, Peryt = *Pagellus erythrinus*, Scabr = *Serranus cabrilla*, Scant = *Spondyliosoma cantharus*, Scine = *Symphodus cinereus*, Spilc = *Sardina pilchardus*, Ssard = *Sarda sarda*, Styph = *Syngnathus typhle*

.




Figure S23: **Population size count for the 50 iterations of the 16 species for set 7 of life tables (age at first maturity at 1 year old, constant age-specific survival rate, increasing age-specific fecundity and sex-specific differences in life tables)**. Adult population size was count for each iteration every 100 years during the 25000 years of the simulation. Each grey line represents population size fluctuations for one iteration. For each species, orange and red line, respectively, represents the median and the mean for all the 50 iterations. Mean over all 50 iterations is written on the plot for each species. Cgale = *Coryphoblennius galerita*, Cjuli = *Coris julis*, Dlabr = *Dicentrarchus labrax*, Dpunt = *Diplodus puntazzo*, Hgutt = *Hippocampus guttulatus*, Lbude = *Lophius budegassa*, Lmorm = *Lithognathus mormyrus*, Mmerl = *Merluccius merluccius*, Msurm = *Mullus surmuletus*, Peryt = *Pagellus erythrinus*, Scabr = *Serranus cabrilla*, Scant = *Spondyliosoma cantharus*, Scine = *Symphodus cinereus*, Spilc = *Sardina pilchardus*, Ssard = *Sarda sarda*, Styph = *Syngnathus typhle*

.




Figure S24: **Population size count for the 50 iterations of the 16 species for set 8 of life tables (age at first maturity at 1 year old, increasing age-specific survival rate, increasing age-specific fecundity and sex-specific differences in life tables)**. Adult population size was count for each iteration every 100 years during the 25000 years of the simulation. Each grey line represents population size fluctuations for one iteration. For each species, orange and red line, respectively, represents the median and the mean for all the 50 iterations. Mean over all 50 iterations is written on the plot for each species. Cgale = *Coryphoblennius galerita*, Cjuli = *Coris julis*, Dlabr = *Dicentrarchus labrax*, Dpunt = *Diplodus puntazzo*, Hgutt = *Hippocampus guttulatus*, Lbude = *Lophius budegassa*, Lmorm = *Lithognathus mormyrus*, Mmerl = *Merluccius merluccius*, Msurm = *Mullus surmuletus*, Peryt = *Pagellus erythrinus*, Scabr = *Serranus cabrilla*, Scant = *Spondyliosoma cantharus*, Scine = *Symphodus cinereus*, Spilc = *Sardina pilchardus*, Ssard = *Sarda sarda*, Styph = *Syngnathus typhle*

.

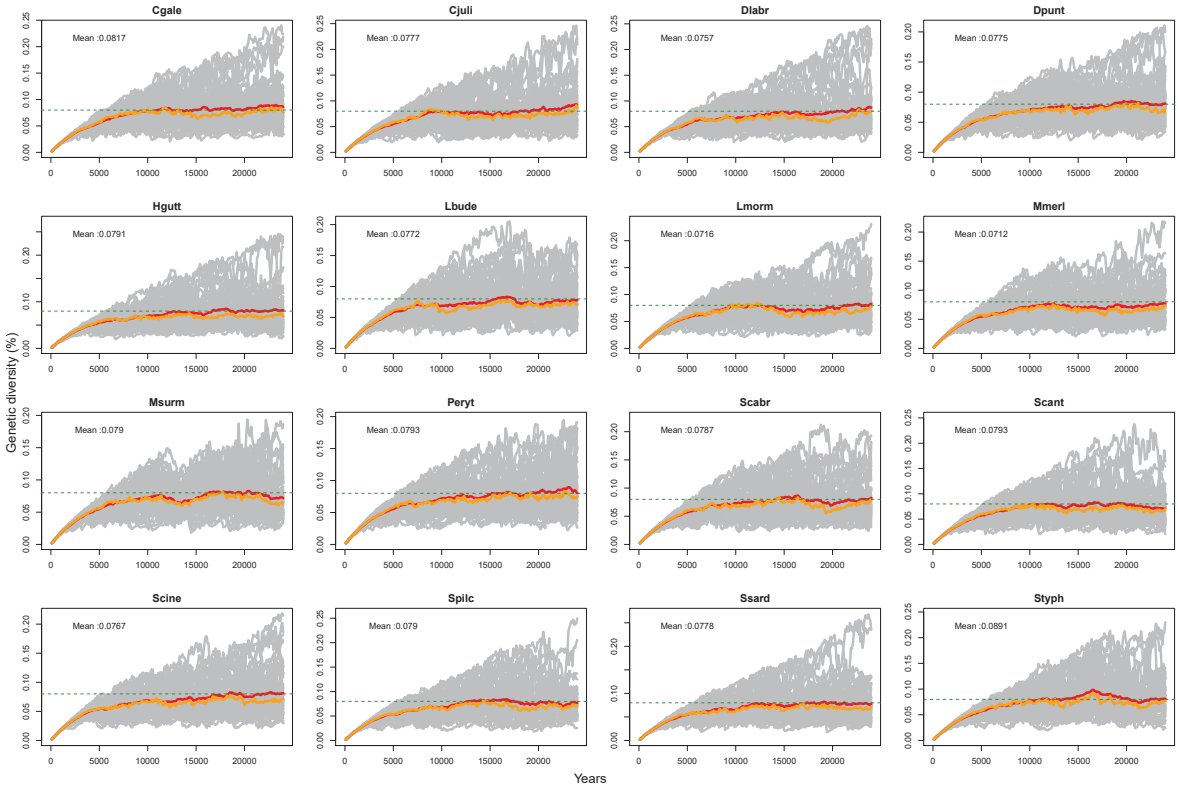



Figure S25: **Genetic diversity simulated for each of the 16 species for set 1 of life tables (age at first maturity at 1 year old, constant age-specific survival rate, constant age-specific fecundity and no differences between sex-specific life tables)**. Simulation were conducted with `SLiM` v.3.3.1 for 25000 years, with mutation rate, $\mu = 1e^{-7}$ on a 1 Mb non recombining loci, and carrying capacity equals $N = 2000$. Genetic diversity was estimated for each iteration every 100 years during the 25000 years of the simulation. Each grey line represents genetic diversity fluctuation for one iteration. For each species, orange and red line, respectively, represents the median and the mean for all the 50 iterations. Green dashed line represents the genetic diversity expected at mutation-drift equilibruim in a Wright-Fisher model ($4N\mu = 4 \times 2000 \times 1e^{-7} = 0.0008 = 0.08\%$). Mean over all 50 iterations, calculated with genetic diversity values estimated between 15 000 and 25 000 years after the beginning of the simulation, is written on the plot for each species. Cgale = *Coryphoblennius galerita*, Cjuli = *Coris julis*, Dlabr = *Dicentrarchus labrax*, Dpunt = *Diplodus puntazzo*, Hgutt = *Hippocampus guttulatus*, Lbude = *Lophius budegassa*, Lmorm = *Lithognathus mormyrus*, Mmerl = *Merluccius merluccius*, Msurm = *Mullus surmuletus*, Peryt = *Pagellus erythrinus*, Scabr = *Serranus cabrilla*, Scant = *Spondyliosoma cantharus*, Scine = *Symphodus cinereus*, Spilc = *Sardina pilchardus*, Ssard = *Sarda sarda*, Styph = *Syngnathus typhle*

.

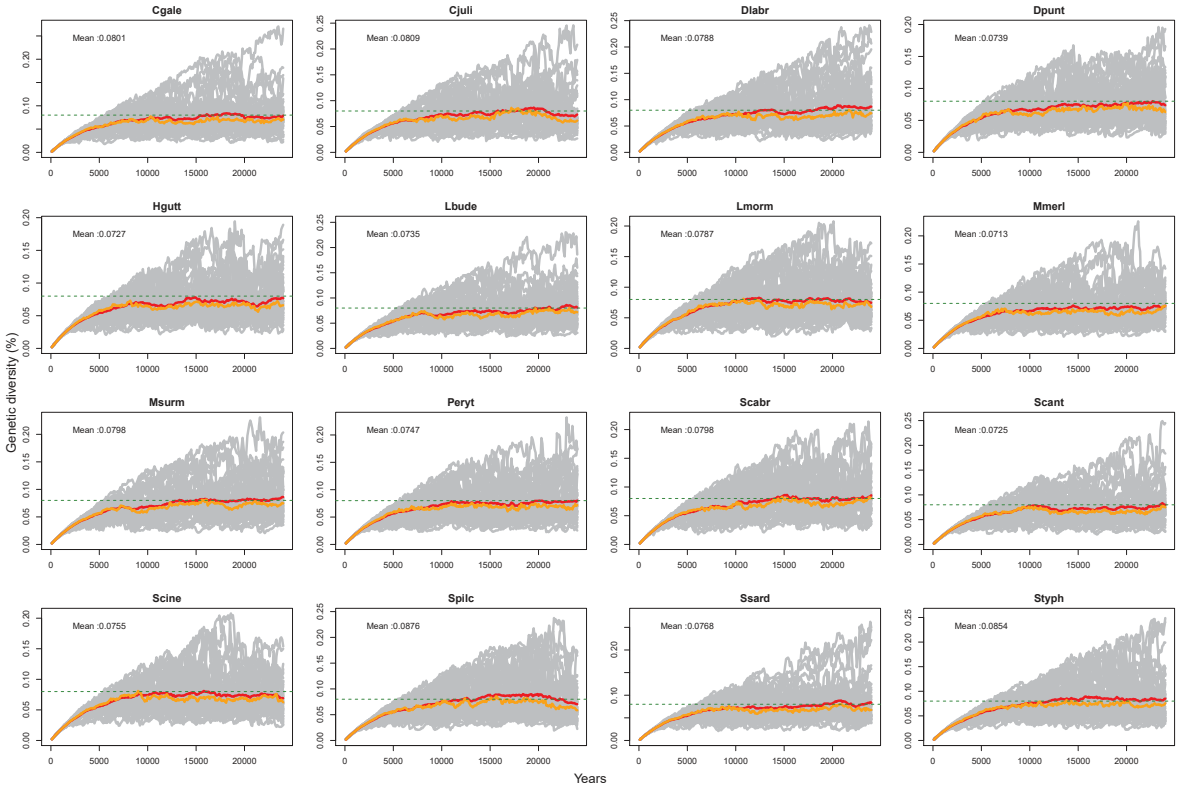



Figure S26: **Genetic diversity simulated for each of the 16 species for set 2 of life tables (age at first maturity at 1 year old, increasing age-specific survival rate, constant age-specific fecundity and no differences between sex-specific life tables)**. Simulation were conducted with `SLiM` v.3.3.1 for 25000 years, with mutation rate, $\mu = 1e^{-7}$ on a 1 Mb non recombining loci, and carrying capacity equals $N = 2000$. Genetic diversity was estimated for each iteration every 100 years during the 25000 years of the simulation. Each grey line represents genetic diversity fluctuation for one iteration. For each species, orange and red line, respectively, represents the median and the mean for all the 50 iterations. Green dashed line represents the genetic diversity expected at mutation-drift equilibruim in a Wright-Fisher model ($4N\mu = 4 \times 2000 \times 1e^{-7} = 0.0008 = 0.08\%$). Mean over all 50 iterations, calculated with genetic diversity values estimated between 15 000 and 25 000 years after the beginning of the simulation, is written on the plot for each species. Cgale = *Coryphoblennius galerita*, Cjuli = *Coris julis*, Dlabr = *Dicentrarchus labrax*, Dpunt = *Diplodus puntazzo*, Hgutt = *Hippocampus guttulatus*, Lbude = *Lophius budegassa*, Lmorm = *Lithognathus mormyrus*, Mmerl = *Merluccius merluccius*, Msurm = *Mullus surmuletus*, Peryt = *Pagellus erythrinus*, Scabr = *Serranus cabrilla*, Scant = *Spondyliosoma cantharus*, Scine = *Symphodus cinereus*, Spilc = *Sardina pilchardus*, Ssard = *Sarda sarda*, Styph = *Syngnathus typhle*

.

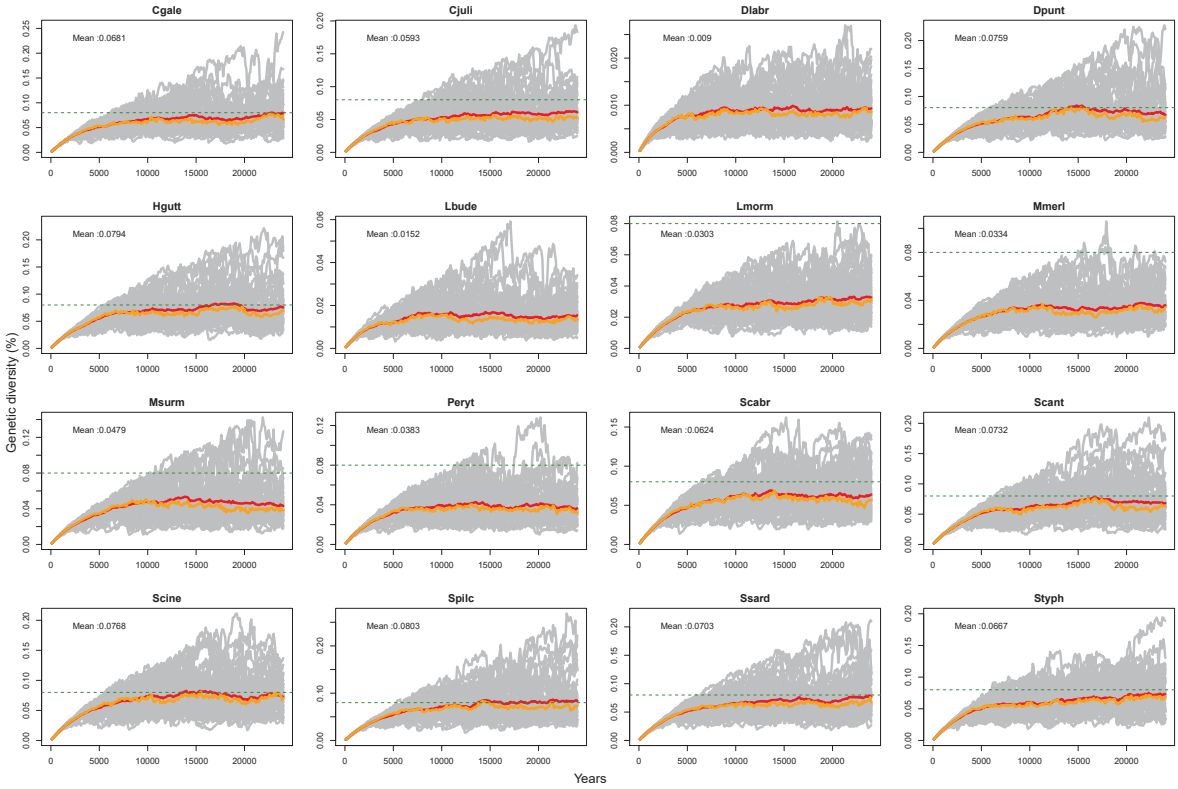



Figure S27: **Genetic diversity simulated for each of the 16 species for set 3 of life tables (age at first maturity at 1 year old, constant age-specific survival rate, increasing age-specific fecundity and no differences between sex-specific life tables)**. Simulation were conducted with `SLiM` v.3.3.1 for 25000 years, with mutation rate, $\mu = 1e^{-7}$ on a 1 Mb non recombining loci, and carrying capacity equals $N = 2000$. Genetic diversity was estimated for each iteration every 100 years during the 25000 years of the simulation. Each grey line represents genetic diversity fluctuation for one iteration. For each species, orange and red line, respectively, represents the median and the mean for all the 50 iterations. Green dashed line represents the genetic diversity expected at mutation-drift equilibruim in a Wright-Fisher model ($4N\mu = 4 \times 2000 \times 1e^{-7} = 0.0008 = 0.08\%$). Mean over all 50 iterations, calculated with genetic diversity values estimated between 15 000 and 25 000 years after the beginning of the simulation, is written on the plot for each species. Cgale = *Coryphoblennius galerita*, Cjuli = *Coris julis*, Dlabr = *Dicentrarchus labrax*, Dpunt = *Diplodus puntazzo*, Hgutt = *Hippocampus guttulatus*, Lbude = *Lophius budegassa*, Lmorm = *Lithognathus mormyrus*, Mmerl = *Merluccius merluccius*, Msurm = *Mullus surmuletus*, Peryt = *Pagellus erythrinus*, Scabr = *Serranus cabrilla*, Scant = *Spondyliosoma cantharus*, Scine = *Symphodus cinereus*, Spilc = *Sardina pilchardus*, Ssard = *Sarda sarda*, Styph = *Syngnathus typhle*

.

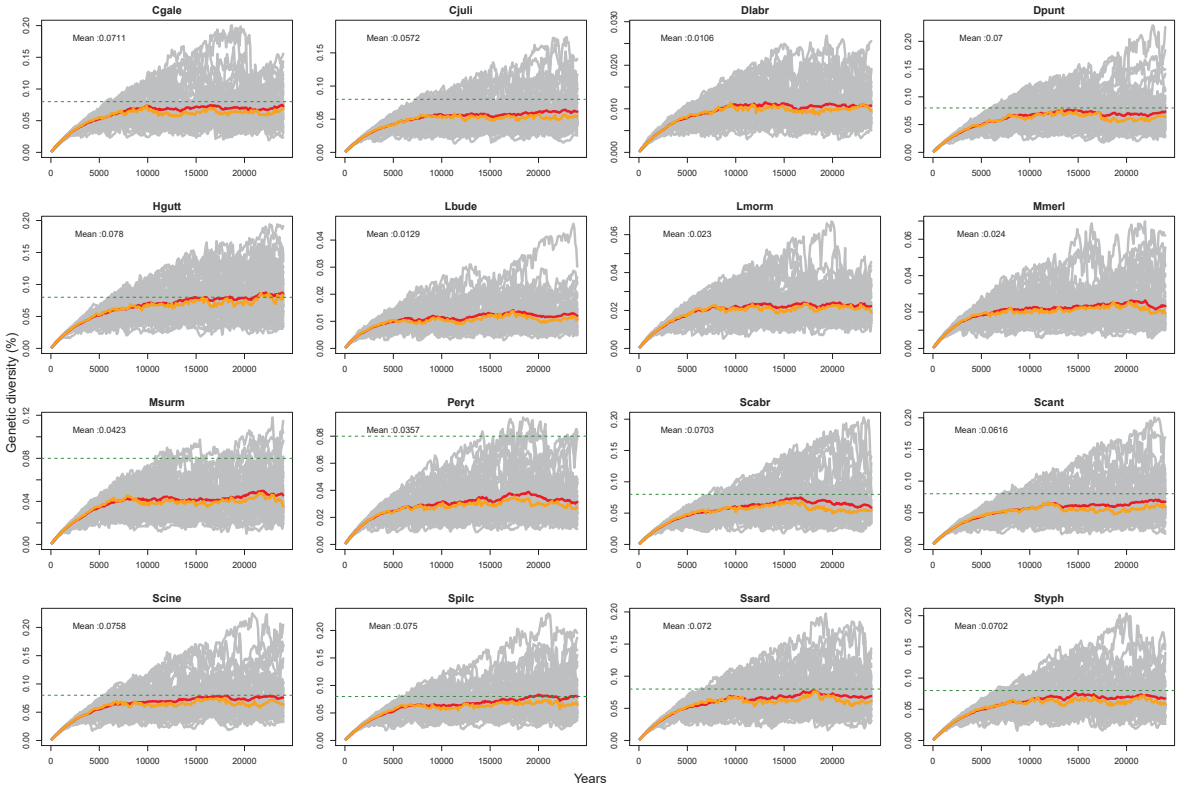



Figure S28: **Genetic diversity simulated for each of the 16 species for set 4 of life tables (age at first maturity at 1 year old, increasing age-specific survival rate, increasing age-specific fecundity and no differences between sex-specific life tables)**. Simulation were conducted with `SLiM` v.3.3.1 for 25000 years, with mutation rate, $\mu = 1e^{-7}$ on a 1 Mb non recombining loci, and carrying capacity equals $N = 2000$. Genetic diversity was estimated for each iteration every 100 years during the 25000 years of the simulation. Each grey line represents genetic diversity fluctuation for one iteration. For each species, orange and red line, respectively, represents the median and the mean for all the 50 iterations. Green dashed line represents the genetic diversity expected at mutation-drift equilibruim in a Wright-Fisher model ($4N\mu = 4 \times 2000 \times 1e^{-7} = 0.0008 = 0.08\%$). Mean over all 50 iterations, calculated with genetic diversity values estimated between 15 000 and 25 000 years after the beginning of the simulation, is written on the plot for each species. Cgale = *Coryphoblennius galerita*, Cjuli = *Coris julis*, Dlabr = *Dicentrarchus labrax*, Dpunt = *Diplodus puntazzo*, Hgutt = *Hippocampus guttulatus*, Lbude = *Lophius budegassa*, Lmorm = *Lithognathus mormyrus*, Mmerl = *Merluccius merluccius*, Msurm = *Mullus surmuletus*, Peryt = *Pagellus erythrinus*, Scabr = *Serranus cabrilla*, Scant = *Spondyliosoma cantharus*, Scine = *Symphodus cinereus*, Spilc = *Sardina pilchardus*, Ssard = *Sarda sarda*, Styph = *Syngnathus typhle*

.

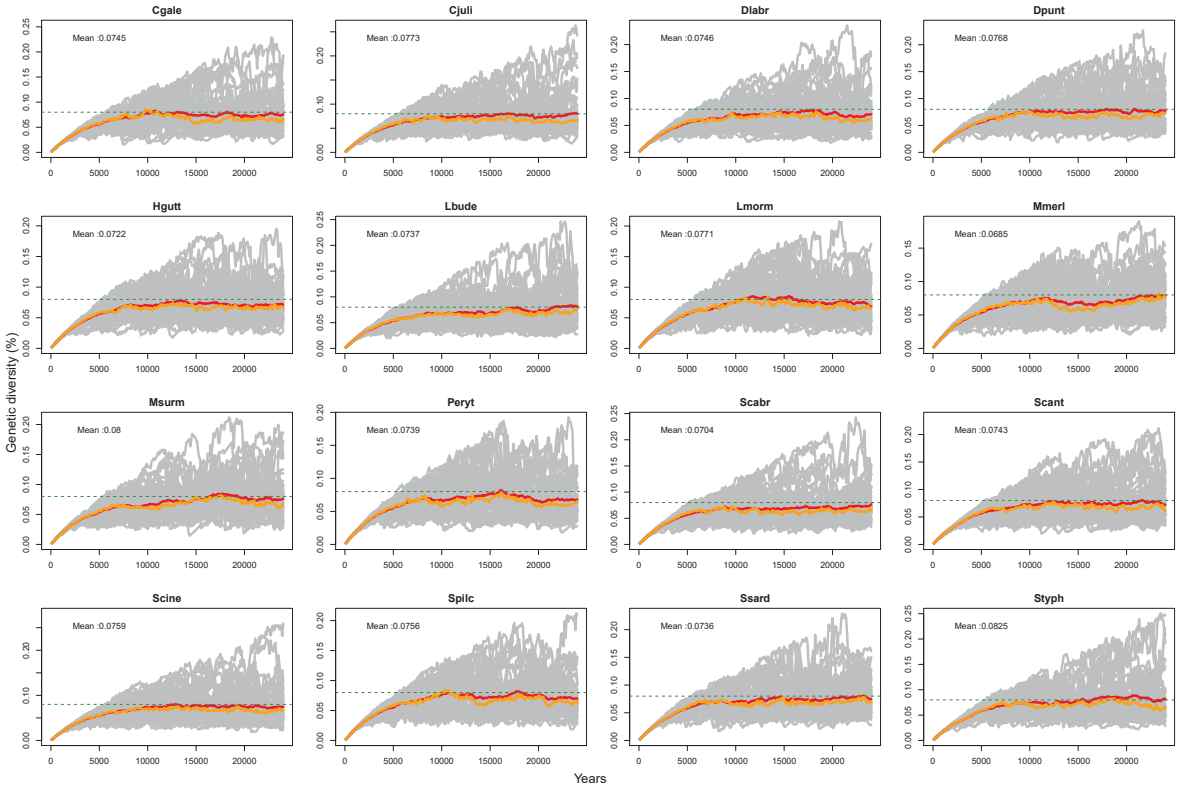



Figure S29: **Genetic diversity simulated for each of the 16 species for set 5 of life tables (age at first maturity at 1 year old, constant age-specific survival rate, constant age-specific fecundity and sex-specific differences in life tables)**. Simulation were conducted with `SLiM` v.3.3.1 for 25000 years, with mutation rate, $\mu = 1e^{-7}$ on a 1 Mb non recombining loci, and carrying capacity equals $N = 2000$. Genetic diversity was estimated for each iteration every 100 years during the 25000 years of the simulation. Each grey line represents genetic diversity fluctuation for one iteration. For each species, orange and red line, respectively, represents the median and the mean for all the 50 iterations. Green dashed line represents the genetic diversity expected at mutation-drift equilibruim in a Wright-Fisher model ($4N\mu = 4 \times 2000 \times 1e^{-7} = 0.0008 = 0.08\%$). Mean over all 50 iterations, calculated with genetic diversity values estimated between 15 000 and 25 000 years after the beginning of the simulation, is written on the plot for each species. Cgale = *Coryphoblennius galerita*, Cjuli = *Coris julis*, Dlabr = *Dicentrarchus labrax*, Dpunt = *Diplodus puntazzo*, Hgutt = *Hippocampus guttulatus*, Lbude = *Lophius budegassa*, Lmorm = *Lithognathus mormyrus*, Mmerl = *Merluccius merluccius*, Msurm = *Mullus surmuletus*, Peryt = *Pagellus erythrinus*, Scabr = *Serranus cabrilla*, Scant = *Spondyliosoma cantharus*, Scine = *Symphodus cinereus*, Spilc = *Sardina pilchardus*, Ssard = *Sarda sarda*, Styph = *Syngnathus typhle*

.

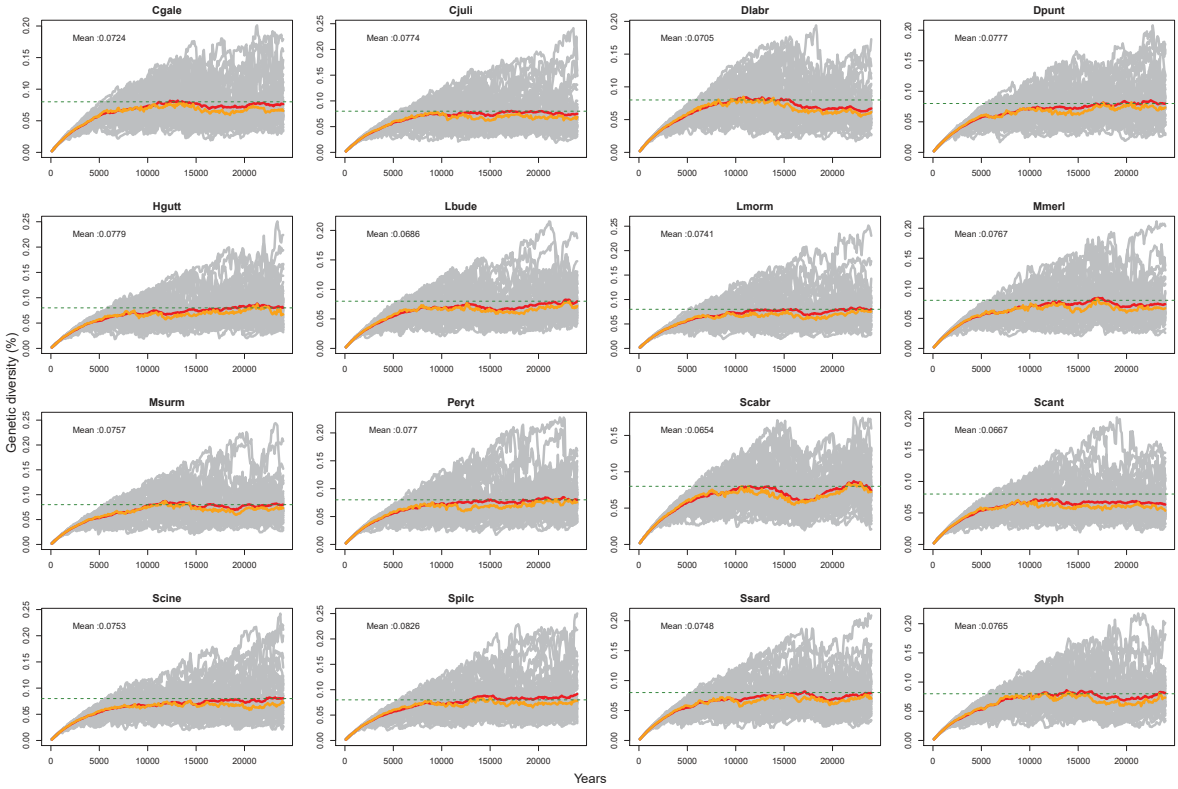



Figure S30: **Genetic diversity simulated for each of the 16 species for set 6 of life tables (age at first maturity at 1 year old, increasing age-specific survival rate, constant age-specific fecundity and sex-specific differences in life tables)**. Simulation were conducted with `SLiM` v.3.3.1 for 25000 years, with mutation rate, $\mu = 1e^{-7}$ on a 1 Mb non recombining loci, and carrying capacity equals $N = 2000$. Genetic diversity was estimated for each iteration every 100 years during the 25000 years of the simulation. Each grey line represents genetic diversity fluctuation for one iteration. For each species, orange and red line, respectively, represents the median and the mean for all the 50 iterations. Green dashed line represents the genetic diversity expected at mutation-drift equilibruim in a Wright-Fisher model ($4N\mu = 4 \times 2000 \times 1e^{-7} = 0.0008 = 0.08\%$). Mean over all 50 iterations, calculated with genetic diversity values estimated between 15 000 and 25 000 years after the beginning of the simulation, is written on the plot for each species. Cgale = *Coryphoblennius galerita*, Cjuli = *Coris julis*, Dlabr = *Dicentrarchus labrax*, Dpunt = *Diplodus puntazzo*, Hgutt = *Hippocampus guttulatus*, Lbude = *Lophius budegassa*, Lmorm = *Lithognathus mormyrus*, Mmerl = *Merluccius merluccius*, Msurm = *Mullus surmuletus*, Peryt = *Pagellus erythrinus*, Scabr = *Serranus cabrilla*, Scant = *Spondyliosoma cantharus*, Scine = *Symphodus cinereus*, Spilc = *Sardina pilchardus*, Ssard = *Sarda sarda*, Styph = *Syngnathus typhle*

.

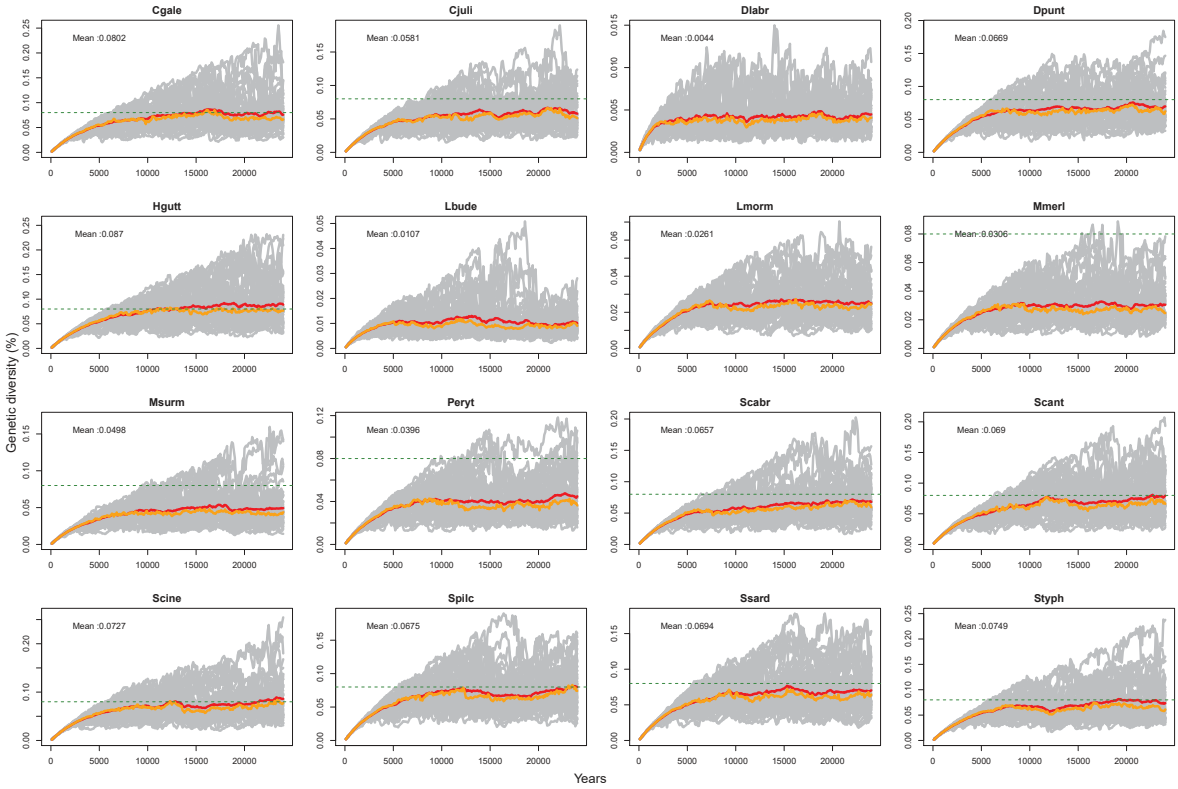



Figure S31: **Genetic diversity simulated for each of the 16 species for set 7 of life tables (age at first maturity at 1 year old, constant age-specific survival rate, increasing age-specific fecundity and sex-specific differences in life tables)**. Simulation were conducted with `SLiM` v.3.3.1 for 25000 years, with mutation rate, $\mu = 1e^{-7}$ on a 1 Mb non recombining loci, and carrying capacity equals $N = 2000$. Genetic diversity was estimated for each iteration every 100 years during the 25000 years of the simulation. Each grey line represents genetic diversity fluctuation for one iteration. For each species, orange and red line, respectively, represents the median and the mean for all the 50 iterations. Green dashed line represents the genetic diversity expected at mutation-drift equilibruim in a Wright-Fisher model ($4N\mu = 4 \times 2000 \times 1e^{-7} = 0.0008 = 0.08\%$). Mean over all 50 iterations, calculated with genetic diversity values estimated between 15 000 and 25 000 years after the beginning of the simulation, is written on the plot for each species. Cgale = *Coryphoblennius galerita*, Cjuli = *Coris julis*, Dlabr = *Dicentrarchus labrax*, Dpunt = *Diplodus puntazzo*, Hgutt = *Hippocampus guttulatus*, Lbude = *Lophius budegassa*, Lmorm = *Lithognathus mormyrus*, Mmerl = *Merluccius merluccius*, Msurm = *Mullus surmuletus*, Peryt = *Pagellus erythrinus*, Scabr = *Serranus cabrilla*, Scant = *Spondyliosoma cantharus*, Scine = *Symphodus cinereus*, Spilc = *Sardina pilchardus*, Ssard = *Sarda sarda*, Styph = *Syngnathus typhle*

.

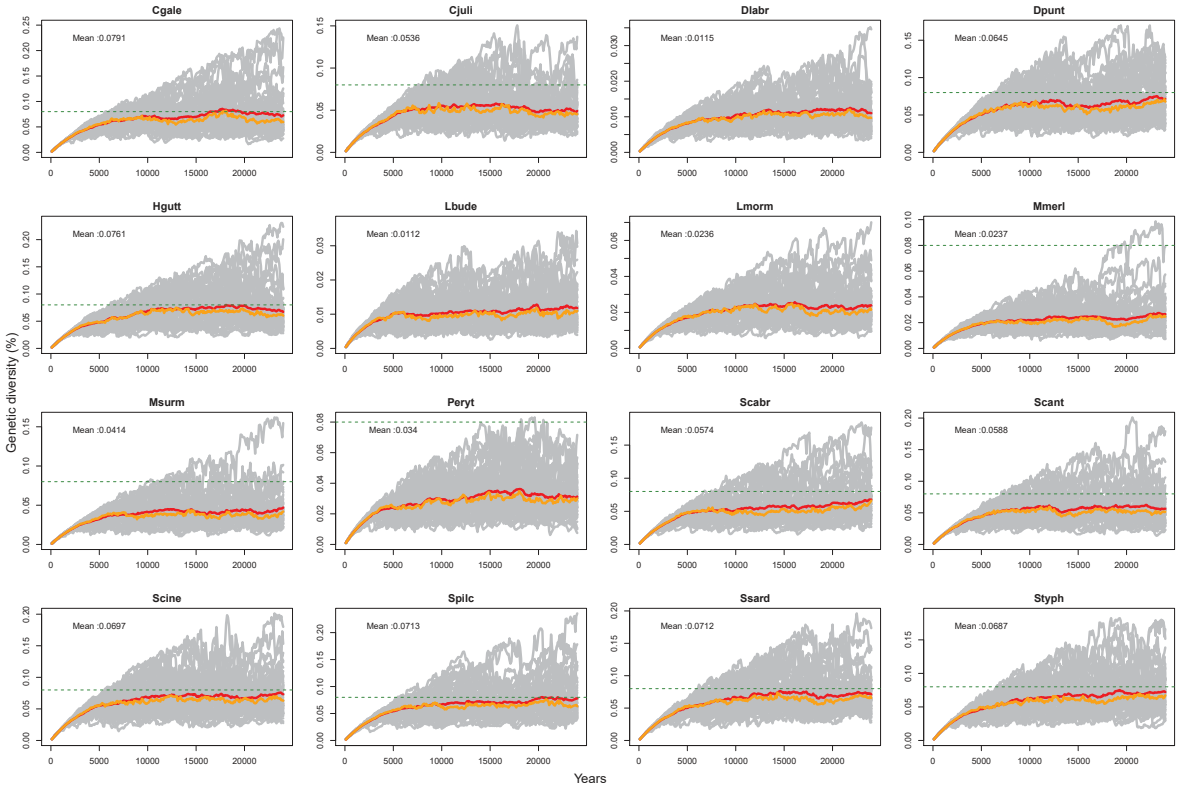



Figure S32: **Genetic diversity simulated for each of the 16 species for set 8 of life tables (age at first maturity at 1 year old, increasing age-specific survival rate, increasing age-specific fecundity and sex-specific differences in life tables)**. Simulation were conducted with `SLiM` v.3.3.1 for 25000 years, with mutation rate, $\mu = 1e^{-7}$ on a 1 Mb non recombining loci, and carrying capacity equals $N = 2000$. Genetic diversity was estimated for each iteration every 100 years during the 25000 years of the simulation. Each grey line represents genetic diversity fluctuation for one iteration. For each species, orange and red line, respectively, represents the median and the mean for all the 50 iterations. Green dashed line represents the genetic diversity expected at mutation-drift equilibruim in a Wright-Fisher model ($4N\mu = 4 \times 2000 \times 1e^{-7} = 0.0008 = 0.08\%$). Mean over all 50 iterations, calculated with genetic diversity values estimated between 15 000 and 25 000 years after the beginning of the simulation, is written on the plot for each species. Cgale = *Coryphoblennius galerita*, Cjuli = *Coris julis*, Dlabr = *Dicentrarchus labrax*, Dpunt = *Diplodus puntazzo*, Hgutt = *Hippocampus guttulatus*, Lbude = *Lophius budegassa*, Lmorm = *Lithognathus mormyrus*, Mmerl = *Merluccius merluccius*, Msurm = *Mullus surmuletus*, Peryt = *Pagellus erythrinus*, Scabr = *Serranus cabrilla*, Scant = *Spondyliosoma cantharus*, Scine = *Symphodus cinereus*, Spilc = *Sardina pilchardus*, Ssard = *Sarda sarda*, Styph = *Syngnathus typhle*

.

# ANNEX OF CHAPTER 2

# Supplementary Material to

## Contents

**Additional methods** **2**
High Molecular Weight DNA extraction . . . 2
Reference genome sequencing and assembly . . . 3
Mapping to reference genome . . . 4
Variant Calling . . . 4
Variant Calling Filtering . . . 6
Genome annotation . . . 6
Phylogeny . . . 6
$F_{ST}$ . . . 7
PCA . . . 8
$\pi$, $d_{XY}$ and $d_a$ statistics . . . 9
$f_3$ statistics . . . 10
$D$ and $f_d$ . . . 11

**Additional results** **11**

**Additional tables** **12**

**Additional figures** **17**

# Additional methods

## High Molecular Weight DNA extraction

For each of the 17 species that lacked a reference genome, we collected fresh gill tissue or a portion of the entire body (for small species < 100mm) from a single individual and submitted it to slow lysis at room temperature in a 50 ml falcon tube containing 25 ml of TNES-Urea solution (in sterile water, 10 mM Tris-HCl pH 7.4, 120 mM NaCl, 10 mM EDTA pH 8.0, 0.5% SDS, 4 M urea, final pH adjusted to pH 8). Tissue lysis was performed without vortexing the tube to avoid DNA fragmentation. The solution was temporarily mixed gently by inverting the Falcon tube several times. At least three to four weeks were necessary to achieve good tissue solubilization before DNA extraction. High Molecular Weight genomic DNA (HMW gDNA) extraction followed a previously established phenol-chloroform protocol (Nakayama et al., 1994). The TNES-Urea solution containing lysed tissues was treated with proteinase K at a final concentration of 150 $\mu g/ml$, and the solution was incubated overnight at 37 °C in a rolling oven at a very slow speed. The solution was then extracted by 2 phenol-chloroform followed by 2 chloroform extractions with Phase-lock gel (`https://us.vwr.com/store/product/17942399/phase-lock-gel-quantabio`) to prevent shearing the gDNA by pipetting. After the last extraction, the aqueous supernatant was precipitated with 2 volumes of EtOH 100% and the pellet was hooked from the solution using a sterile Pasteur pipette. The pellet was then rinsed several times in EtOH 80%, moderately dried and stored in TE buffer at 4 °C for 4 days for resuspension of HMW gDNA.

Resuspended HWM gDNA was then submitted to treatment for repairing basic DNA damages such as single-strand nicks using the NEBNext FFPE DNA Repair Mix and following the instruction manual. Repaired HMW gDNA was quantified by microfluorimetry (QubiT) and diluted to a standardized concentration of 1 ng/l before being sized by capillary electrophoresis on a Fragment Analyzer (Agilent) and a TapeStation Genomic DNA ScreenTape Analysis (Agilent) to check the presence of a peak size around 60,000 bp, corresponding to the maximal resolution size of these instruments.

Repaired HMW gDNA was submitted to high throughput size selection for Next-Gen Sequencing to remove low molecular weight DNA fragments following the Pippin HT (Sage Science) High-Pass protocol, using a 0.75% agarose cassette and setting the high pass threshold value to 40 kb. Size-selected DNA was sized by capillary electrophoresis on a Fragment Analyzer (Agilent). The amount of material of size >40kb were quantified to estimate the total quantity of material suitable for library preparation.

Linked-read 10X Genomics library preparation was performed using the Chromium Genome Library kit and Gel Bead Kit V2 protocol to generate long-range DNA sequences (`https://support.10xgenomics.com/genome-exome/sample-prep/doc/user-guide-chromium-genome-reagent-kit-v2-chemistry`). The Chromium Genome Library kit Protocol produced Illumina sequencing-ready libraries with the following standard paired-end structure (Figure SX):

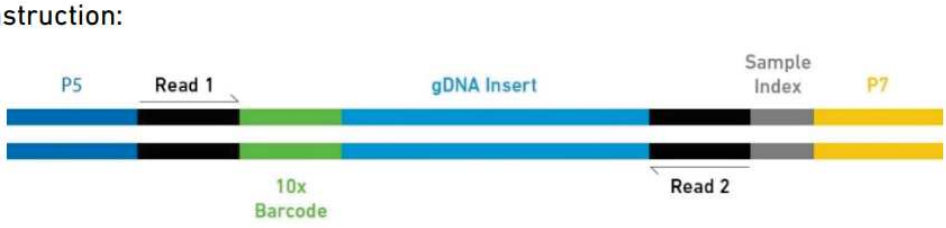



Figure S1: **Structure of Chromium Genome libraries.**

The 16 bp 10x Barcode (one of 737,280 barcode versions) was encoded inline at the start of Read 1, while the sample index sequence was incorporated as the i7 index read using 4 different sequences for each species, in order to balance across all 4 nucleotides (Table SX). Post library construction Quality Control was performed by capillary electrophoresis on a Fragment Analyzer (Agilent). The 17 species libraries were finally multiplexed in equimolar conditions in the final pool library, before being sequenced.

Table S1: **Indexing of the 17 species Chromium Genome libraries.**

| Species library | Index | Sequences |
|---|---|---|
| Lib_Scabr | SI-GA-A10 | GAAACCCT, TTTCTGTC, CCGTGTGA, AGCGAAAG |
| Lib_Eencr | SI-GA-B10 | ACCGTATG, GATTAGAT, CTGACTGA, TGACGCCC |
| Lib_Peryt | SI-GA-C10 | TCTCAGTG, GAGACTAT, CGCTTAGC, ATAGGCCA |
| Lib_Cjuli | SI-GA-D10 | CAATACCC, TGTCTATG, ACCACGAA, GTGGGTGT |
| Lib_Lbude | SI-GA-E10 | AAATGTGC, GGGCAAAT, TCTATCCG, CTCGCGTA |
| Lib_Mmerl | SI-GA-F10 | GCTTGGCT, AAACAAAC, CGGGCTTA, TTCATCGG |
| Lib_Msurm | SI-GA-G10 | TCGCCAGC, AATGTTAG, CGATAGCT, GTCAGCTA |
| Lib_Hgutt | SI-GA-A12 | AGTGGAAC, GTCTCCTT, TCACATCA, CAGATGGG |
| Lib_Styph | SI-GA-B12 | TACCACCA, CTAAGTTT, GGGTCAAG, ACTGTGGC |
| Lib_Scine | SI-GA-D12 | GCACAATG, CTTGGTAC, TGCACCGT, AAGTTGCA |
| Lib_Scant | SI-GA-C12 | TCTCGTTT, GGCTAGCG, ATGACCGC, CAAGTAAA |
| Lib_Dpunt | SI-GA-E12 | ACCGGCTC, GAGTTAGT, CGTCCTAG, TTAAAGCA |
| Lib_Lmorm | SI-GA-F12 | TGATGCAT, GCTACTGA, CACCTGCC, ATGGAATG |
| Lib_Aboye | SI-GA-A1 | GGTTTACT, CTAAACGG, TCGGCGTC, AACCGTAA |
| Lib_Ssard | SI-GA-B1 | GTAATCTT, TCCGGAAG, AGTTCGGC, CAGCATCA |
| Lib_Cgale | S I-GA-C1 | CCACTTAT, AACTGGCG, TTGGCATA, GGTAACGC |
| Lib_Gnige | SI-GA-D1 | CACTCGGA, GCTGAATT, TGAAGTAC, ATGCTCCG |

## Reference genome sequencing and assembly

Sequencing of pooled Genome libraries was performed following recommendations for both sequencing coverage and the total number of reads, using a targeted deduped depth of $> 60X$ per species. To reach this target, the library pool comprising 17 species was first sequenced on one lane of a NovaSeq6000 S1 flowcell. Then, a new pool containing 15 species libraries (i.e. *H. guttulatus* and *S. typhle* were not sequenced twice due to smaller genome size in Syngnathidae) was sequenced on one lane of a NovaSeq6000 S4 flowcell. Sequencing was performed in 150pb paired-end mode.

Raw demultiplexed sequencing reads were de-duplicated without reference genome using the Nubeam-dedup program (Dai and Guan, 2020):

```
nubeam-dedup -i1 input_Species_R1.fastq.gz -i2 input_Species_R2.fastq.gz -o1 \\
Species_DEDUP_R1.fastq.gz -o2 Species_DEDUP_R2.fastq.gz -z 6 r 0
```

Deduplicated reads were then processed with the proc10xG set of python scripts (`https://github.com/ucdavis-bioinformatics/proc10xG`), which aim at extracting and trimming GEM barcode information and primer sequence, and filtering reads that are contained in the whitelist of barcodes used for library preparation.

The GEM barcodes were first extracted (16 first bp of reads) and compared to a whitelist of 737 280 1715287 barcodes using process_10xReads.py:

```
process_10xReads.py -a -o Species_DEDUP_PROC -1 Species_DEDUP_R1.fastq.gz -2 \\
Species_DEDUP_R2.fastq.gz
```

Then, the processed reads were filtered for having a matching barcode in the whitelist, allowing at most 1 mismatch (i.e. keeping only MATCH and MISMATCH1 annotations) using filter_10xReads.py:

```
filter_10xReads.py -o Species_DEDUP_FILTERED -L barcode_list.txt -1 \\
Species_DEDUP_PROC_R1.fastq.gz -2 Species_DEDUP_PROC_R2.fastq.gz
```

The filtered reads were finally regenerated using regen_10xReads.py in the original fastq file format suitable for processing in supernova:

```
regen_10xReads.py -o Species_DEDUP_REGEN -1 Species_DEDUP_FILTERED_R1.fastq.gz -2 \\
Species_DEDUP_FILTERED_R2.fastq.gz
```

We used the software package Supernova-2.1.1 (https://support.10xgenomics.com/de-novo-assembly/software/release-notes/2-1) for de novo assembly of reference genomes from Chromium Linked-Reads (Weisenfeld et al., 2017).

```
supernova run {id Species_DEDUP_FILTERED {outprefix=Species_DEDUP_FILTERED \\
{maxreads=all {accept-extreme-coverage
```

The pseudohap style was used to generate a reference genome with a single record per scaffold, which usually consists of a mix of maternal and paternal alleles within each scaffold.

```
supernova mkoutput {style=pseudohap {asmdir=./Species_DEDUP_FILTERED/outs/assembly \\
{outprefix=Species_DEDUP_FILTERED
```

## Mapping to reference genome

We first index reference as:

```
bwa index -p $REFERENCE_GENOME_SPECIES_OUTPUT -a bwtsw $REFERENCE GENOME
```

and then we mapped each individual to the reference as:

```
bwa mem -M -t 16 $REFERENCE_GENOME $FASTP_R1 $FASTP_R2 > $SAM
```

We convert each sam file as binary files (bam):

```
picard SortSam I=$SAM O=$BAM SO=coordinate CREATE_INDEX=true
VALIDATION_STRINGENCY=LENIENT
```

PCR duplicates are defined as any two reads that came from the same original amplified DNA fragment. It is a recurrent problem in Next-Generation Sequencing and can create false signals of homozygosity and bias variant calling. In Illumina sequencing technologies, after amplification, reads from the same DNA fragment went into different primer lawns on the Illumina flowcell creating PCR duplicates. We identified and mark duplicates with picard tools, MarkDuplicates:

```
picard MarkDuplicates I=$BAM O=$MARKDUP_BAM ASSUME_SORTED=TRUE
REMOVE_DUPLICATES=FALSE CREATE_INDEX=TRUE METRICS_FILE=$SAMPLE_DUPLICATE_METRICS.txt
VALIDATION_STRINGENCY=LENIENT
```

We finally add read groups on each sample bam, as this is required to variant calling with GATK:

```
picard AddOrReplaceReadGroups I=$MARKDUP_PICARD O=$MARKDUP_RG_PICARD RGPL=ILLUMINA
RGLB=lib RGPU=genewiz RGSM=$SAMPLE
```

## Variant Calling

First, we created sequence dictionary, which is mandatory for all subsequent analyses with GATK:

```
picard CreateSequenceDictionary R=REFERENCE_GENOME O=REF_DICT &&
samtools faidx REFERENCE GENOME
```

Then, we inferred germline SNPs per individual using `HaplotypeCaller`. It works in 3 steps: i) defined regions where there is evidence of genetic variation, ii) discard mapping made by `bwa` and make a local reassembly in each of these regions, iii) infer the likelihood of the variant position and of the genotype. The chromosome, the genomic position, the alternate nucleotide and the genotype likelihood is stored in a single-sample GenotypeVCFs file (GVCF):

```
samtools index $MARKDUP_RG_PICARD &&
gatk HaplotypeCaller -R $REFERENCE_GENOME -I $MARKDUP_RG_PICARD -O $GVCF_SAMPLE
-bamout $REALIGN_BAM -ERC GVCF -G StandardAnnotation -G AS_StandardAnnotation
-G StandardHCAnnotation
```

Then, we performed joint genotyping on all samples to produce a VCF (Variant Call Format) per species. However, `GATK` can not perform joint genotyping on multiple GVCF files. We produced a genomic database from all single-sample GVCFs with `GenomicsDBImport`. As we created one database per genomic scaffold, we first had to create a list with all the names of the genomic scaffolds as they are named in the reference genome fasta file, and create a file per interval containing the name of one genomic scaffold.

```
grep '>' $REFERENCE_GENOME | cut -d '>' -f 2 >> inter &&
cut -f1 -d' ' inter >> $INTERVAL_LIST &&
rm inter &&
mkdir interval_split/ &&
INT=$(wc -l $INTERVAL_LIST | awk '{print $1}') &&
cd interval_split/ &&
split -l 1 ../$INTERVAL_LIST &&
i=0 &&
for fi in x*;do mv "$fi" $i.list ; i=$((i+1)) ; done
```

We also create a file containing the name and the absolute path of the corresponding GVCFs with a custom `R` script:

```
R --vanilla --slave --args $SPECIES < get_name.R
```

Then we created one genomic database per genomic scaffold with `GenomicsDBImport`:

```
gatk GenomicsDBImport --sample-name-map $NAME_SAMPLE
--genomicsdb-workspace-path $JOINT_GENOTYPING_INTERVALS_DATABASE
--overwrite-existing-genomicsdb-workspace -L $INTERVAL
```

We perform joint genotyping separately for each genomic scaffold to create a VCF per genomic scaffold:

```
gatk GenotypeGVCFs -R REFERENCE_GENOME
-V gendb://$JOINT_GENOTYPING_INTERVALS_DATABASE -G StandardAnnotation
-O single_vcf/$INTERVALS_joint_gvcf.vcf -L $INTERVAL
```

We finally merge all VCFs to create a single-species VCFs with `vcf-concat` from `vcftools`:

```
ls single_vcf/ | grep 'vcf$' | awk '{print length, $0}' | sort -n
| cut -d' ' -f2- > single_vcf/list_vcf &&
cd single_vcf/ &&
vcf-concat -f single_vcf/list_vcf | bgzip -c > $VCF_SPECIES.vcf.gz
```

## Variant Calling Filtering

We first remove all variant around 5bp of insertion and deletion (indels) as variant calling is difficult around indels even we local-reassembly implemented in `GATK` as in (Fuller et al., 2020):

```
bcftools filter -g 5 --output $VCF_indel5bp.vcf.gz $VCF_SPECIES.vcf.gz
```

Then, we remove indels and multi-allelic variants:

```
vcftools --gzvcf $VCF_indel5bp.vcf.gz --remove-indels --max-alleles 2
--recode --stdout | bgzip > $VCF_indel5bp_snponly.vcf.gz
```

We evaluate the quality and the statistics of the variant calling with `bcftools`, `vcftools` and `rtg` tools:

```
bcftools stats $VCF_indel5bp_snponly.vcf.gz > $VCF_STATS_SPECIES &&
rtg vcfstats $VCF_indel5bp_snponly.vcf.gz > $VCF_STATS_SPECIES_RTG &&
vcftools --depth --gzvcf $VCF_indel5bp_snponly.vcf.gz --out indiv_species &&
vcftools --site-mean-depth --gzvcf $VCF_indel5bp_snponly.vcf.gz --out $SPECIES &&
vcftools --site-quality --gzvcf $VCF_indel5bp_snponly.vcf.gz --out $SPECIES
```

## Genome annotation

We used BUSCO to estimate the pourcentage of present, fragmented and missing genes in each reference genome. We used the actinopterygian database as:

```
busco -i REFERENCE_GENOME -o OUT_SPECIES_BUSCO -m genome -l actinopterygii_odb10
--metaeuk_parameters=--disk-space-limit=5G,--remove-tmp-files=1
--metaeuk_rerun_parameters=--disk-space-limit=5G,--remove-tmp-files=1
```

## Phylogeny

We estimated species phylogeny from all single copy genes found with `BUSCO` for all 20 studied species and for one outgroup, *Lepisosteus oculatus* (Braasch et al., 2016). We found 87 single-copy orthologs for all the 21 species with `Orthofinder v.2.3.8` (Emms and Kelly, 2019).

```
for sp in Aboye Afall Cgale Cjuli Dlabr Dpunt Eencr Gnige Hgutt Lbude Lmorm Mmerl Msurm
Peryt Scabr Scant Scine Spilc Ssard Styph Lepisosteus_oculatus
do
cat /DATA/sdb1/Pierre/BUSCO/${sp}_BUSCO/run_actinopterygii_odb10/busco_sequences/
single_copy_busco_sequences/*.faa >
/home/labosea1/busco/proteome_outgroup/tmp.fas
sed "s/>.*/&_$sp/" /home/labosea1/busco/proteome_outgroup/tmp.fas >
/home/labosea1/busco/proteome_outgroup/${sp}.fas
rm /home/labosea1/busco/proteome_outgroup/tmp.fas
done
orthofinder -f proteome_outgroup/
```

We align each of orthologs with `mafft` (Katoh et al., 2019), concatenate each gene alignment and infer phylogenetic tree with `iqtree v.2.1.2` (Minh et al., 2020). We first determine the best protein evolution model with:

```
iqtree -s all.fa -m MF -o Lepisosteus_oculatus -T 4
```

The best model was JTT+F+R4 and we infer 1000 ultrafast bootstrap with:

```
iqtree -s all.fa -m JTT+F+R4 -o Lepisosteus_oculatus -B 1000 -T 4
```

We run another with UFBoot facilities which reduce overestimation of branch support due to model violations:

```
iqtree -s all.fa -m JTT+F+R4 -o Lepisosteus_oculatus -B 1000 -bnni -T 4
```

We finally assess branch support with standard non parametric boostrap:

```
iqtree -s all.fa -m JTT+F+R4 -o Lepisosteus_oculatus -b 100 -T 4
```

## $F_{ST}$

We estimated genome-wide $F_{ST}$ between pairs of populations using two methods. First, we used `vcftools` (Danecek et al., 2011) to estimate unweighted and weight $F_{ST}$'s Weir and Cockerham (Weir and Cockerham, 1984):

```
bcftools query -l $VCF | grep 'Li' > Li &&
bcftools query -l $VCF | grep 'Mu' > Mu &&
bcftools query -l $VCF | grep 'Fa' > Fa &&
bcftools query -l $VCF | grep 'Ga' > Ga &&
sudo vcftools --gzvcf $VCF --weir-fst-pop Li --weir-fst-pop Mu
--fst-window-size 10000 --out Li_Mu &&
sudo vcftools --gzvcf $VCF --weir-fst-pop Li --weir-fst-pop Fa
--fst-window-size 10000 --out Li_Fa &&
sudo vcftools --gzvcf $VCF --weir-fst-pop Li --weir-fst-pop Ga
--fst-window-size 10000 --out Li_Ga &&
sudo vcftools --gzvcf $VCF --weir-fst-pop Mu --weir-fst-pop Fa
--fst-window-size 10000 --out Mu_Fa &&
sudo vcftools --gzvcf $VCF --weir-fst-pop Mu --weir-fst-pop Ga
--fst-window-size 10000 --out Mu_Ga &&
sudo vcftools --gzvcf $VCF --weir-fst-pop Fa --weir-fst-pop Ga
--fst-window-size 10000 --out Fa_Ga
```

We also used `scikit-allel` (Miles et al., 2020) to estimate Hudon's $F_{ST}$ (Hudson et al., 1992):

```
import allel
vcf=allel.read_vcf($VCF,fields=['samples', 'calldata/GT', 'variants/ALT',
'variants/CHROM', 'variants/FILTER_PASS', 'variants/ID', 'variants/POS',
'variants/QUAL', 'variants/REF', 'variants/DP','variants/AF'],log=sys.stdout)
gt = allel.GenotypeArray(vcf['calldata/GT'])
h = gt.to_haplotypes
ac=gt.count_alleles_subpops(subpops)
fst_hudson, fst_hudson_se, fst_hudson_vb, _=
allel.blockwise_hudson_fst(ac[POP1],ac[POP2],blen=10000)
print("A-B Hudson 's Fst: %.3f +/- %.3f" % (fst_hudson,fst_hudson_se))
```

We also estimated $F_{ST}$ 's Weir and Cockerham:

```
import allel
vcf=allel.read_vcf($VCF,fields=['samples', 'calldata/GT', 'variants/ALT',
'variants/CHROM', 'variants/FILTER_PASS', 'variants/ID', 'variants/POS',
'variants/QUAL', 'variants/REF', 'variants/DP','variants/AF'],log=sys.stdout)
gt = allel.GenotypeArray(vcf['calldata/GT'])
h = gt.to_haplotypes
ac=gt.count_alleles_subpops(subpops)
fst_weirandcockerham, fst_weirandcockerham_se, fst_weirandcockerham_vb, _=
allel.blockwise_weir_cockerham_fst(gt,subpops=
[subpops[A],subpops[B]],blen=10000,max_allele=1)
print("A-B Weir & Cockerham 's Fst:
%.3f +/- %.3f" % (fst_weirandcockerham,fst_weirandcockerham_se))
```

In the two cases, we used the block-jackknife method with a window size of 10kbp to estimate $F_{ST}$ standard error. All $F_{ST}$ estimators gave similar estimations.

# PCA

We ran PCA analyses from the filtered biallelic locus-only VCF with `SNPRelate` tools in `R`. We evaluated the effect of two parameters: minor allele frequency (maf) and linkage disequilibrium between loci. We evaluated the effect of low-frequency variants on population structure by filtering maf from 0 to 0.5 by 0.05. We also evaluated the importance of loci in strong linkage disequilibrium on population structure by pruning the dataset to only keep loci in low linkage disequilibrium ($LD <= 0.2$).

```
library(SNPRelate)
showfile.gds(closeall=TRUE)
VCF_PATH=paste("/DATA/sdb1/Pierre/VCF/",sp,"/",sp,"_indel5bp_snponly.vcf.gz",sep="")

if (file.exists(paste(sp,".gds",sep=""))==F){
vcf.fn <- VCF_PATH
seqVCF2GDS(vcf.fn,paste(sp,".gds",sep=""))
}
```

```
# OPEN GDS AND RUN PCA
genofile <- seqOpen(paste(sp,".gds",sep=""))

genofile_ld02 <- snpgdsLDpruning(genofile,autosome.only=F)

maf_stat=seq(0,0.5,by=0.05)

pca_noprune=vector('list',length(maf_stat))
names(pca_noprune)=paste("maf",maf_stat,sep="")
for (j in 1:length(maf_stat)){
pca <- snpgdsPCA(genofile, num.thread=5,autosome.only = FALSE,maf=maf_stat[j])
pca_noprune[[j]]=list(pca[-2])
}

pca_prune=vector('list',length(maf_stat))
names(pca_prune)=paste("maf",maf_stat,sep="")
for (j in 1:length(maf_stat)){
pca <- snpgdsPCA(genofile, num.thread=5,autosome.only = FALSE,snp.id=as.vector(unlist(ge
pca_prune[[j]]=list(pca[-2])
}

PCA_list[[row]]=list(pca_noprune,pca_prune,
as.vector(unlist(genofile_ld02)))
names(PCA_list[[row]])=c("Raw","Prune","SNP_Prune")
```

## $\pi$, $d_{XY}$ and $d_a$ statistics

We generated VCF with all sites, including non variant ones to estimate unbiased $d_{XY}$. From genomic database created with `GenomicsDBImport`, we re-ran `GenotypeGVCFs` to perform joint genotyping on each genomic database but adding the option `--include-non-variant-sites`:

```
gatk GenotypeGVCFs -R REFERENCE_GENOME
-V gendb://$JOINT_GENOTYPING_INTERVALS_DATABASE -G StandardAnnotation
-O single_vcf/$INTERVALS_allsites_joint_gvcf.vcf --include-non-variant-sites
-L $INTERVAL
```

We estimated $\pi$ following Korunes and Samuk (2021) (see figures in `https://pixy.readthedocs.io/en/latest/about.html`): briefly, we discarded site where there is no nucleotide on the reference genome ($N$) and missing individual genotypes on other sites (.). Then, for each site, we made all the pairwise genotypes comparison (e.g., for 10 individual genotypes, $10 * 9/2 = 45$ possible comparisons, called the denominator) and count the number of comparisons with different bases (called the numerator). We repeat this operation for every sites and summing all site-specific numerators and denominators. Then, $\pi$ for a given locus is the ratio of the sum of all sites-specific numerators and denominators. Although this method is much longer to run, because of the necessity to output another VCF, that can be very large, it does not underestimate $\pi$ because of not considering missing data in the reference genome as invariant or not considering missing genotypes as homozygous to the compared genotypes. To visualize this bias, consider no missing data on a reference genome and no missing individual genotypes and randomly replace some base in the reference genome or some genotypes as missing data. If you consider this missing data in the estimation of $\pi$ this will inflate the denominator (number of

pairwise comparisons) without affecting the numerator (because all genotype pairwise comparisons with missing data will be considered as homozygous in any case). But some of the false genotypes might harbor genetic polymorphism that will be falsely considered as homozygous.

Another bias in the estimation come from the filtering on biallelic sites only: variant sites with more than one allele (multiallelic sites) are frequently filtered for population genomics analyses. However, it can also underestimate $\pi$ as this means to non-randomly discard sites based on their genetic diversity, and to deliberately remove variants with high polymorphism.

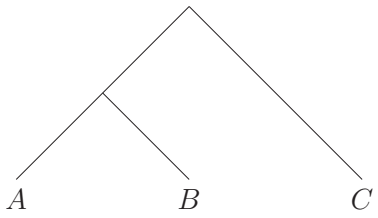
Our estimations of $\pi$ for the European pilchard (*S. pilchardus*) with classic estimation method were about 0.7%, half of the estimation by `GenomeScope` (Barry et al., 2022). By taking into account the 3 previous factors (missing nucleotide in the reference genome, missing genotypes, and multiallelic sites), we retrieved the same results as `GenomeScope`, around 1.4%. We believe that these biases can be important especially for species with low-quality reference genomes and high genetic diversity as for *S. pilchardus*.

This holds also for absolute genetic divergence $d_{XY}$ but in this case, we computed the numerator and the denominator only between pairwise comparisons that consider a genotype from population X and a genotype from population Y.

## $f_3$ statistics

We estimated $f_3$ statistics from allele frequency data, which measure the branch length between a focal population (here $C$) and two other populations (here, $A$ and $B$). Given the following topology and $p_x$, the frequency of a given locus in population $x$, $f_3$ is estimated as the correlation between allele frequencies as:

$$f_3 = (p_c - p_a)(p_c - p_b)$$




A negative value of $f_3$ means that $C$ population is an admixture of the populations $A$ and $B$. We estimated whole-genome $f_3$ for all possible topologies among the 4 sampled populations. We tested if each of the four populations is the result of admixture between 2 out of the 3 other populations, resulting in 12 different topologies and $f_3$ values. We normalized $f_3$ values by dividing the sum of the product by the sum of genetic heterozygosity in population $C$. We used the block-jackknife resampling method to test the significativity of $f_3$ and put a threshold of significativity of $Z$ at -3 following :

```
import allel
vcf=allel.read_vcf($VCF,fields=['samples', 'calldata/GT', 'variants/ALT',
'variants/CHROM', 'variants/FILTER_PASS', 'variants/ID', 'variants/POS',
'variants/QUAL', 'variants/REF', 'variants/DP','variants/AF'],log=sys.stdout)
gt = allel.GenotypeArray(vcf['calldata/GT'])
h = gt.to_haplotypes
ac=gt.count_alleles_subpops(subpops)
```

```
num,den=allel.average_patterson_f3(ac[C],ac[A],ac[B],10000)
f3_C_A_B_snp_num = num
f3_C_A_B_snp_den = den
```

All topologies tested are represented below (Mediterranean locations: Med-out = Gulf of Lion; Med-in = Costa Calida; Atlantic locations: Atl-in = Algarve; Atl-out = Bay of Biscay). For example, the $f_3$ first topology on the top-left tests if Gulf of Lion is the result of admixture between Costa Calida and Algarve. Topologies on the first, second, third and fourth row test if Gulf of Lion, Costa Calida, Algarve and Bay of Biscay locations, respectively, are the results of admixture between the two populations represented in the respective topology.




## $D$ and $f_d$

We infer $D$ with `Dsuite` (Malinsky et al., 2021):

```
Dsuite Dtrios VCF_WITH_ANCESTRAL.vcf.gz POP_FILE.txt DSUITE_SPECIES
```

and then $f_d$ following:

```
Dsuite Dinvestigate VCF_WITH_ANCESTRAL.vcf.gz POP_FILE.txt test_trios.txt
```

# Additional tables

Table S2: **Species de novo reference genome assembly statistics for 17 species.**

| Species | Sample Code | NUMBER OF SCAFFOLDS >= 10 kb | CONTIG N50 (kb) | PHASEBLOCK N50 (kb) | SCAFFOLD N50 (kb) | ASSEMBLY SIZE in scaffolds >= 10 kb | mean molecule length (bp) |
|---|---|---|---|---|---|---|---|
| *A. boyeri* | Aboye_MU_07 | 25330 | 22.07 | 56.34 | 38.13 | 731.8 | 6733.45 |
| *C. galerita* | Cgale_MU_02 | 17580 | 18.17 | 71.48 | 40.6 | 553.08 | 9538.62 |
| *C. julis* | Cjuli_MU_03 | 9840 | 32.72 | 282.04 | 150.8 | 630.73 | 14201.71 |
| *D. puntazzo* | Dpunt_FA_07 | 18470 | 24.4 | 75.58 | 46.61 | 607.58 | 8381.23 |
| *E. encrasicolus* | Eencr_FA_07 | 9840 | 13.08 | 66.53 | 20.36 | 192.02 | 14926.14 |
| *G. niger* | Gnige_MU_07 | 5320 | 14.59 | 3.96 | 16.79 | 89.28 | 4946.24 |
| *H. guttulatus* | Hgutt_GA_13 | 353 | 66.12 | 3890 | 20570 | 425.12 | 35122.52 |
| *L. budegassa* | Lbude_FA_07 | 1460 | 84.92 | 2440 | 7290 | 696.45 | 18212.73 |
| *L. mormyrus* | Lmorm_LI_07 | 17190 | 23.91 | 54.45 | 42.91 | 523.39 | 8381.23 |
| *M. merluccius* | Mmerl_FA_13 | 10790 | 13.11 | 152.38 | 47.59 | 364.4 | 19526.51 |
| *M. surmuletus* | Msurm_MU_01 | 3340 | 43.14 | 1260 | 1020 | 464.28 | 23125.33 |
| *P. erythrinus* | Peryt_LI_06 | 9770 | 38.21 | 247.69 | 126.13 | 631.79 | 11755.32 |
| *S. cabrila* | Scabr_LI_01 | 1680 | 126.1 | 1030 | 1190 | 590.74 | 31481.15 |
| *S. cantharus* | Scant_FA_07 | 5800 | 52.76 | 359.09 | 501.19 | 710.84 | 11872.88 |
| *S. cinereus* | Scine_MU_02 | 8170 | 31.78 | 158.21 | 97.92 | 474.85 | 13378.65 |
| *S. sarda* | Ssard_LI_07 | 4030 | 52.65 | 553.72 | 624.55 | 669.2 | 14487.16 |
| *S. typhle* | Styph_GA_01 | 1970 | 44.09 | 384.19 | 2120 | 308.67 | 17677.09 |

Table S5: **Test for phylogenetic independence for every life-history trait. We tested whether the variation of every life-history trait can be explained by the corresponding phylogeny using two methods: Pagel's $\lambda$ and Bloomberg's K (H0 = independence of phylogeny with life-history traits variation).**

| Traits | Pagel's $\lambda$ | | BLoomberg 's $K$ | |
|---|---|---|---|---|
| | $\lambda$ | $p$-value | $K$ | $p$-value |
| Body size | 6.6107e-5 | 1 | 0.517074 | 0.1923 |
| Trophic level | 0.999934 | 0.240271 | 0.567211 | 0.1473 |
| Fecundity | 6.6107e-5 | 1 | 0.356581 | 0.512 |
| Propagule size | 0.999934 | 0.00542777 | 1.12688 | 0.0544 |
| Age at maturity | 0.999934 | 0.211406 | 0.65942 | 0.1129 |
| Lifespan | 0.999934 | 0.152496 | 0.663148 | 0.0962 |
| Adult lifespan | 0.844909 | 1 | 0.535233 | 0.1569 |
| PLD | 0.99934 | 0.17233 | 0.672768 | 0.0976 |

Table S6: **Test of difference of admixture between non-BUSCO and BUSCO genes**. We inferred $f_3$ statistics taking the Med-in or the Atl-in populations as the admixed populations and compared the value from BUSCO genes and non-BUSCO (mean of 10kb windows outside BUSCO genes). Positive values of diff signify higher $f_3$ for non-BUSCO compared to BUSCO.

| Species | Mediterranean admixed population | | Atlantic admixed population | |
|---|---|---|---|---|
| | diff | $p$-value | diff | $p$-value |
| *C. julis* | 0.026 | $< 2e^{-16}$ | 0.024 | $< 2e^{-16}$ |
| *D. labrax* | -0.012 | 0.29 | 0.006 | $5.40e^{-6}$ |
| *D. puntazzo* | 0.029 | $3.13e^{-15}$ | 0.031 | 0.0003 |
| *H. guttulatus* | 0.0008 | 0.79 | 0.100 | $3.77e^{-8}$ |
| *L. budegassa* | 0.003 | 0.028 | 0.0036 | 0.0008 |
| *L. mormyrus* | -0.0025 | 0.75 | 0.008 | 0.377 |
| *M. merluccius* | 0.0062 | 0.001 | 0.0014 | 0.629 |
| *M. surmuletus* | 0.0007 | 0.70 | 0.017 | $2.78e^{-15}$ |
| *P. erythrinus* | 0.017 | $< 2e^{-16}$ | 0.024 | $< 2e^{-16}$ |
| *S. cabrilla* | 0.022 | $< 2e^{-16}$ | 0.0009 | 0.76 |
| *S. cantharus* | 0.0087 | 0.035 | 0.010 | $2.64e^{-6}$ |
| *S. cinereus* | -0.004 | 0.14 | -0.0189 | 0.32 |
| *S. sarda* | 0.011 | $< 2e^{-16}$ | 0.0125 | $< 2e^{-16}$ |
| *S. typhle* | 0.11 | 0.00005 | 0.037 | $1.92e^{-10}$ |

Table S3: **Species characteristics**. Sampling = number of sampled and sequenced invididuals in total and per locations (Med-out + Med-in + Atl-in + Atl-out); Gen. = generation time inferred by `AgeNe` from life tables in (Barry et al., 2022); $F_{ST}$, $d_{XY}$ and $d_a$ between inner and outer populations; $f_3$ = is inner populations admixed between the two outer populations ?, Med = only Med-in is admixed, Atl = only Atl-in is admixed, Med + Atl = both Med-in and Atl-in are admixed, No = none of Med-in and Atl-in are admixed see Fig. S12; $ABC_{in}$ and $ABC_{out}$ = presence of contemporary gene flow between inner and outer populations respectively, see Fig. S24; $D$ = is one population the donor of a differential introgression between 2 other populations ? Med = the donor is a Mediterranean populations; Atl = the donor is a Atlantic population, Med + Atl = the donor are either a Mediterranean or Atlantic population, No = no differential introgression, see Fig. S13 for further informations

| Species | Charac. | | $F_{ST}$ | | $d_{XY}$ | | $d_a$ | | Introgression | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Sampling | Gen. | In | Out | In | Out | In | Out | $f_3$ | $ABC_{in}$ | $ABC_{out}$ | $D$ |
| *A.boyeri* | 20(5+5+5+5) | - | 0.35 | 0.61 | 2.59 | 3.77 | 3.77 | 0.21 | No | Yes | Yes | - |
| *A.fallax* | 14(5+0+5+4) | - | 0.20 | - | 0.63 | 0.- | 0.23 | - | No | Yes | Yes | - |
| *C. galerita* | 20(5+5+5+5) | 2.346 | 0.60 | 0.61 | 1.52 | 1.51 | 0.87 | 0.87 | No | No | No | - |
| *C. julis* | 20(5+5+5+5) | 3.674 | 0.29 | 0.27 | 2.05 | 2.04 | 0.56 | 0.52 | - | Yes | Yes | - |
| *D. labrax* | 20(5+5+5+5) | 8.952 | 0.14 | 0.09 | 0.47 | 0.44 | 0.09 | 0.07 | Atl | Yes | Yes | Med + Atl |
| *D. puntazzo* | 20(5+4+5+5) | 4.938 | 0.59 | 0.52 | 1.40 | 1.33 | 0.86 | 0.68 | Med+Atl | Yes | No | - |
| *G. niger* | 20(5+5+5+5) | - | 0.12 | 0.11 | 1.61 | - | 0.18 | - | No | - | - | - |
| *H. guttulatus* | 20(5+5+5+11) | 2.034 | 0.20 | 0.23 | 0.31 | 0.32 | 0.02 | 0.02 | No | Yes | Yes | Atl |
| *L. budegassa* | 20(5+4+5+5) | 10.803 | 0.005 | 0.002 | 0.25 | 0.24 | 0.0004 | 0.001 | Med + Atl | Yes | Yes | Atl |
| *L. mormyrus* | 20(5+5+5+5) | 5.821 | 0.64 | 0.61 | 1.67 | 1.66 | 1.05 | 1.03 | Med | Yes | No | - |
| *M. merluccius* | 20(5+5+5+5) | 6.588 | 0.04 | 0.02 | 0.51 | 0.50 | 0.02 | 0.01 | No | Yes | Yes | - |
| *M. surmuletus* | 20(5+5+5+5) | 3.942 | 0.05 | 0.03 | 1.22 | 1.19 | 0.05 | 0.02 | Med | Yes | Yes | Atl |
| *P. erythrinus* | 20(5+5+5+5) | 4.925 | 0.006 | 0.001 | 1.29 | 1.28 | 0.02 | 0.007 | No | Yes | Yes | - |
| *S. cabrilla* | 20(5+5+5+5) | 3.724 | 0.13 | 0.04 | 1.78 | 1.66 | 0.26 | 0.004 | Med | Yes | Yes | Med + Atl |
| *S. cinereus* | 20(5+5+5+5) | 1.904 | 0.27 | 0.06 | 0.70 | 0.69 | 0.09 | 0.02 | Med | Yes | No | Atl + Med |
| *S. cantharus* | 20(5+5+5+5) | 4.779 | 0.42 | 0.41 | 0.90 | 0.89 | 0.42 | 0.40 | No | Yes | No | - |
| *S. pilchardus* | 20(5+5+5+5) | 1.903 | 0.01 | 0.01 | 1.38 | 1.40 | 0.04 | 0.06 | No | Yes | Yes | Atl |
| *S. sarda* | 20(5+5+5+5) | 2.347 | 0.0004 | 0.001 | 0.85 | 0.85 | 0.006 | 0.004 | No | Yes | Yes | - |
| *S. typhle* | 20(5+5+5+5) | 1.675 | 0.39 | 0.019 | 1.27 | 1.26 | 0.11 | 0.16 | No | No | No | Med |

Table S4: **Genome outgroups were chosen to polarized variants and corresponding GenBank access**

| Ingroup | First Outgroup | | Second Outgroup | | Third Outgroup | |
|---|---|---|---|---|---|---|
| | Species | Gen bank access | Species | Gen bank access | Species | Gen bank access |
| *D. labrax* | *Lates calcarifer* | GCA_001640805.1 | *Morone saxatilis* | GCA_004916995.1 | *Gasterosteus aculeatus* | GCA_016920845.1 |
| *H. guttulatus* | *Hippocampus comes* | GCA_001891065.1 | *Hippocampus whitei* | GCA_901007805.1 | *Hippocampus kuda* | GCA_901007745.1 |
| *S. typhle* | *Syngnathus acus* | GCA_901709675.2 | *Syngnathus floridae* | GCA_010014945.1 | *Syngnathus scovelli* | |
| *L. budegassa* | *Lophius piscatorius* | GCA_009660295.1 | *Antennarius maculatus* | GCA_013358685.1 | *Antennarius striatus* | GCA_900303275.1 |
| *S. cabrilla* | *Hypolectrus puella* | GCA_900610375.1 | *Epinephelus moara* | GCA_006386435.1 | *Plectropomus leopardus* | GCA_011397275.1 |
| *M. surmuletus* | *Dactylopterus volitans* | GCA_901007715.1 | *Aeoliscus strigatus* | GCA_901007665.1 | *Callionymus lyra* | GCA_016630915.1 |
| *S. sarda* | *Thunnus albacares* | GCA_900302625.1 | *Thunnus orientalis* | GCA_021601225.1 | *Thunnus thynnus* | GCA_003231725.1 |
| *S. cantharus* | *Sparus aurata* | GCA_900880675.1 | *Diplodus sargus* | GCA_903131615.1 | *Acanthopagrus latus* | GCA_904848185.1 |
| *S. cinereus* | *Symphodus melops* | GCA_002819105.1 | *Labrus bergylta* | GCA_900080235.1 | *Notolabrus celidotus* | GCA_009762535.1 |
| *S.pilchardus* | *Clupea harengus* | GCA_900700415.2 | *Limnothrissa miodon* | GCA_017657215.1 | *Alosa alosa* | GCA_017589495.1 |

# Additional figures

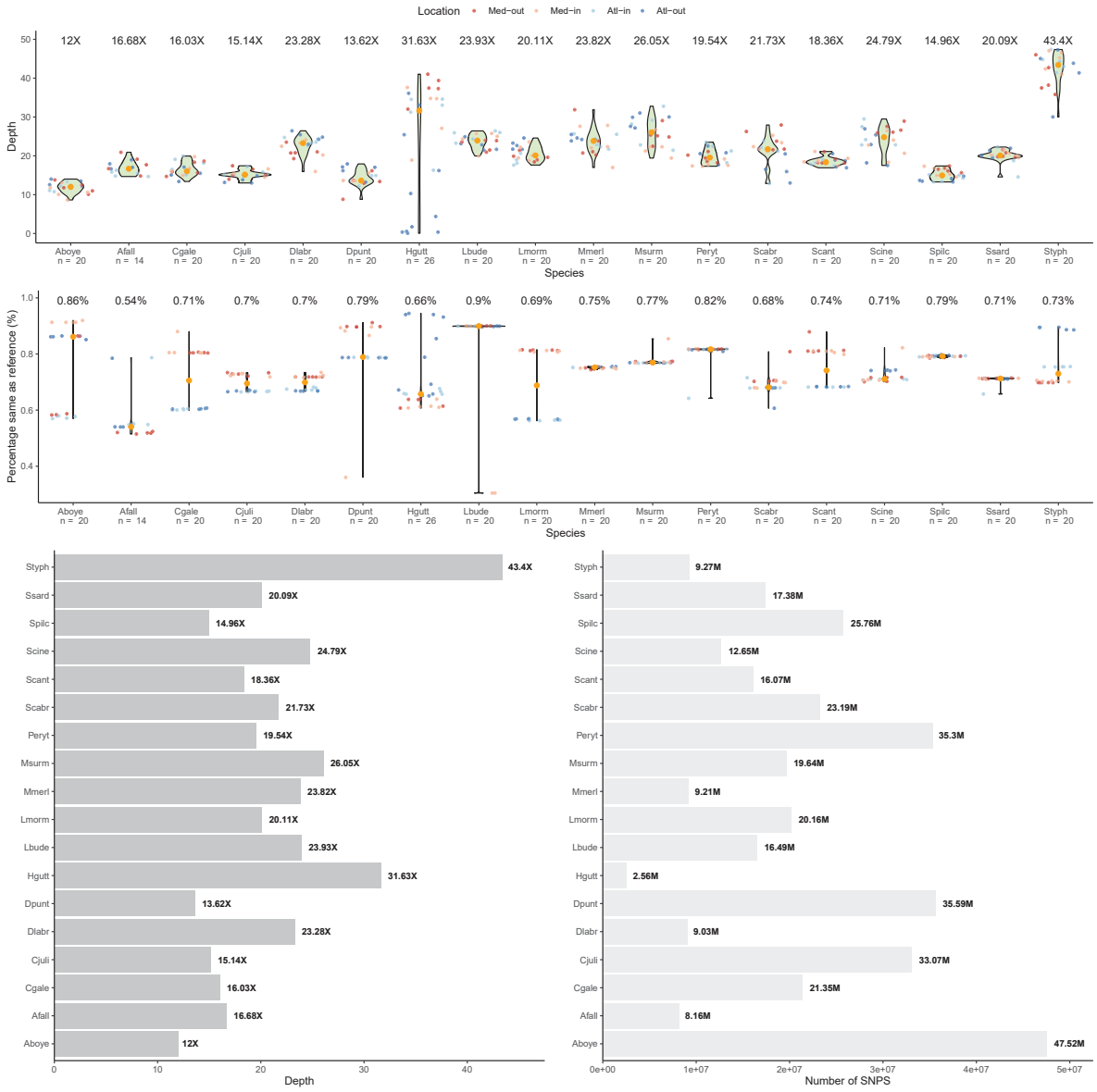



Figure S2: **Statistics of each species VCF**. Top: individual and mean-species depth, Middle: percentage of genotypes identical to the reference genome, Bottom-left: Mean depth per species; Bottom-right = Number of SNPs per species. For top and middle panels, color of the points show the population of the individual (red = Med-out ; orange = Med-in, light-blue = Atl-in, dark-blue = Atl-out).

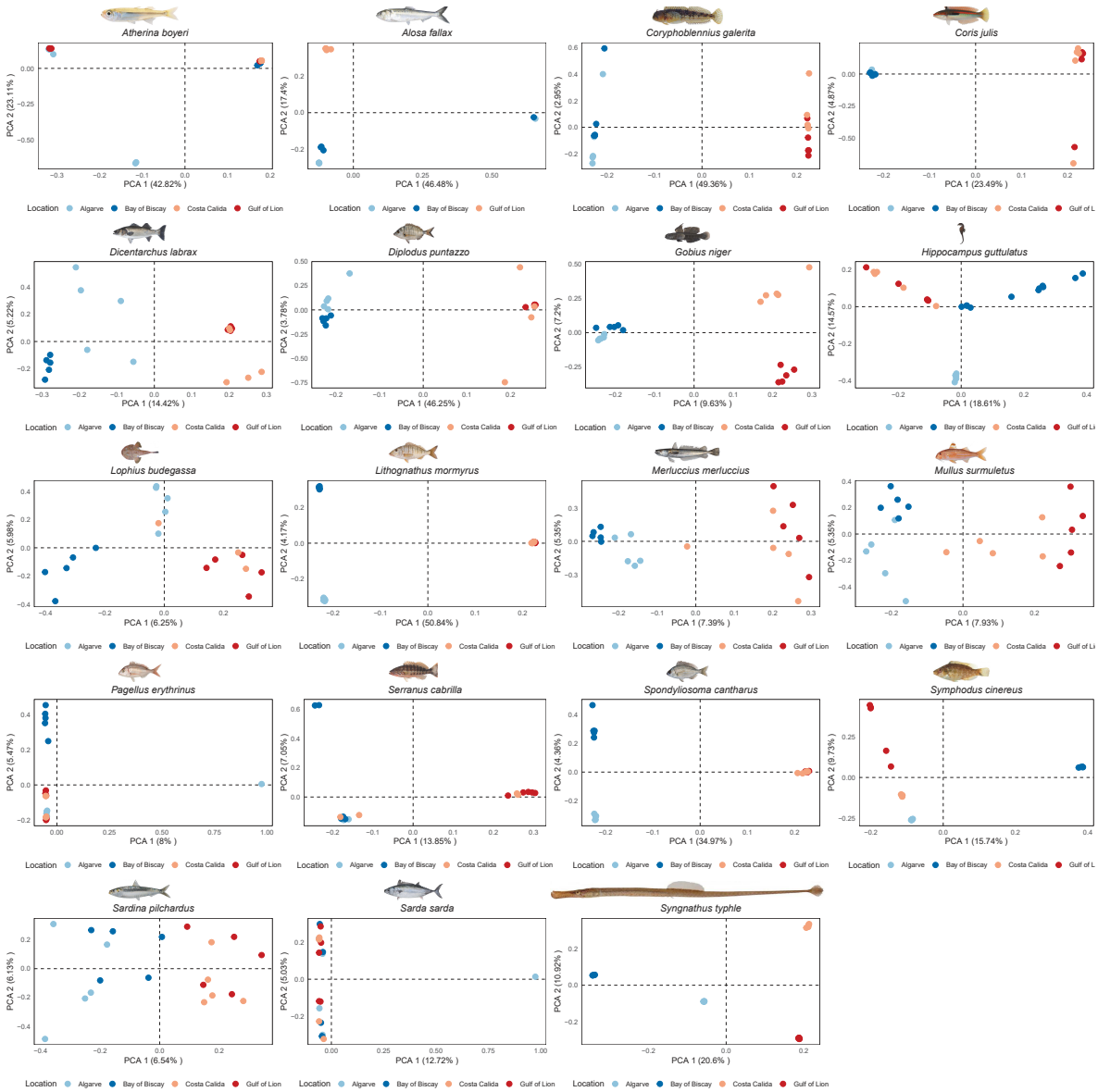



Figure S3: **PCA of all 19 studied species on whole data set with maf = 5%**

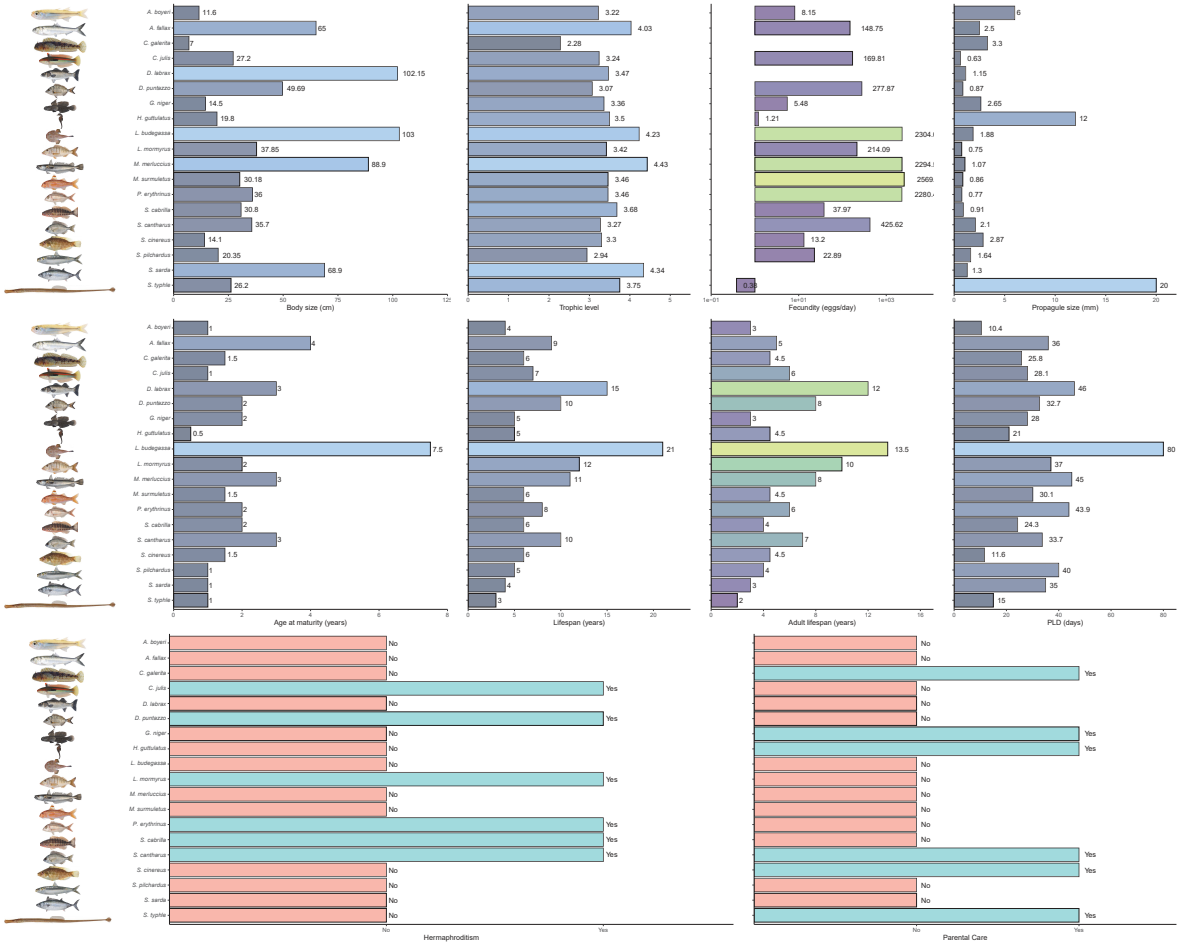



Figure S4: **10 life-history traits variability of the studied species**.

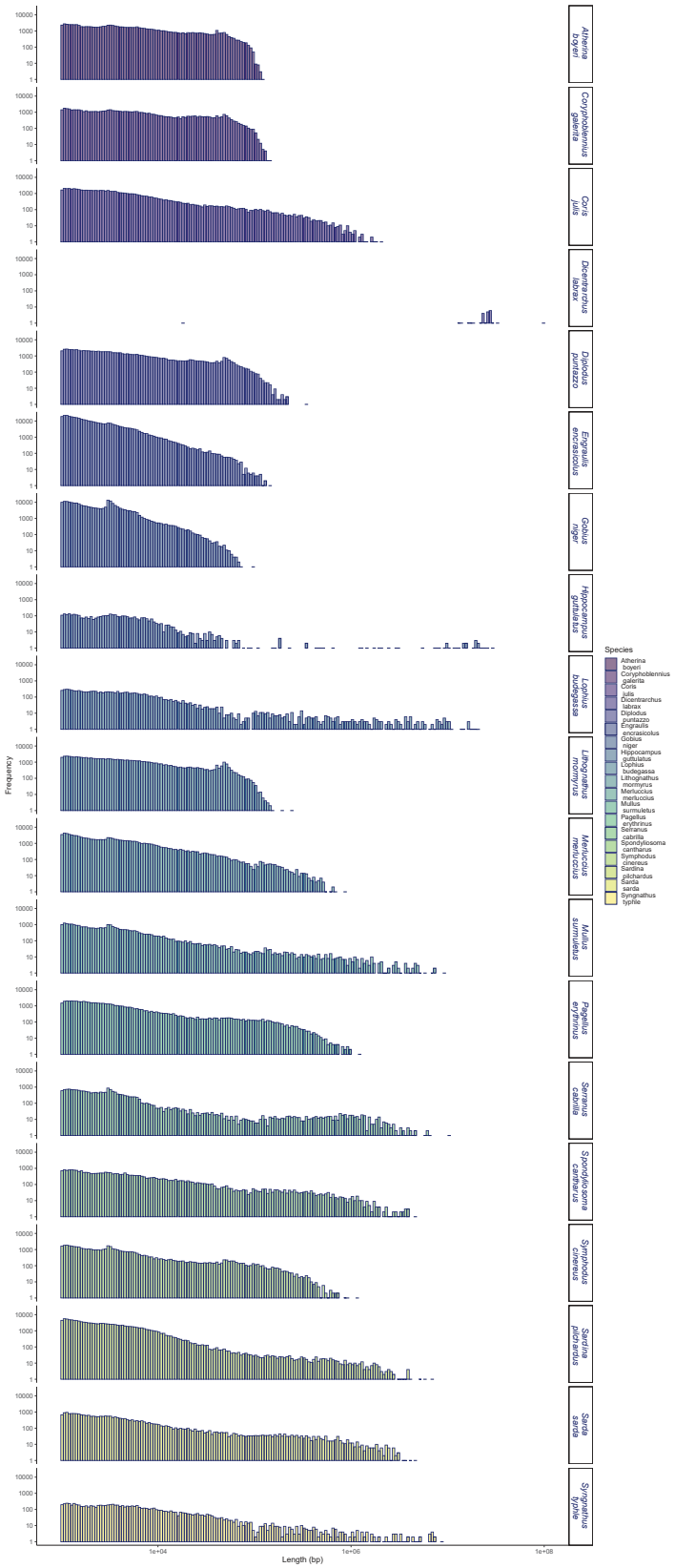



Figure S5: **Length distribution of scaffolds of de-novo references genomes assembly of 19 species**. For reference, chromosome-level assembly of *D. labrax* reference genome was retrieved from Tine et al. (2014). All other reference genomes were assembled during this study (see Methods for supplementary details) except for *S. pilchardus* retrieved from Louro et al. (2019)

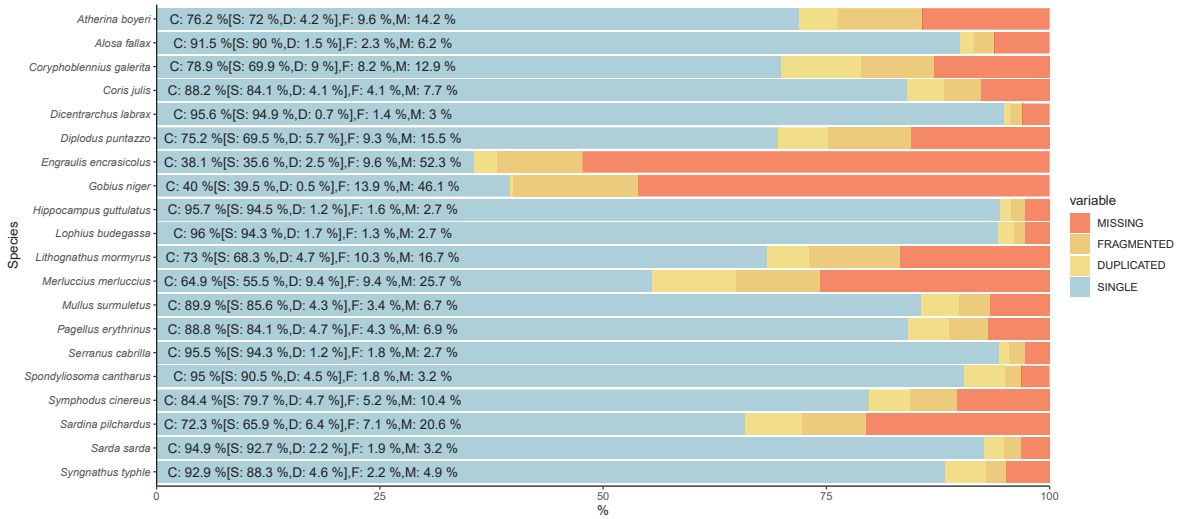



Figure S6: BUSCO **annotation summaries for the 20 studied species**. In blue and light orange are the percentage of respectively single and duplicated complete genes; in dark orange the percentage of fragmented genes and in red the percentage of missing genes using the actinopterygian database. The total genes searched are $n = 3640$.

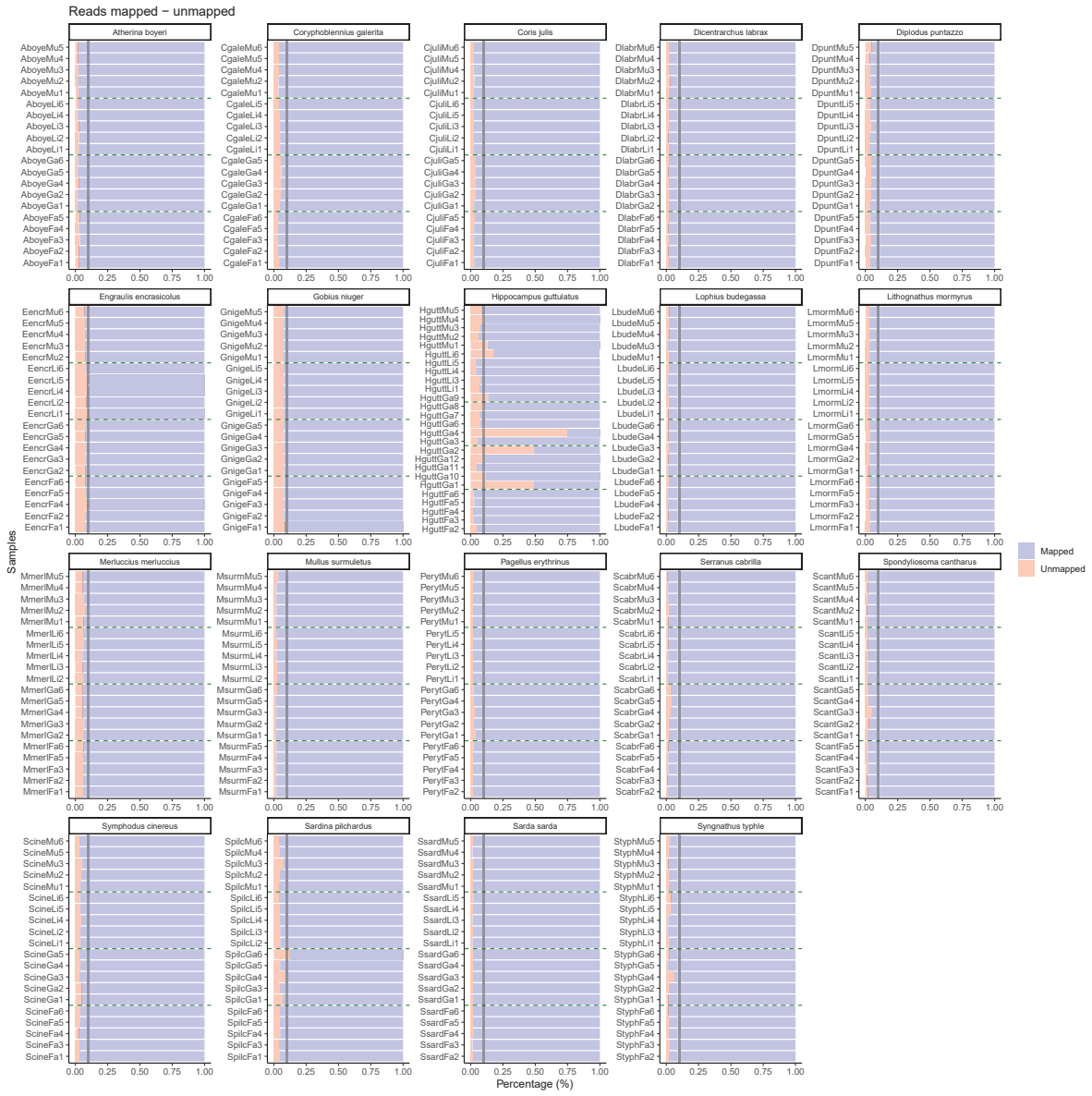



Figure S7: **Percentage of reads mapped per individuals**. Each individual of a given species is represented in row; the percentage of reads mapped and unmapped to the species reference genome are represented in blue and red respectively. The vertical line represents the 10% threshold.

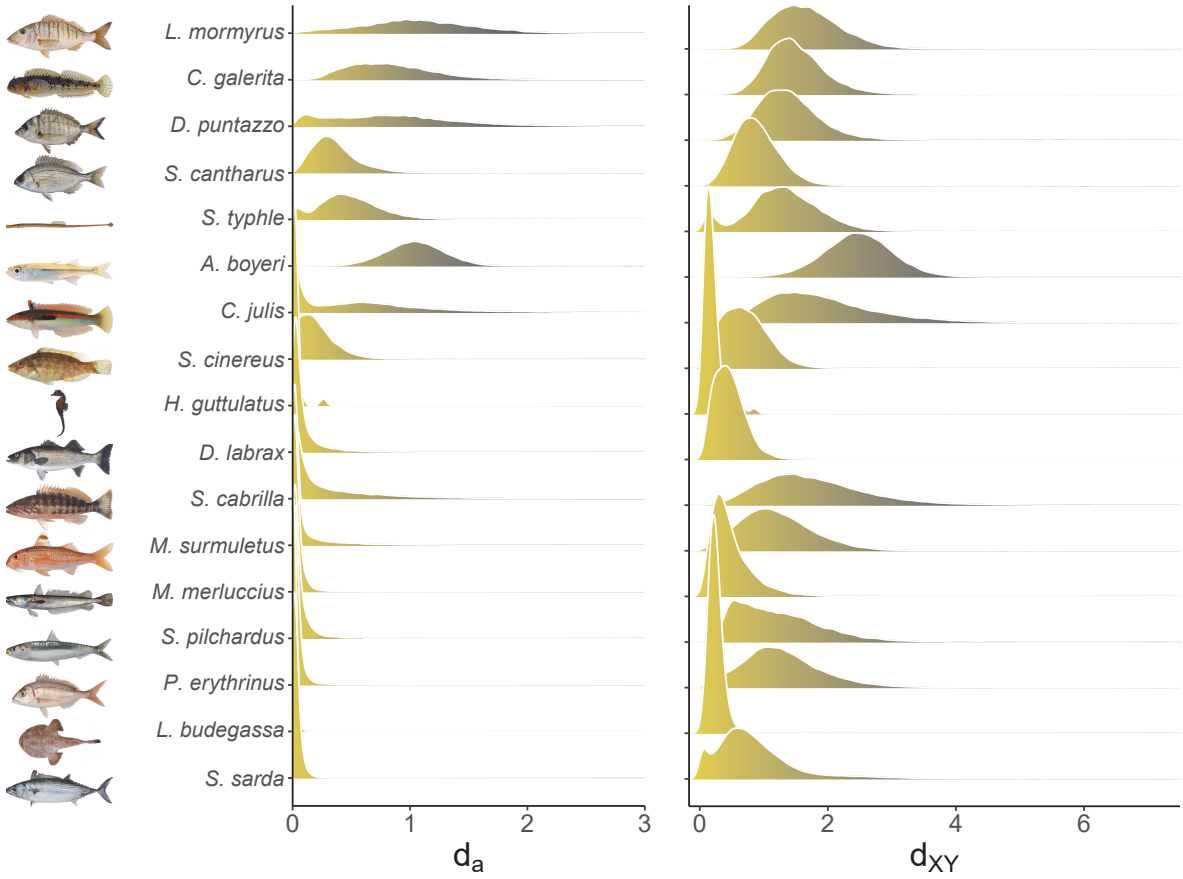



Figure S10: **Distribution in 50kb windows of $d_{XY}$ and $d_a$ for 17 species**. Species are ordered from lower (bottom) to higher (top) genetic differentiation

.

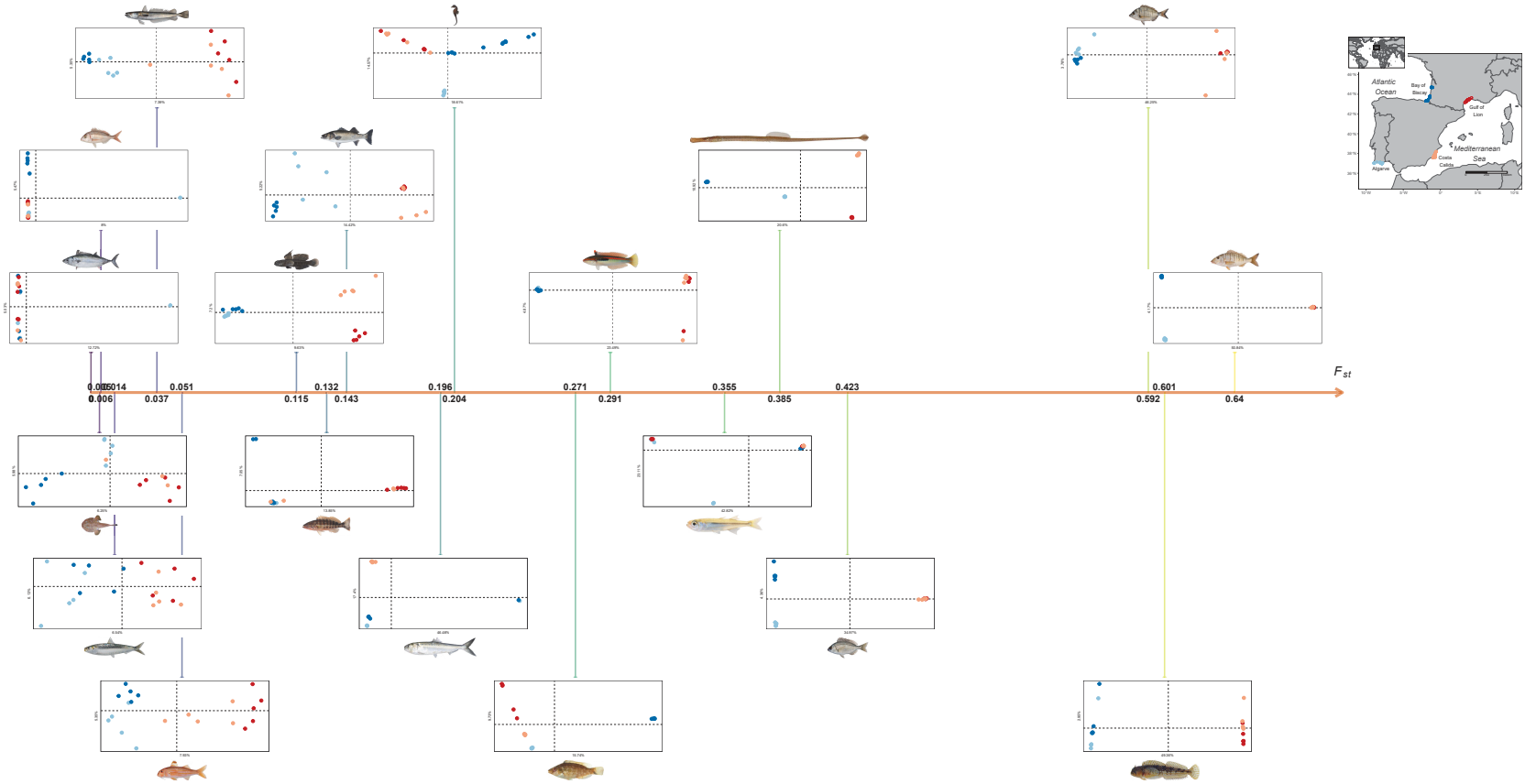



Figure S8: **Continuum of genetic differentiation for 19 marine fish species**. Each species is ordered along an x-axis representing genetic differentiation estimated by $F_{ST}$ between Gulf of Lion (Mediterranean Sea) and Bay of Biscay (Atlantic Ocean) populations. Corresponding values are written in black around the x-axis. Perpendicular to each value, PCA analyses represent genetic population structure: Atlantic Ocean populations: Algarve (light blue), Bay of Biscay (dark blue). Mediterranean Sea populations: Costa Calida (pink), Gulf of Lion (red). PCA components are estimated with biallelic loci only filtering by minor allele frequency superior to 0.05%

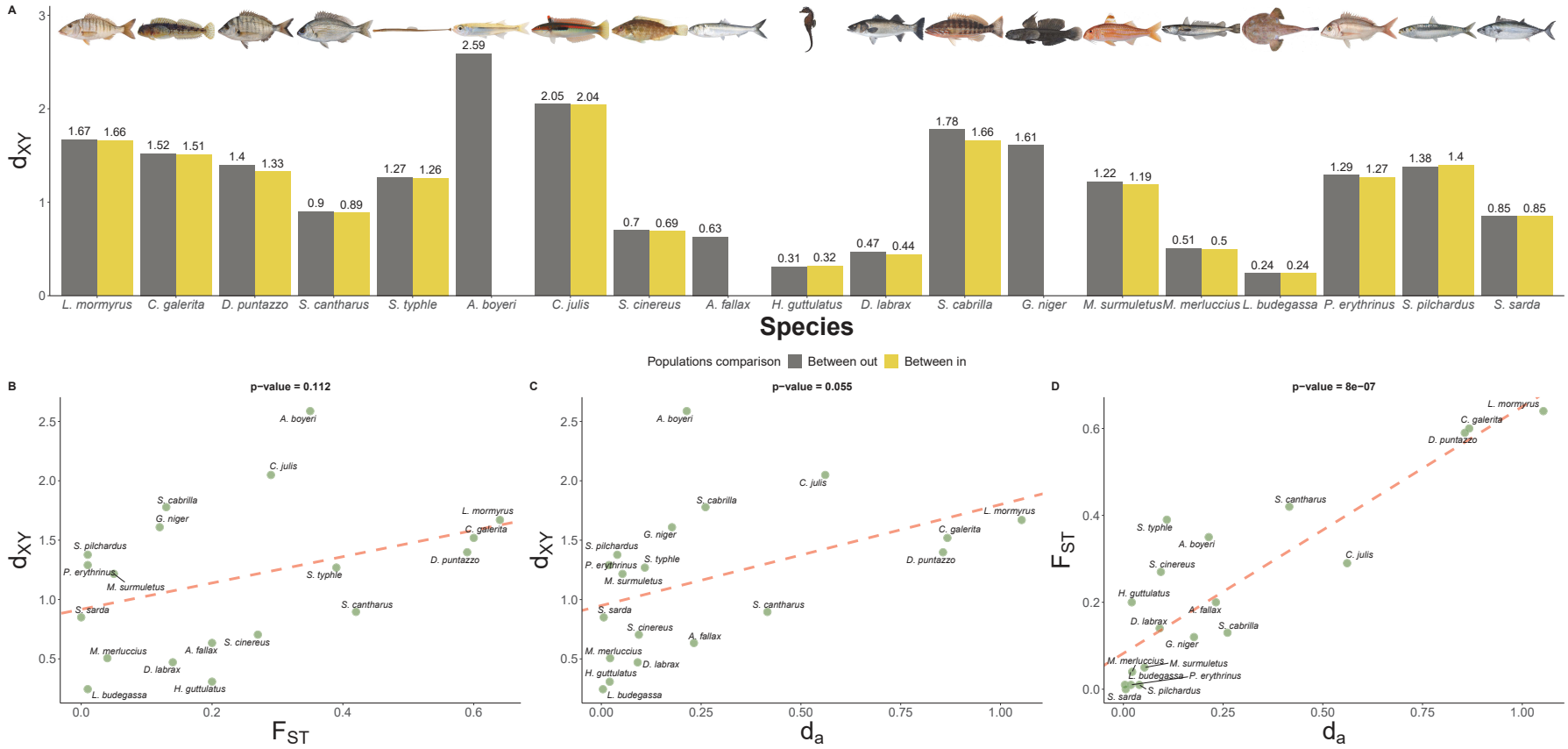



Figure S9: **Gradient of $d_{XY}$ and correlation between $F_{ST}$, $d_{XY}$ and $d_a$.** A) Species are ordered from low (right) to high (left) genetic differentiation, with bars indicating $d_{XY}$ between inner (yellow) and outer (gray) populations. B) Correlation between $d_{XY}$ and $F_{ST}$, C) $d_{XY}$ and $d_a$ and D) $F_{ST}$ and $d_a$

.

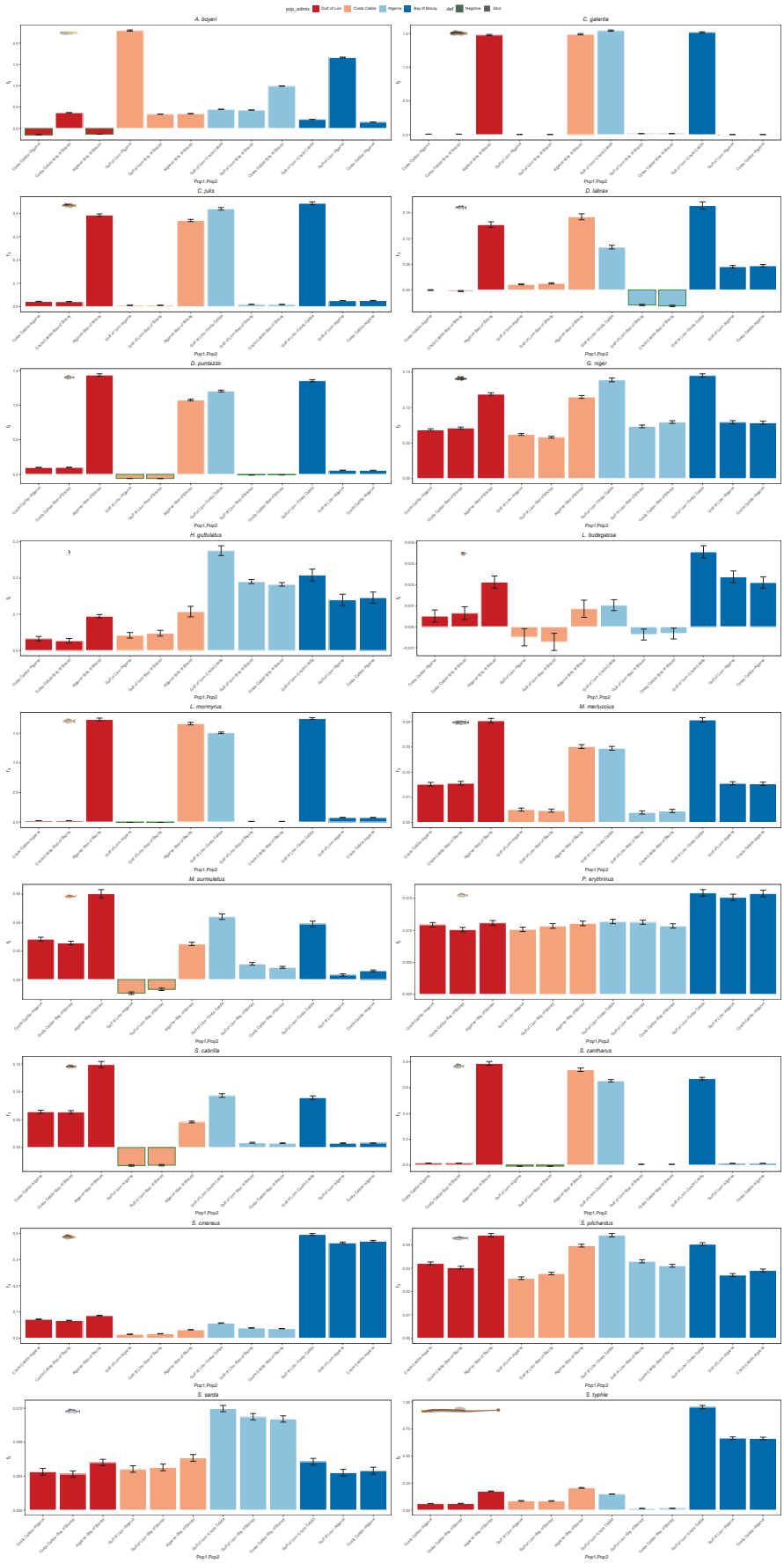


Figure S12: $f_3$ **estimation for all 12 topologies for 18 species**. Colors indicate which population is tested to be the result of admixture between two other populations that are shown on $y$-axis; red = Med-out, orange = Med-in ; light blue = Atl-in , dark blue = Atl-out

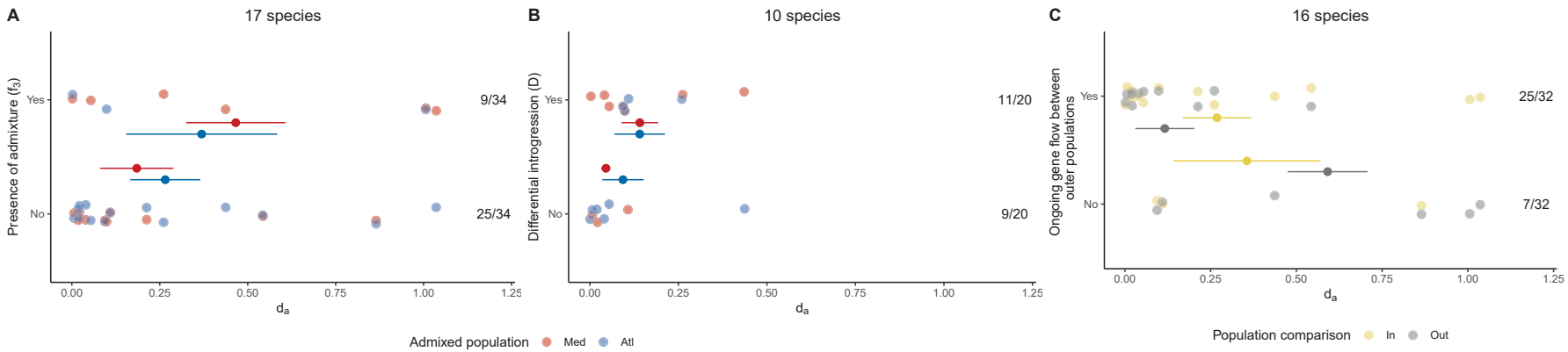



Figure S11: **Admixture, introgression and ongoing gene flow along the gradient of net divergence** $d_a$. Panel A: $f_3$ test showing that several inner populations (either Med-in in violet or Atl-in in blue) result from genetic admixture between the outer populations of the two basins. Panel B: Results of the $D$ statistics showing differential introgression between populations of the same basin (either Med-in violet or Atl-in blue), originating from a population of the other basin. Panel C: Results of ABC inference of contemporary gene flow between either inner (yellow) or outer (grey) populations. In each panel, gene flow results are plotted against the gradient of net divergence measured by $d_a$ between the two outer populations. The mean and standard deviation of $d_a$ across species is reported for each gene flow outcome, as well as the number of observations for each outcome over the total number of tests performed. Gene flow between outer populations is only tested in panel C (grey points), showing a limit to gene flow beyond a certain level of divergence.

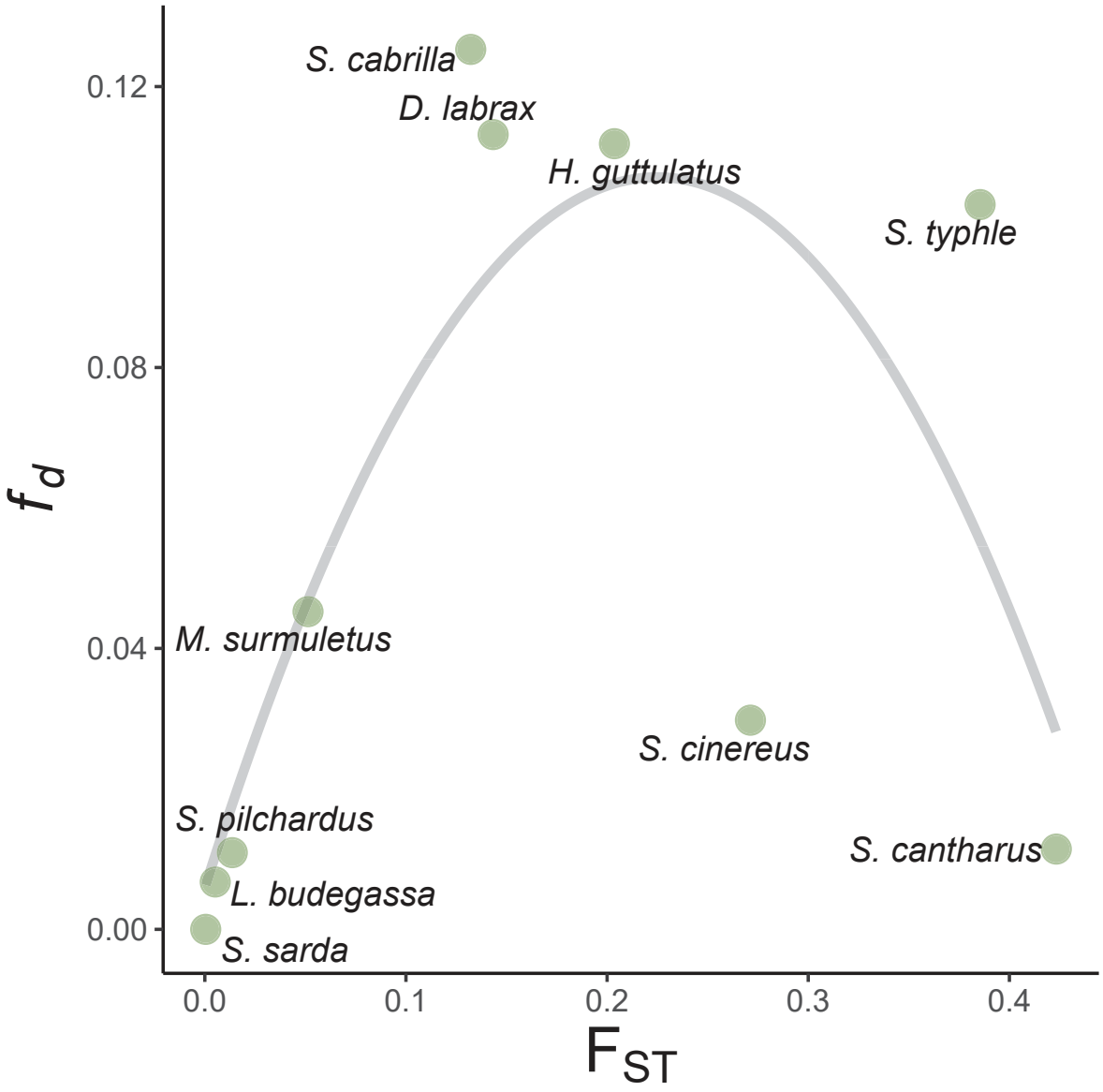



Figure S14: **Inverse polynomial relationship between whole-genome $F_{ST}$ and $f_d$ for 10 species**.

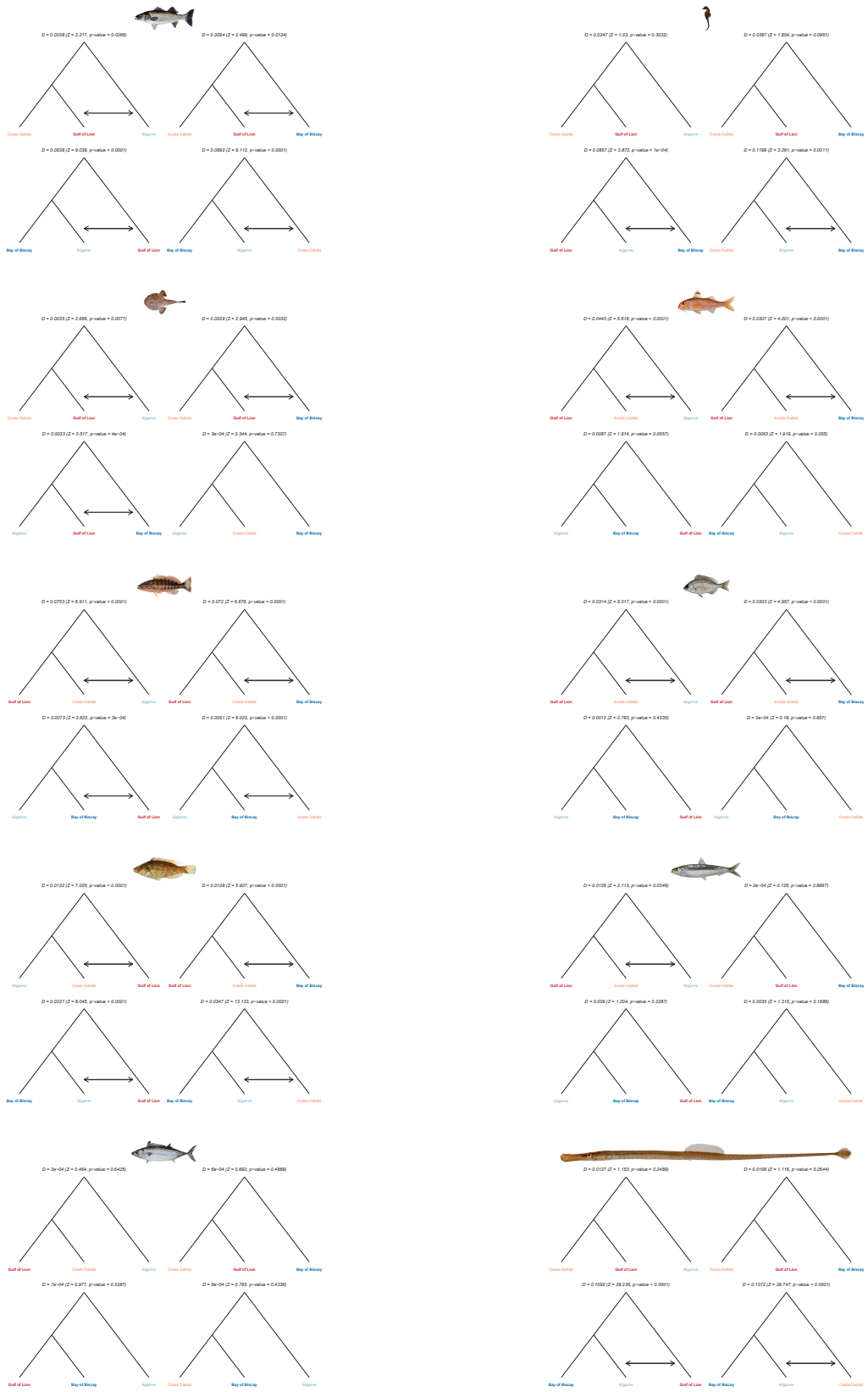


Figure S13: *D* **statistics estimation of 10 species**. *D* was inferred for 4 different topologies per species. *D*, *Z* and *p* value are shown above each topology. A double-sided arrow show significant differential introgression. Species from left to right and top to bottom: *D. labrax, H. guttulatus, L. budegassa, M. surmuletus, S. cabrilla, S. cantharus, S. cinereus, S. pilchardus, S. sarda* and *S. typhle*.

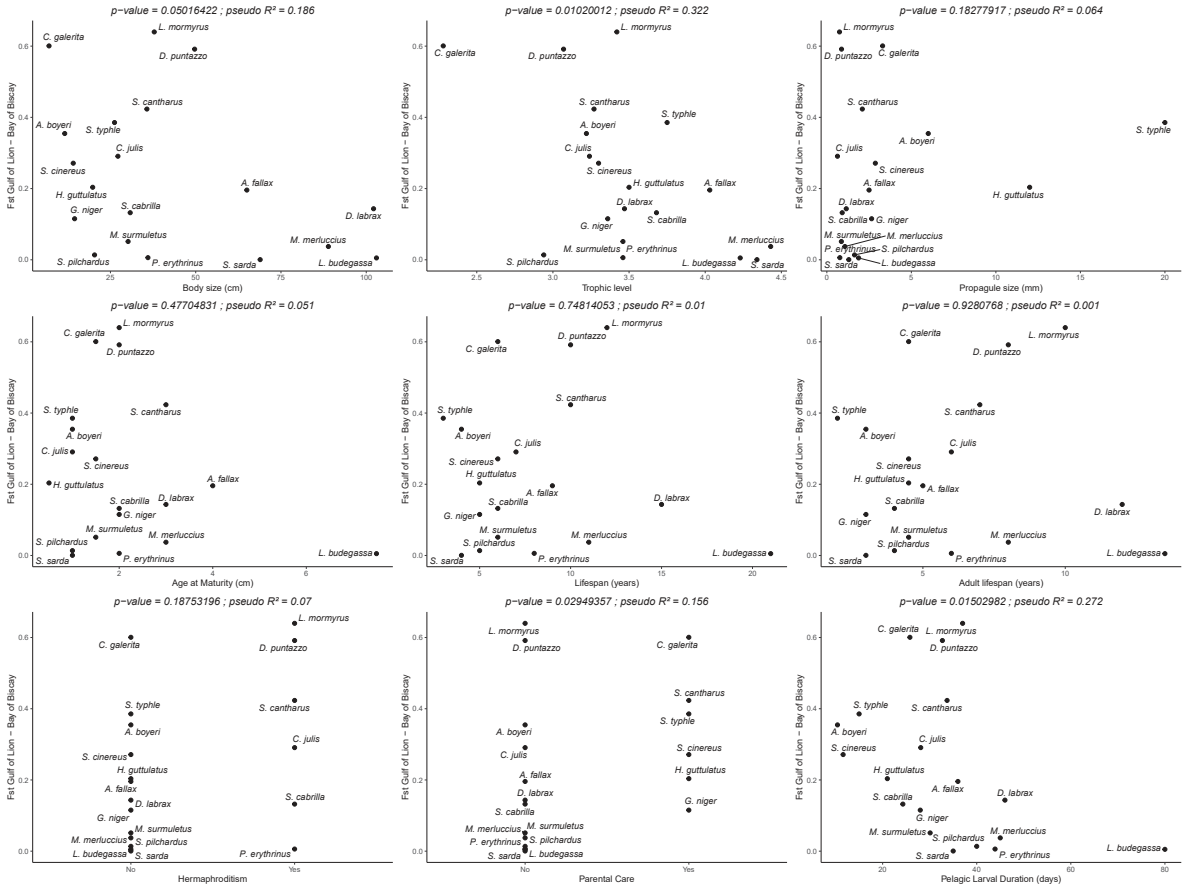



Figure S15: **Correlation between 9 life-history traits and $F_{ST}$ between outer populations for 19 species**. p-value and $R^2$ are shown above each plot.

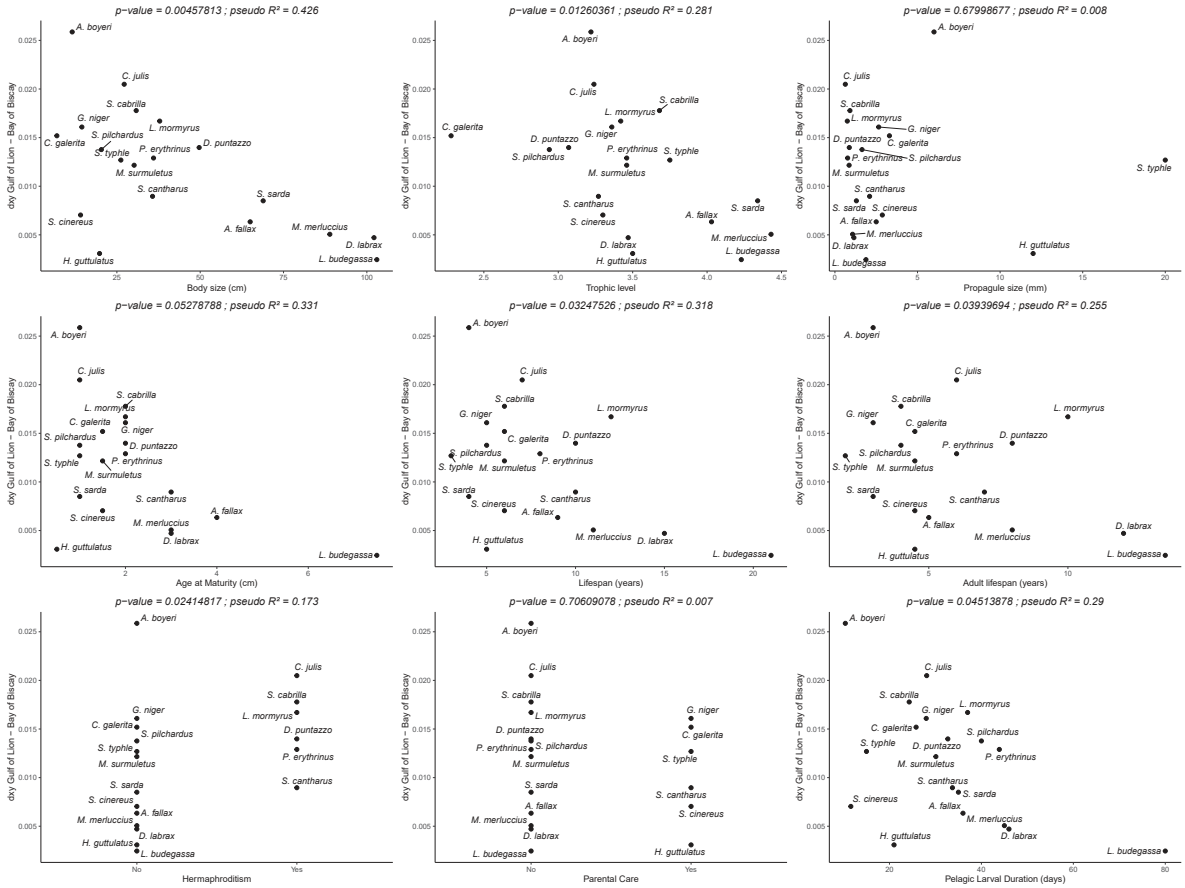



Figure S16: **Correlation between 9 life-history traits and $d_{XY}$ between outer populations for 19 species**. p-value and $R^2$ are shown above each plot.

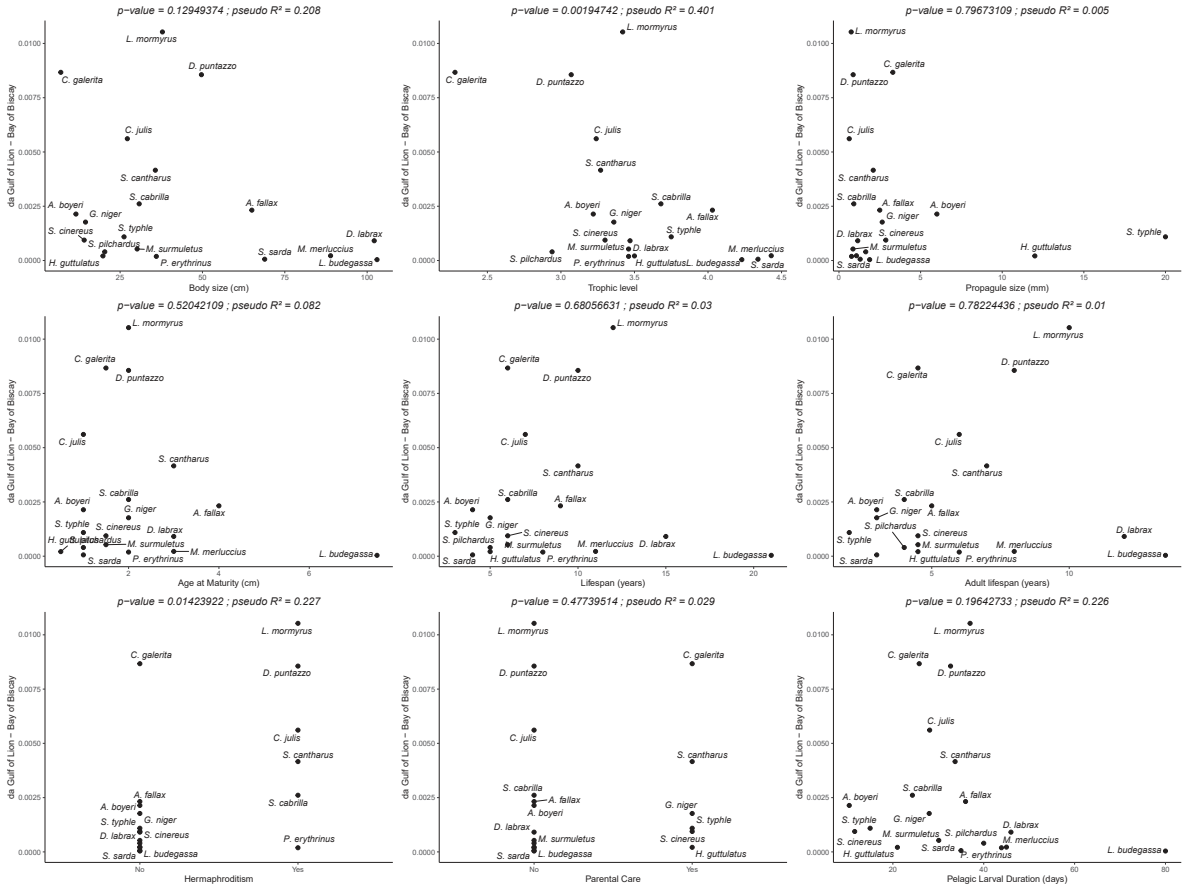



Figure S17: **Correlation between 9 life-history traits and $d_a$ between outer populations for 19 species**. p-value and $R^2$ are shown above each plot.

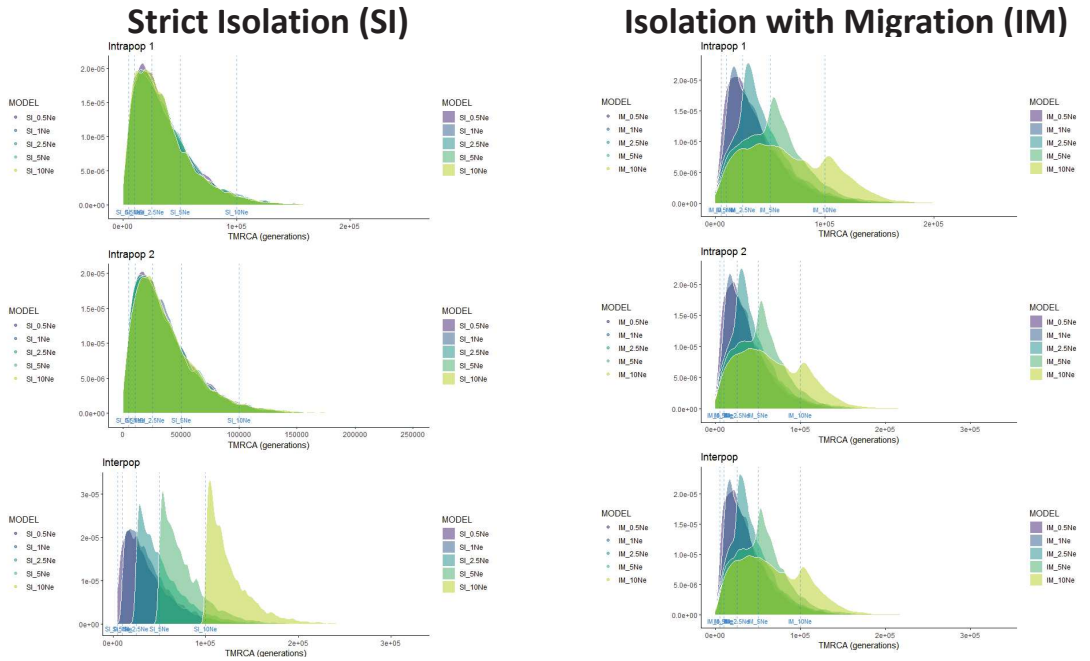



Figure S18: **Evaluation of the impact of demographic scenarios on TMRCA distributions of 2 populations**. Left: strict isolation, Right : Isolation with Migration. Top and middle panels represent within-population TMRCA distributions and bottom panel between population TMRCA distributions. Each different colors represent different time of ancestral split between populations of $0.5N_e$, $1N_e$, $2.5N_e$, $5N_e$ and $10N_e$.

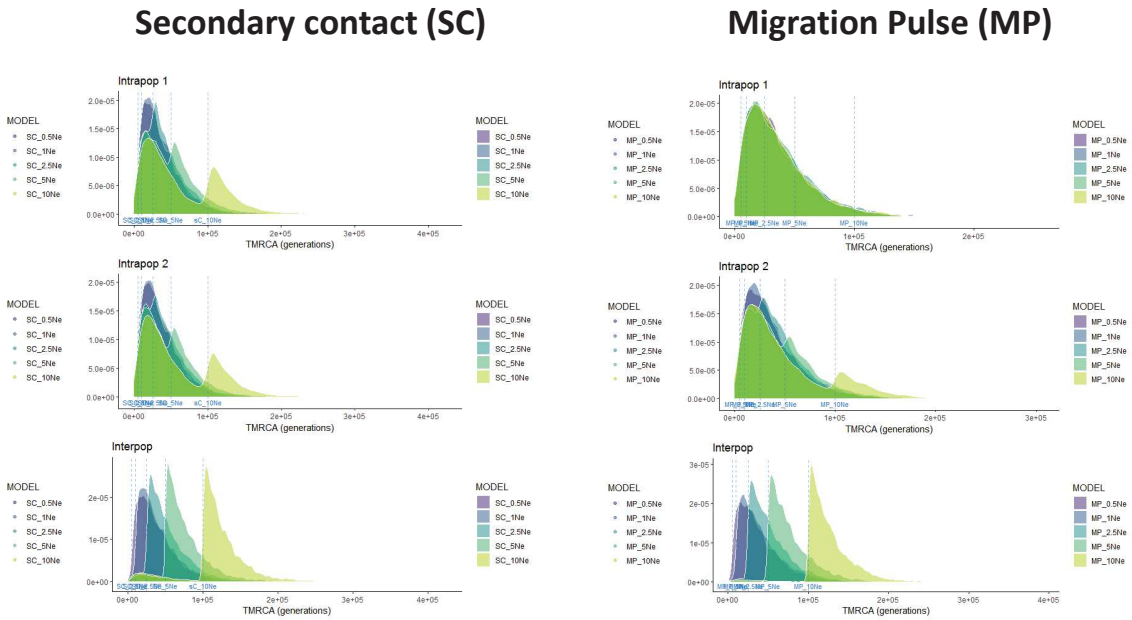



Figure S19: **Evaluation of the impact of demographic scenarios on TMRCA distributions of 2 populations**. Left: secondary contact, Right : migration pulse. Top and middle panels represent within-population TMRCA distributions and bottom panel between population TMRCA distributions. Each different colors represent different time of ancestral split between populations of $0.5N_e$, $1N_e$, $2.5N_e$, $5N_e$ and $10N_e$.

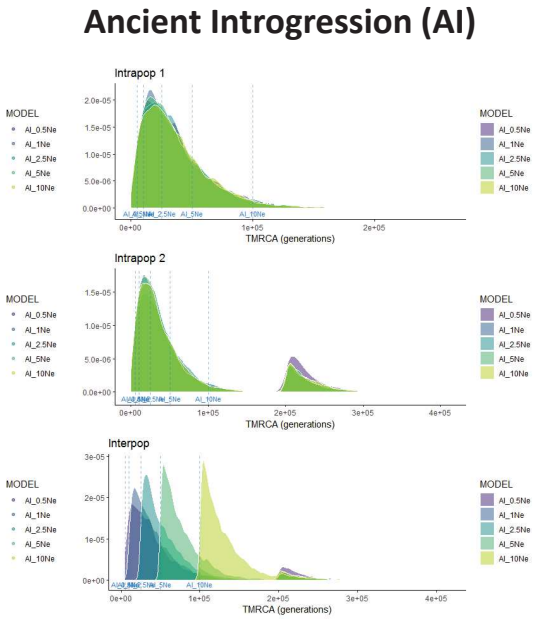



Figure S20: **Evaluation of the impact of demographic scenarios on TMRCA distributions of 2 populations**. Middle: ancient introgression. Top and middle panels represent within-population TMRCA distributions and bottom panel between population TMRCA distributions. Each different colors represent different time of ancestral introgression of $0.5N_e$, $1N_e$, $2.5N_e$, $5N_e$ and $10N_e$.

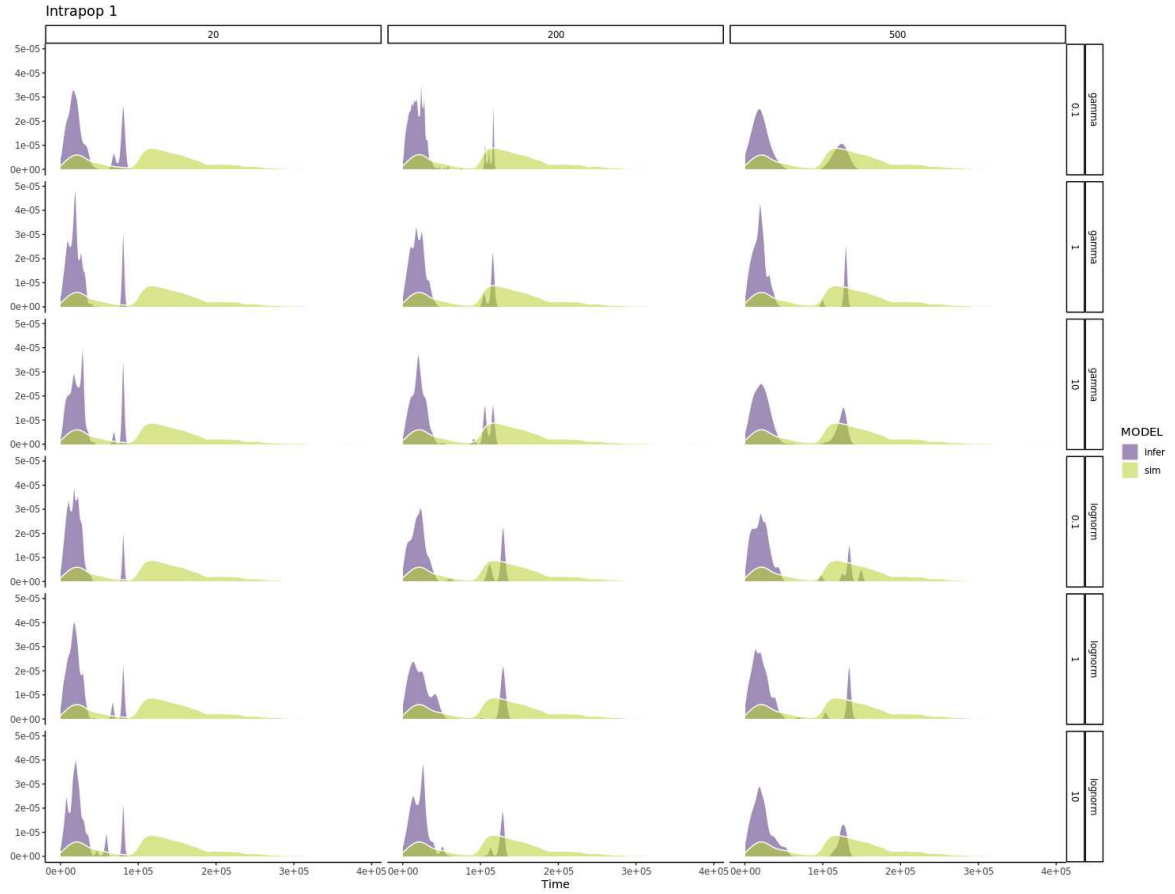



Figure S21: **Evaluation of `tsinfer` and `tsdate` performance within-population**. Comparison of simulated (yellow) and inferred (purple) TMRCA distributions from a 1Mb contiguous scaffolds in a secondary contact model for 3 number of sampled haplotypes (20, 200, 500), two prior distributions for the coalescent distributions (gamma and lognormal) and 3 different mismatch ratio (0.1, 1, 10).

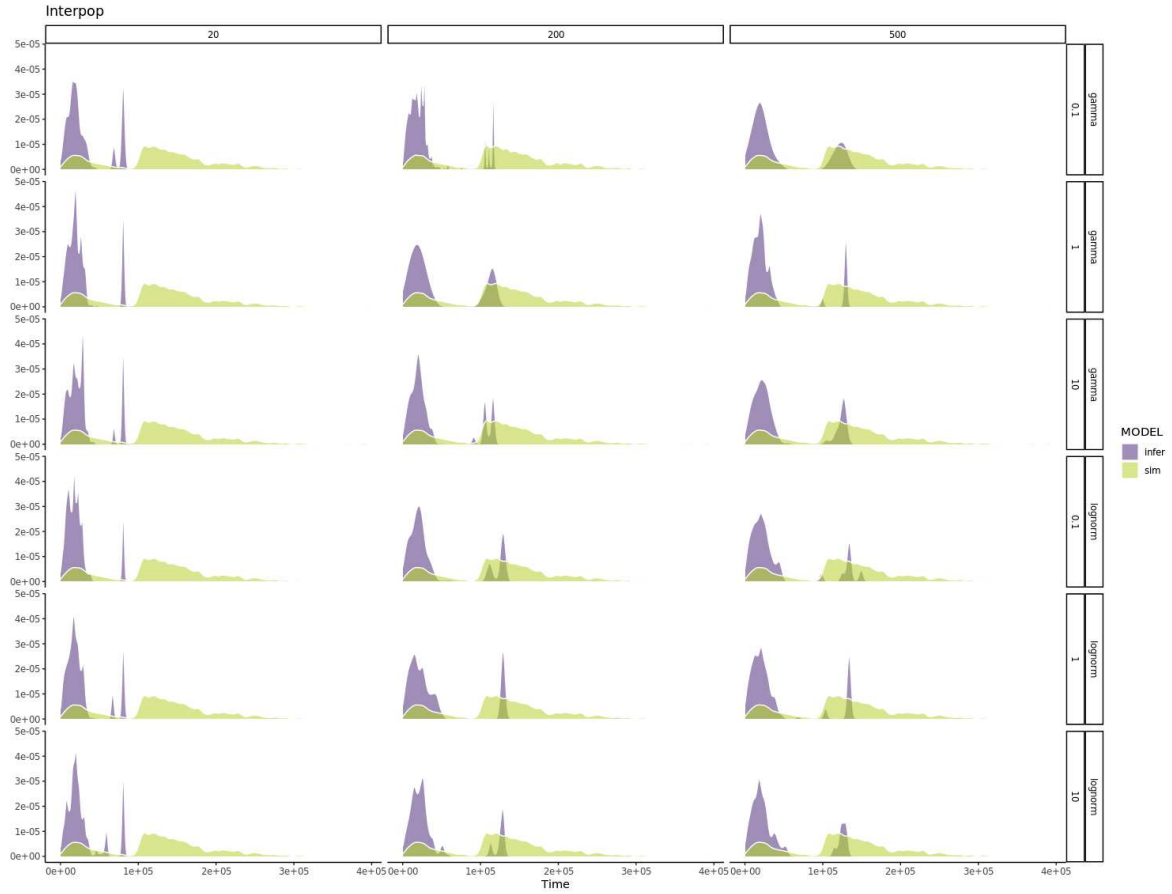



Figure S22: **Evaluation of `tsinfer` and `tsdate` performance between populations**. Comparison of simulated (yellow) and inferred (purple) TMRCA distributions from a 1Mb contiguous scaffolds in a secondary contact model for 3 number of sampled haplotypes (20, 200, 500), two prior distributions for the coalescent distributions (gamma and lognormal) and 3 different mismatch ratio (0.1, 1, 10).



Figure S24: **ABC parameter estimates of 18 species**. N1 = Mediterranean population size, N2 = Atlantic population size; Na = Ancestral population size, N1:Na = ratio of mediterranean to ancestral population size, N2/Na = ratio of atlantic to ancestral population size, Tam = time of stop of ancient migration, Tsc = time of secondary contact, Tsplit = time of ancestral split, Tam/Tsplit = time of stop of ancestral migration to ancestral split, Tsc/Tsplit = time of secondary contact to ancestral split, M12 = gene flow from Atlantic to Mediterranean populations ; M21 = gene flow from Mediterranean to Atlantic populations, m12 = migration rate from Atlantic to Mediterranean populations, m21 = migration rate from Mediterranean to Atlantic populations, shape_N_a and shape_N_b = $\alpha$ and $\beta$ parameters from the beta distribution of $N_e$, shape_M12_a and shape_M12_b = $\alpha$ and $\beta$ parameters from the beta distribution of $M_{12}$ and shape_M21_a and shape_M21_b = $\alpha$ and $\beta$ parameters from the beta distribution of $M_{21}$. Green and blue show comparison betwee inner and outer populations respectively.

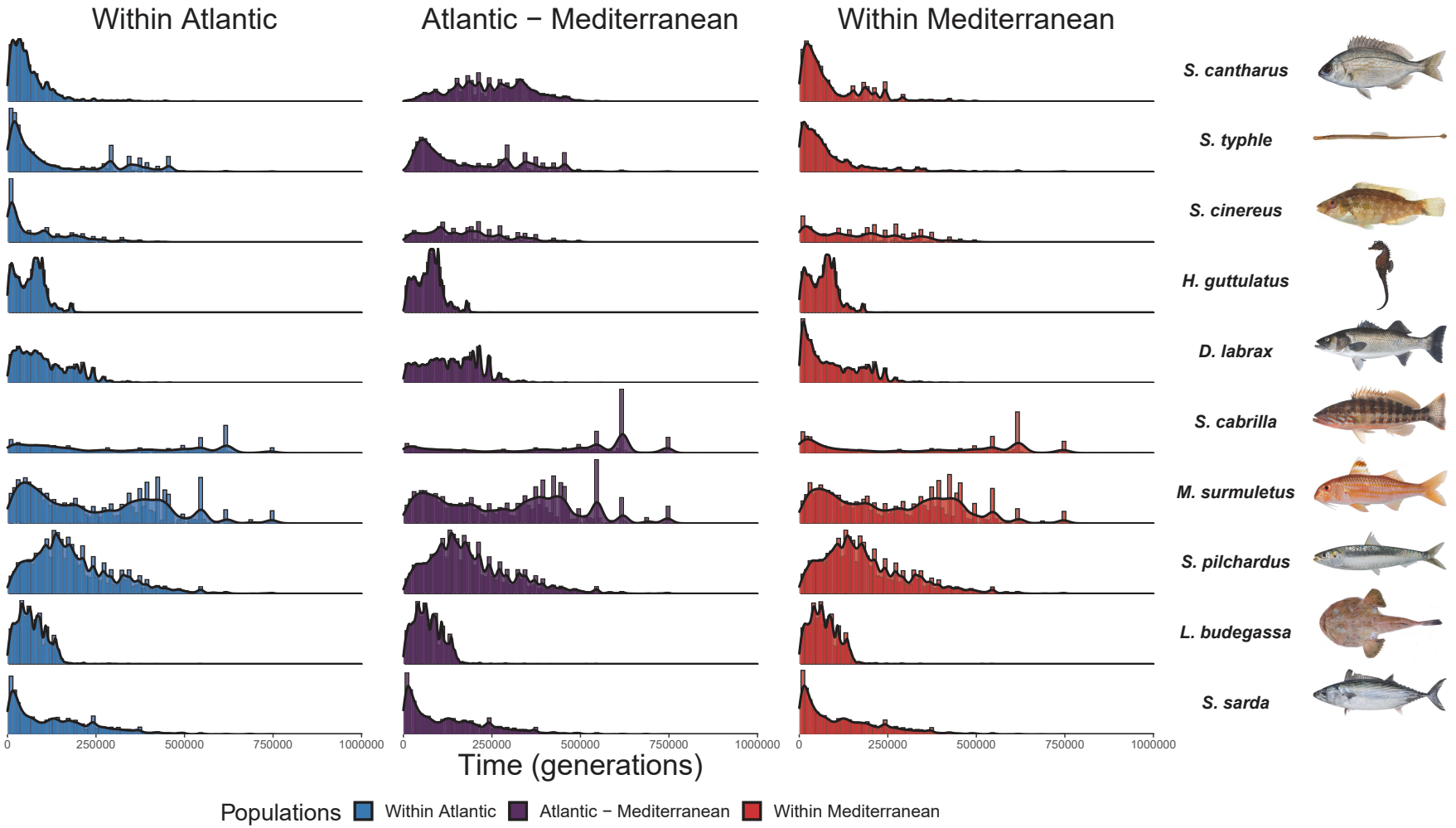



Figure S23: **Distribution of coalescent times within and between basins as in Fig.3 but not scaled by generation time**. The genome-wide tree sequence of non-recombining genomic segments was inferred with `tsinfer` and branch lengths were estimated with `tsdate` using individual haplotypes within contiguous scaffolds longer than 10kb. Each panel represents the genome-wide distribution of time to the most recent common ancestor (TMRCA) converted to years, using all haplotype pairs taken either within the Atlantic (left, blue), the Mediterranean (right, red) or between Atlantic and Mediterranean samples (middle, purple), for ten species in rows. Branch lengths are in the units of generation.

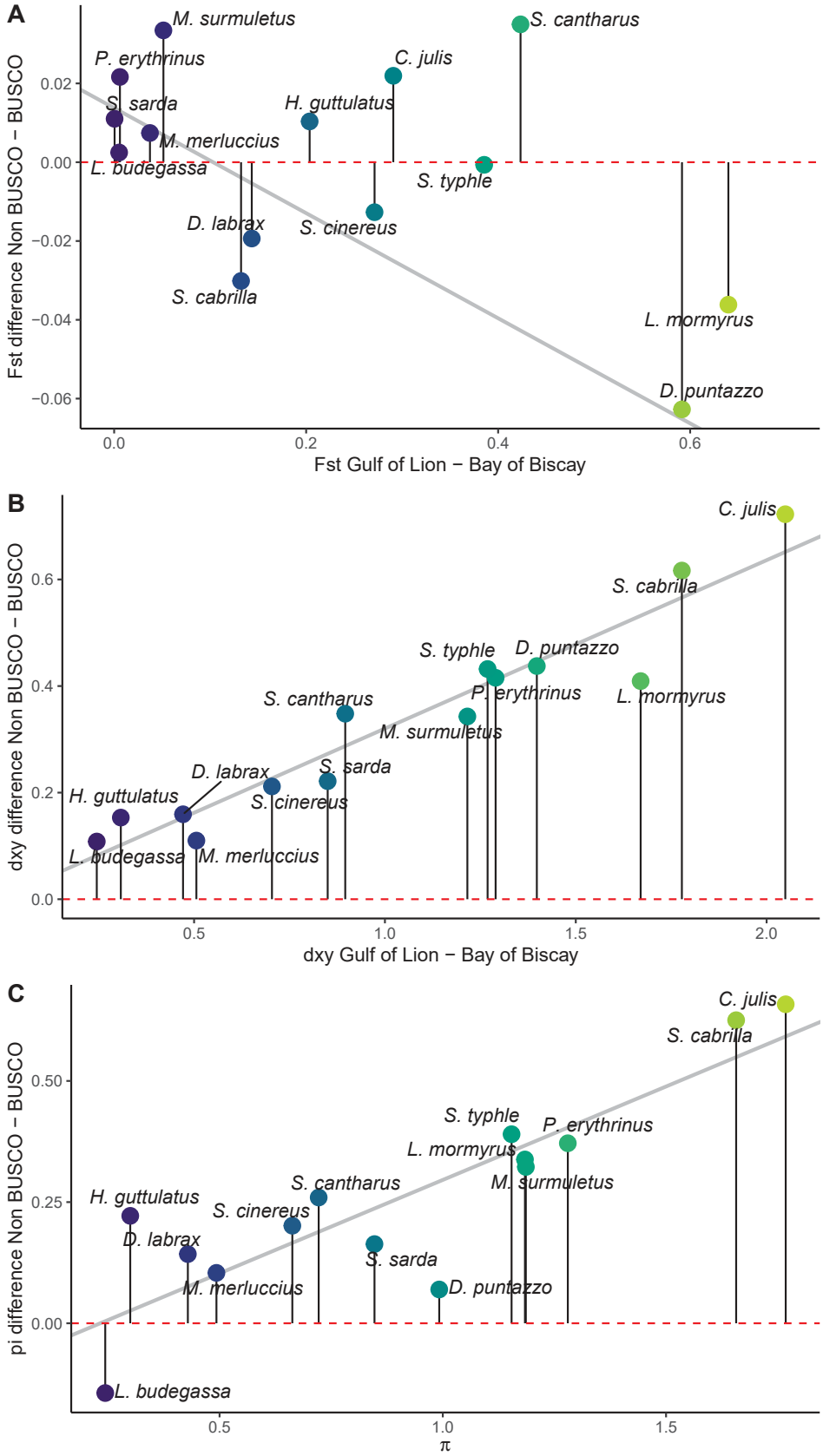



Figure S25: **Difference of $F_{ST}$ (A), $d_{XY}$ (B) and $\pi$ (C) between non-BUSCO and BUSCO genes for 14 species function of $F_{ST}$, $d_{XY}$ and $\pi$ respectively.** Non-BUSCO are the mean of 10kb windows located outside of BUSCO genes. BUSCO genes are highly conserved genes in the Actinoptertygian clade. An excess of each of these statistics of non-BUSCO regions corresponds to positive values. The red dot line represents $y = 0$ and the gray line the fitted linear model. Colors represent the gradient of each corresponding statistic.

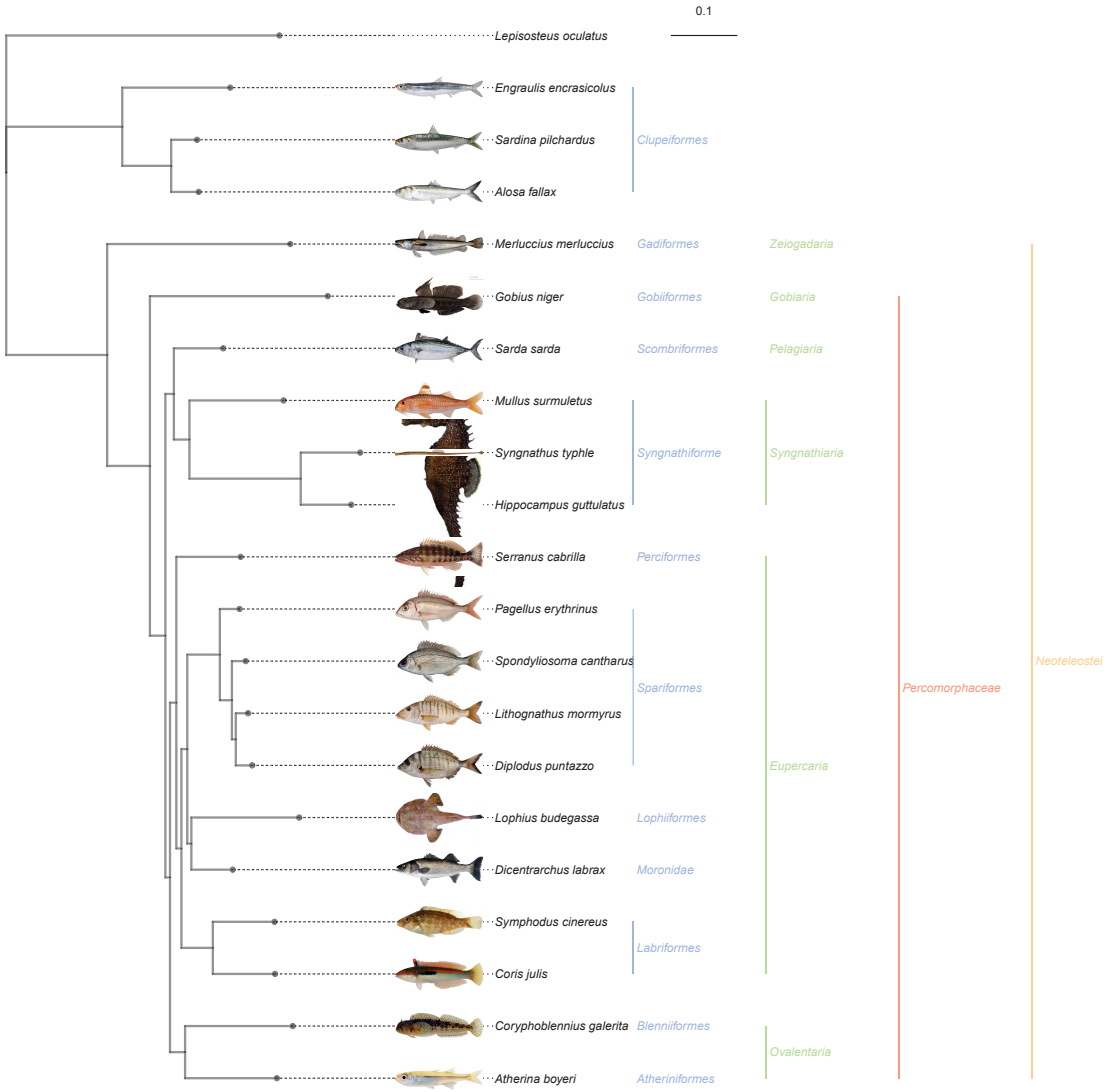



Figure S26: **Phylogeny of studied species**. 87 single-copy orthologs were found with `BUSCO` aligned with `mafft` and concatenated in one single fasta file. The phylogeny was inferred with this set of orthologs with `iqtree`, with a JTT+F+R4 protein evolution model. We chose *Lepisosteus oculatus* as an outgroup. The major clade is represented in color on the left of the tree.

# ANNEX OF CHAPTER 3

# Supplementary Material to Correlated landscapes of mitochondrial and divergence in marine fishes

## Contents

**Additional methods** **1**
MitoZ assembly . . . . . 1

**Additional tables** **2**

**Additional figures** **5**

## Additional methods

### MitoZ assembly

We first assembled individual whole-genome resequencing short reads with `MitoZ` running:

```
singularity exec \\
-B :/Data \\
-B :/Results \\
ok_mitoz/mitoz.simg \\
snakemake -s /workflow/Snakefile all \\
--configfile ok_mitoz/params.yml \\
--cores 12 || true
```

We reorder each mitochondrial scaffold to a species mitochondrial scaffolds reference:

```
python3 \\
Mitogenome_reorder.py \\
-f individual.fasta \\
-r reference.fasta
```

and align each scaffold using `mafft`:

```
/mafft \\
SPECIES_mitochondrial.fasta > \\
SPECIES_mitochondrial_align
```

We then trim corresponding alignment with `trimAl` and infer phylogenetic tree with `iqtree`

```
iqtree -s SPECIES_mitochondrial_align.fa -m MF -T 4 \\
iqtree -s SPECIES_mitochondrial_align.fa -m TN+F+I -T 4
```

# Additional tables

| Species | Length | Biallelic | Polyallelic | Gaps | GC content |
|---|---|---|---|---|---|
| *A. boyeri* | 17339 | 3219 | 275 | 870 | 0.497 |
| *A. fallax* | 16776 | 433 | 2 | 128 | 0.473 |
| *C. galerita* | 17789 | 1479 | 36 | 2127 | 0.433 |
| *C. julis* | 16898 | 1061 | 1 | 473 | 0.483 |
| *D. labrax* | 18143 | 693 | 11 | 893 | 0.462 |
| *D. puntazzo* | 16996 | 511 | 29 | 431 | 0.457 |
| *E. encrasicolus* | 16677 | 819 | 23 | 4 | 0.457 |
| *G. niger* | 16510 | 774 | 26 | 10 | 0.498 |
| *H. guttulatus* | 16534 | 136 | 1 | 12 | 0.382 |
| *L. budegassa* | 16577 | 201 | 2 | 53 | 0.480 |
| *L. mormyrus* | 16921 | 909 | 17 | 829 | 0.468 |
| *M. merluccius* | 17844 | 417 | 23 | 1117 | 0.483 |
| *M. surmuletus* | 16582 | 407 | 9 | 11 | 0.483 |
| *P. erythrinus* | 16671 | 396 | 2 | 7 | 0.452 |
| *S. cabrilla* | 16516 | 568 | 15 | 4 | 0.449 |
| *S. cantharus* | 17214 | 856 | 8 | 568 | 0.459 |
| *S. cinereus* | 16490 | 354 | 7 | 5 | 0.498 |
| *S. pilchardus* | 17526 | 583 | 10 | 34 | 0.498 |
| *S. sarda* | 16960 | 695 | 12 | 58 | 0.458 |
| *S. typhle* | 17160 | 547 | 26 | 1394 | 0.443 |

Table S1: Characteristics of whole-mitochondrial genome alignment for 20 species including length (in base pairs), biallelic sites, polyallelic sites, gaps in the alignment (position where at least one individual have gaps) and GC content.

| Species | $\pi_{tot}$ | $\pi_{li}$ | $\pi_{mu}$ | $\pi_{fa}$ | $\pi_{ga}$ |
|---|---|---|---|---|---|
| *A. boyeri* | 8.053 | 8.320 | 0.261 | 8.752 | 0.071 |
| *A. fallax* | 0.795 | 0.083 | - | 1.016 | 0.816 |
| *C. galerita* | 3.815 | 0.568 | 0.217 | 0.103 | 0.151 |
| *C. julis* | 3.259 | 0.055 | 0.086 | 0.063 | 0.058 |
| *D. labrax* | 1.494 | 0.099 | 0.100 | 1.715 | 0.184 |
| *D. puntazzo* | 0.721 | 0.434 | 0.360 | 1.135 | 0.753 |
| *E. encrasicolus* | 1.489 | 1.790 | 0.423 | 1.727 | 1.627 |
| *G. niger* | 0.748 | 0.801 | 1.182 | 0.307 | 0.260 |
| *H. guttulatus* | 0.164 | 0.175 | 0.177 | 0.000 | 0.064 |
| *L. budegassa* | 0.232 | 0.324 | 0.496 | 0.108 | 0.087 |
| *L. mormyrus* | 2.405 | 2.081 | 0.365 | 0.186 | 0.093 |
| *M. merluccius* | 0.366 | 0.393 | 0.408 | 0.364 | 0.375 |
| *M. surmuletus* | 0.547 | 0.553 | 0.690 | 0.519 | 0.457 |
| *P. erythrinus* | 0.540 | 0.482 | 0.532 | 0.179 | 0.770 |
| *S. cabrilla* | 1.130 | 1.207 | 0.993 | 1.249 | 0.890 |
| *S. cantharus* | 2.365 | 0.096 | 0.118 | 0.171 | 0.208 |
| *S. cinereus* | 0.384 | 0.460 | 0.341 | 0.280 | 0.053 |
| *S. pilchardus* | 0.519 | 0.577 | 0.436 | 0.576 | 0.506 |
| *S. sarda* | 1.757 | 2.163 | 1.522 | 1.520 | 2.108 |
| *S. typhle* | 0.926 | 0.304 | 0.622 | 0.478 | 0.152 |

Table S2: Genetic diversity of 20 species whole-mitochondrial genome including total genetic diversity $\pi_{total}$, within out Mediterranean ($\pi_{li}$), within in Mediterranean ($\pi_{mu}$), within in Atlantic ($\pi_{fa}$) and within out Atlantic ($\pi_{ga}$) populations.

| Species | $F_{ST,li,mu}$ | $F_{ST,li,fa}$ | $F_{ST,li,ga}$ | $F_{ST,mu,fa}$ | $F_{ST,mu,ga}$ | $F_{ST,fa,ga}$ |
|---|---|---|---|---|---|---|
| *A. boyeri* | 0.496 | 0.068 | 0.494 | 0.680 | 0.700 | 0.685 |
| *A. fallax* | 0.000 | 0.348 | 0.396 | 0.000 | 0.000 | -0.031 |
| *C. galerita* | 0.066 | 0.820 | 0.944 | 0.962 | 0.955 | 0.045 |
| *C. julis* | 0.037 | 0.990 | 0.991 | 0.987 | 0.988 | 0.058 |
| *D. labrax* | 0.306 | 0.442 | 0.940 | 0.426 | 0.861 | 0.252 |
| *D. puntazzo* | 0.096 | 0.101 | 0.056 | 0.132 | -0.043 | -0.044 |
| *E. encrasicolus* | 0.353 | -0.109 | -0.093 | 0.158 | 0.180 | -0.127 |
| *G. niger* | 0.017 | 0.271 | 0.279 | 0.215 | 0.221 | 0.021 |
| *H. guttulatus* | 0.015 | 0.638 | 0.310 | 0.569 | 0.067 | 0.804 |
| *L. budegassa* | -0.191 | 0.055 | 0.139 | 0.027 | 0.061 | 0.135 |
| *L. mormyrus* | -0.015 | 0.683 | 0.698 | 0.935 | 0.949 | -0.050 |
| *M. merluccius* | -0.071 | -0.066 | -0.061 | -0.037 | -0.075 | -0.093 |
| *M. surmuletus* | -0.033 | -0.063 | -0.097 | -0.073 | 0.076 | 0.050 |
| *P. erythrinus* | -0.012 | 0.139 | 0.082 | -0.005 | 0.048 | 0.333 |
| *S. cabrilla* | 0.070 | -0.094 | -0.087 | -0.069 | 0.316 | 0.079 |
| *S. cantharus* | -0.011 | 0.966 | 0.966 | 0.958 | 0.958 | 0.312 |
| *S. cinereus* | 0.198 | 0.208 | 0.462 | 0.218 | 0.514 | 0.245 |
| *S. pilchardus* | 0.045 | -0.043 | -0.006 | -0.003 | -0.035 | -0.033 |
| *S. sarda* | -0.114 | -0.133 | -0.135 | -0.215 | 0.104 | 0.089 |
| *S. typhle* | 0.645 | 0.465 | 0.642 | 0.572 | 0.698 | 0.554 |

Table S3: Hudson's genetic differentiation $F_{ST}$ of 20 species whole-mitochondrial genome between Mediterranean ($F_{ST,li,mu}$) and Atlantic ($F_{ST,fa,ga}$) populations, between in ($F_{ST,mu,fa}$) and out ($F_{ST,li,ga}$) populations and between Mediterranean in and atlantic out ($F_{ST,mu,ga}$) and between Mediterranean out and Atlantic in populations ($F_{ST,li,fa}$).

| Species | $d_{XY,li,mu}$ | $d_{XY,li,fa}$ | $d_{XY,li,ga}$ | $d_{XY,mu,fa}$ | $d_{XY,mu,ga}$ | $d_{XY,fa,ga}$ |
|---|---|---|---|---|---|---|
| *A. boyeri* | 8.485 | 9.124 | 8.273 | 14.017 | 0.552 | 0.140 |
| *A. fallax* | - | 0.842 | 0.742 | - | - | 0.009 |
| *C. galerita* | 0.385 | 7.128 | 7.099 | 7.017 | 7.025 | 0.001 |
| *C. julis* | 0.073 | 6.149 | 6.142 | 6.124 | 6.117 | 0.001 |
| *D. labrax* | 0.128 | 1.702 | 2.879 | 1.677 | 2.877 | 0.013 |
| *D. puntazzo* | 0.415 | 0.912 | 0.661 | 0.851 | 0.533 | 0.009 |
| *E. encrasicolus* | 1.709 | 1.586 | 1.563 | 1.276 | 1.250 | 0.015 |
| *G. niger* | 1.009 | 0.760 | 0.736 | 0.948 | 0.926 | 0.003 |
| *H. guttulatus* | 0.179 | 0.242 | 0.174 | 0.206 | 0.129 | 0.002 |
| *L. budegassa* | 0.340 | 0.229 | 0.239 | 0.310 | 0.310 | 0.001 |
| *L. mormyrus* | 1.207 | 3.537 | 3.562 | 4.347 | 4.396 | 0.001 |
| *M. merluccius* | 0.386 | 0.356 | 0.363 | 0.368 | 0.362 | 0.003 |
| *M. surmuletus* | 0.602 | 0.505 | 0.462 | 0.564 | 0.622 | 0.005 |
| *P. erythrinus* | 0.501 | 0.384 | 0.683 | 0.353 | 0.683 | 0.007 |
| *S. cabrilla* | 1.183 | 1.122 | 0.965 | 1.049 | 1.376 | 0.012 |
| *S. cantharus* | 0.106 | 4.309 | 4.359 | 4.312 | 4.362 | 0.003 |
| *S. cinereus* | 0.499 | 0.467 | 0.477 | 0.397 | 0.405 | 0.002 |
| *S. pilchardus* | 0.530 | 0.555 | 0.539 | 0.504 | 0.455 | 0.005 |
| *S. sarda* | 1.654 | 1.625 | 1.882 | 1.252 | 2.027 | 0.020 |
| *S. typhle* | 1.231 | 0.711 | 0.738 | 1.224 | 1.271 | 0.007 |

Table S4: Net genetic divergence $d_{XY}$ of 20 species whole-mitochondrial genome between Mediterranean ($d_{XY,li,mu}$) and Atlantic ($d_{XY,fa,ga}$) populations, between in ($d_{XY,mu,fa}$) and out ($d_{XY,li,ga}$) populations and between Mediterranean in and Atlantic out ($d_{XY,mu,ga}$) and between Mediterranean out and Atlantic in populations ($d_{XY,li,fa}$).

| Species | $d_{a,li,mu}$ | $d_{a,li,fa}$ | $d_{a,li,ga}$ | $d_{a,mu,fa}$ | $d_{a,mu,ga}$ | $d_{a,fa,ga}$ |
|---|---|---|---|---|---|---|
| *A. boyeri* | 4.195 | 0.589 | 4.078 | 9.511 | 0.386 | 0.052 |
| *A. fallax* | - | 0.293 | 0.293 | - | - | -0.001 |
| *C. galerita* | -0.008 | 6.792 | 6.739 | 6.856 | 6.841 | 0.000 |
| *C. julis* | 0.003 | 6.091 | 6.086 | 6.050 | 6.045 | 0.000 |
| *D. labrax* | 0.029 | 0.795 | 2.738 | 0.770 | 2.735 | -0.005 |
| *D. puntazzo* | 0.018 | 0.128 | 0.068 | 0.104 | -0.023 | -0.002 |
| *E. encrasicolus* | 0.603 | -0.172 | -0.146 | 0.201 | 0.225 | -0.002 |
| *G. niger* | 0.017 | 0.206 | 0.205 | 0.204 | 0.204 | 0.000 |
| *H. guttulatus* | 0.003 | 0.154 | 0.054 | 0.117 | 0.009 | 0.002 |
| *L. budegassa* | -0.070 | 0.012 | 0.033 | 0.008 | 0.019 | 0.000 |
| *L. mormyrus* | -0.016 | 2.403 | 2.475 | 4.072 | 4.167 | -0.001 |
| *M. merluccius* | -0.015 | -0.022 | -0.022 | -0.019 | -0.030 | 0.000 |
| *M. surmuletus* | -0.019 | -0.031 | -0.043 | -0.041 | 0.049 | 0.000 |
| *P. erythrinus* | -0.006 | 0.053 | 0.056 | -0.002 | 0.033 | 0.005 |
| *S. cabrilla* | 0.083 | -0.106 | -0.083 | -0.072 | 0.435 | -0.001 |
| *S. cantharus* | -0.001 | 4.175 | 4.207 | 4.167 | 4.199 | 0.001 |
| *S. cinereus* | 0.099 | 0.097 | 0.220 | 0.087 | 0.208 | -0.001 |
| *S. pilchardus* | 0.024 | -0.022 | -0.002 | -0.002 | -0.016 | -0.001 |
| *S. sarda* | -0.188 | -0.216 | -0.254 | -0.269 | 0.212 | 0.005 |
| *S. typhle* | 0.768 | 0.320 | 0.510 | 0.674 | 0.884 | 0.003 |

Table S5: Absolute genetic divergence $d_a$ of 20 species whole-mitochondrial genome between Mediterranean ($d_{a,li,mu}$) and Atlantic ($d_{a,fa,ga}$) populations, between in ($d_{a,mu,fa}$) and out ($d_{a,li,ga}$) populations and between Mediterranean in and Atlantic out ($d_{a,mu,ga}$) and between Mediterranean out and Atlantic in populations ($d_{a,li,fa}$).

# Additional figures

| | Aboye | Afall | Cgale | Cjuli | Dlabr | Dpunt | Eencr | Gnige | Hgutt | Lbude | Lmorm | Mmerl | Msurm | Peryt | Scabr | Scant | Scine | Spilc | Ssard | Styph |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Aboye | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
| Afall | **$1.17e^{-5}$** | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
| Cgale | **$2.09e^{-4}$** | **$1.66e^{-3}$** | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
| Cjuli | **$4.42e^{-5}$** | **$1.74e^{-5}$** | **$1.20e^{-5}$** | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
| Dlabr | **$4.86e^{-4}$** | **$1.48e^{-6}$** | **$2.12e^{-4}$** | **$2.33e^{-5}$** | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
| Dpunt | **$2.61e^{-3}$** | **$2.45e^{-3}$** | **$1.96e^{-4}$** | **$3.46e^{-3}$** | **$2.97e^{-3}$** | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
| Eencr | **$3.60e^{-5}$** | **$8.59e^{-8}$** | **$2.12e^{-7}$** | **$2.23e^{-10}$** | **$5.01e^{-9}$** | **$1.86e^{-4}$** | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
| Gnige | **$7.19e^{-5}$** | **$7.18e^{-7}$** | **$1.87e^{-6}$** | **$1.15e^{-7}$** | **$2.86e^{-8}$** | **$8.79e^{-7}$** | **$1.89e^{-9}$** | - | - | - | - | - | - | - | - | - | - | - | - | - |
| Hgutt | 0.926 | 0.257 | 0.124 | 0.076 | 0.233 | 0.412 | 0.145 | 0.675 | - | - | - | - | - | - | - | - | - | - | - | - |
| Lbude | **$2.06e^{-4}$** | **$1.14e^{-4}$** | **$1.89e^{-3}$** | **$8.85e^{-4}$** | **$3.39e^{-4}$** | **$4.30e^{-7}$** | **$2.41e^{-5}$** | **$5.24e^{-5}$** | 0.129 | - | - | - | - | - | - | - | - | - | - | - |
| Lmorm | **$1.15e^{-4}$** | **$7.69e^{-5}$** | **$2.16e^{-7}$** | **$8.67e^{-8}$** | **$3.24e^{-7}$** | **$4.80e^{-5}$** | **$2.61e^{-10}$** | **$5.63e^{-7}$** | 0.107 | **$2.19e^{-3}$** | - | - | - | - | - | - | - | - | - | - |
| Mmerl | 0.794 | 0.273 | 0.349 | 0.293 | 0.369 | 0.95 | 0.591 | 0.115 | 0.764 | 0.769 | 0.061 | - | - | - | - | - | - | - | - | - |
| Msurm | **$3.07e^{-3}$** | **$1.71e^{-7}$** | **$5.30e^{-7}$** | **$9.55e^{-9}$** | **$4.35e^{-8}$** | **$4.06e^{-5}$** | **$1.70e^{-11}$** | **$1.04e^{-8}$** | 0.195 | **$9.45e^{-5}$** | **$4.61e^{-11}$** | 0.442 | - | - | - | - | - | - | - | - |
| Peryt | **$1.52e^{-4}$** | **$4.90e^{-6}$** | **$9.05e^{-9}$** | **$1.26e^{-8}$** | **$6.52e^{-7}$** | **$3.99e^{-6}$** | **$1.09e^{-12}$** | **$5.74e^{-9}$** | 0.210 | **$5.23e^{-5}$** | **$1.70e^{-11}$** | 0.256 | **$6.59e^{-12}$** | - | - | - | - | - | - | - |
| Scabr | **$4.84e^{-3}$** | **$1.44e^{-3}$** | **$3.44e^{-5}$** | **$5.79e^{-6}$** | **$4.18e^{-5}$** | **$7.99e^{-3}$** | **$4.37e^{-7}$** | **$2.53e^{-4}$** | 0.466 | **$9.06e^{-3}$** | **$6.28e^{-7}$** | 0.381 | **$2.50e^{-6}$** | **$7.26e^{-9}$** | - | - | - | - | - | - |
| Scant | **$1.83e^{-3}$** | 0.184 | 0.093 | **0.035** | **0.017** | 0.380 | **$6.71e^{-3}$** | **0.016** | 0.772 | 0.462 | **$9.37e^{-4}$** | **0.011** | 0.11 | **0.033** | **0.027** | - | - | - | - | - |
| Scine | 0.023 | **0.0165** | 0.116 | 0.105 | 0.121 | 0.183 | **0.026** | **0.034** | 0.794 | 0.078 | 0.076 | **0.017** | **0.016** | 0.093 | 0.41 | **$2.07e^{-2}$** | - | - | - | - |
| Spilc | **$1.59e^{-3}$** | **$4.22e^{-3}$** | **$1.12e^{-3}$** | **$4.27e^{-3}$** | **$4.11e^{-4}$** | **$1.81e^{-6}$** | **$4.96e^{-6}$** | **$7.96e^{-6}$** | 0.480 | **$1.55e^{-4}$** | **$1.57e^{-4}$** | 0.599 | **$5.34e^{-5}$** | **$2.84e^{-4}$** | **0.012** | **$1.70e^{-3}$** | **$4.08e^{-4}$** | - | - | - |
| Ssard | **$1.40e^{-4}$** | **$5.39e^{-6}$** | **$2.20e^{-7}$** | **$1.78e^{-9}$** | **$1.41e^{-8}$** | **$6.73e^{-4}$** | **$2.36e^{-13}$** | **$2.34e^{-8}$** | 0.472 | **$3.23e^{-4}$** | **$8.03e^{-11}$** | 0.139 | **$4.02e^{-10}$** | **$4.09e^{-12}$** | **$5.11e^{-12}$** | **0.013** | 0.087 | **$2.45e^{-4}$** | - | - |
| Styph | 0.313 | 0.443 | 0.687 | **0.0125** | 0.453 | 0.66 | 0.774 | 0.232 | **$7.96^{-5}$** | 0.624 | 0.597 | 0.291 | 0.727 | 0.864 | 0.833 | 0.518 | 0.255 | 0.552 | 0.624 | - |

Table S6: **Correlation between pairwise species total genetic diversity ($\pi$)** - Each entry shows the $p-value$ of the linear regression modeling of gene genetic diversity between corresponding species on each row and column. Significant correlations at $p-value < 0.05$ are shown in bold.

| | Aboye | Afall | Cgale | Cjuli | Dlabr | Dpunt | Eencr | Gnige | Hgutt | Lbude | Lmorm | Mmerl | Msurm | Peryt | Scabr | Scant | Scine | Spilc | Ssard | Styph |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Aboye | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
| Afall | $9.82e-01$ | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
| Cgale | $4.23e-01$ | $1.07e-01$ | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
| Cjuli | $4.60e-01$ | $3.96e-01$ | $6.92e-01$ | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
| Dlabr | $7.99e-01$ | $\mathbf{1.40e-05}$ | $2.74e-01$ | $5.34e-01$ | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
| Dpunt | $3.06e-01$ | $7.84e-01$ | $5.77e-01$ | $7.01e-01$ | $5.04e-02$ | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
| Eencr | $5.06e-01$ | $\mathbf{3.68e-02}$ | $\mathbf{2.98e-02}$ | $7.56e-01$ | $1.47e-01$ | $1.51e-01$ | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
| Gnige | $9.97e-01$ | $\mathbf{7.66e-03}$ | $6.02e-01$ | $4.27e-01$ | $\mathbf{5.54e-04}$ | $8.61e-01$ | $6.65e-01$ | - | - | - | - | - | - | - | - | - | - | - | - | - |
| Hgutt | $5.35e-01$ | $\mathbf{1.32e-02}$ | $3.52e-01$ | $5.13e-01$ | $\mathbf{5.84e-03}$ | $5.96e-01$ | $1.32e-01$ | $\mathbf{3.83e-04}$ | - | - | - | - | - | - | - | - | - | - | - | - |
| Lbude | $5.47e-01$ | $6.37e-02$ | $3.78e-01$ | $6.03e-01$ | $\mathbf{1.21e-02}$ | $\mathbf{1.96e-04}$ | $6.47e-02$ | $\mathbf{9.60e-03}$ | $2.29e-01$ | - | - | - | - | - | - | - | - | - | - | - |
| Lmorm | $3.43e-01$ | $\mathbf{1.17e-03}$ | $7.18e-01$ | $1.38e-01$ | $\mathbf{6.56e-05}$ | $1.01e-01$ | $3.79e-01$ | $1.35e-01$ | $\mathbf{5.80e-03}$ | $\mathbf{3.78e-02}$ | - | - | - | - | - | - | - | - | - | - |
| Mmerl | $8.62e-01$ | $1.21e-01$ | $6.48e-01$ | $8.04e-01$ | $6.79e-01$ | $8.72e-01$ | $9.54e-01$ | $4.56e-01$ | $7.78e-01$ | $7.91e-01$ | $6.28e-01$ | - | - | - | - | - | - | - | - | - |
| Msurm | $6.58e-01$ | $\mathbf{2.26e-03}$ | $9.65e-02$ | $4.23e-01$ | $\mathbf{4.35e-06}$ | $7.89e-01$ | $\mathbf{1.78e-02}$ | $\mathbf{2.92e-03}$ | $\mathbf{6.61e-03}$ | $\mathbf{2.97e-02}$ | $\mathbf{2.22e-04}$ | $1.37e-01$ | - | - | - | - | - | - | - | - |
| Peryt | $6.19e-01$ | $\mathbf{3.79e-05}$ | $5.92e-01$ | $4.02e-01$ | $\mathbf{5.45e-05}$ | $8.50e-01$ | $6.46e-01$ | $\mathbf{5.02e-03}$ | $\mathbf{3.64e-03}$ | $\mathbf{4.43e-02}$ | $\mathbf{1.07e-03}$ | $9.75e-01$ | $\mathbf{3.22e-05}$ | - | - | - | - | - | - | - |
| Scabr | $1.73e-01$ | $\mathbf{4.46e-05}$ | $8.32e-02$ | $3.63e-01$ | $\mathbf{9.62e-04}$ | $9.21e-01$ | $8.69e-02$ | $\mathbf{3.21e-02}$ | $9.86e-02$ | $1.20e-01$ | $\mathbf{1.31e-02}$ | $6.58e-01$ | $\mathbf{4.47e-02}$ | $2.94e-01$ | - | - | - | - | - | - |
| Scant | $5.02e-02$ | $9.75e-01$ | $3.22e-01$ | $4.42e-01$ | $6.23e-02$ | $8.67e-01$ | $5.52e-01$ | $2.69e-01$ | $3.19e-01$ | $2.41e-01$ | $3.02e-01$ | $3.90e-01$ | $2.02e-01$ | $8.69e-01$ | $1.78e-01$ | - | - | - | - | - |
| Scine | $\mathbf{1.58e-02}$ | $8.69e-02$ | $8.81e-01$ | $4.58e-01$ | $4.29e-01$ | $7.39e-01$ | $4.41e-01$ | $1.38e-01$ | $2.24e-01$ | $7.74e-01$ | $5.89e-01$ | $\mathbf{7.85e-03}$ | $1.54e-01$ | $1.50e-01$ | $7.49e-01$ | $2.52e-01$ | - | - | - | - |
| Spilc | $8.79e-02$ | $3.13e-01$ | $7.70e-01$ | $9.66e-01$ | $1.37e-01$ | $7.89e-01$ | $5.61e-01$ | $4.25e-01$ | $5.33e-01$ | $9.37e-01$ | $1.23e-01$ | $7.72e-01$ | $\mathbf{4.46e-02}$ | $2.96e-01$ | $5.16e-01$ | $5.87e-01$ | $5.17e-01$ | - | - | - |
| Ssard | $8.92e-01$ | $\mathbf{1.55e-05}$ | $3.01e-01$ | $2.88e-01$ | $\mathbf{1.71e-07}$ | $\mathbf{6.50e-03}$ | $\mathbf{5.42e-03}$ | $\mathbf{8.92e-04}$ | $\mathbf{1.13e-03}$ | $\mathbf{5.71e-04}$ | $\mathbf{1.43e-04}$ | $1.23e-01$ | $\mathbf{1.06e-06}$ | $\mathbf{7.31e-05}$ | $\mathbf{3.25e-03}$ | $1.36e-01$ | $1.96e-01$ | $1.43e-01$ | - | - |
| Styph | $4.90e-01$ | $\mathbf{2.97e-03}$ | $6.67e-01$ | $1.92e-01$ | $\mathbf{3.78e-05}$ | $4.07e-02$ | $6.63e-02$ | $1.67e-01$ | $\mathbf{1.31e-02}$ | $6.10e-02$ | $\mathbf{2.87e-04}$ | $7.59e-01$ | $\mathbf{6.52e-04}$ | $\mathbf{1.71e-03}$ | $\mathbf{4.84e-02}$ | $6.18e-01$ | $6.74e-01$ | $2.99e-01$ | $\mathbf{1.68e-06}$ | - |

Table S7: **Correlation between pairwise species Hudson's** $F_{ST}$ - Each entry shows the $p-value$ of the linear regression modeling of gene genetic diversity between corresponding species on each row and column. Significant correlations at $p-value < 0.05$ are shown in bold.

| | Aboye | Afall | Cgale | Cjuli | Dlabr | Dpunt | Eencr | Gnige | Hgutt | Lbude | Lmorm | Mmerl | Msurm | Peryt | Scabr | Scant | Scine | Spilc | Ssard | Styph |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Aboye | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
| Afall | **$3.18e-05$** | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
| Cgale | **$2.19e-04$** | **$3.56e-05$** | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
| Cjuli | **$1.02e-03$** | **$4.38e-08$** | **$2.47e-06$** | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
| Dlabr | **$2.19e-03$** | **$1.74e-08$** | **$1.27e-04$** | **$1.12e-05$** | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
| Dpunt | **$2.43e-02$** | **$7.77e-04$** | **$2.09e-03$** | **$2.78e-02$** | **$1.63e-02$** | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
| Eencr | **$3.24e-04$** | **$2.12e-09$** | **$4.53e-07$** | **$3.69e-09$** | **$2.41e-09$** | **$3.74e-03$** | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
| Gnige | **$2.67e-04$** | **$7.76e-06$** | **$5.16e-05$** | **$5.00e-05$** | **$5.25e-07$** | **$5.10e-05$** | $1.29e-05$ | - | - | - | - | - | - | - | - | - | - | - | - | - |
| Hgutt | $6.89e-01$ | $8.55e-01$ | $6.91e-01$ | $5.80e-01$ | $9.77e-01$ | $9.90e-01$ | $6.44e-01$ | $5.08e-01$ | - | - | - | - | - | - | - | - | - | - | - | - |
| Lbude | **$2.67e-03$** | **$1.34e-04$** | **$1.66e-03$** | **$2.08e-02$** | **$7.79e-04$** | **$3.87e-08$** | **$2.34e-04$** | **$2.01e-03$** | $7.73e-01$ | - | - | - | - | - | - | - | - | - | - | - |
| Lmorm | **$8.41e-04$** | **$9.88e-08$** | **$7.86e-07$** | **$1.22e-08$** | **$5.48e-08$** | **$1.63e-03$** | **$2.10e-10$** | **$1.42e-05$** | $5.13e-01$ | **$1.21e-02$** | - | - | - | - | - | - | - | - | - | - |
| Mmerl | $7.61e-01$ | $4.26e-01$ | $5.08e-01$ | $2.68e-01$ | $2.97e-01$ | $9.29e-01$ | $7.07e-01$ | **$5.10e-03$** | $6.72e-01$ | $4.81e-01$ | $5.43e-02$ | - | - | - | - | - | - | - | - | - |
| Msurm | **$1.69e-03$** | **$3.90e-11$** | **$6.37e-06$** | **$2.37e-09$** | **$3.10e-08$** | **$2.25e-03$** | **$3.22e-10$** | **$3.37e-06$** | $9.43e-01$ | **$1.18e-03$** | **$3.45e-09$** | $8.60e-01$ | - | - | - | - | - | - | - | - |
| Peryt | **$6.06e-03$** | **$4.81e-06$** | **$1.46e-08$** | **$3.71e-07$** | **$2.35e-06$** | **$2.53e-04$** | **$2.07e-09$** | **$7.57e-05$** | $8.95e-01$ | **$9.42e-04$** | **$1.52e-09$** | $3.01e-01$ | **$5.82e-09$** | - | - | - | - | - | - | - |
| Scabr | $2.77e-01$ | **$4.88e-02$** | **$2.07e-02$** | **$1.23e-02$** | **$2.14e-02$** | $3.47e-01$ | **$4.13e-03$** | $1.07e-01$ | $8.02e-01$ | $2.15e-01$ | **$6.80e-03$** | $6.07e-01$ | **$1.15e-02$** | **$2.44e-04$** | - | - | - | - | - | - |
| Scant | **$2.76e-04$** | **$4.51e-02$** | $6.99e-02$ | **$6.47e-03$** | **$3.60e-03$** | $2.86e-01$ | **$4.87e-04$** | **$2.22e-02$** | $9.48e-01$ | $1.19e-01$ | **$3.55e-04$** | $8.03e-02$ | $7.51e-02$ | **$3.81e-02$** | $1.32e-01$ | - | - | - | - | - |
| Scine | $1.89e-01$ | **$4.66e-02$** | $4.36e-01$ | $1.57e-01$ | $2.90e-01$ | $5.49e-01$ | $2.20e-01$ | $3.67e-01$ | $8.47e-01$ | $5.91e-01$ | $1.73e-01$ | **$5.92e-03$** | $2.01e-01$ | $1.15e-01$ | $8.13e-01$ | $6.74e-02$ | - | - | - | - |
| Spilc | **$3.19e-03$** | **$2.57e-06$** | **$1.18e-05$** | **$1.42e-06$** | **$3.65e-05$** | **$4.03e-05$** | **$2.67e-07$** | **$1.33e-05$** | $8.02e-01$ | **$2.34e-03$** | **$2.08e-07$** | $1.29e-01$ | **$3.25e-07$** | **$4.15e-07$** | **$4.55e-02$** | **$1.20e-02$** | **$4.08e-03$** | - | - | - |
| Ssard | **$7.54e-03$** | **$3.34e-06$** | **$3.54e-05$** | **$4.46e-07$** | **$1.50e-06$** | **$4.80e-02$** | **$1.15e-08$** | **$2.86e-05$** | **$7.11e-01$** | **$1.02e-02$** | **$4.27e-07$** | $1.65e-01$ | **$3.72e-07$** | **$6.01e-08$** | **$1.87e-09$** | **$1.10e-02$** | $4.42e-01$ | **$5.92e-05$** | - | - |
| Styph | $2.99e-01$ | $4.32e-01$ | $6.44e-01$ | $5.67e-02$ | $3.68e-01$ | $6.01e-01$ | $6.04e-01$ | $2.97e-01$ | **$4.08e-08$** | $4.57e-01$ | $8.55e-01$ | $6.42e-01$ | $6.44e-01$ | $7.86e-01$ | $9.99e-01$ | $7.75e-01$ | $3.90e-01$ | $4.51e-01$ | $5.29e-01$ | - |

Table S8: **Correlation between pairwise species net genetic divergence ($d_{XY}$)** - Each entry shows the $p-value$ of the linear regression modeling of gene genetic diversity between corresponding species on each row and column. Significant correlations at $p-value < 0.05$ are shown in bold.

| | Aboye | Afall | Cgale | Cjuli | Dlabr | Dpunt | Eencr | Gnige | Hgutt | Lbude | Lmorm | Mmerl | Msurm | Peryt | Scabr | Scant | Scine | Spilc | Ssard | Styph |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Aboye | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
| Afall | **$2.73e-04$** | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
| Cgale | **$1.75e-05$** | **$9.30e-09$** | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
| Cjuli | **$2.55e-03$** | **$5.64e-10$** | **$2.51e-06$** | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
| Dlabr | **$3.89e-03$** | **$3.08e-09$** | **$3.37e-04$** | **$2.28e-05$** | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
| Dpunt | **$3.46e-04$** | $1.18e-01$ | $1.97e-01$ | $2.10e-01$ | **$2.60e-02$** | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
| Eencr | **$1.38e-02$** | **$1.76e-07$** | **$1.71e-04$** | **$1.02e-05$** | **$6.69e-06$** | **$2.23e-03$** | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
| Gnige | **$3.10e-03$** | **$5.45e-07$** | **$5.56e-04$** | **$1.00e-04$** | **$1.05e-04$** | $1.47e-01$ | **$1.09e-03$** | - | - | - | - | - | - | - | - | - | - | - | - | - |
| Hgutt | $9.91e-02$ | **$5.21e-03$** | **$2.81e-02$** | $1.95e-01$ | $5.37e-02$ | $1.98e-01$ | **$5.79e-03$** | **$1.97e-03$** | - | - | - | - | - | - | - | - | - | - | - | - |
| Lbude | **$5.32e-03$** | **$2.66e-04$** | **$1.37e-02$** | **$1.09e-04$** | **$6.49e-04$** | **$1.53e-02$** | **$1.67e-04$** | **$1.08e-02$** | $7.73e-01$ | - | - | - | - | - | - | - | - | - | - | - |
| Lmorm | **$1.77e-03$** | **$1.29e-11$** | **$2.78e-05$** | **$2.15e-08$** | **$1.52e-07$** | $1.62e-01$ | **$1.49e-05$** | **$5.90e-04$** | **$4.56e-03$** | **$7.72e-03$** | - | - | - | - | - | - | - | - | - | - |
| Mmerl | $3.15e-01$ | $3.64e-01$ | $6.33e-01$ | $6.95e-01$ | $2.17e-01$ | $6.67e-01$ | $8.64e-01$ | $4.46e-01$ | $6.02e-01$ | **$7.95e-01$** | $8.56e-01$ | - | - | - | - | - | - | - | - | - |
| Msurm | **$4.68e-03$** | **$9.43e-11$** | **$1.41e-06$** | **$5.14e-09$** | **$7.02e-08$** | $1.83e-01$ | **$3.91e-07$** | **$1.41e-05$** | **$3.49e-02$** | **$4.40e-05$** | **$1.34e-10$** | $5.99e-01$ | - | - | - | - | - | - | - | - |
| Peryt | **$2.07e-02$** | **$9.05e-09$** | **$3.32e-03$** | **$5.54e-06$** | **$2.08e-06$** | $2.91e-01$ | **$4.43e-04$** | **$1.41e-05$** | **$3.84e-03$** | **$2.22e-02$** | **$7.48e-09$** | $8.04e-01$ | **$4.76e-07$** | - | - | - | - | - | - | - |
| Scabr | **$6.13e-05$** | **$4.51e-07$** | **$7.99e-06$** | **$1.28e-07$** | **$9.71e-07$** | $3.17e-01$ | **$2.21e-04$** | **$5.45e-03$** | $1.24e-01$ | **$1.39e-03$** | **$3.43e-09$** | $2.91e-01$ | **$4.94e-05$** | **$1.51e-03$** | - | - | - | - | - | - |
| Scant | **$1.27e-03$** | **$4.25e-02$** | **$2.84e-02$** | **$1.20e-02$** | **$1.36e-02$** | **$4.31e-02$** | **$1.67e-02$** | $5.31e-02$ | $1.79e-01$ | **$1.21e-05$** | **$1.16e-02$** | $8.27e-01$ | $5.92e-02$ | $1.22e-01$ | **$1.77e-04$** | - | - | - | - | - |
| Scine | $8.04e-01$ | $8.41e-01$ | $6.47e-01$ | $6.66e-01$ | $7.54e-01$ | $7.95e-01$ | $8.14e-01$ | $9.76e-01$ | $8.73e-01$ | $8.74e-01$ | $9.67e-01$ | **$1.35e-04$** | $7.45e-01$ | $7.45e-01$ | $3.90e-01$ | $5.75e-01$ | - | - | - | - |
| Spilc | $3.15e-01$ | $7.19e-01$ | $8.02e-01$ | $8.14e-01$ | $4.05e-01$ | $5.37e-01$ | $6.01e-01$ | $7.67e-01$ | $9.06e-01$ | $7.99e-01$ | $1.62e-01$ | $3.51e-01$ | $7.53e-01$ | $4.20e-01$ | $4.60e-01$ | $6.55e-01$ | **$3.66e-03$** | - | - | - |
| Ssard | **$1.79e-04$** | **$7.07e-16$** | **$3.19e-07$** | **$1.14e-10$** | **$4.69e-10$** | **$1.16e-02$** | **$1.78e-10$** | **$8.22e-07$** | **$3.62e-03$** | **$2.58e-06$** | **$3.38e-12$** | $8.43e-01$ | **$9.92e-11$** | **$7.18e-09$** | **$1.31e-07$** | **$1.03e-03$** | $9.16e-01$ | $5.89e-01$ | - | - |
| Styph | $8.83e-01$ | $5.91e-01$ | $4.40e-01$ | **$2.35e-02$** | $5.49e-01$ | $7.63e-01$ | $2.01e-01$ | $7.30e-01$ | $6.83e-01$ | $7.88e-01$ | $5.70e-01$ | $9.54e-01$ | $7.03e-01$ | $6.57e-01$ | $5.68e-01$ | $8.94e-01$ | $2.43e-01$ | $9.58e-01$ | $5.33e-01$ | - |

Table S9: **Correlation between pairwise species absolute genetic divergence ($d_a$)** - Each entry shows the $p-value$ of the linear regression modeling of gene genetic diversity between corresponding species on each row and column. Significant correlations at $p - value < 0.05$ are shown in bold.

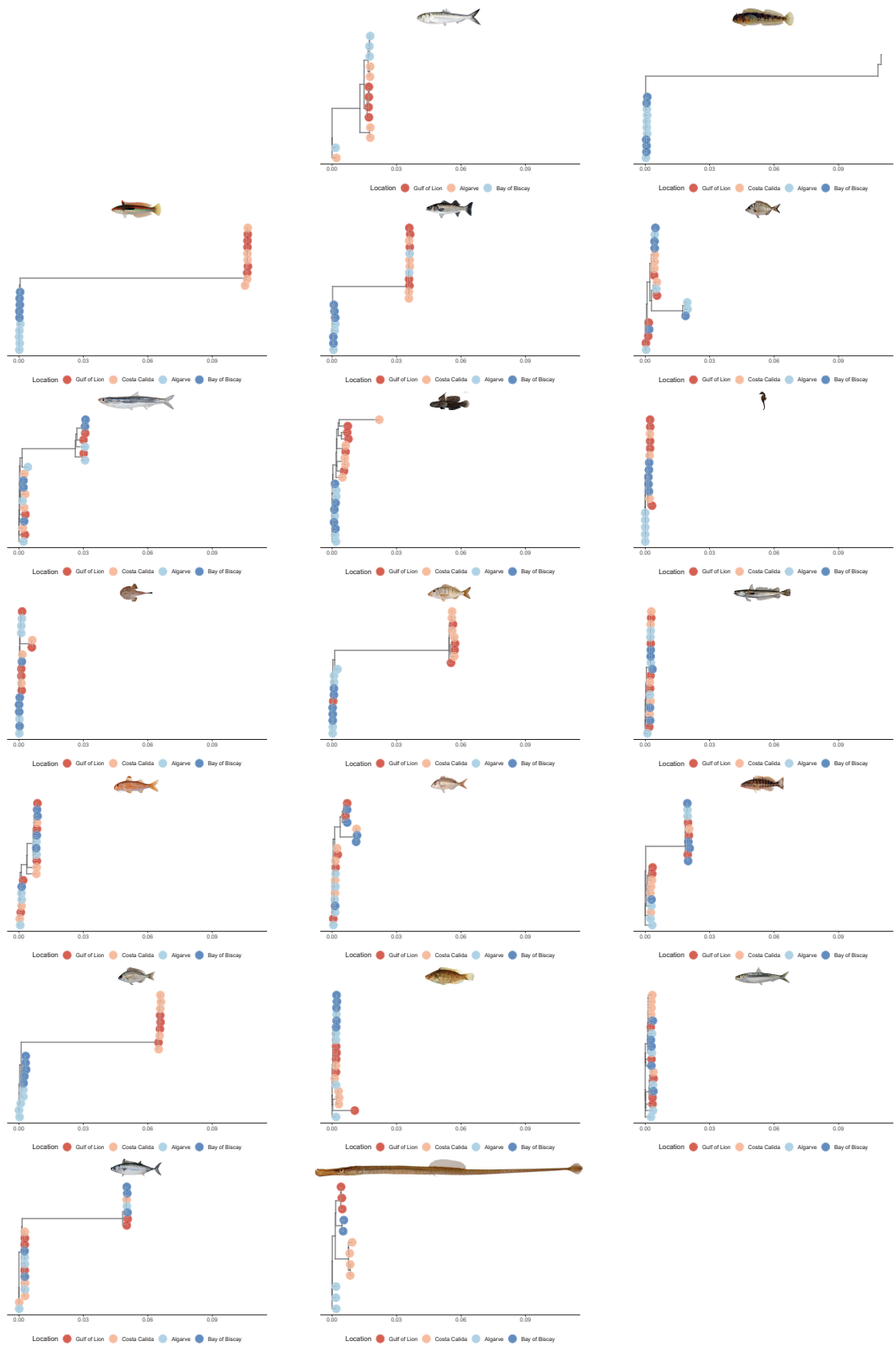



Figure S1: **Unrooted phylogenetic trees inferred from whole-mitochondrial genome for 20 marine teleostan fish species.**

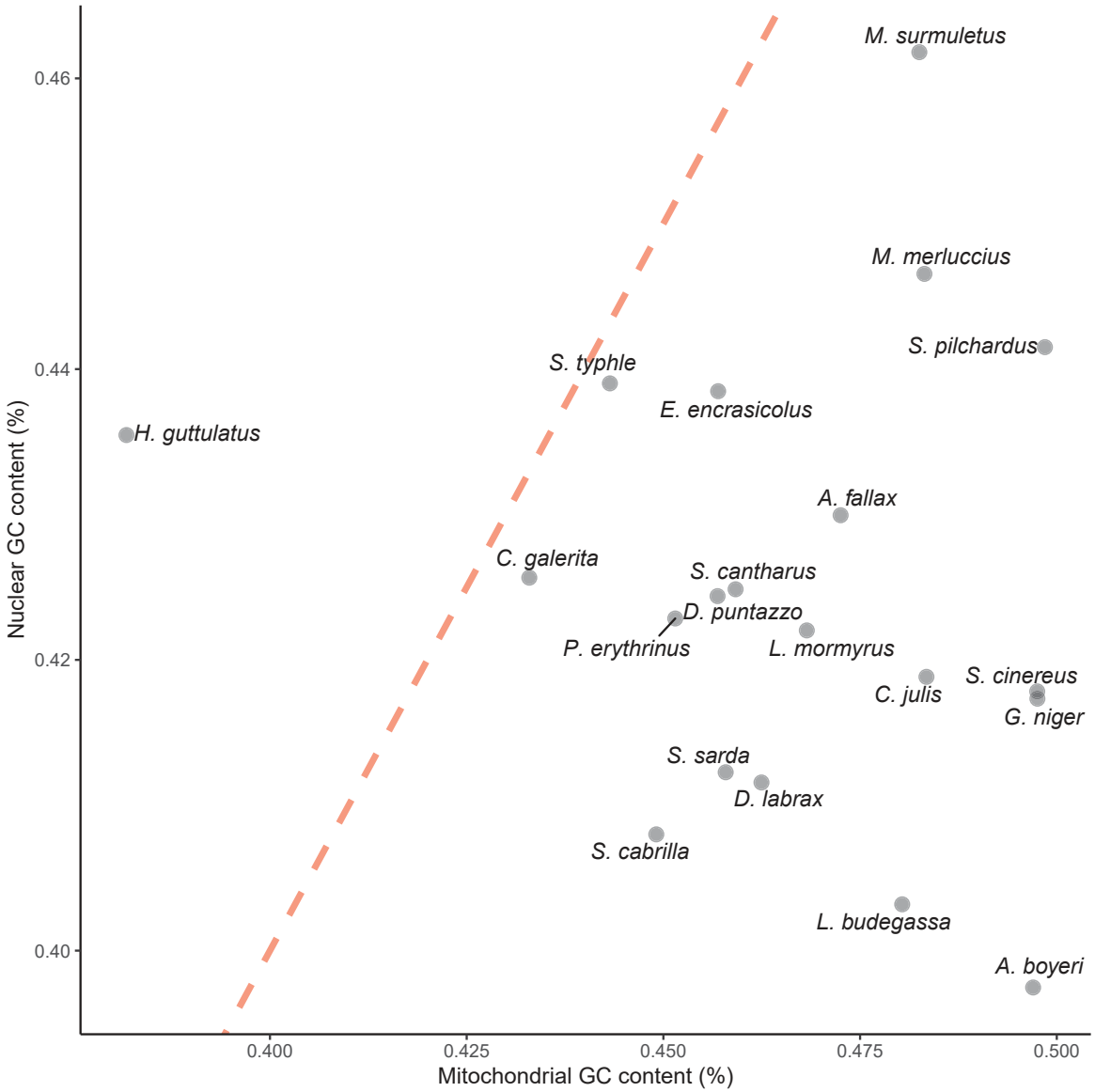



Figure S2: **No correlation between mitochondrial and nuclear GC content (%) across 20 fish species ($t$-test, $t = -0.56$, $p - value = 0.58$, $R^2 = 0.0169$). The red dotted line shows the $y = x$ relationship.**

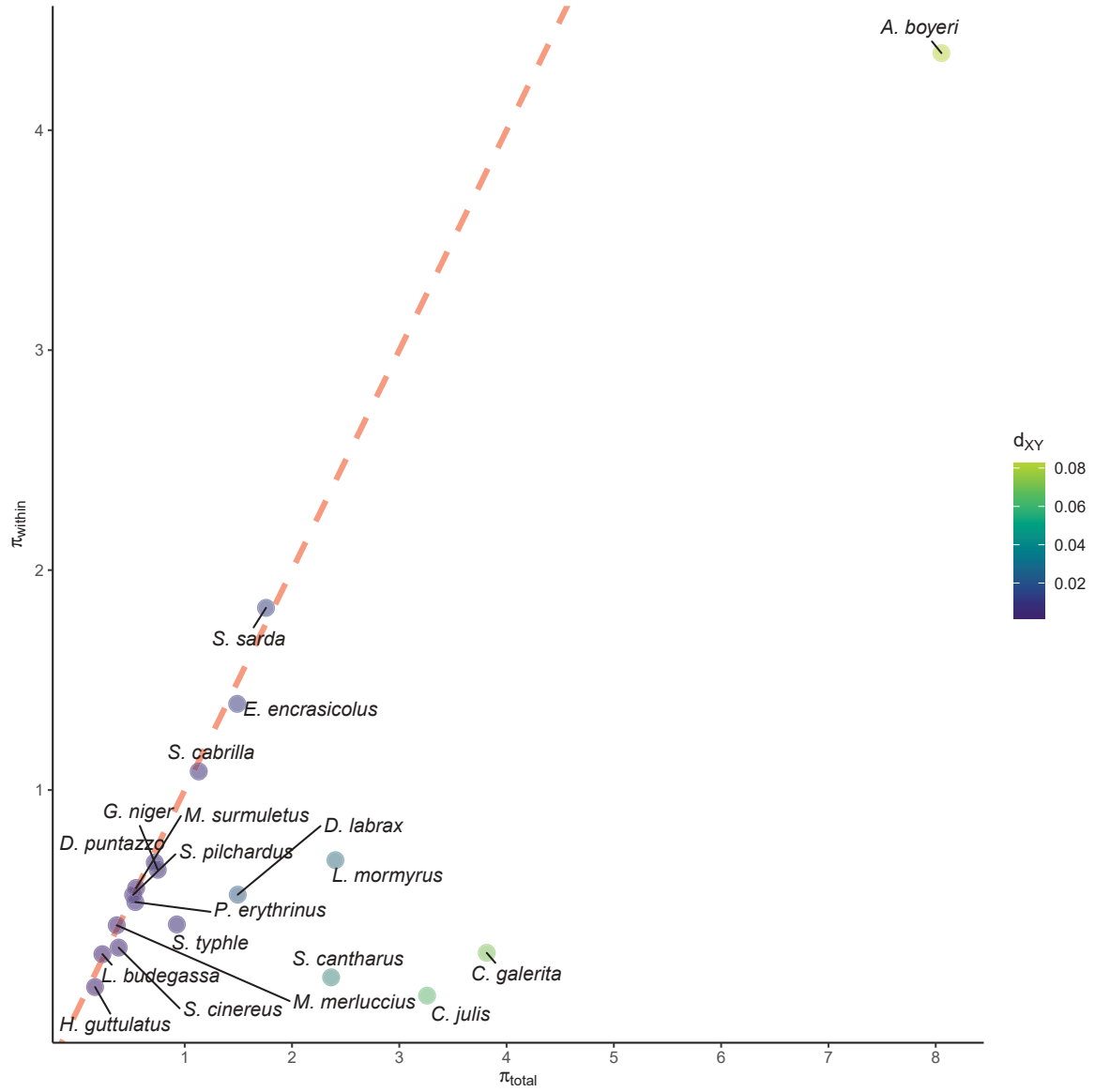



Figure S3: **Relationship between total genetic diversity $\pi_{total}$ and mean within-population genetic diversity, $\pi_{within}$.** Light and dark colors represent respectively high and low absolute genetic divergence ($d_{XY}$) between Atlantic and Mediterranean outer populations. The red dotted line shows the $y = x$ relationship.

# ANNEX 4

# Annexe 4
# The determinants of linked selection in marine fishes

## Abstract

Under the neutral theory, genetic diversity ($\pi$) is expected to be equal to the product of the effective population size ($N_e$) and the mutation rate ($\mu$) (Kimura, 1983). If $N_e$ reflects the species abundance ($N$), we would expect to see a large variability in observed $\pi$ among species. However, the first empirical estimates of the genetic diversity show that the variability of $\pi$ is much more reduced compared to the variability of $N$. This discrepancy was coined as the Lewontin paradox and it still remains a mystery among evolutionary biologists (Lewontin, 1974).

Several mechanisms have been proposed to explain this paradox. Neutral explanations involve more frequent fluctuations in population or higher variance in reproductive success in high $N$ species that cause a higher reduction to $N_e$. On the other hand, selective explanations mainly focused on the role of linked selection on the depletion of diversity. The removal of deleterious mutations through background selection (Charlesworth et al., 1993) or fixation of advantageous mutations through selective sweeps (Maynard Smith and Haigh, 1974) affect also neutral linked variation through linked selection. As selection is stronger in high $N_e$ species, reduction of linked genetic diversity is supposed to be higher in high $N$ species. This effect can be measured by analyzing the intragenomic variation in genetic diversity with local recombination rate. In high $N_e$ species, local genetic diversity is mainly structured by the local recombination rate because low-recombining regions will affect more linked neutral variations compared to high-recombining regions. In contrast, local genetic diversity is supposed to be poorly correlated to the local recombination rate in low $N_e$ species, as selection is less prevalent. By applying this approach, Corbett-Detig et al. (2015) found that linked selection is indeed negatively correlated to body size and positively to range size, concordant with low selection intensity in high $N$ species. However, in a meta-analysis, Buffalo (2021) show that linked selection only is not able to explain the Lewontin paradox.

Here, we conducted the same approach as Corbett-Detig et al. (2015) and evaluated whether linked selection intensity is determined by life-history traits in marine fishes. This work was conducted by Marion Talbi during its M2 internship at the University of Toulouse from January to June 2021 supervised by Pierre-Alexandre Gagnaire and I. Starting from the VCFs of 10 species with high reference genome quality, she wrote the pipeline to perform: i) physical phasing from individual BAM files with `WHatsHap` (Martin et al., 2016), ii) statistical phasing from pre-phased VCF with `SHAPEit4` (Delaneau et al., 2019), iii) estimation of the ancestral state with `est-sfs` (Keightley and Jackson, 2018), iv) estimation of local recombination rate with `LDHelmet` (Chan et al., 2012). Then, she inferred local variation in genetic diversity and she developed a method to infer the strength of linked selection from a comparison between local recombination rate and genetic diversity in 10kb windows. Then, she compared these estimations with several life-history traits that are correlated to $N_e$ and that might affect the strength of linked selection.

By comparing the estimated local recombination landscapes of *D. labrax* with that obtained of a previous studiy with phased trios, she found a high accuracy of `LDHelmet` to infer the local recombination rate. She found variability in the intensity of linked selection and global mean recombination rate betwee species. However, no life-history traits were correlated to linked selection intensity, including those related to $N$, such as body size and lifespan. Adding the 10 other species in the dataset might help further address these questions in the future.

# References

Buffalo, V. (2021). Quantifying the relationship between genetic diversity and population size suggests natural selection cannot explain Lewontin's Paradox. *eLife*, 10:e67509.

Chan, A. H., Jenkins, P. A., and Song, Y. S. (2012). Genome-Wide Fine-Scale Recombination Rate Variation in Drosophila melanogaster. *PLOS Genetics*, 8(12):e1003090.

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

Corbett-Detig, R. B., Hartl, D. L., and Sackton, T. B. (2015). Natural Selection Constrains Neutral Diversity across A Wide Range of Species. *PLOS Biology*, 13(4):e1002112.

Delaneau, O., Zagury, J.-F., Robinson, M. R., Marchini, J. L., and Dermitzakis, E. T. (2019). Accurate, scalable and integrative haplotype estimation. *Nature Communications*, 10(1):5436.

Keightley, P. D. and Jackson, B. C. (2018). Inferring the Probability of the Derived vs. the Ancestral Allelic State at a Polymorphic Site. *Genetics*, 209(3):897–906.

Kimura, M. (1983). *The Neutral Theory of Molecular Evolution*. Cambridge University Press, Cambridge.

Lewontin, R. C. (1974). *The Genetic Basis of Evolutionary Change*. Number 25 in Columbia Biological Series. Columbia Univ. Pr, New York.

Martin, M., Patterson, M., Garg, S., Fischer, S. O., Pisanti, N., Klau, G. W., Schöenhuth, A., and Marschall, T. (2016). WhatsHap: Fast and accurate read-based phasing.

Maynard Smith, J. M. and Haigh, J. (1974). The hitch-hiking effect of a favourable gene. *Genetics Research*, 23(1):23–35.

# Influence des traits d'histoire de vie sur l'érosion de la diversité neutre par la sélection en liaison chez les poissons marins

TALBI Marion

Sous la direction de :

Pierre-Alexandre Gagnaire & Pierre Barry

Institut des Sciences et de l'Evolution de Montpellier

Master 2 – Ecologie et Evolution

2020 - 2021

# Résumé

La théorie neutre de l'évolution prédit que la diversité génétique d'une population est proportionnelle à sa taille efficace ($N_e$). Les données moléculaires montrent cependant une magnitude plus faible des variations de diversité génétique neutre entre espèces par rapport à leurs variations de tailles de populations. Cet écart nommé « paradoxe de Lewontin » peut être en partie expliqué par l'effet de la sélection en liaison, qui érode relativement plus de diversité neutre dans les populations de grandes tailles. Alors que l'impact de la sélection en liaison sur la diversité génétique a été mesuré chez plusieurs espèces, les déterminants de l'intensité de la sélection en liaison restent mal connus. Dans cette étude, les génomes de vingt individus chez dix espèces de poissons marins ont été utilisés pour comparer l'effet de la sélection en liaison entre espèces aux traits de vie contrastés. Un paysage de recombinaison populationnelle ($\rho$) a été estimé pour chaque espèce à l'aide d'une méthode combinant un pré-phasage physique des données à une inférence statistique de la phase haplotypique. Les paysages obtenus ont été comparés avec ceux issus de données à phase connue pour valider la méthode. Les paysages de diversité nucléotidique ($\pi$) ont également été déterminés afin d'estimer l'effet de la sélection en liaison via l'analyse de la covariation entre $\pi=4N_e\mu$ et $\rho=4N_er$ le long des génomes. La comparaison de l'effet de la sélection en liaison entre espèces a enfin été réalisée pour tenter d'identifier des déterminants potentiels. Les résultats suggèrent une forte variabilité de l'intensité de la sélection en liaison entre espèces. Bien qu'aucun facteur expliquant ces différences n'ait pu être identifié, certains effets comme le niveau de recombinaison moyen, la taille du génome, et des traits d'histoire de vie resteront à préciser. En permettant d'évaluer l'intensité de la sélection en liaison à partir du polymorphisme populationnel, la méthodologie développée ouvre de nouvelles possibilités pour mieux identifier ses déterminants.

Diversité génétique – Recombinaison – Phasage – Sélection en liaison – Traits de vie

The neutral theory of evolution predicts that a population's genetic diversity is proportional to its effective size ($N_e$). However, molecular data shows a discrepancy between the narrow range of variation in genetic diversity and the large range of population size variation among species. This gap, known as the "Lewontin's paradox", can be partly explained by the effect of linked selection, supposed to erode higher amounts of neutral diversity in highly abundant species. While the effect of linked selection on genome-wide genetic diversity has been evaluated in several species, the main determinants of linked selection intensity remain largely unknown. Here, we used genome sequence data from 20 individuals in 10 species of marine fish to compare the effect of linked selection with species' traits diversity. We determined the population recombination landscape ($\rho$) of each species by combining physical pre-phasing with classic statistical inference of haplotype phase. To evaluate this method, inferred landscapes were compared with the ones obtained from known haplotype phase data. Genetic diversity landscapes ($\pi$) were also measured to capture the effect of linked selection through the analysis of the covariation between $\pi=4N_e\mu$ and $\rho=4N_er$ along the genomes. The estimated strength of linked selection was then compared among species to identify possible determinants. Our results suggest that linked selection intensity is highly variable among species. Although no factors explaining these differences could be identified, some effects such as average recombination rate, genome size, and life history traits remain to be clarified. By making it possible to assess the intensity of linked selection based on polymorphism data, the methodology developed opens new possibilities for better identifying its main determinants.

Genetic diversity – Population recombination rate – Phasing – Linked selection – Life history

# Contribution personnelle

Mes encadrants ont été présents dès le début du stage afin de me guider dans les différentes étapes.

Une grande partie de ce stage a été la mise en place puis la réalisation du pipeline bio-informatique dans le but d'obtenir les paysages de recombinaison populationnels à partir de fichiers au format *Variant Calling Format* (VCF) pour chacune des dix espèces disponibles en début de stage, et préparées en amont par mer encadrants.

J'ai dans un premier temps effectué un pré-phasage physique des données grâce au programme *Whatshap*, puis un phasage statistique inféré à partir des données populationnelles via le programme *SHAPEIT4* afin d'obtenir des séquences phasées pour l'analyse des paysages de recombinaison avec *LDhelmet*. J'ai réalisé cette étape en utilisant le cluster de calcul hébergé par MBB, et géré la parallélisation de mes tâches grâce à des scripts Python développés pendant le stage et des pipelines d'analyses construits grâce à l'outil *Snakemake*.

Pour chacune des dix espèces, j'ai ensuite dû déterminer pour chaque SNP l'état ancestral et dérivée de chaque variant. A partir de trois groupes externes que mes encadrants m'ont aidé à déterminer, j'ai procédé à des étapes de tri et de transformation des données requérant entre autres l'utilisation de scripts en langage *R* et *bash* pour récupérer l'état allélique aux sites variants, puis j'ai calculé la probabilité de l'état de l'allèle ancestral à l'aide du programme *est-sfs*.

Une fois les séquences phasées et les probabilités ancestrales formatées, j'ai calculé les matrices de transition pour chacune des espèces à l'aide d'un script *R* avant de déterminer les paramètres pour pouvoir lancer l'inférence des paysages de recombinaison avec *LDhelmet*. Quatre machines de calculs à fortes capacités ont été mises à ma disposition pour que je puisse avancer en parallèle et à vitesse suffisante pour obtenir les résultats en temps voulu pour un maximum d'espèces.

J'ai ensuite estimé la diversité génétique pour chaque fragment génomique disponible grâce à un script développé par Pierre Barry, et mis en lien les données de polymorphisme et de recombinaison populationnelles grâce à mes scripts *R*.

J'ai participé avec mes encadrants à la conception d'une méthode pour pouvoir capturer l'effet de la sélection en liaison à travers l'analyse des covariations entre $\pi=4N_e\mu$ et $\rho=4N_e r$ le long du génome.

Enfin, j'ai effectué les analyses comparatives intra et inter-spécifiques sur les résultats obtenus grâce à mes scripts *R*.

# COVID-19

Le COVID-19 n'a pas eu d'impact particulièrement lourd sur le bon déroulement de mon stage. J'ai moi-même été malade, et ai par conséquent dû commencer mon stage avec quelques jours de retards. Après cela, j'ai pu me rendre sur le lieu du stage presque quotidiennement, et ai ainsi pu bénéficier de la vie du laboratoire, des discussions avec mes encadrants ainsi qu'avec les chercheurs et les doctorants, ce qui m'a permis d'échanger facilement sur de nombreux points.

Le COVID-19 a cependant impacté la présence d'autres personnes spécialisées dans le traitement informatique et bio-informatique dans le laboratoire, et a donc allongé le temps de traitement en cas de demandes spécifiques (e.g installation de logiciels requérant une manipulation particulière, problème au niveau du cluster).

Lors du confinement, mon tuteur a pu continuer de m'encadrer et de m'apporter l'aide et le support nécessaire au bon déroulement du stage, en présentiel tout comme en distanciel le cas échéant, avec notamment des appels vidéo.

D'un point de vue technique, ayant effectué une grande partie de mon stage dans le traitement bio-informatique, la possibilité de travailler à distance grâce au réseau VPN permettant l'accès en distanciel aux nombreuses machines de calculs et ordinateur utilisés a été très pratique en ces circonstances particulières.

.

# Remerciements

Mes premiers remerciements vont pour mon premier encadrant, Pierre-Alexandre Gagnaire, qui s'est rendu disponible tout au long du stage pour que je puisse apprendre et avancer dans les meilleures conditions. En plus de m'accorder sa confiance pour que je puisse réaliser ce projet, il a fait preuve de beaucoup de pédagogie, de gentillesse, et d'encouragements, et j'ai énormément appris au cours de ce stage passionnant.

Je remercie mon second encadrant, Pierre Barry, qui m'a fait découvrir le langage Python, et a également été à l'écoute et de bons conseils face à toutes mes questions. Merci également pour sa bienveillance.

Je remercie toutes les personnes de l'équipe BEM de l'ISEM qui ont tous fait preuve de gentillesse: Nicolas qui m'a co-accueilli dans le bureau, les doctorants, notamment Marie et Laura avec qui j'ai pris plaisir à dessiner les dimanches, et les stagiaires, dont Salomé avec qui on se sera soutenu tout du long.

Merci enfin à ma famille, mes amis, et mon colocataire, qui ont fait preuve comme toujours d'un soutien inconditionnel.

# Table des matières

Résumé ..... 1

Contribution personnelle ..... 2

COVID-19 ..... 3

Remerciements ..... 3

I- Introduction ..... 6

II- Matériel et méthodes ..... 10

1- Données ..... 10

2- Estimation du taux de recombinaison populationnel ..... 10

3- Mesure de la diversité nucléotidique locale (pi) ..... 13

4- Validation de la méthode et définition de la taille des fenêtres ..... 14

5- Estimation de l'effet de la sélection en liaison ..... 15

6- Impact des traits d'histoire de vie sur l'intensité de l'érosion de la diversité nucléotidique par la sélection en liaison ..... 16

III- Résultats ..... 17

1- Filtre des VCF ..... 17

2- Pré-phasage des haplotypes ..... 18

3- Taux de recombinaison populationnel ..... 18

4- Estimation de l'effet de la sélection en liaison : ..... 20

5- Impact des traits d'histoire de vie sur l'intensité de l'érosion de la diversité nucléotidique par la sélection en liaison ..... 24

Discussion ..... 25

I/Inférence des paysages de recombinaison populationnelle : Limites et améliorations ..... 26

II/ Variation interspécifique du $\rho$ inter espèces ..... 28

III/Erosion du polymorphisme par la sélection en liaison ..... 29

IV/Déterminants de la sélection en liaison entre espèces ..... 31

Conclusion :.....32

Bibliographie.....32

# I- Introduction

Dans une population idéale stable à reproduction panmictique, où chaque individu a la même espérance de contribuer à la génération suivante (modèle de Wright-Fisher), la diversité génétique neutre (notée $\theta$) résulte d'un équilibre entre l'introduction de nouvelles mutations à un taux $\mu$ et leur perte par dérive génétique à un taux proportionnel à $1/N_e$ (Lynch, 2007). La diversité génétique neutre attendue à l'équilibre mutation-dérive est égale à $\theta=4*N_e*\mu$ (Kimura and Crow, 1964), et reflète donc en partie la taille efficace de la population. Dans les années 1970, avec l'arrivée des premières données de polymorphisme moléculaire, on constate cependant que les différences de niveaux de diversité génétique entre espèces sont largement inférieures à leurs différences de tailles de populations. Cette différence de magnitude, appelée « paradoxe de Lewontin » (Lewontin, 1974), et récemment quantifiée comme égale à un facteur dix (Buffalo, 2021), a été et reste encore aujourd'hui une énigme majeure en génétique des populations.

Deux principaux facteurs ont été proposés pour expliquer ce paradoxe : les fluctuations historiques des tailles de population et la sélection.

Les variations temporelles de taille efficace provoquées par les changements environnementaux et les oscillations climatiques ont un effet direct sur la diversité génétique (Nei, 1975). En effet, celle-ci est supposée refléter la moyenne harmonique de la taille efficace sur le long terme, elle-même sensible aux épisodes de taille efficace réduite survenus au cours du temps, comme les goulots d'étranglement (Charlesworth, 2009).

L'effet d'érosion du polymorphisme par la sélection en liaison, qui affecte la diversité génétique des locus neutres via leur liaison aux locus voisins sous sélection, est une autre explication proposée pour résoudre le paradoxe de Lewontin (Roberts, 2015). En modifiant la fréquence d'un allèle dans une population, la sélection change aussi la fréquence de l'haplotype sur lequel repose l'allèle sous sélection. Les mutations neutres présentes sur cet haplotype subissent ainsi l'effet indirect de la sélection, qui diminue progressivement avec la distance de recombinaison au locus sélectionné (Comeron, 2014).

La sélection peut être positive, et en favorisant un variant, emporter avec lui les mutations neutres environnantes. Cet effet d'autostop peut ainsi conduire à un balayage sélectif qui gomme la diversité génétique neutre liée (Josephs and Wright, 2016). La force du balayage peut être plus ou

moins intense, suivant que la mutation soit plus fortement (ex : mutation adaptative d'un trait monogénique) ou plus faiblement sélectionnée (ex : traits quantitatifs polygéniques) (Josephs and Wright, 2016). La sélection peut également être négative, et en éliminant des allèles délétères de la population, retirer au passage les allèles neutres liés (Charlesworth et al., 1993). Dans les deux cas, il en résulte une baisse de la diversité génétique neutre à proximité des locus sous sélection, et donc une érosion du polymorphisme par la sélection en liaison (Rettelbach et al., 2019).

L'arrivée des séquenceurs de nouvelle génération a permis l'accès du polymorphisme de génomes entiers chez des espèces non modèles, permettant ainsi d'étudier les paysages génomiques de diversité chez de nombreuses espèces (Martin et al., 2016). Ces technologies, couplées à des développements statistiques spécifiques, permettent aujourd'hui d'estimer avec finesse la force et la direction de la sélection afin d'évaluer la part de la sélection en liaison dans le paradoxe de Lewontin (Roberts, 2015).

L'érosion de la diversité génétique neutre par la sélection en liaison peut être capturée par l'étude de la corrélation entre le polymorphisme et le taux local de recombinaison. Une méta-analyse portant sur plus de 40 espèces animales et végétales sexuées a montré une érosion plus forte de la diversité génétique par la sélection en liaison chez les espèces les plus abondantes(Corbett-Detig et al., 2015), confirmant ainsi l'attendu d'une sélection plus efficace dans les populations de grande taille (Kimura, 1979). Un effet similaire a également été montré dans une étude récente portant sur plus de 140 espèces, dans laquelle les relations entre la diversité nucléotidique ($\pi$), la taille contemporaine de la population ($N_c$) et la longueur de la carte de combinaison, ont été quantifiées pour étudier l'impact de la sélection en liaison (Buffalo, 2021). Cette étude a révélé une relation négative entre le $N_c$ et la longueur de la carte de recombinaison, indiquant un impact potentiellement plus élevé de la sélection en liaison chez les espèces abondantes dont la carte de recombinaison est plus condensée.

Ces études mettent ainsi en avant le rôle central de la recombinaison dans les variations de la diversité génétique. Celle-ci est généralement diminuée dans les régions du génome où la recombinaison est faible, alors qu'elle tend vers la diversité attendue sous un modèle neutre dans les régions où la recombinaison est forte (Begun and Aquadro, 1992 ; Comeron et al., 2014) Des modèles plus réalistes ont également permis de mesurer l'impact de la sélection en liaison sur la diversité génétique en prenant en compte la densité locale en éléments fonctionnels, en plus de

l'environnement recombinationnel (Corbett-Detig et al., 2015; Elyashiv et al., 2016; Wang et al., 2016)Cependant, malgré des mesures de plus en plus intégratives des caractéristiques de l'environnement génomique des mutations, la sélection en liaison seule ne semble pas suffisante pour expliquer les différences de magnitude entre les tailles de populations (qui varient de plusieurs ordres de grandeurs entre espèces) et les niveaux de diversité génétique observés (qui diffèrent seulement d'un facteur 100) (Buffalo, 2021; Coop, 2016)

Afin de mieux caractériser l'intensité d'érosion du polymorphisme par la sélection en liaison et d'en comprendre les déterminants majeurs, d'autres études comparatives multi-espèces comme celle de Corbett-Detig et al (2015) semblent nécessaires. Ces études restent néanmoins exigeantes, car elles nécessitent d'avoir accès à des génomes de référence et à des cartes de recombinaison obtenues via des croisements expérimentaux souvent difficiles à mettre en place (Peñalba and Wolf, 2020). Ainsi, l'impact de la sélection en liaison reste encore inconnu chez de nombreux groupes taxonomiques.

L'inférence des paysages génomiques de recombinaison populationnelle ($\rho$=4$N_e r$) à partir des données de reséquençage de génomes peut être utilisée pour accéder aux variations de recombinaison en l'absence de cartes de liaison (McVean et al., 2004), Chen et al. 2012). De même, en l'absence d'annotation des génomes de référence chez les espèces non-modèles, il reste possible d'estimer l'effet d'érosion du polymorphisme par la sélection en liaison via une simple comparaison empirique des co-variations entre les paramètres $\pi$ et $\rho$ (Vijay et al., 2017)Ces estimations peuvent ensuite permettre d'explorer l'effet de facteurs biologiques et écologiques comme les traits d'histoire de vie sur les différences d'intensité de la sélection en liaison entre espèces (Ellegren and Galtier, 2016). Une méta-analyse de la diversité nucléotidique (Romiguier et al., 2014) a en effet montré que les traits d'histoire de vie (THV), et en particulier les stratégies reproductives, sont de bons prédicteurs de la diversité génétique neutre. Les espèces à stratégies reproductives de type *r*, produisant moins de soins parentaux et plus de progénitures, affichent des niveaux de diversité génétique plus élevés que les espèces à stratégies *K*. Cependant, l'effet de ces traits d'histoire de vie passe par des mécanismes complexes, dont les liens avec la taille efficace demeurent encore mal compris (Ellegren et Galtier, 2016). Par exemple, les effets des fluctuations passées de taille efficace pourraient être plus ou moins visibles en fonction des traits d'histoire de vie, et imprimer plus fortement le polymorphisme que ne le fait la taille efficace actuelle. Chen et

al., (2017) ont mis en lumière un effet des THV dans la variation du ratio piN/piS entre espèces, et une étude sur les pinnipèdes a révélé une corrélation entre l'habitat de reproduction et l'efficacité de la sélection (Peart et al., 2020). Néanmoins, la quantification de l'effet des THV sur la diversité via la sélection en liaison demeure un champ de recherche en développement (e.g. Corbett-Detig et al 2015, Buffalo 2021). Comprendre quels sont les déterminants de l'effet de la sélection en liaison sur le polymorphisme reste donc un enjeu majeur en génomique des populations.

Dans le but de comprendre et déterminer l'influence des traits d'histoire de vie sur l'érosion de la diversité génétique neutre par la sélection en liaison, une analyse comparative a été réalisée lors de ce stage chez dix espèces de poissons marins aux traits de vie contrastés. Bien que les poissons représentent à eux seuls près de la moitié des espèces de vertébrés, seules quelques rares études de la sélection en liaison ont été réalisées chez les Téléostéens (Gante et al., 2016; Samuk et al., 2017; Tine et al., 2014)), et aucune étude multi-espèces n'existe à ce jour. Les rares méta-analyses de la sélection en liaison ont porté sur des espèces présentant des architectures génomiques très différentes entre elles (mais voir Vijay et al. 2017 chez les oiseaux). Chez les poissons, cette architecture est en revanche assez stable puisque la plupart des espèces présentent un génome de taille assez peu variable, un caryotype comprenant 23 à 24 paires de chromosomes (Mank and Avise, 2006), une très bonne conservation de synténie, et des taux de recombinaison plus élevé dans les régions subtélomériques que centrochromosomiques (Roesti et al., 2013)Les paysages de recombinaison en « U » qui en résultent à large échelle sont relativement bien conservés entre espèces (Haenel et al., 2018)Une étude comparative de la sélection en liaison chez les poissons permettrait donc de minimiser les potentiels effets confondants des variations d'architecture génomique entre espèces. De plus, le choix a été fait d'étudier des espèces marines présentant des distributions géographiques similaires en Atlantique Nord-Est et Méditerranée, afin de minimiser les différences d'histoires démographiques entre espèces. L'étude comparative des niveaux de diversité génétique et de recombinaison populationnelle devrait donc permettre d'étudier l'influence des THV sur la sélection en liaison avec un contrôle relatif des effets architecturaux et démographiques confondants.

Grâce à ce jeu de données, l'influence de l'investissement parental (qui distingue les stratégies reproductives *r* et *K*) sera testée en comparant des espèces à investissement parental fort (ex : *Hippocampus guttulatus, Syngnatus thyphle,* (Wilson et al., 2001)) à des espèces à investissement

parental faible (*Sardina pilchardus, Sarda sarda*). L'influence de la variance interindividuelle du succès reproducteur sera également testée et capturée par la durée de vie adulte. On peut s'attendre à un effet moins fort de la sélection chez les espèces les plus longévives (e.g *Lophius budegassa, Dcentrarchus labrax*), chez qui le ratio $N_e/N_c$ est plus faible (Barry et al,. 2020). En effet, les poissons ont des courbes de survie de type III (c'est-à-dire une mortalité concentrée aux stades juvéniles) combinée à une fertilité fortement croissante avec l'âge. Certains individus qui atteignent des âges élevés participent donc de manière démesurée à la reproduction, ce qui induit une réduction du rapport $N_e/N_c$ (Coop, 2016). Enfin, l'influence de la taille du corps et du niveau trophique, largement utilisés comme proxys de la taille efficace (Corbett-Detig et al,. 2015 ; Buffalo et al,. 2021; Chen et al,. 2017 ), seront également testés, avec une intensité de la sélection en liaison plus forte chez les espèces à grande taille de corps donc de faible taille efficace.

# II- Matériel et méthodes

## 1- Données

Les données initiales sont des données de polymorphisme génomique populationnel de 10 espèces de poissons téléostéens, stockées dans un fichier VCF (*Variant Call Format*) par espèce. Un total de 20 individus ont été séquencés pour chaque espèce en amont du stage avec une profondeur moyenne d'environ 20x. L'identification des variants génétiques ainsi que le génotypage individuel ont été réalisés grâce à GATK v4 (Poplin et al., 2017) à partir des alignements des lectures d'ADN (*paired-end reads*) de 2x150 paires de bases sur le génome de référence de chaque espèce.

## 2- Estimation du taux de recombinaison populationnel

### *a- Filtre des VCF*

Un premier filtre a été réalisé sur le VCF de chaque espèce à l'aide de VCFtools (Danecek et al., 2011) dans le but d'améliorer la qualité des jeux de données de polymorphisme. Les insertions-délétions ont été retirées en conservant les positions nucléotides directement adjacentes afin de ne pas retirer de possibles SNPs (*Singe Nucleotide Polymorphism*). Ce choix semble justifié

car le module *HaplotypeCaller* de GATK est performant pour résoudre les alignements locaux autour des indels et ainsi limiter l'inférence de SNPs artefactuels.

Seul les sites bialléliques ont été conservés, et à ce stade le VCF a été divisé en sous-VCFs représentant chacun la variation présente au sein d'un segment contigu du génome de référence, appelé *scaffold*. Le nombre et la taille des scaffolds étant variable d'une espèce à l'autre en raison des qualités d'assemblage variables entre espèces, le choix a été fait de ne conserver que les *scaffolds* d'une longueur supérieure à 10 000 paires de bases. Ce seuil permet de garder une fraction représentative du génome pour chaque espèce, tout en évitant les problèmes d'estimations du *rho* dans des fragments génomiques trop courts pour être suffisamment informatifs (Table 1).

### b- *Phasage des haplotypes*

La reconstruction de la phase des génotypes nécessite de déterminer les combinaisons alléliques (ou haplotypes) qui se trouvent sur chacune des deux copies chromosomiques parentales de chaque individu. Cette étape clé en génomique des populations peut faire appel à plusieurs sources d'information.

Premièrement l'information de liaison contenue dans les lectures obtenues lors du séquençage. Les positions des *SNPs* les plus proches (c.a.d. voisins de moins de moins de 600 pb, soit la taille des inserts séquencés) peuvent se trouver physiquement sur une même paire de *reads*, et conférer ainsi une information directe sur la phase. Ainsi, la phase haplotypique de deux génotypes hétérozygotes peut être résolue s'ils sont suffisamment proches, permettant de réaliser un pré-phasage individuel à échelle chromosomique locale. Ce pré-phasage physique a été réalisé grâce au développement d'un pipeline Snakemake (Mölder et al., 2021) utilisant le programme Whatshap (Patterson et al., 2015), qui utilise les *reads* de séquençage alignés (fichiers *bam*) pour résoudre chez chaque individu la phase des positions hétérozygotes à l'échelle de blocs de petite taille (quelques centaines de pb à quelques kb). L'information de pré-phasage générée à cette étape a été incorporée dans le VCF.

Le phasage des blocs pré-phasés à l'échelle des scaffolds entiers a ensuite été réalisé par une méthode d'inférence statistique basée sur des attendus populationnels. Le logiciel SHAPEIT4 (Delaneau et al., 2019)a été utilisé pour phaser statistiquement les blocs d'haplotypes pré-phasés en suivant une procédure itérative. A chaque étape, le programme sélectionne parmi les haplotypes

présents dans la population ceux qui sont les plus proches des deux haplotypes focaux considérés, puis utilise le modèle de Li et Stephens (Li and Stephens, 2003) pour mettre à jour ces deux haplotypes à partir des haplotypes les plus similaires dans la population, jusqu'à converger vers les haplotypes optimaux. Il impose également une contrainte statistique sur le choix des haplotypes en fonction des blocs pré-phasés établis par Whatshap. A l'issue de cette étape, l'information de la phase haplotypique inférée est inscrite dans le VCF pour chaque individu à l'échelle des scaffolds entiers.

### c- *Polarisation des variants*

En plus de l'information de la phase, l'inférence de la recombinaison nécessite de connaître l'état ancestral et dérivé des sites polymorphes.

L'étape de polarisation des variants consiste à estimer une probabilité sur l'état ancestral ou dérivé d'un variant. Pour ce faire, les séquences flanquantes de chaque *SNP* ont été extraites du génome de référence (en prenant 100pb de part et d'autre) de l'espèce focale (appelée *ingroup*). Les séquences orthologues ont ensuite été récupérées par recherche d'homologie dans les génomes de référence de trois groupes externes (*outgroups*) suffisamment proches phylogénétiquement de *l'ingroup* pour présenter une homologie de séquence (cf Annexe). L'identification des meilleurs groupes externes pour chaque espèce étudiée s'est basée sur la comparaison des génomes de référence disponibles au format fasta sur Genbank et des phylogénies de téléostéens les plus complètes (e.g. Near et al, 2012). Les séquences orthologues identifiées par recherche blastn sur les génomes de référence ont été récupérées grâce à un script R. Cette information sur les séquences des *outgroups* a ensuite été combinée à celle des fréquences alléliques de *l'ingroup* pour chaque site polymorphe, afin d'inférer les état ancestraux avec le programme est-sfs (Keightley and Jackson, 2018). Cette méthode par maximum de vraisemblance détermine la probabilité que l'allèle majoritaire chez *l'ingroup* soit l'allèle ancestral, ainsi que les probabilités des compositions nucléotidiques possibles à chaque nœud ancestral de l'arbre des espèces. A partir de la sortie d'est-sfs, la probabilité que chacun des allèles soit l'allèle ancestral a été calculée en prenant en compte l'information conjointe de *l'ingroup* et des *outgroups* afin de déterminer la base la plus vraisemblablement présente chez l'ancêtre commun.

### *d- LDhelmet:*

Le logiciel LDhelmet a enfin été utilisé pour estimer les variations du taux de recombinaison populationnel ($\rho$=4$N_e r$) le long du génome à partir des patrons de déséquilibre de liaison présents dans les données de polymorphisme. LDhelmet calcule d'abord la vraisemblance pour différentes valeurs de $\rho$ et du taux de mutation populationnel ($\theta$) pour chaque paire de SNPs sous un modèle coalescent avec recombinaison pour approximer la vraisemblance totale. Les vraisemblances deux à deux sont stockées pour chaque valeur de $\theta$ et grille de $\rho$ dans une look-up table calculée en amont. Ces étapes nécessitent une matrice de transition d'états nucléotidiques calculée pour chaque espèce en suivant la méthode de Chan et al (2012), grâce à un script développé à cet effet.

La grille des valeurs de $\rho$ explorées a été fixée à 0 0.01 1.0 0.1 10 1.0 100, afin de calculer les vraisemblances pour chaque paire de SNPs à une très fine résolution autour du taux de mutation populationnel, et à une résolution plus large pour des valeurs de $\rho$ plus extrêmes, laissant ainsi la possibilité de détecter des valeurs très hautes du taux de recombinaison populationnel, pouvant refléter la présence de points chauds de recombinaison (McVean et al., 2004).

LDhelmet utilise également une distribution a priori pour le nombre et la position des évènements de recombinaison suivant une loi de Poisson de paramètres $S$ (nombre de SNP dans les données) et $\xi$ (*block penalty*). Ce dernier contrôle la variance à fine échelle de la variation dans l'estimation du paysage de recombinaison. Afin de capturer les variations à fines échelles, il a été fixé à 5. La taille des fenêtres analysées a été fixée à 50 SNPs adjacents, comme recommandé dans le manuel.

Sous un modèle neutre à l'équilibre mutation-dérive, le paramètre de mutation populationnel $\theta$, est égal à la diversité nucléotidique ($\pi$) moyenne, ici estimée à partir de l'hétérozygotie médiane des 20 individus. Cette valeur précédemment estimée par Barry et al (2020) pour chaque espèce a donc été utilisée comme le $\theta$ de chaque espèce pour réaliser les inférences du $\rho$ avec LDhelmet.

## 3- Mesure de la diversité nucléotidique locale (pi)

Estimer avec justesse la diversité nucléotidique ($\pi$), c'est-à-dire le nombre de différences moyen observé entre deux séquences d'ADN tirées de manière aléatoire parmi différents individus d'une même espèce, nécessite de prendre en compte différents facteurs :

Premièrement, les données manquantes, sous la forme de sites manquants dans le génome de référence (Ns), ou d'un génotype manquant chez un (ou plusieurs) individu à un SNP donné dans le VCF. Ces données manquantes sont souvent considérées à tort dans l'estimation de la diversité comme site non variants, et sont ajoutées dans le compte du dénominateur de la diversité ce qui la sous-estime (Korunes and Samuk, 2020)

Les sites pluri-alléliques (plus de deux allèles différents à un locus donné), souvent retirés du VCF dans l'estimation de $\pi$, participent pourtant à la diversité et doivent être pris en compte pour ne pas biaiser également biaiser vers le bas l'estimation de la diversité.

Ici, la diversité nucléotidique ($\pi$), a dans un premier temps été calculée grâce à vcftools (Danecek et al., 2011) pour *Dicentrarchus labrax.* En effet, la diversité nucléotidique calculée par différentes méthodes (vcftools et pixy (Korunes and Samuk, 2020)) a été comparé pour chaque espèce à celle obtenue par une méthode basée sur des distributions de k-mer pour ces mêmes espèces (Barry et al., 2020). Si pour le bar commun (*Dicentrarchus labrax*), la diversité mesurée par vcftools était concordante avec celle obtenue par méthode des k-mer, ce n'était pas le cas pour les autres espèces au génome plus fragmenté. Pour ces dernières, c'est un script Python issu du package scikit-allele (Miles, A. (2020). cggh/scikit-allel: v1. 3.2.) qui a été utilisé.

Ce script calcule la diversité génétique par fenêtre à partir d'un VCF comportant à la fois les sites variants et invariants. Il prend en compte les données manquantes sans les considérer comme invariables, et considère les sites pluri-alléliques. La diversité a été calculée par fenêtre de 10 000 paires de bases.

## 4- Validation de la méthode et définition de la taille des fenêtres

Pour une des espèces, le bar européen *Dicentrarchus labrax,* des données de recombinaison populationnelle étaient déjà disponibles, et ont donc permis d'évaluer la méthode déployée pour inférer les paysages de recombinaison. Ces données de recombinaison disponibles ont été obtenues avec LDhelmet (paramétrage identique) séparément dans les populations atlantiques et méditerranéennes de bar, à partir d'haplotypes entièrement phasés via une approche pedigree (séquençage de trios père-mère-enfant) (Duranton et al., 2018), considérée comme une des plus

précises (Browning and Browning, 2011). L'évaluation du paysage inféré ici chez un mélange d'individus atlantiques et méditerranéens phasés statistiquement nous permet donc de tester la robustesse de notre approche face aux éventuels problèmes liés aux erreurs de phasages et à la structure populationnelle. Des tests de corrélation entre paysages de recombinaison ont été réalisés sous différentes tailles de fenêtres (10 et 100kb), afin d'optimiser l'échelle chromosomique de nos analyses (celle qui maximise les corrélations entre paysages chez le bar). L'analyse de l'impact de la sélection en liaison sur l'érosion de la diversité neutre a ensuite été réalisée chez l'ensemble des espèces en utilisant cette taille de fenêtre.

## 5- Estimation de l'effet de la sélection en liaison

Le taux de recombinaison populationnel estimé entre deux sites polymorphes adjacents a été moyenné dans des fenêtres de 10 000 paires de bases grâce à un script R, en pondérant par la taille de l'intervalle physique entre deux SNPs. Ces valeurs ont ensuite été associées avec celles de la diversité génétique dans la même fenêtre. Le jeu de donnée à ensuite été modifié pour travailler sur des fenêtres de 100 000 paires de base en calculant la moyenne des 10 fenêtres de 10 000 pb.

La capture de l'effet de la sélection en liaison chez chacune des dix espèces a été réalisée empiriquement. Le but étant ici de comparer l'effet de la sélection en liaison de manière standardisée entre les dix espèces afin de tester d'éventuels effets des THV.

Comme la recombinaison vient moduler l'intensité de la sélection en liaison via la rupture du déséquilibre de liaison entre un locus sous sélection et les locus neutres voisins, la sélection devrait éroder davantage de diversité dans les régions à faible taux de recombinaison par rapport aux régions à fort taux de recombinaison. La diversité devrait donc être maximale dans les fenêtres avec un fort taux de recombinaison, et au contraire plus faible dans les fenêtres avec un faible taux de recombinaison.

Pour capture cet effet, les données ont donc été subdivisées en deux parties égales comprenant respectivement les 50% des valeurs les plus faibles et les plus fortes du taux de recombinaison populationnel, afin d'appliquer une régression diversité-recombinaison par intervalle de valeurs de $\rho$. En effet, en l'absence de sélection en liaison (attendu neutre), la relation entre la diversité nucléotidique et le taux de recombinaison populationnel devrait être une droite d'équation $\pi = a * \rho + b$, de pente $a = \mu/r$. Or, en présence de sélection en liaison, la diversité est d'autant plus érodée

que le taux de recombinaison est faible. Ainsi, la pente de la relation entre $\pi$ et $\rho$ pour les 50% des mesures de $\rho$ les plus faibles devrait être supérieure à celle calculée dans l'intervalle des 50% des mesures de $\rho$ les plus fortes en présence de sélection en liaison. De plus, le ratio entre les pentes du premier et du second quantile devrait refléter l'intensité de l'effet de la sélection en liaison sur l'érosion du polymorphisme dans le génome. Afin de capturer cet effet, le rapport des pentes de la régression $\pi = a * \rho + b$ entre de la première moitié (quantile 0 - 50) et de la seconde moitié (quantile 50 - 100) de la distribution des données de $\rho$ sera mesurée. Cette méthode n'a pas pour but de quantifier précisément l'effet de la sélection en liaison, mais de mesurer un effet comparable entre espèces de manière standardisée. De ce fait, ce rapport de pentes sera utilisé comme une méthode de quantification empirique de l'effet de la sélection en liaison.

## 6- Impact des traits d'histoire de vie sur l'intensité de l'érosion de la diversité nucléotidique par la sélection en liaison

Dans le but de tester l'impact des traits d'histoire de vie sur l'intensité de la sélection en liaison, les traits d'histoire de vie suivant ont été considérés :

(i) La durée de vie adulte, qui influence la variance du ratio Ne/Nc, et (ii) l'investissement parental, à travers la variable qualitative « couveur » vs « non couveur » (e.g *brooders)*. Ces deux variables ont un impact sur la diversité neutre et pourrait donc influencer également la sélection en liaison (Barry et al., 2020).

Comme les déterminants du Ne long terme ne sont pas forcément les plus pertinents pour comprendre les variations dans la sélection entre espèces (e.g la sélection agit plus rapidement que la dérive (Kimura, 1979), (iii) le niveau trophique et (iv) la taille du corps, qui reflètent possiblement une abondance plus contemporaine ont également été testés.

Enfin, (v) la fécondité a été testée puisque les stratégies reproductives ont été montrées comme ayant un impact sur la diversité génétique (Romiguier et al., 2014).

Une simple régression linéaire d'équation y = ax + b a été utilisée pour mesurer l'effet éventuel de chaque variable quantitative indépendamment sur le rapport des pentes utilisé pour capturer l'intensité de la sélection en liaison. Concernant l'effet de la variable qualitative de type

catégorique relatif à l'investissement parental, c'est une comparaison de moyenne simple qui sera utilisée grâce à un test de Student.

# III- Résultats

## 1- Filtre des VCF

Après le tri des VCFs, chaque espèce dispose d'un jeu de données de variants répartis sur un nombre de scaffolds de plus de 10 kb allant de 26 (*Dicentrarchus labrax*) à 8202 (*Symphodus cinereus*) et représentant 67% (*Sardina pilchardus*) à 97% (*Hippocampus guttulatus*) du génome de référence. Le nombre de SNPs totaux retenus pour chaque espèce s'étend de 2 775 611 pour *Hippocampus guttulatus* à 39 525 826 pour *Sarda sarda*.

| | Taille du génome (paires de base) | Nombre de SNP après filtre | Nombre de scaffolds | Fraction du génome retenu | Diversité génétique médiane | Fraction des sites htz phasés | N50 des blocks phasés (en SNPs) | Taux de recombinaison population nel médian |
|---|---|---|---|---|---|---|---|---|
| **Dlabr** | 675938161 | 9789388 | 26 | 0,83 | 0,0038 | 0,92 | 3,96 | 0,014 |
| **Hgutt** | 451180666 | 2775611 | 358 | 0,97 | 0,0029 | 0,69 | 11,7 | |
| **Lbude** | 732179437 | 17763225 | 1572 | 0,95 | 0,0023 | 0,57 | 14,2 | 0,015 |
| **Msurm** | 562799217 | 45541516 | 3403 | 0,87 | 0,0113 | 0,9 | 9,18 | 0,196 |
| **Scabr** | 655172454 | 55480114 | 1717 | 0,93 | 0,0119 | 0,96 | 16,09 | 0,017 |
| **Scant** | 781552183 | 35637188 | 5862 | 0,92 | 0,0047 | 0,79 | 13,12 | |
| **Scine** | 600619616 | 14369698 | 8243 | 0,81 | 0,0072 | 0,18 | 13,46 | 0,003 |
| **Spilc** | 949617276 | 15350454 | 8760 | 0,67 | 0,0142 | 0,84 | 7,74 | |
| **Ssard** | 738249543 | 39525826 | 4083 | 0,92 | 0,0086 | 0,68 | 15,49 | 0,011 |
| **Styph** | 338723578 | 10541299 | 1987 | 0,93 | 0,0089 | 0,18 | 22,8 | 0,044 |

**Table 1: Statistiques descriptives pour chacune des dix espèces de l'étude. Ce tableau reprend les principaux résultats descriptifs obtenus dans cette étude.**

## 2- Pré-phasage des haplotypes

Sur les dix espèces étudiées, Whatshap a permis de phaser en moyenne 67,13 % des sites hétérozygotes, avec des pourcentages par espèce allant de 18% (*Syngnathus typhle* et *Symphodus cinereus*) à plus de 90% (*Dicentrarchus labrax* et *Serranus cabrilla*) (Table 1).

Comme Whatshap utilise l'information de la position physique de SNPs présents sur une même paire de read, l'efficacité du phasage par cette méthode devrait dépendre de deux paramètres principaux : (i) le niveau de diversité génétique et (ii) le niveau de fragmentation du génome. En effet, plus la densité en SNP est haute et plus le nombre de positions hétérozygotes couvertes par une même paire de reads est susceptible d'augmenter. De plus, un génome moins fragmenté devrait permettre à Whatshap de récupérer des plus longues chaînes d'informations et d'ainsi former des blocs de phase plus longs.

Cependant, les résultats ne semblent pas refléter ces prédictions chez les espèces de l'étude. En effet, le syngnathe (*Syngnathus typhle*) et le crénilabre (*Symphodus cinereus*), qui affichent une proportion de sites hétérozygotes phasés faible (18%), présentent cependant une diversité génétique moyenne (Hétérozygotie médiane par paire de base de respectivement 0,8964 et 0,7192%) relativement à d'autres espèces (e.g. la baudroie, *Lophius budegassa*, 89 % de sites hétérozygotes phasés et hétérozygotie médiane de 0,23%), et n'ont pas un génome particulièrement fragmenté comparativement à la sardine (*Sardina pilchardus*) qui affiche une fraction de sites hétérozygotes phasés de 84,11 % et le génome le plus fragmenté (117259 *scaffolds*).

## 3- Taux de recombinaison populationnel

Afin de s'assurer de la fiabilité des paysages de recombinaison populationnelle inférés par LDhelmet, malgré les possibles erreurs de phasage statistique et l'existence d'une structure populationnelle, une comparaison a été réalisée avec des paysages de recombinaison obtenus chez *Dicentrarchus labrax* à partir d'haplotypes phasés par transmission (phasage par trios), et inférés séparément dans les populations d'Atlantique et de Méditerranée ouest (données issues de la thèse de Maud Duranton). Cette comparaison a été limitée au chromosome LG1A dont les variations du taux de recombinaison sont représentatives du génome du bar.

Les paysages de $\rho$ inférés à large échelle se sont révélés assez similaires entre les deux approches, avec des taux de recombinaison populationnelle plus élevés aux extrémités comparés au centre du chromosome. Les niveaux de corrélations entre les taux de recombinaison populationnels inférés ici et ceux obtenus à l'échelle intra-populationnelle à partir de données phasées par approche trio sont relativement bons à la fois à fine échelle (fenêtres de 10 kb : $r^2$ de Pearson = 0,35, p-value < 2,2e-16 avec MEDW ; $r^2$ = 0,28, p-value < 2,2e-16 avec ATL) et à large échelle (fenêtres 100 kb : $r^2$ = 0,75, p-value < 2,2e-16 avec MEDW ; $r^2$ = 0,44, p-value = 1,851e-15 avec MEDW). L'amélioration des corrélations obtenues en utilisant des fenêtres de 100 kb nous a conduit à poursuivre les analyses en utilisant des fenêtres de 100 kb, et donc à ne pas considérer les *scaffolds* de moins de 100 kb par la suite.

Ces comparaisons suggèrent que (i) le mélange de deux populations différenciées dans notre analyse ne cause pas de biais d'estimation des paysages de recombinaison et que notre approche est donc robuste à l'existence d'une structure populationnelle. (ii) L'inférence statistique de la phase ne semble pas perturber l'estimation des paysages de recombinaison populationnelle malgré les possibles erreurs de phasage (e,g inférence de faux événements de recombinaison, (O'Connell et al., 2014)).

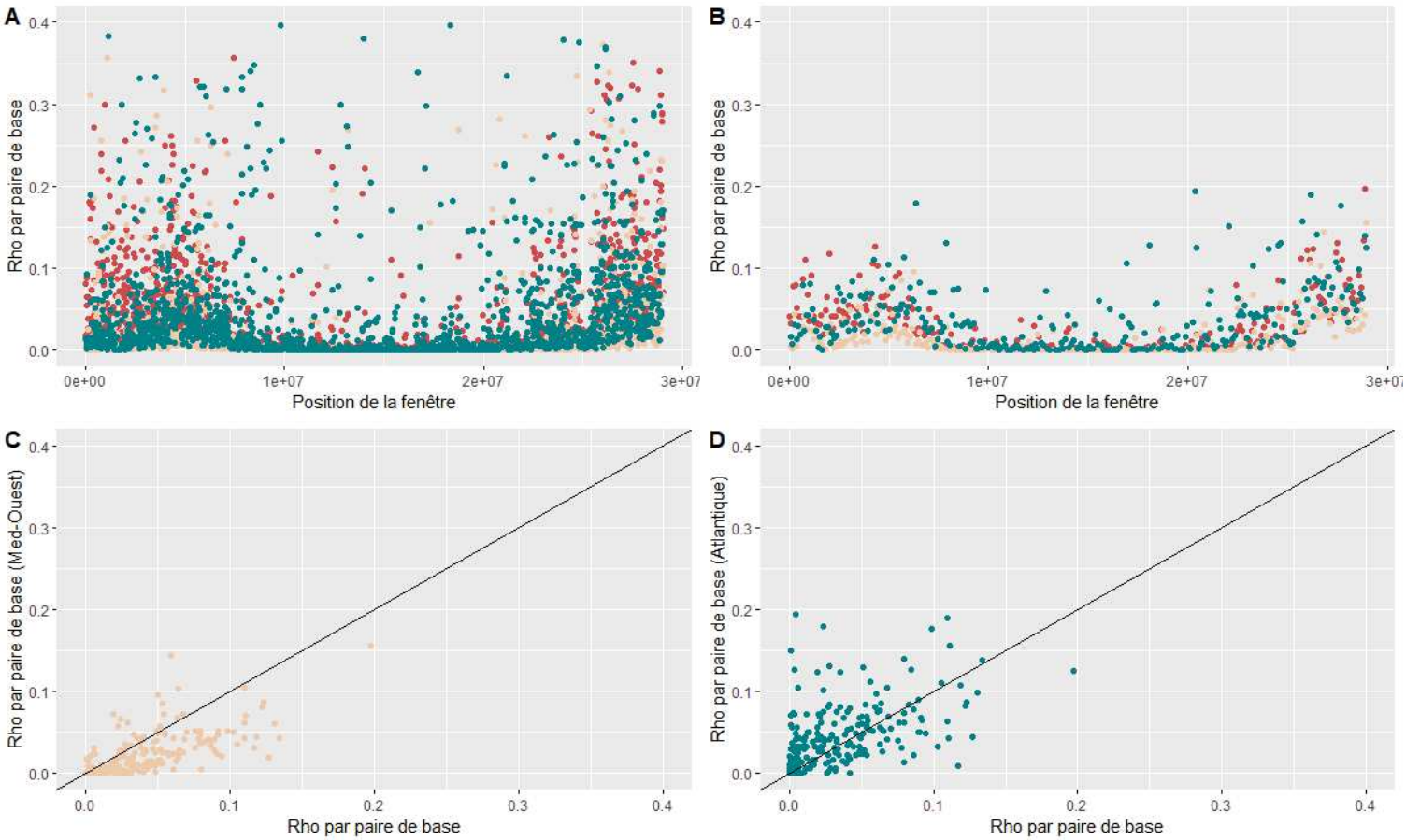



**Figure 1 : Comparaison des inférences du paysage de recombinaison le long du chromosome LG1A chez *Dicentrarchus labrax.* Pour A- et B-, les estimations de $\rho$ obtenues via la méthode déployée dans cette étude sont représentées par des points rouges. Les estimations de $\rho$ obtenues à partir de données phasées par trio pour la population Méditerranée-Ouest et Atlantique sont représentées respectivement en beige et en bleu et par fenêtre génomique de 10 kb (A) et 100 kb (B). Corrélation par fenêtre de 100 kb entre les inférences du taux de recombinaison populationnel du bar commun avec (C) les données Méditerranée-Ouest et (D) les données Atlantique.**

## 4- Estimation de l'effet de la sélection en liaison :

Suite à une limite de temps dû au temps de calcul du programme LDhelmet, l'inférence du taux de recombinaison populationnel a été faite pour les plus gros *scaffolds* de chaque espèce uniquement.

Pour toutes les fenêtres de 100kb se trouvant sur un même *scaffold*, la relation entre la diversité génétique et le taux de recombinaison populationnel a été reporté sur un graphique (Figure 2). Le nombre de fenêtres sur lesquels la diversité génétique et le taux de recombinaison populationnel a été calculé allaient de 125 pour *Symphodus cinereus* à 5480 pour *Dicentrarchus labrax* et représente assez bien le niveau de fragmentation du génome.

| | Moitié 0-50 | | | Moitié 50-100 | | | Rapport |
|---|---|---|---|---|---|---|---|
| | **Pente** | **r²** | **P-value** | **Pente** | **r²** | **P-value** | |
| **Dlabr** | 8.91 | 0.15 | <2.2e-16 | 1.37 | 0.24 | <2.2e-16 | 6.50 |
| **Hgutt** | 10.84 | 0.06 | <2.2e-16 | 0.68 | 0.01 | 3.95E-06 | 15.94 |
| **Lbude** | 4.44 | 0.16 | <2.2e-16 | 1.20 | 0.23 | <2.2e-16 | 3.70 |
| **Msurm** | 1.82 | 0.14 | 1.15E-09 | 1.38 | 0.14 | 1.90E-09 | 1.32 |
| **Scabr** | 19.56 | 0.06184 | 4.19E-07 | 4.73 | 0.23 | <2.2e-16 | 4.14 |
| **Scant** | 39.37439 | 0.24 | <2.2e-16 | 8.85 | 0.33 | <2.2e-16 | 4.45 |
| **Scine** | 77.95 | 0.27 | 1.89E-05 | 6.22 | 0.19 | 0.0004 | 12.53 |
| **Spilc** | 4.81 | 0.06 | 0.006 | -0.16 | <0.01 | NS (0.652) | -30.06 |
| **Ssard** | 30.49 | 0.17 | 3.39E-13 | 8.78 | 0.05 | 7.86E-05 | 3.47 |
| **Styph** | 7.34569 | 0.09426 | 6.17E-12 | 6.62 | 0.24 | <2.2e-16 | 1.11 |

**Table 2 : Valeurs des pentes de la première et de la deuxième moitié des valeurs de $\rho$, ainsi que le rapport entre les deux, pour chacun des jeux de données de la distribution de la diversité nucléotidique fonction du taux de recombinaison populationnel. Le r² représente la partie de la variance dans la diversité expliquée par les variations du taux de recombinaison populationnel. Les *p-value* non significatives sont libellées NS.**

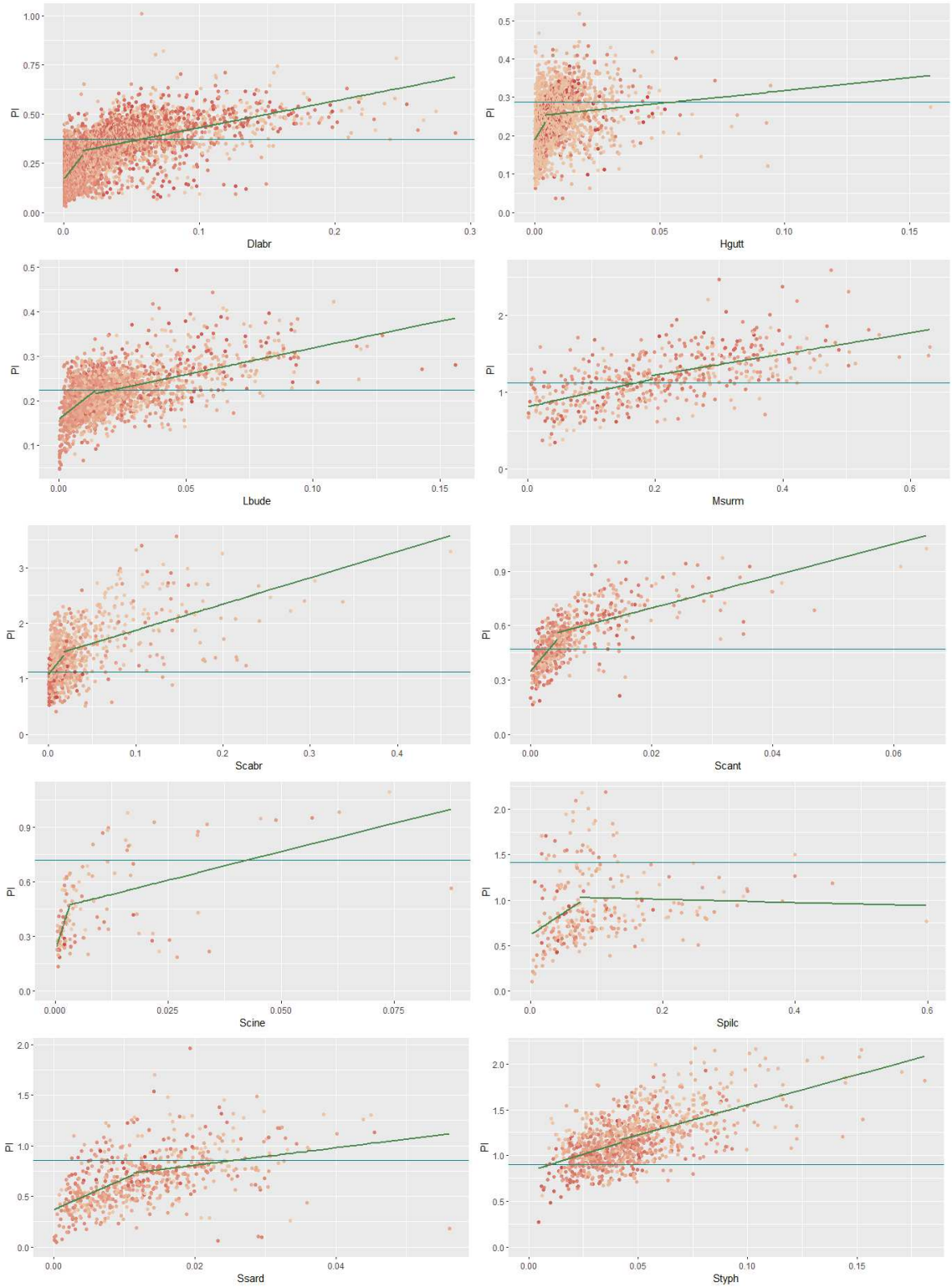



**Figure 2 : Relation entre la diversité nucléotidique et le taux de recombinaison populationnel chez chacune des dix espèces de l'étude. La droite horizontale bleue représente l'hétérozygotie médiane à l'échelle du génome entier calculée par méthode des *k*-mers. Les pentes dans le premier et le dernier quartile sont représentées en vert. Le dégradé de couleur des points correspond à des chromosomes différents.**

Toutes les espèces exceptée *Sardina pilchardus* affichent une relation positive dans la première et la deuxième moitié des données de $\rho$ entre le niveau de diversité génétique et le taux de recombinaison populationnel, ce qui est concordant avec la littérature (Hasan and Ness, 2020; Vijay et al., 2017).

*Sardina pilcardus* présente une pente dans la deuxième moitié des données inférieure à zéro, mais cette relation n'est pas significative. Pour ne pas impacter faussement les interprétations, cette espèce ne sera pas prise en compte dans la suite de l'étude.

A partir de ces relations, le rapport de la pente de la première moitié des données par rapport à la pente de la deuxième moitié des données s'étend d'environ 1 chez *Syngnathus typhe* et *Mullus surmuletus* à plus de 10 chez *Hippocampus guttulatus et Symphodus cinereus* (Table 2).

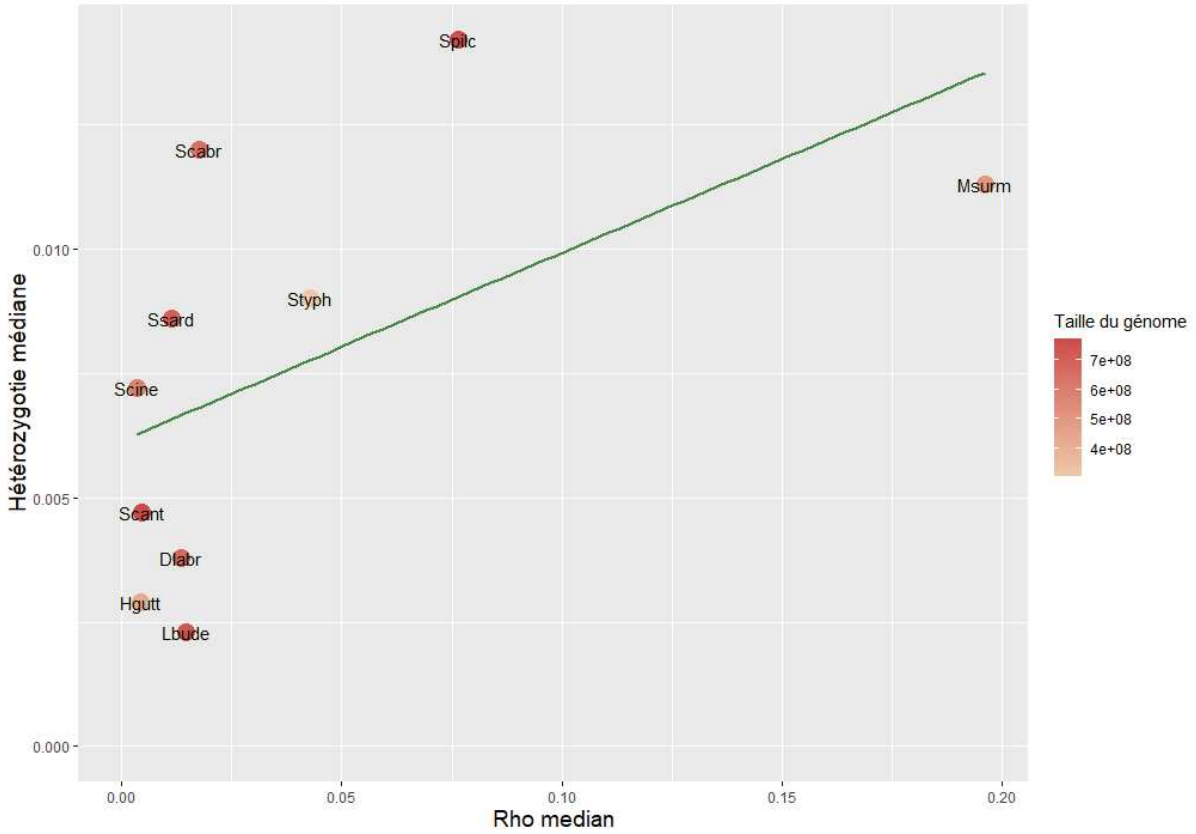



**Figure 3 : Hétérozygotie médiane en fonction du $\rho$ médian inféré par LDhelmet. Chaque point représente une espèce, et le gradient de couleurs représente la variabilité des tailles de génome, les plus foncés représentant les génomes les plus grands.**

Le $\rho$ médian inféré par espèce s'étend de 0.003 pour *Symphodus cinereus* à 0.1961 pour *Mullus surmuletus (*Table 1). Le facteur de variation est de 65, ce qui semble déjà grand pour une étude portant sur des taxons aussi proches.

La relation $\pi$ - $\rho$ inter entre espèces a été étudiée mais semble assez peu linéaire (Figure 3). *Mullus surmuletus* se détache de la tendance en affichant une valeur élevée du taux de recombinaison populationnel. C'est également une des espèces qui présente la taille du génome la plus petite comparativement aux autres (Table 1). *Sardina pilchardus* affiche quant à elle un niveau de diversité génétique très haut.

## 5- Impact des traits d'histoire de vie sur l'intensité de l'érosion de la diversité nucléotidique par la sélection en liaison

Aucun des traits d'histoire de vie testé n'a montré d'effet significatif sur le rapport des pentes utilisé pour mesurer l'intensité de la sélection en liaison.

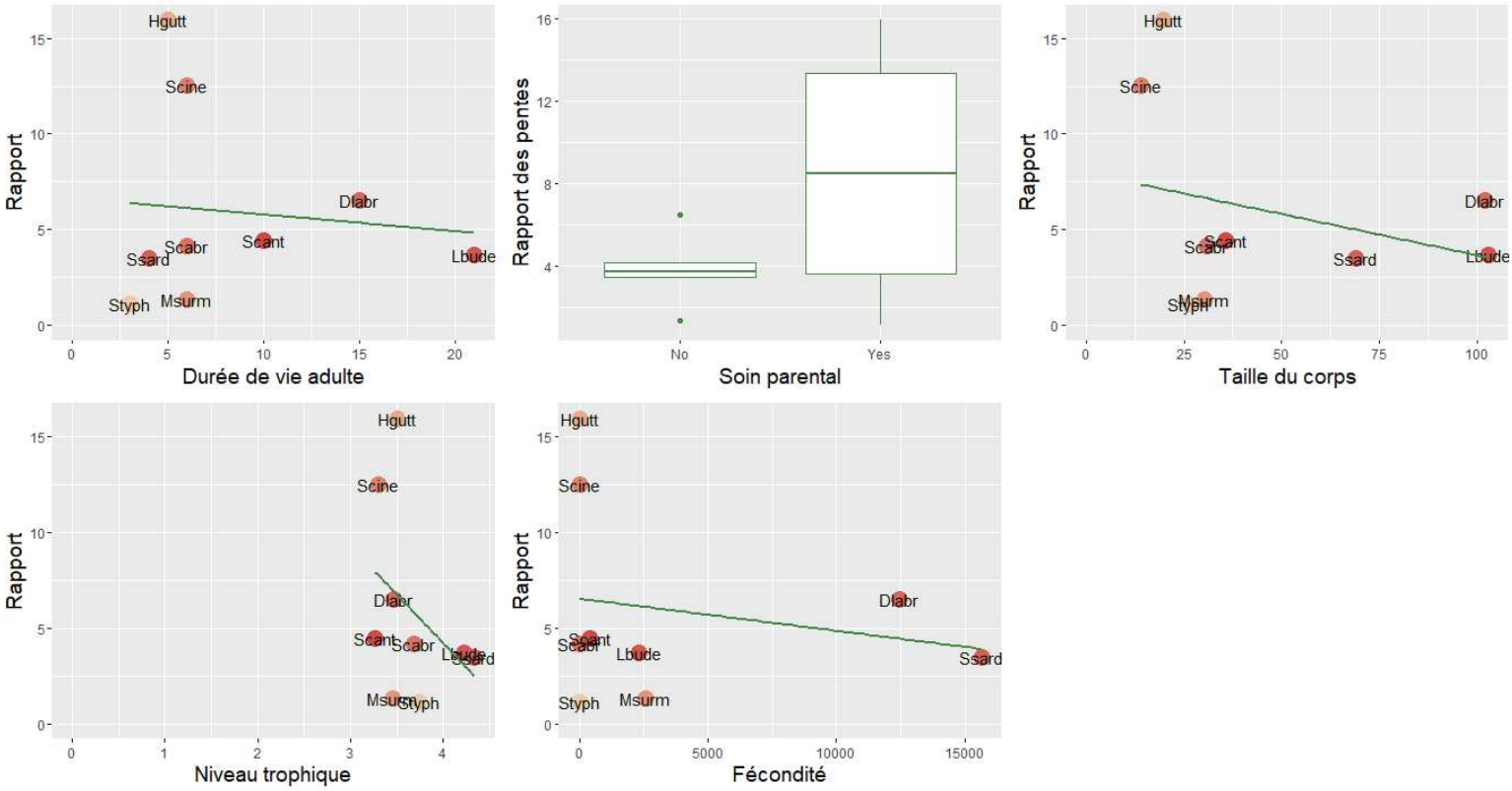



**Figure 4 : Rapport entre les deux pentes (e.g. intensité de la sélection en liaison intraspécifique) en fonction des traits d'histoire de vie testés dans cette étude. Chaque point représente une espèce, et le gradient de couleur représente la variabilité des tailles de génome, les plus foncés représentant les génomes les plus grands.**

Les pentes des régressions linéaires ainsi que les p-value associées aux modèles ont été reportés (Table 3).

Bien que non significatives, toutes les relations apparaissent légèrement négatives. Pour la taille du corps et le niveau trophique, proxys courants de la taille efficace *Ne*, ces tendances semblent concordantes avec les attendus (Buffalo et al, 2021 ; Corbett-Detig et al., 2015). Il en est de même pour la relation avec la durée de vie adulte. En effet, une durée de vie plus longue reflète un ratio entre taille efficace *Ne* et abondance contemporaine *Nc* plus faible, et on devrait observer un effet moins fort de la sélection en liaison chez les espèces les plus longévives. La tendance de la relation avec la fécondité irait en revanche contre les attendus, puisque des études passées ont montré que les espèces à stratégie *r* affichaient un niveau plus élevé de diversité. On aurait donc pu s'attendre à observer un effet supérieur de la sélection en liaison sur l'érosion de la diversité chez les espèces avec une forte fécondité.

| Traits de vie | Effet sur la pente | p-value |
|---|---|---|
| Durée de vie adulte | -0.09 | NS(0.79) |
| Niveau trophique | -5.09 | NS(0.305) |
| Taille du corps | -0.04 | NS(0.441) |
| Fécondité | -0.0002 | NS(0.608) |

| | Test-t de Student | |
|---|---|---|
| Investissement parental | Valeur de la moyenne | p-value |
| Oui | 3.83 | 0.2697 |
| Non | 8.51 | |

**Table 3: Résultats statistiques des tests de l'effets des traits d'histoire de vie sur le rapport des pentes. Aucuns des résultats n'apparaît comme significatif. A gauche, les résultats des modèles de régression linéaire, et à droite, le résultat de la comparaison de moyenne.**

# IV- Discussion

Dans cette étude, les paysages de variation de la diversité génétique et du taux de recombinaison populationnel ainsi que de leurs covariations ont été étudiés chez dix espèces de poissons marins afin d'évaluer l'impact de la sélection en liaison sur l'érosion du polymorphisme. L'objectif était d'identifier les déterminants écologiques et biologiques des différences d'intensité de la sélection en liaison entre espèces avec des caractéristiques différentes (e.g traits d'histoire de vie), grâce à une comparaison multi-espèces. Les aspects méthodologiques liés à l'inférence des paysages de recombinaison et les interprétations possibles des résultats entre espèces seront ici discutés.

# 1- Inférence des paysages de recombinaison populationnelle : Limites et améliorations

Chez de nombreuses espèces, il est encore impossible d'effectuer des croisements et donc d'avoir accès à des cartes de liaison pour estimer la recombinaison (Peñalba and Wolf, 2020). Les approches populationnelles de la détermination du déséquilibre de liaison permettent cependant de reconstituer les paysages de recombinaison chez n'importe quelle espèce pour laquelle des données de polymorphisme génomique sont disponibles.

Malgré l'existence de plusieurs méthodes d'inférence du taux de recombinaison populationnel (e.g LDhat (Auton and McVean, 2007), LDhelmet (Chan et al., 2012), les études qui utilisent ces paysages inférés restent encore rares dans la littérature et limitées à quelques études de cas basés sur une seule espèce (Vijay et al., 2017).

Une des explications possibles est la méthodologie de l'analyse, qui nécessite en amont des étapes difficiles à mettre en œuvre. La première est la détermination de la phase des haplotypes dans les données populationnelles.

Les approches de phasage statistiques exploitent en effet le signal de déséquilibre de liaison dans les données populationnelles (O'Connell et al., 2014; Delaneau et al., 2019). Si une erreur de phasage survient, il y a un risque que cette erreur se répercute dans les inférences des paysages de recombinaison populationnels. Les modèles statistiques sur lesquels se basent ces méthodes sont souvent des modèles simples, sensibles aux écarts dans les données réelles, notamment avec une difficulté dans la gestion des variants rares, ainsi que des possibles problèmes en présence de populations structurés (Browning and Browning, 2011).

Dans la méthode déployée ici, l'ajout aux deux approches populationnelles (SHAPEIT4 et LDhelmet) d'une étape préliminaire de pré-phasage par le logiciel Whatshap, qui fonctionne sans a priori sur les attendus et qui se base uniquement sur la phase physique des paires de reads, devrait donc permettre d'améliorer la phase des haplotypes sans avoir recours à des méthodes comme LDpop qui utilise la même information présente dans les déséquilibres de liaisons populationnels et qui nécessite également un a priori sur les fluctuations démographiques passées (Kamm et al., 2016).

La comparaison des paysages de recombinaison obtenus à la suite de cette méthode avec ceux issus d'inférences obtenus à partir de données phasées par trio a montré des résultats qualitativement et quantitativement proches (Figure 1). Cela permet de s'assurer de la robustesse des résultats malgré la structure de notre population test de bars, composée à la fois d'individus provenant de l'océan Atlantique et de la mer Méditerranée. Utiliser le phasage physique en amont a permis de limiter les erreurs d'inférence au niveau du phasage avec des répercussions positives au niveau des paysages de recombinaison (Booker et al., 2017). Ajouter cette étape de pré-phasage par Whatshap dans un pipeline bio-informatique est relativement peu coûteux, et rapide à mettre en place. Comparativement, le phasage par trio est à la fois onéreux en budget et en temps (besoin de réaliser des croisements), et n'est pas réalisable chez toutes les espèces.

Même si l'approche empirique semble valider cette méthode, effectuer des simulations de données de séquençage permettraient d'aller plus loin en comparant l'inférence des taux de recombinaison avec et sans phasage physique, et serait un bon moyen de quantifier l'amélioration de la méthode d'inférence. Ce type de simulations est cependant coûteux (besoin de simuler des données de types *reads,* pour prendre en considération les erreurs au niveau des zones d'assemblages de *reads* et la variation dans la couverture génomique et pourrait être l'objet d'une étude complémentaire.

L'efficacité du pré-phasage effectué par Whatshap semble peu sensible à la qualité des données. La fraction de sites phasés et la longueur moyenne des blocs de phase ne semblent pas dépendre du niveau de diversité et de la fragmentation du génome. Des espèces comme *Spondyliosoma cantharus*, avec une faible hétérozygotie relativement aux autres espèces, arrivent à de très bonnes proportions de sites hétérozygotes phasés (Plus de 75%) (Figure 5). De même, une espèce comme *Sardina pilchardus*, avec un génome très fragmenté, arrive à une quantité de sites phasés de 84%. Là où il est déjà possible d'obtenir des segments de plusieurs kilobases (e.g. Pacbio, Oxford nanopore), et avec la démocratisation du séquençage *longread* qui va s'intensifier, l'utilisation de logiciels comme Whatshap utilisant l'information physique présente sur ces *reads* pourraient bientôt permettre le phasage d'haplotypes entiers et à terme permettre de s'émanciper des inférences statistiques et des potentiels biais les accompagnant (Booker et al., 2017).

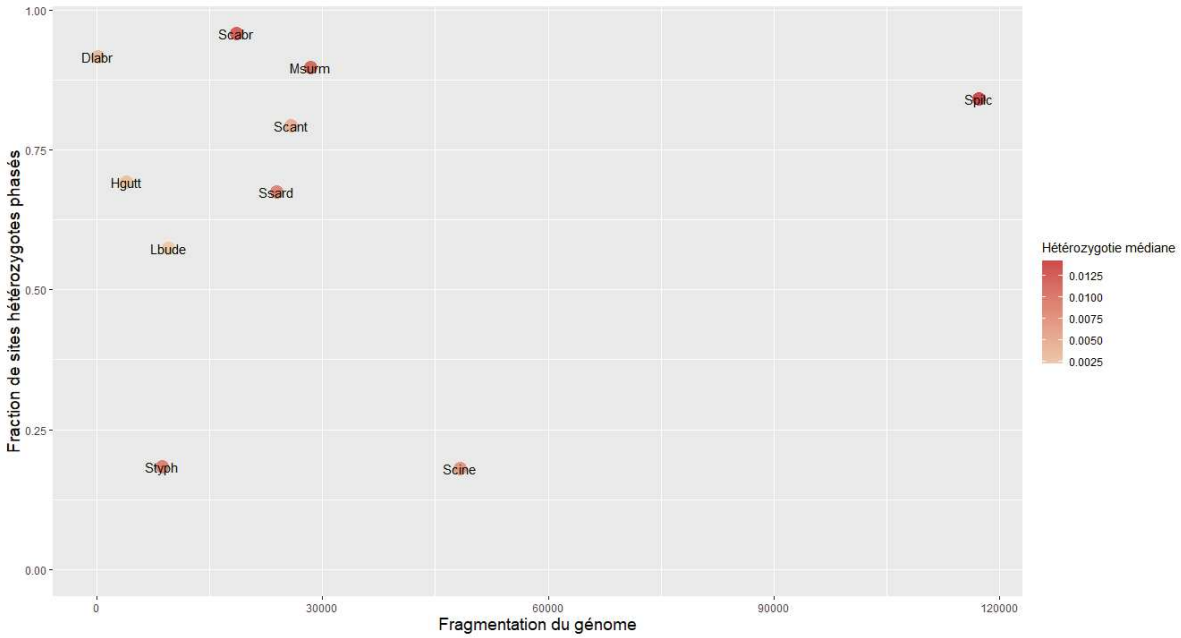



**Figure 5: Fraction des sites hétérozygotes phasés en fonction de deux facteurs : la fragmentation du génome ainsi que le niveau de diversité. Chaque point représente une espèce, et la couleur des points représente le niveau d'hétérozygotie de l'espèce associée. Les points clairs correspondent aux espèces avec le moins de diversité comparativement aux points les plus foncés.**

## 2- Variation interspécifique du $\rho$ inter espèces

Peu de méthodes comparatives standardisés ont été réalisées entre espèces différentes dans le but de mesurer les variations dans le taux de recombinaison populationnel, et concernent souvent des espèces proches (Shen et al., 2019; Stukenbrock and Dutheil, 2018), conférant peu de recul quant aux variations de $\rho$ entre espèces.

Théoriquement, comme $\pi = 4\ Ne\ \mu$ et $\rho = 4\ Ne\ r$, on s'attend à une relation linéaire entre ces deux valeurs, modulé par le rapport $\mu/r$, et influencé également par *Ne* (e.g. la taille efficace).

Si toutes les espèces partageaient le même rapport $\mu/r$, la distribution de la diversité génétique médiane en fonction du taux de recombinaison populationnel médian entre espèces devrait être linéaire. Cependant, ce n'est pas ce qui est observé (Figure 3) : *Sardina pilchardus* dispose d'un rho médian environ dix fois plus élevée que *Hippocampus guttulutus*, et il existe un facteur de 65 entre *Mullus surmuletus* et *Symphodus cinereus* ($\rho$ médian respectifs de 0.196 et 0.0034). Ces résultats semblent indiquer qu'il existe un rapport $\mu/r$ différent d'une espèce à l'autre (Figure 3),

ce qui devrait avoir un impact direct sur la forme de la relation. Cette différence inter espèces dans le ratio $\mu/r$ a un impact direct sur la forme de la relation entre $\pi$ et $\rho$.

Les variations du taux de mutation $(\mu)$ sont encore difficilement accessibles chez les eucaryotes (Baer et al., 2007) et donc difficiles à appréhender. Les variations dans le taux de recombinaison populationnel $\rho$ pourraient être le reflet de l'architecture du génome : en effet, bien que le caryotype soit stable entre les différentes espèces de poissons (Mank and Avise, 2006), la longueur du génome est moins stable d'une espèce à l'autre (338 Mégabases pour *Hippocampus guttulatus* et 949 Mégabases pour *Sardina pilchardus*, soit un rapport de trois (Table 1)).

Lorsque la taille du génome est plus faible, chaque chromosome et alors plus court. Le nombre de *crossing-over* par chromosome est en revanche contraint à un seul évènement de recombinaison par méiose et par chromosome : c'est ce que l'on appelle l'interférence chromosomique (Hillers, 2004). De ce fait, les espèces ayant un génome de plus petite taille, à nombre de chromosomes équivalent, afficheront également une densité en événement recombinationnel supérieure aboutissant à plus de recombinaison dans les génomes plus condensés. Ici, cet effet pourrait s'observer chez *Syngnathus typhle, Hippocampus guttulatus*, et aussi *Mullus surmuletus*, affichant tout trois un génome de taille inférieur (Table 1). Le rouget présente d'ailleurs un $\rho$ médian particulièrement élevé, qui pourrait en partie s'expliquer par cette contrainte architecturale. La taille du génome pourrait donc également avoir un impact sur le niveau de diversité, de par sa relation avec la recombinaison.

Une étude basée sur l'étude des taux de recombinaison estimés à partir de cartes génétiques avait également trouvé une grande variance chez les poissons, conformément à nos résultats (Stapley et al., 2017).

De fortes contraintes architecturales semblent agir sur ces mécanismes, et les différences dans l'estimation du $\rho$ auront un impact direct sur l'estimation de la sélection en liaison. Ces résultats appuient l'importance de la prise en compte de l'architecture génomique dans l'estimation de la sélection en liaison (Buffalo et al., 2021).

## 3- Erosion du polymorphisme par la sélection en liaison

La majeure difficulté de l'étude provient de la volonté d'estimer les variations intragénomiques du $\rho$ et non du $r$. Or, la sélection en liaison dépend des variations du taux de recombinaison local $r$, et

non directement de $\rho$ ($4 * Ne * r$). Une question vient alors naturellement : Est-il possible d'évaluer l'efficacité de la sélection en liaison à travers la comparaison entre $\pi$ et $\rho$ ?

$\pi$ et $\rho$ sont reliés théoriquement par le ratio $\mu/r$. De par la dépendance des deux mesures à la taille efficace *Ne*, on s'attend de fait à une covariance entre la diversité génétique neutre $\pi$ et le taux de recombinaison populationnelle $\rho$ due à la réponse commune aux variations stochastiques du *Ne*. Sous un modèle neutre, à $\mu$ constant, les variations génomiques du $\pi$ sont censées être reliés aux variations du $\rho$ par la pente $\mu/r$, car les deux quantités sont localement influencées par les mêmes variations du *Ne* dues à la stochasticité de la coalescence. Une covariance (e.g. une relation linéaire) entre la diversité génétique neutre $\pi$ et le taux de recombinaison populationnelle $\rho$ est de fait attendu.

Cet attendu neutre pourrait être vérifié par des approches de simulations en utilisant par exemple Msprime pour générer des données génomiques sans sélection en liaison et tester la covariation de ces deux valeurs**,** et servir à terme de modèle neutre et de point de comparaison pour de futurs études comparatives (Kelleher et al., 2016).

Sous l'effet de la sélection en liaison, la perte de diversité devrait être encore plus grande en présence d'un petit taux de recombinaison $\rho$. Cette érosion supérieure en région de faible $\rho$ devrait être à l'origine d'une modification de la pente entre ce dernier et la diversité nucléotidique. En intraspécifique, il semble que ce soit la sélection en liaison qui impact la relation entre $\pi$ et $\rho$ le long du génome.

La comparaison des pentes entre les deux moitiés des données a chez toutes les espèces donné un rapport positif qui semble capturer empiriquement un effet de la sélection en liaison intra-spécifique. Il est intéressant de constater que l'amplitude de variation de la seconde pente est inférieure à celle de la première (facteurs de dix et vingt), et une majorité des pentes tourne autour de 1 (Table 2) ce qui pourrait correspondre à l'attendu neutre. Les différences dans le rapport de pente a permis de capturer une intensité de la sélection sur l'érosion de la diversité variables entre espèces (faible pour *Syngnathus typhle* et *Mullus surmuletus* (rapport proche de 1) et à l'inverse plus fort pour *Hippocampus guttulatus* et *Symphodus cinereus* (rapport supérieur à 10). L'amplitude mesurée au sein des dix espèces de l'étude couvre déjà un facteur 15, et amène à se questionner sur les possibles déterminants de cette amplitude entre espèces proches.

## 4- Déterminants de la sélection en liaison entre espèces

L'effet de la sélection en liaison a été avancé pour expliquer le paradoxe de Lewontin (Roberts, 2015) avec une érosion du polymorphisme plus prononcée chez les organismes de grande taille efficace (Corbett-Detig et al., 2015). Comme le $\pi$ est un bon prédicteur de l'abondance *Ne* long terme (calculée comme la moyenne harmonique du *Ne*, avec un poids supérieur accordé aux périodes de plus faible taille efficace), la relation entre le rapport des pentes et la diversité nucléotidique devrait permettre de voir si la diversité nucléotidique est comme attendu un bon prédicteur de l'effet de la sélection en liaison. Comme la sélection opère cependant plus vite que la dérive (Kimura, 1979) la sélection en liaison devrait être majoritairement impactée par l'abondance récente, et pourrait donc être davantage capturée par certains traits d'histoire de vie, comme la taille du corps ou le niveau trophique, reflétant un *Ne* plus récent que celui reflété par le $\pi$. De plus, comme on sait que la durée de vie adulte est un bon prédicteur de la diversité génétique chez les poissons (Barry et al., 2020), il est légitime ici de tester son effet sur l'intensité de la sélection en liaison.

Cependant, même des traits d'histoire de vie comme le niveau trophique ou la taille du corps peuvent ne pas être adaptés, car ils reflètent un *Ne* encore trop ancien par rapport à l'abondance contemporaine. Une comparaison de l'intensité de la sélection en liaison avec l'abondance contemporaine aurait été intéressante et a été envisagée, cependant, bien que l'accès à l'abondance aurait pu être possible pour des espèces pêchées (e.g *Sardina pilchardus* ou *Dicentrarchus labrax*), estimer l'abondance d'espèces côtières est bien plus compliqué. Des tentatives d'approximation existent à partir de données combinant la densité des espèces avec leurs aire de répartition (Buffalo, 2021) mais n'a pas été réalisé ici car ne reflète pas la réalité du terrain pour les espèces de l'étude : en effet, les microhabitats disponibles sont présents de manière hétérogène dans le milieu (e.g herbiers à zostères naines utilisées par les syngnathes).

Bien que non significatives, les tendances vont toutes dans le sens des prédictions, sauf pour la fécondité. Le manque de significativité peut provenir du faible nombre de points (neuf retenus), mais semble encourageant, et appelle à des études supplémentaires pour étoffer la relation.

Les résultats ont été obtenues en toute fin de stage, et plus de recul serait nécessaire quant à l'analyse des résultats, pour délier les possibles effets combinés de différents facteurs comme la taille du génome, le taux de recombinaison moyen, ou encore la variance dans le taux de

recombinaison populationnel, qui peuvent interférer de manière complexe avec l'effet des traits d'histoire de vie sur l'intensité de la sélection en liaison.

# Conclusion :

A une époque où les données de diversité génétique sont de plus en plus utilisées en génomique de la conservation et dans la gestion des espèces, avec des espèces affichant des niveaux de diversité supérieurs considérées comme moins menacées, force est de constater que les déterminants de cette diversité sont encore peu connus (Ellegren and Galtier, 2016).

Chez les poissons, l'effet de traits d'histoire de vie influençant le *Ne* long terme a été mis en évidence (Barry et al., 2020), avec notamment un fort effet de la durée de vie adulte comme médiateur de la variance dans le succès reproducteur. Cette étude, bien qu'empirique, vient compléter ces connaissances en mettant en lumière un effet de la sélection en liaison qui varie d'un facteur 15 entre espèces, alors que la diversité génétique ne varie que d'un facteur 7. C'est un début, et d'autres études permettront l'ajout de points supplémentaires issus de groupes taxonomiques différents pour affiner la relation entre la diversité nucléotidique $\pi$ et le taux de recombinaison populationnel $\rho$, ainsi que la prise en compte d'autres traits de vie. De plus, des avancées méthodologiques comme un modèle intégrant d'autres variables (comme la densité d'éléments fonctionnels, ou la distance aux sites sous sélection) et modélisant la distribution conjointe de $\pi$ et $\rho$ permettrait d'aboutir à des valeurs quantitatives plus précises de l'effet de la sélection en liaison. Ces améliorations pourront ainsi permettre d'apporter des réponses quant aux déterminants de la sélection en liaison.

# Bibliographie

Auton, A., McVean, G., 2007. Recombination rate estimation in the presence of hotspots. Genome Res. 17, 1219–1227. https://doi.org/10.1101/gr.6386707

Baer, C.F., Miyamoto, M.M., Denver, D.R., 2007. Mutation rate variation in multicellular eukaryotes: causes and consequences. Nat. Rev. Genet. 8, 619–631. https://doi.org/10.1038/nrg2158

Barry, P., Broquet, T., Gagnaire, P.-A., 2020. Life tables shape genetic diversity in marine fishes. bioRxiv. *** Mise en évidence d'effets de trait d'histoire de vie sur la variation de la diversité génétique neutre, avec notamment un fort impact de la durée de vie à travers le ratio *Ne/Nc* chez les poissons marins.**

Begun, D.J., Aquadro, C.F., 1992. Levels of naturally occurring DNA polymorphism correlate with recombination rates in D. melanogaster. Nature 356, 519–520. https://doi.org/10.1038/356519a0
Booker, Ness, Keightley, 2017. The Recombination Landscape in Wild House Mice Inferred Using Population Genomic Data. Genetics 207.
Browning, S.R., Browning, B.L., 2011. Haplotype phasing: existing methods and new developments. Nat. Rev. Genet. 12, 703–714. https://doi.org/10.1038/nrg3054
Buffalo, V., 2021. Why do species get a thin slice of π? Revisiting Lewontin's Paradox of Variation | bioRxiv. BioRxiv.***Une méta-analyse récente sur le paradoxe de Lewontin qui met en évidence l'impact de la longueur de la carte de recombinaison sur l'effet de la sélection en liaison.**
Chan, Jenkins, Song, 2012. Genome-Wide Fine-Scale Recombination Rate Variation in Drosophila melanogaster. PLOS Genet. 8.
Charlesworth, B., 2009. Effective population size and patterns of molecular evolution and variation. Nat. Rev. Genet. 10, 195–205. https://doi.org/10.1038/nrg2526
Charlesworth, B., Morgan, M.T., Charlesworth, D., 1993. The Effect of Deleterious Mutations on Neutral Molecular Variation 15.
Chen, J., Glémin, S., Lascoux, M., 2017. Genetic Diversity and the Efficacy of Purifying Selection across Plant and Animal Species. Mol. Biol. Evol. 34, 1417–1428. https://doi.org/10.1093/molbev/msx088
Comeron, J., 2014. Background Selection as Baseline for Nucleotide Variation across the Drosophila Genome. PLOS Genet. 10.
Coop, G., 2016. Does linked selection explain the narrow range of genetic diversity across species? https://doi.org/10.1101/042598
Corbett-Detig, R.B., Hartl, D.L., Sackton, T.B., 2015. Natural Selection Constrains Neutral Diversity across A Wide Range of Species. PLOS Biol. 13, e1002112. https://doi.org/10.1371/journal.pbio.1002112 ***Première méta-analyse de l'effet de la sélection en liaison mettant en évidence une érosion plus marqu"e de la diversité chez les espèces les plus abondantes.**
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, 2156–2158. https://doi.org/10.1093/bioinformatics/btr330
Delaneau, O., Zagury, J.-F., Robinson, Marchini, J., Dermitzakis, 2019. Accurate, scalable and integrative haplotype estimation | Nature Communications. Nat. Commun. 10.
Duranton, M., Allal, Fraisse, Bierne, N., Gagnaire, P.-A., 2018. The origin and remolding of genomic islands of differentiation in the European sea bass | Nature Communications. Nat. Commun. 9.
Ellegren, H., Galtier, N., 2016. Determinants of genetic diversity. Nat. Rev. Genet. 17, 422–433. https://doi.org/10.1038/nrg.2016.58
Elyashiv, E., Sattath, S., Hu, T.T., Strutsovsky, A., McVicker, G., Andolfatto, P., Coop, G., Sella, G., 2016. A Genomic Map of the Effects of Linked Selection in Drosophila. PLOS Genet. 12, e1006130. https://doi.org/10.1371/journal.pgen.1006130 ***Etude de la sélection en liaison chez la drosophile à travers une approche méthodologique explicitant l'effet de la sélection positive et négative ainsi que la distance aux gènes.**
Gante, Matschiner, Jakobsen, Salzburger, 2016. Genomics of speciation and introgression in Princess cichlid fishes from Lake Tanganyika. Mol. Ecol. 25.

Haenel, Q., Roesti, M., Laurentino, T.G., Berner, D., 2018. Meta-analysis of chromospme-scale crossover rate variation in eukaryotes and its significance to evolutionnary genomics. Mol. Ecol. 27, 2477–2497.
Hasan, A.R., Ness, R.W., 2020. Recombination Rate Variation and Infrequent Sex Influence Genetic Diversity in Chlamydomonas reinhardtii. Genome Biol. Evol. 12, 370–380. https://doi.org/10.1093/gbe/evaa057
Hillers, K., 2004. Crossover interference. Curr. Biol. 14.
Josephs, E.B., Wright, S.I., 2016. On the Trail of Linked Selection. PLOS Genet. 12, e1006240. https://doi.org/10.1371/journal.pgen.1006240
Kamm, Spence, Song, 2016. Two-Locus Likelihoods Under Variable Population Size and Fine-Scale Recombination Rate Estimation | Genetics | Oxford Academic. Genetics 203.
Keightley, Jackson, 2018. Inferring the Probability of the Derived versus the Ancestral Allelic State at a Polymorphic Site. Genetics 209.
Kelleher, Etheridge, Mcvean, 2016. Efficient Coalescent Simulation and Genealogical Analysis for Large Sample Sizes. Plos Comput. Biol. 10.
Kimura, M., 1979. The Neutral Theory of Molecular Evolution. Sci. Am. 35.
Kimura, M., Crow, J.F., 1964. The number of allele that can be maintained in a finite population.
Korunes, K.L., Samuk, K., 2020. Unbiased estimation of nucleotide diversity and divergence in the presence of missing data. https://doi.org/10.1101/2020.06.27.175091
Lewontin, R.C., 1974. The genetic basis of evolutionnary change. Columbia University Press.
Li, N., Stephens, M., 2003. Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. Genetics 165, 2213–2233.
Lynch, M., 2007. THE ORIGINS OF GENOME ARCHITECTURE.
Mank, Avise, 2006. Phylogenetic conservation of chromosome numbers in Actinopterygiian fishes | SpringerLink. Genetica 127, 321–327.
Martin, S.H., Möst, M., Palmer, W.J., Salazar, C., McMillan, W.O., Jiggins, F.M., Jiggins, C.D., 2016. Natural Selection and Genetic Diversity in the Butterfly Heliconius melpomene. Genetics 203, 525–541. https://doi.org/10.1534/genetics.115.183285
McVean, G.A.T., Myers, S.R., Hunt, S., Deloukas, P., Bentley, D.R., Donnelly, P., 2004. The fine-scale structure of recombination rate variation in the human genome. Science 304, 581–584. https://doi.org/10.1126/science.1092500
Mölder, F., Jablonski, K.P., Letcher, B., Hall, M.B., Tomkins-Tinch, C.H., Sochat, V., Forster, J., Lee, S., Twardziok, S.O., Kanitz, A., Wilm, A., Holtgrewe, M., Rahmann, S., Nahnsen, S., Köster, J., 2021. Sustainable data analysis with Snakemake. F1000Research 10. https://doi.org/10.12688/f1000research.29032.2
Nei, M., 1975. Molecular population genetics and evolution. Mol. Popul. Genet. Evol.
O'Connell, J., Gurdasani, D., Delaneau, O., Pirastu, N., Ulivi, S., Cocca, M., Traglia, M., Huang, J., Huffman, J.E., Rudan, I., McQuillan, R., Fraser, R.M., Campbell, H., Polasek, O., Asiki, G., Ekoru, K., Hayward, C., Wright, A.F., Vitart, V., Navarro, P., Zagury, J.-F., Wilson, J.F., Toniolo, D., Gasparini, P., Soranzo, N., Sandhu, M.S., Marchini, J., 2014. A General Approach for Haplotype Phasing across the Full Spectrum of Relatedness. PLOS Genet. 10, e1004234. https://doi.org/10.1371/journal.pgen.1004234
Patterson, Marschall, Pisanti, 2015. WhatsHap: Weighted Haplotype Assembly for Future-Generation Sequencing Reads | Journal of Computational Biology. Comput. Biol. 22.
Peart, C.R., Tusso, S., Pophaly, S.D., Botero-Castro, F., Wu, C.-C., Aurioles-Gamboa, D., Baird, A.B., Bickham, J.W., Forcada, J., Galimberti, F., Gemmell, N.J., Hoffman, J.I., Kovacs, K.M., Kunnasranta, M., Lydersen, C., Nyman, T., de Oliveira, L.R., Orr, A.J., Sanvito, S., Valtonen, M.,

Shafer, A.B.A., Wolf, J.B.W., 2020. Determinants of genetic variation across eco-evolutionary scales in pinnipeds. Nat. Ecol. Evol. 4, 1095–1104. https://doi.org/10.1038/s41559-020-1215-5
Peñalba, J.V., Wolf, J.B.W., 2020. From molecules to populations: appreciating and estimating recombination rate variation. Nat. Rev. Genet. 21, 476–492. https://doi.org/10.1038/s41576-020-0240-1 ***Review synthétisant l'importance de l'étude de la recombinaison en génomique évolutive, proposant notamment une approche populationnelle pour l'étude de la sélection en liaison.**
Poplin, R., Ruano-Rubio, V., DePristo, M.A., Fennell, T.J., Carneiro, M.O., Van der Auwera, G.A., Kling, D.E., Gauthier, L.D., Levy-Moonshine, A., Roazen, D., Shakir, K., Thibault, J., Chandran, S., Whelan, C., Lek, M., Gabriel, S., Daly, M.J., Neale, B., MacArthur, D.G., Banks, E., 2017. Scaling accurate genetic variant discovery to tens of thousands of samples (preprint). Genomics. https://doi.org/10.1101/201178
Rettelbach, A., Ellegren, H., Nater, A., 2019. How Linked Selection Shapes the Diversity Landscape in Ficedula Flycatchers | Genetics | Oxford Academic. Genetics 212.
Roberts, R.G., 2015. Lewontin's Paradox Resolved? In Larger Populations, Stronger Selection Erases More Diversity. PLOS Biol. 13, e1002113. https://doi.org/10.1371/journal.pbio.1002113
Roesti, Moser, Berner, 2013. Recombination in the threespine stickleback genome—patterns and consequences. Mol. Ecol. 22.
Romiguier, J., Gayral, P., Ballenghien, M., Bernard, A., Cahais, V., Chenuil, A., Chiari, Y., Dernat, R., Duret, L., Faivre, N., Loire, E., Lourenco, J.M., Nabholz, B., Roux, C., Tsagkogeorga, G., Weber, A.A.-T., Weinert, L.A., Belkhir, K., Bierne, N., Glémin, S., Galtier, N., 2014. Comparative population genomics in animals uncovers the determinants of genetic diversity. Nature 515, 261–263. https://doi.org/10.1038/nature13685
Samuk, K., Owens, Delmore, Miller, Rennison, Schluter, 2017. Gene flow and selection interact to promote adaptive divergence in regions of low recombination. Mol. Ecol. 26.
Shen, C., Wang, N., Huang, C., Wang, M., Zhang, X., Lin, Z., 2019. Population genomics reveals a fine-scale recombination landscape for genetic improvement of cotton. Plant J. 99, 494–505. https://doi.org/10.1111/tpj.14339
Stapley, J., Feulner, P.G.D., Johnston, S.E., Santure, A.W., Smadja, C.M., 2017. Variation in recombination frequency and distribution across eukaryotes: patterns and processes. Philos. Trans. R. Soc. B Biol. Sci. 372, 20160455. https://doi.org/10.1098/rstb.2016.0455
Stukenbrock, E.H., Dutheil, J.Y., 2018. Fine-Scale Recombination Maps of Fungal Plant Pathogens Reveal Dynamic Recombination Landscapes and Intragenic Hotspots. Genetics 208, 1209–1229. https://doi.org/10.1534/genetics.117.300502
Tine, M., Kuhl, H., Gagnaire, P.-A., Louro, B., Desmarais, E., Martins, R.S.T., Hecht, J., Knaust, F., Belkhir, K., Klages, S., Dieterich, R., Stueber, K., Piferrer, F., Guinand, B., Bierne, N., Volckaert, F.A.M., Bargelloni, L., Power, D.M., Bonhomme, F., Canario, A.V.M., Reinhardt, R., 2014. European sea bass genome and its variation provide insights into adaptation to euryhalinity and speciation. Nat. Commun. 5, 5770. https://doi.org/10.1038/ncomms6770
Vijay, N., Weissensteiner, M., Burri, R., Kawakami, T., Ellegren, H., Wolf, J.B.W., 2017. Genomewide patterns of variation in genetic diversity are shared among populations, species and higher-order taxa. Mol. Ecol. 26, 4284–4295. https://doi.org/10.1111/mec.14195
Wang, J., Street, N.R., Sco, D.G., 2016. Natural Selection and Recombination Rate Variation Shape Nucleotide Polymorphism Across the Genomes of Three Related Populus Species.
Wilson, Vincent, Ahnesjö, Meyer, 2001. Male Pregnancy in Seahorses and Pipefishes (Family Syngnathidae): Rapid Diversification of Paternal Brood Pouch Morphology Inferred From a Molecular Phylogeny | Journal of Heredity | Oxford Academic. J. Hered. 92.

# Annexes

Table 1 :

Outgroup utilisés pour orienter les variants pour chacune des espèces du projet.

| | Outgroup_1 | | Outgroup_2 | | Outgroup_3 | |
|---|---|---|---|---|---|---|
| Ingroup | espèce | genbankaccess | espèce | espèce | espèce | genbankaccess |
| Dlabr | Lates calcarifer | GCA_001640805.1 | Morone saxatilis | Morone saxatilis | Gasterosteus aculeatus | GCA_016920845.1 |
| Hgutt | Hippocampus comes | | Hippocampus whitei | Hippocampus whitei | Hippocampus kuda | |
| Styph | Syngnathus acus | | Syngnathus floridae | Syngnathus floridae | Syngnathus scovelli | |
| Lbude | Lophius piscatorius | GCA_009660295.1 | Antennarius maculatus | Antennarius maculatus | Antennarius striatus | GCA_900303275.1 |
| Scabr | Hypolectrus puella | GCA_900610375.1 | Epinephelus moara | Epinephelus moara | Plectropomus leopardus | GCA_011397275.1 |
| Msurm | Dactylopterus volitans | GCA_901007715.1 | Aeoliscus strigatus | Aeoliscus strigatus | Callionymus lyra | GCA_016630915.1 |
| Ssard | Thunnus albacares | GCA_900302625.1 | Thunnus orientalis | Thunnus orientalis | Thunnus thynnus | GCA_003231725.1 |
| Scant | Sparus aurata | GCA_900880675.1 | Diplodus sargus | Diplodus sargus | Acanthopagrus latus | GCA_904848185.1 |
| Scine | Symphodus melops | GCA_002819105.1 | Labrus bergylta | Labrus bergylta | Notolabrus celidotus | GCA_009762535.1 |
| Spilc | Clupea harengus | GCA_900700415.2 | Limnothrissa miodon | Limnothrissa miodon | Alosa alosa | GCA_017589495.1 |

**Résumé** : La spéciation est le processus évolutif au cours duquel une espèce se scinde en deux lignées qui divergent en accumulant des barrières reproductives, jusqu'à l'acquisition d'un isolement reproductif total. Durant ce processus, les lignées divergentes peuvent toujours s'échanger des gènes par hybridation, mais le flux génique est progressivement limité par l'accumulation des barrières. Il en résulte une semi-perméabilité des génomes, où certains locus s'échangent librement entre lignées et restent indifférenciés tandis que d'autres n'introgressent pas, contribuant ainsi à l'établissement de régions génomiques divergentes, appelées îlots génomiques de spéciation. Ces locus barrières peuvent être impliqués dans différents types de mécanismes d'isolement, incluant le choix de partenaire, l'adaptation à différents environnements, ou des incompatibilités génétiques entre plusieurs gènes coadaptés. L'étude de l'établissement, l'accumulation, l'érosion et la maintenance de ces barrières et de leurs effets sur la semiperméabilité des génomes de lignées en cours de spéciation permet de comprendre comment de nouvelles espèces se forment. L'avènement des techniques de séquençage à haut débit a permis de caractériser le paysage génomique de divergence chez de multiples lignées en cours de spéciation à travers l'arbre du vivant. Ces études ont permis de mesurer l'influence de l'histoire démographique et de l'architecture génomique comme déterminants majeurs du paysage génomique de divergence. Toutefois, d'autres facteurs pourraient intervenir et expliquer la diversité des trajectoires évolutives pouvant conduire ou non à la spéciation. Le principal objectif de cette thèse est d'évaluer l'impact des traits d'histoire de vie des espèces sur la spéciation. Nous avons choisi d'étudier 20 espèces de poissons marins subdivisées en deux lignées (Atlantique et Méditerranéenne), et présentant une large diversité de niveaux de divergence et de traits d'histoire de vie. Ces traits sont supposés impacter l'intensité de la dérive génétique, les capacités de dispersion et le temps de génération des espèces. Le contrôle par une histoire biogéographique et une architecture génomique commune à toutes les espèces nous permet de tester spécifiquement le rôle des traits d'histoire de vie sur plusieurs mécanismes évolutifs intervenant dans la spéciation. Dans le premier chapitre, nous avons étudié les déterminants de la diversité génétique, substrat sur lequel s'établit la divergence lors de la séparation initiale des lignées. Nous avons observé que la longévité adulte des poissons marins est corrélée négativement à la diversité génétique, et nous avons démontré que cette relation pouvait s'expliquer par une plus grande variance du succès reproducteur chez les espèces longévives à cause de stratégies reproductives particulières aux poissons marins (forte mortalité juvénile, faible mortalité adulte et augmentation de la fécondité avec l'âge). Puis, dans un second chapitre, nous avons détecté une grande diversité d'histoires évolutives entre espèces, caractérisée par un fort gradient de divergence génétique entre lignées atlantiques et méditerranéennes. Ce gradient reflète en partie le niveau de semi-perméabilité des génomes. Les espèces à faible différentiation présentent un isolement reproductif faible, alors que les espèces les plus fortement différenciées montrent un isolement reproductif quasi-complet. Les traits d'histoire de vie des espèces expliquent en partie cette diversité de niveaux d'isolement via différents mécanismes. La durée de vie larvaire influence négativement la différenciation génétique en modulant les capacités de dispersion, l'effet de la taille du corps indique un effet négatif de l'abondance long-terme sur la divergence, et la longévité semble impacter le nombre de générations écoulées depuis la séparation ancestrale. Enfin, dans un dernier chapitre, nous avons montré que les patrons de divergence détectés sur le génome nucléaire se reflétaient en partie sur les génomes mitochondriaux. En conclusion, les 20 espèces étudiées présentent une variabilité surprenante d'histoires évolutives au regard des similitudes de leur histoire biogéographique et leur architecture génomique. Les relations entre traits d'histoire de vie et histoire évolutive des espèces sont complexes, mais nous avons pu éclairer certaines d'entre elles en décomposant l'implication des traits dans les différentes étapes de la spéciation. L'application de l'approche de génomique comparative développée au cours de cette thèse dans d'autres zones de suture permettra d'étendre nos connaissances des déterminants du tempo et du mode de la spéciation.

**Mots-clés:** spéciation, poissons marins, traits d'histoire de vie, diversité et divergence génétique, zone de suture atlantico-méditerranéenne.

**Abstract** : Speciation is the evolutionary process through which a species splits into two lineages that diverge and accumulate reproductive barriers, until complete reproductive isolation is achieved. During this process, the diverging lineages can still exchange genes by hybridisation, but gene flow is progressively restricted by the accumulation of barriers. This results in semi-permeable genomes, whereby some loci exchange freely between lineages and remain undifferentiated while others do not introgress, thus contributing to the establishment of divergent genomic regions, called genomic islands of speciation. These barrier loci may be involved in different types of isolating mechanisms, including mate choice, adaptation to different environments, or genetic incompatibilities between co-adapted genes. The study of the establishment, accumulation, erosion and maintenance of these barriers and their effects on the semipermeability of the genomes of lineages undergoing speciation helps to understand how new species are formed. The advent of high-throughput sequencing techniques has made it possible to characterise the genomic landscape of divergence in multiple lineages undergoing speciation across the tree of life. These studies have shown the influence of the demographic history and genomic architecture as major determinants of the genomic landscape of divergence. However, other factors could intervene and explain the diversity of evolutionary trajectories that may or may not lead to speciation. The main objective of this thesis is to assess the impact of species' life history traits on speciation. We have chosen to study 20 marine fish species subdivided into two lineages (Atlantic and Mediterranean), and presenting a wide diversity of degrees of divergence and life history traits. These traits are thought to impact on the intensity of genetic drift, dispersal abilities and generation time of the species. Controlling for a shared biogeographic history and genomic architecture across species allowed us to specifically test the role of life history traits on several evolutionary mechanisms involved in speciation. In the first chapter, we studied the determinants of genetic diversity, the substrate on which divergence is built during the initial separation of lineages. We observed that adult longevity of marine fishes is negatively correlated with genetic diversity, and we demonstrated that this relationship could be explained by a greater variance in reproductive success in long-lived species due to reproductive strategies specific to marine fishes (high juvenile mortality, low adult mortality and increased fecundity with age). Then, in a second chapter, we discovered a great diversity of evolutionary histories between species, characterised by a strong gradient of genetic divergence between Atlantic and Mediterranean lineages. This gradient partly reflects the level of semi-permeability of the genomes. Species with low differentiation show low reproductive isolation, whereas the most highly differentiated species show almost complete reproductive isolation. Species' life history traits partly explain this diversity in isolation levels via different mechanisms. Larval duration negatively influences genetic differentiation by modulating dispersal capacities, the effect of body size indicates a negative effect of long-term abundance on divergence, while longevity seems to impact the number of generations elapsed since ancestral separation. Finally, in a last chapter, we showed that the divergence patterns detected on the nuclear genome were partly reflected on the mitochondrial genomes. In conclusion, the 20 species studied show a surprising variability of evolutionary histories considering the similarities of their biogeographic history and genomic architecture. The relationships between life-history traits and the evolutionary history of the species proved to be complex, but we were nevertheless able to shed light on some of them by decomposing the involvement of traits in the different stages of speciation. The application of the comparative genomics approach developed in this thesis to other suture zones will further extend our knowledge of the determinants of the tempo and mode of speciation.

**Key words:** speciation, marine fishes, life history traits, genetic diversity and divergence, Atlantic - Mediterranean suture zone.