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

En génétique et génomique

École doctorale GAIA

Unité de recherche ISEM

Génomique comparative de la divergence écotypique associée au gradient laguno-marin chez cinq espèces de poissons

Comparative genomic analysis of ecotype divergence across the marine-lagoon ecological gradient in different species of fish

Présentée par Laura MEYER
Le 6 décembre 2023

Sous la direction de Bruno GUINAND
et Pierre-Alexandre GAGNAIRE

Devant le jury composé de

Roger BUTLIN, Professeur, School of Biosciences, University of Sheffield, UK	Rapporteur
Violaine LLAURENS, DR CNRS, Muséum national d'Histoire naturelle, Paris, France	Rapportrice
Carole KERDELHUE, DR2 INRAE, CBGP, Montpellier, France	Examinatrice
Claire MEROT, CR CNRS, ECOBIO, Université de Rennes, France	Examinatrice
Thibault LEROY, CRCN INRAE, GenPhySE, Toulouse, France	Examinateur

UNIVERSITÉ DE MONTPELLIER

Acknowlemerciements

Merci de me pardonner mon franglais. Please excuse my Frenglish.

First and foremost, thank you to Roger Butlin, Violaine Llaurens, Claire Mérot, Thibault Leroy and Carole Kerdelhué for agreeing to evaluate my thesis. It has been a pleasure writing while imagining your big brains reading my future manuscript.

Merci à mes directeurs de thèse, Pierre-Alexandre et Bruno, qui m'ont pris sous leurs ailes dès le début. Vous étiez bienveillants et complémentaires dans votre encadrement, et j'ai passé trois super années grâce à vous. Un grand merci !

Pierre-Alex, dès mon premier Skype avec toi je savais que ce serait top de travailler ensemble. Il y a trois ans, j'étais à Copenhague pour mon stage de M2 et j'avais demandé à une copine si elle connaissait des gens sympas dans la génétique des pops en France. À son tour, elle avait demandé à un autre copain (je ne connais même pas son nom) et il avait répondu "Pierre-Alexandre Gagnaire, young group with some money" (haha). Eh bien, merci à cette personne inconnue d'avoir dirigé mon chemin futur vers le Sud de la France et vers Montpellier et vers l'ISEM. Je veux te remercier d'avoir sauté dans le bain et d'avoir fait confiance à une tête sur un écran qui à la base venait de l'autre bout du monde. C'était super de discuter avec toi au fil de ces trois années et de te convoquer devant mon écran à chaque fois pour te montrer mes derniers graphiques (des résultats tout chauds !). Ça a été super de travailler avec toi, je n'aurais pas pu espérer mieux, je suis si bien tombée. Merci pour tout ce que tu fais - ça se sent que tu fais tout pour mettre tes étudiants et leur travaux en avant. Merci d'être à l'écoute de nos idées et nos ressentis, et d'être enthousiaste et constructif. C'est parce que tu crois en nous qu'ensuite on peut commencer à le faire nous-mêmes. Merci infiniment.

Merci Bruno de m'avoir aidé et guidé, et pour tout le temps que tu m'as accordé, même pour des choses admin (oh les maints formulaires...) où tu m'as pleinement aidé (quasi-immédiatement) à chaque fois que je demandais (et tu allais toujours au-delà de ce que je demandais). Merci pour les gâteaux et les petits gestes. Je trouve que tu es une personne intéressante, créative, d'une profondeur de pensée. Merci aussi pour ton calme et ton recul qui sont des qualités appréciées pendant une thèse courte.

Merci à tous les collègues de l'équipe BEM et du bâtiment 24 ! Merci à Nicolas pour les discussions, merci à Fred Cerqueira pour sa BD sur le féminisme et son air rassurant. Merci à Iago pour son énergie et sa théorie de l'Univers et des biscuits croquants/moelleux. Adrien, Benjamin, Khalid pour leur gentillesse. Merci à mes chères collègues et mes co-bureaux, Fanny et Marie - je suis contente d'avoir partagé cette expérience avec vous dès le début. Vous êtes trop mims et je vous aime. En dernier, d'après une étude exhaustive, l'ISEM c'est le meilleur labo du monde. C'est un endroit formateur, plein de bienveillance, ce qui est une chose rare. Merci à Campus France de m'avoir permis de rejoindre ce labo.

Progress is a slow race, that is for sure. And especially at the beginning, oh how slow a race it was. Arriving in a PhD about evolution after doing a Master's in marine biology (I just remember plankton really). In the beginning it took me so much time to make links between concepts, I was so lost. But piece by piece it started to make a bit more sense, and I have come to appreciate even more the complexity of this research field. Evolutionary biology is so abstract and the intricacies of the way in which concepts intertwine in all directions is staggering. There are so many things that we study that we will simply never know or even be able to see (genomes? speciation? evolution?) or understand. Ça me passionne.

At a base level, this is what I have learnt: it is important to keep flossing. Also sleeping enough hours per night (nine is optimal). And taking breaks (faire une pause c'est travailler aussi). When under pressure, it is our good routines that keep us going. Dance class was a life saver. But despite all of these good habits, I dreamt of work almost every night these last few months - I was coding my dreams, cd /media/laura/brain/subconscious && touch somestuff somemorestuff. I truly did feel emptier when my last VCF file was done, three years later. No more analyses to check up on?

Merci Pierre pour tes scripts, grâce à toi j'ai su apprivoiser snakemake. Comme toi j'aimerais aussi ne PAS remercier le covid (tu m'as fait rire avec ça).

Maurine j'ai envie de t'appeler notre mentor women in science. Merci pour les mots de motivation quand il en fallait et pour ton humanité. Merci pour tes dance moves de Beyoncé.

Lila and Hedvig, you girls are the bomb - having you at work was the funnest thing ever. Thank you for that touch of lightheartedness. I could share anything with you and our friendship feels true and boundless.

Merci à Manu et à Gaël, merci pour nos sessions de jazz. Vous êtes des personnes super stylées.

Merci à Arthur du 22, sans lequel j'aurais sûrement perdu la tête. Merci pour tes idées et ton oreille qui écoute sans aucun jugement, j'ai adoré chaque discussion sur la vie, l'anarchisme et le polyamour.

Merci aux doctorant.e.s et collègues de l'ISEM entier pour les apéros et les week-ends d'intégration !! Vivement l'édition 2023 !

Dankie Mamma en Pappa vir al julle ondersteuning vandat ek so klein is. Julle het my altyd die vryheid gelos om te wees wie ek wil wees en ek sou nooit hier gekom het sonder julle liefde nie. Ek dink ook aan my sussies, Marlene en Anika, van wie ek ver weg is vir 'n lang tyd (nee ek huil nie, dis net te veel blou lig van al die skerms). Ek mis julle en dit is moeilik, elke keer wat ons baai sê op die lughawe. Dankie dat julle almal verstaan dat hierdie avontuur vir my sovéél beteken. Ons bel mekaar nie altyd so gereeld nie, maar ek is altyd só bly om huis toe te kom en in my

kamer te slaap asof niks verander het nie. Hoe het dit alles so vinnig gebeur? Dit voel soos gister wat ek met my groot roltas uit die huis uit is lughawe toe.

J'ai une pensée pour mon pays, dans toute sa complexité.

Thank you to Alex, my oldest and bestest friend. Your influences on me are here in my writing.

La communauté du Madre, vous êtes les meilleur.e.s de Montpellier.

Thank you Camille for going pipefishing with me, even though we didn't manage to catch them at first, those were beautiful days. Thank you for always believing in me and calming me down.

And thank you to Ambre for all of your support and love through all of this. Thank you for listening to my ramblings and dreams about inversions and for taking sparknotes when I explained my work to you. Merci pour les week-end travail, merci d'avoir été compréhensif.

Abstract

Many species show subdivision into phenotypically and genetically differentiated forms that are associated with fine-scale habitat variation. These ecotypes may represent an intermediate stage to the formation of new species, and thus offer key models for understanding the process of speciation. Open questions remain with respect to how local adaptations, historical contingencies and components of genome architecture interact in ecotype formation. The current thesis aimed to study ecotypic subdivision in a comparative framework controlling for a similar biogeographic context. We studied five species of marine fishes from the North East Atlantic and Mediterranean Sea: the European anchovy (*Engraulis encrasicolus*), the long-snouted seahorse (*Hippocampus guttulatus*), the big-scale sand smelt (*Atherina boyeri*), the grey wrasse (*Symphodus cinereus*), and the broadnosed pipefish (*Syngnathus typhle*). These species occur in a variety of different habitats along the marine-lagoon ecological gradient, and comparing their evolutionary histories has the potential to reveal important aspects related to ecotype formation. We wished to characterise the relative roles of ecology, historical contingencies and genomic architecture in determining the evolutionary trajectories of ecotype pairs in each species. Using whole-genome sequencing data, we aimed to test (*i*) whether genetic differences were associated with different habitat types, and (*ii*) how these are maintained in the presence of gene flow. (*iii*) We evaluated the extent to which the genomic architecture participates in maintaining ecotypic differentiation, and (*iv*) whether these differences originated from new mutations, standing genetic variation, or introgressed variation. Finally, we aimed (*v*) to characterise the historical context of ecotypic divergence. In Chapter I, we study ecotypic structure in *E. encrasicolus* - a highly mobile pelagic species showing marine and coastal ecotypes at a wide geographic scale. We identified multiple structural variants (SVs) that underlie ecotypic differentiation and which were likely introgressed from a third lineage in the Southern Atlantic Ocean. In Chapter II, we study two SVs segregating in *H. guttulatus*, which differentiate geographical and ecotype lineages. Our results show that these correspond to large chromosomal inversions representing ancient intraspecific polymorphisms, which are subject to different evolutionary dynamics and contribute differently to ecotype formation. Finally, in Chapter III, we compare the eco-geographic patterns and associated genome architectures of ecotypes in all five species. We found that ecotype structure was generally more pronounced in the Mediterranean as compared to the Atlantic, likely indicating the influence of a shared biogeographic history. Moreover, the comparison of divergence landscapes across species revealed that large SVs, such as chromosomal inversions, are consistently involved in ecotypic differentiation. Due to their suppressive effects on recombination, SVs maintain allelic combinations and could act as barriers to gene flow between diverging lineages experiencing gene flow. Although a single SV might not be sufficient for ensuring reproductive isolation, the build-up of linkage disequilibrium among multiple SVs could help strengthen reproductive isolation, although it remains unclear whether this is a sufficient condition for speciation to complete.

Résumé

De nombreuses espèces se subdivisent en formes phénotypiquement et génétiquement différenciées qui sont associées à des variations d'habitat à fine échelle. Ces écotypes peuvent représenter une étape intermédiaire dans la formation de nouvelles espèces et offrent donc des modèles utiles pour comprendre le processus de spéciation. Des questions importantes restent en suspens quant à la manière dont les adaptations locales, les contingences historiques et les composantes de l'architecture du génome interagissent dans la formation des écotypes. Cette thèse avait pour objectif d'étudier la subdivision écotypique à travers un cadre comparatif réalisé dans contexte biogéographique similaire. Nous avons ainsi étudié cinq espèces de poissons marins de l'Atlantique Nord-Est et de la Méditerranée: l'anchois européen (*Engraulis encrasicolus*), l'hippocampe moucheté (*Hippocampus guttulatus*), l'athérine (*Atherina boyeri*), le crénilabre cendré (*Symphodus cinereus*) et le syngnathe siphonostome (*Syngnathus typhle*). Ces espèces occupent une variété d'habitats différents le long du gradient écologique mer-lagune, et la comparaison de leurs histoires évolutives peut révéler des aspects importants liés à la formation des écotypes. Nous avons cherché à caractériser les rôles relatifs de l'écologie, des contingences historiques et de l'architecture génomique dans la détermination des trajectoires évolutives des paires d'écotypes chez chaque espèce. En utilisant des données de séquençage du génome entier, nous avons cherché à tester (*i*) si les différences génétiques sont associées aux différents types d'habitat et (*ii*) comment celles-ci sont maintenues en présence de flux génique. (*iii*) Nous avons évalué dans quelle mesure l'architecture génomique participe au maintien de la différenciation écotypique et (*iv*) si ces différences proviennent de nouvelles mutations, de variations génétiques pré-existantes ou de variations introgressées. Enfin, nous avons cherché (*v*) à caractériser le contexte historique de la divergence écotypique. Dans le chapitre I, nous étudions la structure écotypique chez *E. encrasicolus* - une espèce pélagique très mobile présentant des écotypes marins et côtiers à une large échelle géographique. Nous avons identifié de multiples variants structuraux (VSs) qui sous-tendent la différenciation écotypique et qui ont probablement été introgressés à partir d'une troisième lignée présente dans le sud de l'océan Atlantique. Dans le chapitre II, nous étudions deux VSs qui différencient les lignées géographiques et écotypiques chez *H. guttulatus*. Nos résultats montrent qu'ils correspondent à d'anciens polymorphismes intraspécifiques d'inversions chromosomiques, soumis à des dynamiques évolutives différentes et contribuant différemment à la différentiation entre écotypes. Enfin, dans le chapitre III, nous comparons les patrons éco-géographiques et les architectures génomiques associées des écotypes des cinq espèces. Nous avons constaté que la structure écotypique était généralement plus prononcée dans la Méditerranée que dans l'Atlantique, ce qui indique probablement l'influence d'une histoire biogéographique commune. De plus, la comparaison des paysages de divergence entre espèces a révélé que les grands VSs, tels que les inversions chromosomiques, sont régulièrement impliqués dans la différenciation écotypique. En raison de leur effet suppresseur sur la recombinaison, les VSs maintiennent les combinaisons alléliques impliquées dans différentes formes d'adaptations, et pourraient ainsi agir comme des barrières au flux génique entre des lignées. Bien que la présence d'un seul VS ne permette pas l'isolement reproductif, l'évolution d'un déséquilibre de liaison entre plusieurs VSs pourrait contribuer à renforcer l'isolement reproductif, même s'il n'est pas certain que cette condition soit suffisante pour achever la spéciation.

Contents

INTRODUCTION

1. Reproductive isolation and speciation

The concept of reproductive isolation is one of the most important ideas in evolutionary biology, and has profound implications for understanding what species are and how new species arise (Coyne & Orr, 2004; Seehausen et al., 2014). Biological species are defined as members of populations that interbreed in nature (or which potentially have that capacity), producing viable and fertile offspring while being reproductively isolated from other such groups (Mayr, 1942). Despite the central importance of reproductive isolation in the way we think about species, this term has been used in many different ways and its precise meaning has remained surprisingly elusive. In a review that addresses this predicament precisely, Westram et al. (2022) proposed that reproductive isolation (RI) should be considered as a quantitative measure of the extent to which gene flow is reduced due to genetic differences between populations. Some of these genetic differences act as genetic barriers to gene flow between populations, reducing the exchange of neutral alleles to a level that would not occur in the absence of these differences. RI barriers thus hinder the flow of alleles that otherwise would have been exchanged, and their accumulation is what defines the process of speciation.

Speciation studies have aimed to understand what kind of RI barriers are involved in lineage divergence, and to dissect the order in which they arise and the conditions under which they build up (Coyne & Orr, 2004). RI can be caused by a variety of reproductive isolating mechanisms acting at different stages in the reproductive cycle (Dobzhansky 1937; 1951; Seehausen et al., 2014). These include premating isolation mechanisms such as allochronic isolation, habitat preference, immigrant inviability (Nosil et al., 2005), assortative mating or sexual selection against hybrids (Servedio, 2004), postmating prezygotic mechanisms such as mechanical isolation or gamete incompatibility (Lessios & Cunningham, 1990; Turissini et al., 2018), and postzygotic mechanisms such as postmating immigrant inviability, hybrid inviability and sterility (Orr, 1995). The conditions that favour the establishment of these different types of RI barriers have been linked to different modes of speciation, which differ in the geographical arrangement of diverging populations and in the role and type of divergent selection involved (Coyne & Orr, 2004; Endler, 1977; Schluter, 2001; Sobel et al., 2009; Seehausen et al., 2014).

The geographical context that accompanied speciation has traditionally received much attention (Mayr, 1942). The classical distinction between allopatric, parapatric and sympatric speciation essentially boils down to the extent to which the absence or presence of gene flow, driven by the rate of gene exchange (m=0 in allopatry, 0<m<0.5 in parapatry and m=0.5 sympatry), opposes divergence and the evolution of RI. Although this is a useful distinction that highlights the central role of gene flow in speciation, it also tends to distract from the other major evolutionary forces underlying the accumulation of RI. Moreover, it provides a static view that does not take into account the possible alternation of phases of allopatric, parapatric and sympatric divergence, with varying levels of gene flow through time (Butlin et al., 2008). For these reasons, a more continuous vision that includes variation in gene flow over time and space has been proposed as a more comprehensive framework for understanding how gene flow counteracts speciation and what conditions may facilitate divergence (Smadja & Butlin, 2011).

Despite the aforementioned limitations, understanding the historical spatial context that accompanied the evolution of RI may reveal something about the relative importance of selection and the type of selection. For example, if RI between populations appears to have evolved in sympatry, we would argue that this must have involved some form of divergent selection. At the opposite extreme, RI could have evolved as a by-product of divergence following a period of geographic isolation, in which case the neutral process of genetic drift could simply be invoked. In cases where selection would appear to have played an important role, the selective establishment of RI could still be qualified in terms of its dependence (extrinsic selection) or independence (intrinsic selection) from the environment. Here again, things may not be simple to categorise. Allopatric divergence may involve ecologically-based divergent selection, while evolutionary changes in response to ecological selection (i.e. local adaptation) may also lead to the emergence of intrinsic barriers (e.g. Dobzhansky-Muller incompatibilities) as a by-product (Kulmuni & Westram, 2017).

The second aspect we need to describe concerning the genetic basis of speciation, is the architecture of the barriers that are reducing gene flow. A handful of large-effect loci causing speciation is not a prevalent situation, and typically the evolution of RI involves the participation of many genes (Nosil et al., 2021). The way in which these barriers are distributed along the genome can interact with the mode of speciation, since tightly linked loci will more readily resist swamping in the face of gene flow, forming genomic islands of divergence (Feder et al., 2012; Ravinet et al., 2017). In the allopatric model, genetic incompatibilities may accumulate genome-wide, leaving us with a more homogeneous landscape of divergence. However, secondary contact between lineages that diverged in allopatry might also produce a pattern of islands of differentiation under certain conditions (Bierne et al., 2013; Duranton et al., 2018; Yeaman et al., 2016).

As evolutionary biologists we are left with these different puzzle pieces that we may attempt to put together in order to get a picture of how speciation took place. One of the more important things we have learned from 30 years of speciation genomics studies is that there are plenty of alternative paths to speciation and that no two cases are the same. Similar to the limitations pointed out for the spatial dichotomy between allopatric and sympatric speciation (Butlin et al., 2008), every aspect of speciation is difficult to classify into discrete categories. To describe the continuous nature of this multifarious process, we need to assess all of the different facets developed above, and additionally, we need to keep in mind that the balance between all of these ingredients may change as speciation progresses (**Fig. 1**). The relative importance of each mechanism is expected to shift as diverging lineages move (not necessarily linearly and not always in the same direction) along the speciation continuum, from weakly differentiated populations to locally adapted forms with partial RI to strongly reproductively isolated species (Feder et al., 2012; Stankowski & Ravinet, 2021).

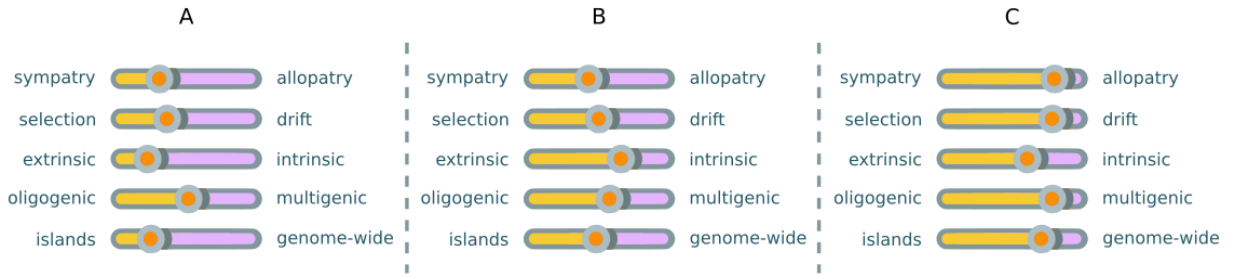


Fig. 1. The multiple facets of speciation. Speciation is a multidimensional process that depends on evolutionary mechanisms related to multiple aspects. Every particular case of speciation lies somewhere in between two extreme configurations along each dimension, including the spatial context of divergence (sympatry-parapatry-allopatry gradient), the main evolutionary forces involved in divergence (drift-selection gradient), the nature of the selective effects regarding their dependence to the environment (intrinsic-extrinsic gradient), the nature of the barrier to gene flow in terms of the number of loci involved (oligogenic-multigenic gradient) and their effect on the genomic landscape of divergence (few small islands to genome-wide divergence). Understanding the evolutionary history of a given system requires that we assess these different facets to document the diversity of alternative paths to speciation, as for instance here between three imaginary cases of speciating lineages or different time points depicted in panels A, B and C.

2. From ecotype formation to speciation

Some early models described how speciation could take place sympatrically in a panmictic population under specific conditions (Felsenstein, 1981; Smith, 1966). The process of ecological speciation assumes ecologically-based divergent selection that operates in contrasting directions and results in adaptation to different environments, habitats, or resources (Nosil, 2012). If alleles that provide a local selective advantage in a given habitat couple with other genes (e.g. loci that cause assortative mating), it could lead to the evolution of reproductive isolation (Felsenstein, 1981; Hendry, 2009; Rundle & Nosil, 2005; Schluter, 2000). The literature has seen a recent resurgence of interest in the role of ecology in speciation, which is usually associated with divergence in the presence of gene flow (reviewed in Lowry, 2012). However, the conditions required for maintaining adaptive polymorphism in full sympatry are extremely restrictive (e.g. requiring very strong selection) (Smith, 1966). Instead, most lineages could be expected to diverge in parapatric conditions ($0<m<0.5$) (Smadja & Butlin, 2011), potentially also involving non-ecological processes of divergence. These non-ecological processes might be especially important in the later stages of speciation, since RI barriers that are not ecologically based would be necessary to prevent the reversal of speciation when environmental conditions change (Hendry, 2009).

To study the mechanisms of ecological speciation, we may look to focus on the early stages of speciation between conspecific populations that are partly reproductively isolated (Hendry, 2009). Ecotypes represent such an intermediate stage along the continuum between panmictic populations that show continuous adaptive variation, and reproductively isolated species with different ecological characteristics. In this view, the mechanism of local adaptation can give rise to ecotypes that differ in their phenotypic characteristics and genetic makeup. Ecotypic lineages were traditionally described based on their morphological traits, but speciation genomics studies

have now started to shed light on the genetic basis of ecotypic differentiation (Andrew & Rieseberg, 2013; Jones et al., 2012; Ravinet et al., 2016). For instance, model systems for the study of ecological speciation between ecotype pairs include phytophagous insects showing adaptation to different host plants (Nosil et al., 2002; Via, 1999), wing mimicry in butterflies (Jiggins, 2008), or morphotypes associated with different habitats in fishes (Lu & Bernatchez, 1999; Schluter et al., 1997). Empirical work has often focused on studying replicate ecotype pairs in different locations, which in some cases may be sufficiently distant to assume that the different replicates are not connected by gene flow. This absence of gene flow is sometimes taken as an argument for parallel, repeated divergence that has accumulated in a short amount of time (Schluter, 2009). Some authors have documented the presence of ecotypic variation associated with a new habitat that has become available only recently (e.g. new volcanic islands, crater lakes or postglacial habitats) and proposed that this illustrates rapid ecotypic divergence (Lescak et al., 2015; Machado-Schiaffino et al., 2017). Such a scenario would call for very strong divergent selection, and even then, requires that the necessary (standing) genetic variation already exists as a raw substrate for selection (Welch & Jiggins, 2014).

While *de novo* mutations might accumulate at a very slow rate, genetic diversity in the form of standing variation (inherited from ancestral lineages) or introgressed variants (from gene flow with divergent populations or species) is immediately available in populations that are under selective pressures. Anciently diverged variants that are segregating in a population (or that were introduced into one of the diverging lineages) can be spatially redistributed and contribute to contemporary divergence (Belleghem et al., 2018). Similarly, if populations diverged in geographic isolation in the past, they may have come into secondary contact and spread into sympatry (e.g. into different habitats) after RI was already acquired (Hendry, 2009; Rundell & Price, 2009). These scenarios provide alternative explanations for how speciation could have taken place between present-day ecotypes, in a way that does not require strong selection for local adaptation in a short amount of time, or divergence in strict sympatry despite the homogenising effects of gene flow. Importantly, these mechanisms do not preclude the contribution of a form of ecologically based divergent selection, but simply include a role for non-ecological processes such as chance events and historical contingencies in ecotype formation. Specifically, ancient variants that contribute to the build-up and maintenance of divergence may be located in genomic islands of differentiation that have been preserved by locally reduced levels of recombination and maintained by a form of balancing selection. Structural variants (SV) that suppress recombination present an ideal solution in this case and have been suggested to be important for ecotype formation (see the next section). This is consistent with theoretical expectations that suggest that concentrated architectures are more likely to resist homogenisation in the face of gene flow and could explain the frequency of SVs underlying the differentiation of ecotypes .

3. The role of structural variation in speciation

3.1. SVs contribute to reducing gene flow

Over the past few decades, there has been a significant increase in our general awareness regarding the important role that SVs play in evolution. This growing understanding has come from many empirical and theoretical studies insisting that SVs are ubiquitous drivers of evolutionary processes, such as local adaptation and speciation (Mérot et al., 2020). This work has, perhaps to the detriment of gaining a greater macroevolutionary understanding, greatly been carried out at the microevolutionary scale where investigations have focused on the role of SVs in population divergence (Lucek et al., 2023). In these cases of incomplete speciation, it has been observed that the genetic barriers involved in RI between populations are often concentrated in chromosomal rearrangements. At first, these observations were ascribed to the idea that heterokaryotypes (i.e. individuals that carry both the ancestral and the rearranged allele) might present reduced fitness as compared to the homokaryotes (Barton & Bengtsson, 1986; Coyne et al., 1993). In such Hybrid-Sterility models, problems during meiotic pairing in heterokaryotypes produce nonfunctional gametes (e.g. through the formation of acentric or bicentric chromosomes) and hence lead to structural underdominance. However, empirical results have failed to confirm whether sterility caused by underdominant SVs are common in nature (Zhang et al., 2021). In addition, these models do not account for the fact that underdominant variants would have a low probability of fixation due to selective disadvantage while rare (Lande, 1979; Walsh, 1982). Thus, for understanding how SVs contribute to reducing gene flow, the field's focus has shifted from their direct effects on hybrid fitness, to their suppressive effect on recombination (Suppressed-Recombination Models).

If we consider the speciation process to be, in essence, the build-up of LD between barrier loci involved in RI (N. H. Barton & de Cara, 2009; Butlin et al., 2021; Felsenstein, 1981), it is not hard to envision that any mechanisms inhibiting recombination would contribute to speciation. For instance, if multiple locally adapted alleles and/or co-adaptive epistatic alleles have accumulated on a given haplotype, any recombination event taking place in this region could break up these advantageous combinations and re-homogenise the genetic variation. This is why it was suggested that SVs that locally reduce recombination might protect diverging haplotypes and promote the evolution of barriers to gene flow (Kirkpatrick & Barton, 2006; Navarro & Barton, 2003; Rieseberg, 2001). In these models, recombination suppression subsequently allows for the accumulation of intrinsic incompatibilities that enhance RI between populations (Connallon & Olito, 2022; Smadja & Butlin, 2011). These properties of SVs should also contribute to preserving differentiation between lineages after they enter into secondary contact (Noor et al., 2001; Rafajlović et al., 2021). Specifically, SVs may play a pivotal role in bringing about speciation in the presence of gene flow (Guerrero & Kirkpatrick, 2014; Rieseberg, 2001), as has already been evidenced in various organisms (see the next section for various examples). It is also a theoretical prediction that, under high-gene-flow conditions, selection tends to favour large-effect loci that better resist genetic swamping (Lenormand, 2002; Yeaman & Whitlock, 2011). This prediction also applies to cases of speciation between ecotype pairs which typically have a patchy distribution and present significant levels of gene flow (Savolainen et al., 2013). Moreover, many

studies have found a strong link between structural variants (mostly large chromosomal inversions) and certain traits that play a role in local adaptation or RI between ecotypes (Campbell et al., 2021; Gould et al., 2017; Hager et al., 2022; Jay et al., 2022; Lundberg et al., 2023). This could be because SVs spanning extensive chromosome segments tend to bring different types of mutations from distant sites into linkage disequilibrium, enhancing the likelihood of generating and conserving haplotype combinations of large effect on the phenotype (Schwander et al., 2014; Thompson & Jiggins, 2014).

3.2. The diversity of SVs involved in speciation

SVs can be defined as any type of genomic change that alters chromosomal location or orientation, or brings about changes in copy number (Zhang et al., 2021). These variants exist along a size continuum, encompassing smaller copy number variants (duplications, insertions, and deletions) as well as large chromosomal rearrangements (typically fusion/fissions, inversions and translocations) (Berdan et al., 2023) which may reach tens of megabases in size. Initially, such chromosomal rearrangements were detected using cytogenetic techniques and microscopy, but recent technological advances have allowed us to identify SVs using either direct or indirect genomic methods (Mérot et al., 2020). For example, SVs such as inversions, may be detected indirectly by the typical patterns of high differentiation and high LD that they produce (e.g. in F_{ST} landscapes, PCA and local PCA). However such methods may be biassed to detecting large (>1 Mb) rearrangements that show significant levels of divergence, while smaller, younger or neutral variants might be overlooked (Mérot et al., 2020). New sequencing technologies (i.e. long-read sequencing) and bioinformatic methods hold promise for the direct detection of a wider variety of SVs, for example, through breakpoint localisation using long reads or linked reads, or through the comparison of *de novo* genome assemblies from multiple individuals sampled over different populations. These methodological advances should help researchers to better understand the full diversity of SVs, as well as their consequences for evolution.

Chromosomal inversions

Research into SVs found its beginnings with the study of chromosomal inversions, a type of rearrangement that was first described just over 100 years ago in *Drosophila* (Sturtevant, 1917, 1921). Subsequently, a large part of classical work on SVs has been based on inversions (Dobzhansky & Sturtevant, 1938; Kirkpatrick, 2010; Kirkpatrick & Barton, 2006; Noor et al., 2001). These are rearrangements that are produced by two breakpoints occurring within the same chromosome, followed by subsequent inversion and reinsertion of this fragment. We may also distinguish between pericentric (including a centromere) and paracentric (not including a centromere) inversions, with the former being more likely to show pronounced underdominance due to production of unbalanced gametes (Kirkpatrick, 2010). Recombination suppression has been well characterised in chromosomal inversions, and results from failure of inverted regions to synapse in heterokaryotes, while recombination in homokaryotes proceeds normally. However, very low levels of recombination may sometimes take place between the arrangements through double crossover and noncrossover gene conversion, inducing what is called ‘gene flux’ (Navarro et al., 1997). It remains true that the general paucity of recombination and genetic homogenisation between arrangements effectively leads to the independent evolution of the inverted and ancestral

haplotypes (Farré et al., 2013). The result is that each arrangement is characterised by its own level of intra-haplotype diversity and its own suite of mutations which may include co-adapted loci, locally favoured alleles and recessive deleterious mutations.

Large chromosomal inversions are the most common type of SV to have been linked to the maintenance of differentiation between lineages as well as adaptive trait variation (Mérot et al., 2020). We find classical examples in multiple taxa where inversions play a significant role in divergence between intra-specific lineages or closely related species. A well-known case is that of marine and freshwater stickleback, where inversions are believed to confer rapid adaptation to new environments from selection on standing genetic variation (Jones et al., 2012). Similarly, in *Littorina saxatilis*, the wave and crab ecotypes are differentiated by several inversions with clinal frequency distributions (Faria et al., 2019b). In *Helianthus* sunflowers, Todesco et al. (2020) identified 37 haploblocks that mainly represent putative inversions, 18 of which were involved in ecotype differentiation. As for mammals, forest and prairie ecotypes of deer mice (*Peromyscus maniculatus*) are differentiated at 13 inversions likely involved in local adaptation (Harringmeyer & Hoekstra, 2022). It was recently shown that four supergenes associated with migratory lifestyle and environmental adaptations in Atlantic cod correspond to chromosomal inversions (Matschiner et al., 2022). Inversions may also show more complex architectures, such as was described for supergene *P* that controls wing colour patterns in *Heliconius numata* (Joron et al., 2006). Multiple inversion events at this locus have extended the region of recombination suppression and resulted in three distinct haplotype classes (Jay et al., 2021). Inversions may also occur within an already inverted region (“nested inversions”), such as was found in the macaque genome, where 83 nested inversions without breakpoint reuse were identified (Maggiolini et al., 2020).

Chromosomal fissions/fusions and translocations

A different class of SVs comprises rearrangements that bring about changes in chromosomal location. Less is known about chromosomal fusions, fissions and translocations than about inversions. Early theoretical studies suggested that chromosomal fusions could contribute to RI (Rieseberg, 2001), especially in scenarios where LD is advantageous (Charlesworth, 1985). Furthermore, recombination suppression in these regions could facilitate local adaptation (Guerrero & Kirkpatrick, 2014), similar to what has been found for inversions. More generally, chromosomal fusions can be thought of as a way of suppressing inter-chromosomal recombination, just as inversion suppresses intra-chromosomal recombination between physically linked loci. It is therefore likely that this kind of SV plays a similar role in speciation as inversions. However, empirical data has only begun to corroborate such hypotheses, with one such example in the Atlantic salmon. Wellband et al. (2018) found that a polymorphic chromosomal fusion in this species showed strong differences between populations that contrasted with otherwise weak population structure. Fusion frequencies were associated with a range of environmental parameters, suggesting that this SV is important for local adaptation. Consistent with theoretical models, the two fused chromosomes showed reduced recombination and genetic diversity near the fusion point, possibly pointing to similar mechanisms that have linked chromosomal inversions to local adaptation.

Some models of chromosomal speciation have put forward that, while a single chromosomal fusion might not cause pronounced underdominance, multiple fusion events could increase underdominant effects in hybrids (Rieseberg, 2001). One such example from further along the speciation continuum, is multiple fusion and fission rearrangements between two *Brenthis* butterfly species (Mackintosh et al., 2023). Using a demographic model of divergence, the authors showed that rearranged chromosomes experienced less effective migration compared to other chromosomes, highlighting their role in RI. It has further been suggested that alternative centric fusion rearrangements (fusion of two acrocentric chromosomes to form a metacentric chromosome) could promote speciation (Coyne & Orr, 2004). In such a scenario, different centric fusions that take place in various lineages could involve the same three chromosomes but would result in the production of sterile hybrids (Zhang et al., 2021). This corresponds to what has been observed in *Pristionchus* nematodes, where two independent fusion events involving the same chromosome differentiate two closely related species (Yoshida et al., 2023). Another example of centric fusion playing a role in speciation has been observed in the European house mouse. This species complex shows remarkable differences in chromosome number, caused by the accumulation of centric fusions as well as whole-arm reciprocal translocations (Garagna et al., 2014; Giménez et al., 2017)

Insertions, deletions and duplications

Thirdly, another type of SV comprises unbalanced rearrangements that affect copy number (i.e. insertions, deletions and duplications). Despite being the most common source of structural variation, this type of smaller SV has often been overlooked in evolutionary studies (Mérot et al., 2020). The main difference between these SVs and those described in the previous sections, might be that rearrangements such as chromosomal inversions and fusions are expected to cause significant recombination suppression along large portions of a chromosome, while copy number variants do not necessarily have this effect. Instead, they are expected to modify gene function, structure or dosage. In terms of their effects on adaptation and divergence, the best described phenomenon is probably that of gene duplication followed by evolutionary diversification of gene paralogs. A strong example was demonstrated by Bikard et al. (2009), who found that divergent evolution caused incompatibility amongst paralogs of a duplicated gene in wild strains of *Arabidopsis thaliana*. Similarly, it was shown that multiple copies of a gene were associated with flowering time in *Mimulus guttatus*, with implications for survival and seed production (Nelson et al., 2019). Deletions may also cause reproductive incompatibility, for example, as observed between populations of mountain pine beetles with different deletions on their Y chromosome (Dowle et al., 2017). Different types of SVs may also tend to co-occur, for instance, chromosomal inversions that lead to deletions at their breakpoints (e.g. causing green colouration in *Timema* stick insects; Villoutreix et al., 2020). SV formation (for example inversions and duplications) may alternatively be induced by the presence of transposable elements (TEs) through non-allelic homologous recombination (Klein & O'Neill, 2018). TEs may directly affect speciation processes themselves, for example, a retrotransposon was shown to affect prezygotic isolation among songbirds by modulating gene expression (Weissensteiner et al., 2020).

3.3. The origin and maintenance of structural variation

Various types of SVs could be important drivers that affect evolutionary outcomes in many species. In order to better understand the effects of such SVs on the process of speciation, we might look to characterise the context surrounding their evolutionary origin, as well as the mechanisms that have maintained structural polymorphism since their emergence. For example, for a large chromosomal rearrangement that results in recombination suppression, we may want to characterise two components related to its origin: (i) Which mutational event led to the observed change in chromosome location or orientation? What molecular mechanisms accompanied this rearrangement, and when did it take place? And (ii), in which evolutionary, demographic and genomic context did divergence evolve between the two alternate arrangements? Importantly, the answer to the first question gives an estimate for the age of the SV, providing us with the necessary information to place its observed level of divergence into a phylogenetic (i.e., interspecific) or demographic history (i.e., intraspecific) perspective. Obtaining an age estimate may further lead us to ask why both arrangements have been maintained since this time. The answer could be linked to (ii), since different evolutionary scenarios evoke different forms of selective or neutral processes that promote structural polymorphism either within or between populations.

The mutational event causing rearrangement

It has been suggested that the frequency of the formation of SVs is not randomly distributed across the genome (Eldridge & Johnston, 1993). Many studies have found that rearrangements tend to take place in genomic regions that contain many repetitive sequences (e.g. Carta & Escudero, 2023) or segmental duplications (e.g. Catacchio et al., 2018). A recent study in *Peromyscus* mice made a special focus on the molecular mechanisms underlying the formation of inversions involved in ecotypic differentiation (Harringmeyer & Hoekstra, 2022). It was found that the breakpoints of these inversions were often associated with long inverted repeats and that they were located in centromeric and telomeric regions. This points to SV formation through the mechanism of ectopic recombination, a phenomenon where recombination takes place atypically at nonallelic positions on a chromosome due to sequence similarity (Casals & Navarro, 2007; Kent et al., 2017). This could explain why SV occurrence is dependent on chromosomal structure and repeat landscape, and why certain types of SVs could be more common within a given lineage (Lucek et al., 2023).

As for dating the mutation that led to chromosomal rearrangement, our current methods are still insufficient for determining absolute age. Many studies have based such estimates on the level of absolute nucleotide divergence or shared polymorphism between arrangements (Corbett-Detig & Hartl, 2012; Fang et al., 2012). This method could potentially underestimate divergence (and hence the age of the SV) due to the limitations associated with mapping divergent sequences to the same reference. This leads to the non-discovery of certain variants that are not present on the reference and lower divergence estimates. Secondly, divergence-based age estimates might be biassed towards lower values in the presence of gene flow, since even low levels of recombination taking place between arrangements could partially erode past accumulated differentiation. Certain types of SVs may also present issues that are specific to a given

mechanism of rearrangement. For example, one avenue of thought that has not yet been fully explored, concerns the formation of a chromosomal inversion through the reversal of a single haplotype that was present in the population. The singletons and rare alleles contained in this randomly sampled haplotype subsequently become fixed in the inverted lineage, whereas we expect that these same mutations would most likely be lost in the lineage of the ancestral arrangement. This could potentially inflate the number of fixed differences between arrangements even for a young inversion. Discriminating between existing variation that was captured upon initial reversal of the segment, and new mutations that have accumulated after the inversion event, could help us to move toward more accurate age estimation, but is hard to do in practice. Furthermore, we could expect this phenomenon to be exacerbated in certain demographic conditions, for example, an inversion that appears in a population that has experienced a recent expansion could be expected to capture more rare mutations.

Establishment after first appearance

Following the initial mutational event that gives rise to an SV, we can distinguish an intermediate phase where the rearranged allele either manages to establish in the population, or is lost. This process is, as for other types of mutations, governed both by neutral and selective processes (Kirkpatrick, 2010). In the case of small and selectively neutral SVs, loss or fixation depends solely on genetic drift. Classically, such an important role for drift has also been highlighted to explain the spread of underdominant inversions that disrupt meiosis (Kirkpatrick, 2010; Mayr, 1978). However, in addition to their structural effects, SVs may also capture multiple loci that are subject to selection (Berdan et al., 2023). These mutations may impact the establishment of an SV, specifically if we assume that they were already captured on the first appearance of the rearrangement (Charlesworth & Barton, 2018; Kirkpatrick & Barton, 2006), rather than accumulating subsequently. As a general rule, the strength of selection on a haplotype can be expected to scale with its length, since longer SVs would contain more variants that are potentially under selection (Berdan et al., 2023). Longer SVs would be more likely to be selected for local adaptation alleles or co-evolved genes they carry, but simultaneously, are more likely to contain deleterious mutations. The loss or spread of a new SV thus results from a balance between all of these processes, and every case should be expected to differ. The demographic context in which a rearrangement takes place is also important in determining its fate. For example, it has been proposed that temporal fluctuations between situations of allopatry and secondary contact can favour the fixation of SVs (Feder et al., 2011). Furthermore, contrary to an intra-specific rearrangement that appears as a single copy in the lineage under study, if an SV is introduced through introgression with another lineage/species (inter-specific), it might start off with multiple copies and have a greater chance of spreading.

In which context did divergence evolve between arrangements?

SVs are fertile ground for the accumulation of divergence, since recombination is reduced between the alternate arrangements. From the point of view of an evolutionary biologist, there are two ways in which the divergence observed between arrangements could have accumulated (illustrated in **Fig. 2**): (i) the mutational event at the origin of the SV took place within the lineage under study, and divergence has accumulated *in situ* over time, or (ii) the mutational event

occurred in a related lineage, and by the time it was introgressed into our study system, it was already divergent.

In many study systems we have found evidence for the second scenario, which describes an inter-specific origin for structural variation (Della Torre et al., 1997; Feder et al., 2003). The introduction of such “pre-packaged” divergence provides an elegant solution to many of the puzzles we faced regarding the maintenance and divergence of non-recombining haplotypes after rearrangement. If two lineages/species that carry different arrangements at an SV hybridise, the rearranged nature of this region will locally impede gene flow and protect the haplotypes that have diverged in (genomic) allopatry. This could result in heterogeneous levels of divergence along the genome, where high differentiation is concentrated in rearranged regions, which is a common feature encountered in population genomic studies. One well-known case for a chromosomal rearrangement of inter-specific origin, is that of an inversion polymorphism in the butterfly *Heliconius numata*. The formation of the supergene controlling for wing colouration patterns in this species involved the introgression of a divergent and inverted segment from a non-sister species (Jay et al., 2018). To determine this, the authors had constructed a phylogeny of the supergene region and observed that *H. numata* and *H. pardalinus* (the donor species) group together despite their non-sister status. The authors could dismiss the possibility that both of these species had fixed the same state of an ancient polymorphic inversion (i.e. incomplete lineage sorting), because divergence time between these species was younger in the inverted region than in the rest of the genome. They proposed that the introgressed haplotype necessarily presented a selective advantage for it to have established in *H. numata*, and that its subsequent maintenance as a polymorphism could be explained by balancing selection.

If a divergent and rearranged haplotype is not inherited from a closely related species, divergence between arrangements might alternatively have evolved within the same lineage. Kirkpatrick (2010) compared this situation to two biological species that co-exist in sympatry. With the exception of a few rare occasions allowing for the exchange of genetic material (corresponding either to hybridisation between species or gene flux between arrangements), as well as potential competition existing between the two entities (inter-specific competition or relative fitness of each haplotype), they do not interact and their evolutionary trajectories are independent. For two arrangements segregating in a species, they start off by being largely genetically homogeneous, and then diverge due to selection, mutation accumulation and genetic drift. During this process, the two lineages (which recombine freely within each arrangement) develop their own sets of co-adapted genes and deleterious mutations. In a case where the species shows population structure and the arrangements are distributed differently across populations or habitats, they might experience different forms of extrinsic selection. This could lead to a situation where each arrangement presents a fitness advantage over the other in a given environment (i.e. local adaptation). The accumulation of such high levels of divergence necessitates that the SV remains polymorphic over a long period of time, perhaps even from before the last speciation event (retention of ancestral polymorphism, e.g. Dobigny et al., 2010; Kapun et al., 2023). For both haplotypes to have been maintained for such an extensive period, we would argue that certain mechanisms must actively have favoured polymorphism and prevented the fixation of either arrangement.

The long-term maintenance of SVs

Studies across multiple taxa have suggested that chromosomal inversions could remain polymorphic for thousands or even millions of generations (reviewed in Faria et al., 2019a). The mechanism that has traditionally been proposed to explain the long-term maintenance of SVs is a form of balancing selection. However, at the same time, we have found growing evidence for the important role that SVs may play in speciation. Taken together, these two observations appear paradoxical, since balancing selection is expected to maintain polymorphism and to oppose divergence and speciation. An alternative view describes how a polymorphic SV could also be maintained at the multi-populational level through a form of divergent selection. These different mechanisms are described by Faria et al. (2019a) who distinguish between two types of inversion polymorphisms (Type I and Type II, outlined below). The authors propose a framework that studies the lifetime evolution of inversions, where an inversion is not considered to be static but is expected to undergo significant change from its first appearance to its establishment and long-term fate. By extending this framework to include all SVs that suppress recombination, we might attempt to dissect the processes that impact the evolutionary dynamics of such rearrangements and to better understand their potential role in speciation.

Divergent selection may promote the maintenance of polymorphism at the species-level if alternate haplotypes confer local adaptation or if there is bistable selection involving incompatibility selection, frequency dependence or assortative mating. These mechanisms maintain Type I polymorphisms, which show marked frequency differences or even differential fixation between populations, although a certain degree of within-level polymorphism might be observed due to migration between populations. Type I polymorphisms are expected to show structural or genic underdominance, for example, due to the accumulation of DM incompatibilities. These types of polymorphisms could promote reinforcement and coupling with other RI mechanisms, suggesting that they could be important for speciation (Kulmuni et al., 2020; Navarro & Barton, 2003).

On the other hand, Type II polymorphisms are not necessarily expected to lead to the evolution of RI barriers. The mechanisms that maintain Type II polymorphisms are linked to balancing selection, corresponding to what was previously proposed (e.g. Jay et al., 2021; Wellenreuther & Bernatchez, 2018; Yeaman & Whitlock, 2011). This could include negative frequency-dependence, antagonistic pleiotropy, disassortative mating or overdominance. Heterokaryote advantage may be conferred by overdominant loci that are present in the inversion, or by recessive deleterious mutations (pseudo-overdominance) that are likely to accumulate in low recombination regions (Barton, 2010; Faria et al., 2019a). Gene flux could present a remedy to the build-up of such mutation load, since heterokaryotes are expected to be common for Type II SVs due to intermediate allele frequencies.

In their review on the maintenance of SV polymorphism, Faria et al. (2019a) highlight that the mechanisms associated with Type I (between-population polymorphism) and Type II (within-population polymorphism) polymorphisms are not mutually exclusive. This is because large SVs capture many mutations that could potentially be under different types of selection. For example, alternate haplotypes at an SV might be under selection for local adaptation, while simultaneously

presenting heterosis or frequency-dependence. This could lead to varying equilibrium frequencies or fixation in some populations. Moreover, the authors point out that there could be interactions between multiple SV polymorphisms segregating in the same species, but that this remains to be studied. We could expect to find epistatic interactions between SVs, making certain haplotype combinations inviable, or an enhanced barrier to gene flow through coupling.

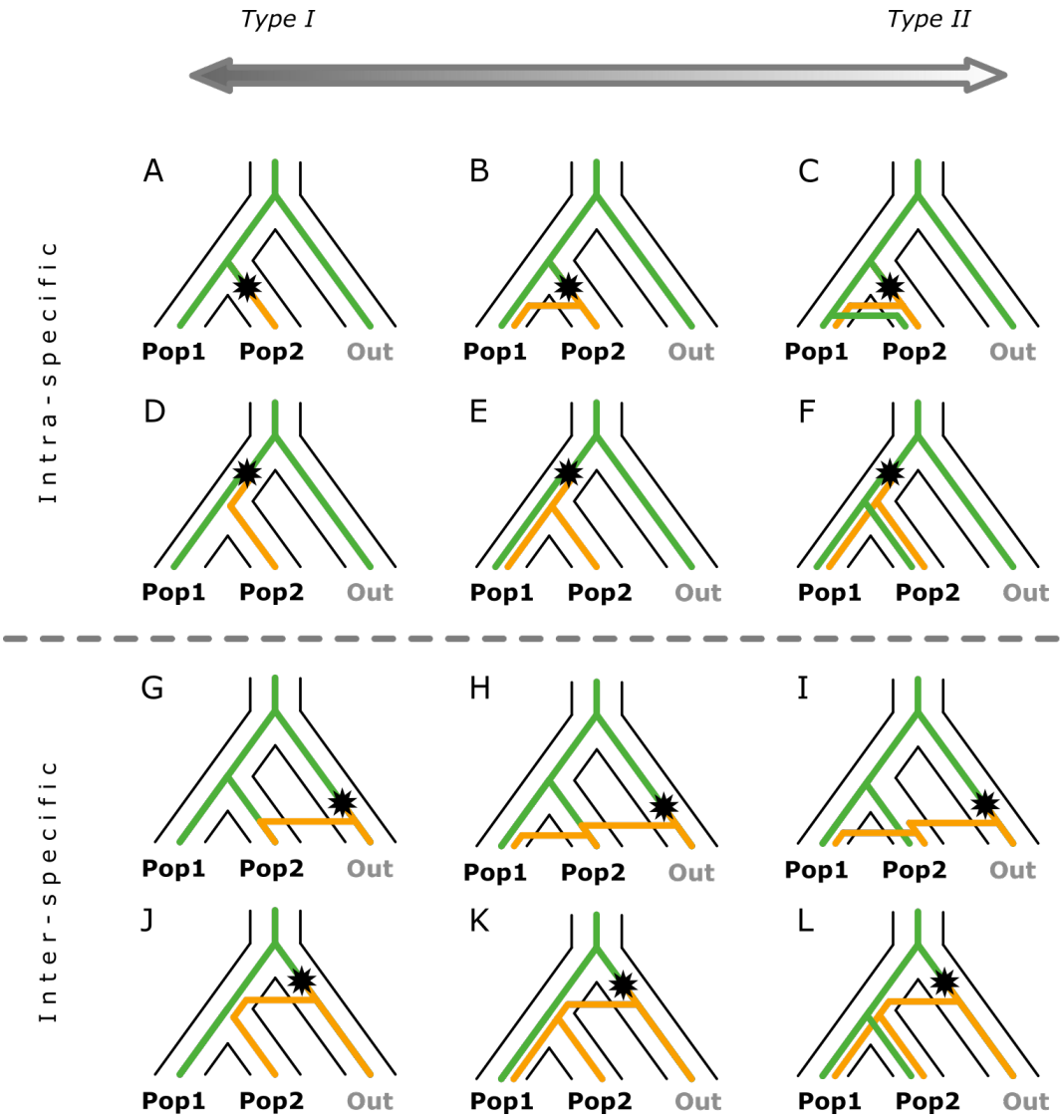


Fig. 2. Possible scenarios showing different origins for an SV segregating in a focal species with two populations (*Pop1* and *Pop2*). For the same given pattern of divergence between populations (A, D, G, J), private polymorphism (B, E, H, K) or shared polymorphism (C, F, I, L), different scenarios may lead to the same result. The appearance of the SV (star) may be relatively recent (A, B, C and G, H, I) or may predate the split between populations (D, E, F and J, K, L) and might have arisen within the species (above the dashed line) or in a different species (below the dashed line). From its initial establishment in the focal species, the SV polymorphism may be maintained by mechanisms associated with Type I (left) to Type II (right) polymorphisms that vary along a continuous gradient.

4. Structure and objectives of the current thesis

Many different factors can be expected to influence the formation of ecotypes. This includes the ecological context of divergence, but also the geographic and demographic history of the involved lineages. To be able to progress towards speciation, diverging ecotypic lineages need to evolve multiple genetic barriers against gene flow and to establish linkage disequilibrium between these barriers. SVs combine these different properties by repressing recombination amongst allelic combinations, and could present an ideal solution for maintaining ecotypic differentiation in the face of gene flow. However many questions remain unanswered regarding their role in speciation, including their evolutionary origin, long-term maintenance and contribution to the completion of speciation.

The current thesis aimed to study ecotypic variation in different species that are exposed to a similar environmental gradient. To accomplish this, we studied the genetic structure in different marine fish that are distributed across a variety of habitats. It is quite commonly observed that certain marine species display phenotypic and genetic variations associated with habitat variations (e.g. depth gradient, foreshore top/bottom, sea/lagoon gradient), on a spatial scale that is often much smaller than the species' dispersal capacity. Several studies have already demonstrated the existence of locally adapted populations within different marine fish species, despite a high dispersal potential and the absence of apparent geographical barriers (e.g. Barth et al., 2017; Clarke et al., 2010; Limborg et al., 2012; Milano et al., 2014; Therkildsen et al., 2013). For this study, we selected five species that occur along the marine-lagoon ecological gradient and which present similar Atlantic-Mediterranean distributions. These species include the big-scale sand smelt (*Atherina boyeri*), the European anchovy (*Engraulis encrasicolus*), the long-snouted seahorse (*Hippocampus guttulatus*), the grey wrasse (*Symphodus cinereus*), and the broadnosed pipefish (*Syngnathus typhle*).

By studying ecotypic differentiation in a comparative framework, we aimed to characterise the respective roles of local adaptation, genome architecture and historical contingencies in the formation of ecotypes. We aimed to describe these different aspects and the evolutionary trajectories of ecotypes across our different study species, to highlight their diversity or alternatively, to identify potential convergences. Our first aim was to test whether genetic structure is correlated with habitat structure. For example, if phenotypic differences exist between populations inhabiting different environments, does this reflect pure phenotypic plasticity or genetic adaptation mechanisms? If ecotypes showed genetic differentiation, we wanted to test whether the whole genome was concerned, or if neutral alleles could flow freely between populations. We wished to characterise the genomic architecture and its role in maintaining 'co-adapted' or selected allelic combinations in the same environment, in the face of gene flow and recombination. We further aimed to characterise the contribution of past demographic fluctuations and admixture events to the situation observed today, with the hypothesis that ecotype divergence could be contingent on past evolutionary history.

We specifically asked the following main questions:

- *Do genetic differences exist between ecotypes? If so, how are they maintained in the face of gene flow?*
- *What is the genome architecture underlying ecotypic divergence? Are the genetic barriers to gene flow dispersed throughout the genome, or are they concentrated in genomic islands of differentiation or SVs?*
- *What are the origins of these variants? Do they represent new mutations, standing variation, or variants that were introduced through introgression?*
- *In which demo-historical context was ecotypic divergence established?*

To address these questions, this thesis is divided into three chapters that focus on different components of ecotypic differentiation. **Chapter I** investigates ecotypic structure in a highly mobile fish species, the European anchovy (*Engraulis encrasicolus*). Here, we present a study that focuses on the spatial distribution of ecotypes and their history of introgression with a divergent lineage. **Chapter II** studies the different genomic components of lineage differentiation in *H. guttulatus* and the history of their establishment. The two SVs differentiating seahorse ecotypes present different characteristics, and we discuss the possible mechanisms responsible for their long-term maintenance. **Chapter III** offers a comparative perspective on eco-geographic patterns across all five species. We compare the presence/absence of ecotype pairs in different locations, and ask whether variation in each species employs similar genomic architectures.

5. References

Andrew, R. L., & Rieseberg, L. H. (2013). Divergence is focused on few genomic regions early in speciation: Incipient speciation of sunflower ecotypes. *Evolution; International Journal of Organic Evolution*, *67*(9), 2468-2482. https://doi.org/10.1111/evo.12106

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

Barton, N., & Bengtsson, B. O. (1986). The barrier to genetic exchange between hybridising populations. *Heredity*, *57*(3). https://doi.org/10.1038/hdy.1986.135

Barton, N. H. (2010). Mutation and the evolution of recombination. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *365*(1544), 1281–1294. https://doi.org/10.1098/rstb.2009.0320

Barton, N. H., & de Cara, M. A. R. (2009). The evolution of strong reproductive isolation. *Evolution*, *63*(5), 1171–1190.

Belleghem, S. M. V., Vangestel, C., Wolf, K. D., Corte, Z. D., Möst, M., Rastas, P., Meester, L. D., & Hendrickx, F. (2018). Evolution at two time frames: Polymorphisms from an ancient singular divergence event fuel contemporary parallel evolution. *PLOS Genetics*, *14*(11), e1007796. https://doi.org/10.1371/journal.pgen.1007796

Berdan E, Aubier TG, Cozzolino S, Faria R, Feder JL, Giménez MD, Joron M, Searle JB & Mérot C. (202x). Structural variants and speciation: Multiple processes at play. In CL Peichel, R Safran, Å Brännström, D Bolnick & U Dieckmann (Eds.), *Speciation*. Cold Spring Harbor (in press)

Bierne, N., Gagnaire, P.-A., & David, P. (2013). The geography of introgression in a patchy environment and the thorn in the side of ecological speciation. *Current Zoology*, *59*(1), 72–86. https://doi.org/10.1093/czoolo/59.1.72

Bikard, D., Patel, D., Le Metté, C., Giorgi, V., Camilleri, C., Bennett, M. J., & Loudet, O. (2009). Divergent evolution of duplicate genes leads to genetic incompatibilities within *A. thaliana*. *Science*, *323*(5914), 623–626. https://doi.org/10.1126/science.1165917

Butlin, R. K., Galindo, J., & Grahame, J. W. (2008). Sympatric, parapatric or allopatric: The most important way to classify speciation? *Philosophical Transactions of the Royal Society B: Biological Sciences*, *363*(1506), 2997–3007. https://doi.org/10.1098/rstb.2008.0076

Butlin, R. K., Servedio, M. R., Smadja, C. M., Bank, C., Barton, N. H., Flaxman, S. M., Giraud, T., Hopkins, R., Larson, E. L., Maan, M. E., Meier, J., Merrill, R., Noor, M. A. F., Ortiz-Barrientos, D., & Qvarnström, A. (2021). Homage to Felsenstein 1981, or why are there so few/many species? *Evolution*, *75*(5), 978–988. https://doi.org/10.1111/evo.14235

Campbell, M. A., Anderson, E. C., Garza, J. C., & Pearse, D. E. (2021). Polygenic basis and the role of genome duplication in adaptation to similar selective environments. *Journal of Heredity*, *112*(7), 614–625. https://doi.org/10.1093/jhered/esab049

Carta, A., & Escudero, M. (2023). Karyotypic diversity: A neglected trait to explain angiosperm diversification? *Evolution*, *77*(4), 1158–1164. https://doi.org/10.1093/evolut/qpad014

Casals, F., & Navarro, A. (2007). Chromosomal evolution: Inversions: the chicken or the egg? *Heredity*, *99*(5), Article 5. https://doi.org/10.1038/sj.hdy.6801046

Catacchio, C. R., Maggiolini, F. A. M., D'Addabbo, P., Bitonto, M., Capozzi, O., Signorile, M. L., Miroballo, M., Archidiacono, N., Eichler, E. E., Ventura, M., & Antonacci, F. (2018). Inversion variants in human and primate genomes. *Genome Research*, *28*(6), 910–920. https://doi.org/10.1101/gr.234831.118

Charlesworth B. (1985). Recombination, genome size and chromosome number. In Cavalier-Smith, T (Eds). *The evolution of genome size* (489–513). John Wiley & Sons.

Charlesworth, B., & Barton, N. H. (2018). The spread of an inversion with migration and selection. *Genetics*, *208*(1), 377–382. https://doi.org/10.1534/genetics.117.300426

Clarke, L. M., Munch, S. B., Thorrold, S. R., & Conover, D. O. (2010). High connectivity among locally adapted populations of a marine fish (*Menidia menidia*). *Ecology*, *91*(12), 3526–3537. https://doi.org/10.1890/09-0548.1

Connallon, T., & Olito, C. (2022). Natural selection and the distribution of chromosomal inversion lengths. *Molecular Ecology*, *31*(13), 3627–3641. https://doi.org/10.1111/mec.16091

Corbett-Detig, R. B., & Hartl, D. L. (2012). Population genomics of inversion polymorphisms in *Drosophila melanogaster*. *PLoS Genetics*, *8*(12), e1003056. https://doi.org/10.1371/journal.pgen.1003056

Coyne, J. A., Meyers, W., Crittenden, A. P., & Sniegowski, P. (1993). The fertility effects of pericentric inversions in *Drosophila melanogaster*. *Genetics*, *134*(2), 487–496. https://doi.org/10.1093/genetics/134.2.487

Coyne, J. A., & Orr, H. A. (2004). *Speciation*. Sinauer.

della Torre, A., Merzagora, L., Powell, J. R., & Coluzzi, M. (1997). Selective introgression of paracentric inversions between two sibling species of the *Anopheles gambiae* complex. *Genetics*, *146*(1), 239–244. https://doi.org/10.1093/genetics/146.1.239

Dobigny, G., Catalan, J., Gauthier, P., O'Brien, P. C. M., Brouat, C., Bâ, K., Tatard, C., Ferguson-Smith, M. A., Duplantier, J. M., Granjon, L., & Britton-Davidian, J. (2010). Geographic patterns of inversion polymorphisms in a wild African rodent, *Mastomys erythroleucus*. *Heredity*, *104*(4), Article 4. https://doi.org/10.1038/hdy.2009.119

Dobzhansky, Th., & Sturtevant, A. H. (1938). Inversions in the chromosomes of *Drosophila pseudoobscura*. *Genetics*, *23*(1), 28–64.

Dobzhansky, T. (1937). *Genetics and the origin of species* (1st ed.). Columbia University Press.

Dobzhansky, T. (1951). *Genetics and the origin of species* (3rd ed.). Columbia University Press.

Dowle, E. J., Bracewell, R. R., Pfrender, M. E., Mock, K. E., Bentz, B. J., & Ragland, G. J. (2017). Reproductive isolation and environmental adaptation shape the phylogeography of mountain pine beetle (*Dendroctonus ponderosae*). *Molecular Ecology*, *26*(21), 6071–6084. https://doi.org/10.1111/mec.14342

Duranton, M., Allal, F., Fraïsse, C., Bierne, N., Bonhomme, F., & Gagnaire, P.-A. (2018). The origin and remolding of genomic islands of differentiation in the European sea bass. *Nature Communications*, *9*(1). https://doi.org/10.1038/s41467-018-04963-6

Eldridge, M. D., & Johnston, P. G. (1993). Chromosomal rearrangements in rock wallabies, Petrogale (Marsupialia: Macropodidae). VIII. An investigation of the nonrandom nature of karyotypic change. *Genome*, *36*(3), 524–534. https://doi.org/10.1139/g93-072

Endler, J. A. (1977). *Geographic Variation, Speciation and Clines. (MPB-10), Volume 10*. Princeton University Press. https://doi.org/10.2307/j.ctvx5wbdg

Fang, Z., Pyhäjärvi, T., Weber, A. L., Dawe, R. K., Glaubitz, J. C., González, J. de J. S., Ross-Ibarra, C., Doebley, J., Morrell, P. L., & Ross-Ibarra, J. (2012). Megabase-scale inversion polymorphism in the wild ancestor of maize. *Genetics*, *191*(3), 883–894. https://doi.org/10.1534/genetics.112.138578

Faria, R., Johannesson, K., Butlin, R., & Westram, A. (2019a). Evolving Inversions. *Trends in Ecology & Evolution*, *34*. https://doi.org/10.1016/j.tree.2018.12.005

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., & Butlin, R. K. (2019b). Multiple chromosomal rearrangements in a hybrid zone between *Littorina saxatilis* ecotypes. *Molecular Ecology*, *28*(6), 1375–1393. https://doi.org/10.1111/mec.14972

Farré, M., Micheletti, D., & Ruiz-Herrera, A. (2013). Recombination rates and genomic shuffling in human and chimpanzee—a new twist in the chromosomal speciation theory. *Molecular Biology and Evolution*, *30*(4), 853–864. https://doi.org/10.1093/molbev/mss272

Feder, J. L., Berlocher, S. H., Roethele, J. B., Dambroski, H., Smith, J. J., Perry, W. L., Gavrilovic, V., Filchak, K. E., Rull, J., & Aluja, M. (2003). Allopatric genetic origins for sympatric host-plant shifts and race formation in Rhagoletis. *Proceedings of the National Academy of Sciences*, *100*(18), 10314–10319. https://doi.org/10.1073/pnas.1730757100

Feder, J. L., Egan, S. P., & Nosil, P. (2012). The genomics of speciation-with-gene-flow. *Trends in Genetics*, *28*(7), 342–350. https://doi.org/10.1016/j.tig.2012.03.009

Feder, J. L., Gejji, R., Powell, T. H. Q., & Nosil, P. (2011). Adaptive chromosomal divergence driven by mixed geographic mode of evolution. *Evolution; International Journal of Organic Evolution*, *65*(8), 2157–2170. https://doi.org/10.1111/j.1558-5646.2011.01321.x

Felsenstein, J. (1981). Skepticism towards Santa Rosalia, or why are there so few kinds of animals? *Evolution*, *35*(1), 124–138. https://doi.org/10.1111/j.1558-5646.1981.tb04864.x

Garagna, S., Page, J., Fernandez-Donoso, R., Zuccotti, M., & Searle, J. B. (2014). The Robertsonian phenomenon in the house mouse: Mutation, meiosis and speciation. *Chromosoma*, *123*(6), 529–544. https://doi.org/10.1007/s00412-014-0477-6

Giménez, M. D., Förster, D. W., Jones, E. P., Jóhannesdóttir, F., Gabriel, S. I., Panithanarak, T., Scascitelli, M.,

Merico, V., Garagna, S., Searle, J. B., & Hauffe, H. C. (2017). A half-century of studies on a chromosomal hybrid zone of the house mouse. *Journal of Heredity*, *108*(1), 25–35. https://doi.org/10.1093/jhered/esw061

Gould, B. A., Chen, Y., & Lowry, D. B. (2017). Pooled ecotype sequencing reveals candidate genetic mechanisms for adaptive differentiation and reproductive isolation. *Molecular Ecology*, *26*(1), 163–177. https://doi.org/10.1111/mec.13881

Guerrero, R. F., & Kirkpatrick, M. (2014). Local adaptation and the evolution of chromosome fusions. *Evolution; International Journal of Organic Evolution*, *68*(10), 2747–2756. https://doi.org/10.1111/evo.12481

Hager, E. R., Harringmeyer, O. S., Wooldridge, T. B., Theingi, S., Gable, J. T., McFadden, S., Neugeboren, B., Turner, K. M., Jensen, J. D., & Hoekstra, H. E. (2022). A chromosomal inversion contributes to divergence in multiple traits between deer mouse ecotypes. *Science*, *377*(6604), 399–405. https://doi.org/10.1126/science.abg0718

Harringmeyer, O. S., & Hoekstra, H. E. (2022). Chromosomal inversion polymorphisms shape the genomic landscape of deer mice. *Nature Ecology & Evolution*, 1–15. https://doi.org/10.1038/s41559-022-01890-0

Hendry, A. P. (2009). Ecological speciation! Or the lack thereof?This Perspective is based on the author's J.C. Stevenson Memorial Lecture delivered at the Canadian Conference for Fisheries Research in Halifax, Nova Scotia, January 2008. *Canadian Journal of Fisheries and Aquatic Sciences*, *66*(8), 1383–1398. https://doi.org/10.1139/F09-074

Jay, P., Chouteau, M., Whibley, A., Bastide, H., Parrinello, H., Llaurens, V., & Joron, M. (2021). Mutation load at a mimicry supergene sheds new light on the evolution of inversion polymorphisms. *Nature Genetics*, *53*(3), 288–293. https://doi.org/10.1038/s41588-020-00771-1

Jay, P., Leroy, M., Le Poul, Y., Whibley, A., Arias, M., Chouteau, M., & Joron, M. (2022). Association mapping of colour variation in a butterfly provides evidence that a supergene locks together a cluster of adaptive loci. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *377*(1856), 20210193. https://doi.org/10.1098/rstb.2021.0193

Jay, P., Whibley, A., Frézal, L., Rodríguez de Cara, M. Á., Nowell, R. W., Mallet, J., Dasmahapatra, K. K., & Joron, M. (2018). Supergene evolution triggered by the introgression of a chromosomal inversion. *Current Biology*, *28*(11), 1839-1845.e3. https://doi.org/10.1016/j.cub.2018.04.072

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

Joron, M., Papa, R., Beltrán, M., Chamberlain, N., Mavárez, J., Baxter, S., Abanto, M., Bermingham, E., Humphray, S. J., Rogers, J., Beasley, H., Barlow, K., ffrench-Constant, R. H., Mallet, J., McMillan, W. O., & Jiggins, C. D. (2006). A conserved supergene locus controls colour pattern diversity in *Heliconius* butterflies. *PLOS Biology*, *4*(10), e303. https://doi.org/10.1371/journal.pbio.0040303

Kapun, M., Mitchell, E. D., Kawecki, T. J., Schmidt, P., & Flatt, T. (2023). An ancestral balanced inversion polymorphism confers global adaptation. *Molecular Biology and Evolution*, *40*(6), msad118. https://doi.org/10.1093/molbev/msad118

Kent, T. V., Uzunović, J., & Wright, S. I. (2017). Coevolution between transposable elements and recombination. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *372*(1736), 20160458. https://doi.org/10.1098/rstb.2016.0458

Kirkpatrick, M. (2010). How and why chromosome inversions evolve. *PLOS Biology*, *8*(9), e1000501. https://doi.org/10.1371/journal.pbio.1000501

Kirkpatrick, M., & Barton, N. (2006). Chromosome inversions, local adaptation and speciation. *Genetics*, *173*(1), 419–434. https://doi.org/10.1534/genetics.105.047985

Klein, S. J., & O'Neill, R. J. (2018). Transposable elements: Genome innovation, chromosome diversity, and centromere conflict. *Chromosome Research*, *26*(1), 5–23. https://doi.org/10.1007/s10577-017-9569-5

Kulmuni, J., Butlin, R. K., Lucek, K., Savolainen, V., & Westram, A. M. (2020). Towards the completion of speciation: The evolution of reproductive isolation beyond the first barriers. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *375*(1806), 20190528. https://doi.org/10.1098/rstb.2019.0528

Kulmuni, J., & Westram, A. M. (2017). Intrinsic incompatibilities evolving as a by-product of divergent ecological selection: Considering them in empirical studies on divergence with gene flow. *Molecular Ecology*, *26*(12), 3093–3103. https://doi.org/10.1111/mec.14147

Lande, R. (1979). Effective deme sizes during long-term evolution estimated from rates of chromosomal

rearrangement. *Evolution*, *33*(1), 234–251. https://doi.org/10.2307/2407380
Lenormand, T. (2002). Gene flow and the limits to natural selection. *Trends in Ecology & Evolution*, *17*(4), 183–189. https://doi.org/10.1016/S0169-5347(02)02497-7
Lescak, E. A., Bassham, S. L., Catchen, J., Gelmond, O., Sherbick, M. L., von Hippel, F. A., & Cresko, W. A. (2015). Evolution of stickleback in 50 years on earthquake-uplifted islands. *Proceedings of the National Academy of Sciences*, *112*(52), E7204–E7212. https://doi.org/10.1073/pnas.1512020112
Lessios, H. A., & Cunningham, C. W. (1990). Gametic incompatibility between species of the sea urchin *Echinometra* on the two sides of the isthmus of panama. *Evolution; International Journal of Organic Evolution*, *44*(4), 933–941. https://doi.org/10.1111/j.1558-5646.1990.tb03815.x
Limborg, M. T., Helyar, S. J., De Bruyn, M., Taylor, M. I., Nielsen, E. E., Ogden, R., Carvalho, G. R., FPT Consortium, & Bekkevold, D. (2012). Environmental selection on transcriptome-derived SNPs in a high gene flow marine fish, the Atlantic herring (*Clupea harengus*). *Molecular Ecology*, *21*(15), 3686–3703. https://doi.org/10.1111/j.1365-294X.2012.05639.x
Lowry, D. B. (2012). Ecotypes and the controversy over stages in the formation of new species. *Biological Journal of the Linnean Society*, *106*(2), 241–257. https://doi.org/10.1111/j.1095-8312.2012.01867.x
Lu, G., & Bernatchez, L. (1999). Correlated trophic specialization and genetic divergence in sympatric lake whitefish ecotypes (*Coregonus clupeaformis*): support for the ecological speciation hypothesis. *Evolution*, *53*(5), 1491-1505. https://doi.org/10.1111/j.1558-5646.1999.tb05413.x
Lucek, K., Giménez, M. D., Joron, M., Rafajlović, M., Searle, J. B., Walden, N., Westram, A. M., & Faria, R. (2023). The impact of chromosomal rearrangements in speciation: from micro- to macroevolution. *Cold Spring Harbor Perspectives in Biology*, a041447. https://doi.org/10.1101/cshperspect.a041447
Lundberg, M., Mackintosh, A., Petri, A., & Bensch, S. (2023). Inversions maintain differences between migratory phenotypes of a songbird. *Nature Communications*, *14*(1). https://doi.org/10.1038/s41467-023-36167-y
Machado-Schiaffino, G., Kautt, A. F., Torres-Dowdall, J., Baumgarten, L., Henning, F., & Meyer, A. (2017). Incipient speciation driven by hypertrophied lips in Midas cichlid fishes? *Molecular Ecology*, *26*(8), 2348–2362. https://doi.org/10.1111/mec.14029
Mackintosh, A., Vila, R., Laetsch, D. R., Hayward, A., Martin, S. H., & Lohse, K. (2023). Chromosome fissions and fusions act as barriers to gene flow between Brenthis fritillary butterflies. *Molecular Biology and Evolution*, *40*(3), msad043. https://doi.org/10.1093/molbev/msad043
Maggiolini, F. A. M., Sanders, A. D., Shew, C. J., Sulovari, A., Mao, Y., Puig, M., Catacchio, C. R., Dellino, M., Palmisano, D., Mercuri, L., Bitonto, M., Porubský, D., Cáceres, M., Eichler, E. E., Ventura, M., Dennis, M. Y., Korbel, J. O., & Antonacci, F. (2020). Single-cell strand sequencing of a macaque genome reveals multiple nested inversions and breakpoint reuse during primate evolution. *Genome Research*, *30*(11), 1680–1693. https://doi.org/10.1101/gr.265322.120
Matschiner, M., Barth, J. M. I., Tørresen, O. K., Star, B., Baalsrud, H. T., Brieuc, M. S. O., Pampoulie, C., Bradbury, I., Jakobsen, K. S., & Jentoft, S. (2022). Supergene origin and maintenance in Atlantic cod. *Nature Ecology & Evolution*, *6*(4), Article 4. https://doi.org/10.1038/s41559-022-01661-x
Mayr, E. (1942). *Systematics and the origin of species*. Columbia Univ. Press, New York
Mayr, E. (1978). Michael J. D. White., Modes of Speciation. *Systematic Biology*, *27*(4), 478–482. https://doi.org/10.1093/sysbio/27.4.478
Mérot, C., Oomen, R. A., Tigano, A., & Wellenreuther, M. (2020). A roadmap for understanding the evolutionary significance of structural genomic variation. *Trends in Ecology & Evolution*, *35*(7), 561–572. https://doi.org/10.1016/j.tree.2020.03.002
Milano, I., Babbucci, M., Cariani, A., Atanassova, M., Bekkevold, D., Carvalho, G., Espiñeira, M., Fiorentino, F., Garofalo, G., Geffen, A., Hansen, J., Helyar, S., Nielsen, E., Ogden, R., Patarnello, T., Stagioni, M., Consortium, F., Tinti, F., & Bargelloni, L. (2014). Outlier SNP markers reveal fine-scale genetic structuring across European hake populations (*Merluccius merluccius*). *Molecular Ecology*, *23*. https://doi.org/10.1111/mec.12568
Navarro, A., & Barton, N. H. (2003). Accumulating postzygotic isolation genes in parapatry: A new twist on chromosomal speciation. *Evolution; International Journal of Organic Evolution*, *57*(3), 447–459. https://doi.org/10.1111/j.0014-3820.2003.tb01537.x
Navarro, A., Betrán, E., Barbadilla, A., & Ruiz, A. (1997). Recombination and gene flux caused by gene conversion and crossing over in inversion heterokaryotypes. *Genetics*, *146*(2), 695–709. https://doi.org/10.1093/genetics/146.2.695

Nelson, T. C., Monnahan, P. J., McIntosh, M. K., Anderson, K., MacArthur-Waltz, E., Finseth, F. R., Kelly, J. K., & Fishman, L. (2019). Extreme copy number variation at a tRNA ligase gene affecting phenology and fitness in yellow monkeyflowers. *Molecular Ecology*, *28*(6), 1460–1475. https://doi.org/10.1111/mec.14904

Noor, M. A. F., Grams, K. L., Bertucci, L. A., & Reiland, J. (2001). Chromosomal inversions and the reproductive isolation of species. *Proceedings of the National Academy of Sciences*, *98*(21), 12084–12088. https://doi.org/10.1073/pnas.221274498

Nosil, P., Crespi, B. J., & Sandoval, C. P. (2002). Host-plant adaptation drives the parallel evolution of reproductive isolation. *Nature*, *417*(6887). https://doi.org/10.1038/417440a

Nosil, P. (2012). *Ecological Speciation*. Oxford University Press.

Nosil, P., Feder, J. L., & Gompert, Z. (2021). How many genetic changes create new species? *Science*, *371*(6531), 777–779. https://doi.org/10.1126/science.abf6671

Nosil, P., Vines, T. H., & Funk, D. J. (2005). Perspective: Reproductive isolation caused by natural selection against immigrants from divergent habitats. *Evolution; International Journal of Organic Evolution*, *59*(4), 705–719.

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

Rafajlović, M., Rambla, J., Feder, J. L., Navarro, A., & Faria, R. (2021). Inversions and genomic differentiation after secondary contact: When drift contributes to maintenance, not loss, of differentiation. *Evolution*, *75*(6), 1288–1303. https://doi.org/10.1111/evo.14223

Ravinet, M., Westram, A., Johannesson, K., Butlin, R., André, C., & Panova, M. (2016). Shared and nonshared genomic divergence in parallel ecotypes of *Littorina saxatilis* at a local scale. *Molecular Ecology*, *25*(1), 287-305. https://doi.org/10.1111/mec.13332

Ravinet, M., Faria, R., Butlin, R. K., Galindo, J., Bierne, N., Rafajlović, M., Noor, M. a. F., Mehlig, B., & 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. https://doi.org/10.1111/jeb.13047

Rieseberg, L. H. (2001). Chromosomal rearrangements and speciation. *Trends in Ecology & Evolution*, *16*(7), 351–358. https://doi.org/10.1016/s0169-5347(01)02187-5

Rundell, R. J., & Price, T. D. (2009). Adaptive radiation, nonadaptive radiation, ecological speciation and nonecological speciation. *Trends in Ecology & Evolution*, *24*(7), 394–399. https://doi.org/10.1016/j.tree.2009.02.007

Rundle, H. D., & Nosil, P. (2005). Ecological speciation. *Ecology Letters*, *8*(3), 336–352. https://doi.org/10.1111/j.1461-0248.2004.00715.x

Savolainen, O., Lascoux, M., & Merilä, J. (2013). Ecological genomics of local adaptation. *Nature Reviews Genetics*, *14*(11). https://doi.org/10.1038/nrg3522

Schluter, D., Rambaut, A., Clarke, B. C., & Grant, P. R. (1997). Ecological speciation in postglacial fishes. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, *351*(1341), 807-814. https://doi.org/10.1098/rstb.1996.0075

Schluter, D. (2000). *The Ecology of Adaptive Radiation*. OUP Oxford.

Schluter, D. (2001). Ecology and the origin of species. *Trends in Ecology & Evolution*, *16*(7), 372–380. https://doi.org/10.1016/s0169-5347(01)02198-x

Schwander, T., Libbrecht, R., & Keller, L. (2014). Supergenes and complex phenotypes. *Current Biology*, *24*(7), R288–R294. https://doi.org/10.1016/j.cub.2014.01.056

Seehausen, O., Butlin, R. K., Keller, I., Wagner, C. E., Boughman, J. W., Hohenlohe, P. A., Peichel, C. L., Saetre, G.-P., Bank, C., Brännström, A., Brelsford, A., Clarkson, C. S., Eroukhmanoff, F., Feder, J. L., Fischer, M. C., Foote, A. D., Franchini, P., Jiggins, C. D., Jones, F. C., … Widmer, A. (2014). Genomics and the origin of species. *Nature Reviews. Genetics*, *15*(3), 176–192. https://doi.org/10.1038/nrg3644

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. https://doi.org/10.1111/j.0014-3820.2004.tb00425.x

Smadja, C. M., & Butlin, R. K. (2011). A framework for comparing processes of speciation in the presence of gene flow. *Molecular Ecology*, *20*(24), 5123–5140. https://doi.org/10.1111/j.1365-294X.2011.05350.x

Smith, J. M. (1966). Sympatric Speciation. *The American Naturalist*, *100*(916), 637–650.

Sobel, J., Chen, G., Watt, L., & Schemske, D. (2009). The Biology of Speciation. *Evolution; International Journal of Organic Evolution*, *64*, 295–315. https://doi.org/10.1111/j.1558-5646.2009.00877.x

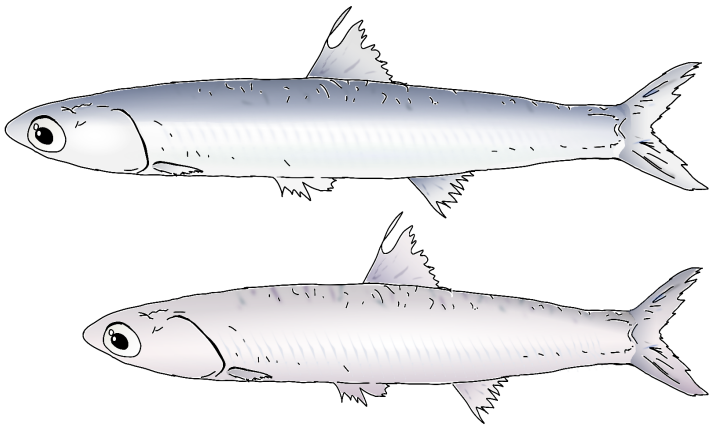
Stankowski, S., & Ravinet, M. (2021). Defining the speciation continuum. *Evolution*, *75*(6), 1256–1273.

https://doi.org/10.1111/evo.14215
Sturtevant, A. H. (1917). Genetic factors affecting the strength of linkage in *Drosophila*. *Proceedings of the National Academy of Sciences of the United States of America*, *3*(9), 555–558.
Sturtevant, A. H. (1921). A case of rearrangement of genes in *Drosophila*. *Proceedings of the National Academy of Sciences*, *7*(8), 235–237. https://doi.org/10.1073/pnas.7.8.235
Therkildsen, N. O., Hemmer-Hansen, J., Hedeholm, R. B., Wisz, M. S., Pampoulie, C., Meldrup, D., Bonanomi, S., Retzel, A., Olsen, S. M., & Nielsen, E. E. (2013). Spatiotemporal SNP analysis reveals pronounced biocomplexity at the northern range margin of Atlantic cod *Gadus morhua*. *Evolutionary Applications*, *6*(4), 690–705. https://doi.org/10.1111/eva.12055
Thompson, M. J., & Jiggins, C. D. (2014). Supergenes and their role in evolution. *Heredity*, *113*(1), Article 1. https://doi.org/10.1038/hdy.2014.20
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., … Rieseberg, L. H. (2020). Massive haplotypes underlie ecotypic differentiation in sunflowers. *Nature*, *584*(7822), Article 7822. https://doi.org/10.1038/s41586-020-2467-6
Turissini, D. A., McGirr, J. A., Patel, S. S., David, J. R., & Matute, D. R. (2018). The rate of evolution of postmating-prezygotic reproductive isolation in *Drosophila*. *Molecular Biology and Evolution*, *35*(2), 312–334. https://doi.org/10.1093/molbev/msx271
Via, S. (1999). Reproductive isolation between sympatric races of pea aphids. I. gene flow restriction and habitat choice. *Evolution; International Journal of Organic Evolution*, *53*(5), 1446-1457. https://doi.org/10.1111/j.1558-5646.1999.tb05409.x
Villoutreix, R., de Carvalho, C. F., Soria-Carrasco, V., Lindtke, D., De-la-Mora, M., Muschick, M., Feder, J. L., Parchman, T. L., Gompert, Z., & Nosil, P. (2020). Large-scale mutation in the evolution of a gene complex for cryptic coloration. *Science*, *369*(6502), 460–466. https://doi.org/10.1126/science.aaz4351
Walsh, J. B. (1982). Rate of accumulation of reproductive isolation by chromosome rearrangements. *The American Naturalist*, *120*(4), 510–532. https://doi.org/10.1086/284008
Weissensteiner, M. H., Bunikis, I., Catalán, A., Francoijs, K.-J., Knief, U., Heim, W., Peona, V., Pophaly, S. D., Sedlazeck, F. J., Suh, A., Warmuth, V. M., & Wolf, J. B. W. (2020). Discovery and population genomics of structural variation in a songbird genus. *Nature Communications*, *11*(1). https://doi.org/10.1038/s41467-020-17195-4
Wellband, K., Mérot, C., Linnansaari, T., Elliott, J., Curry, R., & Bernatchez, L. (2018). Chromosomal fusion and life history-associated genomic variation contribute to within-river local adaptation of Atlantic salmon. *Molecular Ecology*, *28*. https://doi.org/10.1111/mec.14965
Wellenreuther, M., & Bernatchez, L. (2018). Eco-evolutionary genomics of chromosomal inversions. *Trends in Ecology & Evolution*, *33*(6), 427–440. https://doi.org/10.1016/j.tree.2018.04.002
Westram, A. M., Stankowski, S., Surendranadh, P., & Barton, N. (2022). What is reproductive isolation? *Journal of Evolutionary Biology*, *35*(9), 1143–1164. https://doi.org/10.1111/jeb.14005
Yeaman, S., Aeschbacher, S., & Bürger, R. (2016). The evolution of genomic islands by increased establishment probability of linked alleles. *Molecular Ecology*, *25*(11), 2542–2558. https://doi.org/10.1111/mec.13611
Yeaman, S., & Whitlock, M. C. (2011). The genetic architecture of adaptation under migration-selection balance. *Evolution; International Journal of Organic Evolution*, *65*(7), 1897–1911. https://doi.org/10.1111/j.1558-5646.2011.01269.x
Welch, J. J., & Jiggins, C. D. (2014). Standing and flowing: The complex origins of adaptive variation. *Molecular Ecology*, *23*(16), 3935-3937. https://doi.org/10.1111/mec.12859
Yoshida, K., Rödelsperger, C., Röseler, W., Riebesell, M., Sun, S., Kikuchi, T., & Sommer, R. J. (2023). Chromosome fusions repatterned recombination rate and facilitated reproductive isolation during *Pristionchus* nematode speciation. *Nature Ecology & Evolution*, *7*(3), Article 3. https://doi.org/10.1038/s41559-022-01980-z
Zhang, L., Reifová, R., Halenková, Z., & Gompert, Z. (2021). How important are structural variants for speciation? *Genes*, *12*(7), 1084. https://doi.org/10.3390/genes12071084

Chapter I

Multiple structural variants introgressed from a Southern Atlantic lineage differentiate European anchovy ecotypes

---



Context

It has long been known that the European anchovy (*Engraulis encrasicolus*) presents two morphologically distinct forms (for a review see **Annex 2**). These morphotypes have been shown to correspond to a marine and coastal ecotype inhabiting different environments in the Bay of Biscay, as well as in the Mediterranean and Black Seas (Borsa 2002; Karahan et al., 2014; Oueslati et al. 2014). Le Moan et al. (2016) studied pairs of anchovy ecotypes in the Atlantic and Mediterranean Sea, showing that they demonstrated genetic parallelism and heterogeneous levels of divergence between markers. This study further detected the presence of multiple F1- and later-generation hybrids, demonstrating that gene flow between ecotypes takes place to a detectable degree. Despite the relative frequency of hybridisation, genetic differentiation between ecotypes is maintained, indicating that marine and coastal anchovies are partly reproductively isolated. In light of these findings, we recently published a taxonomic note in the *Journal of Fish Biology*, proposing that coastal anchovies be considered as a separate species, namely *E. maeoticus* (**Annex 2**). Even though it is not known to which degree ecotypes show irreversible reproductive isolation, current stock management policies do not take ecotypic subdivision into account, arguing for the need of species delimitation and the recognition of two separate fishing stocks.

Here, we use genome-scale data for the first time to address unanswered questions on the ecotypic structure in *E. cf. encrasicolus*. We present a large-scale study covering most of the species' distribution range, and leverage RAD-seq as well as whole-genome resequencing data to study genetic structure in this species. We examined the spatial arrangements of different pairs of ecotypes, in an attempt to reveal important elements for understanding the origin of the variation underlying ecotypic differentiation. We wished to shed light on the genomic architecture of divergence, and specifically, to test for the presence of structural variants. Furthermore, some studies have reported a particular genetic signature in populations around the Iberian peninsula, the Canary Islands (Zarraonaindia et al., 2012) and evoked the possibility of past exchanges with anchovies from the southern hemisphere (Grant et al., 2005). To test these hypotheses we included samples from South Africa (*E. capensis*) and considered the potential role of ancient admixture events in initiating ecotype speciation.

We provide an HTML report file with an interactive version of some of our results, which can be downloaded at the following link:

https://cloud.isem-evolution.fr/nextcloud/index.php/s/s48bM8RQnz3ELMQ

Abstract

In the marine realm, the absence of strong geographic barriers to dispersal can sometimes promote genetic homogeneity across large distances. Extreme cases of taxa with trans-equatorial distributions may however encounter suitable habitat discontinuities, leading to the emergence of sister species with anti-tropical distributions. The European anchovy (*Engraulis encrasicolus*) presents such a pattern, with its closely related species, *E. capensis*, distributed along the African coast in the southern hemisphere. Despite its large dispersal capacities, *E. encrasicolus* has also been shown to present fine-scale ecological structure, with the existence of a marine and coastal ecotype in the Atlantic and Mediterranean Sea. These ecotypes are known to co-occur in parapatry and hybridise, yet, genetic differentiation is maintained between ecotype lineages. This poses the question of the evolutionary mechanisms and genomic architecture underlying ecotypic differentiation. Here, we present the first genome-scale study investigating genetic structure in *E. encrasicolus*. We generated a reference genome for the species and produced whole-genome resequencing data for anchovies from the North-East Atlantic and Mediterranean Sea, as well as from South Africa. We complimented this approach with the analysis of RAD-seq data in order to study ecotypic structure across the entire distribution range. We found that genetic diversity is not only characterised by the presence of two genetic clusters, namely the marine and coastal ecotypes, but also by a third ancestry which corresponds to a Southern Atlantic lineage. This lineage occurs off South Africa but also in Morocco and the Canary Islands, and shows a gradient of admixture with European anchovies nearing the Atlantic-Mediterranean transition zone. In terms of the genomic architecture of differentiation, our analyses showed highly heterogeneous divergence landscapes between the Southern lineage and the European lineages, as well as between *E. encrasicolus* ecotypes. These landscapes showed evidence for large regions of high linkage disequilibrium, likely representing multiple structural variants that differentiate the three anchovy lineages. Furthermore, some haplotypes carried by the coastal ecotype, which distinguish it from the marine ecotype, showed similarity to the southern lineage, indicating the contribution of introgressed variants in the formation and divergence of anchovy ecotypes.

Introduction

The long-held view of genetic homogeneity in marine species promoted by large dispersal capabilities and high environmental connectivity has been challenged by observational evidence. An ever-increasing number of studies have identified the existence of genetic structure within marine species at different spatial scales (Hellberg, 2009; Palumbi, 1994). At first, oceanographic connectivity was considered to be a major driver of these patterns (Selkoe et al., 2016), as many marine species disperse with the currents primarily during their pelagic larval stage. However, this hypothesis is now being challenged by the analysis of genome-scale data. A salient result of marine population genomics has been to show that, contrary to prevailing demographic models that predict genetic differentiation across the entire genome, marine populations instead tend to show highly heterogeneous genomic landscapes of differentiation (Bradbury et al., 2013; De Jode et al., 2023; Duranton et al., 2018). In the presence of high gene flow erasing differentiation in neutral regions of the genome, divergence concentrated in specific loci indicates that they could

be under some form of selection, and that genomic islands of divergence are evidence of ongoing speciation (Ravinet et al., 2017).

Population genomics studies have now demonstrated that numerous marine species are actually subdivided into partially reproductively isolated entities, taking the form of geographic lineages, cryptic species or ecotypes (Bierne et al., 2011; Gagnaire et al., 2015; Johannesson, Le Moan, et al., 2020). Among these, cases of ecotypic differentiation have been studied with particular attention given to phenotypic adaptations, genomic architecture, and evolutionary history (Berg et al., 2016; Han et al., 2020; Johannesson, 2016; Moan et al., 2016). These studies have collectively provided a far more complex picture of differentiation and adaptation in the sea, which often depends on a mix of factors including ancestral divergence, structural variation, ecological selection and the acquisition of genetic variation through introgression (Duranton et al., 2020; Foote et al., 2019; Johannesson, Butlin, et al., 2020). This raises an important question about the evolutionary origins of marine ecotypes: how do historical, spatial, and ecological factors, as well as genome architecture interact in the emergence of partial reproductive isolation between ecotypes?

The European anchovy (*Engraulis encrasicolus*) is a small pelagic fish that occupies the warm-temperate near-shore and marine habitats of the North-East Atlantic, Mediterranean and Black seas. Ecotype subdivision between a marine and coastal anchovy form has been described based on body coloration, morphology, and allozyme data (Borsa, 2002). This ecotypic structure has been confirmed by other studies using molecular markers such as microsatellites (Huret et al., 2020; Karahan et al., 2014; Oueslati et al., 2014) and single-nucleotide polymorphisms (SNPs) (Catanese et al., 2017; Zarraonaindia et al., 2012). Using restriction-site associated DNA sequencing (RAD-seq), Moan et al. (2016) showed parallel genetic divergence of outlier loci between replicate ecotype pairs from the Bay of Biscay and the northwestern Mediterranean Sea. Combined with the presence of hybrids and admixed genotypes, these results indicated that differentiation between ecotypes is maintained by genetic barriers to gene flow. Moreover, historical demographic inference showed that replicate ecotype pairs share a common history of divergence, which has been followed by postglacial secondary contact. Given that anchovy ecotypes appear sufficiently reproductively isolated to maintain their genetic integrity despite gene flow, taxonomical recognition of the two ecotypes as separate species has been proposed, namely as the marine *E. encrasicolus* and the coastal *E. maeoticus* (Bonhomme et al., 2021).

Despite these significant advances in our understanding of the genetic structure of anchovies along the European coasts, the population structure at the southern margin of the species range remains unclear. Classically, the southern range limit of *E. encrasicolus* along the western African coast is considered as the south of the Gulf of Guinea. However, some authors consider the Southern Benguela system off South Africa as the southern distribution limit, thus equating the two anti-tropical species *E. encrasicolus* and *E. capensis* to a single species (Raybaud et al., 2017). While anti-tropical species distributions are common in the marine realm (Grant & Leslie, 2001; Ludt, 2021), low-resolution molecular studies effectively showed that anchovy genetic structure along Western Africa could potentially be associated to a single species with genetically diverse populations. Grant et al. (2005) showed that *E. capensis* and *E. encrasicolus* shared two mitochondrial lineages that could reflect transequatorial dispersal within- as well as among

species. Nuclear data reported very low genetic differentiation between northern and southern hemisphere anchovies, with the occurrence of *E. capensis*-like genotypes around the Atlantic-Mediterranean transition zone (Silva et al., 2017; Zarraonaindia et al., 2012). These findings could indicate that gene flow occurs between the two hemispheres, with still unknown consequences on the subdivision between *E. encrasicolus* and *E. maeoticus*. Here, we used whole-genome sequencing to provide a description of the genomic architecture associated with divergence between anchovy ecotypes and species. We combined this approach with RAD-seq data to further describe the eco-geographic structure of anchovies across their whole distribution range, and aimed to determine whether and how they have genetically interacted during their evolutionary history.

Materials and methods

Sampling and DNA extraction

Samples were collected from across the species range (Locations table, **Supplementary Table S2**) and were issued from various sampling expeditions and local fisheries (**Supplementary Table S1**, Type==”Tissue”). These samples were collected in different types of habitats, which were classified either as coastal (lagoons or estuaries) or marine habitats. Also included in our sampling scheme were eight samples of *E. capensis* collected off the South African coast (Gqeberha). Whole genomic DNA was extracted from muscle tissue or fin clips using commercial tissue kits (Qiagen and Macherey-Nagel). Extraction quality was checked on agarose gel for the presence of high molecular weight DNA, and double-stranded nucleic acid concentration was measured using Qubit 2.0 and standardised in concentration before library construction.

Reference genome assembly

We performed high-coverage linked-read sequencing of a marine Atlantic *E. encrasicolus* individual from the Faro location (Algarve) to generate a high-quality reference genome assembly (Eencr_V1), following the same methodology as for the seahorse reference genome (Meyer et al., 2023, **Chapter II**). The total length of all assembled scaffolds was ~926 Mb (925,873,119 bp), but scaffold N50 reached only 20.36 Kb, indicating a rather fragmented assembly. We thus restricted all analyses to scaffolds that were longer than 10 kb (9293 scaffolds totalling 177,043,243 bp, i.e. 20% of the assembled genome). These scaffolds were anchored to the chromosome-level assembly of the Japanese grenadier anchovy (*Coilia nasus*) (GenBank assembly accession: GCA_027475355.1), a related Engraulid species from the Northwest Pacific (**Supplementary Fig. S1**). Whole-genome alignment was performed with Minimap2 (Li, 2018) and visualised using D-GENIES (Cabanettes & Klopp, 2018).

Whole-genome resequencing data

Thirty-nine samples (**Supplementary Table S1**, WGS==”yes”) were selected for whole-genome sequencing (WGS), including samples from coastal and marine habitats in the Atlantic (*GAS*) and

the Mediterranean Sea (*GDL*, *SPN*). We also included samples from the Atlantic-Mediterranean transition zone (*PRS*) and *E. capensis* samples to investigate genetic composition in these regions. Individual whole-genome sequencing libraries were prepared following the Illumina TruSeq DNA PCR-Free Protocol and sequenced to an average depth of ~10-30X on an Illumina NovaSeq 6000. Raw demultiplexed reads were processed using fastp (v0.19.05) (Chen et al., 2018) and aligned to our reference genome using BWA-MEM (BWA v0.7.17; Li, 2013). Picard (v2.26.8) ("Picard toolkit", 2019) was used for sorting read alignments, marking duplicates and adding read groups.

Variants were called using the GATK best practices workflow (McKenna et al., 2010; Van der Auwera et al., 2013). Firstly, individual GVCF files were created from bam files with HaplotypeCaller (GATK v.4.3.0.0). This information was then stored in a GVCF database using GenomicsDBImport, and VCF files (one file per scaffold) were generated with GenotypeGVCFs. These files were concatenated into a single VCF file which was filtered to contain only high quality SNPs (VCFtools v0.1.16) (Danecek et al., 2011). This included recoding genotypes as missing for low-quality sites (*--minGQ 20*) and hard-filtering sites based on their normalised variant quality, average genotype quality, mapping quality, strand bias, and average depth (greater than 90X,c corresponding to the 97.5th quantile). The VCF was also filtered for indels, multiallelic SNPs and missing data ("--max-missing 0.85"). The final VCF file (hereafter referred to as the WGS dataset) contained ~5,7 M sites located on 8480 different scaffolds (longer than 10 kb and mapping to *C. nasus*).

RAD sequencing data

RAD-sequencing libraries were prepared for 243 samples (**Supplementary Table S1**, RAD=="yes" and Type=="Tissue") following a similar protocol to Baird et al. (2008). Twenty-five of these samples were also used to produce WGS data, providing a link to understand the genetic structure in both datasets. For the RAD data, sequencing was performed on an Illumina HiSeq2500 sequencer in single-read mode. To complement our sampling, we also included data for 128 samples from Le Moan et al. (2016). The data was demultiplexed using process_radtags (Stacks v2.60) and reads were aligned to our reference genome using BWA-MEM (BWA v0.7.17; Li, 2013). The reference-based Stacks pipeline was constructed using tools from the MBB framework (https://web.mbb.cnrs.fr/subwaw/workflowmanager.php) (Penaud et al., 2020). Gstacks was run using default parameters ("--model marukilow --var-alpha 0.05 --gt-alpha 0.05 --max-clipped 0.2") and minimum PHRED-scaled mapping quality set to 20 ("--min-mapq 20"). Thereafter, genotypes were finally exported in VCF using the populations module ("--min-populations 2 --min-samples-per-pop 0.7 --min-maf 0.05 --max-obs-het 1") and filtered to not contain more than 15% missing data. The VCF was also filtered to only contain sites that were present in the WGS dataset, since our objective was to describe the same genetic variation but at a larger geographic scale. Lastly, all samples from the WGS dataset were integrated into the RAD VCF. The final VCF file (hereafter referred to as the RAD dataset) contained genotype data for 385 samples at 3880 variable sites.

Population structure

To describe the genetic structure in both WGS and RAD datasets, we conducted chromosome-wide Principal Component Analysis (PCA) using the R package SNPRelate (v1.28.0) (Zheng et al., 2012) and calculated individual heterozygosity per chromosome using VCFtools (v0.1.16) (Danecek et al., 2011). We used ADMIXTURE (Alexander et al., 2009) to estimate ancestry proportions in all samples using the RAD dataset. Genetic differentiation (F_{ST}), nucleotide diversity (π) and absolute genetic divergence (d_{XY}) were calculated for the WGS data in non-overlapping 5 kb windows (with "-m 15") using the popgenWindows.py script (Martin, 2018; https://github.com/simonhmartin/genomics_general).

Characterisation of structural variants

Our analyses investigating the population structure in our data revealed evidence for the presence of structural variants (SVs). In order to determine which genomic regions displayed these SV-like patterns, we conducted local PCA for the WGS dataset in non-overlapping windows of 5 kb using lostruct (v0.0.0.9; Li & Ralph, 2019). We detected windows associated with patterns of high linkage disequilibrium (LD), and potentially SVs, by calculating the change in the position on PCA axis 1 from one window to the next for all samples (i.e. the slope of each individual line in a plot of PCA 1 over window number). A window was considered to show LD patterns if the mean slope over all the samples was less than 0.07 (this cut-off was determined visually).

We further wished to characterise haplotype combinations in all samples for chromosomes displaying SV-like patterns. In order to assign SV "genotypes" to individuals in both the WGS and RAD datasets, we based our classifications on the (chromosome-wide) PCA coordinates and heterozygosity values of each sample (**Supplementary Fig. S2 and S3**). For samples that had WGS data, we corroborated these results with the relative positions of each sample in local PCA.

Results

Genetic structure of the European anchovy - not two but three distinct ancestries

We produced RAD-seq data with a mean per sample coverage of 52.7X that was used to describe the overall genetic structure in the full dataset of 385 individuals (**Fig. 1A**). This analysis gives, for the first time, a clear picture of the complex genetic structure of the European anchovy, and gives very similar results to those obtained using our WGS data (10-30X). Firstly, we observed differentiation between samples collected in marine and coastal habitats, corresponding to previously described ecotypes. This could be observed along the second axis of variation (PCA 2), where coastal samples were positioned at the top of the plot and marine samples in the bottom right corner. As for PCA 1, this axis showed a different genetic structure that has been less well described and which reflects geographic structure rather than ecological structure. On this

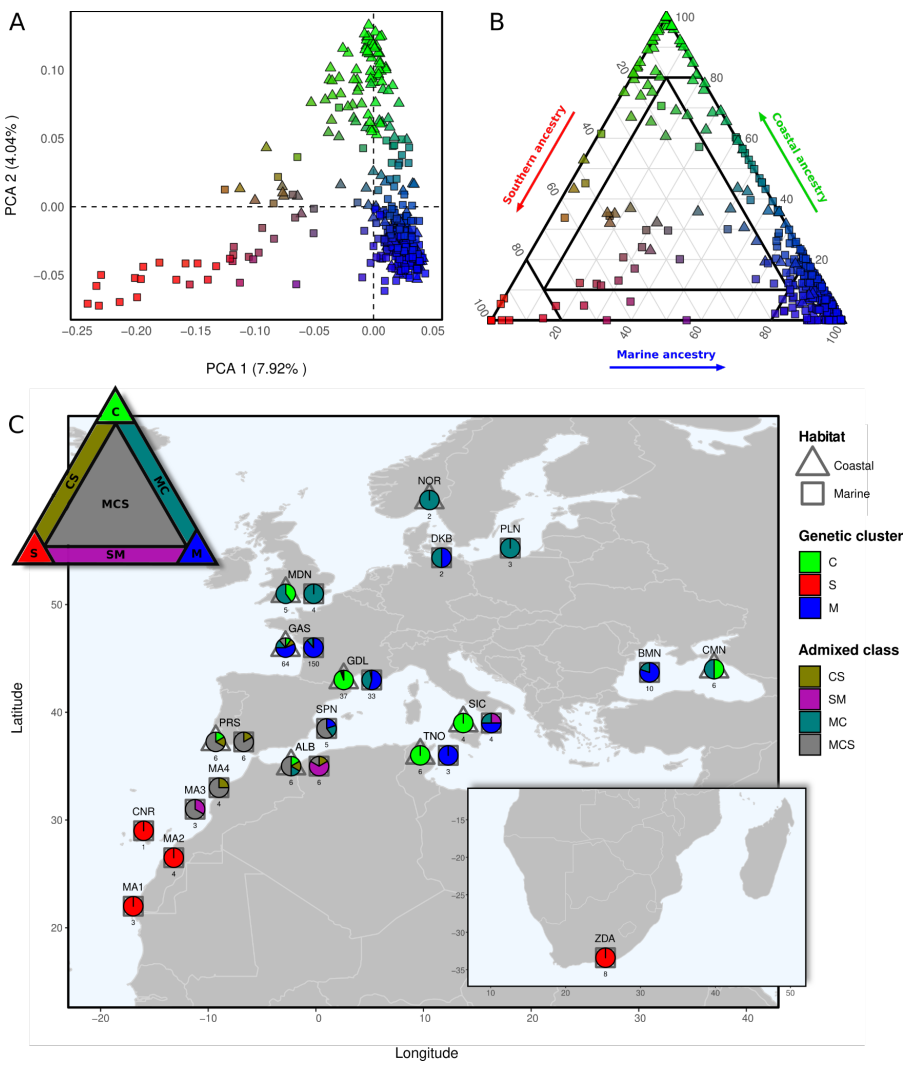


Fig. 1. A) Principal Component Analysis (PCA) performed on the entire dataset of 385 anchovy samples. Sites used in the analysis were high-quality variants present both in the whole-genome data as well as in the RAD data, corresponding to a total of 3881 SNPs. Shapes indicate habitat type and colours reflect ancestry proportions as determined by admixture analysis (see B). B) Ternary plot showing the admixture level between three genetic ancestries: coastal (green), Southern (red), and marine (blue) ancestry. Coordinates, as well as RGB colours, reflect the relative ancestry proportions of samples along each of the three axes. Samples were classified as belonging to a genetic cluster (black lines demarcating seven areas) based on their position in the plot. Clusters *C*, *S* and *M* represent "non-admixed" parental lineage ancestries, while *CS*, *SM*, *MC* and *MCS* represent various levels of admixture. B) Map with sampling locations where symbols represent habitat type and pie charts show the proportions of different genetic clusters present. Numbers beneath pie charts indicate sample sizes. Location are described in **Supplementary Table S2**.

horizontal axis, South African samples and other individuals collected off the African Atlantic coast (e.g. Morocco) were spread out towards the left-hand side of the plot, with the majority of European samples grouping to the right. Similar results were obtained using ADMIXTURE (**Fig. 1B**) with K=3 showing that samples carried varying proportions of coastal (top), marine (right)

and Southern (left) ancestry. Visualised as a ternary plot, this analysis showed considerable levels of admixture between the three ancestries, in particular between the marine and coastal clusters. A large number of samples also fell in the middle of the plot, reflecting relatively balanced proportions of different ancestry components in these individuals.

Based on their ternary coordinates, samples were classified as belonging to one of seven genetic classes which each corresponded to an area on the plot (black demarcations **Fig. 1B**, triangle in **Fig. 1C**). A sample was considered to belong to a given genetic cluster (*C*: green; *S*: red; *M*: blue) if that ancestry reached more than 80% of total genetic ancestry. Secondly, we distinguished three different classes of admixed individuals where ancestry proportions were mainly made up of two ancestries (the third not amounting to more than 10%). These classes were *CS* (admixed between *C* and *S*; khaki), *SM* (admixed between *S* and *M*; purple) and *MC* (admixed between *M* and *C*; seagreen). The last admixed class, *MCS*, consisted of individuals with more balanced proportions of all ancestries (admixed between *M*, *C* and *S*; grey). Subdivision into categories allowed us to summarise the genetic variation present at each sampling location and the resulting map showed the eco-geographical distribution of the three genetic clusters (**Fig. 1C**). We observed that individuals belonging to the *C* cluster (green) were only found in coastal habitats, while *M* individuals (blue) mainly occurred in marine environments, corresponding to the two anchovy ecotypes. This pattern was especially marked in the Mediterranean Sea, where almost all coastal samples were part of the *C* cluster (e.g. triangles at *SIC*, *TNO* and *GDL*). However, this signal of ecotypic differentiation becomes diluted nearer to the Atlantic-Mediterranean boundary, where we observe a gradient of increasing Southern ancestry. This introgression gradient can be seen through the increasing proportion of *MCS* individuals (grey) in the Alboran Sea (*ALB*), off the southern coast of Portugal (*PRS*) and in northern Morocco (MA4 and MA3). Finally, we observed that samples from locations to the south of the Canary islands (*CNR*), including the sampling site in South Africa (*ZDA*, inset map), all belonged to the *S* cluster (red).

Dividing our samples into genetic clusters and admixed classes further allowed us to study divergence between different units and to characterise the genomic architecture of differentiation. Using the WGS dataset, we reconstructed F_{ST} landscapes between individuals from the three genetic clusters (**Fig. 2**). We observed heterogeneous differentiation patterns for comparisons between the marine and coastal clusters (**Fig. 2A**), between the coastal and Southern clusters (**Fig. 2B**) and between the Southern and marine clusters (**Fig. 2C**). Differentiation landscapes were generally similar whether marine and coastal individuals originated from the Atlantic (first row of each comparison) or from the Mediterranean Sea (second row). However some notable differences between the Atlantic and Mediterranean were visible, for example on *CM050217* (**Fig. 2B**) and on *CM050215* (**Fig. 2C**). Furthermore, genetic differentiation was generally lower between the coastal and Southern clusters than between the marine and Southern clusters (*ATL*: mean F_{ST} 17% lower in **Fig. 2B** than in **Fig. 2C**; *MED*: mean F_{ST} 9% lower in **Fig. 2B** than in **Fig. 2C**). Most strikingly, this analysis revealed that 13 chromosomes (stars in **Fig. 2**) showed high differentiation along a substantial portion of their length (> 2.5% of windows with F_{ST} above the 95th quantile) in at least one comparison. We consider that these 13 high-differentiation chromosomes likely play an important role in divergence between different anchovy lineages and they were the main targets of subsequent analyses.

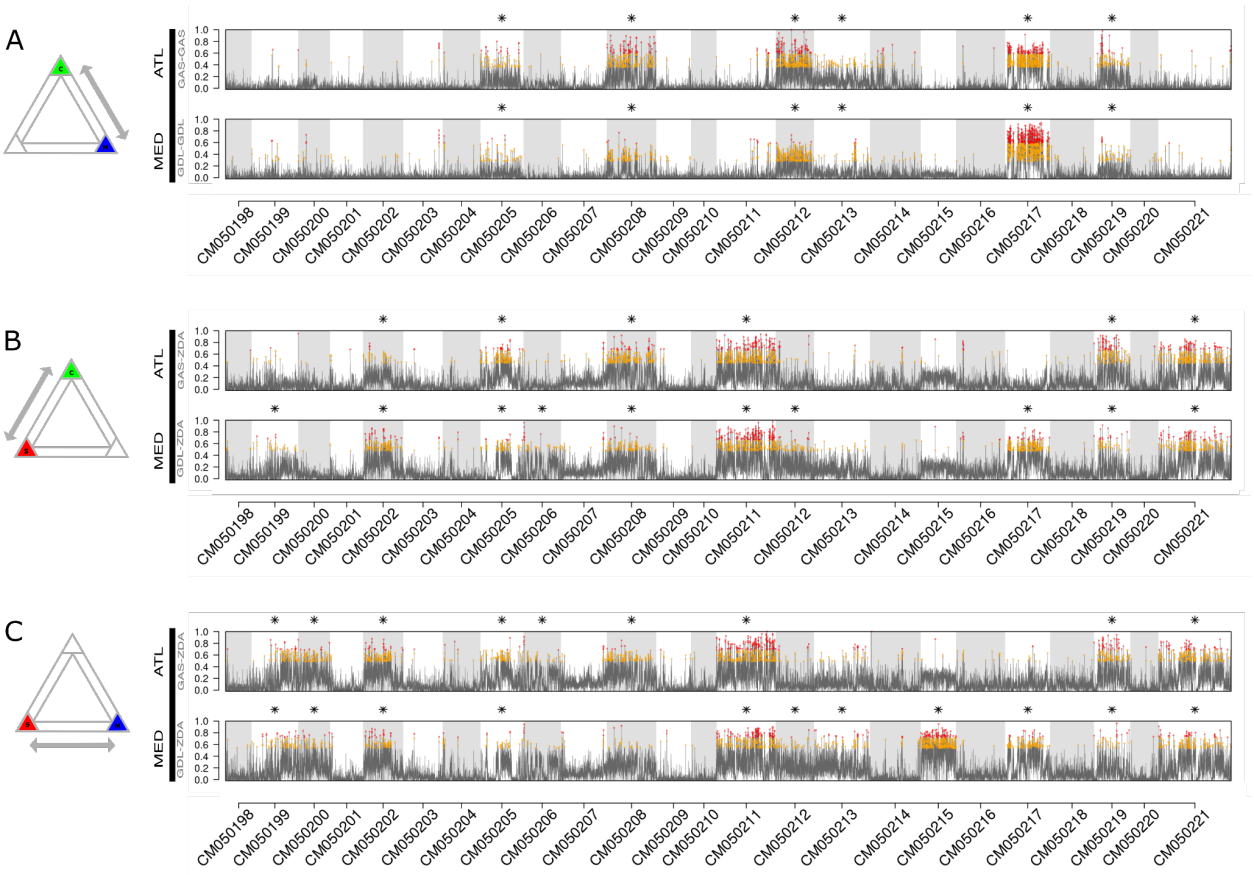


Fig. 2. Genomic landscapes of differentiation (F_{ST}) calculated in 5 kb sliding windows between groups of samples (3 individuals per group) from different genetic clusters (see Fig. 1). Differentiation landscapes are shown for three different comparisons (A: coastal vs. marine; B: coastal vs. Southern; C: Southern vs. marine). Each panel consists of two rows, representing cases where coastal/marine samples either originated from the Atlantic (ATL) or from the Mediterranean Sea (MED). Orange points are windows where F_{ST} was higher than the 95th quantile, while red points are above the 99th quantile. Stars indicate chromosomes where more than 2.5% of windows showed F_{ST} higher than the 95th quantile. Grey and white rectangles delimit 24 pseudo-chromosomes of *C. nasus*.

Genomic architecture of marine-coastal ecotype differentiation

F_{ST} landscapes between marine (cluster *M*) and coastal (cluster *C*) anchovies showed similar patterns in the Atlantic and Mediterranean Sea (top and bottom row of **Fig. 2A**). These results point to strong parallelism between marine and coastal ecotypes in different locations, since their differentiation largely involves the same genomic regions. In the comparison between marine and coastal individuals, we identified six chromosomes that showed high levels of differentiation (stars in **Fig. 2A**), contrasting with generally low F_{ST} on other chromosomes. To further investigate these genomic regions that differentiate ecotypes, we examined relationships amongst samples using PCA conducted at a chromosome-wide scale (**Fig. 3**).

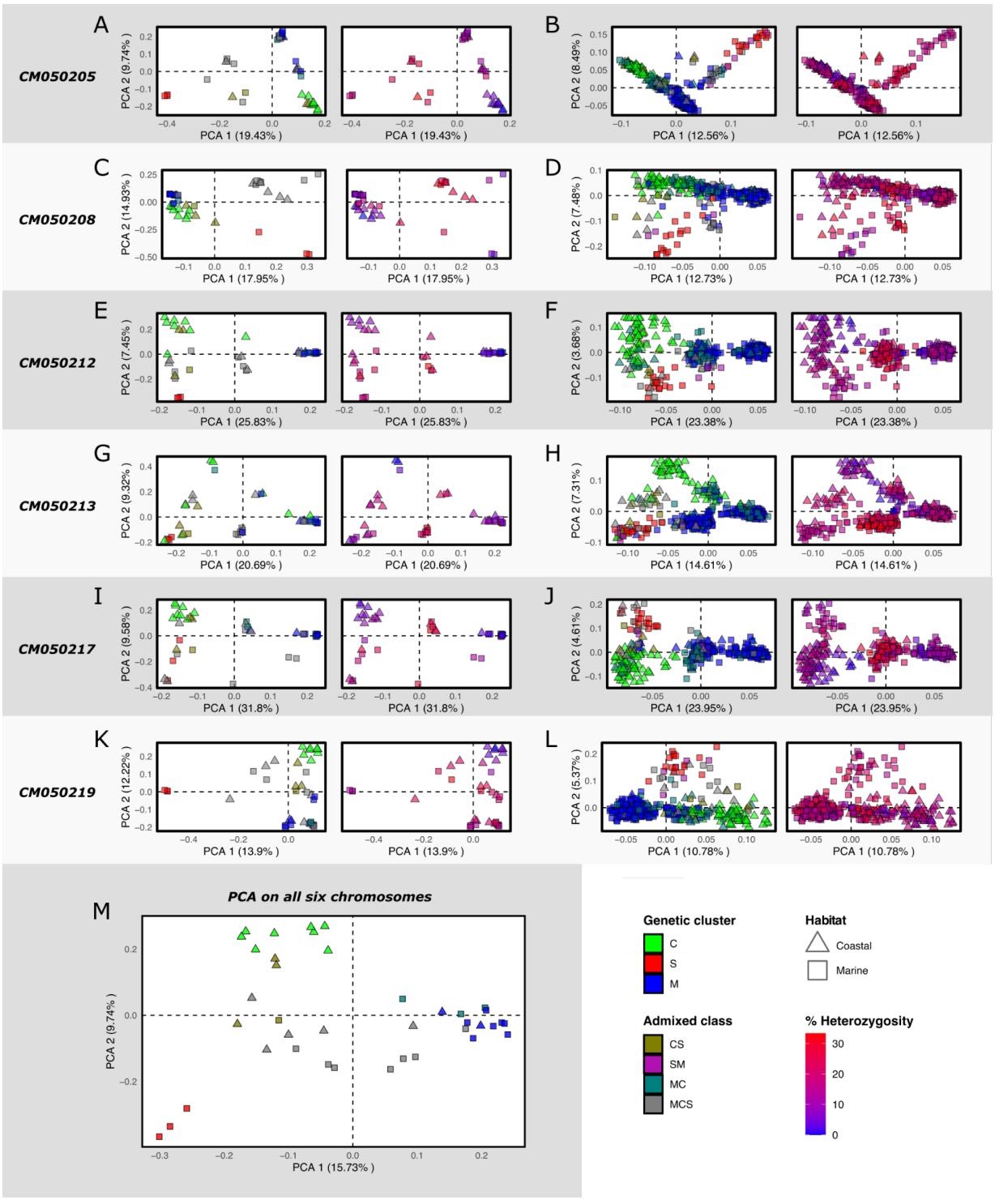


Fig. 3. PCA conducted on six chromosomes that differentiate coastal and marine genetic clusters. The first six lines show PCA at a chromosome-wide scale for each of the chromosomes (A-L), while panel *M* shows PCA performed on all six chromosomes simultaneously. The left-hand column (A, C, E, G, I, K, M) shows PCA for whole-genome data (n=39) while the right-hand column (B, D, F, H, J, L) shows results for RAD data (n=385). Horizontal and vertical axes correspond to PCA 1 and 2. Shapes indicate habitat type, while fill colour indicates either the individual's assigned genetic cluster (left panel of each column) or chromosome-wide heterozygosity (right panel of each column).

We found that PCA performed on the WGS dataset and on the RAD dataset yielded nearly identical results for almost all six chromosomes. We consistently observed differentiation among the coastal, marine and Southern clusters that either separated along PCA axis 1 or 2 in different configurations. The structure along PCA axis 1 either separated Southern individuals from all other samples (e.g. *CM050205*) or placed Southern and coastal individuals together on one side, separated from the majority of the marine samples (*CM050212*, *CM050213* and *CM050217*). The only chromosomes that showed a different structure when comparing results from the WGS and RAD datasets, were *CM050208* and *CM050219*. The latter chromosome showed similar amounts of variance explained by PCA axes 1 and 2, which could account for the fact that the genetic structure on each axis in the WGS and RAD PCA plots were switched. As for *CM050208*, by examining the positions of samples that were present in both datasets, we could conclude that the Southern individuals in the WGS dataset were in the centre of the plot in the RAD PCA. This could mean that the full structure, revealed by the large amount of samples in the RAD dataset, was not present in our WGS VCF (n=39).

The presence of multiple groups in PCA conducted on individual chromosomes (**Fig. 3**) contrasted with more continuous ancestry gradients detected in PCA using all markers (**Fig. 1A**). These multiple PCA clusters indicate that a large number of SNPs are in LD, resulting in the segregation of a number of non-recombining haplotypes. A combination of high F_{ST} and high LD is often associated with the presence of SVs. For the six chromosomes differentiating marine and coastal individuals, the presence of SVs was further supported by high levels of heterozygosity in intermediate groups (**Fig. 3**). These middle groups could represent heterokaryotes that carry two different haplotypes, while the groups of samples at opposite ends represent alternate homokaryotes. Overall, results thus suggest that marine and coastal anchovy ecotypes are differentiated by multiple SVs of large size occurring on multiple chromosomes.

Combinations at multiple high-LD regions differentiate anchovy lineages

In addition to the six chromosomes differentiating marine and coastal ecotypes, a number of other chromosomes showed high F_{ST} in comparisons involving the *S* cluster (stars in **Fig. 2B and 2C**). Since some of our results suggested the presence of SVs (see the previous section), we extended these analyses to screen all 13 high-differentiation chromosomes for similar patterns. PCA and individual heterozygosity in these regions again confirmed a high abundance of SVs. We therefore aimed to characterise SVs on all 13 chromosomes and to assign haplotype combinations in all samples. This methodology is illustrated in **Fig. 4**, where we describe haplotype combinations on *CM050213*. Here, three types of homokaryotes (*00*: gold; *11*: pink; *22*: brown) defined a triangular pattern in the PCA plot (**Fig. 4A**). On the other hand, intermediate groups along each of the three sides represented heterokaryotes (*01*, *12*, *02*) that showed increased heterozygosity levels (**Fig. 4B**). In this way, we assigned haplotype combinations (colours) to the six groups in order to study the distribution of our samples amongst them.

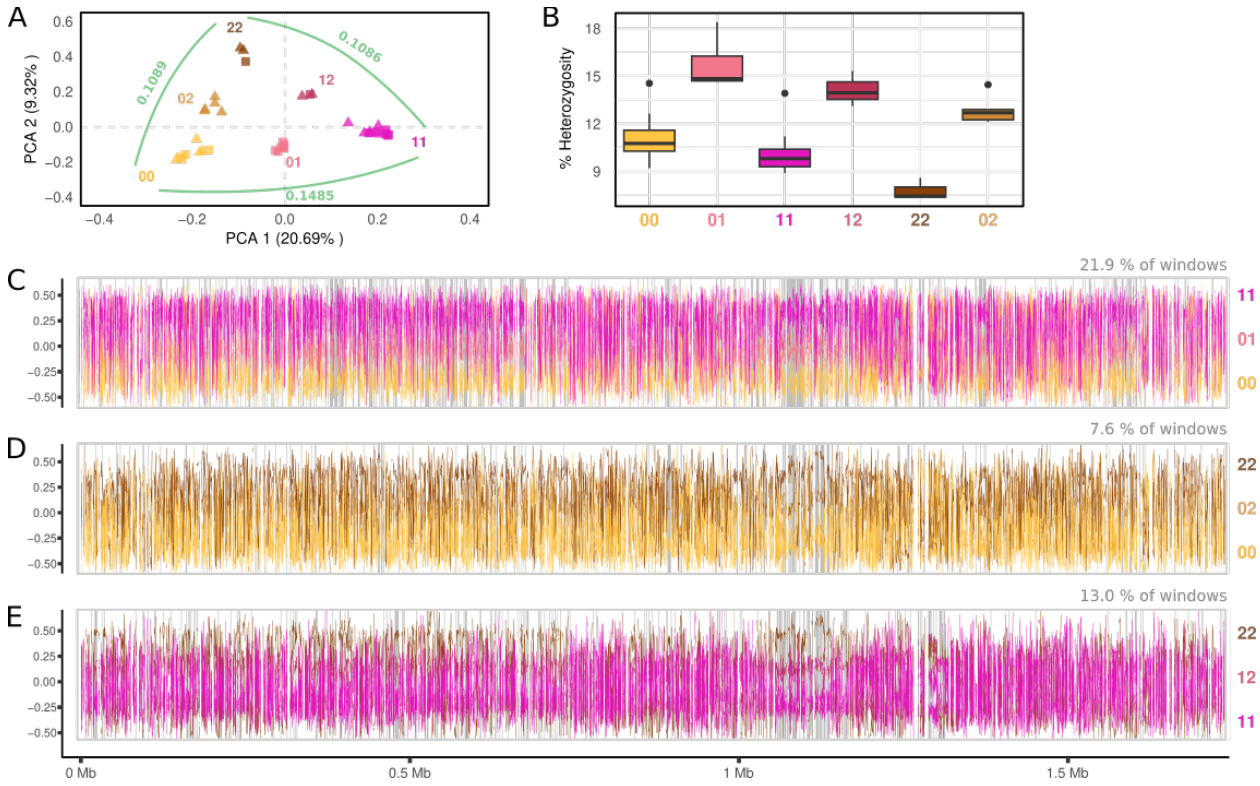


Fig 4. Example of haplotype assignment for WGS data on *C. nasus* chromosome *CM050213*. A) Chromosome-wide PCA showing six groups of individuals, corresponding to six different haplotype combinations: *00* (gold), *11* (pink) and *22* (brown) homokaryotes, and *01*, *12* and *02* heterokaryotes (intermediate colours). Green text shows mean chromosome-wide *Dxy* (%) between opposite homokaryotes (connected by green lines). B) Individual heterozygosity on *CM050213* for samples with different haplotype combinations. C-E) PCA 1 coordinates from local PCA plotted for non-overlapping 5 kb windows along the chromosome. Local PCA was conducted on different subsets of individuals according to their haplotype combination: only *00*, *01* and *11* individuals (C); only *00*, *02* and *22* individuals (D); and only *11*, *12* and *22* individuals (E). Individual lines represent samples and are coloured according to their genotype. Grey rectangles indicate windows displaying patterns associated with SVs and grey text indicates their relative percentage out of all windows on the chromosome.

The same classification was performed for all 13 chromosomes using both the WGS and RAD datasets (**Supplementary Fig. S2 and S3**). We consistently found evidence for three different haplotypes, except on chromosomes *CM050211*, *CM050215* and *CM050221*, which showed only two haplotypes. For consistency, we assigned the haplotype combination *00* to the group with the most Southern samples, *11* to the group with the marine samples, and *22* to the coastal samples. This way of characterising ancestry at the chromosome-level allowed us to study overall patterns in 385 samples (**Fig. 5**). Here, we observe that individuals (vertical bars) belonging to a given genetic cluster mostly showed similar haplotype combinations. For example, the samples in the coastal cluster were mostly *22* homokaryotes (brown) on the six chromosomes involved in ecotype differentiation (**Fig. 3**). One exception to this pattern occurred on *CM050213*, where coastal individuals in the Bay of Biscay (*GAS*) often carried the *0* haplotype. Indeed, *0* haplotypes were generally more common in the Atlantic than in the

Mediterranean. This was true across several different chromosomes and for both Atlantic marine and coastal individuals. As for the Southern cluster, these individuals were differentiated from marine and coastal individuals on largely all chromosomes (except on *CM050206* which displayed shared polymorphism), mainly presenting *00* homokaryotes (gold). In locations such as Southern Portugal (*PRS*) and the Alboran Sea (*ALB*), we observed breakdown of usual interchromosomal associations in the admixed class with *MCS* individuals. These samples presented a mixture of all haplotype combinations (haplotypes *0*, *1* and *2* in all combinations), probably due to extensive admixture between the three lineages.

In order to investigate which types of SVs could explain the presence of non-recombining haplotypes, we performed local PCA in sliding windows. This allowed us to identify which genomic windows showed patterns associated with SVs, for example, on chromosome *CM050213* (**Fig. 4C-E**). In order to specifically discern windows differentiating each of the three haplotypes (i.e. which windows underlie the variation along each side of the triangle in **Fig. 4A**), we included different subsets of individuals in the analysis. In local PCA performed on *00*, *01* and *11* individuals (horizontal axis of variation, **Fig. 4C**), we found that many windows showed a pattern typical of inversions. These windows were characterised by three groups of horizontal lines, representing the three clusters usually observed in a PCA of an inversion region. These windows showed that *01* heterokaryotes occupied intermediate positions between the *00* (bottom) and *11* (top) homokaryotes. This structure was observed in at least ~20% of windows on the chromosome. Different results were obtained when performing local PCA on *00*, *02* and *22* individuals (**Fig. 4D**) and *11*, *12* and *22* individuals (**Fig. 4E**). Here, the percentage of windows showing a three-cluster pattern was reduced compared to **Fig. 4C**. This observation was consistent with d_{XY} values calculated between different types of homokaryotes, where divergence was highest between *00* and *11* individuals (**Fig. 4A**). Taken together, these results suggest that there are multiple SVs present on chromosome *CM050213*, and that these vary in size and/or level of divergence. This could be explained by the presence of nested inversions, since SV windows that underlie the differentiation of *22* individuals (variation along PCA axes *00-02-22* and *11-12-22*) were located within the larger inversion (axis *00-01-11*). However, the quality of our genome assembly and lack of contiguity information limited our ability to make further interpretations about the nature of the SVs involved.

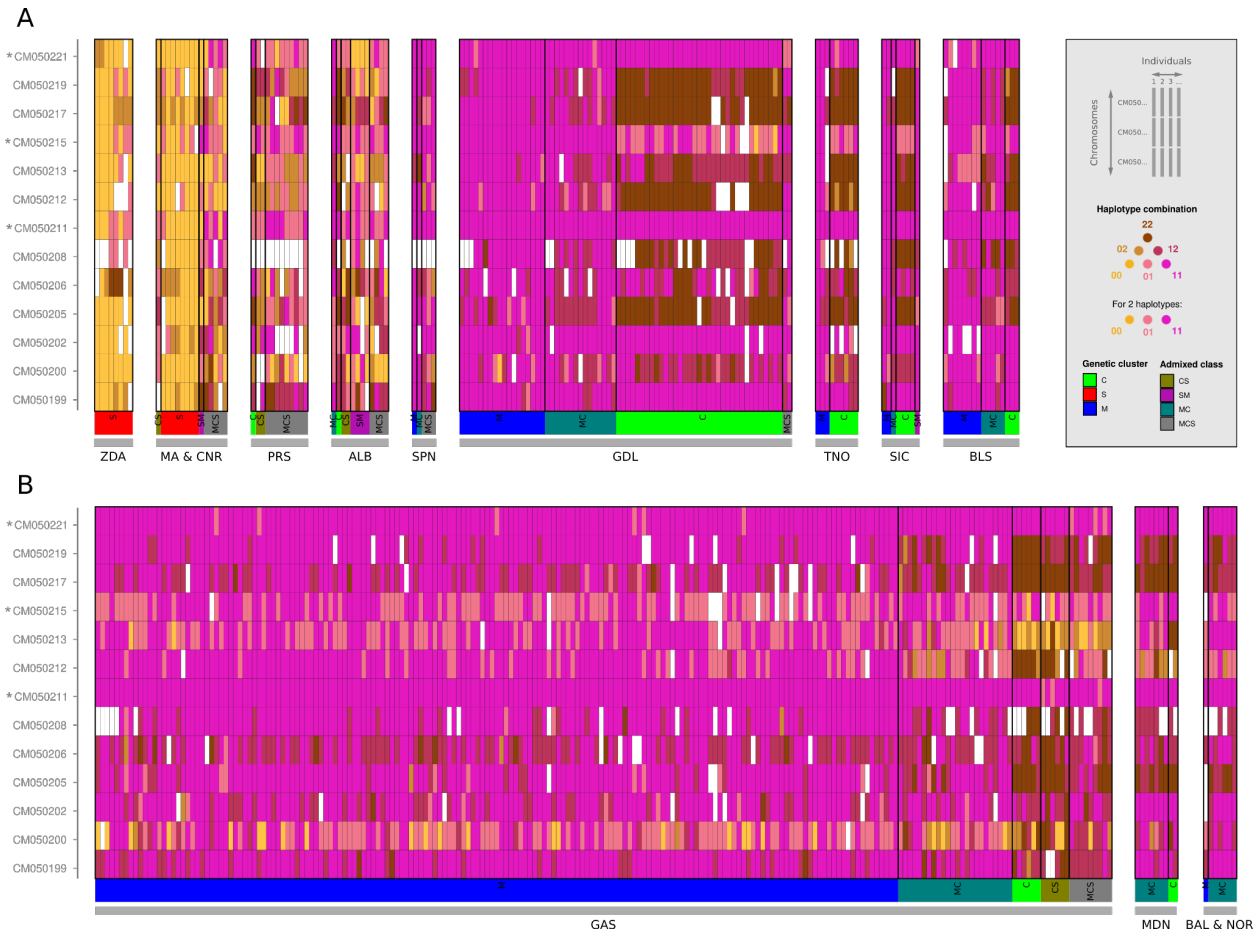


Fig. 5. Haplotype combinations in all samples (WGS and RAD datasets, n=385). Vertical bars correspond to individuals and each row shows the haplotype combination on a specific chromosome (13 in total). Colours correspond to three different homokaryotes (*00*: yellow; *11*: pink; *22*: brown) and three types of heterokaryotes (*01*, *12* and *02*; intermediate colours). Chromosome marked with an asterisk (*) only presented two haplotypes. Unfilled cells (white) indicate that it was not possible to assign the haplotype combination on the relevant chromosome in this sample. Samples are ordered according to their location (grey horizontal bars) and according to their genetic cluster or admixed class (coloured bars at the base). A) Haplotype combinations in locations going from the South Atlantic into the Mediterranean and Black Sea. B) Haplotype combinations in locations from the Bay of Biscay to the Baltic Sea.

Discussion

Three components of anchovy genetic diversity: a Southern lineage in addition to marine and coastal ecotypes

Previous genetic studies have shown that the European anchovy is subdivided into marine and coastal ecotypes that are present from the Bay of Biscay, through the Mediterranean to the Black Sea. Here, we show that there is a third component to genetic structure in this species, corresponding to an Atlantic lineage occurring off the African coastline. This Southern lineage shows genetic homogeneity at a large spatial scale, with genetic similarity between individuals sampled in Morocco, the Canary Islands and even as far as South Africa. However, from Northern Morocco and Southern Portugal into the Alboran Sea, we observe a gradient of decreasing Southern ancestry. Zarraonaindia et al. (2012) also reported the presence of a particular genetic structure in this region, but interpreted this signal as being due to the presence of differentiated populations inhabiting narrow-shelf waters associated with upwelling. We instead propose that this region corresponds to a contact zone between the Southern and European (marine and coastal) lineages. Here, we observe introgressive hybridisation and post-F1 gene flow between the Southern, coastal and marine clusters, as can be seen by gradual ancestry gradients in the PCA plot (**Fig. 1**). Almost all individuals that were identified as belonging to the three-way admixed class (*MCS*) were sampled in this zone (**Fig. 1C**). This pattern of three-way admixture also extends further north into the Bay of Biscay, where a few individuals carrying a *MCS* background are detected, similar to Zarranoianda et al. (2012). The existence of gene flow between ecotypes has already been illustrated in previous work (Le Moan et al., 2016), but our results show that there is also admixture with a third genetic cluster that contributes to diversity patterns. Previous studies reporting strong structure in this species may have unknowingly captured some of this three-way variation, leading to many different and often incoherent interpretations in the literature.

Multiple structural variants underlie ecotypic differentiation

A major aspect of many speciation genomic studies has been to identify the genomic architecture that differentiates lineages. The genomic regions underlying divergence between marine and coastal ecotypes in the European anchovy have remained unknown, but the first clues came from studying differentiation at RAD-derived SNP markers (Le Moan et al., 2016). These authors showed that divergence was heterogeneous across the genome, and that increased differentiation was limited to 20-25% of loci. In the current study, we present the first genome-scale investigation into divergence landscapes using a newly generated genome assembly. Our reference genome showed a high degree of fragmentation, which could be mitigated by anchoring our scaffolds to the chromosomes of a closely related Engraulid species. Furthermore, our analyses were limited to longer scaffolds, so we focused our analyses on large-scale patterns (and not finer intrachromosomal variation). Despite limited resolution, we found that ecotypic differentiation was mainly concentrated on six chromosomes (totalling 25% of the genome) (**Fig. 2A**). These chromosomes showed high differentiation values along a large portion of their length, contrasting with the patterns observed on other chromosomes (a few differentiation peaks in

otherwise homogeneous regions). Furthermore, the comparison of different ecotype pairs in the Atlantic and in the Mediterranean Sea revealed high degrees of parallelism between these distant locations, confirming that ecotypic differentiation largely involved the same genomic regions.

Groupings in PCA (**Fig. 3A-L**) and sustained clustering patterns across many local PCA windows (**Fig. 4C-E**) were indicative of high levels of linkage disequilibrium. These patterns are highly suggestive of SVs occurring on multiple chromosomes. We thus present evidence that large SVs contribute to ecotypic differentiation in anchovies. What is more, we found that divergence with the Southern lineage also involves similar architectures, implying an important role of SVs in lineage differentiation in general. This is further supported by the fact that samples that were classified as being admixed between the Southern and European lineages, presented heterokaryotypes at multiple SVs. In particular, genetic diversity and differentiation patterns on many chromosomes indicated the presence of not two, but three non-recombining haplotypes (Ishigohoka et al., 2021). These patterns could represent nested SVs, with one rearrangement having taken place within another larger rearranged segment (**Fig. 4C-E**). For such chromosomes, we could expect the variation along the first axis of PCA to represent different karyotypes at the largest or most divergent SV. This should also be reflected by higher heterozygosity values for the middle groups on this axis. We find evidence for this pattern on multiple chromosomes, and in addition, note that the Southern and coastal samples tend to be positioned on the same side (left) of the plot (e.g. **Fig. 3E, G and I**). This could imply that coastal individuals sometimes carry similar haplotypes (the SV on PCA 1) as the Southern individuals, differentiating them from the marine samples. Increased similarity between the Southern and coastal clusters as compared to Southern and marine clusters is further supported by genome-wide F_{ST} values and PCA conducted on all six chromosomes differentiating ecotypes (**Fig. 3M**). Here, we observe that the coordinates on PCA 1 show that the coastal samples are closer to the South African samples than the marine samples are (the angle of the triangle is rotated compared to **Fig. 1A**). These different elements and the partial sharing of haplotypes between the coastal and Southern lineages leads us to question which anchovy lineage first diverged from the others (the Southern lineage, or alternatively, the marine lineage) and whether some local genealogies along the genome might be discordant with this scenario.

The role of historical contingencies

The past evolutionary history of lineages can have profound impacts for contemporary processes such as speciation (Gould, 2000). Some studies have highlighted the importance of historical contingencies such as cycles of allopatry and sympatry between ancient lineages for the speciation of ecotypes (Fang et al., 2022; Foote et al., 2011). For anchovy ecotypes, divergence patterns were suggested to have resulted from postglacial secondary contact between ancient lineages that diverged in allopatric isolation (Le Moan et al., 2016). Since these authors found evidence for heterogeneous levels of differentiation between genomic markers, they hypothesised that recent differential gene flow could have eroded divergence around islands of selected loci. This secondary contact hypothesis is consistent with our findings of heterogeneous genome-wide divergence and could largely be explained by the presence of SVs on multiple chromosomes. Chromosomal rearrangements such as large inversions suppress recombination

due to problems during meiotic pairing in heterokaryotes, presenting opportunity for divergence to accumulate on the alternate rearrangements. If divergence built up in allopatric lineages carrying different arrangements, haplotypes at SVs would be protected from rehomogenisation through gene flow upon secondary contact, possibly explaining divergence patterns between anchovy ecotypes.

However, the study by Le Moan et al. (2016) did not include samples from the Southern lineage, and their demographic models did not take gene flow with this third component into account. Some studies have reported genetic similarity between European anchovies and anchovies in South Africa, such as mutual sharing of two divergent mitochondrial lineages present in both hemispheres (Grant et al., 2005). Our results also provide evidence for genetic interactions between the European anchovy and anchovies from the South-East Atlantic. This takes the form of shared diversity at two levels: (i) firstly, we identified the presence of possible F1 hybrids and later-generation backcrosses showing genome-wide introgression in the Atlantic-Mediterranean transition zone. (ii) Secondly, there are haplotypes at many SVs that are shared between the European and Southern lineages. These sometimes distinguish marine and coastal ecotypes, with the latter resembling the Southern lineage. We suggest that these two patterns of shared diversity did not result from the same introgression event. Instead, the pattern in (ii) could be the result of introgression that took place in the more distant past. Our results indicate that present-day admixture in the Atlantic-Mediterranean contact zone mainly occurs through hybridisation with the marine lineage, since no unadmixed marine individuals were observed in this region (**Fig. 1C**). Coastal anchovies were present however, and might have less chances of mating with Southern individuals due to their confinement in coastal areas and potential differences related to spawning behaviour. This suggests that Southern haplotypes carried by the coastal ecotype were not introduced due to current admixture, but rather are the products of past introgression episodes. This would explain why these haplotypes are present at high frequencies in Mediterranean lagoons, whereas Southern ancestry is currently highly rare in these areas. If this introduction of SVs took place a long time ago, it further explains why the coastal haplotypes have had time to diverge from modern-day South African haplotypes as we observed (e.g. appearance of a second nested SV within the larger haplotype).

Anchovies in the Eastern Atlantic Ocean have most probably experienced a complex history of fluctuations between allopatric isolation and increased gene flow during secondary contact. Genetic diversity patterns in many species have been shown to be impacted by glacial cycles modulating gene flow between populations in the northern and southern hemisphere (Burridge, 2002; Hilbish et al., 2000). These cycles have led to the remixing, reshuffling and redistribution of genetic variants between the European and Southern lineages of anchovy. We observe a complex pattern of shared variation (mitochondrial haplotypes and SVs) or divergence (private SVs) that seems to differ on almost every other chromosome. These fluctuations have most probably contributed to speciation between ecotypes through the provisioning of divergent haplotypes and assembling of the demographic conditions suitable for divergence. We find evidence that introgressed haplotypes from the Southern lineage contribute to ecotypic differentiation, consistent with what has been shown for divergence between lineages in other marine systems (Duranton et al., 2020; Fraïsse et al., 2014). Since these SVs occupy a large

fraction of the genome, it is possible that they have captured loci which are involved in incompatibilities and that certain haplotypic combinations are less fit. The question stands whether the three lineages have acquired a sufficient degree of reproductive isolation to tend towards the completion of speciation, or whether their differentiation would be eroded by extensive gene flow. There is minor evidence for this reversal in the Atlantic-Mediterranean contact zone, where divergence between ecotypes has partly been erased, but our current understanding of this system does not allow us to answer all questions.

Acknowledgements

This project benefited from the Montpellier Bioinformatics Biodiversity platform (MBB) supported by the LabEx CeMEB, an ANR "Investissements d'avenir" program (ANR-10-LABX-04-01).

References

Alexander, D. H., Novembre, J., & Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome Research*, *19*(9), 1655-164. https://doi.org/10.1101/gr.094052.109

Baird, N. A., Etter, P. D., Atwood, T. S., Currey, M. C., Shiver, A. L., Lewis, Z. A., Selker, E. U., Cresko, W. A., & Johnson, E. A. (2008). Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLOS ONE*, *3*(10), e3376. https://doi.org/10.1371/journal.pone.0003376

Benjamin Penaud, Rémy Dernat, Iago Bonnici, & Khalid Belkhir. (2020). MBB-Framework: A set of tools to build and run reproducible and portable workflows. Available at https://web.mbb.cnrs.fr/subwaw/workflowmanager.php

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

Bierne, N., Welch, J., Loire, E., Bonhomme, F., & David, P. (2011). The coupling hypothesis: Why genome scans may fail to map local adaptation genes. *Molecular Ecology*, *20*(10), 2044-2072. https://doi.org/10.1111/j.1365-294X.2011.05080.x

Bonhomme, F., Meyer, L., Arbiol, C., BĂnaru, D., Bahri-Sfar, L., Fadhlaoui-Zid, K., Strelkov, P., Arculeo, M., Soulier, L., Jean-pierre, Q., & Gagnaire, P. (2021). Systematics of European coastal anchovies (genus *Engraulis* Cuvier). *Journal of Fish Biology*, *100*. https://doi.org/10.1111/jfb.14964

Borsa, P. (2002). Allozyme, mitochondrial-DNA, and morphometric variability indicate cryptic species of anchovy (*Engraulis encrasicolus*). *Biological Journal of the Linnean Society*, *75*(2), 261-269. https://doi.org/10.1046/j.1095-8312.2002.00018.x

Bradbury, I. R., Hubert, S., Higgins, B., Bowman, S., Borza, T., Paterson, I. G., Snelgrove, P. V. R., Morris, C. J., Gregory, R. S., Hardie, D., Hutchings, J. A., Ruzzante, D. E., Taggart, C. T., & Bentzen, P. (2013). Genomic islands of divergence and their consequences for the resolution of spatial structure in an exploited marine fish. *Evolutionary Applications*, *6*(3), 450-461. https://doi.org/10.1111/eva.12026

Burridge, C. P. (2002). Antitropicality of Pacific Fishes: Molecular Insights. *Environmental Biology of Fishes*, *65*(2), 151-164. https://doi.org/10.1023/A:1020040515980

Cabanettes, F., & Klopp, C. (2018). D-GENIES: Dot plot large genomes in an interactive, efficient and simple way. *PeerJ*, *6*, e4958. https://doi.org/10.7717/peerj.4958

Catanese, G., Watteaux, R., Montes, I., Barra, M., Rumolo, P., Borme, D., Buongiorno Nardelli, B., Botte, V., Mazzocchi, M. G., Genovese, S., Di Capua, I., Iriondo, M., Estonba, A., Ruggeri, P., Tirelli, V., Caputo-Barucchi, V., Basilone, G., Bonanno, A., Iudicone, D., & Procaccini, G. (2017). Insights on the drivers of genetic divergence in the European anchovy. *Scientific Reports*, *7*(1). https://doi.org/10.1038/s41598-017-03926-z

Chen, S., Zhou, Y., Chen, Y., & Gu, J. (2018). fastp : An ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*, *34*(17), i884-i890. https://doi.org/10.1093/bioinformatics/bty560

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., & 1000 Genomes Project Analysis Group. (2011). The variant call format and VCFtools. *Bioinformatics*, *27*(15), 2156-2158. https://doi.org/10.1093/bioinformatics/btr330

De Jode, A., Le Moan, A., Johannesson, K., Faria, R., Stankowski, S., Westram, A. M., Butlin, R. K., Rafajlović, M., & Fraïsse, C. (2023). Ten years of demographic modelling of divergence and speciation in the sea. *Evolutionary Applications*, *16*(2), 542-559. https://doi.org/10.1111/eva.13428

Duranton, M., Allal, F., Fraïsse, C., Bierne, N., Bonhomme, F., & Gagnaire, P.-A. (2018). The origin and remolding of genomic islands of differentiation in the European sea bass. *Nature Communications*, *9*(1). https://doi.org/10.1038/s41467-018-04963-6

Duranton, M., Allal, F., Valière, S., Bouchez, O., Bonhomme, F., & Gagnaire, P.-A. (2020). The contribution of ancient admixture to reproductive isolation between European sea bass lineages. *Evolution Letters*, *4*(3), 226-242. https://doi.org/10.1002/evl3.169

Fang, B., Momigliano, P., Kahilainen, K. K., & Merilä, J. (2022). Allopatric origin of sympatric whitefish morphs with insights on the genetic basis of their reproductive isolation. *Evolution*, *76*(8), 1905-1913. https://doi.org/10.1111/evo.14559

Foote, A. D., Martin, M. D., Louis, M., Pacheco, G., Robertson, K. M., Sinding, M.-H. S., Amaral, A. R., Baird, R. W., Baker, C. S., Ballance, L., Barlow, J., Brownlow, A., Collins, T., Constantine, R., Dabin, W., Dalla Rosa, L.,

Davison, N. J., Durban, J. W., Esteban, R., … Morin, P. A. (2019). Killer whale genomes reveal a complex history of recurrent admixture and vicariance. *Molecular Ecology*, *28*(14), 3427-3444. https://doi.org/10.1111/mec.15099

Foote, A. D., Morin, P. A., Durban, J. W., Willerslev, E., Orlando, L., & Gilbert, M. T. P. (2011). Out of the Pacific and back again: insights into the matrilineal history of pacific killer whale ecotypes. *PLOS ONE*, *6*(9), e24980. https://doi.org/10.1371/journal.pone.0024980

Fraïsse, C., Roux, C., Welch, J. J., & Bierne, N. (2014). Gene-flow in a mosaic hybrid zone: Is local introgression adaptive? *Genetics*, *197*(3), 939-951. https://doi.org/10.1534/genetics.114.161380

Gagnaire, P.-A., Broquet, T., Aurelle, D., Viard, F., Souissi, A., Bonhomme, F., Arnaud-Haond, S., & Bierne, N. (2015). Using neutral, selected, and hitchhiker loci to assess connectivity of marine populations in the genomic era. *Evolutionary Applications*, *8*(8), 769-786. https://doi.org/10.1111/eva.12288

Gould, S. J. (1989). *Wonderful life: the Burgess Shale and the nature of history*. First edition. New York, W.W. Norton & Company.

Grant, W. S., & Leslie, R. W. (2001). Inter-ocean dispersal is an important mechanism in the zoogeography of hakes (Pisces: *Merluccius* spp.). *Journal of Biogeography*, *28*(6), 699-721. https://doi.org/10.1046/j.1365-2699.2001.00585.x

Grant, W. S., Leslie, R. W., & Bowen, B. W. (2005). Molecular genetic assessment of bipolarity in the anchovy genus *Engraulis*. *Journal of Fish Biology*, *67*(5), 1242-1265. https://doi.org/10.1111/j.1095-8649.2005.00820.x

Han, F., Jamsandekar, M., Pettersson, M. E., Su, L., Fuentes-Pardo, A. P., Davis, B. W., Bekkevold, D., Berg, F., Casini, M., Dahle, G., Farrell, E. D., Folkvord, A., & Andersson, L. (2020). Ecological adaptation in Atlantic herring is associated with large shifts in allele frequencies at hundreds of loci. *eLife*, *9*, e61076. https://doi.org/10.7554/eLife.61076

Hellberg, M. E. (2009). Gene flow and isolation among populations of marine animals. *Annual Review of Ecology, Evolution, and Systematics*, *40*(1), 291-310. https://doi.org/10.1146/annurev.ecolsys.110308.120223

Hilbish, T. J., Mullinax, A., Dolven, S. I., Meyer, A., Koehn, R. K., & Rawson, P. D. (2000). Origin of the antitropical distribution pattern in marine mussels (*Mytilus* spp.): Routes and timing of transequatorial migration. *Marine Biology*, *136*(1), 69-77. https://doi.org/10.1007/s002270050010

Huret, M., Lebigre, C., Iriondo, M., Montes, I., & Estonba, A. (2020). Genetic population structure of anchovy (*Engraulis encrasicolus*) in North-western Europe and variability in the seasonal distribution of the stocks. *Fisheries Research*, *229*, 105619. https://doi.org/10.1016/j.fishres.2020.105619

Ishigohoka, J., Bascón-Cardozo, K., Bours, A., Fuß, J., Rhie, A., Mountcastle, J., Haase, B., Chow, W., Collins, J., Howe, K., Uliano-Silva, M., Fedrigo, O., Jarvis, E. D., Pérez-Tris, J., Illera, J. C., & Liedvogel, M. (2021). Recombination suppression and selection affect local ancestries in genomes of a migratory songbird. *bioRxiv*. https://doi.org/10.1101/2021.12.22.473882

Johannesson, K. (2016). What can be learnt from a snail? *Evolutionary Applications*, *9*(1), 153-165. https://doi.org/10.1111/eva.12277

Johannesson, K., Butlin, R. K., Panova, M., & Westram, A. M. (2020). Mechanisms of Adaptive Divergence and Speciation in *Littorina saxatilis*: Integrating Knowledge from Ecology and Genetics with New Data Emerging from Genomic Studies. In M. F. Oleksiak & O. P. Rajora (Eds.), *Population Genomics: Marine Organisms* (p. 277-301). Springer International Publishing. https://doi.org/10.1007/13836_2017_6

Johannesson, K., Le Moan, A., Perini, S., & André, C. (2020). A Darwinian laboratory of multiple contact zones. *Trends in Ecology & Evolution*, S016953472030210X. https://doi.org/10.1016/j.tree.2020.07.015

Karahan, A., Borsa, P., Gucu, A. C., Kandemir, I., Ozkan, E., Orek, Y. A., Acan, S. C., Koban, E., & Togan, I. (2014). Geometric morphometrics, Fourier analysis of otolith shape, and nuclear-DNA markers distinguish two anchovy species (*Engraulis* spp.) in the Eastern Mediterranean Sea. *Fisheries Research*, *159*, 45-55. https://doi.org/10.1016/j.fishres.2014.05.009

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

Li, H. (2018). Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics*, *34*(18), 3094-3100. https://doi.org/10.1093/bioinformatics/bty191

Li, H., & Ralph, P. (2019). Local PCA shows how the effect of population structure differs along the genome. *Genetics*, *211*(1), 289-304. https://doi.org/10.1534/genetics.118.301747

Ludt, W. B. (2021). Missing in the middle: a review of equatorially disjunct marine taxa. *Frontiers in Marine Science*, *8*. https://www.frontiersin.org/articles/10.3389/fmars.2021.660984

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

Meyer, L., Barry, P., Riquet, F., Foote, A., Sarkissian, C. D., Cunha, R., Arbiol, C., Cerqueira, F., Desmarais, E., Bordes, A., Bierne, N., Guinand, B., & Gagnaire, P.-A. (2023). Divergence and gene flow history at two large chromosomal inversions involved in long-snouted seahorse ecotype formation (p. 2023.07.04.547634). *bioRxiv*. https://doi.org/10.1101/2023.07.04.547634

Le Moan, A., Gagnaire, P.-A., & Bonhomme, F. (2016). Parallel genetic divergence among coastal–marine ecotype pairs of European anchovy explained by differential introgression after secondary contact. *Molecular Ecology*, *25*(13), 3187-3202. https://doi.org/10.1111/mec.13627

Oueslati, S., Fadhlaoui-Zid, K., Kada, O., Augé, M. T., Quignard, J. P., & Bonhomme, F. (2014). Existence of two widespread semi-isolated genetic entities within Mediterranean anchovies. *Marine Biology*, *161*(5), 1063-1071. https://doi.org/10.1007/s00227-014-2399-5

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

Picard toolkit. (2019). In *Broad Institute, GitHub repository*. Broad Institute. https://broadinstitute.github.io/picard/

Ravinet, M., Faria, R., Butlin, R. K., Galindo, J., Bierne, N., Rafajlović, M., Noor, M. a. F., Mehlig, B., & 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. https://doi.org/10.1111/jeb.13047

Raybaud, V., Bacha, M., Amara, R., & Beaugrand, G. (2017). Forecasting climate-driven changes in the geographical range of the European anchovy (*Engraulis encrasicolus*). *ICES Journal of Marine Science*, *74*(5), 1288-1299. https://doi.org/10.1093/icesjms/fsx003

Selkoe, K., D'Aloia, C., Crandall, E., Iacchei, M., Liggins, L., Puritz, J., von der Heyden, S., & Toonen, R. (2016). A decade of seascape genetics: Contributions to basic and applied marine connectivity. *Marine Ecology Progress Series*, *554*, 1-19. https://doi.org/10.3354/meps11792

Silva, G., Cunha, R. L., Ramos, A., & Castilho, R. (2017). Wandering behaviour prevents inter and intra oceanic speciation in a coastal pelagic fish. *Scientific Reports*, *7*(1). https://doi.org/10.1038/s41598-017-02945-0

Martin, S. H. (2018). *Genomics_general*. https://github.com/simonhmartin/genomics_general

Van der Auwera, G. A., Carneiro, M. O., Hartl, C., Poplin, R., del Angel, G., Levy-Moonshine, A., Jordan, T., Shakir, K., Roazen, D., Thibault, J., Banks, E., Garimella, K. V., Altshuler, D., Gabriel, S., & DePristo, M. A. (2013). From fastq data to high-confidence variant calls: the genome analysis toolkit best practices pipeline. *Current Protocols in Bioinformatics*, *43*(1), 11.10.1-11.10.33. https://doi.org/10.1002/0471250953.bi1110s43

Zarraonaindia, I., Iriondo, M., Albaina, A., Pardo, M. A., Manzano, C., Grant, W. S., Irigoien, X., & Estonba, A. (2012). Multiple SNP markers reveal fine-scale population and deep phylogeographic structure in european anchovy (*Engraulis encrasicolus* L.). *PLOS ONE*, *7*(7), e42201. https://doi.org/10.1371/journal.pone.0042201

Zheng, X., Levine, D., Shen, J., Gogarten, S. M., Laurie, C., & Weir, B. S. (2012). A high-performance computing toolset for relatedness and principal component analysis of SNP data. *Bioinformatics (Oxford, England)*, *28*(24), 3326-3328. https://doi.org/10.1093/bioinformatics/bts606

Chapter II

Divergence and gene flow history at two large chromosomal inversions underlying ecotype differentiation in the long-snouted seahorse

---

Context

Genetic structure in the long-snouted seahorse (*Hippocampus guttulatus*) has been less thoroughly studied than in *E. encrasicolus*. Using a set of SNP markers, Riquet et al. (2019) showed that *H. guttulatus* was subdivided into a Northern and a Southern lineage in the Atlantic, and into marine and lagoon ecotypes in the Mediterranean Sea. This original structure with both geographic and ecological components further showed a form of genetic parallelism, since the same genomic markers were involved in differentiation between lineages in both ocean basins. These loci showed patterns of high linkage disequilibrium and were interpreted to reflect the presence of a large genomic island of differentiation showing reduced recombination.

In the present chapter, we set out to better characterise genome-wide population structure in *H. guttulatus*, with the specific goal of describing the genomic island differentiating ecotypes, which putatively represented a chromosomal inversion. Where in **Chapter I** we highlighted the spatial patterns of ecotypic structure, we here present a more detailed study of the genomic architecture underlying ecotypic differentiation. We produced a high quality reference genome, which facilitated the detection of not one but two large inversions. We characterise the origins of these inversions and consider the mechanisms potentially impacting their evolutionary dynamics, and the processes that could explain their maintenance over the long term. The study presented in this chapter corresponds to a manuscript that we submitted for publication in *Molecular Ecology*, as part of a special issue which focuses on the role of structural variants in evolution.

Full detailed metadata and results may be consulted in the supplied HTML report:

https://cloud.isem-evolution.fr/nextcloud/index.php/s/XKwNymTJLWiLXrz

Authors

Laura Meyer[1], Pierre Barry[1,2], Florentine Riquet[3], Andrew Foote[4], Clio Der Sarkissian[5], Regina L. Cunha[6], Christine Arbiol[1], Frédérique Cerqueira[1], Erick Desmarais[1], Anaïs Bordes[1], Nicolas Bierne[1]*, Bruno Guinand[1]*, Pierre-Alexandre Gagnaire[1]*

[1] ISEM, Univ Montpellier, CNRS, EPHE, IRD, Montpellier, France
[2] CIBIO-InBIO, Centro de Investigação em Biodiversidade e Recursos Genéticos Universidade do Porto, Vairão, Portugal
[3] IFREMER, RBE-ASIM, Station de La Tremblade, La Tremblade, France
[4] Centre for Ecological and Evolutionary Synthesis (CEES), Department of Biosciences, University of Oslo, 0316 Oslo, Norway
[5] Centre for Anthropobiology and Genomics of Toulouse, CNRS, University of Toulouse Paul Sabatier, Toulouse, France
[6] Centre of Marine Sciences-CCMAR, University of Algarve, Campus de Gambelas, 8005-139 Faro, Portugal
* Co-last authors

Abstract

Chromosomal inversions can play an important role in divergence and reproductive isolation by building and maintaining distinct allelic combinations between evolutionary lineages. Alternatively, they can take the form of balanced polymorphisms that segregate within populations until one arrangement becomes fixed. Many questions remain about how these inversion polymorphisms arise, how they are maintained over the long-term, and ultimately, whether and how they contribute to speciation. The long-snouted seahorse (*Hippocampus guttulatus*) is known to be subdivided into partially isolated lineages and marine-lagoon ecotypes with structural variation likely playing a role in their divergence. Here, we aim to characterise these variants and to reconstruct their history and suspected role in ecotype formation. We generated a near chromosome-level genome assembly and described genome-wide patterns of diversity and divergence through the analysis of 112 whole-genome sequences from Atlantic, Mediterranean, and Black Sea populations. Combined with linked-read sequencing data, we found evidence for two chromosomal inversions that were several megabases in length and showed contrasting allele frequency patterns between lineages and ecotypes across the species range. We reveal that these inversions represent ancient intraspecific polymorphisms, one likely being maintained by divergent selection, and the other by pseudo-overdominance. The possibility for selective coupling between the two inversions is supported by the absence of specific haplotype combinations and by the functional enrichment for two molecular pathways, one present on each inversion, that may interact in reproduction pathways. Lastly, we detected gene flux eroding divergence between inverted alleles at varying levels for the two inversions, with a likely impact on their dynamics and contribution to divergence and speciation.

Introduction

How genetic differences accumulate between nascent species and contribute to the build-up of reproductive isolation remains a fundamental question in evolutionary biology (Westram et al., 2022b). Mechanisms that bring unlinked or distant sites into linkage disequilibrium and that maintain strong allelic associations in the presence of gene flow, are key factors favouring the emergence of reproductive isolation (Butlin, 2005; Ortiz-Barrientos et al., 2016; Tigano & Friesen, 2016). Large structural variants (SVs) such as chromosomal inversions combine these two properties, since they both establish and maintain divergent haplotypes by reducing inter-haplotype recombination (Faria & Navarro, 2010; Hoffmann & Rieseberg, 2008; Mérot et al., 2020a; Wellenreuther et al., 2019; Zhang et al., 2021). For these reasons, large SVs have often received special attention in speciation genomics studies, particularly those investigating ecotype formation (e.g. Atlantic cod, Berg et al., 2016; rough periwinkles, Faria et al., 2019b; deer mice, Harringmeyer & Hoekstra, 2022; seaweed flies, Mérot et al., 2020b; sunflowers, Todesco et al., 2020; monkeyflowers, Lowry & Willis, 2010). Despite the growing number of studies reporting the importance of inversions for divergence between ecotypes in a variety of organisms, many questions remain about the conditions in which these inversions arise, how they are maintained over the long term, and how they contribute to reproductive isolation and the progress towards speciation (Faria et al., 2019a; Kirkpatrick, 2010; Wellenreuther & Bernatchez, 2018; Westram et al., 2022a).

The most immediate role of inversions is generally ascribed to recombination suppression, which protects combinations of locally advantageous or co-adapted alleles locked within an inverted region (Butlin, 2005; Kirkpatrick, 2010; Noor et al., 2001; Rieseberg, 2001). Several studies have directly linked inversions to traits that differ between ecotypes, with functional implications in local adaptation or reproductive isolation (Campbell et al., 2021; Funk et al., 2021; Gould et al., 2017; Hager et al., 2022; Jay et al., 2022; Lundberg et al., 2023). These tend to be complex phenotypes that likely involve several genes inherited as a single block (Lamichhaney et al., 2016; Matschiner et al., 2022; Schwander et al., 2014; Thompson & Jiggins, 2014). Large inversions spanning megabase-sized chromosomal segments can facilitate the emergence of such complex phenotypes by building linkage disequilibrium between multiple distant functional sites (Thompson & Jiggins, 2014). However, inversions can also be expected to carry other types of selected sites, including co-adapted gene complexes and deleterious mutations that are captured by chance (Berdan et al., 2021; Jay et al., 2021; Nei et al., 1967). These different types of selected loci, as well as the potential interactions between them, imply that the dynamics of an inversion can be driven by multiple simultaneous processes (Faria et al., 2019a; Guerrero et al., 2012). Furthermore, the allelic content of an inversion is not fixed and can change over time, for example through the accumulation of new mutations or the exchange of genetic material by gene flux between inverted haplotypes (Cheng et al., 2012; Faria et al., 2019; Matschiner et al., 2022; Navarro et al., 1997; Schaeffer & Anderson, 2005). This can occur through double crossover events with the formation of an inversion loop, or through non-crossover gene conversion during the repair of double-stranded breaks (Korunes & Noor, 2019; Matschiner et al., 2022). The resulting changes in allelic content could affect the processes that control the frequency of an

inversion (Berdan et al., 2021), and ultimately, its evolutionary trajectory and contribution to reproductive isolation.

Considering the lifetime evolution of inversions provides a particularly useful framework for studying their evolutionary dynamics and their potential role in speciation (Faria et al., 2019a). This framework aims to disentangle the roles of different processes acting at different stages – from the appearance of inversions to their maintenance and long-term fate – in order to go beyond a simple inventory of the inversions present in a given system. In their review on the evolution of inversions, Faria et al. (2019a) distinguished two main classes of inversion polymorphisms, which they refer to as Type I and Type II. Type I inversions are expected to show frequency differences and divergence between populations, with the possible presence of polymorphism within populations due to gene flow. This pattern may be caused by divergent selection acting on alternate haplotypes conferring local adaptation and/or a form of bistable selection involving frequency dependence, incompatibility selection or assortative mating. A Type I polymorphism may show signs of underdominance either due to direct effects (loss of unbalanced recombinant gametes during meiosis) or due to the allelic contents of the inversion (Faria et al., 2019a; Kirkpatrick & Barton, 2006). Over time, Type I inversions likely contribute to the accumulation of Dobzhansky-Müller (DM) incompatibilities and reinforcement mechanisms via coupling with other reproductive isolation polymorphisms (Kulmuni et al., 2020; Navarro & Barton, 2003). Alternatively, inversions can be maintained by a form of balancing selection (Jay et al., 2021; Mérot et al., 2020a; Wellenreuther & Bernatchez, 2018; Yeaman, 2013; Yeaman & Whitlock, 2011). Faria et al. (2019a) recognise these Type II inversions as polymorphisms that are maintained by mechanisms such as frequency-dependence, disassortative mating, or antagonistic pleiotropy. Type II inversions might also contain overdominant loci and reversed phase deleterious mutations (i.e. pseudo-overdominance), leading to heterokaryote advantage (Marion & Noor, 2023). Indeed, some inversions can be expected to suffer from high mutation load, due to deleterious mutations hitchhiking with selected genes, or simply accumulating in low recombination regions (Hill & Robertson, 1966; Nei et al., 1967). These various processes associated with Type I and Type II polymorphisms are not necessarily mutually exclusive and might act in concert to shape the evolutionary dynamics of a given system. In particular, the role of potential interactions between different types of inversions remains a key question that needs to be addressed with empirical data.

Here, we aim to disentangle the different processes that affect the lifetime evolution of chromosomal inversions involved in ecotype differentiation in the long-snouted seahorse (*Hippocampus guttulatus*). This species is subdivided into semi-isolated lineages separated by sharp eco-geographical boundaries (Riquet et al., 2019). In the North-East Atlantic, seahorse populations are split into a northern and southern genetic lineage, while Mediterranean populations display ecotypic structure associated with marine and lagoon habitats. These environments are characterised by seagrass beds (*Zostera*, *Cymodocea*, *Posidonia*) that differ in their species community assemblages and in their physico-chemical parameters (Perez-Ruzafa et al. 2017). Despite these habitat differences, lagoon and marine ecotypes are morphologically indistinguishable. However, they show strong genetic differences concentrated in a large genomic island of divergence which also differentiates the two Atlantic lineages (Riquet et al., 2019). This

previous study, although not based on a reference genome for the species, suggested that a large SV such as a chromosomal inversion was involved in ecotype differentiation.

We introduce the first genome-scale study in *H. guttulatus*, aiming to characterise and understand the origin and evolution of genome-wide divergence patterns among lineages and ecotypes. We generate a near chromosome-level assembly of the *H. guttulatus* genome, and describe genetic variation across habitats throughout the range distribution of the species. We found the presence of not one, but two megabase-scale inversions, which may have played an important role in ecotype formation in the Mediterranean Sea. We then attempt to reconstruct the lifetime evolution of these two inversions, especially focusing on their origin, divergence and subsequent evolution, including gene exchange between the inverted alleles. Our study highlights interactions between two inversions in the same system and the impact of gene flux between arrangements.

Materials and methods

Sampling and DNA extraction

A total of 112 samples from different habitats across the species range were used for whole-genome resequencing (**Supplementary Table S1**). Different DNA extraction protocols were applied to three types of samples that either originated from published studies, our own laboratory collection, or museum collections. (*i*) Samples from published studies (Riquet et al., 2019; Barry et al., 2022) consisted of non-lethal fin or tail clips preserved in 95% ethanol at -20 °C. Their genomic DNA was extracted using the Nucleospin Tissue kit ® (Macherey–Nagel, Germany) following the manufacturer's protocol. (*ii*) Dried seahorse samples were obtained from private collections by means of a call to the public in a local newspaper. The approximate date (1935-2010) and collection site were recorded based on information provided by the donors. The dorsal fin of each dried seahorse was scratched to collect c.a. 20 µg of tissue powder. Genomic DNA was extracted using a standard CetylTrimethyl Ammonium Bromide (CTAB) Chloroform:Isoamyl alcohol (24:1) protocol (Doyle & Doyle, 1987). (*iii*) Lastly, four alcohol-preserved museum samples (1856-1898) were provided by the National Museum of Natural History (Paris, France). Tissue fragments (1-2 mm) were obtained by internal needlepoint sampling (Haÿ et al., 2020) and subjected to overnight enzymatic digestion (40 µL proteinase K at 25 mg/mL for 500 µL volume, at 56 °C). DNA was extracted using a phenol-chloroform method (Campos & Gilbert, 2012) in a dedicated clean lab facility located at the Institute of Evolutionary Science of Montpellier (ISEM, France).

Assembly and annotation of the *H. guttulatus* reference genome

We performed high-coverage linked-read sequencing of an Atlantic *H. guttulatus* individual from Hossegor lagoon (Bay of Biscay, Hgutt_GA_13) to generate a high-quality reference genome assembly (hereafter referred to as Hgutt_V1). Fresh gill and muscle tissue were solubilized in a 25 ml solution of TNES-Urea (10 mM Tris-HCl, 120 mM NaCl, 10 mM EDTA, 0.5% SDS, 4 M urea, PH 8) during 4 weeks at 20°C. High molecular weight genomic DNA (HMW gDNA) was

isolated using three phenol-chloroform followed by two chloroform extractions after digestion with proteinase K (150 µg/ml). Final precipitation was performed using two volumes of 100% Ethanol. The resulting pellet was washed several times in 80% ethanol and resuspended in ultrapure water by heating to 65°C, before being kept at 40°C during 4 days. The length distribution of extracted DNA molecules was assessed by electrophoresis on a TapeStation Genomic DNA ScreenTape assay (Agilent Technologies). Single-stranded DNA damage was treated with the NEBNext FFPE DNA Repair mix and repaired DNA was then subjected to size selection to remove fragments shorter than 40 kb using a PippinHT instrument (Sage Science) with a 0.75% Agarose Gel Cassette. HMW gDNA was submitted to the 10x Genomics linked-read library preparation following the Chromium Genome Reagent Kit v2 protocol at the MGX sequencing facility (CNRS, Montpellier, France). The genome library was sequenced to ~100X on a S1 lane of an Illumina NovaSeq6000 in 150 bp paired-end mode by Genewiz Inc (USA), generating ~0.3 billion reads.

Raw demultiplexed reads were deduplicated using *nubeam-dedup* (Dai and Guan 2020) and processed with *process_10xReads.py* (https://github.com/ucdavis-bioinformatics/proc10xG) to extract the barcode sequence of each read pair and the number of read pairs associated to each barcode. The distribution of the number of read pairs per barcode was then analysed to identify rare barcodes potentially generated by sequencing errors and over-represented barcodes (**Supplementary Fig. S1**). A total of 143.7 million paired-end reads carrying 1.455 million retained barcodes were finally extracted using *proc10xG* scripts and used for linked-read-based *de novo* genome assembly using the *Supernova-2.1.1* software package (Weisenfeld et al., 2017). Assembled scaffolds were outputted in pseudohap style, with a minimum size set to 1 kb. Assembly statistics of the Hgutt_V1 reference genome were computed and visualised with the *BlobToolKit v3.5.2* software suite (Challis et al., 2020), combined with an assessment of genome assembly completeness with *BUSCO 5.4.4* (Manni et al., 2021) using the actinopterygii_odb10 fish dataset containing 3640 conserved genes. Whole-genome alignment was performed with *Minimap2* (Li, 2018) and visualised using *D-GENIES* (Cabanettes & Klopp, 2018) to compare and anchor *H. guttulatus* scaffolds to the chromosome-scale assembly of the closely related *H. erectus* (Li et al., 2021), provided by the authors. We also used the genome sequence of a north Atlantic long-snouted seahorse (UK), which was assembled by Iridian Genomes by ordering and orienting pre-assembled contigs based on other fish reference genomes (Accession PRJNA481552, hereafter called HguttRefA).

We used *RepeatModeler2* (Flynn et al., 2020) for *de novo* repeat finding and identification of the unique transposable element families present in the seahorse genome. In addition, we searched for tandem repeats (TRs) following the strategy developed in Melters et al. (2013), using the same parameter values to run *Tandem Repeats Finder* v4.09.1 (Benson, 1999) within the *pyTanFinder* pipeline (Kirov et al., 2018). We then used *RepeatMasker* v4.0.5 (http://repeatmasker.org) to perform repeat annotation and masking of the identified repeat elements. Structural gene annotation was performed using the RNA-Seq pipeline in *Braker2* v2.1.6 (Brůna et al., 2021), which manages the training of the gene prediction tools *GeneMark-ET* (Lomsadze et al., 2014) and *AUGUSTUS* (Stanke et al., 2008). In brief, we used seven RNA-Seq libraries from NCBI's SRA (Accessions SRX565152-57, SRX20881937), totalling 27.8 Gb of Illumina short reads, and mapped them to the soft-masked reference genome with *HISAT2* v2.0.4 (Kim et al., 2019). RNA-

Seq alignment information was used to iteratively train *GeneMark-ET* to generate initial gene structures, which were passed to *AUGUSTUS* along with RNA-Seq mapping information to generate final gene predictions. Functional annotation of the predicted gene coding sequences was finally conducted using *eggNOG-mapper* v2 (Cantalapiedra et al., 2021), which relies on precomputed Orthologous Groups (OGs) to transfer functional information using phylogenetically refined orthology assignments.

Library preparation and whole-genome resequencing

Whole-genome sequencing (WGS) libraries were prepared for 86 samples using the Ovation Ultralow System V2 library preparation kit (NuGEN/Tecan) from 100 ng DNA input (when possible) following the manufacturer's instructions. We used unique dual indexing to minimise the effect of index-hopping and PCR cycles were adapted to the amount of input DNA (9-20 cycles). Libraries were pooled in equimolar ratio and sequenced to different coverage depths on a single S4 flow cell on a NovaSeq6000 instrument (Illumina) to generate 150 bp paired-end reads.

Sequence processing and alignment

To complement our dataset, WGS data for an additional 26 seahorse samples were obtained from Barry et al. (2022) (TruSeq DNA PCR-free libraries sequenced on Illumina NovaSeq6000, GenBank Sequence Read Archive, accession BioProject ID PRJNA777424). Therefore, our final combined dataset consisted of 112 samples (hereafter referred to as the “full dataset”, **Supplementary Table S1**). Raw demultiplexed reads were processed using fastp (v0.23.1) (Chen et al., 2018) with the “*--merge*” option, in order to stitch together paired-end reads with overlapping sections. Both merged and unmerged reads were aligned to our reference genome using BWA-MEM (BWA v0.7.17; (H. Li, 2013). Picard (v2.26.8) (« Picard toolkit », 2019) was used for sorting read alignments, marking duplicates and adding read groups. DNA damage patterns in older dried and museum samples were visualised using PMDtools (v0.60) (Skoglund et al., 2014) (**Supplementary Figure S2**).

Variant calling of medium- to high-coverage samples

Forty-eight samples with sufficient coverage (~5-50X) were selected for variant calling in such a way that this subset contained five individuals in certain key locations (from now on referred to as the “GATK dataset”, **Supplementary Table S1**, variant calling = Yes). Variants were called using the GATK best practices workflow (McKenna et al., 2010; Van der Auwera et al., 2013) without performing Variant and Base Quality Score Recalibration (VQSR and BQSR). Firstly, individual GVCF files were created from bam files with HaplotypeCaller (GATK v.4.1.8.0). This information was then stored in a GVCF database using GenomicsDBImport, and VCF files (one file per scaffold) were generated with GenotypeGVCFs. After concatenation, the resulting VCF was filtered for indels, multiallelic SNPs and missing data (“*--max-missing 0.9*”). Our detailed workflow and commands used are provided in **Supplementary File S2**. An Rmarkdown report produced with flex dashboard (Sievert et al., 2022), showing detailed results across various bioinformatic

steps and analyses is provided in the **Context** section at the beginning of this chapter Chapter_II_Suppl_Report.html).

Population structure

A Principal component analysis (PCA) was carried out on samples with sufficient coverage (>3X) from the full dataset without calling genotypes. To accomplish this, genotype likelihoods were calculated using ANGSD (v0.933) (Korneliussen et al., 2014) with the GATK model ("*-GL 2*") and the following parameters: "*-doMajorMinor 1 -minMapQ 30 -minQ 20 -doMaf 1 -doCounts 1 -minMaf 0.05 -uniqueonly 1 -remove_bads 1 -C 50 -baq 1 -doCov 1 -doIBS 2 -makeMatrix 1 -ref referencegenome_Hgutt_V1.fa*". In addition, the first 5 bp were trimmed off of reads (*-trim 5*) to account for DNA damage patterns due to cytosine deamination in historical samples. The analysis was restricted to sites with a strong probability of being SNPs ("*-SNP_pval 1e-6*"). Finally, sites were not considered if the total depth across samples was greater than twice the sum of mean coverage per sample ("*-setMaxDepth*"), in order to avoid artefacts in problematic regions that received unexpectedly high coverage (low-complexity and duplicated regions), or if less than half of the individuals had data ("*-minInd*"). The same filters were used in subsequent analyses conducted with ANGSD, unless specified otherwise. To perform a PCA while minimising the effect of linkage disequilibrium, we used ANGSD with the "*-sites*" option and provided a file containing one randomly selected SNP per 10 kb window in every 500 kb interval. To complement this genotype likelihood-based approach, a PCA was also performed on called genotypes from the GATK dataset using the R package SNPRelate (v1.28.0) (Zheng et al., 2012). Lastly, we conducted local PCA on the GATK dataset to capture variation in population structure along the genome in order to identify outlying patterns indicating the presence of putative chromosomal inversions. To this end, local PCA was performed in non-overlapping windows of 5 kb using the R package lostruct (v0.0.0.9, Li & Ralph, 2019).

Analysis of large structural variants

Genomic regions that were identified in local PCA as being candidates for the presence of large SVs (multiple groups that persist over many windows) were more specifically tested for evidence of chromosomal inversions in linked-read sequencing data. Firstly, we used abrupt signal shifts flanking outlying regions in the local PCA to determine the location of putative inversion breakpoints. We performed manual curation of the reference genome assembly in those regions following (Rhie et al., 2021), using the alignment to the genome assembly of *H. erectus* to confirm the suspected breakpoints. We then used *MTG-Link* v2.4.1 (Guichard et al., 2023) to perform local re-assembly of the linked-reads mapping in the regions surrounding breakpoints. The 10X linked-reads data were preprocessed with *EMA* v0.6.2 (Shajii et al., 2018) and *LRez* v2.2.4 (Morisse et al., 2021b) before running *MTG-Link* with default settings. The locally reassembled contigs were finally aligned against the Hgutt_V1 reference genome with *Minimap2* (H. Li, 2018) to check for local consistency between assemblies and identify possible mis-joins. We characterised repeat content around the inversion breakpoints, looking specifically for the presence of recombinogenic sequences such as long inverted repeats (LIRs). Finally, we used *Leviathan* v1.0.2 (Morisse et al., 2021a) for calling SVs using linked-read data information

connecting distant regions that share a higher number of barcodes than expected based on their distance. We ran *Leviathan* on the 10X linked-read data from the Hgutt_V1 assembly mapped to the HguttRefA genome.

In addition, we followed the haplotagging library construction protocol described in Meier et al. (2021) to generate complementary linked-read sequencing data for alternate genotypes at the two suspected inversions. Namely, we constructed haplotagging libraries for a DD/AA Mediterranean and a CC/AB Atlantic individual, and sequenced them with 2*50 bp paired-end reads to an average coverage depth of 12X.

Finally, we performed a functional enrichment analysis of the gene sets contained in each inversion using ShinyGO v0.77 (Ge et al., 2020) with a false discovery rate (FDR) threshold of 0.05.

Genomic landscape of divergence and introgression

We characterised the genomic landscape of divergence between localities and habitats using the GATK dataset. Genetic differentiation (F_{ST}), nucleotide diversity (π) and absolute genetic divergence (d_{XY}) were calculated in 25 kb windows using the popgenWindows.py script (Martin, 2018; https://github.com/simonhmartin/genomics_general). We then sought to characterise two highly divergent regions on Chr2 and Chr12, which represented suspected inversions. We used BAMscorer v1.4 (Ferrari et al., 2022) to assign inversion genotypes in all samples (including low-coverage samples). This firstly consisted of classifying all samples in the GATK dataset with regards to their haplotypes, as based on PCA groupings and individual heterozygosity values calculated with VCFtools (v0.1.16) (Danecek et al., 2011). This reference database was then used to score alignments and to ascertain allelic state and haplotype for all samples.

In order to test for introgression with other seahorse species and to determine evolutionary relationships across the genome, we obtained whole-genome resequencing data for *H. erectus*, *H. hippocampus*, *H. zosterae*, *H. capensis* and *H. comes* from Li et al. (2021) (NCBI BioProjects accession code PRJNA612146), including data for one individual from each species. Reads were processed and aligned to the Hgutt_V1 reference genome as described above. ANGSD was used to call genotypes and to produce a VCF ("*-doBcf*") containing seven individuals (two *H. guttulatus* that were homozygous for the four alternate inversion haplotypes, i.e. AA-DD and BB-CC, and five individuals from other species). This VCF was filtered to include only fixed sites (no heterozygous genotypes) and was used as input for performing topology weighting using the Twisst pipeline (Martin & Van Belleghem, 2017). This method performs iterative sampling of subtrees and quantifies the proportion of trees that match a certain taxon topology, allowing to study complex evolutionary relationships along the genome (e.g. patterns of introgression). We generated unrooted phylogenies using PhyML (Guindon et al., 2010) as implemented in the script *phyml_sliding_windows.py* (https://github.com/simonhmartin/twisst) for windows of 50 SNPs, setting a minimum of 25 non-missing genotypes per individual per window ("*--minPerInd 25"*) and performed topology weighting using the *twisst.py* script. Additionally, we produced a genome-

wide consensus tree, as well as consensus trees for the two inversions using the averageTree function from phytools (Revell, 2012).

Inferring inversion history from the Ancestral Recombination Graph

We used tsinfer (v0.3.0) (Kelleher et al., 2019) to describe the diversity of genetic relationships between sequenced individuals along the genome. This method takes ancestral recombination events into account to infer the genome-wide sequence of correlated local genealogies, called the Ancestral Recombination Graph (ARG). Given accurate inference, the ARG provides a complete description of the available information on the evolutionary history of a set of related sequences, represented in terms of recombination and coalescence events. Inferring the ARG with tsinfer requires an input VCF file for a set of phased, diploid genomes, and the ancestral state of each mutation present in the INFO field. We carried out statistical phasing and missing data imputation of per-chromosome VCF files using SHAPEIT (v4.2.2) (Delaneau et al., 2019), assuming constant recombination rate of 1 cM per Mb and an effective population size (N_e) estimated based on genome-wide diversity values (Charlesworth, 2009). To infer ancestral states, we used BLAST searches to determine the allelic state of each SNP position in each of three outgroup species (*H. erectus, H. kuda* [Accession GCA_901007745.1]*, H. comes* [Accession GCA_001891065.2]) using the BLAST+ package (v2.2.28) (Camacho et al., 2009). Flanking regions of 100 bp on either side of each SNP were blasted to the reference genome of each outgroup using *blastn*. The top hit was inspected to determine the state of the homologous position in each outgroup using a program which was developed for this purpose (https://gitlab.mbb.cnrs.fr/ibonnici/snom). Thereafter, est-sfs (v2.04) (Keightley & Jackson, 2018) was run using the Kimura 2-parameter model to infer ancestral state probabilities and phased VCFs were annotated with the most likely ancestral variant using BCFtools (v1.9) (Danecek et al., 2021).

We reconstructed the ARG by running tsinfer on our phased, oriented VCF containing all chromosomes. Sample objects were created using the CYVCF2 library (v0.30.18; Pedersen & Quinlan, 2017) as in the tskit tutorial (https://tskit.dev/tsinfer/docs/stable/tutorial.html). The tree sequence was then inferred using a constant recombination rate of 1e-8 per bp and a mismatch ratio of 1. The ages of the ancestral nodes in the inferred trees were estimated using tsdate (v0.1.5) (Wohns et al., 2022) using parameters "N_e*=200000*", "*timepoints=100*" and "*mutation_rate=1e-8*". The N_e value used was calculated from the nucleotide diversity of the most divergent region of the genome (i.e. Chr12 inversion). This choice did not affect the distribution of inferred coalescent times in the genome background, while enabling tsdate to infer ancient node ages within inversions.

The inferred ARG was used for two purposes. First, we described variation in coalescence times across the genome using time to the most recent common ancestor (TMRCA) for randomly selected combinations of sample haplotypes (i.e. tree leaves) extracted from genealogies along the tree sequence. Secondly, we characterised local ancestry for each individual haplotype within inversion regions. The signal of an inversion was shown to appear in the ARG as clades of samples that persist over a more extended region of the genome than would otherwise be

expected (Ignatieva et al., 2023). Here, we assumed that, within an inversion, the two branches immediately below the oldest node of each local tree represent the two clades of non-inverted and inverted extended haplotypes. To assign local ancestry to each branch, a clade was considered to belong to non-inverted or inverted ancestry if it satisfied one of two empirically established conditions: (*i*) the clade contained more than 75% of the haplotype copies from all of the homokaryotes of a given haplotype, and reciprocally for the alternate haplotype in the other clade; (*ii*) the clade contained 100% of the copies from all of the homokaryotes of that haplotype, and the same clade contained less than 75% of the copies from the alternate haplotype. Local ancestry switches along individual haplotypes were used to identify inter-haplotype recombination events within inversions, indicative of gene conversion or introgression.

Results

Reference genome

We obtained a near chromosome-level genome assembly of the *Hippocampus guttulatus* Hgutt_V1 reference genome. The assembly contained 3,878 scaffolds spanning 451 Mb (424 Mb in scaffolds of at least 10 kb, longest scaffold 28.8 Mb, scaffold N50 = 18.1 Mb, N90 = 5.65 Mb, **Supplementary Fig. S3**), which is close to the genome size of 424 Mb predicted by GenomeScope (Barry et al., 2022). Genome completeness was very high (Busco C: 95.6% [S: 94.4%, D: 1.2%], F: 1.6%, M: 2.8%, n: 3640) and comparable to the HguttRefA and the *H. erectus* assemblies (**Supplementary Fig. S4a and S4b**). All scaffolds showed strong homology and conserved synteny relationships with the chromosome-level genome assembly of *H. erectus* (**Supplementary Fig. S5**), suggesting that Hgutt_V1 is a high-quality assembly. Scaffolds anchoring to the *H. erectus* chromosome assembly produced 22 pseudomolecules that were used to characterise the genomic landscape of divergence (**Fig. 2**).

The repeat landscape showed that interspersed repeats add up to 33% of the genome, with 13% being occupied by unknown repeats (**Supplementary Fig. S6**). Among these unknown repeats, the clustering analysis of similar tandem repeats revealed a highly abundant TR of approximately 500 bp in length, present in more than 5,500 copies across the genome, with an accumulated abundance of 2.7 Mb. The consensus monomer of this TR, hereafter called Hgutt_Tan9, is a short sequence of 37 bp present in all chromosomes (**Supplementary Fig. S7**). Structural gene annotation based on RNA-Seq data predicted a total of 25,770 coding gene sequences, of which 13,346 unique gene names were identified by functional annotation.

Overall genetic structure

We produced genome resequencing data of heterogeneous quality (ranging from 1X to ~50X, **Supplementary Table S1**) and therefore characterised overall genetic structure using a genotype likelihood-based approach after controlling for low-coverage individuals (<3X) (**Fig. 1B and 1C**). We found evidence for pronounced population structure, even between certain sampling locations that were in close geographic proximity (inset map, **Fig. 1A**). This genetic structure could largely

be ascribed to markers that are in tight linkage disequilibrium (LD), since PCA performed on unlinked markers revealed a different pattern (**Fig. 1C**). Here, all Mediterranean marine samples grouped together with Black Sea samples, separately from the different Mediterranean lagoon populations (Bz, Li & Mu). In the Atlantic, we also discern a northern (Br & Ga) and a southern (Fa) cluster of samples. By contrast, when all markers are considered (**Fig. 1B**), samples do not group together based on geographical origin. Instead, we observe that the individuals from a given location are grouped into either 1, 2 or 3 clusters that are organised along two orthogonal axes. The clusters from different locations were sometimes shifted away from each other (e.g. the three groups of Bs and Ma mirror those of Br and Ga), although they likely captured the same underlying variation. Observing these replicated three-cluster patterns in the PCA is consistent with the presence of large chromosomal inversions, with the three groups representing different inversion genotypes - homokaryotes with two inverted alleles, heterokaryotes carrying both haplotypes, and non-inverted homokaryotes. The segregation of samples along two orthogonal axes, each showing replicated three-cluster patterns, is thus indicative of two polymorphic inversions containing large numbers of sites in strong LD.

In line with this finding, we observed heterogeneous landscapes of differentiation along the genome (**Fig. 2**). Background genomic differentiation between northern and southern Atlantic lineages ranged between 0.06 and 0.15 (first and third quartiles of F_{ST} distribution) (**Fig. 2A**). Differentiation was weaker between Mediterranean marine and lagoon locations, with F_{ST} values ranging between 0 and 0.03 (**Fig. 2B**). For other similar contrasts (i.e. northern vs. southern Atlantic lineage, or Mediterranean marine vs. lagoon comparison), the values were not substantially affected by the choice of the specific locations being compared (**Supplementary Fig. S8, S9 & S10**). These patterns of relatively weak genome-wide differentiation contrasted with high F_{ST} and d_{XY} values found on two chromosomes - Chr2 (**Fig. 2C**) and Chr12 (**Fig. 2D**). Chr12 presented an 8.2 Mb long plateau of high F_{ST} and d_{XY} values in both Atlantic and Mediterranean comparisons, a pattern that has previously been associated with inversions segregating at different frequencies between populations. For Chr2, a differentiation plateau was only observed in the Mediterranean marine vs. lagoon comparison. This pattern was also less clear, as the highest values did not occur in one contiguous block, but were split across scaffolds and even showed discontinuities within scaffolds.

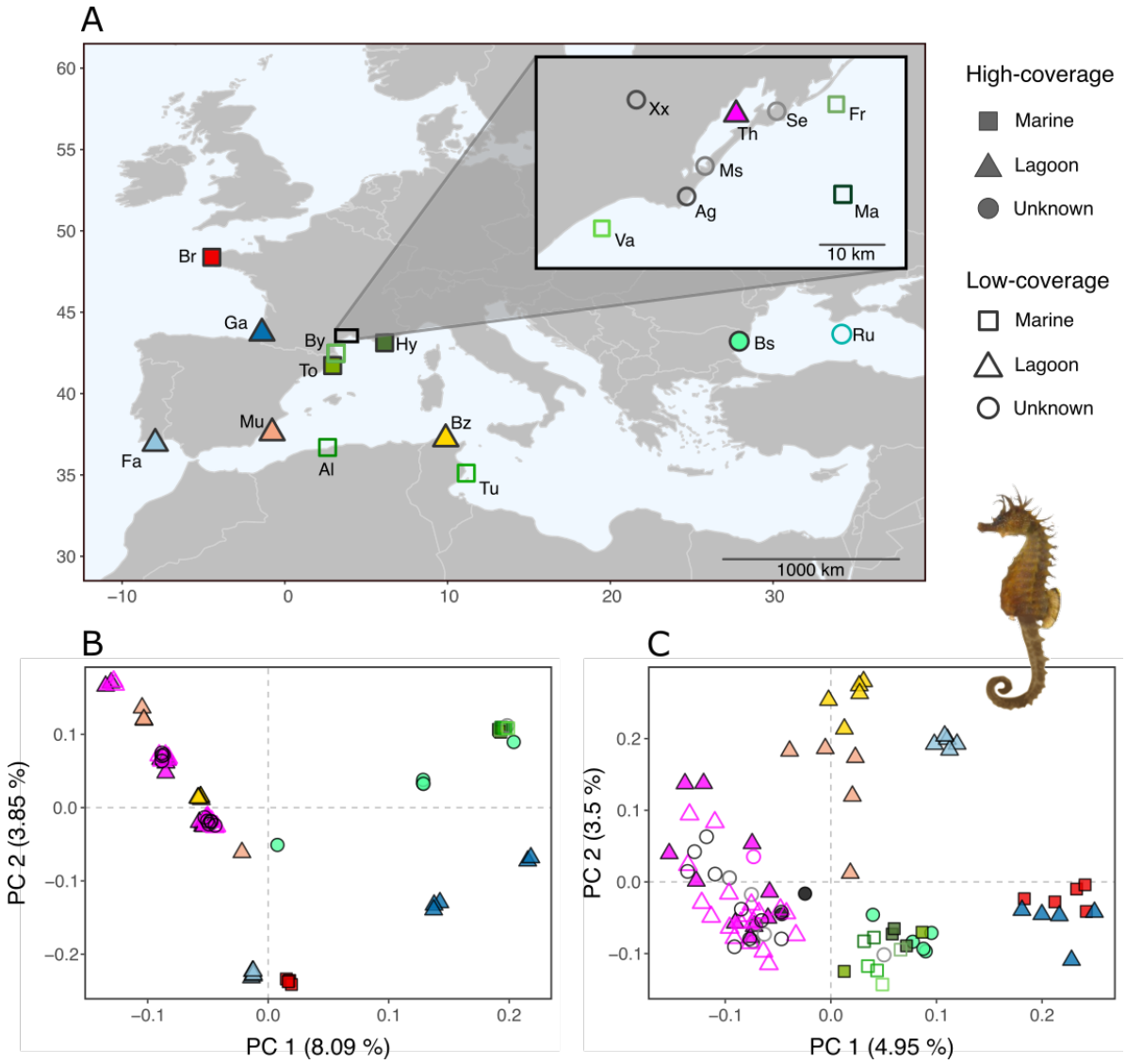


Fig. 1. A) Sampling map for the full dataset of *H. guttulatus* samples (n=112). Triangles: lagoon sites. Squares: marine sites. Circles: unknown habitat type. Filled shapes: medium to high-coverage samples for which individual genotype calling was performed (the GATK dataset). Unfilled shapes: low-coverage samples for which a genotype likelihood-based approach was used. Br: Brest. Ga: Hossegor. Fa: Faro. Mu: Murcia. To: Tossa de mar. By: Banyuls. Va: Valras. Ag: Agde. Ms: Marseillan. Th: Thau lagoon. Se: Sète. Fr: Frontignan. Hy: Hyères. Ma: unknown Mediterranean marine site. Al: unknown Algerian site. Bz: Bizerte lagoon. Tu: unknown Tunisian site. Bs: Varna. Ru: unknown Russian site. Xx: unknown northwestern Mediterranean location. B & C) PCA of 89 individuals with >3X coverage based on IBS distances calculated in ANGSD. PCA was either performed using all genome-wide markers (1,478,955 SNPs) (B), or using a subset of markers at linkage equilibrium (898 SNPs) (C). © Seahorse picture Iglésias 2013.

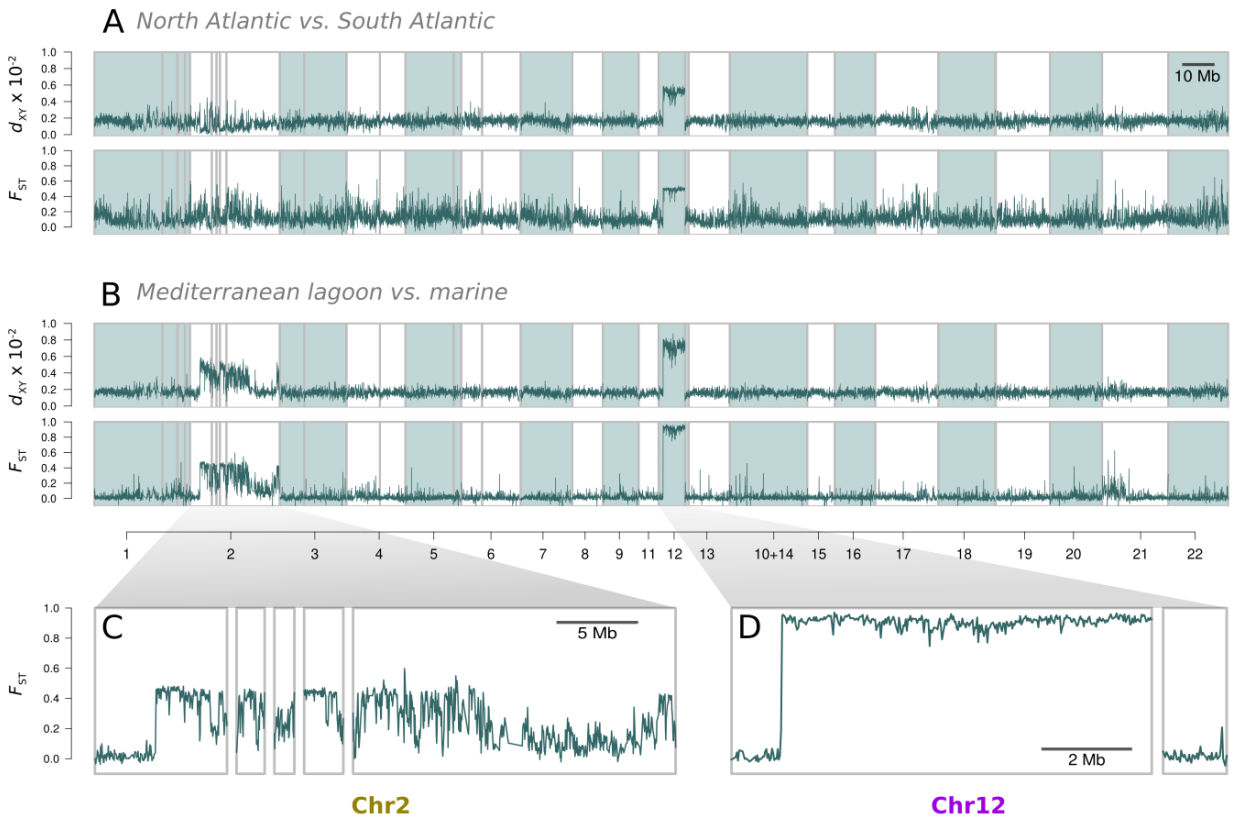


Fig. 2. Genomic landscape of absolute divergence (d_{XY} x 10^{-2}) and differentiation (F_{ST}) calculated in 25 kb non-overlapping sliding windows between a north Atlantic (Ga) and a south Atlantic (Fa) population (A) and between a Mediterranean lagoon (Mu) and a Mediterranean marine (To and Hy) population (B), using 5 high-coverage samples per population. *H. guttulatus* scaffolds (grey rectangles) were aligned to the chromosome-level assembly of *H. erectus* and are displayed according to their homology with the 22 *H. erectus* chromosomes (marked by alternating white and blue rectangles). Regions of high divergence clustered on two chromosomes: Chr2 (~32 Mb) (C) and Chr12 (~11 Mb) (D), which each carry a large chromosomal inversion. Exact ordering and orientation of *H. guttulatus* scaffolds within chromosomes 2 and 12 of *H. erectus* could not be determined due to multiple rearrangements between species (**Supplementary Figure S5**).

Inversions differentiate ecotypes in the Mediterranean Sea

The presence of an inversion for the block of high differentiation on Chr12 was first supported by the alignment of Hgutt_V1 against HguttRefA (generated from the opposite homokaryote, (**Supplementary Fig. S11**), as well as with the *H. erectus* assembly. This analysis showed the presence of an 8.2 Mb long inverted segment between the two *H. guttulatus* assemblies, with a putative inversion breakpoint located near position 1.699 Mb on scaffold 14 of Hgutt_V1, which corresponds to the abrupt signal shift in divergence at the beginning of the block. Linked-read based local reassembly of a 3 kb region centred on the putative breakpoint confirmed the contiguity of the Hgutt_V1 reference. In the middle of this 3 kb region, we found a long inverted repeat (LIR) consisting of two inverted arrays of the Tan9 monomer, with the internal spacer overlapping the inversion breakpoint (**Supplementary Fig. S12**). The other end of the inversion

at the end of scaffold 14 (near 9.877 Mb) also contained a tandem repeat of Tan9 monomers (**Supplementary Fig. S6**). Linked-read sequencing data obtained from three individuals showing the three possible genotypes (AA, AB and BB) showed mapping patterns that were consistent with the presence of a large chromosomal inversion (**Supplementary Fig. S12**). Finally, the analysis of linked-reads from a BB genotype mapped to the HguttRefA assembly allowed the direct detection of a 8.2 Mb inversion at the breakpoints expected from all previous analyses.

Although we were not able to perform the same detailed analysis for Chr2, since the two *H. guttulatus* assemblies both carry the same haplotype, other patterns suggested that a second inversion is highly likely to be present on this chromosome. Performing PCA separately on Chr2 and Chr12, we observed three clusters along PC1 with the middle group presenting higher heterozygosity than the outer groups (**Fig. 3A & 3B**). Using BAMscorer (Ferrari et al., 2022), we characterised inversion genotypes in all samples except for one extremely low-quality individual (Hgutt_Se_1898_94, see **Supplementary Appendix**). These scores were consistent with the groupings observed from the PCA (**Fig. 3A & 3B**) - that is, samples that were classified as heterokaryotes were in the middle PCA group, and homokaryote samples were in the outer groups. These patterns confirm that the multiple blocks of high F_{ST} on Chr2 are in perfect LD, and that they collectively segregate as a single contiguous variant, despite the discontinuous signature observed in the divergence landscape (**Fig. 2C**). Consequently, our results indicate that there are two polymorphic inversions segregating in seahorse populations, and that these play an important role in the differentiation between lineages and ecotypes. As for their gene content, the 8.2 Mb inverted segment on Chr12 contained a total of 409 annotated genes showing a significant 3.7-fold enrichment for the estrogen signalling pathway (Enrichment FDR=0.042, 11 of 134 annotated genes in the pathway). The high-LD region on Chr2 had a total of 419 annotated genes and showed a significant 2.5-fold enrichment for the neuroactive ligand-receptor interaction pathway (Enrichment FDR=0.047, 20 of 372 annotated genes in the pathway).

For each inversion, we characterised the relative frequencies of the three alternative genotypes across sampling locations and habitats (**Fig. 3C & 3F**). We found that the inversion on Chr12 (B12) was present in one of the two homozygous states (either AA or BB) in almost all locations (**Fig. 3F**). Atlantic samples from the northern part of Biscay (Br), the Gulf of Cadiz (Fa) and samples from Mediterranean lagoons (Th, Mu & Bz) exclusively presented homokaryotes for the A allele, while Mediterranean marine sites only presented homokaryotes for the B allele. Fine-scale variation in inversion genotypes was especially pronounced in the Mediterranean Sea, since samples that were collected only a few kilometres apart showed differential fixation for B12 between marine and lagoon habitats (inset map, **Fig. 3F**). Samples for which precise location information was not available (grey circles) were either AA or BB homokaryotes, with A alleles probably hailing from lagoon habitats in these locations and B alleles from the sea. Samples of unknown origin (“Xx”) mostly carried the lagoon haplotype, probably reflecting frequent incidental catches in lagoon habitats during artisanal and non-professional fishing activities. There were only two sampling locations which were polymorphic for B12: the Hossegor marine lake in the Bay of Biscay (Ga) and Varna in the Black Sea (Bs), which presented samples from each of the three inversion genotypes (AA, AB and BB). The inversion located on Chr2 (B2) showed different distribution patterns compared to that of B12, since this polymorphism was only found in

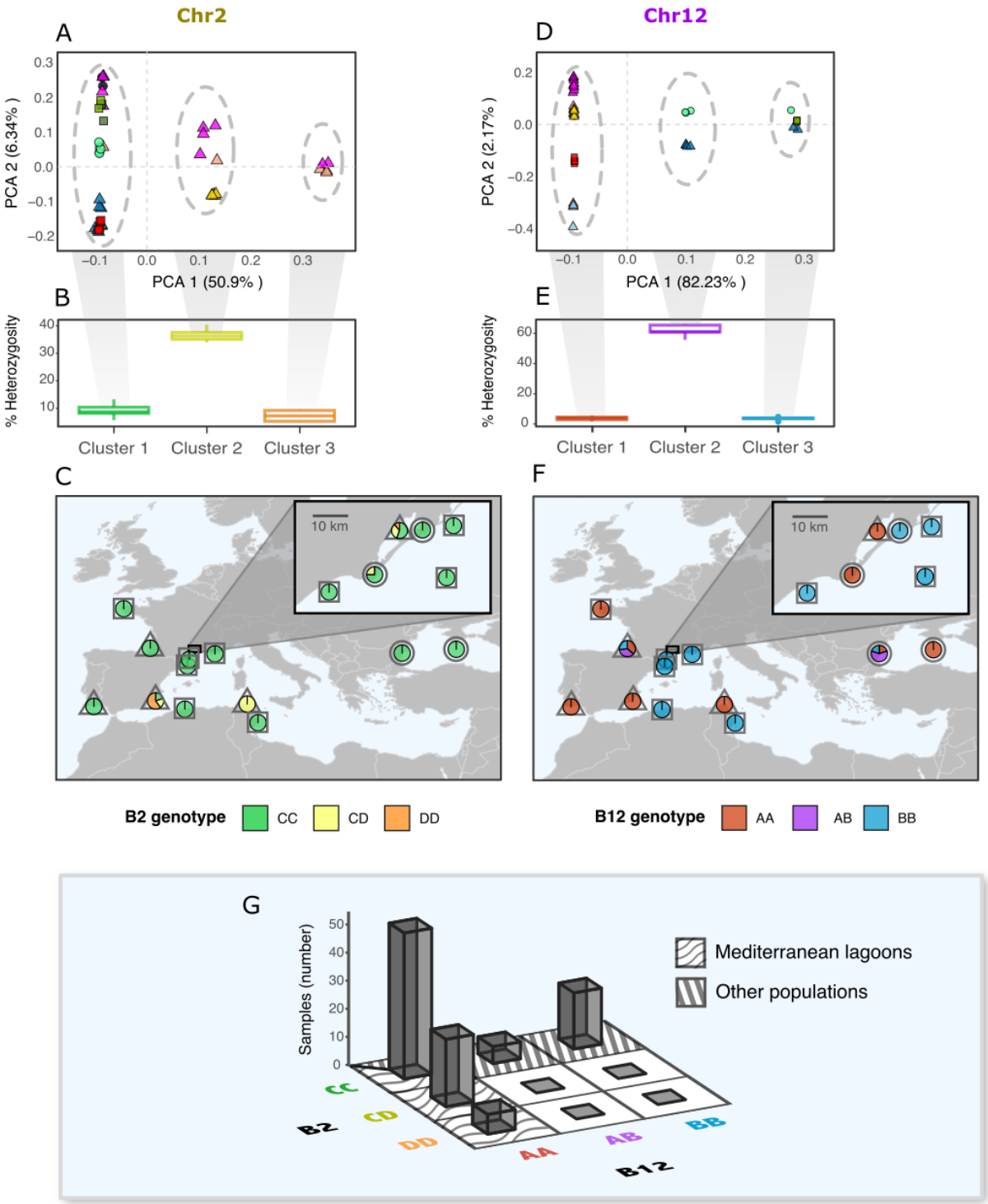

Chr2
A
PCA 2 (6.34%)
PCA 1 (50.9%)
B
% Heterozygosity
Cluster 1
Cluster 2
Cluster 3
Chr12
D
PCA 2 (2.17%)
PCA 1 (82.23%)
E
% Heterozygosity
Cluster 1
Cluster 2
Cluster 3
C
10 km
F
10 km
B2 genotype
CC
CD
DD
B12 genotype
AA
AB
BB
G
Samples (number)
Mediterranean lagoons
Other populations
B2
CC
CD
DD
B12
AA
AB
BB

Fig. 3 (previous page). Molecular diversity patterns observed for Chr2 (A-C) and Chr12 (D-F), which carry large inversions (B2 and B12). A & D) Chromosome-wide Principal Component Analysis (PCA) of 48 medium- to high-coverage samples (the GATK dataset) using the same symbols as in **Fig. 1**. The three clusters for Chr2 (A) and Chr12 (D) are illustrated by grey ellipses. B & E) Boxplots of observed heterozygosity (percentage of heterozygous sites) for each cluster. Heterozygosity was calculated at a chromosome-wide scale and averaged per sample. C & F) Maps of inversion genotype frequencies with respect to habitat type. Pie charts represent inversion genotypes and grey symbols indicate marine (square), lagoon (triangle) or uncharacterised (circle) habitats. G) Observed genotypic combinations at inversions B2 and B12, showing distinct associations between Mediterranean lagoons (B12 fixed for the A haplotype and B2 polymorphic) and other populations (B2 fixed for the C haplotype, B12 locally polymorphic), and the complete absence of 4 genotype combinations.

Mediterranean lagoons (**Fig. 3C**). Both in the Atlantic and in the Black Sea, all samples were fixed for haplotype C. The alternative D haplotype was found at varying frequencies (28% to 70%) in Thau, Bizerte and Murcia lagoons. All samples caught in the Bizerte lagoon were CD heterokaryotes (n=5), while 3 out of 5 samples from Murcia were DD homokaryotes. All samples carrying one or two D alleles had been classified as an AA homokaryote for B12 (including those from unknown habitats), indicating that the Chr2 polymorphism is private to Mediterranean lagoon populations.

Our results thus indicate that two large chromosomal inversions covering roughly 9% of the seahorse genome largely drive ecotype differentiation in the Mediterranean Sea. Lagoon populations are fixed at B12 for the A haplotype, whereas populations in the sea only carry the alternate B haplotype. In addition, there is a second inversion on Chr2 (with haplotypes C or D) which differentiates marine and lagoon ecotypes based on a polymorphism which is private to the lagoon ecotype. Even though the B12 inversion is also polymorphic in the Atlantic, it does not show differential fixation between habitat types in this zone. Instead, Atlantic populations to the north of Hossegor carry the A allele, while both A and B alleles are present from Hossegor to the south of Portugal (see also Riquet et al., 2019). Interestingly, the B haplotype on B12 and the D haplotype on B2 were never found together in the same individual (**Fig. 3G**), suggesting a possible negative interaction between the two alleles. In what follows, we sought to further characterise the respective histories of the two inversions, including their origins and ages.

Evolutionary history of the two inversions

An important question regarding the origin of the inversions is whether the heterogeneous divergence landscape has resulted from the differential erosion of past genome-wide divergence, or alternatively, from the emergence of newly divergent regions (i.e. secondary contact *versus* primary divergence). To address this question, we studied the distribution of coalescence times along the genome as inferred by tsdate (v0.1.5). Outside the inversion regions, we did not find a strong signature of ancient coalescence in the form of deep time to the most recent common ancestor (TMRCA) (**Fig. 4E**). Only the two chromosomes carrying inversions showed the presence of TMRCA values that were up to 5 times older than in the genome background, which also corresponded to their high d_{XY} values (**Fig. 2**). What is more, except for the inversions, regions of high F_{ST} were not shared between Atlantic lineages and Mediterranean ecotype contrasts, suggesting a lack of parallelism in the genome background (**Supplementary Fig. S13**).

Topology weighting conducted using Twisst (Martin & Van Belleghem, 2017) revealed that the most predominant topologies – both in the genome background as well as in the inversions – were consistent with the currently accepted genome-wide phylogeny of the genus (Li et al., 2021; Stiller et al., 2022). However, they simultaneously showed high levels of incomplete lineage sorting (ILS) near the ancestral nodes of the genus. If one of the two haplotypes on either Chr2 or Chr12 had been introduced through introgression from a closely related species, we would have expected the tree topologies in this region to differ from genome-wide topologies. In this case, an introgressed haplotype carried by a *H. guttulatus* individual would have grouped closer to the donor species than with the alternate haplotype of *H. guttulatus*. The topologies that did not group alternate haplotypes together within *H. guttulatus* amounted to only 4.57% and 15.7% of all topologies that were observed in the B2 and B12 regions, respectively (see **Fig. 4C & 4D** for the five most frequent of these topologies). Even when the two *H. guttulatus* haplotypes were not in the same grouping, they were generally not placed in distant positions (i.e. they still occurred in the same sub-branch). These topologies did not show any particular haplotype tending to group with a potential donor species, and both haplotypes showed shifted positions, most probably due to ILS. Given these findings, we conclude that the two inversion polymorphisms in *H. guttulatus* were not likely introduced through introgression with a closely related species.

Our results indicate that the B2 and B12 polymorphisms are most likely explained by chromosomal inversion events that took place within the *H. guttulatus* lineage. These inversions have been maintained as intraspecific polymorphisms for a long enough period of time for divergence to have accumulated between haplotypes. The divergence between A and B haplotypes on Chr12 was particularly high, since d_{XY} values for comparisons between opposite homokaryotes ranged between 0.69 and 0.76% (first and third quartiles). Absolute divergence between opposite B2 homokaryotes was slightly lower and showed more variance (first to third quartile, 0.36 to 0.57%) (**Supplementary Fig. S9**). TMRCA inferred in the B12 region were also higher than for the B2 region, and consensus phylogenies constructed with PhyML showed a deeper split for haplotypes A and B (**Fig. 4C**) than for C and D (**Fig. 4B**). We conclude that the inversions probably do not have the same age, and that the B12 polymorphism emerged before B2. Furthermore, it should be noted that divergence might have been underestimated in our study due to reference bias, since reads from AA individuals were mapped to our B reference (Hgutt_V1).

In addition to these inversions, other intra- and inter-chromosomal rearrangements were evidenced in alignments between our reference assembly (Hgutt_V1), HguttRefA and the *H. erectus* assembly. Due to the abundance of these structural rearrangements, we were not able to determine the ancestral arrangement for inversions B2 and B12 through comparison with the *H. erectus* assembly. These analyses showed that *H. erectus* chromosomes Chr11 and Chr12 were fused in the HguttRefA assembly to form chromosome JAOYMQ010000004.1 (**Supplementary Fig. S11**). This could potentially indicate that the A haplotype (carried by HguttRefA) on Chr12 is involved in a chromosomal fusion with Chr11 (contrary to the B haplotype), or alternatively, that this might represent a technical artefact of the HguttRefA assembly, which is based on short-read data. As for B2, we speculate that the discontinuous

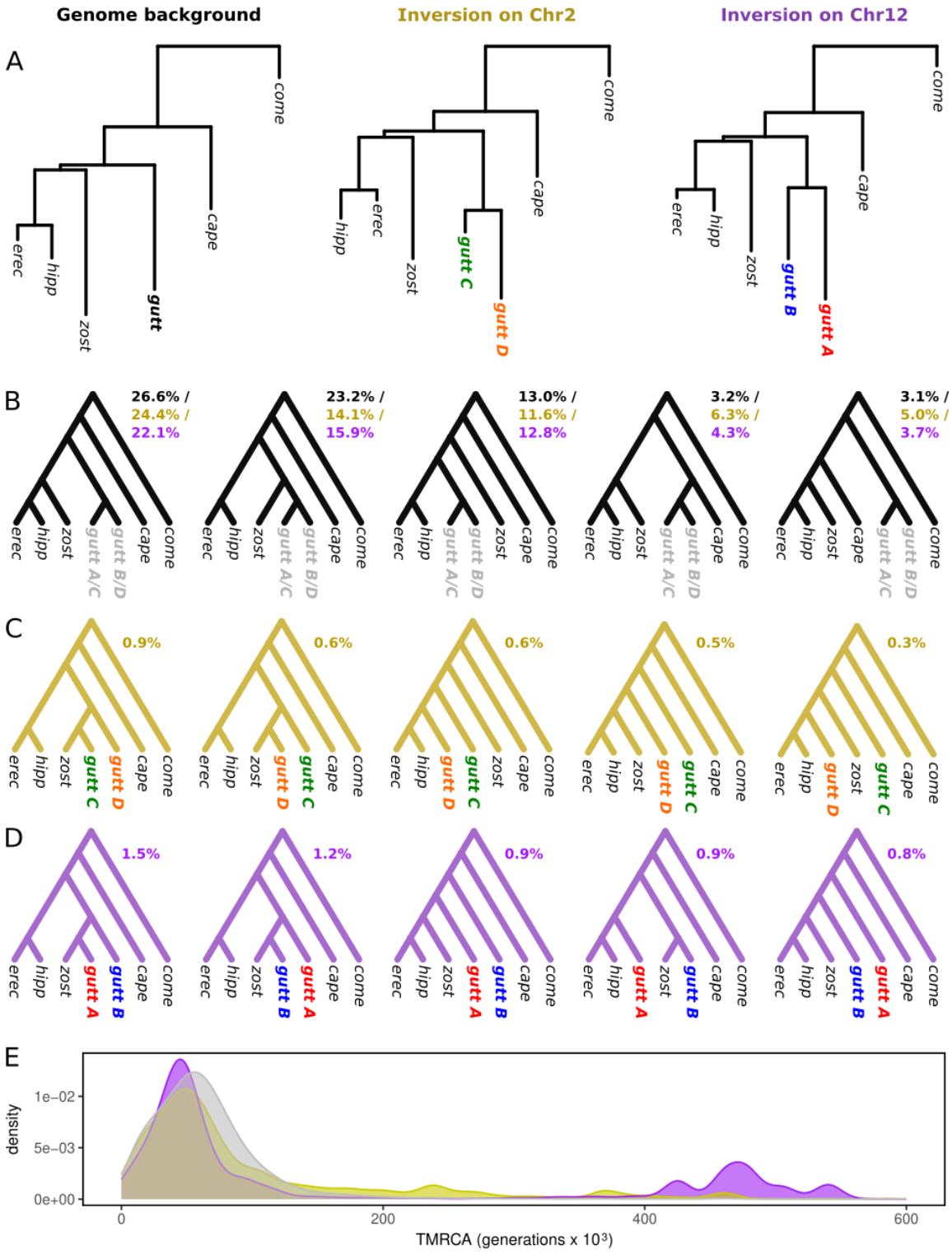

Genome background
Inversion on Chr2
Inversion on Chr12
A
B
C
D
E
erec
hipp
zost
gutt
gutt C
gutt D
gutt B
gutt A
cape
come
26.6% / 24.4% / 22.1%
23.2% / 14.1% / 15.9%
13.0% / 11.6% / 12.8%
3.2% / 6.3% / 4.3%
3.1% / 5.0% / 3.7%
gutt A/C
gutt B/D
0.9%
0.6%
0.6%
0.5%
0.3%
1.5%
1.2%
0.9%
0.9%
0.8%
density
1e−02
5e−03
0e+00
0
200
400
600
TMRCA (generations x 10³)

Fig. 4 (previous page). A) Maximum likelihood phylogenies for different regions of the seahorse genome. B) The five most predominant 7-taxon topologies identified with Twisst in the genome background as well as in the inversions. Relative abundance among all topologies for each specific region is written as a percentage for the genome background (top), for inversion B2 (middle, leaves *gutt C* and *gutt D*) and for inversion B12 (bottom, leaves *gutt A* and *gutt B*). C-D) Amongst topologies not placing alternate haplotypes as sister lineages, the five most common are shown for inversions B2 (C) and B12 (D). E) Distribution of coalescent times estimated with *tsdate* (v0.1.5) in the genome background (grey), on Chr2 (yellow) and on Chr12 (purple). The TMRCA (in units of 1k generations) was extracted for any random pair of haplotypes in every local genealogy along the genome. Abbreviations: *Come*: *H. comes. Cape*: *H. capensis. Zost*: *H. zosterae. Hipp*: *H. hippocampus. Erec*: *H. erectus*.

signal of high F_{ST} associated with this inversion (**Fig. 2C**) might be due to additional rearrangements that have altered the collinearity between the C and D haplotypes. The D haplotype might present a different intra-chromosomal structure to the C haplotype, resulting in a heterogeneous divergence landscape when mapped to the C reference. However, our analyses did not allow us to confirm this hypothesis, since both *H. guttulatus* assemblies were CC homokaryotes. Alternatively, regions of low F_{ST} and d_{XY} could be due to local erosion of divergence through recombination between inverted haplotypes.

We tested for gene flux between inverted and non-inverted haplotypes, which can be expected in the form of double crossover events or gene conversion. Plotting PC1 coordinates from local PCA along inversion regions allowed us to locate genome windows which were candidates for gene flux. This was based on local deviations from the three-cluster pattern typical of inversions (three groups of horizontal lines representing the three genotypes, **Fig. 5A and 5D**). B12 showed the classical three-cluster pattern at a large scale, whereas B2 showed a more "choppy" landscape, with long stretches of "irregular" patterns separating three-cluster windows. For both inversions, these deviated windows were further examined for evidence of local gene flux. We used inferred ARG trees to determine local ancestry along both inversions (**Fig. 5B, 5C & 5E**). We determined which trees showed the expected topology of reciprocal monophyly between non-inverted and inverted haplotypes grouping in two different clades (**Fig. 5B**), and which trees showed a discordant topology (**Fig. 5C**). Local trees grouping opposite haplotypes in the same clade indicated regions which were locally introgressed with a tract from the alternate haplotype. This allowed us to perform chromosome painting for each non-recombinant block associated with a particular tree. This approach revealed large-scale haplotypes corresponding to inverted and non-inverted ancestry, interspersed with local ancestry variation at a finer scale. Chromosome painting strongly reflected the patterns observed in the local PCA and F_{ST} landscapes (see **Supplementary Fig. S14, S15a and S15b** for painted chromosomes of the entire GATK dataset). For lower quality samples, we found evidence for phasing errors in the form of switches between maternal and paternal haplotypes, which were visible in heterokaryotic samples. These errors did not prevent us from locating introgressed tracts in homokaryotes, which spanned up to ~100 kb for B12 and up to ~500 kb for B2.

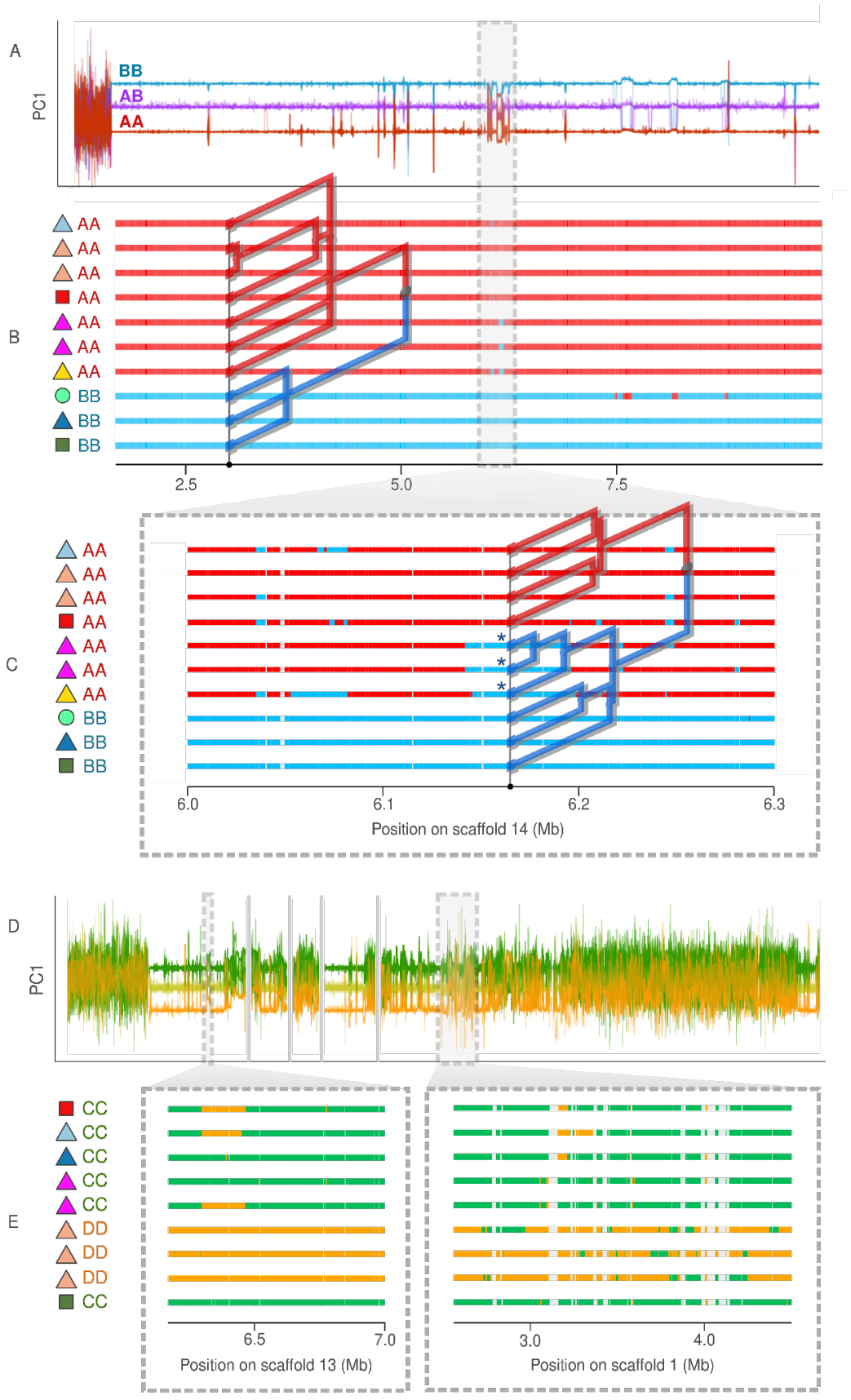

A
PC1
BB
AB
AA
B
AA
AA
AA
AA
AA
AA
AA
BB
BB
BB
2.5
5.0
7.5
C
AA
AA
AA
AA
AA
AA
AA
BB
BB
BB
6.0
6.1
6.2
6.3
Position on scaffold 14 (Mb)
D
PC1
E
CC
CC
CC
CC
CC
DD
DD
DD
CC
6.5
7.0
Position on scaffold 13 (Mb)
3.0
4.0
Position on scaffold 1 (Mb)

Fig. 5 (previous page). Gene flux between the inverted and non-inverted haplotypes on Chr12 (A-C) and on Chr2 (D-E). A & D) PC1 coordinates from local PCA plotted for non-overlapping 5 kb window along the inversions. Individual lines represent samples and are coloured according to inversion genotype. Vertical grey lines indicate the edges of scaffolds composing Chr2. B & C & E) Chromosome painting showing local ancestry within inversion regions as determined by our ARG-based approach. The superposed trees were constructed with *tsinfer* (v0.3.0) and illustrate how local ancestry at a given position was determined for each haplotype leaf. Horizontal coloured bars represent a subsample of all haplotypes and colours indicate inverted or non-inverted ancestry. Symbols on the left of each chromosome indicate sampling location and habitat type (see **Fig. 1A**), as well as the large-scale inversion genotype for each individual. C) Enlarged view of an introgressed region of B12, where certain A haplotypes (red) locally show B ancestry (blue). The tree at position 6.175 Mb shows that three recombinant A haplotypes (marked by asterisks) are grouped in the same clade as the B haplotypes. E) Enlarged views of two regions on Chr2 showing different levels of introgression.

Discussion

We present the first genome-wide study of intraspecific diversity in *H. guttulatus*, investigating multiple aspects of population structure related to geography, habitat type and genome architecture. We further characterise divergence between two previously described geographical lineages in the Atlantic and between marine-lagoon ecotypes in the Mediterranean Sea (Riquet et al., 2019). Our results reveal a sharp contrast between low levels of genome background differentiation and the substantial haplotype frequency differences observed at two megabase-scale chromosomal inversions differentiating Mediterranean ecotypes. We characterise the origin of these inversions and discuss the possible mechanisms responsible for their long-term maintenance. Lastly, we find evidence for gene flux between inverted alleles and address its role in shaping the dynamics and evolutionary fate of seahorse inversions.

Evolutionary origin of seahorse inversions

Our results indicate that inversions B2 and B12 have been maintained as polymorphisms for hundreds of thousands of generations, that is, well beyond the mean coalescence time inferred within populations (<100k generations, **Fig. 4E**). Estimating coalescence times with *tsdate* is known to be hampered by technical limitations (Brandt et al., 2022) that prevented our ability to precisely determine and compare the ages of B2 and B12. Among the reported limitations is the observation that *tsdate* tends to underestimate the oldest coalescence times. It is therefore conservative to assume that the ages we inferred for the inverted alleles greatly predate average within-population coalescence times. Consistent with this expectation, these regions also showed remarkably high levels of raw divergence compared to the rest of the genome (d_{XY} up to 1%, *vs.* maximum of 0.25% elsewhere). Moreover, the levels of nucleotide diversity associated with each haplotype were, given their respective frequencies, consistent with genome-wide average values (**Supplementary Fig. S10**). It is therefore likely that sufficient time has elapsed since the appearance of the inverted haplotypes for them to have progressed towards mutation-drift equilibrium.

Our finding that the two seahorse inversions represent ancient polymorphisms thus raises the question of their origin. One way that a divergent inverted haplotype may be introduced is through

hybridisation with a closely related species (e.g., Hsieh et al., 2019; Jay et al., 2018). We did not find evidence for introgressed ancestry at either of the inversions, since topology analysis did not support similarity between any inversion haplotypes and a potential donor species (**Fig. 4 B-E**). Instead, all major sources of genealogical conflict could be explained by incomplete lineage sorting (ILS) occurring at deep nodes in the *Hippocampus* genus phylogeny. Another scenario that could have given rise to similar heterogenous divergence landscapes, is secondary contact between previously isolated, divergent lineages. In this case, the inversions could have resisted re-homogenisation by gene flow due to recombination suppression (Lundberg et al., 2023; Rafajlović et al., 2021; Yeaman, 2013). It could have been expected that an erosion of past differentiation outside of the inversions would remain detectable by the presence of short divergent haplotypes produced by recombination between diverged ancestries. However, we did not find any such regions of deep coalescence comparable to that found in the inverted regions. It is thus most likely that the inversions have emerged within the *H. guttulatus* lineage and that they have remained polymorphic at the species level for long periods of time. The observation that north Atlantic and Mediterranean lagoon populations share the same haplotype at the B12 inversion suggests a shared history of divergence or past contact of these geographically isolated populations (see also Riquet et al., 2019). As for similar systems (Barrett et al., 2008; Belleghem et al., 2018; Le Moan et al., 2021), cycles of isolation and secondary contact can be prone to the long-term maintenance of inversion polymorphisms by their recurrent reuse at different moments in time.

Molecular mechanisms facilitating the emergence of inversions have been studied in detail using long-read sequencing in deer mice, showing that inversion breakpoints tend to occur in centromeric and telomeric regions and to be flanked by LIRs (Harringmeyer & Hoekstra, 2022). Here, we were able to directly demonstrate, using linked-read data, that the 8.2 Mb inversion B12 occurs near a chromosome extremity and has its breakpoints in an approximately 1 kb LIR bearing a Tan9 monomer. Tandem repeats containing Tan9 are widespread throughout the *H. guttulatus* genome and may therefore facilitate ectopic recombination, leading to an increased rate of formation of new structural variants. Although we could not perform such a detailed analysis for the breakpoint regions of B2, it would be interesting to test whether the emergence of this inversion has also been promoted by the presence of recombinogenic elements (Wang & Leung, 2006), such as Tan9 LIRs.

Maintenance of polymorphisms B12 and B2

How ancient inversion polymorphisms persist over time has been a long-standing question in evolutionary biology (reviewed in Wellenreuther & Bernatchez, 2018). In their review outlining the different stages characterising the lifetime evolution of inversions, Faria et al. (2019a) discuss the evolutionary changes that an inversion might undergo from the moment of its appearance to its loss or fixation. By placing our results in this framework, we sought to distinguish between the mechanisms that might maintain the two inversion polymorphisms in our system, and to provide possible explanations for their respective patterns. We propose that the B12 polymorphism is most comparable in its characteristics to Type I inversion polymorphisms (i.e. divergent between

populations), while B2 probably calls on the interaction of forces linked to both Type I and Type II (i.e. within-population variation) polymorphisms.

The B12 polymorphism corresponds to the genomic island identified by Riquet et al. (2019) (**Supplementary Fig. S16**) as the main region differentiating geographical lineages in the Atlantic and ecotypes in the Mediterranean Sea. B12 shows local fixation for a given haplotype within almost all populations (**Fig. 3F**), in accordance with the pattern expected for Type I polymorphisms (Faria et al. 2019a). In this scenario, within-population polymorphism is not maintained unless there is sufficient gene flow, but among-population variation is conferred by divergent selection on alternate haplotypes. Given the differential fixation of B12 haplotypes between adjacent lagoon (A haplotype) and marine (B haplotype) sites in the Mediterranean Sea, it is possible that the inversion is either underdominant, under negative epistasis, or affected by selection for locally favoured alleles. Regardless of the specific form of selection at work, lack of recombination between inverted haplotypes ensures that the genes in the A and B haplotypes are inherited together, collectively forming a barrier to gene flow. Over a long period of time, the accumulation of new mutations between inverted haplotypes leads to high levels of divergence, as is reflected by high d_{XY} values observed between AA and BB homokaryotes (**Fig. 2B**). Although there is little opportunity for gene flow between alternate haplotypes, recombination can proceed normally within every population that is fixed for a given haplotype, since crossing over is unimpeded in homokaryotes. In this way, it has been suggested that the accumulation of mutation load in inverted regions would not be substantially higher than in the collinear genome (Berdan et al. 2021). Besides being a theoretical prediction, empirical examples are provided by inversions in deer mice that show local fixation and lack of mutation load (Harringmeyer & Hoekstra, 2022), as well as inversions in sunflower populations fixed for one arrangement having lower mutation load compared to polymorphic populations (Huang et al., 2022). We can therefore speculate that the long-term subsistence of B12 inversion as a Type I polymorphism has not been hampered by the accumulation of deleterious mutations.

Regarding the possible selective mechanisms that are at play, lower fitness of heterozygotes (i.e., underdominance) is expected at Type I polymorphisms, since divergent selection and independent evolution of haplotypes may lead to the accumulation of DM incompatibilities within the inversion. We identified only seven samples (out of 112) that carried the AB genotype at B12, possibly pointing to a deficit of heterokaryotes at the inter-population level. These seven samples originated from only two populations (Hossegor, Bay of Biscay; and Varna, Black Sea) where both haplotypes A and B were present and where heterokaryotes were locally common (4 out of 11 and 3 out of 5 samples, respectively). If B12 heterokaryotes are selected against, these populations might represent underdominant clines, as was put forward by Riquet et al. (2019) for polymorphic populations in the contact zone in the Basque country. Furthermore, we also observe a breakdown of the perfect association that exists between habitat and B12 genotype outside of the Mediterranean Sea. Atlantic and Black Sea populations do not carry different inversion haplotypes in marine and lagoon environments, as already observed by Riquet et al. (2019). This partial decoupling from habitat type could be explained by predominantly intrinsic rather than extrinsic incompatibility between A and B. The difference in patterns observed between the Atlantic and the Mediterranean Sea may also involve epistatic interactions with loci on other

chromosomes, such as the B2 inversion (see below). In any case, it remains possible that the B12 inversion is to some extent directly involved in differential adaptation to marine and lagoon environments, particularly in the Mediterranean Sea.

In contrast to B12, the B2 inversion is maintained as a polymorphism which is private to Mediterranean lagoons, rather than being differentially fixed between marine and lagoon environments. Classical work that has addressed the long-term maintenance of inversion polymorphisms has evoked an important role of balancing selection in maintaining within-population variation (e.g. reviewed in Llaurens et al., 2017; Wellenreuther & Bernatchez, 2018). This results in Type II polymorphisms *sensu* Faria et al. (2019a), which could explain the distribution of the B2 polymorphism in the Mediterranean Sea. We observed relatively high frequencies of CD genotypes in Mediterranean lagoons (Th: 12 out of 36; Mu: 3 out of 5; Bz: 5 out of 5), possibly pointing to a form of heterozygote advantage. However, when grouping these locations to test for Hardy-Weinberg deviation, we were unable to detect any significant excess of heterozygotes (p=0.344). This lack of significant deviation has been reported in other studies (reviewed in Waples, 2015) and does not necessarily disprove the existence of overdominance. In theory, the limited distribution of the B2 polymorphism combined with intermediate haplotype frequencies at the populational level, might result in a reduced effective population size and a higher mutation load associated with both or only the minor haplotype (Berdan et al., 2021). Additionally, the B2 inversion is nearly three times longer than B12, suggesting that it is more likely to have captured multiple recessive deleterious mutations that are inherited as a single block (Connallon & Olito, 2022). For inversions that suffer from high mutation load, theory predicts that polymorphism could be maintained by pseudo-overdominance (Berdan et al., 2021), which is also supported by empirical studies in insects (Jay et al., 2021; Yang et al., 2002). For these reasons, we suggest that compensation of the recessive mutation load in heterokaryotes could contribute to the persistence of the B2 polymorphism in Mediterranean lagoons.

The simultaneous action of forces associated with both Type I and Type II polymorphisms has been shown to potentially give rise to a range of different inversion equilibrium frequencies (Faria et al., 2019a). For instance, selection for heterokaryotes combined with local adaptation can result in polymorphism in one environment and fixation in the other. It is therefore plausible that a similar commixture of forces could be at work to explain the fixation of B2 haplotype C in marine *H. guttulatus* populations, while polymorphism is maintained in Mediterranean lagoons. If the B2 polymorphism is maintained by pseudo-overdominance, we could ask why the D haplotype is not present outside of Mediterranean lagoons. The answer may potentially be connected to local adaptation alleles that are carried in the haplotypes of B2, or in epistatic association with B12. If D provides local advantage in lagoons, or if D/A combinations are favoured over C/A, it could help maintain polymorphism in these environments. On the other hand, D would be outcompeted elsewhere (e.g. in Mediterranean marine habitats) due to lack of selective advantage. Although current knowledge does not allow us to answer all questions regarding the maintenance of the B2 polymorphism, we suspect that it involves an interaction of forces which could include divergent selection of differing strengths and a form of balancing selection (e.g. pseudo-overdominance). All of these processes should also be affected by the level and direction of effective migration connecting different populations.

Long-term fate of inversion polymorphisms in the long-snouted seahorse

Inversion haplotypes are not expected to persist indefinitely if the evolutionary mechanisms underlying polymorphism are subject to change. An inversion might eventually fix one arrangement throughout the species range, or alternatively, couple with other genomic components of reproductive isolation and fix differentially between incipient species (Faria et al., 2019a). The long-term fate of an inversion thus depends on the processes that drive its dynamics throughout its lifetime. For example, Berdan et al. (2021) studied the "feedback loop" between allelic content and haplotype frequency, caused by the accumulation of deleterious mutations, which in turn affects the frequencies of the different karyotypes. Another process that can influence the long-term fate of an inversion, is the exchange of genetic material between alternate haplotypes through gene conversion or double crossover during meiosis (i.e., gene flux). In the current section we discuss these various points and attempt to make predictions about the long-term fates of the B2 and B12 polymorphisms segregating in *H. guttulatus*.

Theoretical (reviewed in, e.g. Hoffmann & Rieseberg, 2008) and empirical studies (e.g. Huang et al., 2020; Lohse et al., 2015; Noor et al., 2001) have found that, under certain conditions, the presence of (Type I) inversion polymorphisms may facilitate speciation. In fish, multiple empirical studies have highlighted the role of chromosomal inversions in local adaptation, ecotype formation and speciation (Berg et al., 2016; Cayuela et al., 2020; Jones et al., 2012; Le Moan et al., 2021; Matschiner et al., 2022; Pettersson et al., 2019; Tigano et al., 2021). Mediterranean ecotypes of long-snouted seahorse show differential fixation at inversion B12, which is under divergent selection (Type I polymorphism, see previous section) and could eventually facilitate speciation between marine and lagoon ecotypes. We found evidence for very low levels of gene flux taking place in this inversion, suggesting that gene exchange in heterokaryotes is either rare or selected against. Since exchanged segments were generally longer than a few kilobases and occurred in the middle of the inversion, they most likely resulted from double crossovers rather than gene conversion. This was contrary to what was found for cod inversions where gene conversion predominated (Matschiner et al., 2022), but was in line with theoretical expectations for long inversions (Navarro et al., 1997). Introgressed tracts of B ancestry within A haplotypes were found across the species range, but were nonetheless generally small in size (up to 50 kb) (**Supplementary Fig. S14**). This suggests that introgression took place a long time ago and that the remaining tracts represent recombined segments that have passed the filter of selection. Evidence for introgression in the opposite direction (A into B) was observed only in the Black Sea population, where both haplotypes segregate and opportunities for gene flux are increased. These tracts of A ancestry within B haplotypes were slightly longer (up to 100 kb) and more numerous, potentially suggesting more recent gene flux.

Whatever the direction of gene flux, the fact that introgressed segments occupy only few and relatively narrow genomic regions in B12, argues for selection acting against introgressed ancestry. This is consistent with the existence of multiple selected mutations that have accumulated within the inversion over the long term (e.g., DM incompatibilities; Navarro & Barton, 2003). If B12 is responsible for strong underdominance and further becomes involved in

reinforcement by coupling with premating isolating mechanisms, it may eventually strengthen reproductive isolation between lagoon and marine populations (Faria et al., 2019a). It can be noted that some level of genome-wide differentiation is already observed between Mediterranean ecotypes, thus indicating a significant reduction in effective gene flow at a small spatial scale relative to dispersal. However, the contribution of B12 to gene flow reduction is more uncertain in the Atlantic and the Black Sea. Although a weak association has been observed between B12 and habitat type along the Portuguese coasts, no associated genetic structure was detected in the genomic background (Riquet et al., 2019). Surprisingly, we found an increase in the frequency of the A haplotype in Faro lagoon over the last ten years, but further study will be necessary to confirm whether this observation is a true temporal trend, an indication of cryptic microhabitat variation, or sampling noise.

If the B12 polymorphism lends itself more to speciation than to universal fixation, the long-term fate of the B2 (i.e. Type II) polymorphism is less straightforward to predict. Since the frequencies of this inversion probably result from a balance between different processes, namely a form of divergent along with balancing selection, it is unclear whether B2 will eventually undergo differential fixation between habitats, or universal fixation of one arrangement. Furthermore, our large-scale assembly of Chr2 is still only partly resolved, and we cannot rule out the possibility that additional chromosomal rearrangements have affected its divergence landscape and evolutionary trajectory. For example, multiple inversions occurring in the same region might have extended the block of high LD, as is the case for adjacent (Jay et al., 2021) or nested inversions (Maggiolini et al., 2020). In contrast to B12, we also found ample evidence for the erosion of divergence in B2 through gene flux, as illustrated by the detection of many introgressed segments (**Supplementary Fig. S15a and S15b**) and “suspension bridge” patterns in the F_{ST} landscape (**Fig. 2C**). The intensity of gene flux was clearly heterogeneous along this chromosome, since we found evidence for relatively recent double crossover events (i.e. larger introgressed segments than in B12) as well as one region where low F_{ST} values (comparable to the genome background) are bordered by high differentiation segments, indicating that divergence has been eroded over time. Gene flux is therefore likely to impact the fate of this inversion, especially if the dynamics of B2 are largely driven by mutation load, as we discussed in the previous section. Recombination between haplotypes might favour the removal of deleterious mutations, as has been shown through simulations, where even low levels of gene conversion were sufficient to mitigate mutation load (Berdan et al., 2021). Given enough time, gene flux between B2 haplotypes could thus weaken the pseudo-overdominance that is suspected of maintaining them.

The framework laid out by Faria et al. (2019a) does not specifically consider the dynamics of multiple inversions co-existing in the same system, but leaves such interactions as an outstanding question. The presence of other polymorphic inversions could indeed impact the establishment, maintenance, and long-term fate of a given inversion, for example through epistatic interactions or coupling. The two inversion polymorphisms segregating in *H. guttulatus* might present a case study for such potential interactions. Timing and demographic context could be important in determining these dynamics, since these factors could influence the allelic contents of a new inversion and its probability to establish. For example, if the B12 polymorphism was already established and fixed between certain populations by the time B2 emerged, it could have affected

the frequency trajectory of the new B2 polymorphism. Furthermore, it is possible that there are epistatic interactions between B2 and B12, which co-occur in lagoon populations. Epistatic interactions could potentially explain why there is a strong association between B12 and habitat type in the Mediterranean, while the absence of B2 polymorphism in the Atlantic would only produce weak associations of B12 with habitat. The question therefore remains as to whether B2 contributes to speciation between marine and lagoon lineages in the Mediterranean through coupling with B12. Interestingly, the two molecular pathways that were enriched within the inversions have interconnected functions in reproduction. The estrogen signalling pathway enriched in B12 has been shown to play a role in seahorse sexual dimorphism, gonad development, and stimulation of parturition in pregnant seahorses (Qin et al., 2019; Whittington et al., 2015; H. Zhang et al., 2022). This involves the regulation of expression of the two estrogen receptor genes, *esr1* and *esr2*, which are both present on B12. In mice, estrogenic regulation by *esr1* induces expression changes in a population of neurons that mediate estrogen feedback mechanisms affecting the neuroactive ligand-receptor interaction pathway (Göcz et al., 2022), corresponding to the pathway that was enriched in B2. Functional incompatibilities among genes that interact in the regulation of these pathways may potentially disrupt the reproductive function in incompatible genotypes at the two inversions involved in seahorse ecotype differentiation. Future directions for research could address these aspects by focusing on polymorphic populations, attempting to characterise the fitnesses of different karyotypes in different environments, and by quantifying gene expression and mutation load. Studying gene flux may also reveal more about the loci contained in each arrangement, as we should expect erosion of divergence in regions carrying selectively neutral or disadvantageous mutations, and maintenance in regions that are under divergent selection.

Acknowledgements

The data were partly produced and analysed with the support of the *GenSeq* genotyping and sequencing platform and the *MBB* Montpellier Bioinformatics Biodiversity platform, both being supported by ANR program ”Investissements d’avenir” (ANR-10-LABX-04-01). We thank Rémy Dernat and Khalid Belkhir for their assistance in data storage, management and processing, and Iago Bonnici for support on variant orientation. DNA extraction of historical samples was carried out in a clean lab at ISEM, Montpellier (Plateforme d’ADN dégradé, LabEx CeMEB). We thank the Montpellier GenomiX platform for constructing the reference genome's 10X Chromium library. We are grateful to the National Museum of Natural History (MNHN Paris, Agnès Dettaï) and colleagues who provided us with samples, as well as those who facilitated or participated in sampling: Jorge Palma and Rita Castilho (CCMAR, Portugal), Cristina Mena (Hippocampus association, Spain), Patrick Louisy (peau-bleue association, CPIE Bassin de Thau, France), Lucy Woodall (University of Oxford, UK), and citizens from the Sète region who provided dried seahorses samples. We thank Huixian Zhang of the Key Laboratory of Tropical Marine Bio-Resources and Ecology, South China Sea Institute of Oceanology, for providing the *Hippocampus erectus* reference genome. Finally, we thank Thomas Broquet for his valuable inputs on the manuscript. This work was supported by the ANR grant CoGeDiv ANR-17-CE02-0006-01 to PAG, and by a Languedoc-Roussillon Region “Chercheur(se)s d'avenir” grant to NB (Connect7 project), with the support of the Occitanie Regional Council’s program «Key challenge BiodivOc». Mobility

during the project was partly funded by an ESEB Godfrew Hewitt Mobility Award as well as a Laura Corrigan conservation grant.

Data accessibility and benefit-sharing statement

Our reference genome assembly will be deposited in GenBank, and raw sequence reads from whole-genome re-sequencing data (86 individuals) will be deposited in the GenBank Sequence Read Archive under the accession code BioProject ID XXXXXX. **Supplementary File S2** contains scripts and commands used for all bioinformatic analyses. Full detailed metadata and results may be consulted in the supplied Rmarkdown report in the **Context** section (Chapter_II_Suppl_Report.html). Regarding the benefits generated, we reanalysed genetic samples from countries (other than France) that were published in a previous study (Riquet et al., 2019) where collaborators who provided samples were already included as co-authors. The results of this research have been shared with the sample providers and the broader scientific community, in particular through the sharing of genomic sequences on NCBI public databases. Other genetic resources will be made available on request.

Author contributions

LM, BG, NB and P-AG conceived the study. LM wrote the manuscript with inputs from all co-authors. Sampling was conducted by NB, FR, PB, RC and P-AG. CA performed HMW DNA extraction and preparation for the reference genome 10X library. DNA extraction for WGS libraries was performed by LM, PB, and FR. Aspects of degraded/historical DNA extraction and analysis were handled by AF, CDS, BG and LM. WGS library preparation was done by LM and FC. FC, AB and ED were responsible for haplotagging library construction and sequencing. P-AG produced the reference genome and analysed linked-read sequencing data. PB provided scripts for genotyping high-coverage samples. LM performed all other bioinformatics and population genomic analyses. P-AG managed financial support.

References

Barrett, R. D. H., Rogers, S. M., & Schluter, D. (2008). Natural selection on a major armor gene in threespine stickleback. *Science (New York, N.Y.)*, *322*(5899), 255-257. https://doi.org/10.1126/science.1159978

Barry, P., Broquet, T., & Gagnaire, P.-A. (2022). Age-specific survivorship and fecundity shape genetic diversity in marine fishes. *Evolution letters*, *6*(1), 46-62.

Belleghem, S. M. V., Vangestel, C., Wolf, K. D., Corte, Z. D., Möst, M., Rastas, P., Meester, L. D., & Hendrickx, F. (2018). Evolution at two time frames: Polymorphisms from an ancient singular divergence event fuel contemporary parallel evolution. *PLOS Genetics*, *14*(11), e1007796. https://doi.org/10.1371/journal.pgen.1007796

Benson, G. (1999). Tandem repeats finder: A program to analyze DNA sequences. *Nucleic Acids Research*, *27*(2), 573. https://doi.org/10.1093/nar/27.2.573

Berdan, E. L., Blanckaert, A., Butlin, R. K., & Bank, C. (2021). Deleterious mutation accumulation and the long-term fate of chromosomal inversions. *PLOS Genetics*, *17*(3), e1009411. https://doi.org/10.1371/journal.pgen.1009411

Berg, P. R., Star, B., Pampoulie, C., Sodeland, M., Barth, J. M. I., Knutsen, H., Jakobsen, K. S., & Jentoft, S. (2016). Three chromosomal rearrangements promote genomic divergence between migratory and stationary ecotypes of Atlantic cod. *Scientific Reports*, *6*(1), Article 1. https://doi.org/10.1038/srep23246

Brůna, T., Hoff, K. J., Lomsadze, A., Stanke, M., & Borodovsky, M. (2021). BRAKER2: Automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein database. *NAR Genomics and Bioinformatics*, *3*(1), lqaa108. https://doi.org/10.1093/nargab/lqaa108

Butlin, R. K. (2005). Recombination and speciation. *Molecular Ecology*, *14*(9), 2621-2635. https://doi.org/10.1111/j.1365-294X.2005.02617.x

Cabanettes, F., & Klopp, C. (2018). D-GENIES: Dot plot large genomes in an interactive, efficient and simple way. *PeerJ*, *6*, e4958. https://doi.org/10.7717/peerj.4958

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., & Madden, T. L. (2009). BLAST+: Architecture and applications. *BMC Bioinformatics*, *10*(1), 421. https://doi.org/10.1186/1471-2105-10-421

Campos, P. F., & Gilbert, T. M. P. (2012). DNA extraction from formalin-fixed material. *Methods in Molecular Biology (Clifton, N.J.)*, *840*, 81-85. https://doi.org/10.1007/978-1-61779-516-9_11

Cantalapiedra, C. P., Hernández-Plaza, A., Letunic, I., Bork, P., & Huerta-Cepas, J. (2021). eggNOG-mapper v2: Functional annotation, orthology assignments, and domain prediction at the metagenomic scale. *Molecular Biology and Evolution*, *38*(12), 5825-5829. https://doi.org/10.1093/molbev/msab293

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., & 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. https://doi.org/10.1111/mec.15499

Challis, R., Richards, E., Rajan, J., Cochrane, G., & Blaxter, M. (2020). BlobToolKit – Interactive quality assessment of genome assemblies. *G3 Genes|Genomes|Genetics*, *10*(4), 1361-1374. https://doi.org/10.1534/g3.119.400908

Charlesworth, B. (2009). Effective population size and patterns of molecular evolution and variation. *Nature Reviews Genetics*, *10*(3), 195-205. https://doi.org/10.1038/nrg2526

Chen, S., Zhou, Y., Chen, Y., & Gu, J. (2018). fastp: An ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*, *34*(17), i884-i890. https://doi.org/10.1093/bioinformatics/bty560

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

Connallon, T., & Olito, C. (2022). Natural selection and the distribution of chromosomal inversion lengths. *Molecular Ecology*, *31*(13), 3627-3641. https://doi.org/10.1111/mec.16091

Dai, H., & Guan, Y. (2020). The Nubeam reference-free approach to analyze metagenomic sequencing reads. *Genome Research*, *30*(9), 1364-1375. https://doi.org/10.1101/gr.261750.120

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., & 1000 Genomes Project Analysis Group. (2011). The variant call format and VCFtools. *Bioinformatics*, *27*(15), 2156-2158. https://doi.org/10.1093/bioinformatics/btr330

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

Delaneau, O., Zagury, J.-F., Robinson, M. R., Marchini, J. L., & Dermitzakis, E. T. (2019). Accurate, scalable and integrative haplotype estimation. *Nature Communications*, *10*(1), Article 1. https://doi.org/10.1038/s41467-019-13225-y

Doyle, J. J., & Doyle, J. L. (Eds.). (1987). A rapid DNA isolation procedure for small quantities of fresh leaf tissue. *Phytochemical Bulletin*, *19*, 11-15.

Faria, R., Johannesson, K., Butlin, R., & Westram, A. (2019a). Evolving inversions. *Trends in Ecology & Evolution*, *34*(3), 239-248. https://doi.org/10.1016/j.tree.2018.12.005

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., & Butlin, R. K. (2019b). Multiple chromosomal rearrangements in a hybrid zone between *Littorina saxatilis* ecotypes. *Molecular Ecology*, *28*(6), 1375-1393. https://doi.org/10.1111/mec.14972

Faria, R., & Navarro, A. (2010). Chromosomal speciation revisited : Rearranging theory with pieces of evidence. *Trends in Ecology & Evolution*, *25*(11), 660-669. https://doi.org/10.1016/j.tree.2010.07.008

Ferrari, G., Atmore, L. M., Jentoft, S., Jakobsen, K. S., Makowiecki, D., Barrett, J. H., & Star, B. (2022). An accurate assignment test for extremely low-coverage whole-genome sequence data. *Molecular Ecology Resources*, *22*(4), 1330-1344. https://doi.org/10.1111/1755-0998.13551

Flynn, J. M., Hubley, R., Goubert, C., Rosen, J., Clark, A. G., Feschotte, C., & Smit, A. F. (2020). RepeatModeler2 for automated genomic discovery of transposable element families. *Proceedings of the National Academy of Sciences*, *117*(17), 9451-9457. https://doi.org/10.1073/pnas.1921046117

Ge, S. X., Jung, D., & Yao, R. (2020). ShinyGO: A graphical gene-set enrichment tool for animals and plants. *Bioinformatics*, *36*(8), 2628-2629. https://doi.org/10.1093/bioinformatics/btz931

Guichard, A., Legeai, F., Tagu, D., & Lemaitre, C. (2023). MTG-Link: Leveraging barcode information from linked-reads to assemble specific loci. *BMC Bioinformatics*, *24*(284). https://doi.org/10.1186/s12859-023-05395-w

Guindon, S., Dufayard, J.-F., Lefort, V., Anisimova, M., Hordijk, W., & Gascuel, O. (2010). New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Systematic Biology*, *59*(3), 307-321. https://doi.org/10.1093/sysbio/syq010

Harringmeyer, O. S., & Hoekstra, H. E. (2022). Chromosomal inversion polymorphisms shape the genomic landscape of deer mice. *Nature Ecology & Evolution*, 1-15. https://doi.org/10.1038/s41559-022-01890-0

Haÿ, V., Mennesson, M. I., Dettaï, A., Bonillo, C., Keith, P., & Lord, C. (2020). Needlepoint non-destructive internal tissue sampling for precious fish specimens. *Cybium: Revue Internationale d'Ichtyologie*, *44*(1), 73-79. https://doi.org/10.26028/cybium/2020-441-010

Hill, W. G., & Robertson, A. (1966). The effect of linkage on limits to artificial selection. *Genetics Research*, *8*(3), 269-294. https://doi.org/10.1017/S0016672300010156

Hoffmann, A. A., & Rieseberg, L. H. (2008). Revisiting the impact of inversions in evolution: from population genetic markers to drivers of adaptive shifts and speciation? *Annual review of ecology, evolution, and systematics*, *39*, 21-42. https://doi.org/10.1146/annurev.ecolsys.39.110707.173532

Hsieh, P., Vollger, M. R., Dang, V., Porubsky, D., Baker, C., Cantsilieris, S., Hoekzema, K., Lewis, A. P., Munson, K. M., Sorensen, M., Kronenberg, Z. N., Murali, S., Nelson, B. J., Chiatante, G., Maggiolini, F. A. M., Blanché, H., Underwood, J. G., Antonacci, F., Deleuze, J.-F., & Eichler, E. E. (2019). Adaptive archaic introgression of copy number variants and the discovery of previously unknown human genes. *Science*, *366*(6463), eaax2083. https://doi.org/10.1126/science.aax2083

Huang, K., Andrew, R. L., Owens, G. L., Ostevik, K. L., & Rieseberg, L. H. (2020). Multiple chromosomal inversions contribute to adaptive divergence of a dune sunflower ecotype. *Molecular Ecology*, *29*(14), 2535-2549. https://doi.org/10.1111/mec.15428

Huang, K., Ostevik, K. L., Elphinstone, C., Todesco, M., Bercovich, N., Owens, G. L., & Rieseberg, L. H. (2022). Mutation load in sunflower inversions is negatively correlated with inversion heterozygosity. *Molecular Biology and Evolution*, *39*(5), msac101. https://doi.org/10.1093/molbev/msac101

Ignatieva, A., Favero, M., Koskela, J., Sant, J., & Myers, S. R. (2023). The distribution of branch duration and detection of inversions in ancestral recombination graphs. *bioRxiv* doi: 10.1101/2023.07.11.548567

Jay, P., Whibley, A., Frézal, L., Rodríguez de Cara, M. Á., Nowell, R. W., Mallet, J., Dasmahapatra, K. K., & Joron, M. (2018). Supergene evolution triggered by the introgression of a chromosomal inversion. *Current Biology*,

28(11), 1839-1845.e3. https://doi.org/10.1016/j.cub.2018.04.072
Jay, P., Chouteau, M., Whibley, A., Bastide, H., Parrinello, H., Llaurens, V., & Joron, M. (2021). Mutation load at a mimicry supergene sheds new light on the evolution of inversion polymorphisms. *Nature Genetics*, *53*(3), 288-293. https://doi.org/10.1038/s41588-020-00771-1
Jones, F. C., Grabherr, M. G., Chan, Y. F., Russell, P., Mauceli, E., Johnson, J., Swofford, R., Pirun, M., Zody, M. C., White, S., Birney, E., Searle, S., Schmutz, J., Grimwood, J., Dickson, M. C., Myers, R. M., Miller, C. T., Summers, B. R., Knecht, A. K., … Kingsley, D. M. (2012). The genomic basis of adaptive evolution in threespine sticklebacks. *Nature*, *484*(7392), Article 7392. https://doi.org/10.1038/nature10944
Keightley, P. D., & Jackson, B. C. (2018). Inferring the probability of the derived vs. the ancestral allelic state at a polymorphic site. *Genetics*, *209*(3), 897-906. https://doi.org/10.1534/genetics.118.301120
Kelleher, J., Wong, Y., Wohns, A. W., Fadil, C., Albers, P. K., & McVean, G. (2019). Inferring whole-genome histories in large population datasets. *Nature Genetics*, *51*(9), Article 9. https://doi.org/10.1038/s41588-019-0483-y
Kim, D., Paggi, J. M., Park, C., Bennett, C., & Salzberg, S. L. (2019). Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nature Biotechnology*, *37*(8), 907-915. https://doi.org/10.1038/s41587-019-0201-4
Kirkpatrick, M., & Barton, N. (2006). Chromosome inversions, local adaptation and speciation. *Genetics*, *173*(1), 419-434. https://doi.org/10.1534/genetics.105.047985
Kirkpatrick, M. (2010). How and why chromosome inversions evolve. *PLOS Biology*, *8*(9), e1000501. https://doi.org/10.1371/journal.pbio.1000501
Kirov, I., Gilyok, M., Knyazev, A., & Fesenko, I. (2018). Pilot satellitome analysis of the model plant, *Physcomitrella patens*, revealed a transcribed and high-copy IGS related tandem repeat. *Comparative Cytogenetics*, *12*(4), 493-513. https://doi.org/10.3897/CompCytogen.v12i4.31015
Korneliussen, T. S., Albrechtsen, A., & Nielsen, R. (2014). ANGSD: Analysis of next generation sequencing data. *BMC Bioinformatics*, *15*(1), 356. https://doi.org/10.1186/s12859-014-0356-4
Korunes, K. L., & Noor, M. A. F. (2019). Pervasive gene conversion in chromosomal inversion heterozygotes. *Molecular Ecology*, *28*(6), 1302-1315. https://doi.org/10.1111/mec.14921
Kulmuni, J., Butlin, R. K., Lucek, K., Savolainen, V., & Westram, A. M. (2020). Towards the completion of speciation: The evolution of reproductive isolation beyond the first barriers. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *375*(1806), 20190528. https://doi.org/10.1098/rstb.2019.0528
Le Moan, A., Bekkevold, D., & Hemmer-Hansen, J. (2021). Evolution at two time frames: Ancient structural variants involved in post-glacial divergence of the European plaice (*Pleuronectes platessa*). *Heredity*, *126*(4), Article 4. https://doi.org/10.1038/s41437-020-00389-3
Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM (arXiv:1303.3997). *arXiv*. http://arxiv.org/abs/1303.3997
Li, H. (2018). Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics*, *34*(18), 3094-3100. https://doi.org/10.1093/bioinformatics/bty191
Li, H., & Ralph, P. (2019). Local PCA shows how the effect of population structure differs along the genome. *Genetics*, *211*(1), 289-304. https://doi.org/10.1534/genetics.118.301747
Li, C., Olave, M., Hou, Y., Qin, G., Schneider, R. F., Gao, Z., Tu, X., Wang, X., Qi, F., Nater, A., Kautt, A. F., Wan, S., Zhang, Y., Liu, Y., Zhang, H., Zhang, B., Zhang, H., Qu, M., Liu, S., … Lin, Q. (2021). Genome sequences reveal global dispersal routes and suggest convergent genetic adaptations in seahorse evolution. *Nature Communications*, *12*(1). https://doi.org/10.1038/s41467-021-21379-x
Llaurens, V., Whibley, A., & Joron, M. (2017). Genetic architecture and balancing selection: The life and death of differentiated variants. *Molecular Ecology*, *26*(9), 2430-2448. https://doi.org/10.1111/mec.14051
Lohse, K., Clarke, M., Ritchie, M. G., & Etges, W. J. (2015). Genome-wide tests for introgression between cactophilic *Drosophila* implicate a role of inversions during speciation. *Evolution*, *69*(5), 1178-1190. https://doi.org/10.1111/evo.12650
Lomsadze, A., Burns, P. D., & Borodovsky, M. (2014). Integration of mapped RNA-Seq reads into automatic training of eukaryotic gene finding algorithm. *Nucleic Acids Research*, *42*(15), e119. https://doi.org/10.1093/nar/gku557
Lowry, D. B., & Willis, J. H. (2010). A widespread chromosomal inversion polymorphism contributes to a major life-history transition, local adaptation, and reproductive isolation. *PLOS Biology*, *8*(9), e1000500. https://doi.org/10.1371/journal.pbio.1000500
Lundberg, M., Mackintosh, A., Petri, A., & Bensch, S. (2023). Inversions maintain differences between migratory

phenotypes of a songbird. *Nature Communications*, *14*(1), Article 1. https://doi.org/10.1038/s41467-023-36167-y

Maggiolini, F. A. M., Sanders, A. D., Shew, C. J., Sulovari, A., Mao, Y., Puig, M., Catacchio, C. R., Dellino, M., Palmisano, D., Mercuri, L., Bitonto, M., Porubský, D., Cáceres, M., Eichler, E. E., Ventura, M., Dennis, M. Y., Korbel, J. O., & Antonacci, F. (2020). Single-cell strand sequencing of a macaque genome reveals multiple nested inversions and breakpoint reuse during primate evolution. *Genome Research*, *30*(11), 1680-1693. https://doi.org/10.1101/gr.265322.120

Manni, M., Berkeley, M. R., Seppey, M., & Zdobnov, E. M. (2021). BUSCO: Assessing genomic data quality and beyond. *Current Protocols*, *1*(12), e323. https://doi.org/10.1002/cpz1.323

Marion, S. B., & Noor, M. A. F. (2023). Interrogating the roles of mutation–selection balance, heterozygote advantage, and linked selection in maintaining recessive lethal variation in natural populations. *Annual Review of Animal Biosciences*, *11*(1), 77-91. https://doi.org/10.1146/annurev-animal-050422-092520

Martin, S. H., & Van Belleghem, S. M. (2017). Exploring evolutionary relationships across the genome using topology weighting. *Genetics*, *206*(1), 429-438. https://doi.org/10.1534/genetics.116.194720

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

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

Meier, J. I., Salazar, P. A., Kučka, M., Davies, R. W., Dréau, A., Aldás, I., Box Power, O., Nadeau, N. J., Bridle, J. R., Rolian, C., Barton, N. H., McMillan, W. O., Jiggins, C. D., & Chan, Y. F. (2021). Haplotype tagging reveals parallel formation of hybrid races in two butterfly species. *Proceedings of the National Academy of Sciences of the United States of America*, *118*(25), e2015005118. https://doi.org/10.1073/pnas.2015005118

Melters, D. P., Bradnam, K. R., Young, H. A., Telis, N., May, M. R., Ruby, J. G., Sebra, R., Peluso, P., Eid, J., Rank, D., Garcia, J. F., DeRisi, J. L., Smith, T., Tobias, C., Ross-Ibarra, J., Korf, I., & Chan, S. W. (2013). Comparative analysis of tandem repeats from hundreds of species reveals unique insights into centromere evolution. *Genome Biology*, *14*(1), R10. https://doi.org/10.1186/gb-2013-14-1-r10

Mérot, C., Llaurens, V., Normandeau, E., Bernatchez, L., & Wellenreuther, M. (2020a). Balancing selection via life-history trade-offs maintains an inversion polymorphism in a seaweed fly. *Nature Communications*, *11*(1), Article 1. https://doi.org/10.1038/s41467-020-14479-7

Mérot, C., Oomen, R. A., Tigano, A., & Wellenreuther, M. (2020b). A roadmap for understanding the evolutionary significance of structural genomic variation. *Trends in Ecology & Evolution*, *35*(7), 561-572. https://doi.org/10.1016/j.tree.2020.03.002

Morisse, P., Legeai, F., & Lemaitre, C. (2021a). LEVIATHAN: Efficient discovery of large structural variants by leveraging long-range information from Linked-Reads data. *bioRxiv* doi: 10.1101/2021.03.25.437002

Morisse, P., Lemaitre, C., & Legeai, F. (2021b). LRez: A C++ API and toolkit for analyzing and managing Linked-Reads data. *Bioinformatics Advances*, *1*(1), vbab022. https://doi.org/10.1093/bioadv/vbab022

Navarro, A., Betrán, E., Barbadilla, A., & Ruiz, A. (1997). Recombination and gene flux caused by gene conversion and crossing over in inversion heterokaryotypes. *Genetics*, *146*(2), 695-709. https://doi.org/10.1093/genetics/146.2.695

Navarro, A., & Barton, N. H. (2003). Accumulating postzygotic isolation genes in parapatry: A new twist on chromosomal speciation. *Evolution, 57*(3), 447-459. https://doi.org/10.1111/j.0014-3820.2003.tb01537.x

Nei, M., Kojima, K.-I., & Schaffer, H. E. (1967). Frequency changes of new inversions in populations under mutation-selection equilibria. *Genetics*, *57*(4), 741-750. https://doi.org/10.1093/genetics/57.4.741

Noor, M. A. F., Grams, K. L., Bertucci, L. A., & Reiland, J. (2001). Chromosomal inversions and the reproductive isolation of species. *Proceedings of the National Academy of Sciences*, *98*(21), 12084-12088. https://doi.org/10.1073/pnas.221274498

Ortiz-Barrientos, D., Engelstädter, J., & Rieseberg, L. H. (2016). Recombination rate evolution and the origin of species. *Trends in Ecology & Evolution*, *31*(3), 226-236. https://doi.org/10.1016/j.tree.2015.12.016

Pedersen, B. S., & Quinlan, A. R. (2017). cyvcf2 : Fast, flexible variant analysis with Python. *Bioinformatics*, *33*(12), 1867-1869. https://doi.org/10.1093/bioinformatics/btx057

Perez-Ruzafa, A., Marcos, C., Pérez-Ruzafa, I., & Pérez-Marcos, M. (2011). Coastal lagoons: “Transitional

ecosystems" between transitional and coastal waters. *Journal of Coastal Conservation*, *15*, 369-392. https://doi.org/10.1007/s11852-010-0095-2

Pettersson, M. E., Rochus, C. M., Han, F., Chen, J., Hill, J., Wallerman, O., Fan, G., Hong, X., Xu, Q., Zhang, H., Liu, S., Liu, X., Haggerty, L., Hunt, T., Martin, F. J., Flicek, P., Bunikis, I., Folkvord, A., & Andersson, L. (2019). A chromosome-level assembly of the Atlantic herring genome-detection of a supergene and other signals of selection. *Genome Research*, *29*(11), 1919-1928. https://doi.org/10.1101/gr.253435.119

Picard toolkit. (2019). In *Broad Institute, GitHub repository*. Broad Institute. https://broadinstitute.github.io/picard/

Qin, G., Luo, W., Tan, S., Zhang, B., Ma, S., & Lin, Q. (2019). Dimorphism of sex and gonad-development-related genes in male and female lined seahorse, *Hippocampus erectus*, based on transcriptome analyses. *Genomics*, *111*(3), 260-266. https://doi.org/10.1016/j.ygeno.2018.11.008

Rafajlović, M., Rambla, J., Feder, J. L., Navarro, A., & Faria, R. (2021). Inversions and genomic differentiation after secondary contact: When drift contributes to maintenance, not loss, of differentiation. *Evolution*, *75*(6), 1288-1303. https://doi.org/10.1111/evo.14223

Revell, L. J. (2012). Phytools: An R package for phylogenetic comparative biology (and other things). *Methods in Ecology and Evolution*, *3*(2), 217-223. https://doi.org/10.1111/j.2041-210X.2011.00169.x

Rhie, A., McCarthy, S. A., Fedrigo, O., Damas, J., Formenti, G., Koren, S., Uliano-Silva, M., Chow, W., Fungtammasan, A., Kim, J., Lee, C., Ko, B. J., Chaisson, M., Gedman, G. L., Cantin, L. J., Thibaud-Nissen, F., Haggerty, L., Bista, I., Smith, M., … Jarvis, E. D. (2021). Towards complete and error-free genome assemblies of all vertebrate species. *Nature*, *592*, 737–746. https://doi.org/10.1038/s41586-021-03451-0

Rieseberg, L. H. (2001). Chromosomal rearrangements and speciation. *Trends in Ecology & Evolution*, *16*(7), 351-358. https://doi.org/10.1016/s0169-5347(01)02187-5

Riquet, F., Liautard-Haag, C., Woodall, L., Bouza, C., Louisy, P., Hamer, B., Otero-Ferrer, F., Aublanc, P., Béduneau, V., Briard, O., Ayari, T. E., Hochscheid, S., Belkhir, K., Arnaud-Haond, S., Gagnaire, P.-A., & 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. https://doi.org/10.1111/evo.13696

Schaeffer, S. W., & Anderson, W. W. (2005). Mechanisms of genetic exchange within the chromosomal inversions of *Drosophila pseudoobscura*. *Genetics*, *171*(4), 1729-1739. https://doi.org/10.1534/genetics.105.041947

Shajii, A., Numanagić, I., & Berger, B. (2018). Latent variable model for aligning barcoded short-reads improves downstream analyses. *Research in computational molecular biology : Annual International Conference, proceedings. 10812*, 280-282.

Skoglund, P., Northoff, B. H., Shunkov, M. V., Derevianko, A. P., Pääbo, S., Krause, J., & Jakobsson, M. (2014). Separating endogenous ancient DNA from modern day contamination in a Siberian Neandertal. *Proceedings of the National Academy of Sciences*, *111*(6), 2229-2234. https://doi.org/10.1073/pnas.1318934111

Stanke, M., Diekhans, M., Baertsch, R., & Haussler, D. (2008). Using native and syntenically mapped cDNA alignments to improve de novo gene finding. *Bioinformatics (Oxford, England)*, *24*(5), 637-644. https://doi.org/10.1093/bioinformatics/btn013

Stiller, J., Short, G., Hamilton, H., Saarman, N., Longo, S., Wainwright, P., Rouse, G. W., & Simison, W. B. (2022). Phylogenomic analysis of Syngnathidae reveals novel relationships, origins of endemic diversity and variable diversification rates. *BMC Biology*, *20*(1), 75. https://doi.org/10.1186/s12915-022-01271-w

Thompson, M. J., & Jiggins, C. D. (2014). Supergenes and their role in evolution. *Heredity*, *113*(1). https://doi.org/10.1038/hdy.2014.20

Tigano, A., & Friesen, V. L. (2016). Genomics of local adaptation with gene flow. *Molecular Ecology*, *25*(10), 2144-2164. https://doi.org/10.1111/mec.13606

Tigano, A., Jacobs, A., Wilder, A. P., Nand, A., Zhan, Y., Dekker, J., & Therkildsen, N. O. (2021). Chromosome-level assembly of the Atlantic silverside genome reveals extreme levels of sequence diversity and structural genetic variation. *Genome Biology and Evolution*, *13*(6), evab098. https://doi.org/10.1093/gbe/evab098

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., … Rieseberg, L. H. (2020). Massive haplotypes underlie ecotypic differentiation in sunflowers. *Nature*, *584*, 602-607. https://doi.org/10.1038/s41586-020-2467-6

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

Protocols in Bioinformatics, *43*(1), 11.10.1-11.10.33. https://doi.org/10.1002/0471250953.bi1110s43
Wang, Y., & Leung, F. C. C. (2006). Long inverted repeats in eukaryotic genomes: Recombinogenic motifs determine genomic plasticity. *FEBS Letters*, *580*(5), 1277-1284. https://doi.org/10.1016/j.febslet.2006.01.045
Weisenfeld, N. I., Kumar, V., Shah, P., Church, D. M., & Jaffe, D. B. (2017). Direct determination of diploid genome sequences. *Genome Research*, *27*(5), 757-767. https://doi.org/10.1101/gr.214874.116
Wellenreuther, M., & Bernatchez, L. (2018). Eco-evolutionary genomics of chromosomal inversions. *Trends in Ecology & Evolution*, *33*(6), 427-440. https://doi.org/10.1016/j.tree.2018.04.002
Wellenreuther, M., Mérot, C., Berdan, E., & Bernatchez, L. (2019). Going beyond SNPs: The role of structural genomic variants in adaptive evolution and species diversification. *Molecular Ecology*, *28*(6), 1203-1209. https://doi.org/10.1111/mec.15066
Westram, A. M., Faria, R., Johannesson, K., Butlin, R., & Barton, N. (2022a). Inversions and parallel evolution. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *377*(1856). https://doi.org/10.1098/rstb.2021.0203
Westram, A. M., Stankowski, S., Surendranadh, P., & Barton, N. (2022b). What is reproductive isolation? *Journal of Evolutionary Biology*, *35*(9), 1143-1164. https://doi.org/10.1111/jeb.14005
Whittington, C. M., Griffith, O. W., Qi, W., Thompson, M. B., & Wilson, A. B. (2015). Seahorse brood pouch transcriptome reveals common genes associated with vertebrate pregnancy. *Molecular Biology and Evolution*, *32*(12), 3114-3131. https://doi.org/10.1093/molbev/msv177
Wohns, A., Wong, Y., Jeffery, B., Akbari, A., Mallick, S., Pinhasi, R., Patterson, N., Reich, D., Kelleher, J., & McVean, G. (2022). A unified genealogy of modern and ancient genomes. *Science*, 375. https://doi.org/10.1126/science.abi8264
Yang, Y.-Y., Lin, F.-J., & Chang, H. (2002). Comparison of recessive lethal accumulation in inversion-bearing and inversion-free chromosomes in *Drosophila*. *Zoological Studies*, *41*(3): 271-282.
Yeaman, S. (2013). Genomic rearrangements and the evolution of clusters of locally adaptive loci. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(19), E1743-1751. https://doi.org/10.1073/pnas.1219381110
Yeaman, S., & Whitlock, M. C. (2011). The genetic architecture of adaptation under migration-selection balance. *Evolution*, *65*(7), 1897-1911. https://doi.org/10.1111/j.1558-5646.2011.01269.x
Zhang, H., Liu, Y., Qin, G., & Lin, Q. (2022). Identification of neurohypophysial hormones and the role of VT in the parturition of pregnant seahorses (*Hippocampus erectus*). *Frontiers in Endocrinology*, *13*. https://doi.org/10.3389/fendo.2022.923234
Zhang, L., Reifová, R., Halenková, Z., & Gompert, Z. (2021). How important are structural variants for speciation? *Genes*, *12*(7), 1084. https://doi.org/10.3390/genes12071084
Zheng, X., Levine, D., Shen, J., Gogarten, S. M., Laurie, C., & Weir, B. S. (2012). A high-performance computing toolset for relatedness and principal component analysis of SNP data. *Bioinformatics (Oxford, England)*, *28*(24), 3326-3328. https://doi.org/10.1093/bioinformatics/bts606

Chapter III

Eco-geographic patterns of ecotypic structure in five species of marine fish

---

Context

After presenting detailed cases of ecotypic structure in two species (*Engraulis encrasicolus*, **Chapter I**; *Hippocampus guttulatus*, **Chapter II**), we here describe and compare general eco-geographic patterns and associated genomic differentiation landscapes across all five species studied in the current thesis. We include the big-scale sand smelt (*Atherina boyeri*), the grey wrasse (*Symphodus cinereus*), and the broadnosed pipefish (*Syngnathus typhle*), which all present less well characterised cases of ecotypic subdivision.

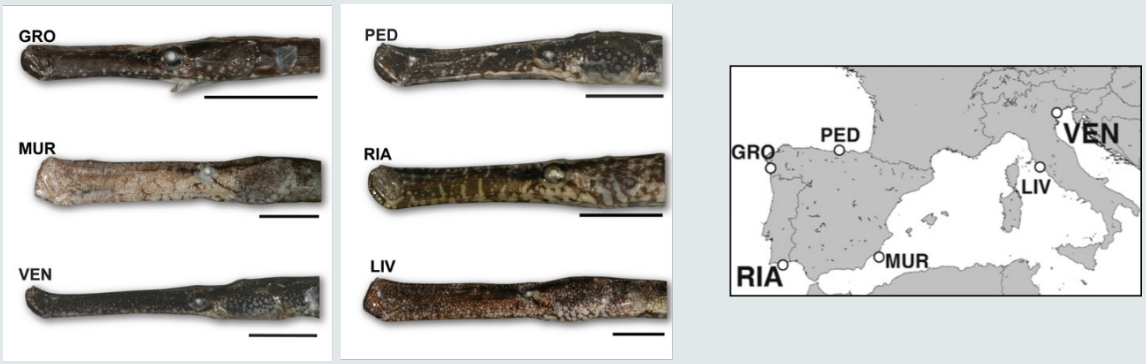


Fig. 1. Morphological variation observed in *S. typhle* (scale: 1 cm) (adapted from Wilson et al. 2020).

S. typhle presents significant morphological variations across its geographical range and habitats (**Fig. 1**), which were studied by Wilson et al. (2020). These authors showed that morphological variations were associated with dietary composition and that differences were maintained when individuals from different populations were reared in common garden experiments. These morphological differences therefore have a heritable component and apparently reflect dietary specialisation. However, to date, no study has attempted to link these morphological variations to genomic differentiation patterns across different types of habitats.

In ***A. boyeri*** coastal, lagoonal and estuarine habtats have been known to harbour a mosaic of isolated and semi-isolated populations, each with their own morphological and meristic characteristics that seem to be linked to environmental variations (Henderson et al., 1988; Kara & Quignard, 2019; Trabelsi et al., 2002). Several studies based on nuclear and mitochondrial markers (Boudinar et al., 2016; Francisco et al., 2011; Klossa-Kilia et al., 2002; Mauro et al., 2007; Milana et al., 2007; Trabelsi et al., 2002) have proposed the subdivision of *A. boyeri* into two or three species: a non-punctated lagoonal species (absence of spots along the lateral line), a non-punctated marine species, and a punctated marine species.

Ecotypic variation in ***S. cinereus*** has remained the most poorly characterised of all five species. Some studies have described morphological characteristics specific to populations inhabiting Thau and Berre lagoon in the south of France (Gourret, 1897; Quignard, 1966). Some authors have even recognised two subspecies, *S. cinereus staitii* and *S. cinereus cinereus* (e.g. Hanel et al., 2002), but this dichotomy is not used systematically and has not been clearly linked to habitat types. The status of lagoon populations of *S. cinereus* therefore remains to be reassessed (Kara & Quignard, 2019).

Chapter III aims to study the diversity and convergence of the evolutionary trajectories of ecotypes across different species. This study forms part of a larger project investigating genetic subdivision in 20 species of marine fish with broadly overlapping Atlantic-Mediterranean distributions (PhD thesis of Pierre Barry). While most species in the project showed geographic structure between the two ocean basins, the five species presented here did not follow this pattern and were suspected of showing ecotypic structure. In this chapter we describe their genetic structure across different locations and habitats in order to compare the geographical range over which ecotype subdivisions seem to exist. We further study genome-wide divergence patterns to investigate the genomic architecture underlying ecotype differentiation. For the 15 other species in the project, no cases of structural variation contributing to divergence were detected, but we wished to test this hypothesis more specifically for species showing ecotypic structure.

The methodological aspects of this chapter were identical to those described in Chapter I and II and were therefore not described here to avoid repetition. For more detailed information about samples and summary statistics related to bioinformatic steps, we provide an HTML report file that can be downloaded at the following link:

https://cloud.isem-evolution.fr/nextcloud/index.php/s/3TfAY4MtbfW5FSk

Abstract

Studying speciation in a comparative framework could help to reveal common factors contributing to divergence between incipient lineages. Although some work has been conducted on species showing geographic lineages that overlap in shared contact zones, few comparative studies have investigated ecotypes occurring in the same biogeographic region. Here, we compare eco-geographic patterns in five species of fish distributed along the marine-lagoon ecological gradient in the North-East Atlantic and Mediterranean Sea. For three of these species, *Atherina boyeri*, *Engraulis encrasicolus*, and *Hippocampus guttulatus*, marine and coastal ecotypes have already been described in genetic studies, whereas *Symphodus cinereus* and *Syngnathus typhle* ecotypes are only suspected to exist based on morphological differences. To address these knowledge gaps with genomic data, we generated reference genomes for all of the species and produced whole-genome resequencing data from samples collected in the same four locations (totalling 25 individuals per species). We showed that marine and coastal ecotypes exist in all species, providing the first description of ecological genetic structure in *Symphodus cinereus* and *Syngnathus typhle*. We further found evidence for more pronounced ecotypic structure across species in Mediterranean locations compared to the Atlantic. This could be due to the particular biogeographic context of the Mediterranean Sea and the availability of many sheltered lagoon habitats presenting characteristic species assemblages. Genetic barriers that initially evolved in other contexts, such as allopatric isolation, could have come to be associated with strong ecological contrasts in the Mediterranean, leading to the emergence of the ecotypic lineages observed today. In line with this, we found that ecotypic differentiation often involved structural variants which could represent ancient standing variation or islands of differentiation that have resisted rehomogenisation upon secondary contact.

Introduction

Ecotypes are often seen as an intermediate stage towards ecologically driven speciation, representing locally adapted forms that are partially reproductively isolated between different habitats (Hendry, 2009). Genetic studies investigating divergence between ecotypes have generally focused on a single species, or a group of closely related species showing ecotypic subdivision (Machado-Schiaffino et al., 2017; Magalhaes et al., 2012). An approach frequently implemented relies on examining replicate ecotype pairs of the same species to identify genomic regions consistently involved in divergence to unravel the evolutionary mechanisms involved in ecotype divergence (Gagnaire et al., 2013; Jones et al., 2012; Kess et al., 2018; Moan et al., 2016). In many cases, it has been demonstrated that the regions showing parallel divergence between ecotype pairs involve large structural variants (SVs), and that these variants are often older than the suspected age of divergence between ecotypes (Le Moan et al., 2021; Todesco et al., 2020). This illustrates how the process of ecotype formation might depend on the contribution of genetic barriers which originated in different contexts, which could be related to more ancient episodes of local adaptation, or alternatively to divergence in geographical isolation or introgression from a different species. Since ecotype divergence could be contingent on past evolutionary history, it is often uncertain whether replicate ecotype pairs of a given species could

truly be considered as independent evolutionary outcomes of ecologically-driven divergence. However, a comparative study of multiple unrelated species could be less prone to such bias, and reveal key information about the evolutionary trajectories of diverging ecotypes. Comparative approaches are generally seen as a powerful means for studying the complex and multifactorial process of speciation (e.g. Johannesson et al., 2020), but very few studies, if any, have compared ecotype pairs across different species in a shared biogeographic context.

The North-Eastern Atlantic and Mediterranean Sea have been subject to environmental and connectivity fluctuations due to glacial cycles during the Pleistocene (Patarnello et al., 2007). The effects of these fluctuations have been particularly pronounced at the Atlantic-Mediterranean transition zone, where variation in sea levels have cyclically impacted connectivity between the two ocean basins. This zone has been described as a phylogeographic break in many marine species, with genetically differentiated lineages distributed on either side of the Strait of Gibraltar (Patarnello et al., 2007). However, not all species in this region show geographic subdivision. Instead, some species show genetic associations with ecological boundaries between spatially heterogeneous habitats. This includes several species of marine fish which have been shown to display fine-scale ecotypic structure associated with the marine/lagoon ecological gradient, despite their high mobility and the perceived lack of physical barriers to movement in the sea (Boudinar et al., 2016; Moan et al., 2016; Riquet et al., 2019). These marine fish species thus offer interesting cases of parallel ecotype divergence associated with a similar ecological gradient. Moreover, teleost fishes are known to present a relatively well conserved genome architecture in terms of chromosome number and length (Almeida et al., 2017; Galvão et al., 2011), making them valuable models for multispecies comparative genomic studies.

Here, we present a comparative study of divergence associated with the marine/lagoon environmental gradient in five species of marine fish, with the objective of investigating ecotypic subdivision in a similar biogeographic context. Previous genetic studies have reported the existence of ecotypic forms in *Atherina boyeri* (Henderson et al., 1988; Kara & Quignard, 2019; Kartas & Trabelsi, 1990; Klossa-Kilia et al., 2002), *Engraulis encrasicolus* (Borsa, 2002; Moan et al., 2016; Montes et al., 2016) and *Hippocampus guttulatus* (Riquet et al., 2019). We also include two additional species for which ecotypes have not been confirmed but are suspected to exist based on reports of morphological variation possibly associated with habitat, namely *Syngnathus typhle* (Wilson et al., 2020) and *Symphodus cinereus* (Gourret, 1897; Quignard, 1966). We generated new reference genomes for each species and produced whole-genome resequencing data in a standardised design including 25 individuals per species, which were used to call variants genome-wide and characterise geographic and ecological components of genetic diversity. Samples were collected across marine and coastal/lagoon habitats (whenever possible) in the four same locations in each species (**Fig. 3A**): two locations in the Atlantic (*Ga*: Bay of Biscay; *Fa*: Southern Portugal) and two in the Mediterranean Sea (*Mu*: Costa Cálida; *Li*: Gulf of Lion). Our first aim was to test whether genetic structure had a component associated with habitat type across the different sampling locations in each species. We further wished to characterise the genomic architecture underlying differentiation between ecotypes, in an attempt to compare their diversity of evolutionary trajectories.

Results and Discussion

We obtained medium- to high-coverage resequencing data (ranging from 10X to ~50X) for most samples, with relatively few lower-quality (~5X) samples in *H. guttulatus* for which we accordingly applied different data filtering steps. We performed Principal Component Analysis (PCA) to investigate genome-wide patterns of population structure and found evidence for pronounced genetic structure in all five species (**Fig. 1**). We found evidence for a dominant geographic signal of differentiation between the Atlantic and Mediterranean in some cases (e.g. along PCA axis 1 in *S. cinereus* and *S. typhle*, **Fig. 1A**), but this was often less prominent than the component of differentiation associated with the marine/lagoon ecological gradient. We observed mild to strong ecotypic subdivision in all species, with genetic structure existing between samples from marine and coastal/lagoon habitats in the same location (mostly in *Li*). Although the ecotype subdivisions have already been described in some species, this is the first study confirming genetic differentiation between ecotypic forms in *S. cinereus* and *S. typhle*. We further confirmed strong structure existing between two different clusters marine samples and on cluster of lagoon-caught individuals in *A. boyeri*. For *E. encrasicolus*, PCA captured ecotypic structure, as well as a signal of admixture (in *Fa*) with a third genetic ancestry corresponding to a lineage off the African Atlantic coastline (**Fig. 1G**). The patterns observed for *H. guttulatus* (**Fig. 1E**) neither reflected pure geographical nor ecological structure, but rather, the presence of large structural variants (SVs) that are known to segregate in this species (Meyer et al., 2023 / Chapter II).

We observed disparities between eco-geographic patterns observed in the Atlantic and in the Mediterranean Sea. We found that ecotypic subdivision was generally more pronounced in the Mediterranean as compared to the Atlantic, with ecotypic structure specifically concentrated around locations in *Li*. Here, we identified the existence of ecotype pairs in all five species (**Fig. 2B**). Ecotypic structure seems to be even quantitatively stronger here for *E. encrasicolus*, consistent with previous studies that have reported less admixture between marine and coastal ecotypes in *Li* than in Atlantic locations (Le Moan et al., 2016; Chapter I). The northern coast of the Western Mediterranean Sea is characterised by the presence of many coastal lagoons with assemblages of fish species that are typical of lagoon environments (Perez-Ruzafa et al., 2011). Our other Mediterranean sampling location (*Mu*) presented a somewhat intermediate situation, with ecotypes being present in some species but not all. Mar Menor is the only large lagoon along the Costa Cálida coastline, and it is more subject to marine influences than many of the lagoons in the *Li* region. Consistent with mixed species assemblages observed here by Perez-Ruzafa et al. (2011), Mar Menor lagoon was inhabited by the lagoon ecotype of some species (*S. typhle*, *H. guttulatus*, *A. boyeri*) and the marine ecotype of one other species (*S. cinereus*). However, the *S. cinereus* samples were captured at the entrance of the lagoon, where habitat type was clearly under marine influence. As for the anchovy, it is possible that lagoon forms of *S. cinereus* may reside in the more protected areas of Mar Menor, but they could not be found during the sampling.

To explain the contrasting results observed between Atlantic and Mediterranean sampling locations, we might look at potential habitat differences that could exist between these regions. Coastal habitats in the sampled parts of the Atlantic are mainly associated with estuarine-like systems and inland sea waters (Ria Formosa, Hossegor lake, Arcachon basin) that are subject

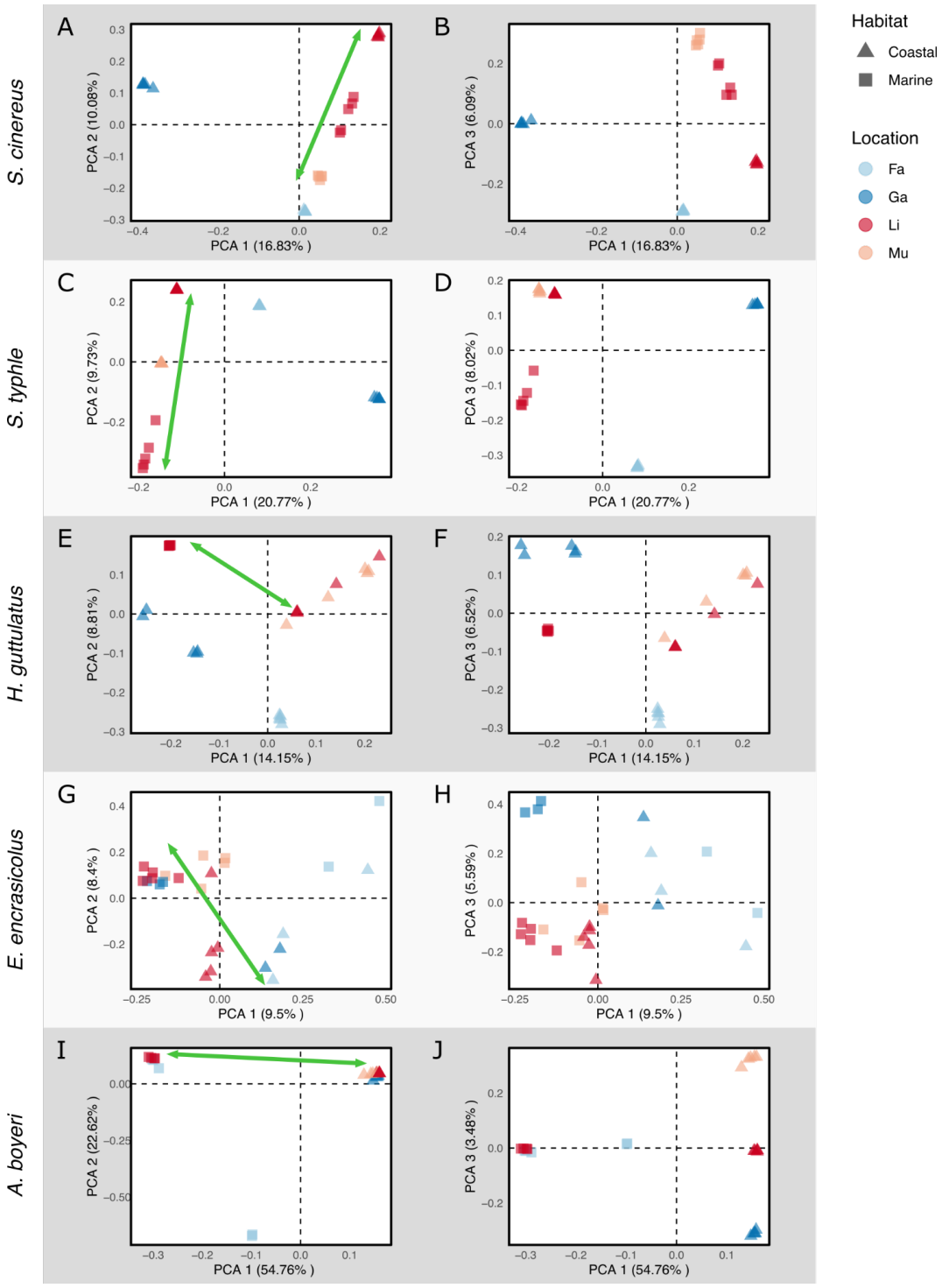


Fig. 1. Principal Component Analysis (PCA) performed in each of our five study species (rows). The first column (A, C, E, G, I) shows PCA axis 1 versus axis 2, whereas the second column (B, D, F, H, J) shows axis 1 versus axis 3. Green arrows illustrate the main axis of differentiation separating marine and lagoon samples. Symbol colours represent sampling location and shapes indicate habitat type.

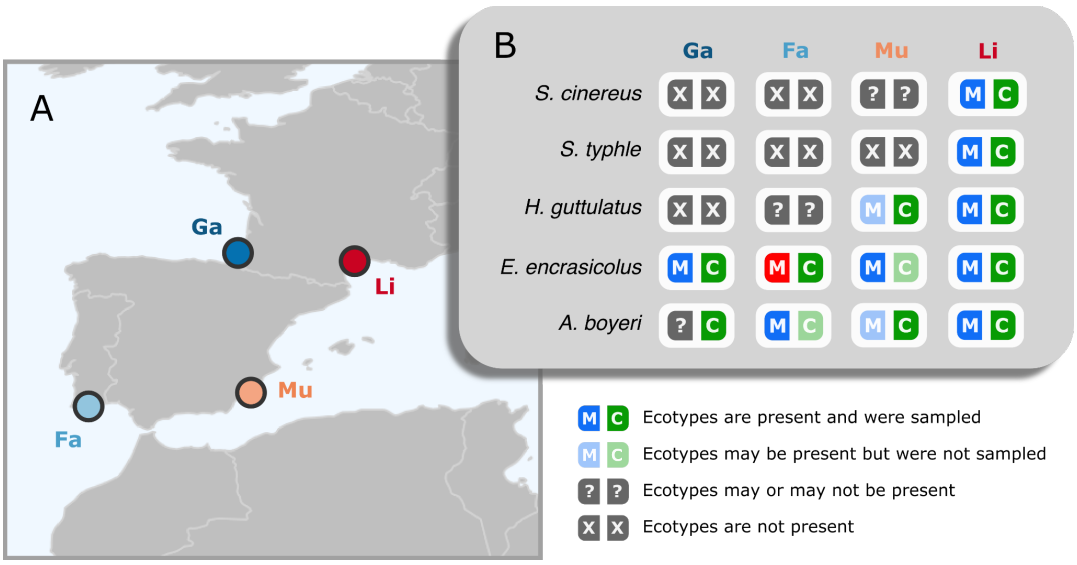


Fig. 2. Summary of ecotypic subdivision observed across five different species. The map (A) shows the four locations which were sampled and characterised in terms of their ecotypic structure (B). Different levels of information were available concerning the presence of the marine (*M*, left) and coastal (*C*, right) ecotype at each location. *E. encrasicolus* presented a particular case at *Fa* since two genetic forms were present, but the marine ecotype (red) showed mixed ancestry with a different genetic lineage and differed from marine individuals in other locations.

to the influence of the tides, whereas lagoon environments in the Mediterranean are much more enclosed areas that are connected to sea by relatively small channels. The Atlantic distributions of our study species are perhaps more limited to estuaries and associated habitats, since they are not well adapted to living in other unsheltered environments. If these fish are not present in open, marine habitats, but only in estuaries, this could explain the absence of ecotypic forms in Atlantic locations. As an alternative explanation for increased ecotypic structure in the Mediterranean, our selection of study species might have been biassed towards species presenting distribution ranges that are centred on the Mediterranean Sea. If our Atlantic sampling locations are on the northern edges of their ranges, lower densities and leading edge dynamics could have impacted evolutionary outcomes and impeded the formation, establishment or maintenance of ecotypes (Angert et al., 2020). Lastly, unsampled populations (e.g. in a given microhabitat) could present an issue if we concluded that ecotypes were absent, when in fact we simply did not sample them.

We characterised the genomic architecture underlying differentiation between ecotypes by reconstructing their genomic differentiation landscapes (**Fig. 3**). We compensated for high fragmentation and lack of contiguity in some of our reference genomes by anchoring our scaffolds to the chromosome-level genome assemblies of a related species (*Symphodus melops*, *Syngnathus acus*, *Hippocampus erectus*, *Coilia nasus* and *Menidia menidia*). In all comparisons, we observed heterogeneous divergence landscapes between marine and coastal ecotypes. Even though intra-chromosomal patterns should be interpreted with caution due to potential rearrangements with the species that was used for scaffold anchoring, we could still observe that

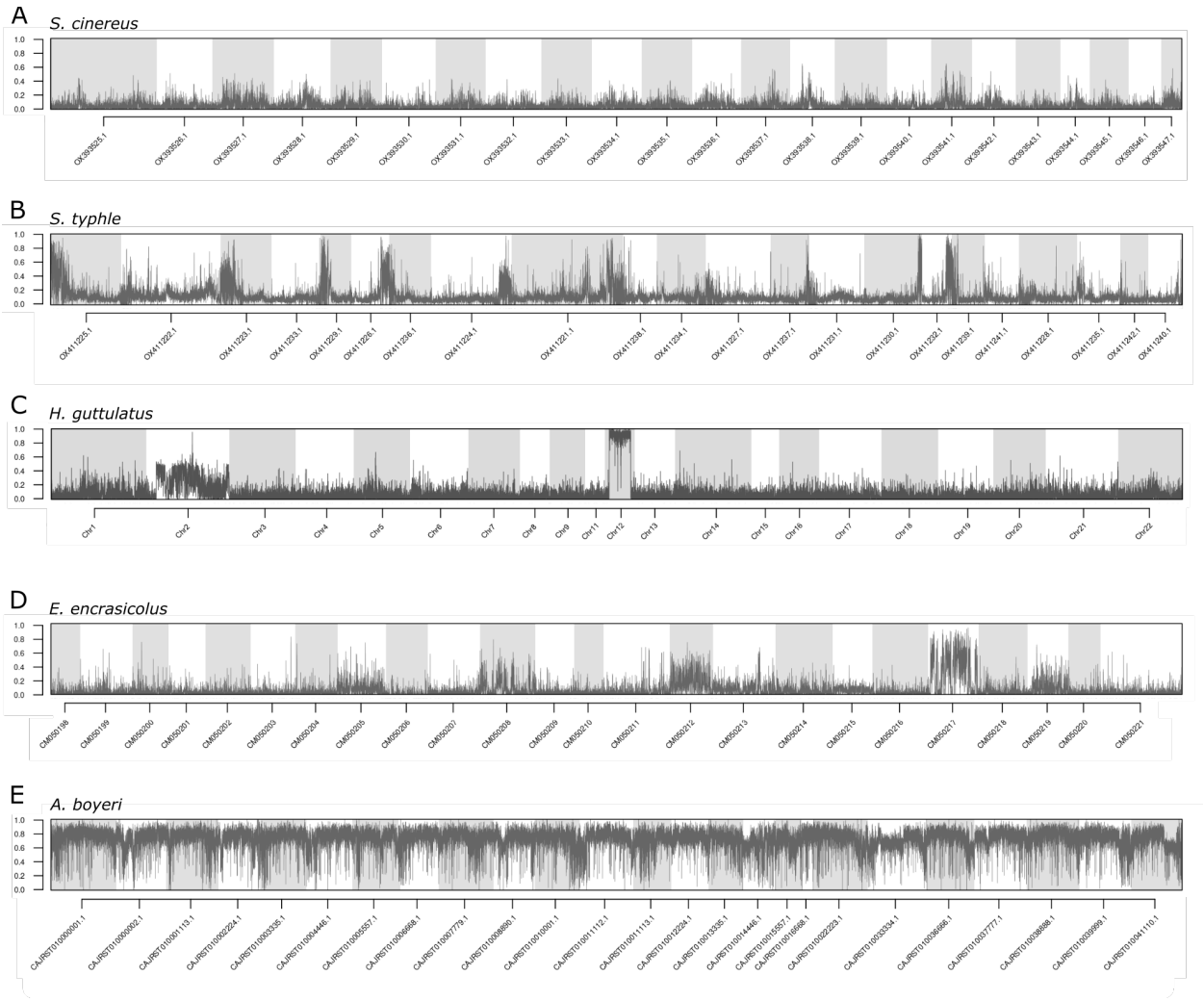


Fig. 3. Genomic landscapes of genetic differentiation (F_{ST}) in five species. For each comparison, F_{ST} was calculated between marine and lagoon samples in the Gulf of Lion (*Li*) in 5 kb non-overlapping sliding windows. The scaffolds of our reference genomes were anchored to closely related species with chromosome-level genome assemblies: *Symphodus melops* (A), *Syngnathus acus* (B), *Hippocampus erectus* (C), *Coilia nasus* (D) and *Menidia menidia* (E).

many of these landscapes showed sudden drops or increases in F_{ST} and patterns associated with strong linkage disequilibrium regions. Such plateaus of differentiation suggest the presence of SVs that differentiate ecotypes not only in seahorse and anchovy, but at least also in the pipefish. For *A. boyeri*, we observed high levels of differentiation genome-wide, but the elevated background differentiation was interspersed with some continuous segments of decreased divergence. These could represent shared SV polymorphisms segregating at different frequencies between ecotypes, which was also supported by differences in mapping statistics between ecotypes (varying number of reads pairs mapping to different scaffolds, see the HTML report file).

We further observed that the level of differentiation between ecotypes varied in strength, with some species showing low overall levels of genome-wide differentiation outside of putative SVs and genomic islands (e.g. *S. cinereus* or *H. guttulatus*). On the other hand, *A. boyeri* presented the highest levels of divergence and its ecotypes might in fact have attained sufficient reproductive isolation to be considered as separate species. The ecotypes observed in our five study species thus occupy different positions along the speciation continuum and we propose that this gradation is further reflected, to some extent, by the spatial distribution patterns of each ecotype. Species with higher levels of ecotypic divergence tended to show ecotypic subdivision in more locations, including Atlantic ones (**Fig. 2B**), such as was observed for *A. boyeri*. Highly divergent ecotypes in this species were present in at least three locations out of four, based on our sampling efforts as well as reports in the literature (Boudinar et al., 2016). We found similar results for *E. encrasicolus,* with the two ecotypes presenting wide distributions in different ocean basins. On the other hand, weakly differentiated ecotypes in *S.cinereus* were only present in one location (*Li*).

Observing replicate ecotype pairs across distant locations in a same species poses the question of their single versus repeated origin. In the latter scenario, *in situ* differentiation could have taken place independently in response to strong selection along an ecological gradient (Schluter, 2000). However, this is not truly compatible with what has been proposed for example in *E. encrasicolus* ecotypes, which are suggested to result from a single divergence event followed by secondary contact and spatial reassortment of ecotype pairs (Le Moan et al., 2016; **Chapter I**). Similarly, the genomic landscape of divergence observed between *S. cinereus* ecotypes (**Fig. 3A**) could reflect recent gene flow between divergent lineages. For this species, we find that high levels of differentiation persist in most centro-chromosomal regions, which are generally characterised by reduced recombination rates. This landscape closely mirrors the one observed between Atlantic and Mediterranean sea bass (*Dicentrarchus labrax*) lineages, where differentiation has been eroded in the peripheric chromosomal regions with increased recombination rates (Duranton et al., 2018). For *A. boyeri*, speciation appears to have progressed to late stage, making it hard to disentangle which evolutionary processes initially led to ecotype formation. However, marine and coastal forms in different ocean basins show extremely pronounced genetic parallelism (superposition of points in **Fig. 1I**), making a common origin highly likely. *H. guttulatus* ecotypes in the Mediterranean Sea are characterised by their karyotypes at two large inversions which represent ancient intraspecific polymorphisms. These inversions also differentiate geographic lineages in the Atlantic, highlighting the role of ancient SVs which originated in a different context and which did not necessarily evolve in response to selection pressures along the marine-lagoon gradient. Although less is known about genetic structure in *S. typhle*, the same SVs also partially seem to differentiate ecotypes in the Mediterranean Sea and geographic lineages in the Atlantic (*Fa* vs. *Ga*, not shown here).

The results of our comparative approach point to the possible importance of historical contingencies and the contribution of ancient SVs in the formation of ecotypes in the fish species. These old genetic variants that originated and diverged in different contexts (e.g. in different lineages, or even in separate species) could act as new barriers to gene flow between incipient ecotypic lineages (Van Belleghem et al., 2018; Le Moan et al., 2021). It has been proposed that

genetic incompatibilities that are environment-independent tend to coincide with exogenous barriers when they are trapped by local adaptation loci (Bierne et al., 2011). This means that lineages that accumulated intrinsic genetic incompatibilities (such as Dobzhansky-Muller incompatibilities) or any other type or barriers while diverging in geographical isolation, could subsequently be geographically redistributed to become associated with different habitats. Interestingly, the coupling hypothesis also predicts that, depending on the conditions of drift, selection and migration, genetic-by-habitat associations can form either at a fine or a large spatial scale. Within the confines of the northwestern Mediterranean Sea, the most pronounced ecological gradient between marine habitat and multitude of coastal lagoons could have allowed establishing a fine-scale association with habitat along the shoreline. This could explain why we currently tend to observe Mediterranean ecotypes, whereas the Atlantic would typically present genetic differentiation distributed along a large-scale latitudinal gradient. Furthermore, this scenario is compatible within a context of glacial cycles causing distributional range shifts and secondary contacts between previously isolated lineages. Future studies will have to investigate more in depth the demographic divergence history of the species studied here, to evaluate the role of past geographic divergence in ecotype formation.

References

Almeida, L. A. H., Nunes, L. A., Bitencourt, J. A., Molina, W. F., & Affonso, P. R. A. M. (2017). Chromosomal Evolution and Cytotaxonomy in Wrasses (Perciformes; Labridae). *The Journal of Heredity*, *108*(3), 239–253. https://doi.org/10.1093/jhered/esx003

Angert, A. L., Bontrager, M. G., & Ågren, J. (2020). What do we really know about adaptation at range edges? *Annual Review of Ecology, Evolution, and Systematics*, *51*(1), 341–361. https://doi.org/10.1146/annurev-ecolsys-012120-091002

Belleghem, S. M. V., Vangestel, C., Wolf, K. D., Corte, Z. D., Möst, M., Rastas, P., Meester, L. D., & Hendrickx, F. (2018). Evolution at two time frames: Polymorphisms from an ancient singular divergence event fuel contemporary parallel evolution. *PLOS Genetics*, *14*(11), e1007796. https://doi.org/10.1371/journal.pgen.1007796

Bierne, N., Welch, J., Loire, E., Bonhomme, F., & David, P. (2011). The coupling hypothesis: Why genome scans may fail to map local adaptation genes. *Molecular Ecology*, *20*(10), 2044–2072. https://doi.org/10.1111/j.1365-294X.2011.05080.x

Borsa, P. (2002). Allozyme, mitochondrial-DNA, and morphometric variability indicate cryptic species of anchovy (*Engraulis encrasicolus*). *Biological Journal of the Linnean Society*, *75*(2), 261–269. https://doi.org/10.1046/j.1095-8312.2002.00018.x

Boudinar, A. S., Chaoui, L., Quignard, J. P., Aurelle, D., & 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. https://doi.org/10.1016/j.ecss.2016.09.019

Duranton, M., Allal, F., Fraïsse, C., Bierne, N., Bonhomme, F., & Gagnaire, P.-A. (2018). The origin and remolding of genomic islands of differentiation in the European sea bass. *Nature Communications*, *9*(1), Article 1. https://doi.org/10.1038/s41467-018-04963-6

Francisco, S., Congiu, L., von der Heyden, S., & Almada, V. (2011). Multilocus phylogenetic analysis of the genus *Atherina* (Pisces: Atherinidae). *Molecular Phylogenetics and Evolution*, *61*, 71–78. https://doi.org/10.1016/j.ympev.2011.06.002

Gagnaire, P.-A., Pavey, S. A., Normandeau, E., & 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. https://doi.org/10.1111/evo.12075

Galvão, T. B., Bertollo, L. A. C., & Molina, W. F. (2011). Chromosomal complements of some Atlantic Blennioidei and Gobioidei species (Perciformes). *Comparative Cytogenetics*, *5*(4), 259–275. https://doi.org/10.3897/CompCytogen5i4.1834

Gourret, P. (1897). Les étangs saumâtres du Midi de la France et leurs pêcheries. *Annuaire Du Muséum d'Histoire Naturelle de Marseille*, *5*(1), 1–386.

Hanel, R., Westneat, M. W., & Sturmbauer, C. (2002). Phylogenetic relationships, evolution of broodcare behavior, and geographic speciation in the wrasse tribe Labrini. *Journal of Molecular Evolution*, *55*(6), 776–789. https://doi.org/10.1007/s00239-002-2373-6

Henderson, P. A., Holmes, R. H. A., & Bamber, R. N. (1988). Size-selective overwintering mortality in the sand smelt, *Atherina boyeri* Risso, and its role in population regulation. *Journal of Fish Biology*, *33*(2), 221–233. https://doi.org/10.1111/j.1095-8649.1988.tb05465.x

Hendry, A. P. (2009). Ecological speciation! Or the lack thereof?This Perspective is based on the author's J.C. Stevenson Memorial Lecture delivered at the Canadian Conference for Fisheries Research in Halifax, Nova Scotia, January 2008. *Canadian Journal of Fisheries and Aquatic Sciences*, *66*(8), 1383–1398. https://doi.org/10.1139/F09-074

Johannesson, K., Le Moan, A., Perini, S., & André, C. (2020). A Darwinian laboratory of multiple contact zones. *Trends in Ecology & Evolution*, S016953472030210X. https://doi.org/10.1016/j.tree.2020.07.015

Jones, F. C., Chan, Y. F., Schmutz, J., Grimwood, J., Brady, S. D., Southwick, A. M., Absher, D. M., Myers, R. M., Reimchen, T. E., Deagle, B. E., Schluter, D., & Kingsley, D. M. (2012). A genome-wide SNP genotyping array reveals patterns of global and repeated species-pair divergence in sticklebacks. *Current Biology: CB*, *22*(1), 83–90. https://doi.org/10.1016/j.cub.2011.11.045

Kara, H., & Quignard, J.-P. (2019). *Fishes in Lagoons and Estuaries in the Mediterranean 2: Sedentary Fish*.

https://doi.org/10.1002/9781119452768
Kartas, F., & Trabelsi, M. (1990). *Sur le polymorphisme de l'athérine,* Atherina boyeri *Risso, 1810 (Atherinidae) des eaux littorales tunisiennes*. https://doi.org/10.26028/CYBIUM/1990-144-002
Kess, T., Galindo, J., & Boulding, E. G. (2018). Genomic divergence between Spanish *Littorina saxatilis* ecotypes unravels limited admixture and extensive parallelism associated with population history. *Ecology and Evolution*, *8*(16), 8311–8327. https://doi.org/10.1002/ece3.4304
Klossa-Kilia, E., Prassa, M., Papasotiropoulos, V., Alahiotis, S., & Kilias, G. (2002). Mitochondrial DNA diversity in *Atherina boyeri* populations as determined by RFLP analysis of three mtDNA segments. *Heredity*, *89*(5), Article 5. https://doi.org/10.1038/sj.hdy.6800144
Le Moan, A., Bekkevold, D., & Hemmer-Hansen, J. (2021). Evolution at two time frames: Ancient structural variants involved in post-glacial divergence of the European plaice (*Pleuronectes platessa*). *Heredity*, *126*(4), Article 4. https://doi.org/10.1038/s41437-020-00389-3
Machado-Schiaffino, G., Kautt, A. F., Torres-Dowdall, J., Baumgarten, L., Henning, F., & Meyer, A. (2017). Incipient speciation driven by hypertrophied lips in Midas cichlid fishes? *Molecular Ecology*, *26*(8), 2348–2362. https://doi.org/10.1111/mec.14029
Magalhaes, I., Lundsgaard-Hansen, B., Mwaiko, S., & Seehausen, O. (2012). Evolutionary divergence in replicate pairs of ecotypes of Lake Victoria cichlid fish. *Evolutionary Ecology Research*, *14*, 381–401.
Mauro, A., Arculeo, M., Mazzola, A., & Parrinello, N. (2007). Are there any distinct genetic sub-populations of sand smelt, *Atherina boyeri* (Teleostei: Atherinidae) along Italian coasts? Evidence from allozyme analysis. *Folia Zoologica -Praha-*, *56*, 194–200.
Meyer, L., Barry, P., Riquet, F., Foote, A., Sarkissian, C. D., Cunha, R., Arbiol, C., Cerqueira, F., Desmarais, E., Bordes, A., Bierne, N., Guinand, B., & Gagnaire, P.-A. (2023). Divergence and gene flow history at two large chromosomal inversions involved in long-snouted seahorse ecotype formation (p. 2023.07.04.547634). *bioRxiv*. https://doi.org/10.1101/2023.07.04.547634
Milana, V., Sola, L., Congiu, L., & Rossi, anna rita. (2007). 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*, 1227. https://doi.org/10.1111/j.1095-8649.2008.01994.x
Moan, A. L., Gagnaire, P.-A., & Bonhomme, F. (2016). Parallel genetic divergence among coastal–marine ecotype pairs of European anchovy explained by differential introgression after secondary contact. *Molecular Ecology*, *25*(13), 3187–3202. https://doi.org/10.1111/mec.13627
Montes, I., Zarraonaindia, I., Iriondo, M., Grant, W. S., Manzano, C., Cotano, U., Conklin, D., Irigoien, X., & Estonba, A. (2016). Transcriptome analysis deciphers evolutionary mechanisms underlying genetic differentiation between coastal and offshore anchovy populations in the Bay of Biscay. *Marine Biology*, *163*(10), 205. https://doi.org/10.1007/s00227-016-2979-7
Patarnello, T., Volckaert, F. a. M. J., & Castilho, R. (2007). Pillars of Hercules: Is the Atlantic–Mediterranean transition a phylogeographical break? *Molecular Ecology*, *16*(21), 4426–4444. https://doi.org/10.1111/j.1365-294X.2007.03477.x
Perez-Ruzafa, A., Marcos, C., Pérez-Ruzafa, I., & Pérez-Marcos, M. (2011). Coastal lagoons: "Transitional ecosystems" between transitional and coastal waters. *Journal of Coastal Conservation*, *15*, 369–392. https://doi.org/10.1007/s11852-010-0095-2
Quignard, J.-P. (1966). *Recherches sur les Labridae (poissons téléostéens perciformes) des côtes européennes: Systématique et biologie*. Institut botanique.
Riquet, F., Liautard-Haag, C., Woodall, L., Bouza, C., Louisy, P., Hamer, B., Otero-Ferrer, F., Aublanc, P., Béduneau, V., Briard, O., Ayari, T. E., Hochscheid, S., Belkhir, K., Arnaud-Haond, S., Gagnaire, P.-A., & 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. https://doi.org/10.1111/evo.13696
Schluter, D. (2000). *The Ecology of Adaptive Radiation*. OUP Oxford.
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., … Rieseberg, L. H. (2020). Massive haplotypes underlie ecotypic differentiation in sunflowers. *Nature*, *584*(7822). https://doi.org/10.1038/s41586-020-2467-6
Trabelsi, M., Gilles, A., Fleury, C., Mâamouri, F., Quignard, J.-P., & Faure, É. (2002). *Atherina punctata* and *Atherina lagunae* (Pisces, Atherinidae), new species found in the Mediterranean Sea. 2. Molecular investigations of

three Atherinid species. *Comptes Rendus Biologies*, *325*(11), 1119–1128. https://doi.org/10.1016/S1631-0691(02)01529-9

DISCUSSION

The main objective of the current thesis was to study ecotypic structure in different species of marine fishes that are distributed along a similar ecological gradient in a shared biogeographic area. In so doing, we wished to shed light on the way in which lineages that segregate and sometimes co-exist in patchy environments may become partially reproductively isolated. The five species that were selected for study have roughly similar Atlantic-Mediterranean distributions, and occur in a wide range of heterogeneous habitats across what might be defined as a marine-lagoon ecological gradient. The first question we wished to address was whether genetic differences existed between populations of the same species in different habitats, pointing to more than phenotypic plasticity underlying habitat use and associated trait differentiation. In all of the species studied, we revealed evidence for a component of genetic structure that was at least partly associated with marine and lagoon habitats. We found that the geographical context also played a role in shaping these patterns, since marine and lagoon ecotypes were not present in all locations across the species' ranges, and geographic structure often showed interaction with ecotype structure. In all five species, ecotypic differentiation was either observed both in the Atlantic and in the Mediterranean but less pronounced in the Atlantic, or completely absent from the Atlantic. It is therefore possible that the particular biogeographic and ecological context of the Mediterranean Sea has provided more opportunity for ecotypic divergence to evolve or be maintained at the genomic level.

The presence of genetic structure and ecotypic differentiation in all of our study species leads to other questions that we wish to discuss in this thesis. The ecotypic lineages identified were often found to be in close geographical proximity, and showed some evidence of being connected by gene flow. If migration effectively takes place to a significant extent, one can ask (*i*) how genetic differentiation is maintained despite these homogenising effects. (*ii*) Is resistance to gene flow linked to the genomic architecture that underlies ecotypic differentiation? (*iii*) What are the origins of the variants underlying divergence? Do they represent new mutations, standing genetic variation that was sorted between the lineages, or foreign diversity that was introgressed from other lineages? (*iv*) Lastly, in which historical demographic context was divergence established? The answers to these questions may allow us to assess the diversity, or perhaps the alikeness, of the evolutionary trajectories of the different ecotypic lineages, and to evaluate their status of progression along the speciation continuum.

1. Genomic architecture of ecotype differentiation

1.1. SVs tend to underlie ecotypic differentiation

We found that genetic differentiation between ecotypes was generally not evenly dispersed across the genome, but rather concentrated in a number of divergent genomic regions. The proportion of the genome that was contained in these divergent regions varied substantially among species. In four of the five species, these regions occupied less than half of the genome. For example, highly differentiated regions covered 9% of the *H. guttulatus* genome, while this fraction reached 25% in *E. encrasicolus*. The opposite pattern was observed in *A. boyeri,* where a minority of regions showed locally reduced differentiation, contrasting with genome-wide

divergence. We implemented a wide range of population genetic and bioinformatic analyses (from analysing LD patterns to direct detection with linked reads) to show that most of these genomic islands of differentiation correspond to polymorphic SVs. The number of SVs differentiating ecotypes varied from only two in *H. guttulatus* to six in *E. encrasicolus*. Although SV detection needs to be confirmed with additional analyses in the three species added to **Chapter III**, well-identified genomic regions also seemed to contribute to the heterogeneous landscapes observed in these species. This was indicated by the presence of sharp blocks of increased differentiation between ecotypes of *S. typhle,* and the presence, among other genomic islands, of regions with contiguous signals of enhanced differentiation in *S. cinereus*. For *A. boyeri*, on the opposite spectrum, outlying regions in the landscape showed reduced divergence relative to the genome background. Here again, the presence of large SVs was indicated by continuous signals of lower divergence contrasting with the remainder of the genome, suggesting that shared structural variation between ecotypes could explain locally reduced divergence.

As for the size of these SVs, we could precisely determine through the identification of linked reads that mapped on either side of its breakpoints, that the chromosomal inversion present on chromosome 12 in *H. guttulatus* spanned a length of 8.2 Mb. We also estimated that the size of the other SV located on chromosome 2 in *H. guttulatus* could be on the order of 30 Mb. Obtaining accurate estimates for the sizes of other SVs identified in the four other species was not straightforward without the availability of linked reads mapped to chromosome-level reference assemblies. Based on reconstructed genomic landscapes, it might well be possible that the SVs found in *E. encrasicolus* show similarly large sizes, since they occupied a large fraction of the chromosomes on which they occurred. Interpreting intra-chromosomal patterns here requires caution, given the frequency of rearrangements between species and the possibility that this could have reshuffled the true order of *E. encrasicolus* scaffolds when mapped onto the chromosomes of the *C. nasus* assembly. However, even in the presence of many rearrangements between *E. encrasicolus* and *C. nasus*, local PCA windows showing a three-cluster pattern covered up to 50% of windows on the chromosome, suggesting that large SVs are involved in anchovy ecotype divergence.

All SVs that we analysed in detail (i.e. mostly in *E. encrasicolus* and *H. guttulatus*) showed patterns that were consistent with the presence of large chromosomal inversions. Evidence for an inverted segment could be confirmed for the 8.2 Mb haploblock in *H. guttulatus* due to the mapping locations and orientations of reads spanning breakpoints. One of the main results of the current thesis is thus that divergence between ecotypes tends to involve a concentrated genomic architecture, specifically involving large SVs like inversions. This type of architecture is expected to provide resistance to the rehomogenisation of allelic combinations in the face of gene flow, contributing to the maintenance of ecotypes despite migration between differentiated populations. Specifically, recombination is suppressed between alternate arrangements of an inversion, allowing divergence to accumulate and to be maintained in these regions. Consistent with this prediction, we observed a ~5-fold increase in divergence within the inversions compared to the genome background in *H. guttulatus*. Nonetheless, the expectation of complete recombination suppression in inversions may need to be nuanced, since our results, as well as an increasing number of studies, have evidenced rare recombination events taking place between inverted and

non-inverted haplotypes (e.g. Matschiner et al., 2022). Using an ARG-based approach to determine local ancestry along inversion regions, we detected introgressed segments of hundreds of kilobases in length in both *H. guttulatus* inversions. This was especially pronounced for the longer inversion (B2), which is consistent with the predicted positive relationship between SV length and gene flux (Berdan et al., 2023). Moreover, inversion type might also influence the likelihood of gene flux taking place between inverted alleles, since frequent heterozygotes at Type II inversions increase the chances of double crossovers occurring. It has further been put forward that the exchange of genetic material through double crossovers and gene conversion might be important for purging deleterious load in old inversions, affecting their dynamics and long-term maintenance (Faria et al., 2019a).

Our results thus join those of two decades of speciation genomic studies, showing that speciation between incipient ecotypic lineages often involves large chromosomal inversions, with dramatic to more subtle effects on phenotypic differences and RI between ecotypes (e.g. wing patterns in butterflies, Joron et al., 2006; life-history transition in monkeyflowers, Lowry & Willis, 2010; dune adaptation in sunflowers, Todesco et al., 2020; life-history trade-off in seaweed flies, Mérot et al., 2020; forest and prairie ecotypes in deer mice, Hager et al., 2022). Interestingly, studies of marine organisms have shown that chromosomal inversions are often involved in cases of ecotype divergence (e.g. threespine stickleback, Jones et al., 2012b; Atlantic cod, Berg et al., 2015; marine snails, Faria et al., 2019b, Le Moan et al., 2023; Atlantic herring, Martinez Barrio et al., 2016; European plaice, Le Moan et al., 2019). Collectively, these observations may lead us to conclude that RI involving inversions is a possibly frequent correlate of ecotypic speciation, especially in marine species (Johannesson et al., 2020). For the time being, however, we cannot completely dismiss that this could partly be due to a detection bias. Inversions contain many markers that are in perfect LD, disproportionately impacting the results of PCA and genomic divergence landscape analyses and potentially obscuring the signal of other smaller genomic islands that could be important for RI. Furthermore, many of these inversions were detected using indirect methods and these SVs might represent other types of rearrangements mistakenly classified as inversions. For example, a chromosomal fusion might be associated with similar tight clusters in PCA and locally increased levels of differentiation. If it is indeed the case that inversions specifically are the main culprits behind ecotypic divergence, we could ask why that is. This could be linked to the fact that an inversion only involves the reordering of chromosomal content, and therefore does not include the removal or insertion of genes, possibly limiting the negative effects of the rearrangement. Furthermore, an inversion event only involves one chromosome, perhaps increasing its chances of taking place through ectopic recombination (as compared to a fusion which needs to assemble two chromosomes).

1.2. Why are karyotypes at SVs associated with habitat differences?

If inversions play an important role in ecotypic differentiation, which processes led to the establishment of frequency differences between habitats in the first place? Are these the same as the mechanisms that maintain the association with the environment today? These questions are related to the order of events in which we expect SVs to contribute to divergence and speciation, and whether local adaptation was key in the evolution of reproductive isolation or whether it evolved as a by-product.

If marine and lagoon ecotypes are currently exchanging genes through migration and recombination, while frequency differences at SVs continue to be maintained, we could argue that this calls for a form of local selection. For example, *E. encrasicolus* ecotypes readily move between habitats and hybridise, and yet the haplotypic combinations at SVs that characterise these two forms remain in association with habitat type. The mechanisms responsible for the maintenance of such between-population divergence correspond to what has been described for Type I inversion polymorphisms (Faria et al., 2019a). Type I inversions showing pronounced frequency differences between habitats could be maintained by extrinsic selection acting on locally adapted alleles that are less fit in the other environment. In *E. encrasicolus*, the marine and coastal ecotypes show significant morphological differences (e.g. body size and colouration, eye size) that could reflect differential adaptations to these environments (e.g. growth-reproduction trade-offs, predator avoidance). For the inversions separating *H. guttulatus* ecotypes, local adaptation is not self-evident, since no clear morphological differences between the two ecotypes have been observed. Furthermore, functional enrichment in the B2 and B12 inversions did not show a direct link with the external environment, but instead showed associations with two molecular pathways that may interact in reproductive functions. This rather points to intrinsic selection possibly acting on incompatibilities between the inversions and the rest of the genomic environment and/or between the inversions themselves. However, we wish to stress that large SVs contain such a large amount of loci (B2 and B12 contain ~400 genes each) that some of them could be under different types of selection, including (but not limited to) local selection driving association with habitat (Faria et al., 2019a). Therefore, the inversions detected in *H. guttulatus* and in *E. encrasicolus* could be expected to contain both DMIs and locally adapted alleles.

If we wish to address how SV frequency differences between habitats were established in the first place, valuable information is to be gleaned from characterising the origin and the age of these rearrangements. Current methods for estimating the age of SVs are not fully reliable, but obtaining more accurate estimates could allow us to date their appearance relative to the formation of the ecotypes. For example, in a simple scenario, an inversion could have appeared after the ecotypic lineages already existed (i.e. after their split in a population tree). If the chromosomal inversion took place in the lagoon lineage, it could have captured locally adapted alleles already segregating in the population upon its initial rearrangement. Due to its selective advantage it could spread in the population, reaching a high frequency and perhaps even spreading to neighbouring lagoons. It might not spread in marine populations, either because gene flow is rare, or because it is actively counterselected in this environment, leading to substantial frequency differences between habitats (Type I polymorphism). This could be because the alleles contained in the inverted arrangement are better adapted to lagoon environments, or because of other intrinsic incompatibilities. In the case where the inversion appeared after ecotype formation, the two lineages were already diverging when the rearrangement took place. We should thus note that the inversion is likely to have captured co-adapted alleles that interact with other loci in the genome. The probability of this would depend on the level of divergence between ecotypes, as the inverted and ancestral arrangement are initially no more divergent than the mean branch length between individuals in the two populations (since the inversion event sampled a single

haplotype). The time since the split between ecotypes would also affect the probability of the inversion capturing recessive deleterious mutations, since a newly formed lagoon lineage could be expected to present an increased number of private rare alleles.

The scenario presented above illustrates how a new inversion might develop and maintain an association with habitat and reinforce reproductive isolation between ecotypes. However, it assumes that the inversion is younger than the ecotypes, with its appearance post-dating the initial divergence event between ecotypic lineages. In reality, however, inversions underlying ecotypic divergence in many systems have been reported to be old and highly divergent (see Johannesson et al. 2020 for a review in marine species). Furthermore, the dynamics of an inversion appearing shortly after the ecotype split could be potentially impacted by mutation load. Inversions carrying many deleterious mutations are either unlikely to establish, or are more prone to forming Type II polymorphisms. The B2 inversion in *H. guttulatus* could present such an example if it first appeared in the (already existent) lagoon lineage. If SVs associated with habitat only appeared after the split between ecotypes, we might even be led to question their allegedly important role in speciation, since other processes led to the initial lineage split.

2. The importance of historical contingencies for ecotypic speciation

"All geographic races are also ecological races,
and all ecological races are also geographical races." (Mayr, 1947)

Speciation between incipient ecotypes is contingent on their past evolutionary history(ies) of demographic fluctuations. Such historical contingencies encompass different forms of demographic events which took place in the past and which did not directly depend upon deterministic genetic processes. This includes demographic fluctuations within populations that could affect the balance between genetic drift, migration and selection, impacting the efficacy of positive selection (i.e. local adaptation and intra-genomic co-adaptation) or the purging of deleterious mutations. Such processes could also have shaped the subdivision between different genetic lineages, causing shifts in their distribution ranges and modulating opportunities for genetic interaction and the exchange of genes between formerly geographically isolated lineages.

Contrary to common perceptions, historical contingencies may sometimes prove to be more important for the speciation of ecotypes than direct ecological adaptation itself (e.g. Bierne et al., 2011). A first step in this direction could be to consider selection acting on existing standing variation that contains locally favoured alleles (e.g. that evolved within the ancestral population or that was introgressed from a divergent lineage). Here, adaptation would have not been possible in the same amount of time if these "prepackaged" variants were not already segregating in the ancestral lineage. What is more, SVs could represent particularly important sources of standing variation for speciating lineages, due to the preservation of ancient variation in non-recombining

haplotypes. For example, an inversion might have captured multiple loci that were already segregating in an ancestral population and which confer local adaptation in a given habitat. This could lead to the differential establishment of alternate arrangements in different environments, giving rise to an SV-habitat association in ecotypic lineages. This might correspond to the situation described between marine and freshwater ecotypes of threespine sticklebacks (Jones et al., 2012a). The transporter hypothesis proposes that haplotypes that provide a selective advantage in freshwater habitats circulate in the marine population at low frequencies (Schluter & Conte, 2009). Rapid adaptation to new freshwater habitats is thus made possible through strong selection on standing variation and does not rely on the accumulation of new mutations. A similar scenario involving repeated selection on standing variation has been proposed for *Littorina saxatilis* ecotypes, although some studies (e.g. Grahame et al., 2006) have argued that certain markers show patterns that are consistent with secondary contact between lineages that diverged in allopatry. In line with these findings, it has moreover been proposed that fast parallel ecological divergence may result from evolution at two time frames: divergence could have taken place in the past, while repeated selection on the divergent variants could explain their rapid reassortment and association with habitat on an ecological timescale (Van Belleghem et al., 2018; Le Moan et al., 2021).

In **Chapter II** we studied the long-term maintenance of the B12 polymorphism in *H. guttulatus*. This inversion represents an ancient intraspecific polymorphism that has been segregating in the species for hundreds of thousands of generations, and plays an important role in differentiating ecotypes in the Mediterranean Sea. However, we did not specifically address which processes originally led to the differential fixation of B12 between habitats. Due to its age, this inversion could have been present as standing variation in the ancestral population before ecotype formation. If it already showed underdominance due to meiotic disruption or DMIs, it could have been sorted differentially into incipient lineages at the following speciation event. The lineage that was the precursor of the lagoon ecotype could have fixed the A haplotype and the proto-marine lineage could have fixed the B haplotype, as this is the association that we observe today. If these lineages were already distributed differently across habitats, the direction of this association may indeed have involved extrinsic selection on genes within the inversion. What we wish to underline, however, is that the fixation of A in lagoons and B in the sea could alternatively have been due to chance or historical contingencies. If these haplotypes were sorted blindly and differential fixation took place to resolve intragenomic conflicts (i.e. intrinsic incompatibilities) and not because of local adaptation, this could explain why B12 is not associated with habitat in the Atlantic. Here, the sorting of incompatible haplotypes would then have taken place geographically between a northern and southern lineage. This highlights that the association between habitat and B12 did not necessarily establish due to repeated selection on local adaptation alleles, although this does not preclude that locally favoured alleles may have accumulated subsequently.

As compared to standing variation that evolved in the ancestral population of incipient lineages, the genetic diversity fueling divergence might alternatively have been introduced through introgression with another species. Some studies have reported that SVs underlying reproductive isolation had an interspecific origin (e.g. Jay et al., 2018), and this scenario differs in various ways from what was described above for intraspecific SVs. Chiefly, an intraspecific SV does not initially

introduce new variation - it only causes rearrangement of existing variation and establishes LD. By contrast, when an introgressed haplotype is introduced into a population, it can already be highly divergent. This is the case even if the mutational event leading to rearrangement took place relatively recently in the donor species, since the alleles it contains have had time to diverge in separate lineages for a certain amount of time. These genetic differences could potentially expose an introgressed SV to strong selection from the very beginning. Introgressed haplotypes could thus spread throughout the entire species, or alternatively could only establish in isolated populations. The other big difference for interspecific SVs is that an introgressed haplotype may be introduced as multiple copies, whereas an intraspecific SV starts off as a single copy. These contrasting properties may prove to be more than anecdotal for the evolutionary dynamics of SVs, and may be determinant for the types of SVs that succeed in being established.

To better understand the processes that accompanied the introgression of a divergent SV, we could attempt to characterise the conditions in which introgressive hybridisation took place. One aspect that could be important to consider for the dynamics of introgressed SVs, is the number of haplotype copies that are introduced through interspecific matings. In a scenario of extensive hybridisation taking place through secondary contact across a wide hybrid zone, we could expect that an SV be introduced as many copies. What is more, if the lineages are highly divergent, we could expect the collinearity of their genomes to differ in several places, perhaps leading to the simultaneous introduction of multiple SVs. Even if these haplotypes do not confer a great selective advantage, they could invade the population through swamping if they are numerous, or due to the resolution of multilocus conflicts involving DMIs (Schumer et al., 2015). Many lines of research have proposed that secondary contact between lineages can facilitate subsequent diversification (Campbell et al., 2018). This could be due to the introduction of divergent variants that are already in LD, representing two important ingredients for speciation. If past secondary contact led to the remixing of two ancient lineages (lineages A and B), these divergent ancestries might subsequently be apportioned differentially into two new lineages (1 and 2) due to incompatibilities between certain combinations of alleles, or due to other external factors (e.g. a new physical barrier to gene flow and the heterogeneous spatial distribution of A and B haplotypes). Incipient lineages 1 and 2 might give rise to ecotypes due to coupling with environmental gradients or due to extrinsic selection on A-derived or B-derived alleles. Alternatively, secondary contact could be observed in the present-day between ecotypes that are directly descended from ancient lineages that diverged in allopatry and which were not associated with habitat in the past.

On the other hand, if hybridisation between lineages was limited to a small part of the distribution range or if interspecific matings were rare, divergent haplotypes might only have been introduced as a few copies. If an introgressed SV managed to spread in a large population despite its rarity, we could suppose that it likely causes large phenotypic effects or a large selective advantage (e.g. supergene *P* in *H. numata*). In this sense it would seem unlikely that multiple SVs are introgressed and all retained in the recipient population due to each having a selective advantage. One explanation for such massive introgression of SVs could be that they present significant positive epistatic effects between them, increasing inter-chromosomal LD and their chances of establishing together. If multiple SVs or even one large SV containing alleles that promote reproductive isolation was introgressed into a population, the divergent haplotypes may have

contributed to the initiation of population subdivision and the formation of ecotypes. Alternatively, introgression could have only happened in one of the ecotypes if they were already diverging, further reinforcing reproductive isolation barriers.

Our results suggested that divergence patterns in *E. encrasicolus* could have resulted from introgression with a divergent lineage taking place at different time frames. On the one hand, we observe sharing of haplotypes at SVs between the coastal ecotype and the Southern Atlantic lineage, suggesting increased exchange between these two lineages as compared with the marine European ecotype. However, we also observe present-day gene flow between the Southern lineage and European populations, which mainly takes place through introgression with marine anchovies. This could suggest that the introgression of SVs from the Southern lineage into the coastal lineage took place further in the past and does not result from contemporary admixture in the Atlantic-Mediterranean transition zone. Present-day admixture probably involves extensive hybridisation, as can be observed by the abundance of three-way admixed individuals in the contact zone. It is less clear which geographic and demographic conditions accompanied the previous episode of contact resulting in the exchange of divergent SVs - for example, did the ecotypic lineages already exist at this time? In which geographic zone did contact take place (e.g. further South than present distribution ranges?), and did it involve many or few interspecific matings? To be able to answer these questions, further study is required (e.g. using demographic models) to elucidate how repeated cycles of isolation and contact between anchovies in the northern and southern hemispheres have shaped divergence patterns in this species complex.

3. Ecotypes as cases of incomplete speciation?

In the current thesis we have presented different cases of diverging ecotypes, but the question remains as to whether these will eventually complete speciation and become sister species occupying different ecological niches. It is not clear how often ecotypes become separate species, or whether they rather have a tendency to repeatedly fail to achieve complete reproductive isolation, leading to rehomogenisation of the accumulated genetic differentiation. Rehomogenisation is all the more likely if ecotypes occur in geographical proximity or if they present patchy distributions, with populations connected by high levels of gene flow. We could argue that, if the role of historical contingencies was preponderant in causing divergence between ecotypes in the first place (compared to a lesser role of direct ecological adaptation), any future change in the level of population connectivity (e.g. due to climatic changes) could potentially throw off the migration-selection balance maintaining differentiation. The outcomes of these processes could be expected to depend on many factors, such as the major evolutionary forces that led to the accumulation of reproductive isolation and the genomic architecture underlying divergence. For instance, concentrated architectures involving SVs could indeed present an immediate solution for reducing gene flow and maintaining allelic combinations, but do they lead to speciation in the long term? In what follows we consider these various points and discuss what it would take for ecotypes to become species.

For speciation to progress towards complete reproductive isolation, the level of gene flow taking place between ecotypes would need to continue to diminish. The first way in which this could be accomplished is if mating between the incipient lineages is disfavoured. Prezygotic barriers to gene flow such as sexual selection could play a role in species displaying mate choice (*S. cinereus*, *S. typhle* and *H. guttulatus*), but could not be expected to contribute as much in species that are reported to be broadcast spawners (*E. encrasicolus* and *A. boyeri*). The second filter could be a postzygotic barrier, where genetic differentiation is still maintained at barrier loci even if lineages hybridise. For LD between genetic differences to be maintained in the face of gene flow, it either requires strong selection or the suppression of recombination. SVs maintain LD between alleles contained in the same arrangement, which presents a useful property for the progression towards speciation. This is supported by theoretical models and empirical studies at the microevolutionary level, but the role of SVs in the completion of the speciation process is less evident (Lucek et al., 2023). It could be the case that the simple presence of SVs is not sufficient for speciation, and that the completion of speciation depends on their type and number. For example, Type II SV polymorphisms show intermediate frequencies in populations and are not expected to contain strong DMIs, which is an important property for the evolution of reproductive isolation. Type I SVs, on the other hand, are expected to show underdominance and to favour divergence between populations. Yet, here we note that individual SVs which are highly underdominant have a reduced probability of establishing, making them rare and unlikely to contribute to all cases of ecotype speciation (Faria & Navarro, 2010; Lucek et al., 2023; Rieseberg, 2001).

The number of SVs differentiating ecotypes could impact the likelihood of moving towards the completion of speciation. A single inversion without the presence of any other additional reproductive isolation loci might not suffice for ensuring reproductive isolation, since most of the genome experiences free gene flow. In *H. guttulatus*, one Type I inversion (B12) differentiates Mediterranean ecotypes, while a second inversion (B2) is a private polymorphism in the lagoon ecotype. These inversions have been segregating in the species for an extended period of time, and yet we only see weak levels of background differentiation outside of these regions, suggesting that they have not served as efficient barriers for the progression towards speciation. Speciation typically requires the presence of multiple barriers that are spread across multiple regions of the genome for strong reproductive isolation to evolve (Nosil et al., 2021). If multiple SVs acted as barriers, could LD and coupling between them be sufficient for speciation? Such SVs would have to be Type I polymorphisms, since the presence of many Type II SVs would result in segregation load and would not promote speciation. Type I polymorphisms on the other hand could accumulate and become associated amongst them to strengthen reproductive isolation between lineages that are characterised by different haplotypic combinations. This is what we observe for divergence between *E. encrasicolus* ecotypes and it has similarly been reported for the 12 inversions that differentiate ecotypes in *Littorina fabalis* (Moan et al., 2023). Alternatively, speciation could also proceed if a small number of SVs couple with other reproductive isolation loci that are dispersed throughout the genome and serve as multiple barriers to gene flow. Such cases of polygenic ecotype speciation are less commonly described, but this could be because large SVs tend to obscure the signal of smaller effect loci (we do not see the forest for the trees).

The process of ecological speciation can be viewed as evolution occurring at two different time frames, with current selection taking place on divergent variants that evolved at some time in the past (Van Belleghem et al., 2018). This view includes two components that are essential for speciation: the evolution of genetic divergence and the buildup of LD (Westram et al., 2022). We have models and expectations for the gradual accumulation of genetic divergence, but which dynamics control the establishment of LD between divergent alleles? In the case of primary sympatric divergence, even if divergent variants were present as standing variation or were introgressed from a different species, it could still take a significant amount of time for these “bricks” to be progressively assembled into haplotypes, and eventually, divergent genomes. What is more, the LD that has accumulated over time might be eroded in a context of fluctuating conditions, repeatedly preventing speciation from proceeding to completion. However, in the case of secondary contact between previously isolated lineages, the multiple genetic differences that cause reproductive isolation are already in phase (Barton & de Cara, 2009; Barton & Hewitt, 1985). These have only to be redistributed across space for the emergence of new lineages or ecotypes, once again highlighting the importance of allopatric divergence for speciation.

4. References

Barton, N. H., & de Cara, M. A. R. (2009). The evolution of strong reproductive isolation. *Evolution*, *63*(5), 1171–1190.

Barton, N. H., & Hewitt, G. M. (1985). Analysis of hybrid zones. *Annual Review of Ecology and Systematics*, *16*(1), 113–148. https://doi.org/10.1146/annurev.es.16.110185.000553

Belleghem, S. M. V., Vangestel, C., Wolf, K. D., Corte, Z. D., Möst, M., Rastas, P., Meester, L. D., & Hendrickx, F. (2018). Evolution at two time frames: Polymorphisms from an ancient singular divergence event fuel contemporary parallel evolution. *PLOS Genetics*, *14*(11), e1007796. https://doi.org/10.1371/journal.pgen.1007796

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

Bierne, N., Welch, J., Loire, E., Bonhomme, F., & David, P. (2011). The coupling hypothesis: Why genome scans may fail to map local adaptation genes. *Molecular Ecology*, *20*(10), 2044–2072. https://doi.org/10.1111/j.1365-294X.2011.05080.x

Campbell, C. R., Poelstra, J. W., & Yoder, A. D. (2018). What is speciation genomics? The roles of ecology, gene flow, and genomic architecture in the formation of species. *Biological Journal of the Linnean Society*, *124*(4), 561–583. https://doi.org/10.1093/biolinnean/bly063

Faria, R., & Navarro, A. (2010). Chromosomal speciation revisited: Rearranging theory with pieces of evidence. *Trends in Ecology & Evolution*, *25*(11), 660–669. https://doi.org/10.1016/j.tree.2010.07.008

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., & Butlin, R. K. (2019). Multiple chromosomal rearrangements in a hybrid zone between *Littorina saxatilis* ecotypes. *Molecular Ecology*, *28*(6), 1375–1393. https://doi.org/10.1111/mec.14972

Grahame, J. W., Wilding, C. S., & Butlin, R. K. (2006). Adaptation to a steep environmental gradient and an associated barrier to gene exchange in *Littorina saxatilis*. *Evolution; International Journal of Organic Evolution*, *60*(2), 268–278.

Hager, E. R., Harringmeyer, O. S., Wooldridge, T. B., Theingi, S., Gable, J. T., McFadden, S., Neugeboren, B., Turner, K. M., Jensen, J. D., & Hoekstra, H. E. (2022). A chromosomal inversion contributes to divergence in multiple traits between deer mouse ecotypes. *Science*, *377*(6604), 399–405. https://doi.org/10.1126/science.abg0718

Jay, P., Whibley, A., Frézal, L., Rodríguez de Cara, M. Á., Nowell, R. W., Mallet, J., Dasmahapatra, K. K., & Joron, M. (2018). Supergene evolution triggered by the introgression of a chromosomal inversion. *Current Biology*, *28*(11), 1839-1845.e3. https://doi.org/10.1016/j.cub.2018.04.072

Johannesson, K., Le Moan, A., Perini, S., & André, C. (2020). A Darwinian laboratory of multiple contact zones. *Trends in Ecology & Evolution*, S016953472030210X. https://doi.org/10.1016/j.tree.2020.07.015

Jones, F. C., Chan, Y. F., Schmutz, J., Grimwood, J., Brady, S. D., Southwick, A. M., Absher, D. M., Myers, R. M., Reimchen, T. E., Deagle, B. E., Schluter, D., & Kingsley, D. M. (2012a). A genome-wide SNP genotyping array reveals patterns of global and repeated species-pair divergence in sticklebacks. *Current Biology: CB*, *22*(1), 83–90. https://doi.org/10.1016/j.cub.2011.11.045

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

Joron, M., Papa, R., Beltrán, M., Chamberlain, N., Mavárez, J., Baxter, S., Abanto, M., Bermingham, E., Humphray, S. J., Rogers, J., Beasley, H., Barlow, K., ffrench-Constant, R. H., Mallet, J., McMillan, W. O., & Jiggins, C. D. (2006). A conserved supergene locus controls colour pattern diversity in *Heliconius* butterflies. *PLOS Biology*, *4*(10), e303. https://doi.org/10.1371/journal.pbio.0040303

Le Moan, A., Bekkevold, D., & Hemmer-Hansen, J. (2021). Evolution at two time frames: Ancient structural variants involved in post-glacial divergence of the European plaice (*Pleuronectes platessa*). *Heredity*, *126*(4), Article 4. https://doi.org/10.1038/s41437-020-00389-3

Lowry, D. B., & Willis, J. H. (2010). A widespread chromosomal inversion polymorphism contributes to a major life-history transition, local adaptation, and reproductive isolation. *PLOS Biology*, *8*(9), e1000500.

https://doi.org/10.1371/journal.pbio.1000500
Lucek, K., Giménez, M. D., Joron, M., Rafajlović, M., Searle, J. B., Walden, N., Westram, A. M., & Faria, R. (2023). The impact of chromosomal rearrangements in speciation: from micro- to macroevolution. *Cold Spring Harbor Perspectives in Biology*, a041447. https://doi.org/10.1101/cshperspect.a041447
Martinez Barrio, A., Lamichhaney, S., Fan, G., Rafati, N., Pettersson, M., Zhang, H., Dainat, J., Ekman, D., Höppner, M., Jern, P., Martin, M., Nystedt, B., Liu, X., Chen, W., Liang, X., Shi, C., Fu, Y., Ma, K., Zhan, X., … Andersson, L. (2016). The genetic basis for ecological adaptation of the Atlantic herring revealed by genome sequencing. *eLife*, *5*, e12081. https://doi.org/10.7554/eLife.12081
Matschiner, M., Barth, J. M. I., Tørresen, O. K., Star, B., Baalsrud, H. T., Brieuc, M. S. O., Pampoulie, C., Bradbury, I., Jakobsen, K. S., & Jentoft, S. (2022). Supergene origin and maintenance in Atlantic cod. *Nature Ecology & Evolution*, *6*(4), Article 4. https://doi.org/10.1038/s41559-022-01661-x
Mayr, E. (1947). Ecological Factors in Speciation. *Evolution*, *1*(4), 263–288. https://doi.org/10.2307/2405327
Mérot, C., Llaurens, V., Normandeau, E., Bernatchez, L., & Wellenreuther, M. (2020). Balancing selection via life-history trade-offs maintains an inversion polymorphism in a seaweed fly. *Nature Communications*, *11*(1), Article 1. https://doi.org/10.1038/s41467-020-14479-7
Le Moan, A., Gaggiotti, O., Henriques, R., Martinez, P., Bekkevold, D., & Hemmer-Hansen, J. (2019). Beyond parallel evolution: When several species colonize the same environmental gradient (p. 662569). *bioRxiv*. https://doi.org/10.1101/662569
Le Moan, A., Stankowski, S., Rafajlovic, M., Ortega-Martinez, O., Faria, R., Butlin, R., & Johannesson, K. (2023). Coupling of 12 chromosomal inversions maintains a strong barrier to gene flow between ecotypes (p. 2023.09.18.558209). *bioRxiv*. https://doi.org/10.1101/2023.09.18.558209
Nosil, P., Feder, J. L., & Gompert, Z. (2021). How many genetic changes create new species? *Science*, *371*(6531), 777–779. https://doi.org/10.1126/science.abf6671
Rieseberg, L. H. (2001). Chromosomal rearrangements and speciation. *Trends in Ecology & Evolution*, *16*(7), 351–358. https://doi.org/10.1016/s0169-5347(01)02187-5
Schluter, D., & Conte, G. L. (2009). Genetics and ecological speciation. *Proceedings of the National Academy of Sciences*, *106*, 9955–9962. https://doi.org/10.1073/pnas.0901264106
Schumer, M., Cui, R., Rosenthal, G. G., & Andolfatto, P. (2015). Reproductive isolation of hybrid populations driven by genetic incompatibilities. *PLOS Genetics*, *11*(3), e1005041. https://doi.org/10.1371/journal.pgen.1005041
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., … Rieseberg, L. H. (2020). Massive haplotypes underlie ecotypic differentiation in sunflowers. *Nature*, *584*(7822). https://doi.org/10.1038/s41586-020-2467-6
Westram, A. M., Stankowski, S., Surendranadh, P., & Barton, N. (2022). What is reproductive isolation? *Journal of Evolutionary Biology*, *35*(9), 1143–1164. https://doi.org/10.1111/jeb.14005

ANNEXES

Annex 1 - Résumé en français

Problématique

La biologie de la spéciation s'intéresse aux facteurs qui promeuvent et modèlent la dynamique de diversification du vivant. A travers des approches génomiques qui ont subi des développements technologiques remarquables au cours des quinze dernières années grâce au progrès et à l'accessibilité du séquençage ainsi que par le développement d'outils d'analyse bioinformatique dédiés, elle cherche à comprendre ce qui rend possible dans le temps et dans l'espace la dynamique évolutive d'entités biologiques – populations, lignées, écotypes… - qui peuvent ultimement mener à des espèces distinctes. Comprendre cette dynamique nécessite principalement :

- de détecter les éléments constitutifs de l'architecture des génomes qui montreraient une certaine rupture des flux de gènes et un arrêt de la recombinaison et qui pourraient ainsi constituer les éléments d'un isolement reproducteur entre unités biologiques ;

- de décrire les caractéristiques majeures de ces éléments (nombre, longueur, et positionnement chromosomique, constitution), mais c'est aussi chercher à décrire les événements relatifs à leur apparition dans un contexte démo-historique qui a pu favoriser la divergence entre entités (par ex. mutation de novo *vs* introgression), à leur établissement et à leur maintien (nature des processus sélectifs s'opposant aux flux géniques), ainsi que leur dynamique future ;

- de comparer ces processus et ces dynamiques chez différents organismes aux aires de distribution largement communes qui, à une échelle plus fine, se répartissent dans et exploitent des environnements et habitats similaires et chez lesquels des indices généralement de nature morpho-anatomique mais parfois moléculaire d'une différenciation écotypique ont été reportés.

Est-il alors possible d'observer une récurrence dans la nature des processus mis en œuvre dans l'apparition de tels écotypes et donc des trajectoires évolutives au moins partiellement similaires ? Quels éléments pourraient alors participer et justifier les similarités observées : une histoire évolutive partagée ? une capacité d'adaptation locale similaire, convergente pour faire face à un même gradient écologique entre habitats ? une architecture génomique ancestrale partagée ?

Contexte de l'étude

Une approche multispécifique - Mon travail s'inscrit dans ce canevas et s'intéresse globalement à cinq espèces de poissons marins côtiers : l'anchois (*Engraulis encrasicolus*), l'hippocampe moucheté (ou à long bec : *Hippocampus guttulatus*), l'athérine (*Atherina boyeri*), le crénilabre

cendré (*Symphodus cinereus*) et enfin un syngnathe (*Syngnathus typhle*). Si dans le domaine des espèces marines dont la dispersion est reconnue comme suffisamment élevée pour promouvoir de flux géniques importants et soumise à peu de barrières écogéographiques (i.e. une forte connectivité), il a été longtemps et trop communément admis que la différenciation génétique devait être limitée avant tout par les capacités physiologiques des espèces et des règles d'assemblage d'espèces qui contraignaient leur niche écologique. Ce paradigme est désormais largement battu en brèche et de très nombreuses études ont démontré que de nombreuses espèces marines – notamment de vertébrés et d'invertébrés – pouvaient posséder une structure génétique marquée, même à une échelle géographique restreinte suggérant des capacités d'adaptation locale non négligeables. Ainsi, chacune des espèces sus-citée possède au moins un écotype marin et un écotype lagunaire. Il est nécessaire en premier lieu de ***(i)*** décrire pour chacune d'entre elles qu'elle est la nature génomique qui sous-tend cette différenciation, voire même si elle existe réellement et ne repose pas uniquement sur de la plasticité phénotypique, et ***(ii)*** de tester si la structure génétique observé est corrélée avec la structure des habitats (i.e. si la structure génétique observée oppose habitats marins et lagunaires).

Pour ces différentes espèces, les connaissances déjà acquises sont distinctes en raison de l'intérêt qu'elles ont suscité, tout particulièrement au niveau moléculaire. Nous pouvons considérer que pour l'anchois et l'hippocampe, un nombre conséquent d'études a été réalisé. Ceci a permis de décrire les structures génétiques de ces espèces et – dans ces ceux cas – une structure génétique basée sur le gradient laguno-marin. Bien moins de données sont disponibles à la fois pour le syngnathe et l'athérine ; elles sont inexistantes chez le crénilabre. Néanmoins, si des structures ont pu être déjà observées, pour aucune d'entre elles la nature structurelle de la différenciation génétique n'a réellement pu être décrite à l'échelle du génome et encore moins comparée.

Enfin, seuls des poissons osseux (ostéichtyens) sont concernés et si les espèces considérées diffèrent par de très nombreux traits de vie, ce groupe se caractérise par une architecture génomique qu'il est possible de qualifier comme étant très conservée au sein de ce groupe, même si la taille des génomes peut largement varier.

Des données de génomes complets pour de nouvelles possibilités - Pour chacune des espèces, des séquençages de génomes complets ont été réalisés (*N*= 25 par espèce) et analysés en mettant en avant leur dimension écotypique. Comme dans d'autres études et notamment pour des espèces marines – notamment littorines, morue et épinoches – j'ai aussi pu acquérir après avoir réalisé l'assemblage de ces génomes de nouvelles données de séquençage qui m'ont permis de réaliser des investigations plus précises. Ces données m'ont tout particulièrement permis d'aller à la recherche de certains variants structuraux au sein de ces génomes et tout particulièrement d'inversions chromosomiques qui ont pu être recherchées chez chacune des espèces. Si les inversions ne sont pas les seuls variants structuraux au sein des génomes (insertions, duplications, …), les données commencent à s'accumuler quant à leur rôle dans la variation écotypique des espèces et leur isolement reproducteur et – plus généralement – le rôle des variants structuraux comme éléments de l'architecture génomique impliqués dans ce type de différenciation.

Une inversion est donc une portion du génome dont la position et la longueur sur un chromosome sont identifiables, qui est protégée de la recombinaison (donc du flux génique) et peut contenir des gènes coadaptés qui confèrent un avantage adaptatif dans un environnement donné, potentiellement contribuent à l'isolement reproducteur et dont chaque nucléotide en faisant partie est en déséquilibre de liaison avec ses voisins. Entre autres choses, une inversion peut être présente dans une lignée/population ou pas, polymorphe (*a minima* un haplotype marin, un haplotype lagunaire) ou pas, et ainsi être analysée à travers des modèles classiques qui permettent d'inférer les formes de sélection à l'origine de la différenciation mer-lagune. La distribution de ces inversions dans le génome ainsi que dans les lignées, leurs interactions, leurs compositions en accueillant elles-mêmes de nouveaux variants ou des événements de recombinaisons limités peuvent contribuer à une meilleure connaissance des processus évolutifs qui sous-tendent la dynamique de différenciation mer-lagune des espèces.

Quand elles étaient présentes, la connaissance de ces inversions m'a permis de mieux analyser les données complémentaires notamment chez l'anchois et l'hippocampe qui sont les deux espèces qui m'ont le plus intéressées.

Un échantillonnage commun - Quatre localités méditerranéennes et atlantiques ont initialement été considérées (**Fig. 1**) pour l'échantillonnage des individus qui ont servi au séquençage complet de génome. A chacune de ces localités, des individus ont été prélevées. En Méditerranée, il s'agit du Golfe du Lion (France : Etang de Thau, LI), Murcia (Espagne : Mar Menor, MU), puis en Atlantique, de Faro (Portugal : Rio Formosa, FA), et de Port d'Albret (France : Lac marin, GA). D'autres localités spécifiques à l'étude de certaines espèces (anchois et hippocampe) ne sont pas données ici.

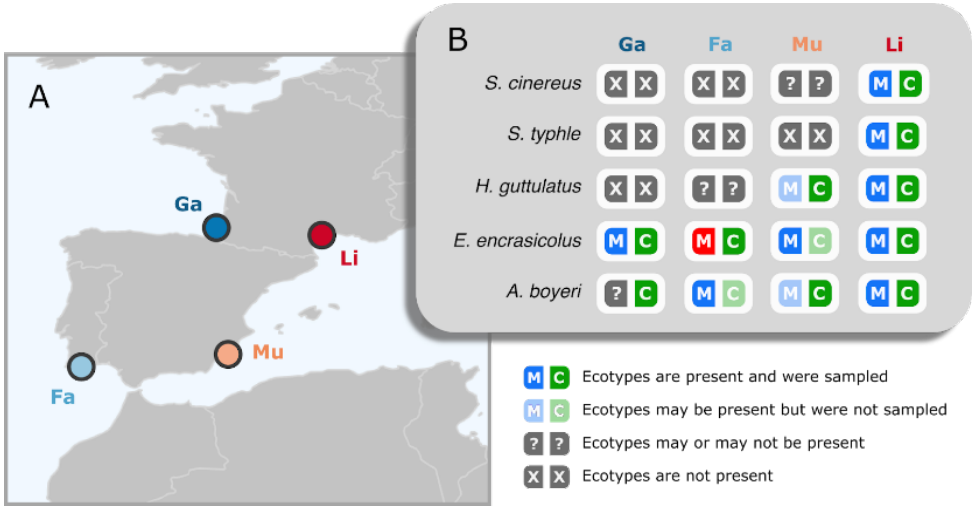


Fig. 1. (A) Localités où les échantillons des 5 espèces ont été prélevées pour analyse de leurs génomes complets (acronymes donnés dans le texte); **(B)** Représentation graphique de la présence des écotypes marins (M) et côtiers/lagunaires (C) dans chacune des localités.

Cas d'étude 1 : Les anchois (*Engraulis spp.*)

Les travaux sur l'anchois ont été réalisés dans le contexte d'une structure écotypique connue qui – pour son versant purement atlantique et méditerranéen (plus Mer Noire) – est établie

depuis un certain temps. Une révision taxonomique récente a été proposée pour définir une « espèce » côtière (*E. maeoticus*; notée C) et une « espèce » marine (*E. encrasicolus*; notée M) qui sont sans doute des lignées distinctes qui ont ensuite connu un contact secondaire. Le génome de l'anchois est supérieur à 2Gb et je n'ai pu sans doute avoir accès à son ensemble et donc réaliser un assemblage parfait, mais j'ai pu pour la première fois caractériser la nature de cette différenciation qui réside bien dans des inversions, dont six inversions principales permettent de différencier anchois côtières et anchois marines européennes. Néanmoins, l'échantillonnage incluant des anchois des côtes ouest-africaines (Afrique du Sud, Canaries, Maroc), je montre également qu'une troisième ligné d'anchois est présente, caractérisée par des inversions qui lui sont propres. Cette troisième lignée (notée C pour 'Sud') pourrait être assimilée à *E. capensis*, présente en Afrique australe. Celle-ci n'a jamais été reportée au-delà du Golfe de Guinée et de l'Equateur.

Si le statut spécifique compte peu ici, je montre avant que des inversions spécifiques de cette troisième lignée sont bien présentes tant chez les anchois côtières et marines européennes, notamment à Faro et au Maroc mais aussi une présence marquée en Méditerranée où ces inversions sont plus présentes chez la forme européenne côtière que marine (**Fig. 2**). Il existe donc des flux géniques variables et des composantes d'un isolement reproducteur entre les lignées. Ceci mène pour le nombre d'inversion auxquelles j'ai eu accès à des combinatoires complexes entre celles-ci et donc à des architectures génomiques particulièrement diversifiées notamment la transition Méditerranée-Atlantique.

Les trois lignées d'anchois ont vraisemblablement connue deux contacts secondaires, dont l'un serait ancien et a affecté autant les lignées européennes marines et côtières générant des patrons d'hybridation entre paires de lignées (CS, CM), puis une seconde vague plus récente, plus limitée spatialement et prépondérante chez les anchois marines qui pourraient ainsi combiner au fil de générations et de barrières reproductives imparfaites des inversions des trois génomes et présenter un génome CMS.

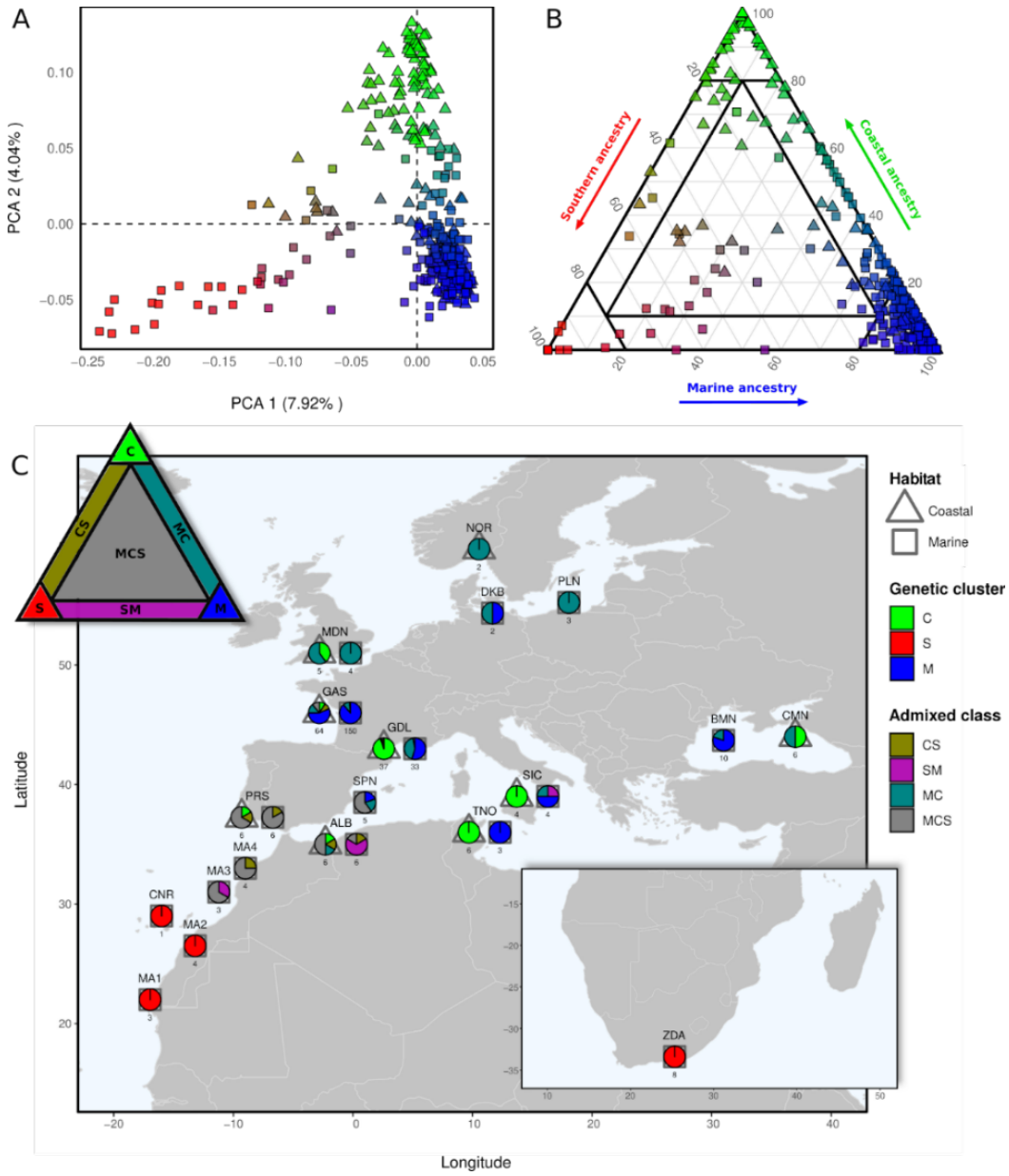


Fig. 2. Illustration de la différenciation écotypique dans trois lignées d'anchois (*Engraulis spp*.) basée sur l'analyse de leurs variants structuraux (inversions) **A –** Plan ½ d'ACP illustrant l'existence parmi 385 individus de trois lignées d'anchois (bleu : lignée européenne marine; vert : lignée européenne côtière; rouge : lignée sud) et des individus admixés en gris. **B –** proportions d'admixture de chacun des individus incluant les individus non-admixés à chacun des pôles du triangle, les individus admixés entre paires de lignées chacun des côtés (formes CM, Cs, SM) et au centre de ce triangle les individus qui portent des traces des trois lignées (CMS). **C** – Distributions des individus dans chacune des populations analysées et habitats. Les couleurs reflètent leur appartenance à une lignée (cluster génétique) ou à une des formes réarrangées/admixées pour leurs inversions.

Cas d’étude 2 - L’hippocampe (*Hippocampus guttulatus*)

L’hippocampe est une espèce côtière qui à la différence de l’anchois est sédentaire et possède un génome compact. J’ai généré un génome de référence (451Mb) au niveau chromosomique et décrit les patrons de diversité de 112 individus à ce niveau de détail. J’ai ainsi pu établir pour cette espèce la présence de deux inversions chromosomiques de plusieurs mégabases présentes dans ce génome et détaillé une structure déjà décrite qui oppose en Atlantique une différenciation géographique des populations selon un axe nord-sud sans composante écotypique, alors qu’en Méditerranée et en Mer Noire, cette différenciation est bien écotypique et structurée sur une différenciation laguno-marine (**Fig. 3**). En positionnant ces résultats dans les attendus d’un modèle définissant différents types de polymorphismes issus de réarrangements structuraux de type inversions, j’ai montré que celles-ci représentent des polymorphismes intraspécifiques anciens dont l’un est maintenu par sélection divergente (sur le chromosome B12, fixé pour des allèles distincts en milieux marin et lagunaire en Méditerranée), alors que l’autre serait maintenu par sur-dominance associée (‘associative overdominance’) en raison de la longueur plus importante de l’inversion qui le rendrait plus susceptible de porter des mutations délétères dont les effets seraient masqués par le polymorphisme (sur chromosome B2). En Méditerranée, les combinaisons haplotypiques de ces inversions suggèrent fortement des interactions pléiotropiques entre celles –ci et pourraient induire des effets environnement-dépendant sur la valeur sélective et la différenciation écotypique méditerranéenne de cette espèce.

J’ai par ailleurs montré que chacune de ces inversions possédait des allèles (ancestral *vs* inversé) hautement divergents, mais qu’il n’existait aucun signe tangible d’érosion de la divergence pour B12 (analyse du ‘gene flux’, des événements recombinaisons internes à l’inversion due dans ce cas à des doubles crossing-overs) alors que celle-ci s’érodait pour B2 et participait sans doute - à un rythme inconnu mais impliquant des événements assez récents d’un point de vue évolutif - à la purge d’un certain nombre de mutations délétères présentes sur B2. Je prédis que cette inversion va abaisser le niveau de sur-dominance associée qui préside à son maintien. Les inversions peuvent donc posséder intrinsèquement des dynamiques qui ont bien été décrites dans la littérature, mais il est possible de se poser la question sur la dynamique conjointe d’inversions multiples qui, elle, reste un champ largement ouvert. L’association habitat-dépendante d’une inversion (B12) en Méditerranée pourrait être potentiellement due à des interactions à B2, alors que sans ces interactions, elle n’est observée ni en Atlantique, ni en Mer Noire.

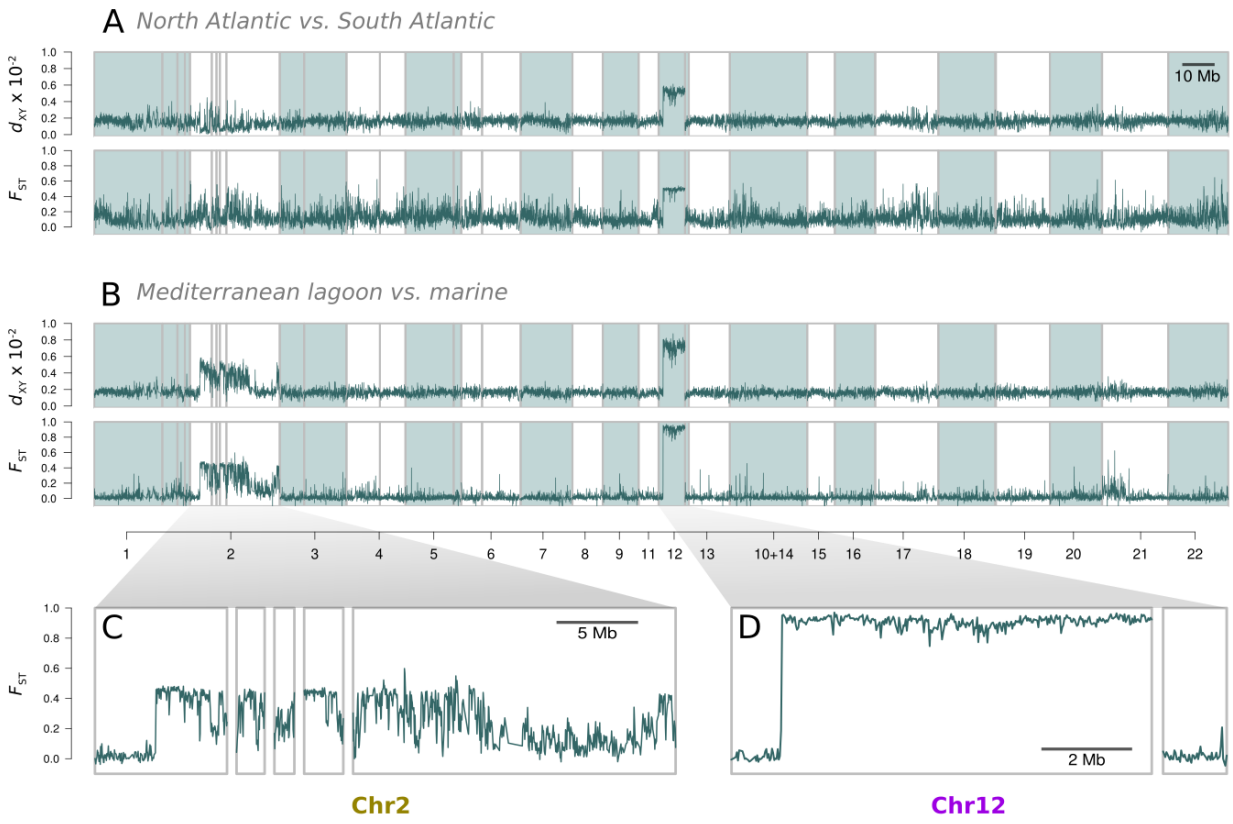


Fig. 3. Illustration des patrons de divergence (d_{XY}) et de différenciation génétique (F_{ST}) dans le génome de l'hippocampe, *H. guttulatus*, dont l'ordre des 22 chromosomes est donné en abscisse. Ceci est présenté pour : **(A)** la différenciation géographique présente en Atlantique (comparaison des populations Fa-Ga (**Fig. 1**), et **(B)** la différenciation écotypique présente en Méditerranée. Deux inversions chromosomiques sont présentes dans ce génome, l'une située **(C)** sur le chromosome (B)2 et commençant à s'éroder en raison d'une recombinaison locale par doubles crossing-overs ('gene flux'), et l'autre **(D)** sur le chromosome (B)12. Remarquer que seule l'inversion située sur le Chr. 12 est présente en Atlantique, avec des niveaux de divergence et de différenciation plus faibles. Informations supplémentaires dans le texte.

Approche comparative

Cette approche comparative vise à proposer des éléments concrets qui permettent de rendre compte de la diversité ou de la convergence des trajectoires évolutives des 5 espèces étudiées dans un gradient laguno-marin. Trois questions ont été plus particulièrement mises en avant :

- Est-ce que la variabilité écotypique observée détectée au sein de chaque espèce – traditionnellement mise sur le compte d'une spéciation écologique parallèle – possède bien une base génétique ?

- Est-ce que celle-ci se reflète à travers un type de modification similaire à l'échelle du génome ? (existe-t-il une prédominance des variants structuraux et plus précisément des inversions ?)

- Est-ce qu'il est possible de tirer des caractéristiques générales à partir des différenciations observées à l'échelle du génome dans un contexte éco-géographique.

J'ai généré des génomes de références et des données de reséquençage et analysé les données dans un pipeline standardisé. Les données obtenues ont globalement été de qualité standard à haute (10X à 50X). Dans une majorité de cas un signal géographique de différenciation a été détecté entre Atlantique et Méditerranée. Ceci reflète une structure bien connue dans cette zone géographique, présente chez une majorité d'espèces, absentes chez d'autres. Néanmoins, une structure de différenciation écotypique forte est détectée pour toutes les espèces, notamment chez *S. cinereus* et *S. typhle* chez lesquelles ceci n'avait pas été montré (**Fig. 4**), ainsi qu'une structure déjà décrite chez *A. boyeri* dans cette zone géographique mais pas à une échelle génomique, à savoir l'existence deux formes/écotypes marins et d'un unique écotype lagunaire. Les structures écotypiques de l'anchois (la présence d'une troisième lignée) et de l'hippocampe (ni structure pleinement géographique, ni écologique, mais la présence de variants structuraux en interaction) ont déjà été abordées. Cette différenciation génomique écotypique est à la fois plus forte en termes de différenciation génomique et en termes de présence chez les différentes espèces en Méditerranée qu'en Atlantique (**Fig. 1B**). Une interprétation serait que les gradients environnementaux associés à l'habitat des espèces y sont plus marqués sans que cela n'implique nécessairement une spéciation écologique.

J'ai ensuite caractérisé les architectures génomiques entre les écotypes de chaque espèce en profitant des génomes assemblés présents pour des espèces proches. Les paysages de divergence détectés entre écotypes marins et lagunaires sont hétérogènes (pics de F_{ST}) et des déséquilibres de liaison élevés le long de certains chromosomes (**Fig. 4**). Ceci suggère la présence de variants structuraux à l'origine de la différenciation écotypique dont ceux présentés chez l'anchois et l'hippocampe, mais aussi le syngnathe. Le cas de *S. cinereus* ne permet pas d'affirmer la présence de variants structuraux au sens strict, mais d'îlots de divergence qui peuvent ou non appartenir à cette catégorie de variants (échelle de résolution à ce stade non suffisante). D'autres investigations seraient nécessaires. Chez l'athérine, une différenciation élevée est observée sur tous les chromosomes ; ces écotypes seraient des espèces fortement différenciées et « vraies » (**Fig. 4**). A travers ces 5 espèces, j'illustre un continuum de spéciation et montre que pour au moins trois de ces espèces, les variants structuraux sont à l'origine de la différenciation écotypique. J'illustre ainsi leur rôle dans une différenciation centrée sur l'habitat physique, dans un contexte de connectivité qui devrait favoriser les flux de gènes et qui semble pourtant assez courante chez les poissons. Ceci mérite d'être étendu à d'autres espèces, vertébrés ou invertébrés, animales ou végétales.

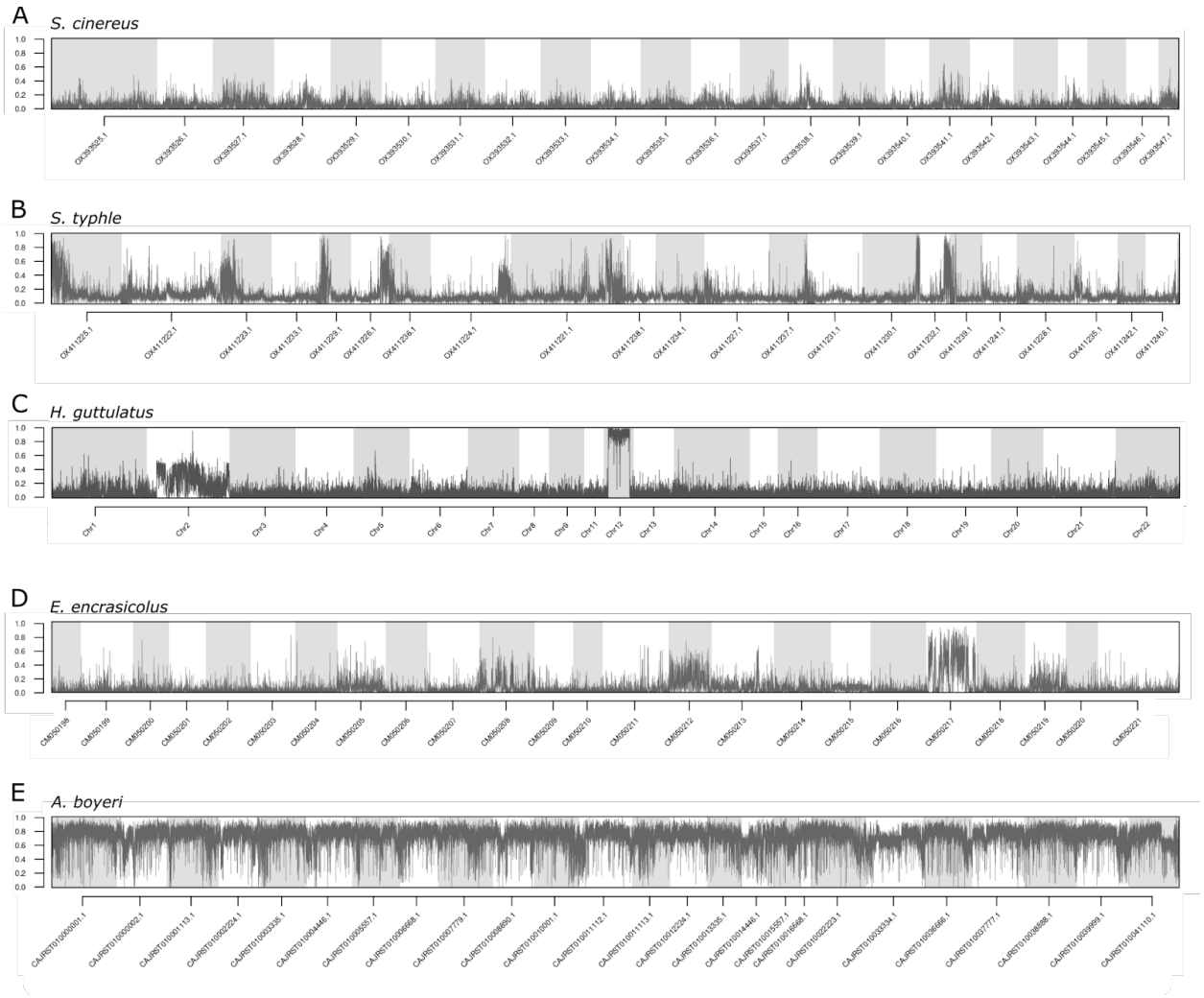


Fig. 4. Niveaux de différenciation génétique (F_{ST}) entre écotypes marins et côtiers/lagunaires sur chacun des chromosomes définis après analyse de leurs génomes complets (25 génomes/espèce) chez les espèces considérées dans mon étude (**Fig. 1**). Un continuum de spéciation est observé entre *S. cinereus* (structuration génétique écotypique faible basée sur quelques îlots de différenciation) et *A. boyeri* où la différenciation écotypique est élevée sur chaque chromosome et reflétant vraisemblablement une spéciation plus aboutie. Les trois autres espèces dont anchois et hippocampe représentent des cas intermédiaires ; arrangées ici selon un gradient de divergence. Pour *S. cinereus* et *S. typhle*, mon travail est la première à illustrer une différenciation écotypique chez ces espèces.

Conclusions – Perspectives

Le but de mes travaux était d'étudier à l'échelle génomique la structure écotypique de différentes espèces de poissons distribuées dans un gradient écologique similaire et partageant tout ou partie d'une même aire biogéographique. Je souhaitais ainsi mieux connaître les mécanismes permettant à des lignées évolutives de ségréger, essayer d'appréhender la

dynamique de cette ségrégation (séquence apparition-établissement-maintien) et ainsi devenir partiellement isolées reproductivement.

Les cinq espèces étudiées sont génétiquement différenciées sur un gradient laguno-marin et, si chaque espèce ciblée possède son individualité, mes résultats ont montré une influence fondamentale de l'architecture génomique dans la différenciation génomique écotype et le modelage des paysages adaptatifs de divergence. Des variants structuraux - notamment les inversions qui sont des barrières actives au flux de gènes et empêchent ou limitent très fortement la recombinaison – structurent largement cette différenciation écotypique chez ces espèces. Ceci est tout particulièrement vérifié en Méditerranée, beaucoup moins en Atlantique. Nos résultats rejoignent ainsi ceux de nombreuses études chez des organises terrestres et marins montrant que les processus de spéciation qui se manifestent dans la différenciation écotypique impliquent des inversions chromosomiques. Ceux-ci sont particulièrement été détaillés dans le cas des anchois et de l'hippocampe, des espèces aux traits de vie très distincts.

Par ailleurs, en m'appuyant plus spécifiquement sur ces deux espèces, je défends l'hypothèse que les variants structuraux remettent au centre du jeu le rôle des processus liés à la contingence historique comme pourvoyeurs de structures génomiques sur lesquelles la sélection naturelle peut agir. En cas de contact(s) entre lignées différenciées comme chez l'anchois, la présence d'inversions dans les génomes peut permettre l'introgression et la mise en place de nouvelles combinaisons génomiques qui se traduisent dans la structuration des écotypes. Chez l'hippocampe, la maintenance d'un polymorphisme intraspécifique ancien sur le chromosome 12 peut aussi bien être expliquée par un scénario de sélection 'extrinsèque' sur cette inversion ou sur les gènes qu'elle contient et ainsi refléterait de l'adaptation locale et un phénomène associé à la spéciation écologique, que par un modèle de résolution des conflits intragénomiques liés à des incompatibilités intrinsèques apparues et présentes dans les génomes qui vont évoluer sans lien avec l'habitat. Par ailleurs, chez cette espèce, le premier cas cité pourrait avoir eu lieu en Méditerranée où les deux écotypes sont trouvés, le second en Atlantique où la différenciation est géographique et non liée à l'habitat.

L'ensemble des variations structurelles présentes dans les génomes se révèle donc à bien des égards majeur dans nos connaissances de l'émergence d'architectures génomiques spécifiques à différents écotypes – ici côtiers et marins - par mise en place et maintien de structures génomiques caractérisées par un fort déséquilibre de liaison, et de la mise en place d'un isolement reproducteur. Si cette information est désormais aisément accessible par séquençage de génomes, ces variants structuraux peuvent aussi être « l'arbre qui cache la forêt », à savoir qu'ils peuvent cacher des éléments génomiques plus discrets impliqués dans la différenciation écotypique, ou dans un sens plus large, la spéciation écologique. Ils agissent ainsi comme des briques dont l'étude permet de construire des scénarios antagonistes intéressants, mais qui peuvent rester globalement irrésolus puisqu'il est difficile de définir la proportion de ce qui relève de processus intrinsèques et extrinsèques dans la différenciation écotypique. Si les premiers semblent plus en phase avec nos données chez l'anchois et l'hippocampe, nous cernons mal ce qui relève plus directement des seconds ou d'interactions possibles entre les deux dans l'adaptation des écotypes et/ou l'isolement reproducteur. Un

effort tout particulier devra donc être mené pour repositionner comment des variants structuraux variables en nombre, en contenu et en taille se combinent à d'autres éléments présents dans les génomes et contribuant également à leur divergence pour soutenir les mécanismes de spéciation. Cela peut passer par la modélisation, des études empiriques de terrain et sans doute des approches expérimentales.

2. Taxonomic note on European anchovies (*Engraulis cf. encrasicolus*)

Received: 19 November 2021 | Accepted: 23 November 2021

DOI: 10.1111/jfb.14964

JOURNAL OF FISH BIOLOGY fsbi

BRIEF COMMUNICATION

Systematics of European coastal anchovies (genus *Engraulis* Cuvier)

François Bonhomme[1] | Laura Meyer[1] | Christine Arbiol[1] | Daniela Bănaru[2] | Lilia Bahri-Sfar[3] | Karima Fadhlaoui-Zid[4] | Petr Strelkov[5] | Marco Arculeo[6] | Laurent Soulier[7] | Jean-Pierre Quignard[8] | Pierre-Alexandre Gagnaire[1]

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

[2]Aix-Marseille Université, Université de Toulon, CNRS, IRD, MIO UM 110, Mediterranean Institute of Oceanography, Marseille, France

[3]Biodiversité, Parasitologie et Ecologie des Écosystèmes Aquatiques, LR18ES05, 2092, Faculté des Sciences de Tunis, Université de Tunis El Manar, Tunis, Tunisia

[4]Institut National des Sciences et Technologie de la Mer, INSTM, La Goulette, Tunisia

[5]Department of Ichthyology & Hydrobiology, St Petersburg State University, St Petersburg, Russia

[6]Department STEBICEF, University of Palermo, Palermo, Italy

[7]Centre pour l'Aquaculture, la Pêche et l'Environnement en Nouvelle-Aquitaine, Bayonne, France

[8]Académie des Sciences et Lettres de Montpellier, Montpellier, France

Correspondence
Laura Meyer, Institut des Sciences de l'Evolution (UMR 5554), CNRS-UM2-IRD, Place Eugène Bataillon, Montpellier, F-34095, France.
Email: laura.meyer@umontpellier.fr

Abstract

Reports of morphological differences between European anchovy (*Engraulis* cf. *encrasicolus*) from coastal and marine habitats have long existed in the ichthyologic literature and have given rise to a long-standing debate on their taxonomic status. More recently, molecular studies have confirmed the existence of genetic differentiation between the two anchovy ecotypes. Using ancestry-informative markers, we show that coastal anchovies throughout the Mediterranean share a common ancestry and that substantial genetic differentiation persists in different pairs of coastal/marine populations despite the presence of limited gene flow. On the basis of genetic and ecological arguments, we propose that coastal anchovies deserve a species status of their own (*E. maeoticus*) and argue that a unified taxonomical framework is critical for future research and management.

KEYWORDS
ancestry-informative markers, anchovy, ecotypes, genetic divergence, partial reproductive barrier, taxonomy

The European anchovy (commonly referred to as *Engraulis encrasicolus* L. – Clupeiform, Engraulidae) is a small pelagic fish with a large geographic distribution spanning the north-eastern Atlantic and Mediterranean regions from the Baltic to the Black Sea. It is now recognized that this polytypic taxon consists of several genetically differentiated populations with contrasting abilities to occupy and forage in coastal environments (Borsa, 2002; Oueslati *et al*., 2014; Le Moan *et al*. 2016; Montes *et al*., 2016; Catanese *et al*., 2017, 2020; Huret *et al*., 2020). Some populations thrive preferentially in shallow coastal lagoons with highly variable salinity, while others are predominantly pelagic, with nevertheless a large overlap in their respective habitats (Le Moan *et al*. 2016; Catanese *et al*., 2017, 2020; Zuev, 2019; Huret *et al*., 2020). In the abundant literature on this species, the former are sometimes referred to as coastal, lagoonal or inshore populations, while the adjectives marine, pelagic or offshore are used for the latter. For the sake of simplicity, we will hereafter use the terms 'coastal' *vs*. 'marine' and relate these forms to morphology-based descriptions from the ichthyological literature. While the question of their taxonomical status as local races, subspecies or species has been pending for over a century (reviewed below), it is now well established by genetic evidence (Borsa, 2002; Oueslati *et al*., 2014; Le Moan *et al*. 2016) that the coastal form constitutes one (or several) separate evolutionarily significant units (ESU) having received several specific Latin binomens in the past. In the present paper, we address the question

of the possible unicity of the coastal form and its taxonomic consequences. We argue that coastal anchovy populations, despite being genetically differentiated from each other, share a common genetic ancestry and can be genetically recognized throughout their range as a single ESU. We further show that despite ample opportunities for gene flow, the coastal form remains genetically distinct from the marine form, implying the existence of (partial) reproductive isolation barriers that justify taxonomical recognition.

Early works on the Atlantic/Mediterranean/Black Sea anchovies went in parallel with very few cross-comparisons. Since the very beginning, it has been suggested that ecological differences between anchovy morphs could point to the existence of separate entities, sometimes referred to as 'races' (Grassi, 1903; Maximov, 1913; Zernov, 1904, 1913). However, the nature of the underlying differences, either being inherited or reflecting plastic growth trajectories in contrasted environments, remained controversial (Grassi, 1903; Lo Giudice, 1911a,b) or were dismissed (Fage, 1911; Tichy, 1914). A clear report of this ecological differentiation dates back 100 years with Lo Giudice (1922), who was the first to use the terms of 'coastal' and 'pelagic' races for anchovy populations occurring in close proximity off the Italian coasts. Shortly thereafter, related work by Pusanov (1923) and Alexandrov (1925) differentiated the anchovies of the Azov Sea from those of the open waters of the Black Sea, and a subspecific status was proposed for the Azov Sea anchovies by Pusanov and Tzeeb (1926) and Alexandrov (1927). Two decades later, in a study of anchovies from the Ionian Sea and Lake Ganzirri, Sicily, Dulzetto (1947) proposed a specific status for the latter population. Subsequent morphological studies confirmed the existence of ecophenotypic differentiation between coastal and marine anchovies in several other locations across the Mediterranean (see, for instance, Quignard *et al*., 1973), while there was still debate as to their eventual taxonomic status.

Before the advent of genetic studies, several questions relative to the evolutionary origin and status of anchovy forms remained unanswered:

1. Is phenotypic differentiation between coastal and marine anchovies a purely plastic response to living in different environmental conditions or does it have a heritable genetic basis? In other words, are the coastal and marine forms freely interbreeding or are they partially or entirely reproductively isolated?
2. In the latter case, are the various geographical populations of the coastal form closely related to one another (and likewise for the marine form) or do they constitute independent entities in each marine basin?
3. What is the phylogeographic history behind this situation?
4. Finally, what should their taxonomical status be?

These questions have now been partly solved by molecular population genetic studies, although the subject has been animated by intense debate. From the late 1970s onwards, many studies had targeted a number of exploited fish species, including anchovies. Multiple papers reported electrophoretic, mitochondrial, microsatellite or single nuclear polymorphism (SNP) variation patterns in anchovies at various geographical scales. However, most studies, surprisingly, seem to have stemmed from a *tabula rasa* with regards to the old morphological literature. It is further interesting to note that the first reports on mitochondrial DNA already evidenced two deeply divergent anchovy lineages. These two mitotypes were found to coexist in the same sampling locations, albeit in variable proportions, and therefore were not interpreted as reflecting the existence of two parapatric or quasi-sympatric entities (*e.g*., Bembo *et al*., 1996a; Magoulas *et al*., 1996; Grant, 2005; Silva *et al*., 2014; Vodyasova & Abramson, 2017). Nevertheless, on the basis of a coupled morphometric and allozymic analysis, Bembo *et al*. (1996b) concluded that there were necessarily two 'stocks' among the Adriatic anchovies, primarily separated according to water depth. Hence, the question of the existence of two ecotypic forms has largely been overlooked, even in relatively recent studies (*e.g*., Borell *et al*., 2012; Zarraonaindia *et al*., 2012; Viñas *et al*., 2014; Silva *et al*., 2014). Most of these studies pointed toward the existence of a relatively strong genetic structure as compared to other highly dispersing broadcast spawners like sardines (*e.g*., Grant *et al*., 1998). This observation was not easy to account for without invoking unrealistic limitations on individual movement or strong environmentally induced selection occurring at each generation [see, for instance, Ruggeri *et al*. (2016) for the Adriatic].

By reanalysing published allozymic data, Borsa (2002) proposed that Mediterranean anchovies present a species complex with at least two forms, one of them corresponding to a coastal form that was later proposed to deserve a species rank on its own (*Engraulis albidus*; Borsa *et al*., 2004). After these first genetic clues, several studies have addressed the extent and evolutionary origin of divergence between anchovy forms with molecular markers (Bouchenak-Khelladi *et al*., 2008; Karahan *et al*., 2014; Oueslati *et al*., 2014; Le Moan *et al*., 2016; Montes *et al*., 2016; Catanese *et al*., 2017). These studies showed that coastal anchovies could be genetically characterized in areas as distant as the Bay of Biscay, Alboran Sea and the near Atlantic, Gulf of Lions, Siculo-Tunisian Strait, Tyrrhenian Sea, Adriatic Sea and Levantine Basin, and that these were genetically more similar to each other than to geographically closer marine anchovies. As for anchovies in the Black and Azov seas and the related literature in Russian, see the review of Zuev (2019) that deals with all points above but point 2.

For the Atlantic and Mediterranean, Le Moan *et al*. (2016) more specifically addressed the question of the unique versus repeated evolutionary origin of the marine–coastal ecotype pairs. This genome-wide investigation revealed that coastal populations from the Bay of Biscay and the Gulf of Lions share a common ancestry that distinguishes them from the marine populations. The current existence of multiple ecotype pairs was thus not attributed to independent, *in situ* differentiation in response to parallel divergent selection, but to a secondary contact that probably took place about 300,000 years ago between two pre-existing evolutionary lineages followed by their spatial redistribution. Since both ecotypes are highly mobile and often hybridize, historical gene flow following secondary contact has been sufficient to partially erode the genetic differences that existed between the two anciently diverged lineages. Some regions of the

genome, such as those involved in eco-phenotypic differentiation, have, however, retained their divergence as a result of selection against unfit hybrid combinations and/or ecological selection. The use of ancestry-informative markers located in those genome regions that resist gene flow is thus crucial to be able to characterize the spatial and ecological structure of the present European anchovy populations, possibly explaining why the genetic distinction between marine and coastal anchovies was not evident in all molecular studies.

Now that the existence of two ecotypes has been widely recognized by several molecular studies, the way is paved for further investigations on the genetic bases of their physiological, behavioural and reproductive characteristics. Anchovies, being polytypic, have been able to occupy a wider range of habitats compared to a monotypic species (Catanese *et al*., 2020; Huret *et al*., 2020; Zuev, 2019). As a first step, which is the aim of this short paper, it is necessary to adopt a common vocabulary and to clarify the present-day taxonomical situation. To this end, we produced genome-wide polymorphism data using a similar methodology as in Le Moan *et al*. (2016) and complemented their sampling design with more individuals throughout the Mediterranean and Black seas. Since reduced-representation genome sequencing generates large numbers of SNPs, we considered that a limited number of individuals per location was sufficient to adequately represent the genomic variability of any given location. Given the precise objectives of the current study, our analysis was limited to taxonomic assignment based on genotypic combinations at ancestry-informative markers.

To briefly summarize the methodology, individual genomic DNA of 30 samples collected from various sampling expeditions and local fisheries was used to generate restriction-site associated DNA (RAD) sequencing libraries following a similar protocol to Baird *et al*. (2008). Sequencing was performed on an Illumina HiSeq2500 sequencer in single-read mode. Demultiplexed reads were matched to the same catalogue of loci as in Le Moan *et al*. (2016) after applying the same quality filters. We then merged the genotypes of the 30 newly sequenced individuals with those of 28 individuals from Le Moan *et al*. (2016), which were used as reference samples. Our final dataset was composed of 58 individuals representing five pairs of coastal/marine anchovy populations from the north-eastern Atlantic, the western Mediterranean and the Black Sea (Figure 1a; Supporting Information Table S1). The filtered variant call format (VCF) file contained 2952 polymorphic loci genotyped in 58 individuals with a maximum rate of 30% of missing data and a minimum allelic frequency (MAF) of 4%. Genetic structure was visualized by principal component analysis (PCA), performed using the R package SNPRelate (Zheng *et al*., 2012). A dendrogram based on an uncorrected nucleotidic similarity (identity by state, IBS) matrix was constructed using the same software. Individual assignment to *K* ancestral populations was inferred with FastSTRUCTURE (Raj *et al*., 2014).

The genomic differentiation of the 58 individuals depicted in the first two PC axes (Figure 1b) and the FastSTRUCTURE diagram (Figure 1e) mainly distinguishes two groups of samples. PC1 (7.56% of total variance) clearly separates the individuals sampled in coastal waters on the right side from those sampled in marine conditions on the left (Figure 1b). The second component (3.38% of total variance) separates coastal individuals from the Gulf of Biscay from their western Mediterranean (Tunisia, Sicily, Gulf of Lions) and Black Sea (Kerch Strait) counterparts. This differentiation along PC2 indicates that the Atlantic and Mediterranean coastal anchovies underwent significant differentiation, while their marine counterparts are less differentiated from each other [see also discussion in Catanese *et al*. (2017)]. Noticeably, some individuals appear in intermediate positions along PC1, consistent with the identification of early-generation hybrids (*e.g*., F1, F2 and backcrosses) in Le Moan *et al*. (2016), as well as later-generation backcrosses between marine and coastal anchovies both in the Atlantic and Mediterranean. Such hybrids were also evidenced from anchovy eggs along the Thyrrenian coast (Catanese *et al*., 2020). Here we observe a similar pattern for some individuals from Crimea (Kerch Strait) which could potentially represent hybrids or admixed genotypes (Figure 1b,e). The FastSTRUCTURE analysis also strongly captured the coastal/marine dichotomy at $K = 2$ (Figure 1e) without any significant changes for higher values of *K*. Individuals with mixed ancestry could correspond to different classes of hybrids as discussed above, an observation also reflected by their intermediate position in a dendrogram based on IBS distances (Supporting Information Figure S2).

The present analysis allowed various geographical populations of coastal anchovy to be related to each other through the identification of common genetic bases that distinguish them from their marine counterparts. Le Moan *et al*. (2016) showed that genetic divergence between coastal and marine ecotypes was restricted to about 20%–25% of the genome. These genomic regions contain ancestry-informative markers that are useful for ecotype assignment and for identifying hybrid genotypes. Although hybrids are relatively common, heterogeneous genome divergence between ecotypes indicates that the barrier to gene flow is sufficiently strong for the two ecotypes to persist in a parapatric/quasi-sympatric (although not entirely syntopic) situation without a complete re-mixing of their genomes. This contrasts with the relative genetic homogeneity among populations of the same ecotype throughout their geographical range. Hence, it can be considered that marine and coastal anchovies fulfil the conditions to be treated as separate species. In line with one of the most fundamental components of the biological species concept, the two anchovy forms are maintained as distinct genotypic clusters despite their spatial overlap (Mallet, 2020). This situation thus calls for a re-examination of the taxonomic status of anchovy ecotypes. According to the rule of anteriority, we discuss in what follows the correct naming of each ecotype.

Concerning the marine or offshore ecotype, we shall follow Borsa *et al*. (2004) who state: 'No type is known for this species and Linnaeus' original description is too vague to allow the distinction between the two species…/…. For the sake of stability, we propose to arbitrarily maintain the specific name *encrasicolus* to the apparently most common and widespread anchovy species in the seas of Europe. …/…also referred to as "oceanic" or "open-sea" anchovy'. This fish is often referred to as 'blue anchovy' or 'green anchovy', depending on location. The genetic homogeneity of marine anchovies has now been

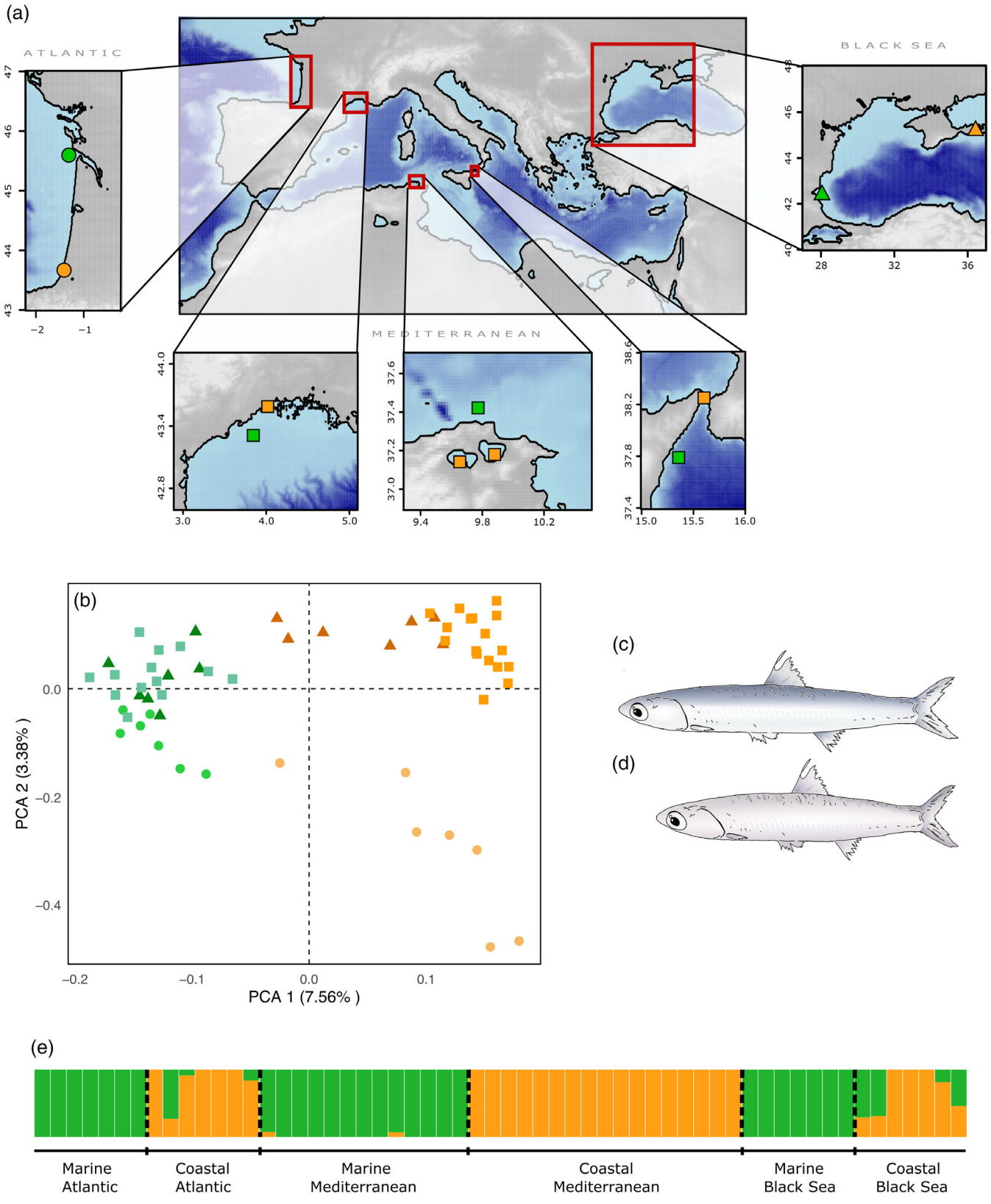


FIGURE 1 (a) Sampling locations of *Engraulis* cf. *encrasicolus*. Symbols represent locations (●, Atlantic; ■, Mediterranean; ▲, Black Sea) while colours represent habitat type (green, marine; orange, coastal). (b) Principal component analysis based on 2952 SNPs in 58 individuals (symbols correspond to those used in (a). Schematic representations are shown for anchovies from (c) marine and (d) coastal habitats. (e) Individual ancestry proportions as determined by FastSTRUCTURE with $K = 2$ clusters identified

confirmed throughout a large part of its range, and hence the numerous subspecific trinomens that were given to local populations should be considered invalid. It should also be noted that, despite being described as marine/offshore/oceanic/pelagic, individual identification by multilocus genotyping has shown that these fish are able to enter continental systems such as estuaries [see, for instance, the individuals of the marine taxon identified in the Adour estuary, Gulf of Biscay, in Le Moan *et al*. (2016)]. Borsa *et al*. (2004) have deposited a neotype and voucher specimens for this species at the Musée National d'Histoire Naturelle (MNHN), Paris (neotype MNHN 2002-1775, vouchers MNHN 2002-1776 to MNHN 2002-1844).

As for the coastal ecotype, which is the focus of the present study, our results point to the genetic homogeneity of this taxon throughout most of its range, although subtle genetic differentiation may exist among coastal populations due to limited genetic connectivity between them (Oueslati *et al*., 2014; Le Moan *et al*., 2016; Catanese *et al*., 2017). Coastal anchovies also display common morphological features that separate them from marine anchovies (Figure 1c,d). Generally, this includes paler dorsal colouration (they are often locally referred to as 'white', 'yellow', 'grey' or 'silver' anchovy in different regional languages), smaller size at maturity, fewer vertebrae, a dorsal fin implanted closer to the tail and a proportionally bigger eye. For more details, see the morphological descriptions in Borsa *et al*. (2004), Quignard *et al*. (1973), Tortonese (1967) and Karahan *et al*. (2014) as well as earlier works. A conspicuous difference in otolith shape has also been reported and used to identify putative 'stocks' (Messaoud *et al*., 2011; Vodyasova & Soldatov, 2017).

Until now, there have been, to our knowledge, three attempts at providing a morphological diagnosis and attributing a Latin binomen or trinomen to coastal anchovy populations. These are *E. e. maeoticus* (Pusanov & Tzeeb, 1926, with a diagnosis in Latin; Supporting Information Figure S3) from the Sea of Azov, *E. russoi* (Dulzetto, 1947) from Sicilian lagoons and lastly *E. albidus* (Borsa *et al*., 2004, with diagnostic features in English) from the Gulf of Lions. A mention should also be made for *E. e. symaetensis* (Dulzetto, 1940) which was collected from the 'beach' near a small estuary near Catania (Sicily). Interestingly, the morphological analysis performed by this last author indicates morphometric characteristics that are apparently intermediate to those of *E. russoi* and those of the marine Ionian Sea *E. encrasicolus* [data reanalysed in Tortonese (1967), who dismissed *symaetensis* as a valid name]). Since these samples have disappeared, it will not be possible to confirm whether they were *bona fide* coastal anchovies that were locally introgressed or a mixed stock containing hybrids.

Given the nomenclatorial rule of antecedence, the only valid name for the coastal species is *E. maeoticus* (Pusanov & Tzeeb, 1926), which applies to all coastal populations that have been found to share a common ancestry. Pusanov and Tzeeb (1926) published a comparative diagnosis for what they considered to be a subspecies and named it after the antique Meotian people that used to inhabit the banks of the Azov Sea. Since, to our knowledge, no type specimens were deposited by Pusanov nor Dulzetto, those secured by Borsa *et al*. (2004) at MNHN (type registered as MNHN 2002-1716, paratypes MNHN 2002-1717 to MNHN 2002-1774) under the name *E. albidus* should be considered as valid type specimens of *E. maeoticus*.

We believe that placing the biological diversity observed for anchovies within a clear and unified taxonomical framework will greatly benefit future research across a variety of disciplines. Although various recent studies have recognized the shared molecular bases associated with the two eco-phenotypically divergent forms, a harmonized nomenclature is critically lacking. We propose that it is time to take this step to make better sense of the future generation of whole-genome sequence data on anchovies. This will aid characterization of the molecular bases and biological functions associated with the species' ecological divergence. Furthered by these molecular advances, eco-physiological studies will hopefully be able to shed some light on the biology of marine (*E.* cf. *encrasicolus*) and coastal (*E. maeoticus*) anchovies, investigating the genetic bases of behavioural, physiological and life-history traits that explain the persistence of the two species despite their large co-occurrence. Such advances would also provide valuable tools to improve current fishery models and to move towards a management of stocks that takes the biological duality of anchovies into account. Last but not least, we hope that this taxonomic recognition in one of the most emblematic fishes in the Mediterranean ecosystem will encourage future consideration of cryptic subdivisions that also exist in other fish species to ultimately better preserve these hidden layers of biodiversity.

ACKNOWLEDGEMENTS

We are grateful to Dr Paola Rinelli (CNR Messina) for the samples from Lake Ganzirri and the Ionian Sea and to Dr M.L. Molino, Director of the Riserva Naturale Orientata Capo Peloro, for authorization PROT. INT. N.7140/VI DEL to collect from Lake Ganzirri (91/12/2017). We are also grateful to Dr Aurelio Mazzè of the University of Palermo and Christine Bibal from ISEM for helping us to find some old papers. Lastly, we would like to thank the Montpellier Bioinformatics Biodiversity platform as well as the GenSeq sequencing platform. This work was financed through the basal annual funding of ISEM.

ORCID

François Bonhomme https://orcid.org/0000-0002-8792-9239
Jean-Pierre Quignard https://orcid.org/0000-0002-1908-3235

REFERENCES

Alexandrov, A. I. (1925). Annual report of Kerch ichthyologic laboratory for the year 1924. *Trudy Kerchenskoy ikhtiologicheskoy laboratorii [Reports of the Kertch Ichthyological Laboratory]*, *1*, 32–34 (in Russian).

Alexandrov, A. I. (1927). Anchovies of the Azov–Black Sea basin, their origin and taxonomic designation. *Trudy Kerchenskoi nauchnoi rybokhozyaistvennoi stantsii [Proceedings of Kerch Scientific Station of Fisheries]*, *1*(2–3), 35–92 (in Russian, summary in French p. 93–100.).

Baird, N. A., Etter, P. D., Atwood, T. S., Currey, M. C., Shiver, A. L., Lewis, Z. A., … Johnson, E. A. (2008). Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS One*, *3*, e3376.

Bembo, D. G., Carvalho, G. R., Cingolani, N., Arneri, E., Giannetti, G., & Pitcher, T. J. (1996a). Allozymic and morphometric evidence for two stocks of the European anchovy Engraulis encrasicolus in Adriatic waters. *Marine Biology*, *126*(3), 529–538.

Bembo, D. G., Carvalho, G. R., Snow, M., Cingolani, N., & Pitcher, T. J. (1996). Stock discrimination among European anchovies, Engraulis

encrasicolus, by means of PCR-amplified mitochondrial DNA analysis. *Oceanographic Literature Review*, *12*(43), 1274.

Borsa, P. (2002). Allozyme, mitochondrial-DNA, and morphometric variability indicate cryptic species of anchovy (*Engraulis encrasicolus*). *Biological Journal of the Linnean Society*, *75*, 261–269.

Borsa, P., Collet, A., & Durand, J. D. (2004). Nuclear-DNA markers confirm the presence of two anchovy species in the Mediterranean. *Comptes Rendus Biologies*, *327*, 1113–1123.

Borrell, Y. J., Pinera, J. A., Sánchez Prado, J. A., & Blanco, G. (2012). Mitochondrial DNA and microsatellite genetic differentiation in the European anchovy Engraulis encrasicolus L. *ICES Journal of Marine Science*, *69*(8), 1357–1371.

Bouchenak-Khelladi Y., Durand J.-D., Magoulas A., & Borsa P. (2008). Geographic structure of European anchovy: A nuclear-DNA study. *Journal of Sea Research*, *59*(4), 269–278. https://doi.org/10.1016/j.seares.2008.03.001

Catanese, G., Watteaux, R., Montes, I., Barra, M., Rumolo, P., Borme, D., … Procaccini, G. (2017). Insights on the drivers of genetic divergence in the European anchovy. *Scientific Reports*, *7*, 1–15.

Catanese, G., Di Capua, I., Iriondo, M., Bonanno, A., Estonba, A., & Procaccini, G. (2020). Application of high-throughput single nucleotide polymorphism genotyping for assessing the origin of *Engraulis encrasicolus* eggs. *Aquatic Conservation: Marine and Freshwater Ecosystems*, *30*, 1313–1324.

Dulzetto, F. (1940). Su una particolare forma di Engraulis della "Plaia" di Catania e del Simeto. *Atti della Reale Accademia d' Italia*, *XI*, 325–400.

Dulzetto, F. (1947). L'Engraulis dei laghi di Ganzirri e del Faro. *Memorie della Società Italiana delle Scienze Detta dei XL*, *3* (XXVI), 5–28.

Fage, L. (1911). Recherches sur la biologie de l'anchois *(Engraulis encrasicolus* LINNÉ): races âge, migrations. *Annales de l'Institut Océanographique*, *2*(4), 1–37.

Grant, W. S. (2005). A second look at mitochondrial DNA variability in European anchovy (Engraulis encrasicolus): assessing models of population structure and the Black Sea isolation hypothesis. *Genetica*, *125* (2), 293–309.

Grant, W. S., & Bowen, B. W. (1998). Shallow population histories in deep evolutionary lineages of marine fishes: insights from sardines and anchovies and lessons for conservation. *Journal of heredity*, *89*(5), 415–426.

Grassi, L., (1903). I pesci dei laghi di Ganzirri e Faro.*Neptunia*, Fasc. 17, 18, 19.

Huret, M., Lebigre, C., Iriondo, M., Montes, I., & Estonba, A. (2020). Genetic population structure of anchovy (*Engraulis encrasicolus*) in north-western Europe and variability in the seasonal distribution of the stocks. *Fisheries Research*, *229*, 105619. https://doi.org/10.1016/j.fishres.2020.105619.

Karahan, A., Borsa, P., Gucu, A. C., Kandemir, I., Ozkan, E., Orek, Y. A., … Togan, I. (2014). Geometric morphometrics, Fourier analysis of otolith shape, and nuclear-DNA markers distinguish two anchovy species (*Engraulis* spp.) in the Eastern Mediterranean Sea. *Fisheries Research*, *159*, 45–55.

Le Moan A., Gagnaire P.-A., & Bonhomme F. (2016). Parallel genetic divergence among coastal-marine ecotype pairs of European anchovy explained by differential introgression after secondary contact. *Molecular Ecology*, *25*(13), 3187–3202. https://doi.org/10.1111/mec.13627

Lo Giudice, P. (1911a). Le Acciughe dei mari italiani. *Rivista Mensile di Pesca e Idrobiologia*, *Pavia*, *6*(13), 83–87.

Lo Giudice, P. (1911b). Ancora sulle diverse razze locali di Acciughe (*Engraulis encrasicholus* L.). *Rivista Mensile di Pesca e Idrobiologia*, *Pavia*, *6*(13), 226–236.

Lo Giudice, P. (1922). Le Acciughe dei mari Italiani. *Bollettino della Società dei naturalisti in Napoli*, *34*, 196–209.

Mallet, J. (2020). Alternative views of biological species: Reproductively isolated units or genotypic clusters? *National Science Review*, *7*(8), 1401–1407.

Magoulas, A., Tsimenides, N., & Zouros, E. (1996). Mitochondrial DNA phylogeny and the reconstruction of the population history of a species: the case of the European anchovy (Engraulis encrasicolus). *Molecular Biology and Evolution*, *13*(1), 178–190.

Maximov, N. E. (1913). Life pattern and commercial fishery in North-Western part of the Black Sea in front the Bulgarian and Romanian coast. *Ezhegodnik Zoologicheskogo muzeya Imperatorskoi akademii nauk* [Annales du Musée Zoologique de l'Académie Impériale des Sciences de St.-Pétersbourg], *18*(1), 1–52 (in Russian).

Messaoud, H., Bouriga, N., Daly Yahia, M. N., Boumaiza, M., Faure, E., Quignard, J. P., & Trabelsi, M. (2011). Discrimination de trois populations d'anchois du genre *Engraulis* (Clupeiforme, Engraulidae) des côtes Tunisiennes par analyse de forme des otolithes. *Bulletin de l'Institut National des Sciences et Techniques de la Mer (Salambô)*, *38*, 21–27.

Montes, I., Zarraonaindia, I., Iriondo, M., Grant, W. S., Manzano, C., Cotano, U., … Estonba, A. (2016). Transcriptome analysis deciphers evolutionary mechanisms underlying genetic differentiation between coastal and offshore anchovy populations in the Bay of Biscay. *Marine Biology*, *163*, 1–13.

Oueslati, S., Fadhlaoui-Zid, K., Kada, O., Augé, M. T., Quignard, J. P., & Bonhomme, F. (2014). Existence of two widespread semi-isolated genetic entities within Mediterranean anchovies. *Marine Biology*, *161*, 1063–1071.

Pusanov, I. I. (1923). Data on commercial ichthyology of the crimea. *Rybnoe khozyaistvo [Fisheries]*, *2*, 10–16 (in Russian).

Pusanov, I., & Tzeeb, Y. (1926). About anchovy races inhabiting the black and Azov seas. *Trudy Krymskogo nauchno-issledovatelskogo instituta [Proceedings of the Crimea Research Institute]*, *1*, 86–95 (in Russian with German summary).

Quignard, J. P., Hamdouni, T., & Zaouli, J. (1973). Données préliminaires sur les caractères biométriques des Anchois *Engraulis encrasicolus* (Linné, 1758) des côtes de Tunisie et du Lac Ichkeul. *Revue des Travaux de l' Institut des Pêches Maritimes*, *37*(2), 191–196.

Raj, A., Stephens, M., & Pritchard, J. K. (2014). fastSTRUCTURE: Variational inference of population structure in large SNP data sets. *Genetics*, *197*, 573–589.

Ruggeri, P., Splendiani, A., Occhipinti, G., Fioravanti, T., Santojanni, A., Leonori, I., … Caputo Barucchi, V. (2016). Biocomplexity in populations of European anchovy in the Adriatic Sea. *PLoS One*, *11*, e0153061. https://doi.org/10.1371/journal.pone.0153061.

Silva, G., Lima, F. P., Martel, P., & Castilho, R. (2014). Thermal adaptation and clinal mitochondrial DNA variation of European anchovy. *Proceedings of the Royal Society of London B: Biological Sciences*, *281*. https://doi.org/10.1098/rspb.2014.1093.

Tichy, M. I. (1914). Some words about anchovies. *Vestnik rybopromyshlennosti [Vestnik fishery]*, *1-2*, 50–68 (in Russian).

Tortonese, E. (1967). Differenziazioni infraspecifiche nelle Acciughe (*Engraulis encrasicholus* L.: Pisces, Clupeiformes) della Sicilia Orientale. *Atti Accadademia Gioenia Scienze Naturali di Catania*, *19*, 57–65.

Viñas, J., Sanz, N., Penarrubia, L., Araguas, R.-M., Garcıa-Marın, J.-L., Roldan, M. I., & Pla, C. (2014). Genetic population structure of European anchovy in the Mediterranean Sea and the Northeast Atlantic Ocean using sequence analysis of the mitochondrial DNA control region. *ICES Journal of Marine Science*, *71*(2), 391–397.

Vodiasova, E. A., & Abramson, N. I. (2017). Genetic variability of anchovy in the Azov-Black Sea basin. *Russian Journal of Genetics*, *53*(6), 680–687.

Vodyasova, E. A., & Soldatov, A. A. (2017). Identification of subspecies of European anchovy Engraulis encrasicolus (Engraulidae) in the wintering aggregations based on morphological parameters of otoliths. *Journal of Ichthyology*, *57*(4), 553–559.

Zarraonaindia I., Iriondo M., Albaina A., Pardo M. A., Manzano C., Grant W. S., Irigoien X., & Estonba A. (2012). Multiple SNP Markers Reveal Fine-Scale Population and Deep Phylogeographic Structure in

European Anchovy (Engraulis encrasicolus L.). *PLoS ONE*, *7*(7), e42201. https://doi.org/10.1371/journal.pone.0042201

Zernov, S. A. (1904). Third report on studying fisheries in Tauride gubernia. *S.M.Spiro Press Sevastopol*, 16p. (in Russian).

Zernov, S. A. (1913). On the study of life of the Black Sea. *Mémoires de l' Académie Impériale des sciences de St.Petersburg VIII*, *32*(1), 167–171 (in Russian).

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

Zuev, G. V. (2019). Current population structure of the European anchovy *Engraulis encrasicolus* L. (Engraulidae; Pisces) in the Black and Azov sea sand history of its formation. *Morskoj Biologicheskij Zhurnal [Marine Biological Journal]*, *4*, 45–62 (in Russian). https://doi.org/10.21072/mbj.2019.04.1.05.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

How to cite this article: Bonhomme, F., Meyer, L., Arbiol, C., Bănaru, D., Bahri-Sfar, L., Fadhlaoui-Zid, K., Strelkov, P., Arculeo, M., Soulier, L., Quignard, J.-P., & Gagnaire, P.-A. (2022). Systematics of European coastal anchovies (genus *Engraulis* Cuvier). *Journal of Fish Biology*, *100*(2), 594–600. https://doi.org/10.1111/jfb.14964

Annex 3 - Supplementary information

3.1. Supplementary information to Chapter I

Supplementary Table S1. All samples used in the study (n=385), including 128 samples obtained from Le Moan et al. (2016). Habitat type was classified as either coastal ("COT", i.e. lagoons and estuaries) or marine ("MAR"). The "WGS" and "RAD" columns indicate whether a sample was included for whole-genome sequencing and/or RAD-sequencing. The last column corresponds to the genetic cluster or admixed class that each sample was assigned to based on ADMIXTURE results (Fig. 1B).

NOM_FINAL	Type	Latitude	Longitude	Location	Habitat	WGS	RAD	Genetic class
MED_MAR_ALB_34_0483	Tissue	35.256	-3.835	ALB	MAR	no	yes	SM
MED_MAR_ALB_34_0493	Tissue	35.256	-3.835	ALB	MAR	no	yes	SM
MED_MAR_ALB_34_0494	Tissue	35.256	-3.835	ALB	MAR	no	yes	SM
MED_LAG_ALB_37_1436	Tissue	35.170	-2.870	ALB	COT	no	yes	MCS
MED_LAG_ALB_38_1439	Tissue	35.170	-2.870	ALB	COT	no	yes	CS
MED_LAG_ALB_37_1431	Tissue	35.170	-2.870	ALB	COT	no	yes	MC
MED_LAG_ALB_37_1432	Tissue	35.170	-2.870	ALB	COT	no	yes	MCS
MED_LAG_ALB_38_1440	Tissue	35.170	-2.870	ALB	COT	no	yes	MCS
MED_LAG_ALB_38_1441	Tissue	35.170	-2.870	ALB	COT	no	yes	C
MED_MAR_ALB_35_0571	Tissue	35.313	-2.661	ALB	MAR	no	yes	MCS
MED_MAR_ALB_35_0584	Tissue	35.313	-2.661	ALB	MAR	no	yes	CS
MED_MAR_ALB_35_0585	Tissue	35.313	-2.661	ALB	MAR	no	yes	SM
BLS_MAR_BMN_39_0789	Tissue	43.701	29.436	BMN	MAR	no	yes	M
BLS_MAR_BMN_39_0790	Tissue	43.701	29.436	BMN	MAR	no	yes	M
BLS_MAR_BMN_39_0791	Tissue	43.701	29.436	BMN	MAR	no	yes	MC
BLS_MAR_BMN_39_0792	Tissue	43.701	29.436	BMN	MAR	no	yes	M
BLS_MAR_BMN_39_0793	Tissue	43.701	29.436	BMN	MAR	no	yes	M
BLS_MAR_BMN_39_0794	Tissue	43.701	29.436	BMN	MAR	no	yes	M
BLS_MAR_BMN_39_0795	Tissue	43.701	29.436	BMN	MAR	no	yes	MC
BLS_MAR_BMN_39_0796	Tissue	43.701	29.436	BMN	MAR	no	yes	M
BLS_MAR_BMN_39_0797	Tissue	43.701	29.436	BMN	MAR	no	yes	M
BLS_MAR_BMN_39_0798	Tissue	43.701	29.436	BMN	MAR	no	yes	M
BLS_EST_CMN_62_1186	Tissue	45.645	36.497	CMN	COT	no	yes	C
BLS_EST_CMN_62_1456	Tissue	45.645	36.497	CMN	COT	no	yes	MC
BLS_EST_CMN_62_1182	Tissue	45.645	36.497	CMN	COT	no	yes	MC
BLS_EST_CMN_62_1183	Tissue	45.645	36.497	CMN	COT	no	yes	MC
BLS_EST_CMN_62_1184	Tissue	45.645	36.497	CMN	COT	no	yes	C
BLS_EST_CMN_62_1185	Tissue	45.645	36.497	CMN	COT	no	yes	C
ATL_MAR_CNR_88_1475	Tissue	27.934	-15.605	CNR	MAR	no	yes	S
BAL_MAR_DKB_65_1486	Tissue	54.347	11.681	DKB	MAR	no	yes	M
BAL_MAR_DKB_65_1487	Tissue	54.347	11.681	DKB	MAR	no	yes	MC
ATL_MAR_FAD_02_0950	Tissue	43.562	-1.517	GAS	MAR	no	yes	M
ATL_MAR_FAD_02_0958	Tissue	43.562	-1.517	GAS	MAR	no	yes	M
ATL_MAR_FAD_02_0961	Tissue	43.562	-1.517	GAS	MAR	no	yes	M
ATL_MAR_FAD_02_0963	Tissue	43.562	-1.517	GAS	MAR	no	yes	M
ATL_MAR_FAD_02_0974	Tissue	43.562	-1.517	GAS	MAR	no	yes	MC
ATL_MAR_FAD_02_0979	Tissue	43.562	-1.517	GAS	MAR	no	yes	M
ATL_MAR_FAD_02_0989	Tissue	43.562	-1.517	GAS	MAR	no	yes	M
ATL_EST_FAD_01_0894	Le Moan et al. (2016)	43.514	-1.494	GAS	COT	no	yes	MCS

ATL_EST_FAD_01_0896	Le Moan et al. (2016)	43.514	-1.494	GAS	COT	no	yes	CS
ATL_EST_FAD_01_0897	Le Moan et al. (2016)	43.514	-1.494	GAS	COT	no	yes	MC
ATL_EST_FAD_01_0899	Le Moan et al. (2016)	43.514	-1.494	GAS	COT	no	yes	M
ATL_EST_FAD_01_0900	Le Moan et al. (2016)	43.514	-1.494	GAS	COT	no	yes	MCS
ATL_EST_FAD_01_0902	Le Moan et al. (2016)	43.514	-1.494	GAS	COT	no	yes	M
ATL_EST_FAD_01_0903	Le Moan et al. (2016)	43.514	-1.494	GAS	COT	no	yes	C
ATL_EST_FAD_01_0904	Le Moan et al. (2016)	43.514	-1.494	GAS	COT	no	yes	M
ATL_EST_FAD_01_0905	Le Moan et al. (2016)	43.514	-1.494	GAS	COT	no	yes	MC
ATL_EST_FAD_01_0907	Le Moan et al. (2016)	43.514	-1.494	GAS	COT	no	yes	MCS
ATL_EST_FAD_01_0910	Le Moan et al. (2016)	43.514	-1.494	GAS	COT	no	yes	MCS
ATL_EST_FAD_01_0912	Le Moan et al. (2016)	43.514	-1.494	GAS	COT	no	yes	CS
ATL_EST_FAD_01_0913	Le Moan et al. (2016)	43.514	-1.494	GAS	COT	no	yes	C
ATL_EST_FAD_01_0914	Le Moan et al. (2016)	43.514	-1.494	GAS	COT	no	yes	M
ATL_EST_FAD_01_0916	Le Moan et al. (2016)	43.514	-1.494	GAS	COT	no	yes	MC
ATL_EST_FAD_01_0920	Le Moan et al. (2016)	43.514	-1.494	GAS	COT	no	yes	M
ATL_EST_FAD_01_0922	Le Moan et al. (2016)	43.514	-1.494	GAS	COT	no	yes	C
ATL_EST_FAD_01_0923	Le Moan et al. (2016)	43.514	-1.494	GAS	COT	no	yes	M
ATL_EST_FAD_01_0924	Le Moan et al. (2016)	43.514	-1.494	GAS	COT	no	yes	MC
ATL_EST_FAD_01_0925	Le Moan et al. (2016)	43.514	-1.494	GAS	COT	no	yes	M
ATL_EST_FAD_01_0926	Le Moan et al. (2016)	43.514	-1.494	GAS	COT	no	yes	M
ATL_EST_FAD_01_0929	Le Moan et al. (2016)	43.514	-1.494	GAS	COT	no	yes	M
ATL_EST_FAD_01_0930	Le Moan et al. (2016)	43.514	-1.494	GAS	COT	no	yes	MC
ATL_EST_FAD_01_0931	Le Moan et al. (2016)	43.514	-1.494	GAS	COT	no	yes	M
ATL_EST_FAD_01_0932	Le Moan et al. (2016)	43.514	-1.494	GAS	COT	no	yes	M
ATL_EST_FAD_01_0934	Le Moan et al. (2016)	43.514	-1.494	GAS	COT	no	yes	MC
ATL_EST_FAD_01_0935	Le Moan et al. (2016)	43.514	-1.494	GAS	COT	no	yes	MCS
ATL_EST_FAD_01_0937	Le Moan et al. (2016)	43.514	-1.494	GAS	COT	no	yes	MC
ATL_EST_FAD_01_0938	Le Moan et al. (2016)	43.514	-1.494	GAS	COT	no	yes	M
ATL_EST_FAD_01_0939	Le Moan et al. (2016)	43.514	-1.494	GAS	COT	no	yes	MC
ATL_EST_FAD_66_1139	Le Moan et al. (2016)	43.526	-1.506	GAS	COT	no	yes	MC
ATL_EST_FAD_66_1145	Le Moan et al. (2016)	43.526	-1.506	GAS	COT	no	yes	MCS
ATL_EST_FAD_66_1147	Le Moan et al. (2016)	43.526	-1.506	GAS	COT	no	yes	CS
ATL_EST_FAD_66_1148	Le Moan et al. (2016)	43.526	-1.506	GAS	COT	no	yes	M
ATL_EST_FAD_66_1149	Le Moan et al. (2016)	43.526	-1.506	GAS	COT	no	yes	M
ATL_EST_FAD_66_1150	Le Moan et al. (2016)	43.526	-1.506	GAS	COT	no	yes	M
ATL_EST_FAD_66_1151	Le Moan et al. (2016)	43.526	-1.506	GAS	COT	no	yes	CS
ATL_EST_FAD_66_1153	Le Moan et al. (2016)	43.526	-1.506	GAS	COT	no	yes	MCS
ATL_MAR_GAS_90_1511	Le Moan et al. (2016)	44.710	-1.480	GAS	MAR	no	yes	MC
ATL_MAR_GAS_90_1512	Le Moan et al. (2016)	44.710	-1.480	GAS	MAR	no	yes	M
ATL_MAR_GAS_90_1513	Le Moan et al. (2016)	44.710	-1.480	GAS	MAR	no	yes	M
ATL_MAR_GAS_90_1514	Le Moan et al. (2016)	44.710	-1.480	GAS	MAR	no	yes	MCS
ATL_MAR_GAS_90_1515	Le Moan et al. (2016)	44.710	-1.480	GAS	MAR	no	yes	M
ATL_MAR_GAS_90_1516	Le Moan et al. (2016)	44.710	-1.480	GAS	MAR	no	yes	M
ATL_MAR_GAS_90_1517	Le Moan et al. (2016)	44.710	-1.480	GAS	MAR	no	yes	M
ATL_MAR_GAS_90_1518	Le Moan et al. (2016)	44.710	-1.480	GAS	MAR	no	yes	M
ATL_MAR_GAS_18_0621	Le Moan et al. (2016)	45.507	-2.869	GAS	MAR	no	yes	M
ATL_MAR_GAS_18_0622	Le Moan et al. (2016)	45.507	-2.869	GAS	MAR	no	yes	M
ATL_MAR_GAS_18_0624	Le Moan et al. (2016)	45.507	-2.869	GAS	MAR	no	yes	M
ATL_MAR_GAS_18_0625	Le Moan et al. (2016)	45.507	-2.869	GAS	MAR	no	yes	M
ATL_MAR_GAS_18_0634	Le Moan et al. (2016)	45.507	-2.869	GAS	MAR	no	yes	M
ATL_MAR_GAS_18_0644	Le Moan et al. (2016)	45.507	-2.869	GAS	MAR	no	yes	M
ATL_MAR_GAS_18_0645	Le Moan et al. (2016)	45.507	-2.869	GAS	MAR	no	yes	M

ATL_MAR_GAS_18_0646	Le Moan et al. (2016)	45.507	-2.869	GAS	MAR	no	yes	M
ATL_MAR_GAS_18_0647	Le Moan et al. (2016)	45.507	-2.869	GAS	MAR	no	yes	M
ATL_MAR_GAS_18_0648	Le Moan et al. (2016)	45.507	-2.869	GAS	MAR	no	yes	M
ATL_MAR_GAS_25_0853	Le Moan et al. (2016)	44.465	-1.514	GAS	MAR	no	yes	M
ATL_MAR_GAS_25_0856	Le Moan et al. (2016)	44.465	-1.514	GAS	MAR	no	yes	M
ATL_MAR_GAS_25_0857	Le Moan et al. (2016)	44.465	-1.514	GAS	MAR	no	yes	M
ATL_MAR_GAS_25_0858	Le Moan et al. (2016)	44.465	-1.514	GAS	MAR	no	yes	M
ATL_MAR_GAS_25_0864	Le Moan et al. (2016)	44.465	-1.514	GAS	MAR	no	yes	M
ATL_MAR_GAS_25_0865	Le Moan et al. (2016)	44.465	-1.514	GAS	MAR	no	yes	M
ATL_MAR_GAS_06_1103	Tissue	45.364	-1.594	GAS	MAR	no	yes	M
ATL_MAR_GAS_06_1098	Tissue	45.364	-1.594	GAS	MAR	no	yes	M
ATL_MAR_GAS_06_1099	Tissue	45.364	-1.594	GAS	MAR	no	yes	M
ATL_MAR_GAS_06_1100	Tissue	45.364	-1.594	GAS	MAR	no	yes	MC
ATL_MAR_GAS_06_1102	Tissue	45.364	-1.594	GAS	MAR	no	yes	M
ATL_EST_GAS_70_1199	Tissue	45.400	-1.380	GAS	COT	no	yes	M
ATL_EST_GAS_70_1200	Tissue	45.400	-1.380	GAS	COT	no	yes	M
ATL_EST_GAS_70_1201	Tissue	45.400	-1.380	GAS	COT	no	yes	M
ATL_EST_GAS_70_1202	Tissue	45.400	-1.380	GAS	COT	no	yes	M
ATL_EST_GAS_70_1203	Tissue	45.400	-1.380	GAS	COT	no	yes	M
ATL_EST_GAS_70_1204	Tissue	45.400	-1.380	GAS	COT	no	yes	M
ATL_EST_GAS_70_1205	Tissue	45.400	-1.380	GAS	COT	no	yes	M
ATL_EST_GAS_70_1206	Tissue	45.400	-1.380	GAS	COT	no	yes	M
ATL_MAR_FAD_02_0995	Tissue	43.562	-1.517	GAS	MAR	no	yes	MCS
ATL_MAR_FAD_02_0969	Tissue	43.562	-1.517	GAS	MAR	no	yes	M
ATL_MAR_FAD_02_0970	Tissue	43.562	-1.517	GAS	MAR	no	yes	M
ATL_MAR_FAD_02_0977	Tissue	43.562	-1.517	GAS	MAR	no	yes	M
ATL_MAR_FAD_02_0978	Tissue	43.562	-1.517	GAS	MAR	no	yes	M
ATL_EST_GAS_71_1207	Tissue	45.400	-1.380	GAS	COT	no	yes	M
ATL_EST_GAS_71_1208	Tissue	45.400	-1.380	GAS	COT	no	yes	M
ATL_EST_GAS_71_1209	Tissue	45.400	-1.380	GAS	COT	no	yes	M
ATL_EST_GAS_71_1210	Tissue	45.400	-1.380	GAS	COT	no	yes	M
ATL_EST_GAS_71_1211	Tissue	45.400	-1.380	GAS	COT	no	yes	M
ATL_EST_GAS_71_1212	Tissue	45.400	-1.380	GAS	COT	no	yes	M
ATL_EST_GAS_71_1213	Tissue	45.400	-1.380	GAS	COT	no	yes	M
ATL_EST_GAS_71_1214	Tissue	45.400	-1.380	GAS	COT	no	yes	M
ATL_MAR_GAS_22_0712	Tissue	44.667	-1.575	GAS	MAR	no	yes	M
ATL_MAR_GAS_22_0717	Tissue	44.667	-1.575	GAS	MAR	no	yes	M
ATL_MAR_GAS_22_0733	Tissue	44.667	-1.575	GAS	MAR	no	yes	M
ATL_MAR_GAS_22_0734	Tissue	44.667	-1.575	GAS	MAR	no	yes	M
ATL_MAR_GAS_22_0743	Tissue	44.667	-1.575	GAS	MAR	no	yes	M
ATL_MAR_GAS_22_0750	Tissue	44.667	-1.575	GAS	MAR	no	yes	M
ATL_MAR_GAS_22_0754	Tissue	44.667	-1.575	GAS	MAR	no	yes	M
ATL_MAR_GAS_73_1254	Tissue	43.477	-1.630	GAS	MAR	no	yes	M
ATL_MAR_GAS_73_1255	Tissue	43.477	-1.630	GAS	MAR	no	yes	M
ATL_MAR_GAS_73_1256	Tissue	43.477	-1.630	GAS	MAR	no	yes	M
ATL_MAR_GAS_73_1257	Tissue	43.477	-1.630	GAS	MAR	no	yes	MC
ATL_MAR_GAS_73_1258	Tissue	43.477	-1.630	GAS	MAR	no	yes	MC
ATL_MAR_GAS_73_1260	Tissue	43.477	-1.630	GAS	MAR	no	yes	M
ATL_MAR_GAS_73_1261	Tissue	43.477	-1.630	GAS	MAR	no	yes	M
ATL_MAR_GAS_73_1263	Tissue	43.477	-1.630	GAS	MAR	no	yes	M
ATL_MAR_GAS_73_1246	Tissue	43.477	-1.630	GAS	MAR	no	yes	M
ATL_MAR_GAS_73_1265	Tissue	43.477	-1.630	GAS	MAR	no	yes	M

ATL_MAR_GAS_73_1266	Tissue	43.477	-1.630	GAS	MAR	no	yes	M
ATL_MAR_GAS_73_1264	Tissue	43.477	-1.630	GAS	MAR	no	yes	M
ATL_MAR_GAS_73_1247	Tissue	43.477	-1.630	GAS	MAR	no	yes	M
ATL_MAR_GAS_73_1248	Tissue	43.477	-1.630	GAS	MAR	no	yes	MC
ATL_MAR_GAS_73_1249	Tissue	43.477	-1.630	GAS	MAR	no	yes	M
ATL_MAR_GAS_73_1250	Tissue	43.477	-1.630	GAS	MAR	no	yes	M
ATL_MAR_GAS_73_1251	Tissue	43.477	-1.630	GAS	MAR	no	yes	M
ATL_MAR_GAS_73_1252	Tissue	43.477	-1.630	GAS	MAR	no	yes	M
ATL_MAR_GAS_73_1253	Tissue	43.477	-1.630	GAS	MAR	no	yes	M
ATL_MAR_GAS_74_1277	Tissue	43.471	-1.658	GAS	MAR	no	yes	M
ATL_MAR_GAS_74_1278	Tissue	43.471	-1.658	GAS	MAR	no	yes	M
ATL_MAR_GAS_74_1279	Tissue	43.471	-1.658	GAS	MAR	no	yes	M
ATL_MAR_GAS_74_1280	Tissue	43.471	-1.658	GAS	MAR	no	yes	M
ATL_MAR_GAS_74_1281	Tissue	43.471	-1.658	GAS	MAR	no	yes	M
ATL_MAR_GAS_74_1282	Tissue	43.471	-1.658	GAS	MAR	no	yes	MC
ATL_MAR_GAS_74_1284	Tissue	43.471	-1.658	GAS	MAR	no	yes	M
ATL_MAR_GAS_74_1285	Tissue	43.471	-1.658	GAS	MAR	no	yes	MC
ATL_MAR_GAS_74_1267	Tissue	43.471	-1.658	GAS	MAR	no	yes	M
ATL_MAR_GAS_74_1286	Tissue	43.471	-1.658	GAS	MAR	no	yes	M
ATL_MAR_GAS_74_1287	Tissue	43.471	-1.658	GAS	MAR	no	yes	M
ATL_MAR_GAS_74_1288	Tissue	43.471	-1.658	GAS	MAR	no	yes	MC
ATL_MAR_GAS_74_1291	Tissue	43.471	-1.658	GAS	MAR	no	yes	M
ATL_MAR_GAS_74_1292	Tissue	43.471	-1.658	GAS	MAR	no	yes	M
ATL_MAR_GAS_74_1294	Tissue	43.471	-1.658	GAS	MAR	no	yes	M
ATL_MAR_GAS_74_1268	Tissue	43.471	-1.658	GAS	MAR	no	yes	M
ATL_MAR_GAS_74_1296	Tissue	43.471	-1.658	GAS	MAR	no	yes	MC
ATL_MAR_GAS_74_1298	Tissue	43.471	-1.658	GAS	MAR	no	yes	M
ATL_MAR_GAS_74_1299	Tissue	43.471	-1.658	GAS	MAR	no	yes	M
ATL_MAR_GAS_74_1300	Tissue	43.471	-1.658	GAS	MAR	no	yes	M
ATL_MAR_GAS_74_1301	Tissue	43.471	-1.658	GAS	MAR	no	yes	MC
ATL_MAR_GAS_74_1302	Tissue	43.471	-1.658	GAS	MAR	no	yes	M
ATL_MAR_GAS_74_1304	Tissue	43.471	-1.658	GAS	MAR	no	yes	M
ATL_MAR_GAS_74_1305	Tissue	43.471	-1.658	GAS	MAR	no	yes	M
ATL_MAR_GAS_74_1269	Tissue	43.471	-1.658	GAS	MAR	no	yes	M
ATL_MAR_GAS_74_1306	Tissue	43.471	-1.658	GAS	MAR	no	yes	M
ATL_MAR_GAS_74_1308	Tissue	43.471	-1.658	GAS	MAR	no	yes	M
ATL_MAR_GAS_74_1309	Tissue	43.471	-1.658	GAS	MAR	no	yes	M
ATL_MAR_GAS_74_1310	Tissue	43.471	-1.658	GAS	MAR	no	yes	M
ATL_MAR_GAS_74_1311	Tissue	43.471	-1.658	GAS	MAR	no	yes	M
ATL_MAR_GAS_74_1312	Tissue	43.471	-1.658	GAS	MAR	no	yes	M
ATL_MAR_GAS_74_1313	Tissue	43.471	-1.658	GAS	MAR	no	yes	MC
ATL_MAR_GAS_74_1314	Tissue	43.471	-1.658	GAS	MAR	no	yes	M
ATL_MAR_GAS_74_1315	Tissue	43.471	-1.658	GAS	MAR	no	yes	M
ATL_MAR_GAS_74_1270	Tissue	43.471	-1.658	GAS	MAR	no	yes	M
ATL_MAR_GAS_74_1316	Tissue	43.471	-1.658	GAS	MAR	no	yes	M
ATL_MAR_GAS_74_1271	Tissue	43.471	-1.658	GAS	MAR	no	yes	M
ATL_MAR_GAS_74_1272	Tissue	43.471	-1.658	GAS	MAR	no	yes	M
ATL_MAR_GAS_74_1273	Tissue	43.471	-1.658	GAS	MAR	no	yes	M
ATL_MAR_GAS_74_1274	Tissue	43.471	-1.658	GAS	MAR	no	yes	MC
ATL_MAR_GAS_74_1275	Tissue	43.471	-1.658	GAS	MAR	no	yes	M
ATL_MAR_GAS_75_1327	Tissue	43.467	-1.671	GAS	MAR	no	yes	M
ATL_MAR_GAS_75_1333	Tissue	43.467	-1.671	GAS	MAR	no	yes	M

ATL_MAR_GAS_75_1338	Tissue	43.467	-1.671	GAS	MAR	no	yes	M
ATL_MAR_GAS_75_1340	Tissue	43.467	-1.671	GAS	MAR	no	yes	M
ATL_MAR_GAS_76_1359	Tissue	43.466	-1.674	GAS	MAR	no	yes	M
ATL_MAR_GAS_76_1360	Tissue	43.466	-1.674	GAS	MAR	no	yes	MC
ATL_MAR_GAS_76_1361	Tissue	43.466	-1.674	GAS	MAR	no	yes	M
ATL_MAR_GAS_76_1362	Tissue	43.466	-1.674	GAS	MAR	no	yes	M
ATL_MAR_GAS_76_1363	Tissue	43.466	-1.674	GAS	MAR	no	yes	M
ATL_MAR_GAS_76_1364	Tissue	43.466	-1.674	GAS	MAR	no	yes	M
ATL_MAR_GAS_76_1365	Tissue	43.466	-1.674	GAS	MAR	no	yes	M
ATL_MAR_GAS_76_1366	Tissue	43.466	-1.674	GAS	MAR	no	yes	M
ATL_MAR_GAS_76_1367	Tissue	43.466	-1.674	GAS	MAR	no	yes	M
ATL_MAR_GAS_76_1369	Tissue	43.466	-1.674	GAS	MAR	no	yes	M
ATL_MAR_GAS_76_1374	Tissue	43.466	-1.674	GAS	MAR	no	yes	M
ATL_MAR_GAS_76_1377	Tissue	43.466	-1.674	GAS	MAR	no	yes	M
ATL_MAR_GAS_76_1350	Tissue	43.466	-1.674	GAS	MAR	no	yes	M
ATL_MAR_GAS_77_1399	Tissue	43.484	-1.650	GAS	MAR	no	yes	M
ATL_MAR_GAS_76_1378	Tissue	43.466	-1.674	GAS	MAR	no	yes	M
ATL_MAR_GAS_76_1381	Tissue	43.466	-1.674	GAS	MAR	no	yes	M
ATL_MAR_GAS_76_1382	Tissue	43.466	-1.674	GAS	MAR	no	yes	M
ATL_MAR_GAS_76_1351	Tissue	43.466	-1.674	GAS	MAR	no	yes	M
ATL_MAR_GAS_76_1352	Tissue	43.466	-1.674	GAS	MAR	no	yes	M
ATL_MAR_GAS_76_1385	Tissue	43.466	-1.674	GAS	MAR	no	yes	M
ATL_MAR_GAS_76_1389	Tissue	43.466	-1.674	GAS	MAR	no	yes	M
ATL_MAR_GAS_76_1354	Tissue	43.466	-1.674	GAS	MAR	no	yes	M
ATL_MAR_GAS_76_1356	Tissue	43.466	-1.674	GAS	MAR	no	yes	M
ATL_MAR_GAS_77_1407	Tissue	43.484	-1.650	GAS	MAR	no	yes	M
ATL_MAR_GAS_77_1392	Tissue	43.484	-1.650	GAS	MAR	no	yes	M
ATL_MAR_GAS_77_1408	Tissue	43.484	-1.650	GAS	MAR	no	yes	M
ATL_MAR_GAS_77_1415	Tissue	43.484	-1.650	GAS	MAR	no	yes	MC
ATL_MAR_GAS_77_1419	Tissue	43.484	-1.650	GAS	MAR	no	yes	M
ATL_MAR_GAS_77_1397	Tissue	43.484	-1.650	GAS	MAR	no	yes	M
ATL_EST_FAD_78_1421	Tissue	43.525	-1.506	GAS	COT	no	yes	M
ATL_EST_FAD_78_1423	Tissue	43.525	-1.506	GAS	COT	no	yes	M
ATL_EST_FAD_78_1424	Tissue	43.525	-1.506	GAS	COT	no	yes	M
ATL_MAR_GAS_25_0821	Tissue	44.465	-1.514	GAS	MAR	no	yes	M
ATL_MAR_GAS_25_0830	Tissue	44.465	-1.514	GAS	MAR	no	yes	M
ATL_MAR_GAS_25_0835	Tissue	44.465	-1.514	GAS	MAR	no	yes	M
ATL_MAR_GAS_25_0838	Tissue	44.465	-1.514	GAS	MAR	no	yes	M
ATL_MAR_GAS_25_0840	Tissue	44.465	-1.514	GAS	MAR	no	yes	M
ATL_MAR_GAS_25_0866	Tissue	44.465	-1.514	GAS	MAR	no	yes	M
ATL_MAR_FAD_02_0953	Tissue	43.562	-1.517	GAS	MAR	yes	yes	M
ATL_MAR_FAD_02_0957	Tissue	43.562	-1.517	GAS	MAR	yes	yes	M
ATL_MAR_FAD_02_0968	Tissue	43.562	-1.517	GAS	MAR	yes	yes	M
ATL_EST_FAD_01_0901	Le Moan et al. (2016)	43.514	-1.494	GAS	COT	yes	yes	C
ATL_EST_FAD_01_0911	Le Moan et al. (2016)	43.514	-1.494	GAS	COT	yes	yes	C
ATL_MAR_GAS_73_1259	Tissue	43.477	-1.630	GAS	COT	yes	yes	M
ATL_MAR_GAS_73_1262	Tissue	43.477	-1.630	GAS	COT	yes	yes	M
ATL_MAR_FAD_02_0947	Tissue	43.437	-0.660	GAS	COT	yes	yes	CS
ATL_MAR_FAD_02_0948	Tissue	43.437	-0.660	GAS	COT	yes	yes	C
ATL_MAR_FAD_02_0966	Tissue	43.437	-0.660	GAS	COT	yes	yes	CS
MED_LAG_GDL_82_1519	Le Moan et al. (2016)	43.576	4.018	GDL	COT	no	yes	C
MED_LAG_GDL_82_1520	Le Moan et al. (2016)	43.576	4.018	GDL	COT	no	yes	C

MED_LAG_GDL_82_1521	Le Moan et al. (2016)	43.576	4.018	GDL	COT	no	yes	C
MED_LAG_GDL_82_1522	Le Moan et al. (2016)	43.576	4.018	GDL	COT	no	yes	C
MED_LAG_GDL_82_1523	Le Moan et al. (2016)	43.576	4.018	GDL	COT	no	yes	C
MED_LAG_GDL_82_1524	Le Moan et al. (2016)	43.576	4.018	GDL	COT	no	yes	C
MED_LAG_GDL_82_1525	Le Moan et al. (2016)	43.576	4.018	GDL	COT	no	yes	C
MED_LAG_GDL_82_1526	Le Moan et al. (2016)	43.576	4.018	GDL	COT	no	yes	C
MED_LAG_GDL_82_1527	Le Moan et al. (2016)	43.576	4.018	GDL	COT	no	yes	C
MED_LAG_GDL_82_1528	Le Moan et al. (2016)	43.576	4.018	GDL	COT	no	yes	C
MED_LAG_GDL_82_1529	Le Moan et al. (2016)	43.576	4.018	GDL	COT	no	yes	C
MED_LAG_GDL_82_1530	Le Moan et al. (2016)	43.576	4.018	GDL	COT	no	yes	C
MED_LAG_GDL_82_1531	Le Moan et al. (2016)	43.576	4.018	GDL	COT	no	yes	C
MED_LAG_GDL_82_1532	Le Moan et al. (2016)	43.576	4.018	GDL	COT	no	yes	C
MED_LAG_GDL_82_1533	Le Moan et al. (2016)	43.576	4.018	GDL	COT	no	yes	C
MED_LAG_GDL_82_1534	Le Moan et al. (2016)	43.576	4.018	GDL	COT	no	yes	C
MED_LAG_GDL_82_1535	Le Moan et al. (2016)	43.576	4.018	GDL	COT	no	yes	C
MED_LAG_GDL_82_1536	Le Moan et al. (2016)	43.576	4.018	GDL	COT	no	yes	C
MED_LAG_GDL_82_1537	Le Moan et al. (2016)	43.576	4.018	GDL	COT	no	yes	C
MED_LAG_GDL_82_1538	Le Moan et al. (2016)	43.576	4.018	GDL	COT	no	yes	C
MED_LAG_GDL_83_1134	Le Moan et al. (2016)	43.576	4.018	GDL	COT	no	yes	C
MED_LAG_GDL_83_1135	Le Moan et al. (2016)	43.576	4.018	GDL	COT	no	yes	C
MED_LAG_GDL_83_1136	Le Moan et al. (2016)	43.576	4.018	GDL	COT	no	yes	C
MED_LAG_GDL_83_1539	Le Moan et al. (2016)	43.576	4.018	GDL	COT	no	yes	C
MED_LAG_GDL_82_1540	Le Moan et al. (2016)	43.576	4.018	GDL	COT	no	yes	C
MED_LAG_GDL_82_1541	Le Moan et al. (2016)	43.576	4.018	GDL	COT	no	yes	C
MED_LAG_GDL_82_1542	Le Moan et al. (2016)	43.576	4.018	GDL	COT	no	yes	C
MED_LAG_GDL_82_1543	Le Moan et al. (2016)	43.576	4.018	GDL	COT	no	yes	C
MED_LAG_GDL_82_1137	Le Moan et al. (2016)	43.576	4.018	GDL	COT	no	yes	C
MED_LAG_GDL_82_1544	Le Moan et al. (2016)	43.576	4.018	GDL	COT	no	yes	C
MED_LAG_GDL_82_1545	Le Moan et al. (2016)	43.576	4.018	GDL	COT	no	yes	C
MED_LAG_GDL_82_1546	Le Moan et al. (2016)	43.576	4.018	GDL	COT	no	yes	MC
MED_MAR_GDL_81_1547	Le Moan et al. (2016)	43.199	3.613	GDL	MAR	no	yes	MC
MED_MAR_GDL_81_1548	Le Moan et al. (2016)	43.199	3.613	GDL	MAR	no	yes	M
MED_MAR_GDL_81_1124	Le Moan et al. (2016)	43.199	3.613	GDL	MAR	no	yes	MC
MED_MAR_GDL_81_1549	Le Moan et al. (2016)	43.199	3.613	GDL	MAR	no	yes	MC
MED_MAR_GDL_81_1550	Le Moan et al. (2016)	43.199	3.613	GDL	MAR	no	yes	MC
MED_MAR_GDL_81_1551	Le Moan et al. (2016)	43.199	3.613	GDL	MAR	no	yes	MC
MED_MAR_GDL_81_1125	Le Moan et al. (2016)	43.199	3.613	GDL	MAR	no	yes	MC
MED_MAR_GDL_81_1127	Le Moan et al. (2016)	43.199	3.613	GDL	MAR	no	yes	MC
MED_MAR_GDL_81_1128	Le Moan et al. (2016)	43.199	3.613	GDL	MAR	no	yes	M
MED_MAR_GDL_81_1130	Le Moan et al. (2016)	43.199	3.613	GDL	MAR	no	yes	M
MED_MAR_GDL_81_1131	Le Moan et al. (2016)	43.199	3.613	GDL	MAR	no	yes	MC
MED_MAR_GDL_81_1133	Le Moan et al. (2016)	43.199	3.613	GDL	MAR	no	yes	M
MED_MAR_GDL_81_0053	Le Moan et al. (2016)	43.329	3.820	GDL	MAR	no	yes	MC
MED_MAR_GDL_81_0056	Le Moan et al. (2016)	43.329	3.820	GDL	MAR	no	yes	M
MED_MAR_GDL_81_0057	Le Moan et al. (2016)	43.329	3.820	GDL	MAR	no	yes	M
MED_MAR_GDL_81_0058	Le Moan et al. (2016)	43.329	3.820	GDL	MAR	no	yes	MC
MED_MAR_GDL_81_0060	Le Moan et al. (2016)	43.329	3.820	GDL	MAR	no	yes	M
MED_MAR_GDL_81_0061	Le Moan et al. (2016)	43.329	3.820	GDL	MAR	no	yes	M
MED_MAR_GDL_81_0062	Le Moan et al. (2016)	43.329	3.820	GDL	MAR	no	yes	M
MED_MAR_GDL_81_0063	Le Moan et al. (2016)	43.329	3.820	GDL	MAR	no	yes	M
MED_MAR_GDL_81_0066	Le Moan et al. (2016)	43.329	3.820	GDL	MAR	no	yes	MC
MED_MAR_GDL_81_0069	Le Moan et al. (2016)	43.329	3.820	GDL	MAR	no	yes	M

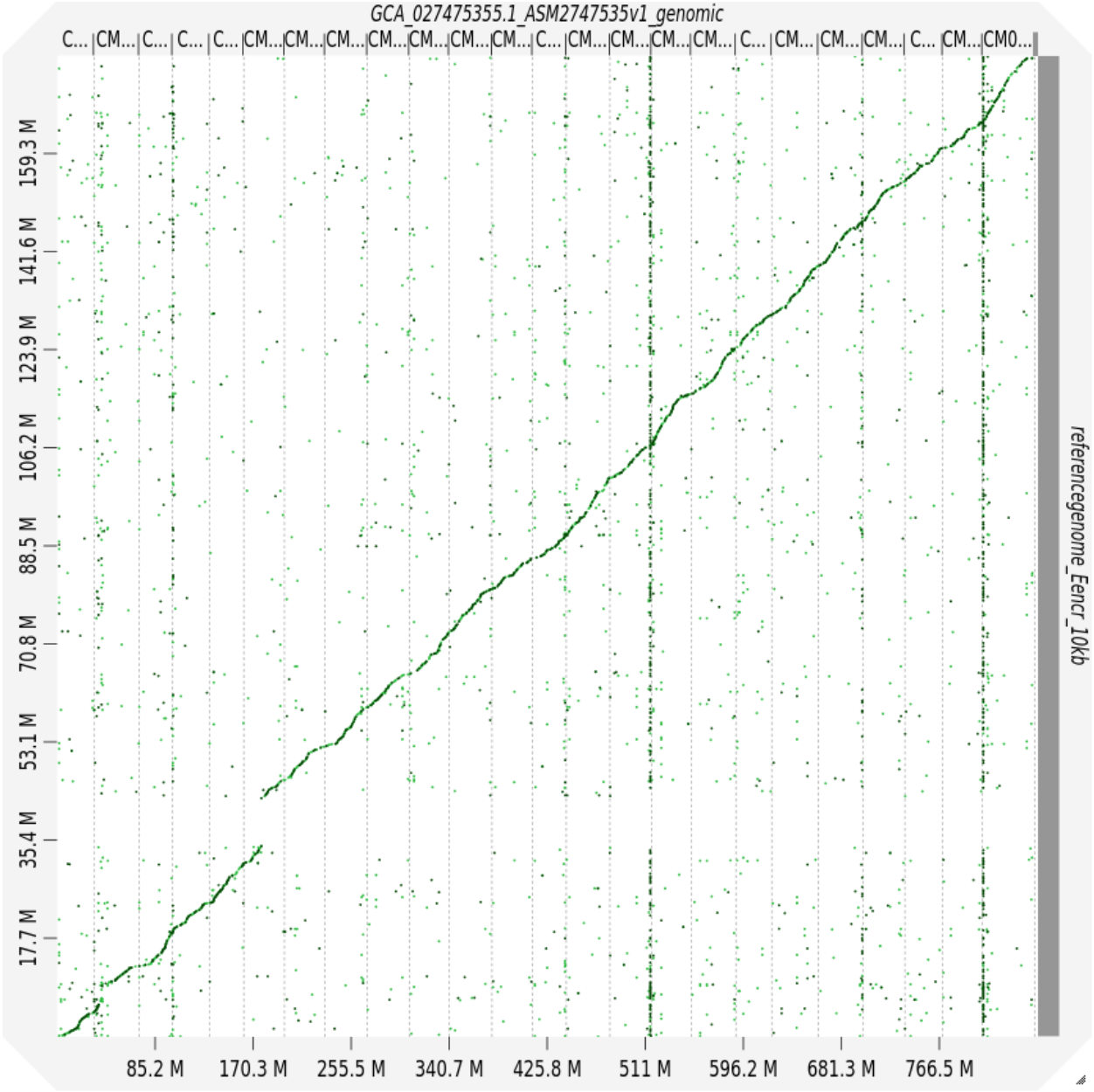
MED_MAR_GDL_81_0071	Le Moan et al. (2016)	43.329	3.820	GDL	MAR	no	yes	M
MED_MAR_GDL_81_0072	Le Moan et al. (2016)	43.329	3.820	GDL	MAR	no	yes	M
MED_MAR_GDL_81_0074	Le Moan et al. (2016)	43.329	3.820	GDL	MAR	no	yes	MCS
MED_MAR_GDL_81_0075	Le Moan et al. (2016)	43.329	3.820	GDL	MAR	no	yes	MC
MED_MAR_GDL_81_0078	Le Moan et al. (2016)	43.329	3.820	GDL	MAR	no	yes	M
MED_MAR_GDL_81_0079	Le Moan et al. (2016)	43.329	3.820	GDL	MAR	no	yes	M
MED_MAR_GDL_81_1122	Le Moan et al. (2016)	43.199	3.613	GDL	MAR	yes	yes	MC
MED_MAR_GDL_81_1129	Le Moan et al. (2016)	43.199	3.613	GDL	MAR	yes	yes	M
MED_MAR_GDL_81_1132	Le Moan et al. (2016)	43.199	3.613	GDL	MAR	yes	yes	M
Eencr_1L	Tissue	43.528	3.886	GDL	COT	yes	no	C
Eencr_2L	Tissue	43.528	3.886	GDL	COT	yes	no	C
EencrLi1	Tissue	43.199	3.613	GDL	MAR	yes	no	MC
MED_MAR_GDL_81_1123	Le Moan et al. (2016)	43.199	3.613	GDL	MAR	yes	yes	M
EencrLi4	Tissue	43.528	3.886	GDL	COT	yes	no	C
EencrLi5	Tissue	43.528	3.886	GDL	COT	yes	no	MCS
EencrLi6	Tissue	43.528	3.886	GDL	COT	yes	no	C
ATL_MAR_MSA_57_1164	Tissue	21.998	-16.965	MA1	MAR	no	yes	S
ATL_MAR_MSA_57_1165	Tissue	21.998	-16.965	MA1	MAR	no	yes	S
ATL_MAR_MSA_57_1169	Tissue	21.998	-16.965	MA1	MAR	no	yes	S
ATL_MAR_MSA_58_1425	Tissue	24.154	-15.997	MA2	MAR	no	yes	S
ATL_MAR_MSA_58_1426	Tissue	24.154	-15.997	MA2	MAR	no	yes	S
ATL_MAR_MSA_58_1427	Tissue	24.154	-15.997	MA2	MAR	no	yes	S
ATL_MAR_MSA_58_1428	Tissue	24.154	-15.997	MA2	MAR	no	yes	S
ATL_MAR_MSA_59_1445	Tissue	28.685	-11.218	MA3	MAR	no	yes	SM
ATL_MAR_MSA_59_1449	Tissue	28.685	-11.218	MA3	MAR	no	yes	MCS
ATL_MAR_MSA_59_1450	Tissue	28.685	-11.218	MA3	MAR	no	yes	MCS
ATL_MAR_MNA_56_1170	Tissue	33.678	-7.698	MA4	MAR	no	yes	MCS
ATL_MAR_MNA_56_1171	Tissue	33.678	-7.698	MA4	MAR	no	yes	CS
ATL_MAR_MNA_56_1172	Tissue	33.678	-7.698	MA4	MAR	no	yes	MCS
ATL_MAR_MNA_56_1173	Tissue	33.678	-7.698	MA4	MAR	no	yes	MCS
ATL_MAR_MDN_63_1215	Tissue	50.786	1.479	MDN	MAR	no	yes	MC
ATL_MAR_MDN_63_1216	Tissue	50.786	1.479	MDN	MAR	no	yes	MC
ATL_MAR_MDN_63_1217	Tissue	50.786	1.479	MDN	MAR	no	yes	MC
ATL_MAR_MDN_63_1218	Tissue	50.786	1.479	MDN	MAR	no	yes	MC
ATL_EST_MDN_91_1510	Tissue	51.500	0.640	MDN	COT	no	yes	C
ATL_EST_MDN_91_1506	Tissue	51.500	0.640	MDN	COT	no	yes	MC
ATL_EST_MDN_91_1507	Tissue	51.500	0.640	MDN	COT	no	yes	MC
ATL_EST_MDN_91_1508	Tissue	51.500	0.640	MDN	COT	no	yes	MC
ATL_EST_MDN_91_1509	Tissue	51.500	0.640	MDN	COT	no	yes	C
ATL_EST_NOR_86_1459	Tissue	59.702	10.551	NOR	COT	no	yes	MC
ATL_EST_NOR_86_1460	Tissue	59.702	10.551	NOR	COT	no	yes	MC
BAL_MAR_PLN_85_1461	Tissue	55.238	18.058	PLN	MAR	no	yes	MC
BAL_MAR_PLN_85_1462	Tissue	55.238	18.058	PLN	MAR	no	yes	MC
BAL_MAR_PLN_85_1463	Tissue	55.238	18.058	PLN	MAR	no	yes	MC
ATL_MAR_PRS_89_1470	Tissue	36.945	-8.550	PRS	MAR	no	yes	MCS
ATL_MAR_PRS_89_1471	Tissue	36.945	-8.550	PRS	MAR	yes	yes	MCS
ATL_MAR_PRS_89_1472	Tissue	36.945	-8.550	PRS	MAR	yes	yes	MCS
ATL_MAR_PRS_89_1473	Tissue	36.945	-8.550	PRS	MAR	yes	no	MCS
ATL_EST_PRS_84_1467	Tissue	37.028	-7.812	PRS	COT	yes	yes	CS
ATL_EST_PRS_84_1468	Tissue	37.028	-7.812	PRS	COT	yes	yes	MCS
ATL_EST_PRS_84_1469	Tissue	37.028	-7.812	PRS	COT	yes	yes	MCS
ATL_EST_PRS_72_1453	Tissue	37.029	-8.003	PRS	COT	yes	yes	C

ATL_EST_PRS_72_1454	Tissue	37.029	-8.003	PRS	COT	yes	yes	MCS
ATL_EST_PRS_72_1455	Tissue	37.029	-8.003	PRS	COT	yes	yes	MCS
EencrFa1	Tissue	36.945	-8.550	PRS	MAR	yes	no	MCS
EencrFa2	Tissue	36.945	-8.550	PRS	MAR	yes	no	CS
MED_LAG_SIC_68_1233	Tissue	38.269	15.637	SIC	COT	no	yes	C
MED_LAG_SIC_68_1234	Tissue	38.269	15.637	SIC	COT	no	yes	C
MED_LAG_SIC_68_1235	Tissue	38.269	15.637	SIC	COT	no	yes	C
MED_LAG_SIC_68_1236	Tissue	38.269	15.637	SIC	COT	no	yes	C
MED_MAR_SIC_67_1227	Tissue	38.148	15.595	SIC	MAR	no	yes	M
MED_MAR_SIC_67_1228	Tissue	38.148	15.595	SIC	MAR	no	yes	MC
MED_MAR_SIC_67_1229	Tissue	38.148	15.595	SIC	MAR	no	yes	SM
MED_MAR_SIC_67_1230	Tissue	38.148	15.595	SIC	MAR	no	yes	M
EencrMu2	Tissue	37.970	-0.682	SPN	MAR	yes	no	MC
EencrMu3	Tissue	37.970	-0.682	SPN	MAR	yes	no	M
EencrMu4	Tissue	37.970	-0.682	SPN	MAR	yes	no	MCS
EencrMu5	Tissue	37.970	-0.682	SPN	MAR	yes	no	MCS
EencrMu6	Tissue	37.970	-0.682	SPN	MAR	yes	no	MCS
MED_MAR_TNO_40_0187	Tissue	37.067	9.000	TNO	MAR	no	yes	M
MED_MAR_TNO_40_0196	Tissue	37.067	9.000	TNO	MAR	no	yes	M
MED_MAR_TNO_40_0197	Tissue	37.067	9.000	TNO	MAR	no	yes	M
MED_LAG_TNO_53_0388	Tissue	37.183	9.850	TNO	COT	no	yes	C
MED_LAG_TNO_55_1075	Tissue	37.167	9.667	TNO	COT	no	yes	C
MED_LAG_TNO_55_1076	Tissue	37.167	9.667	TNO	COT	no	yes	C
MED_LAG_TNO_53_0405	Tissue	37.183	9.850	TNO	COT	no	yes	C
MED_LAG_TNO_55_1079	Tissue	37.167	9.667	TNO	COT	no	yes	C
MED_LAG_TNO_55_1081	Tissue	37.167	9.667	TNO	COT	no	yes	C
ATL_MAR_ZDA_60_1154	Tissue	-34.530	25.660	ZDA	MAR	no	yes	S
ATL_MAR_ZDA_60_1156	Tissue	-34.530	25.660	ZDA	MAR	no	yes	S
ATL_MAR_ZDA_60_1157	Tissue	-34.530	25.660	ZDA	MAR	no	yes	S
ATL_MAR_ZDA_61_1160	Tissue	-34.530	25.660	ZDA	MAR	no	yes	S
ATL_MAR_ZDA_61_1161	Tissue	-34.530	25.660	ZDA	MAR	no	yes	S
ATL_MAR_ZDA_60_1155	Tissue	-34.530	25.660	ZDA	MAR	yes	yes	S
ATL_MAR_ZDA_61_1162	Tissue	-34.530	25.660	ZDA	MAR	yes	yes	S
ATL_MAR_ZDA_61_1159	Tissue	-34.530	25.660	ZDA	MAR	yes	yes	S

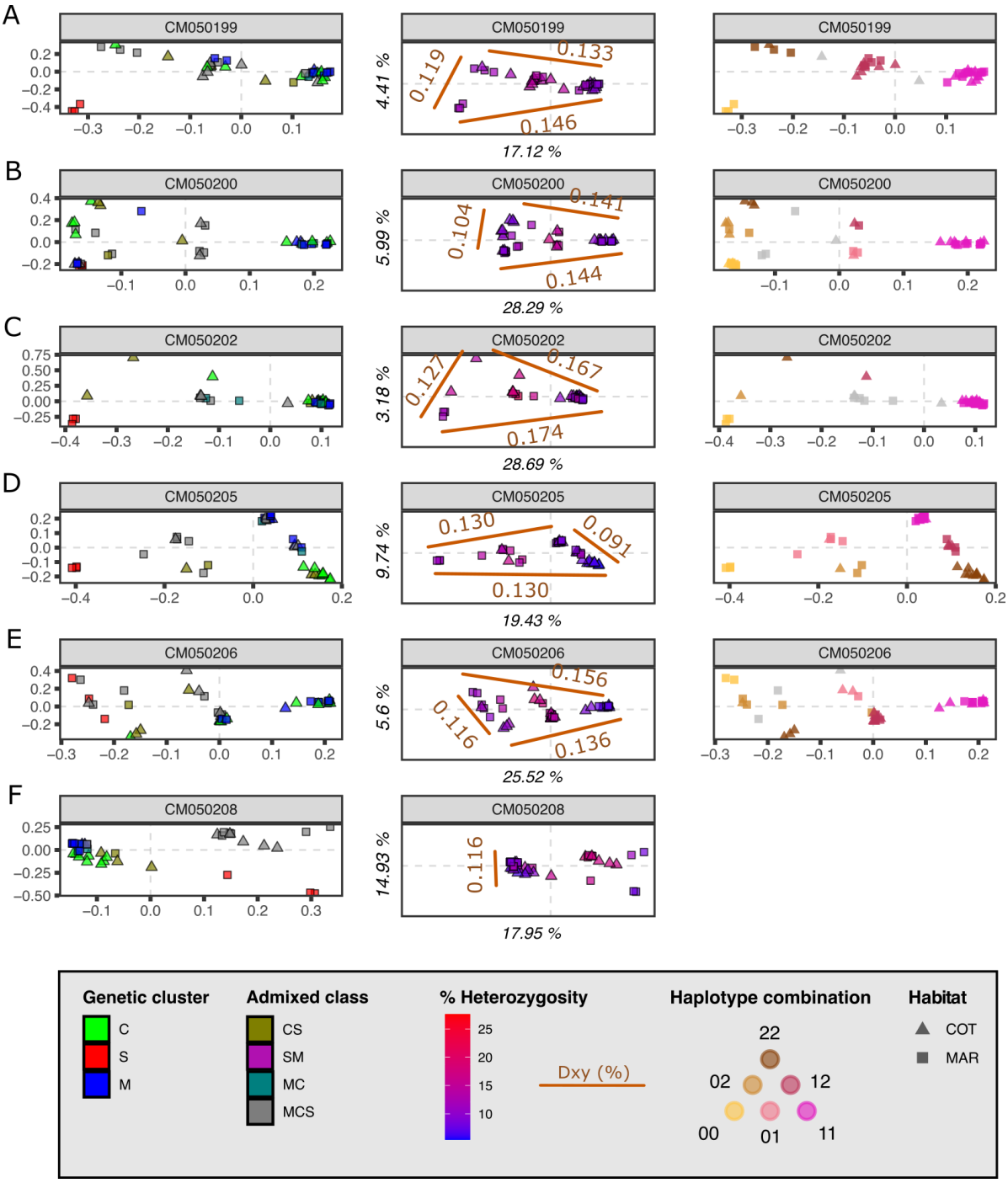
Supplementary Table S2. Sampling locations used in the study. Samples were collected from both coastal and marine habitats in some locations.

Location	Description
ALB	Morocco, Alboran Sea
BMN	Bulgaria, Black Sea
CMN	Kerch Strait, Black Sea
CNR	Canary islands, North-East Atlantic
DKB	Denmark, Baltic Sea
GAS	Bay of Biscay, North-East Atlantic
GDL	Gulf of Lion, Mediterranean Sea
MA1	Southern Morocco, North-East Atlantic
MA2	Southern Morocco, North-East Atlantic
MA3	Nothern Morocco, North-East Atlantic
MA4	Nothern Morocco, North-East Atlantic
MDN	English Channel, North-East Atlantic
NOR	Skagerrak, North-East Atlantic
PLN	Poland, Baltic Sea
PRS	Southern Portugal, North-East Atlantic
SIC	Sicily, Mediterranean Sea
SPN	Spain, Mediterranean Sea
TNO	Tunisia, Mediterranean Sea
ZDA	South Africa, South-East Atlantic

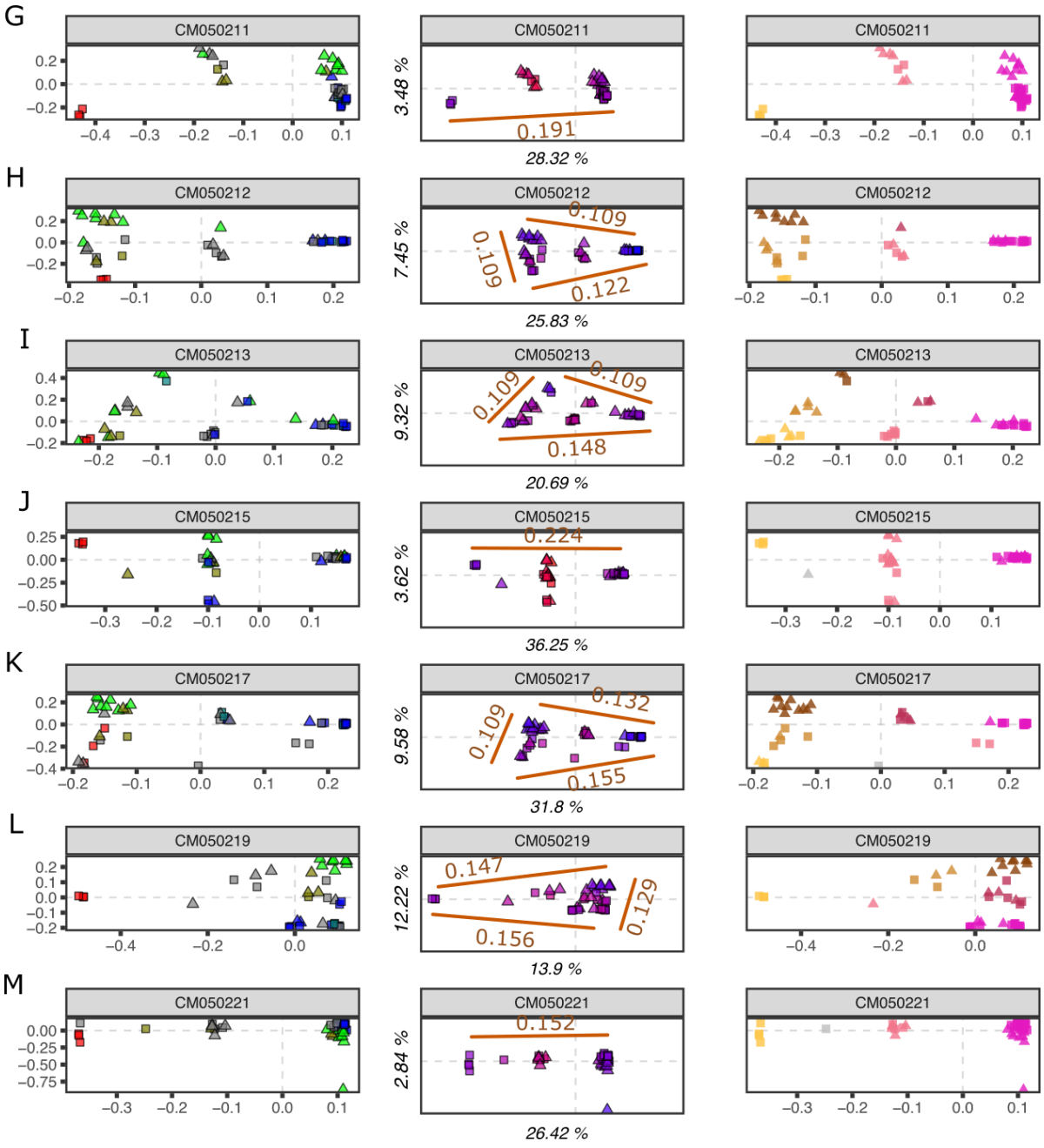
Supplementary Fig. S1. Genomic alignment dot plot showing the comparison between the chromosome-level assembly of *Coilia nasus* (top) and our *Engraulis encrasicolus* genome assembly (right) subset to contain only scaffolds longer than 10 kb. Only shown here are alignment matches with similarity threshold of 50%. The plot was generated using D-GENIES (https://github.com/genotoul-bioinfo/dgenies).



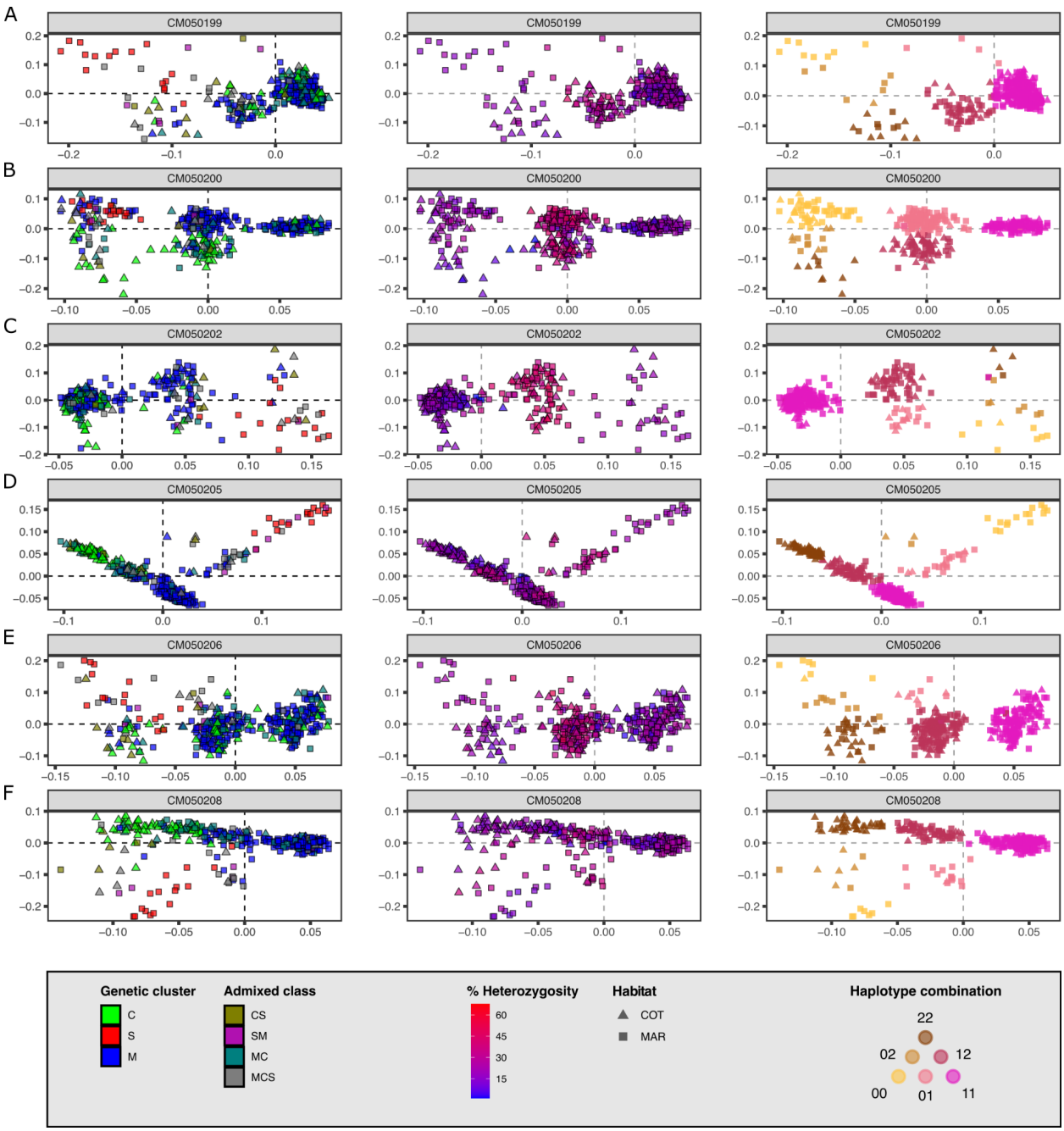
Supplementary Fig. S2. Assignment of haplotype combinations for WGS-sequenced individuals (n=39) (continues on next page). Rows correspond to 13 highly differentiated chromosomes, where horizontal axes show PCA 1 and vertical axes show PCA 2. Shapes indicate the habitat type. In the first column, colours indicate genetic cluster or admixed class. In the second column, colour indicates percentage heterozygosity and axis labels (italics) indicate the amount of variation explained by each axis. Brown text shows the d_{XY} value calculated for individuals forming opposite poles connected by brown lines. In the third column, colours represent the assigned genotype: either *00/01/11* (G, J and M), or *00/11/22/01/02/12* for chromosomes presenting three distinct haplotypes (all other chromosomes). Haplotype combinations on chromosome *CM050208* (F) were not assigned for WGS data, since no pure *00* individuals were present (determined with the RAD dataset).



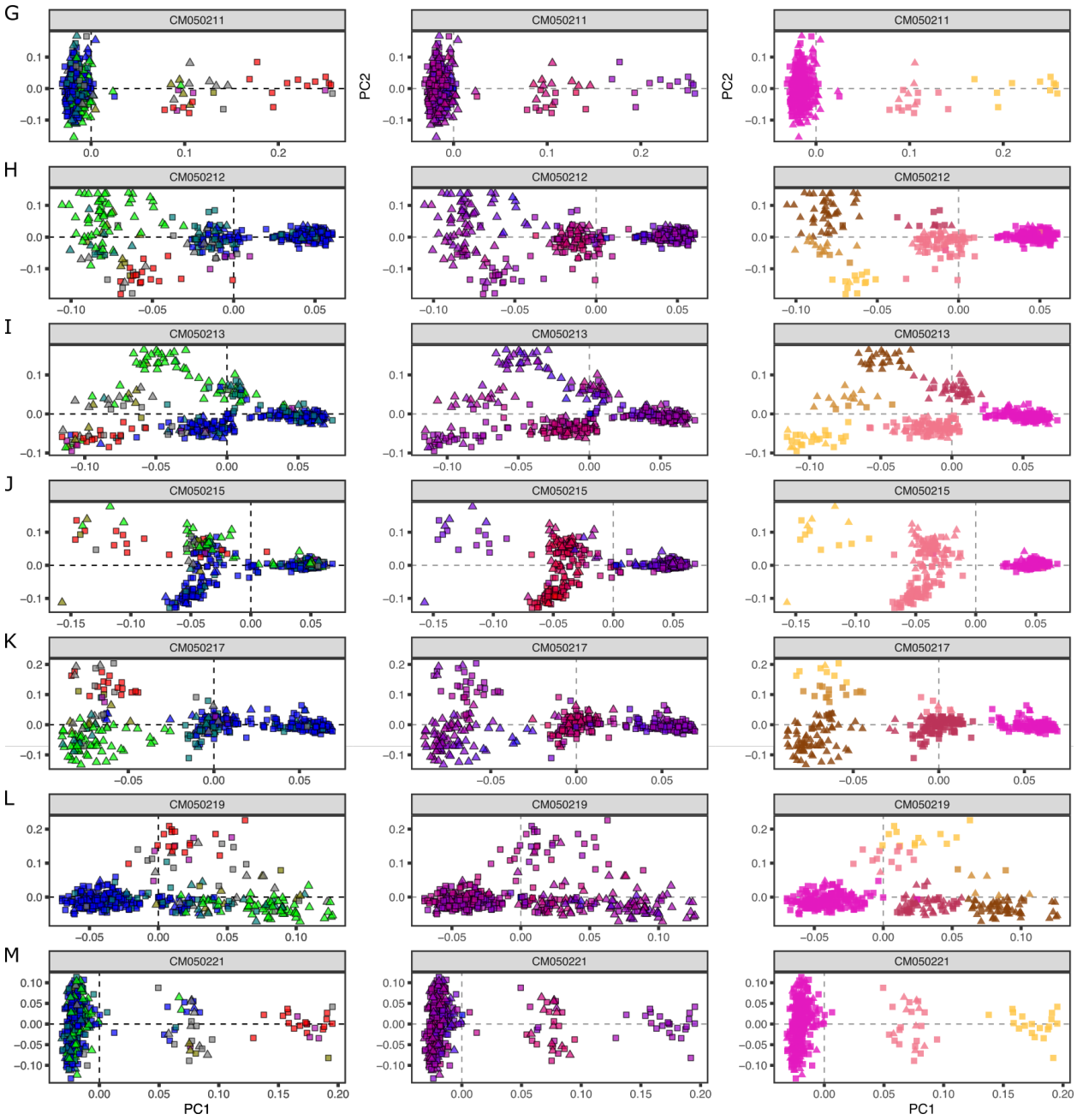
Supplementary Fig. S2. Assignment of haplotype combinations for WGS-sequenced individuals (n=39) (continued).



Supplementary Fig. S3. Assignment of haplotype combinations for RAD-sequenced individuals (n=385) (continues on next page). Symbols and colours are as described for Supplementary Figure S2. Haplotype combinations on chromosome *CM050208* (F) were assigned for RAD data, but not for WGS data.



Supplementary Fig. S3. Assignment of haplotype combinations for RAD-sequenced individuals (n=385) (continued).



3.2. Supplementary information to Chapter II

Supplementary Table S1. Full dataset of samples used in the study. For our sequenced samples, codes are composed of the species name abbreviation, the approximate sampling locality, approximate age and the sample number. The GATK dataset consisted of samples for which variant calling was performed (Variant Calling = “yes”). The last column shows mean coverage depth per individual. Br: Brest. Ga: Hossegor. Fa: Faro. Mu: Murcia. To: Tossa de mar. By: Banyuls. Va: Valras. Ag: Agde. Ms: Marseillan. Th: Thau lagoon. Se: Sète. Fr: Frontignan. Hy: Hyères. Ma: unknown Mediterranean marine site. Al: unknown Algerian site. Bz: Bizerte lagoon. Tu: unknown Tunisian site. Bs: Varna. Ru: unknown Russian site. Xx: unknown location.

Sample	Location	Habitat	Origin	Type	Library	Variant Calling	Mean cov
Hgutt_Ru_1856_77	Ru	unknown	Museum	Ethanol preserved	OVATION	No	5.15
Hgutt_Al_1864_80	Al	marine	Museum	Ethanol preserved	OVATION	No	2.16
Hgutt_Se_1898_81	Se	unknown	Museum	Ethanol preserved	OVATION	No	3.42
Hgutt_Se_1898_94	Se	unknown	Museum	Ethanol preserved	OVATION	No	8.16
Hgutt_Ag_1935_17	Ag	unknown	Citizen collect	Dried	OVATION	No	4.09
Hgutt_Ag_1935_19	Ag	unknown	Citizen collect	Dried	OVATION	No	1.54
Hgutt_Ag_1935_59	Ag	unknown	Citizen collect	Dried	OVATION	No	4.74
Hgutt_Ag_1935_70	Ag	unknown	Citizen collect	Dried	OVATION	No	4.88
Hgutt_Th_1935_66	Th	lagoon	Citizen collect	Dried	OVATION	No	4.14
Hgutt_Va_1935_22	Va	marine	Citizen collect	Dried	OVATION	No	4.00
Hgutt_Ba_1950_41	Th	lagoon	Citizen collect	Dried	OVATION	No	4.16
Hgutt_Ma_1960_60	Ma	marine	Citizen collect	Dried	OVATION	No	1.37
Hgutt_Ms_1960_21	Ms	unknown	Citizen collect	Dried	OVATION	No	4.20
Hgutt_Ro_1960_23	Th	lagoon	Citizen collect	Dried	OVATION	No	1.27
Hgutt_Ro_1960_43	Th	lagoon	Citizen collect	Dried	OVATION	No	4.39
Hgutt_Th_1960_11	Th	lagoon	Citizen collect	Dried	OVATION	No	3.91
Hgutt_Th_1960_24	Th	lagoon	Citizen collect	Dried	OVATION	No	4.30
Hgutt_Th_1960_42	Th	lagoon	Citizen collect	Dried	OVATION	No	3.11
Hgutt_Th_1960_44	Th	lagoon	Citizen collect	Dried	OVATION	No	4.76
Hgutt_Th_1960_95	Th	lagoon	Citizen collect	Dried	OVATION	No	1.61
Hgutt_Th_1964_45	Th	lagoon	Citizen collect	Dried	OVATION	No	4.46
Hgutt_Th_1964_65	Th	lagoon	Citizen	Dried	OVATION	No	3.93

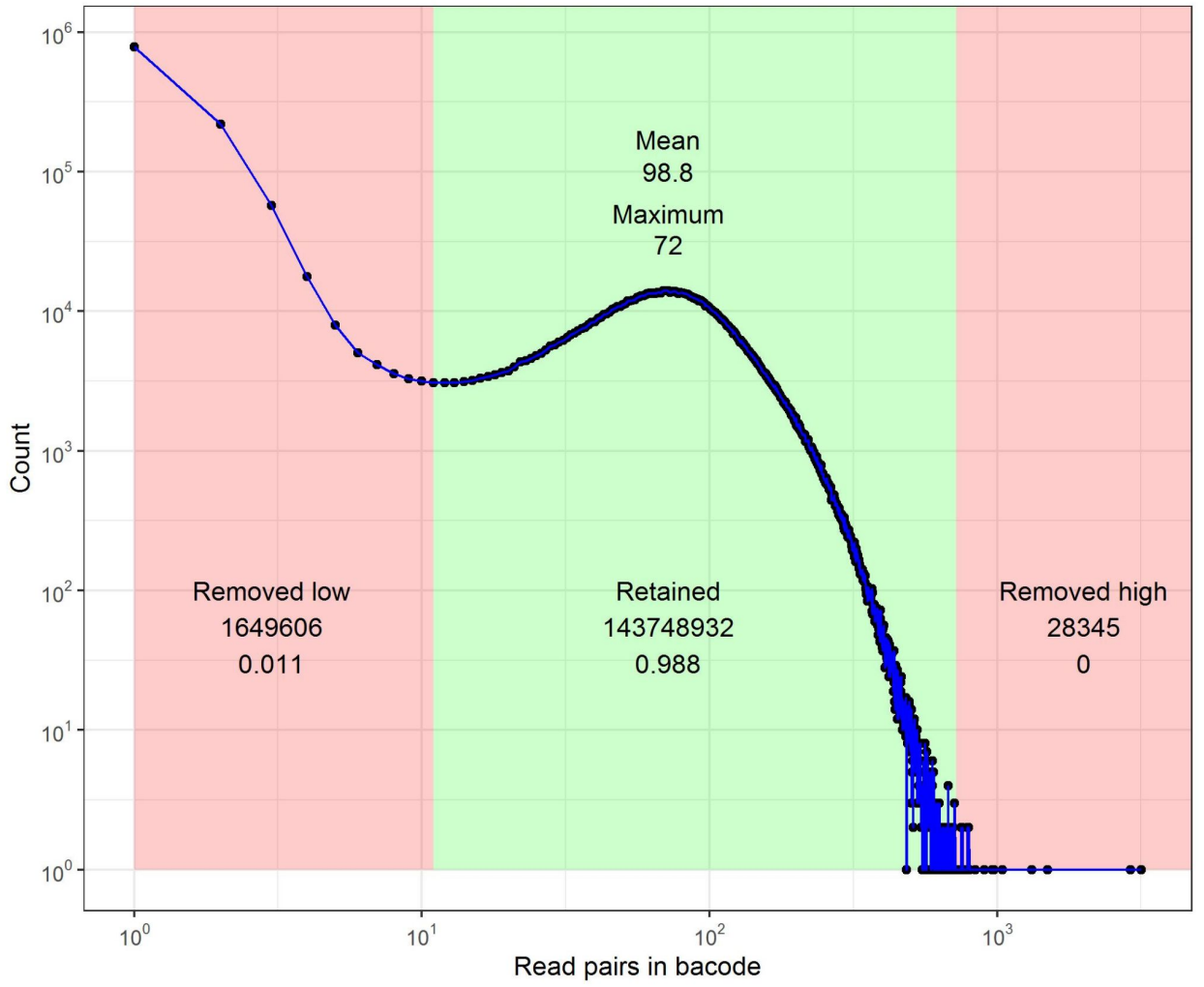
Hgutt_Th_1970_30	Th	lagoon	Citizen collect	Dried	OVATION	No	4.03
Hgutt_Th_1970_46	Th	lagoon	Citizen collect	Dried	OVATION	No	3.82
Hgutt_Il_1975_47	Th	lagoon	Citizen collect	Dried	OVATION	No	4.40
Hgutt_Il_1975_48	Th	lagoon	Citizen collect	Dried	OVATION	Yes	6.66
Hgutt_Il_1975_68	Th	lagoon	Citizen collect	Dried	OVATION	Yes	6.95
Hgutt_Bo_1980_50	Th	lagoon	Citizen collect	Dried	OVATION	No	4.30
Hgutt_By_1980_67	By	marine	Citizen collect	Dried	OVATION	No	1.33
Hgutt_Th_1980_12	Th	lagoon	Citizen collect	Dried	OVATION	No	1.10
Hgutt_Th_1980_49	Th	lagoon	Citizen collect	Dried	OVATION	No	1.58
Hgutt_Th_1980_69	Th	lagoon	Citizen collect	Dried	OVATION	No	4.59
Hgutt_Tu_1980_25	Tu	lagoon	Citizen collect	Dried	OVATION	No	4.78
Hgutt_Tu_1980_31	Tu	lagoon	Citizen collect	Dried	OVATION	No	4.46
Hgutt_Me_1982_51	Th	lagoon	Citizen collect	Dried	OVATION	No	4.63
Hgutt_Fr_1985_52	Fr	marine	Citizen collect	Dried	OVATION	No	6.07
Hgutt_Ma_1985_61	Ma	marine	Citizen collect	Dried	OVATION	No	4.89
Hgutt_Ms_1985_18	Ms	unknown	Citizen collect	Dried	OVATION	No	3.91
Hgutt_Xx_1985_71	Xx	unknown	Citizen collect	Dried	OVATION	No	6.15
Hgutt_Ma_1990_63	Ma	marine	Citizen collect	Dried	OVATION	No	4.62
Hgutt_Th_1990_72	Th	lagoon	Citizen collect	Dried	OVATION	Yes	6.46
Hgutt_Xx_1990_13	Xx	unknown	Citizen collect	Dried	OVATION	No	5.3
Hgutt_Xx_1990_14	Xx	unknown	Citizen collect	Dried	OVATION	No	2.57
Hgutt_Xx_1990_15	Xx	unknown	Citizen collect	Dried	OVATION	No	1.46
Hgutt_Xx_1990_16	Xx	unknown	Citizen collect	Dried	OVATION	No	3.45
Hgutt_Xx_1990_26	Xx	unknown	Citizen collect	Dried	OVATION	No	4.5
Hgutt_Xx_1990_27	Xx	unknown	Citizen collect	Dried	OVATION	No	4.33
Hgutt_Xx_1990_32	Xx	unknown	Citizen collect	Dried	OVATION	No	4.67
Hgutt_Xx_1990_33	Xx	unknown	Citizen collect	Dried	OVATION	No	4.82
Hgutt_Xx_1990_34	Xx	unknown	Citizen	Dried	OVATION	No	4.72

Hgutt_Xx_1990_53	Xx	unknown	Citizen collect	Dried	OVATION	No	4.98
Hgutt_Xx_1990_54	Xx	unknown	Citizen collect	Dried	OVATION	No	1.77
Hgutt_Th_1991_55	Th	lagoon	Citizen collect	Dried	OVATION	No	5.2
Hgutt_Ms_1995_64	Ms	unknown	Citizen collect	Dried	OVATION	No	5.28
Hgutt_Th_1995_56	Th	lagoon	Citizen collect	Dried	OVATION	No	4.80
Hgutt_Th_1995_76	Th	lagoon	Citizen collect	Dried	OVATION	No	1.61
Hgutt_Xx_1995_20	Xx	unknown	Citizen collect	Dried	OVATION	Yes	7.03
Hgutt_Xx_1995_38	Xx	unknown	Citizen collect	Dried	OVATION	No	1.33
Hgutt_Xx_1995_39	Xx	unknown	Citizen collect	Dried	OVATION	No	3.66
Hgutt_Xx_1995_73	Xx	unknown	Citizen collect	Dried	OVATION	Yes	8.35
Hgutt_Th_1996_28	Th	lagoon	Citizen collect	Dried	OVATION	No	4.12
Hgutt_Th_1996_74	Th	lagoon	Citizen collect	Dried	OVATION	No	1.54
Hgutt_Th_1996_75	Th	lagoon	Citizen collect	Dried	OVATION	No	4.97
Hgutt_Bs_2006_83	Bs	unknown	Riquet et al. 2019	Fresh	OVATION	Yes	8.17
Hgutt_Bs_2006_85	Bs	unknown	Riquet et al. 2019	Fresh	OVATION	Yes	11.06
Hgutt_Bs_2006_87	Bs	unknown	Riquet et al. 2019	Fresh	OVATION	Yes	11.2
Hgutt_Bs_2006_89	Bs	unknown	Riquet et al. 2019	Fresh	OVATION	Yes	9.97
Hgutt_Bs_2006_91	Bs	unknown	Riquet et al. 2019	Fresh	OVATION	Yes	11.71
Hgutt_Th_2007_36	Th	lagoon	Citizen collect	Dried	OVATION	Yes	7.39
Hgutt_Th_2007_57	Th	lagoon	Citizen collect	Dried	OVATION	No	5.94
Hgutt_Me_2010_58	Th	lagoon	Citizen collect	Dried	OVATION	Yes	7.85
Hgutt_Br_2014_06	Br	marine	Riquet et al. 2019	Fresh	OVATION	Yes	7.75
Hgutt_Br_2014_07	Br	marine	Riquet et al. 2019	Fresh	OVATION	Yes	8.18
Hgutt_Br_2014_08	Br	marine	Riquet et al. 2019	Fresh	OVATION	Yes	6.68
Hgutt_Br_2014_09	Br	marine	Riquet et al. 2019	Fresh	OVATION	Yes	6.30
Hgutt_Br_2014_10	Br	marine	Riquet et al. 2019	Fresh	OVATION	Yes	7.04
Hgutt_Bz_2014_82	Bz	lagoon	Riquet et al. 2019	Fresh	OVATION	Yes	9.81
Hgutt_Bz_2014_84	Bz	lagoon	Riquet et	Fresh	OVATION	Yes	10.23

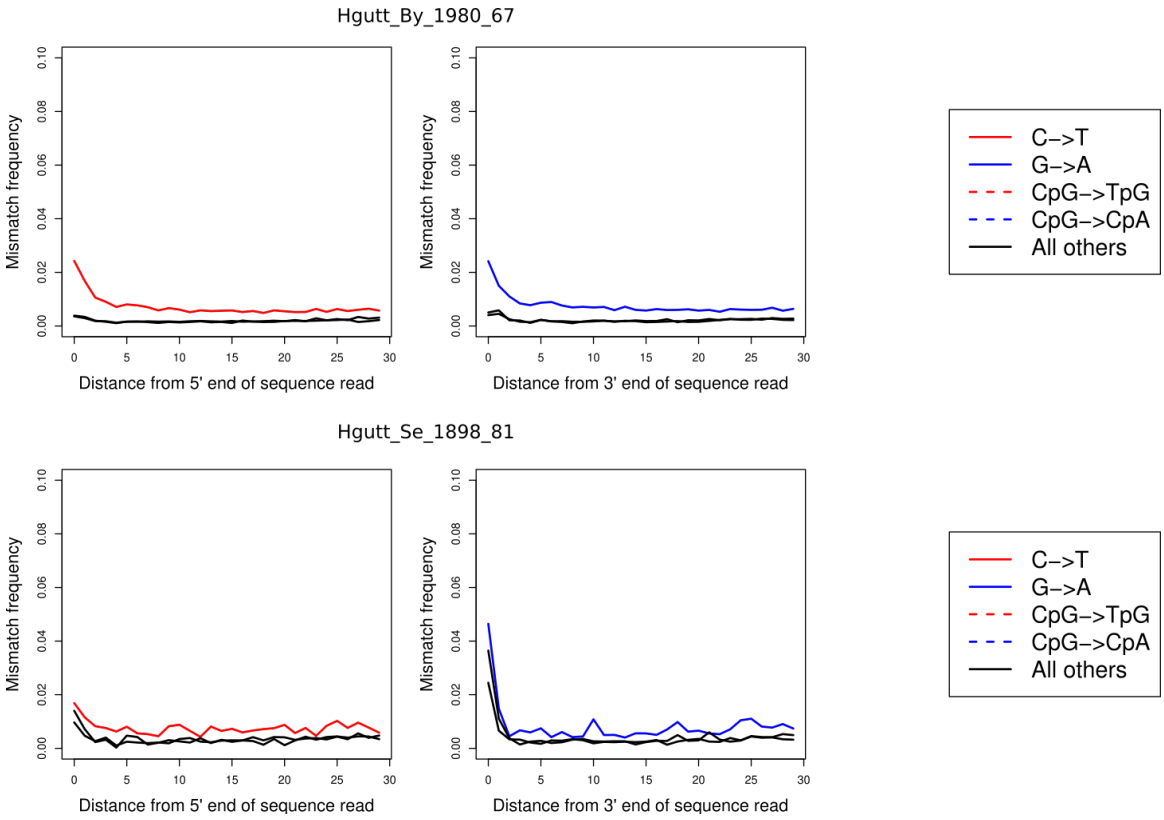
			al. 2019				
Hgutt_Bz_2014_86	Bz	lagoon	Riquet et al. 2019	Fresh	OVATION	Yes	9.40
Hgutt_Bz_2014_88	Bz	lagoon	Riquet et al. 2019	Fresh	OVATION	Yes	11.03
Hgutt_Bz_2014_90	Bz	lagoon	Riquet et al. 2019	Fresh	OVATION	Yes	8.83
Hgutt_Li_2015_1	Th	lagoon	Barry et al. (2022)	Fresh	TruSeq	Yes	39.82
Hgutt_Li_2015_3	Th	lagoon	Barry et al. (2022)	Fresh	TruSeq	Yes	54.64
Hgutt_Li_2015_4	Th	lagoon	Barry et al. (2022)	Fresh	TruSeq	Yes	51.15
Hgutt_Li_2015_5	Th	lagoon	Barry et al. (2022)	Fresh	TruSeq	Yes	47.56
Hgutt_Li_2015_6	Th	lagoon	Barry et al. (2022)	Fresh	TruSeq	Yes	47.57
Hgutt_To_2016_01	To	marine	Riquet et al. 2019	Fresh	OVATION	Yes	7.83
Hgutt_To_2016_02	To	marine	Riquet et al. 2019	Fresh	OVATION	Yes	6.98
Hgutt_Hy_2017_03	Hy	marine	Riquet et al. 2019	Fresh	OVATION	Yes	5.75
Hgutt_Hy_2017_04	Hy	marine	Riquet et al. 2019	Fresh	OVATION	Yes	7.41
Hgutt_Hy_2017_05	Hy	marine	Riquet et al. 2019	Fresh	OVATION	Yes	8.18
Hgutt_Ga_2019_10	Ga	lagoon	Barry et al. (2022)	Fresh	TruSeq	No	2.89
Hgutt_Ga_2019_11	Ga	lagoon	Barry et al. (2022)	Fresh	TruSeq	Yes	50.33
Hgutt_Ga_2019_12	Ga	lagoon	Barry et al. (2022)	Fresh	TruSeq	No	1.77
Hgutt_Ga_2019_1	Ga	lagoon	Barry et al. (2022)	Fresh	TruSeq	No	1.35
Hgutt_Ga_2019_2	Ga	lagoon	Barry et al. (2022)	Fresh	TruSeq	No	1.34
Hgutt_Ga_2019_3	Ga	lagoon	Barry et al. (2022)	Fresh	TruSeq	Yes	31.01
Hgutt_Ga_2019_4	Ga	lagoon	Barry et al. (2022)	Fresh	TruSeq	No	1.28
Hgutt_Ga_2019_6	Ga	lagoon	Barry et al. (2022)	Fresh	TruSeq	Yes	17.7
Hgutt_Ga_2019_7	Ga	lagoon	Barry et al. (2022)	Fresh	TruSeq	Yes	46.32
Hgutt_Ga_2019_8	Ga	lagoon	Barry et al. (2022)	Fresh	TruSeq	Yes	5.74
Hgutt_Ga_2019_9	Ga	lagoon	Barry et al. (2022)	Fresh	TruSeq	Yes	13.36
Hgutt_Mu_2019_1	Mu	lagoon	Barry et al. (2022)	Fresh	TruSeq	Yes	48.45
Hgutt_Mu_2019_2	Mu	lagoon	Barry et al. (2022)	Fresh	TruSeq	Yes	43.72
Hgutt_Mu_2019_3	Mu	lagoon	Barry et al. (2022)	Fresh	TruSeq	Yes	20.79
Hgutt_Mu_2019_4	Mu	lagoon	Barry et	Fresh	TruSeq	Yes	33.14

			al. (2022)				
Hgutt_Mu_2019_5	Mu	lagoon	Barry et al. (2022)	Fresh	TruSeq	Yes	43.94
Hgutt_Fa_2020_2	Fa	lagoon	Barry et al. (2022)	Fresh	TruSeq	Yes	43.61
Hgutt_Fa_2020_3	Fa	lagoon	Barry et al. (2022)	Fresh	TruSeq	Yes	48.43
Hgutt_Fa_2020_4	Fa	lagoon	Barry et al. (2022)	Fresh	TruSeq	Yes	45.89
Hgutt_Fa_2020_5	Fa	lagoon	Barry et al. (2022)	Fresh	TruSeq	Yes	21.86
Hgutt_Fa_2020_6	Fa	lagoon	Barry et al. (2022)	Fresh	TruSeq	Yes	47.96

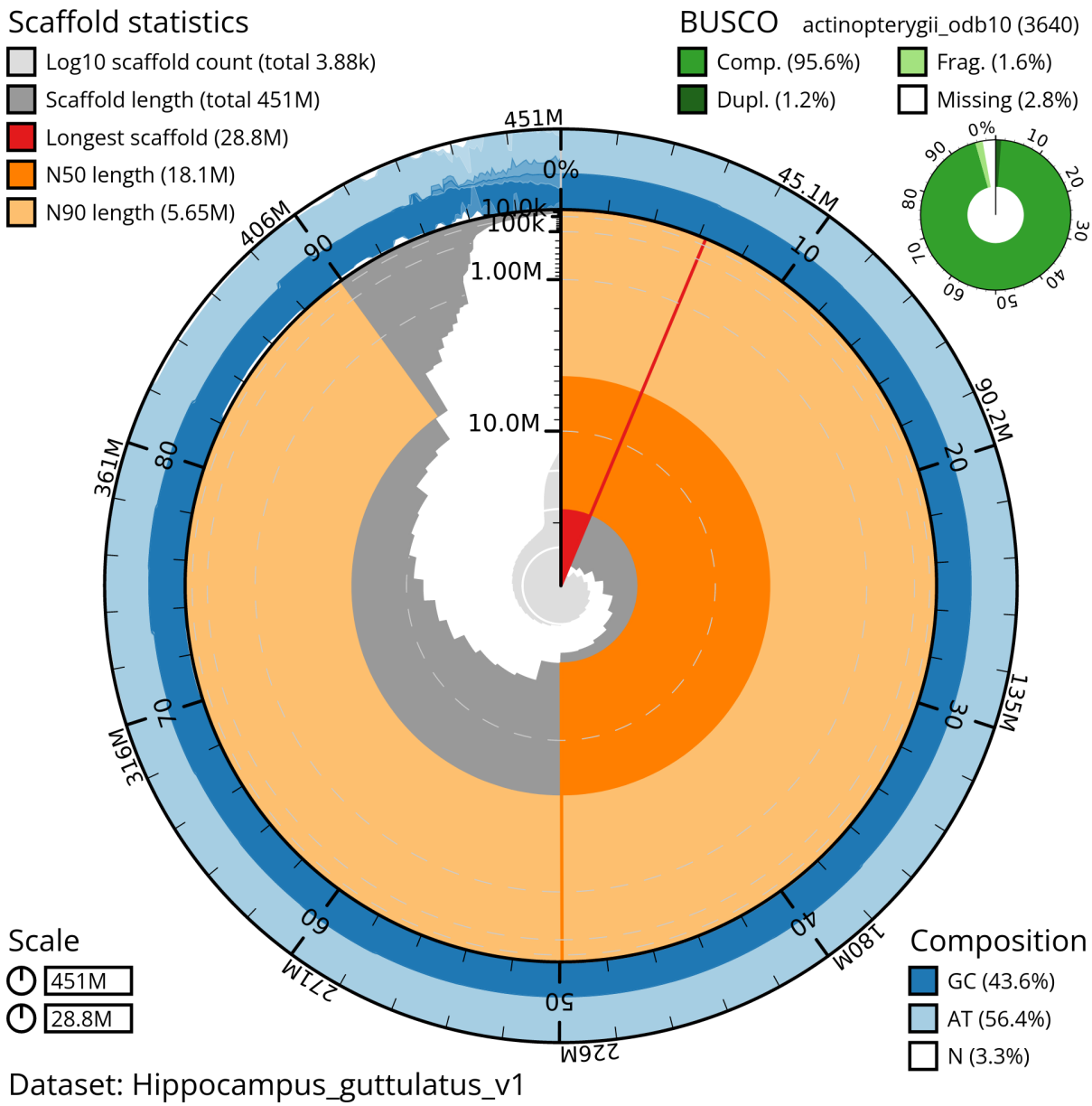
Supplementary Figure S1. Count distribution of the number of read pairs per barcode in the deduplicated 10X Chromium linked-read sequencing data generated for the reference genome assembly (Hgutt_V1). Molecular barcodes associated with low numbers of read pairs (left part of the distribution shaded in red) reflect noise in the 10X data (e.g. sequencing errors within barcodes), and were therefore excluded before performing *de novo* assembly (1,649,606 read pairs, 1.1% of the total sequencing data). Over-represented molecular barcodes on the right end of the distribution (28,345 read pairs) were also excluded. The resulting average number of read pairs per retained barcode was approximately 100, with a maximum occurrence at 72.



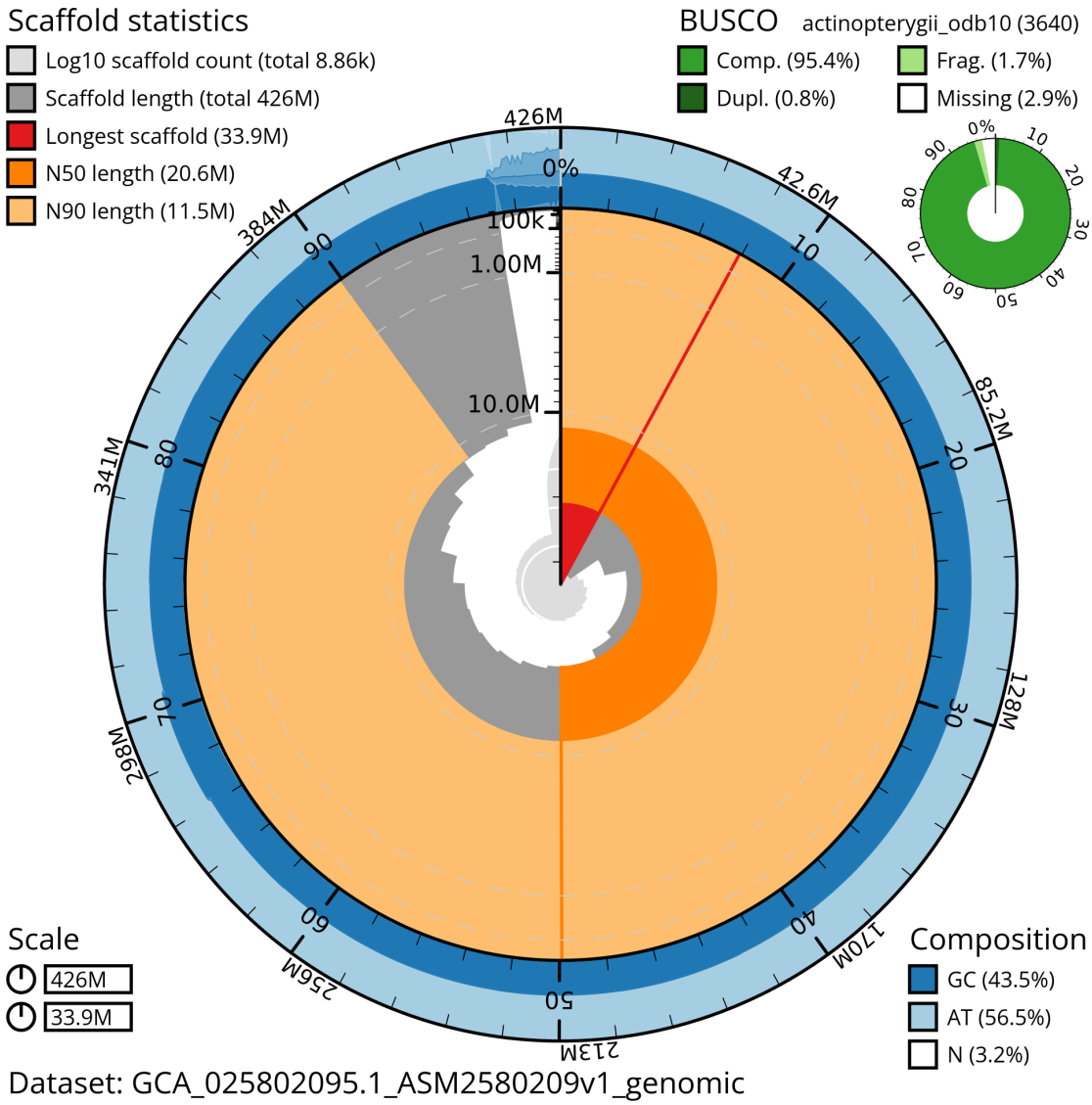
Supplementary Figure S2. DNA damage patterns (cytosine deamination) visualised using PMDtools (v0.60) (Skoglund et al., 2014). Results are shown for two historical samples.



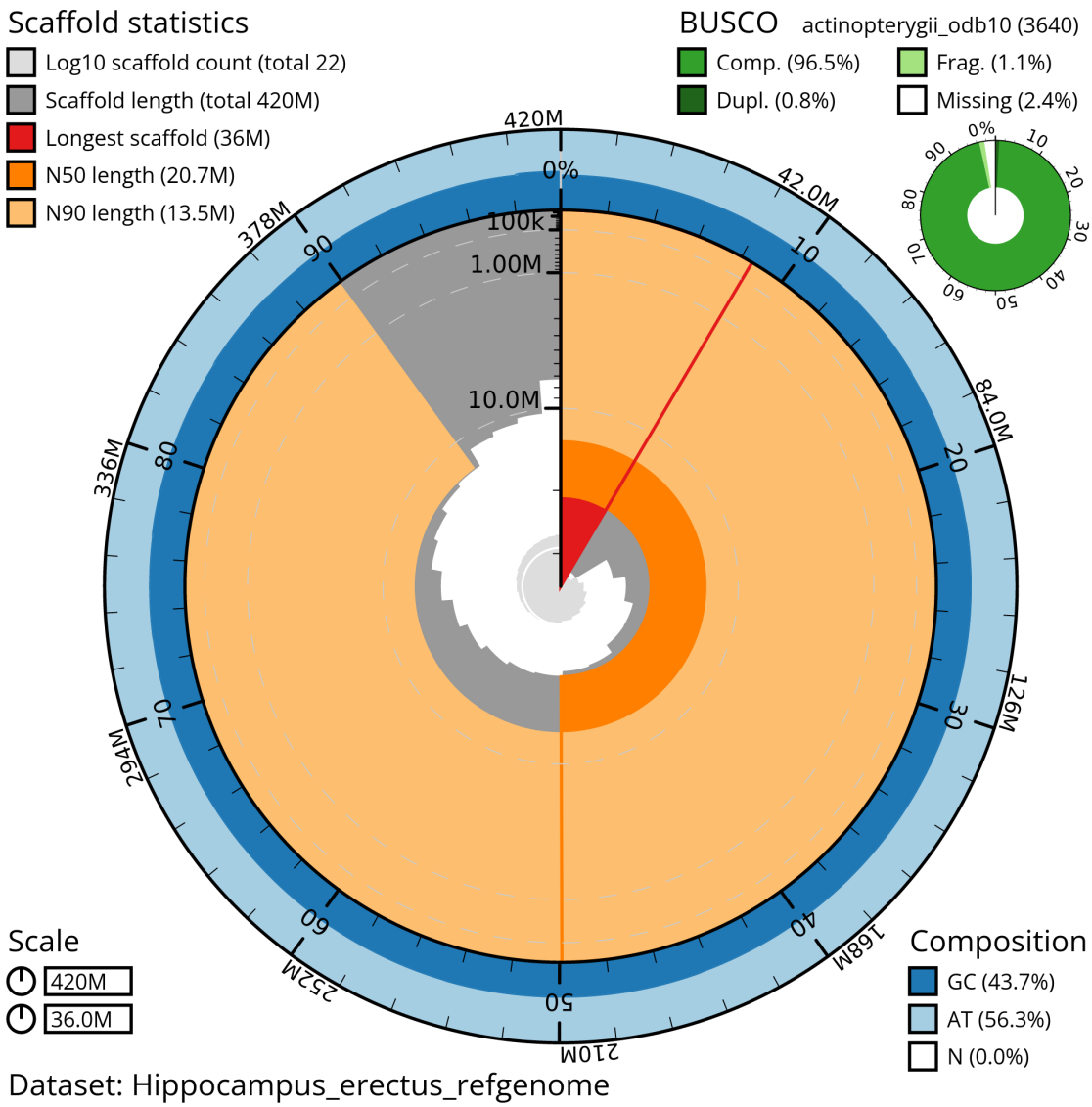
Supplementary Figure S3. Snail plot describing the assembly statistics of the *Hippocampus guttulatus* Hgutt_V1 reference genome. The genome contiguity is shown in a circle representing the full assembly length of 451 Mb, with the N50 length in dark orange and the N90 length in light orange. The longest scaffold was 28.8 Mb. The completeness BUSCO scores (in shades of green) and base composition (percentage of GC in dark blue, AT in light blue, and N in light grey) appear in top right and bottom left panels, respectively. The plot was generated using *Blobtoolkit* (https://blobtoolkit.genomehubs.org/).



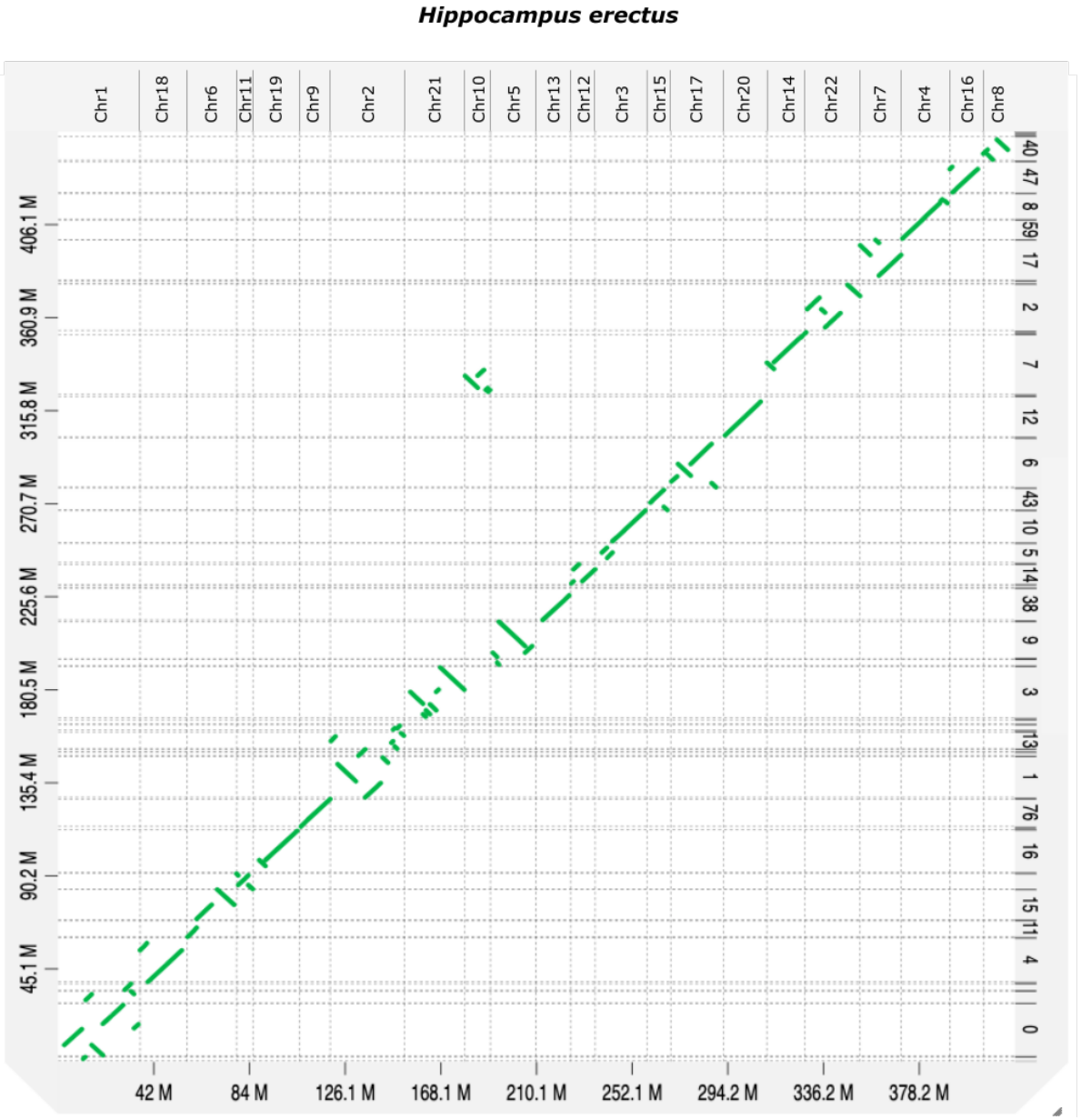
Supplementary Figure S4a. Snail plot describing the assembly statistics of the *Hippocampus guttulatus* GCA_025802095.1 reference genome (HguttRefA, Jones et al in prep).



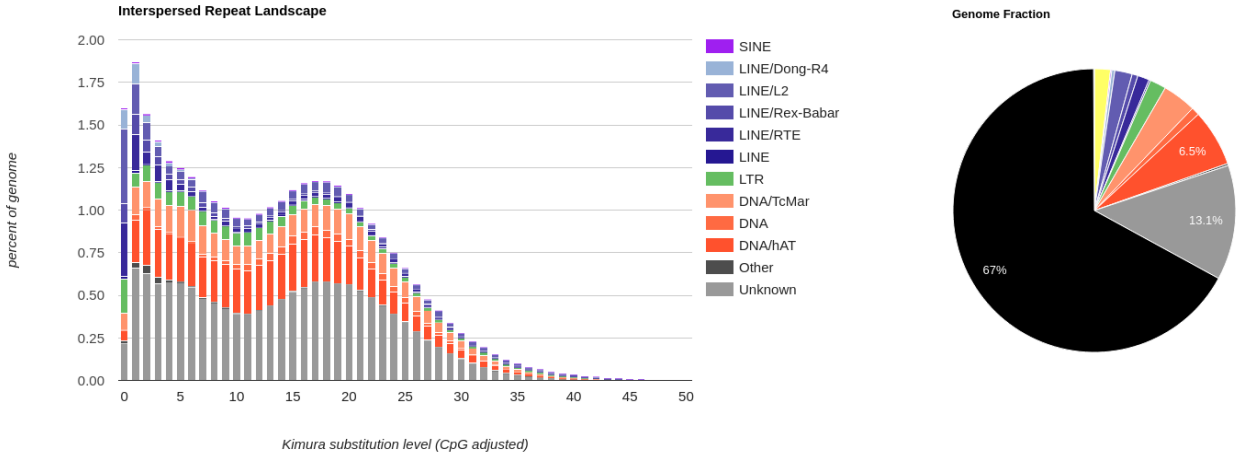
Supplementary Figure S4b. Snail plot describing the assembly statistics of the *Hippocampus erectus* reference genome (Li et al 2021).



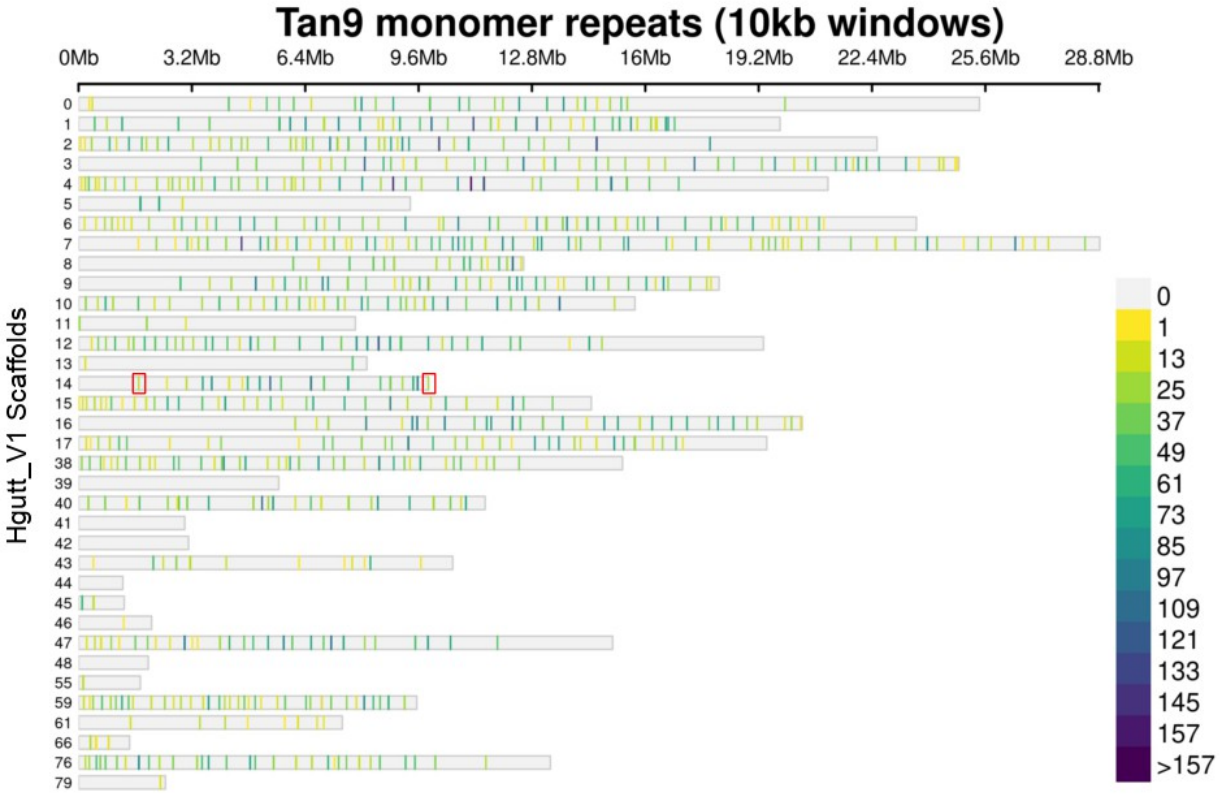
Supplementary Figure S5. Genomic alignment dot plot showing the comparison between the chromosome-level assembly of *Hippocampus erectus* (top) and our *H. guttulatus* Hgutt_V1 genome assembly (right). Only shown here are alignment matches passing a minimum size filter and a similarity threshold of 50%. All *H. guttulatus* scaffolds match a single *H. erectus* chromosome (with varied number of intrachromosomal rearrangements), except for scaffold 7 that matches both Chr10 and Chr14 of the *H. erectus* genome. The plot was generated using *D-GENIES* (https://github.com/genotoul-bioinfo/dgenies).



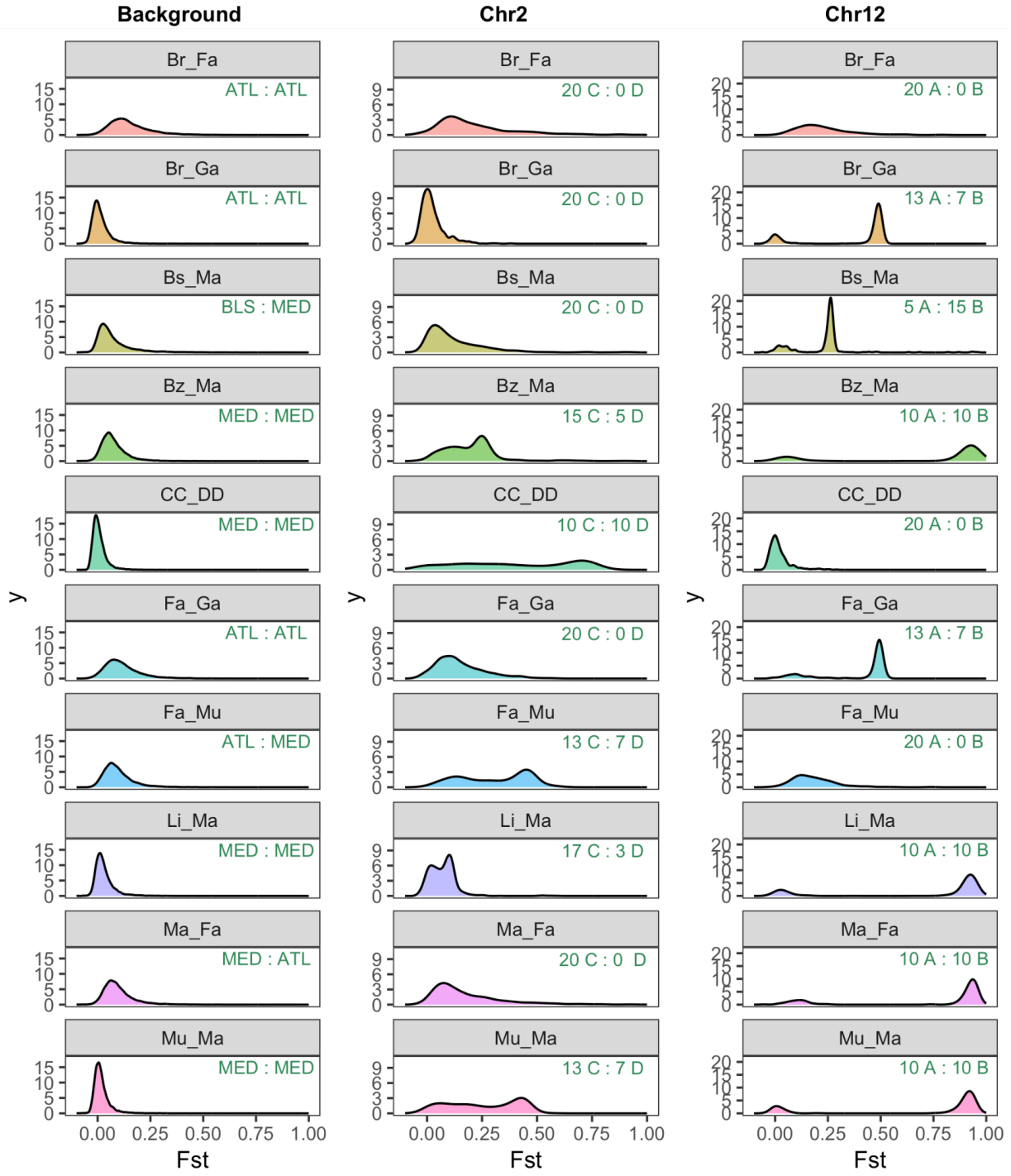
Supplementary Figure S6. Repeat landscape showing the percentage of the Hgutt_V1 genome occupied by different categories of interspersed repeat elements (right), and detailed as a function of weighted average Kimura divergence in alignments for each repeat family (left).



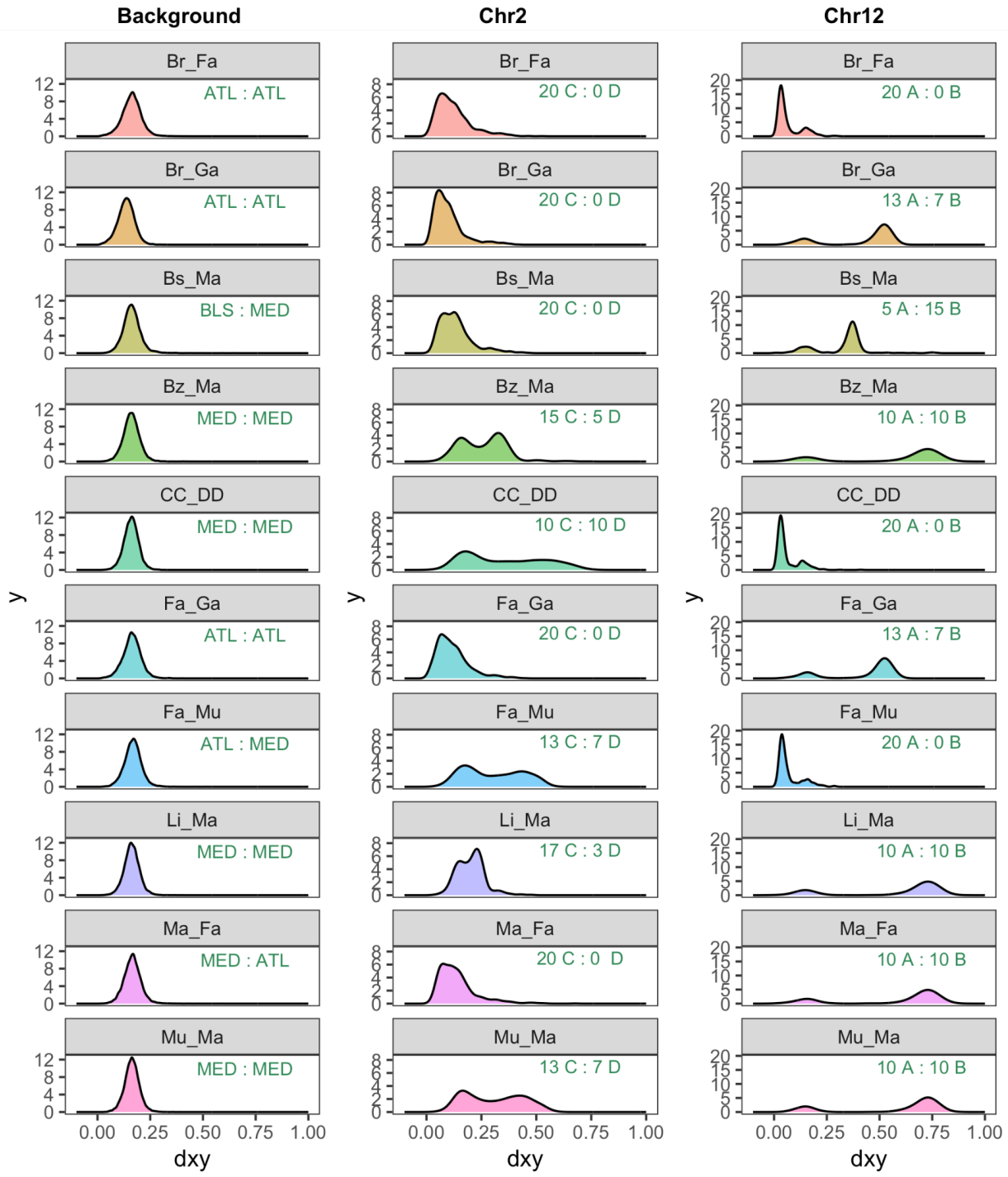
Supplementary Figure S7. Repeat landscape showing the number of Tan9 monomer repeats within 10kb windows across Hgutt_V1 scaffolds longer than 1Mb. The consensus sequence of the Tan9 monomer (37bp, GTCGTTTTTTTCGGCTAAAAACGCCTTACTATACATG) was searched with *blastn* with parameters set to M=1, N=-1, Q=2, R=2, W=8, following (Melters et al 2013). Red boxes indicate the Tan9 repeats where the two inversion breakpoints map on Scaffold 14 (Chr12), delimiting an 8.2 Mb-long inversion.



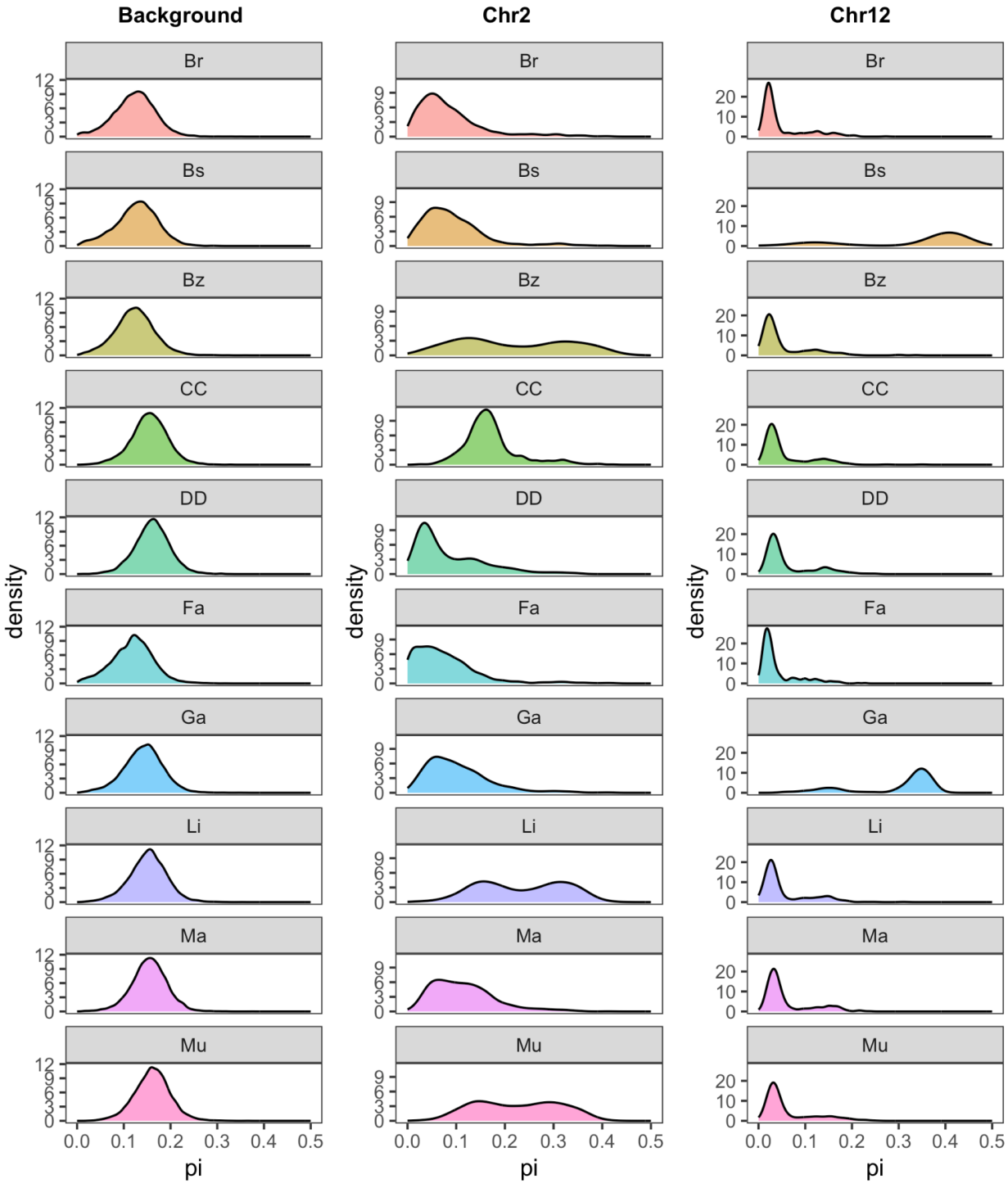
Supplementary Figure S8. Distributions of F_{ST} values displayed separately for the genome background, Chr2 and Chr12. Black text gives information about the populations being compared: oceanic basins ("Background" column), total number of C/D alleles in the comparison ("Chr2" column), or total number of A/B alleles in the comparison ("Chr12" column). F_{ST} was calculated for each population pair using the popgenWindows.py script in 25 kb windows (Martin, 2018; https://github.com/simonhmartin/genomics_general). The "Ma" population is here represented by individuals from the Mediterranean marine sites Tossa de mar and Hyères.



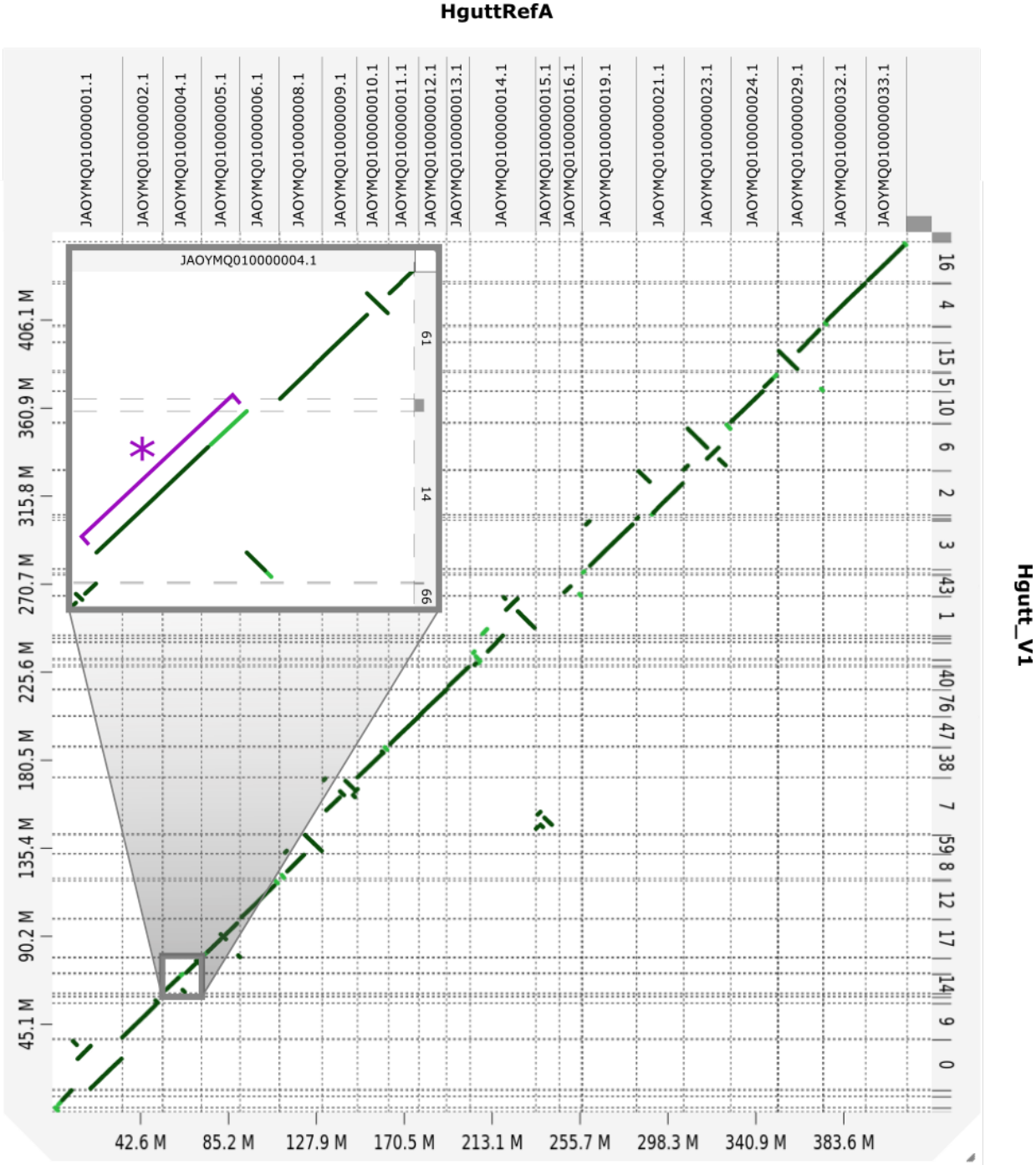
Supplementary Figure S9. Distributions of d_{XY} values displayed separately for the genome background, Chr2 and Chr12 (for details see previous figure).



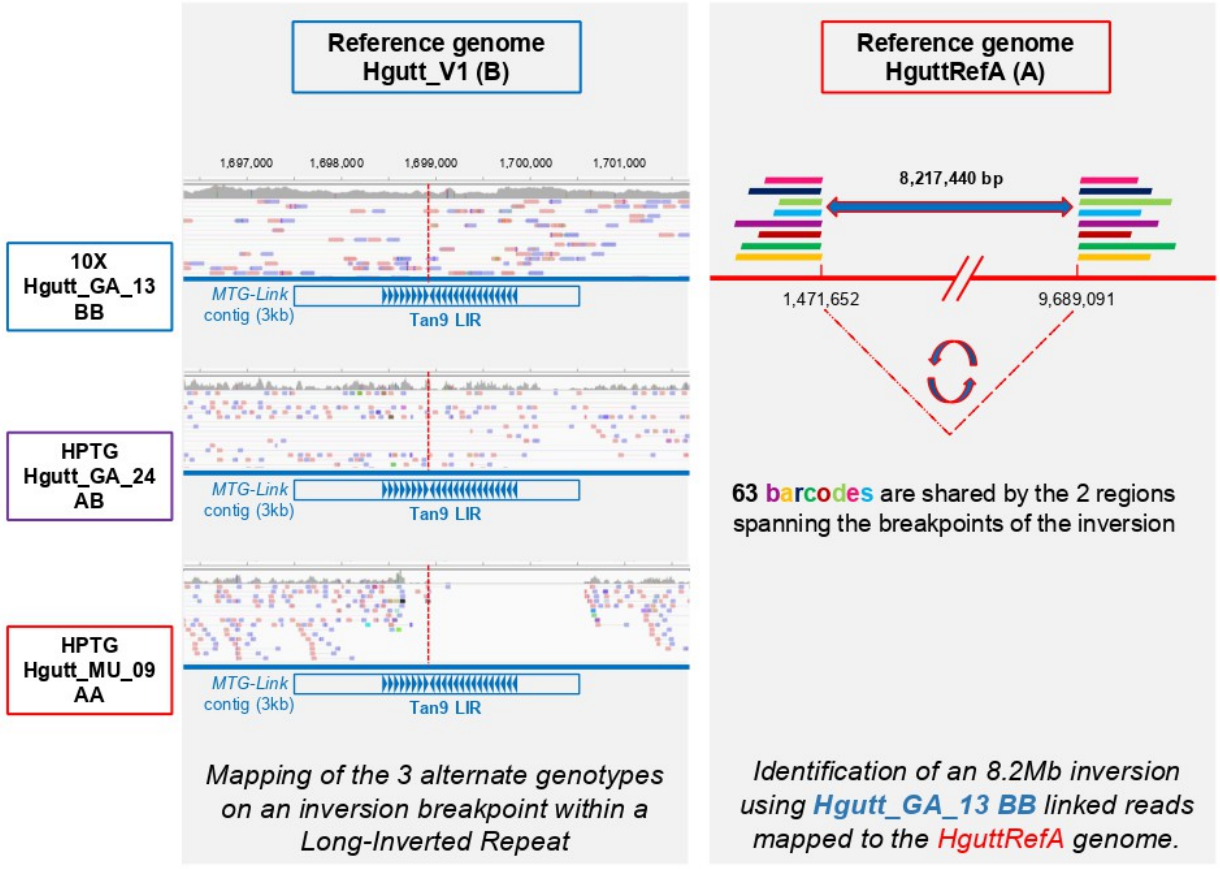
Supplementary Figure S10. Distributions of nucleotide diversity (π) values displayed separately for the genome background, Chr2 and Chr12. Nucleotide diversity was calculated for each population or haplotype group (for details see Supplementary Fig. S8).



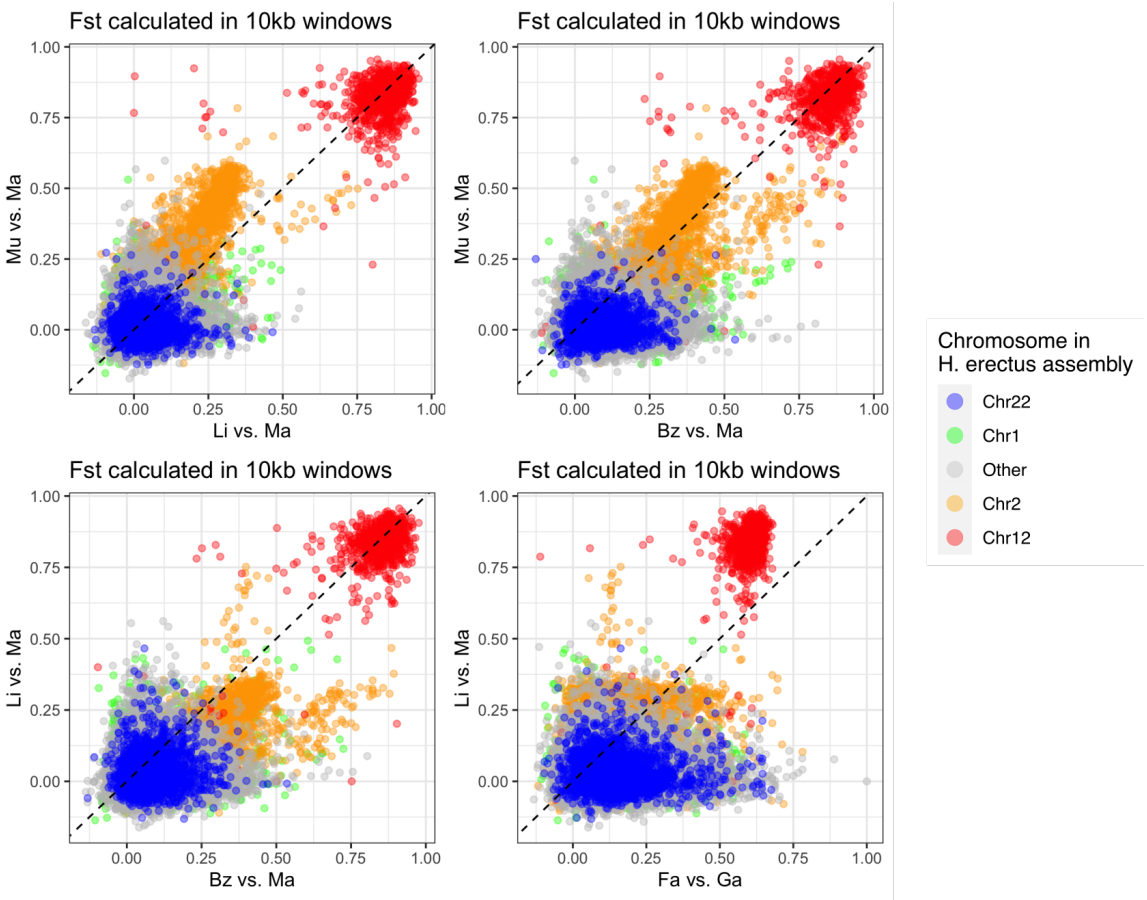
Supplementary Figure S11. Genomic alignment dot plot showing the comparison between the chromosome-level assembly of *Hippocampus guttulatus* from the English Channel (HguttRefA) (top) and our Hgutt_V1 genome assembly (right). Only shown here are alignment matches passing a minimum size filter and a similarity threshold of 50%. The zoomed section shows chromosome JAOYMQ010000004.1 which corresponds to *H. erectus* Chr11 (scaffold 61) and Chr12. The junction between Chr11 and Chr12 within chromosome JAOYMQ010000004.1 is possibly an artefact of the assembly pipeline (i.e. the junction is not supported by reads from *H. guttulatus*). The purple asterisk indicates the 8.2 Mb-long region that is inverted between the two *H. guttulatus* reference assemblies. The plot was generated using *D-GENIES* (https://github.com/genotoul-bioinfo/dgenies).



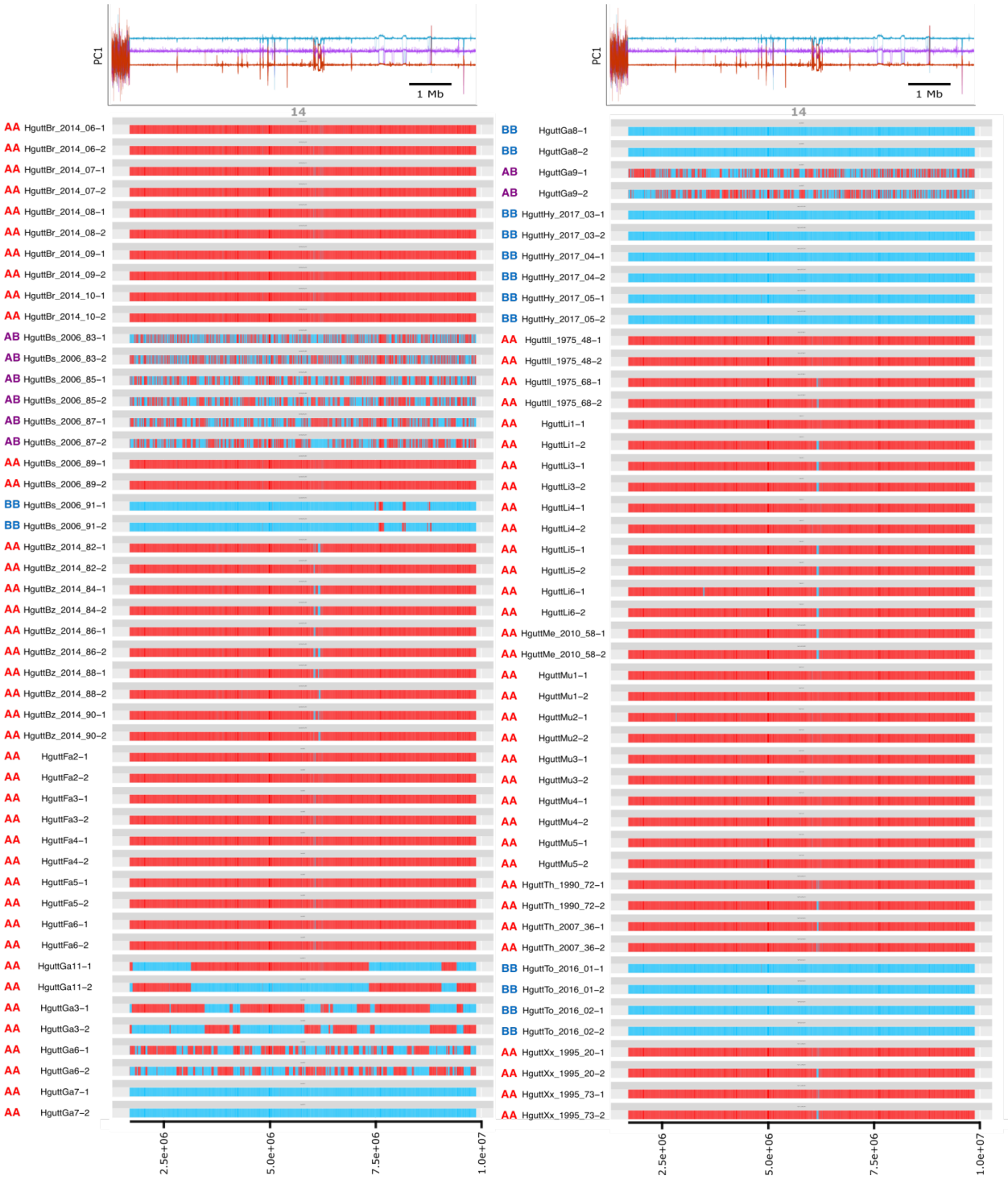
Supplementary Figure S12. Direct identification of an 8.2 Mb inversion on Chr12 and characterisation of inversion breakpoints. *Left panel:* The region around the inversion breakpoint localised near position 1,699 Mb (read dotted line) on scaffold 14 of Hgutt_V1 (Blue horizontal line, B arrangement) was reassembled using *MTG-Link*, resulting in a 3kb contig matching the reference (blue rectangle). A long inverted repeat of Tan9 monomers (blue triangles) was found centred at the breakpoint position. Linked-read sequencing data obtained for three individuals with alternate B12 genotypes (i.e. the BB individual used for reference genome assembly sequenced at >100X coverage, plus an AA and an AB obtained by haplotagging and sequenced at ~12X coverage) were mapped against the Hgutt_V1 genome assembly. Horizontal links in each local alignment show linked paired-end reads sharing the same molecular barcode. Long synthetic molecules spanning the 3kb breakpoint region were found for both the BB and the AB genotypes, but not for the AA genotype that showed a ~1.5kb mapping gap directly after the breakpoint. *Right panel:* Linked-reads from the BB genotype were mapped against the HguttRefA assembly (Red horizontal line, A arrangement), and analysed with *Leviathan* to search for structural variants. A total of 63 long synthetic molecules (coloured segments) were found shared by two regions spanning the inversion breakpoints and separated by 8.2 Mb. Read orientation supported a structural variant of type inversion.



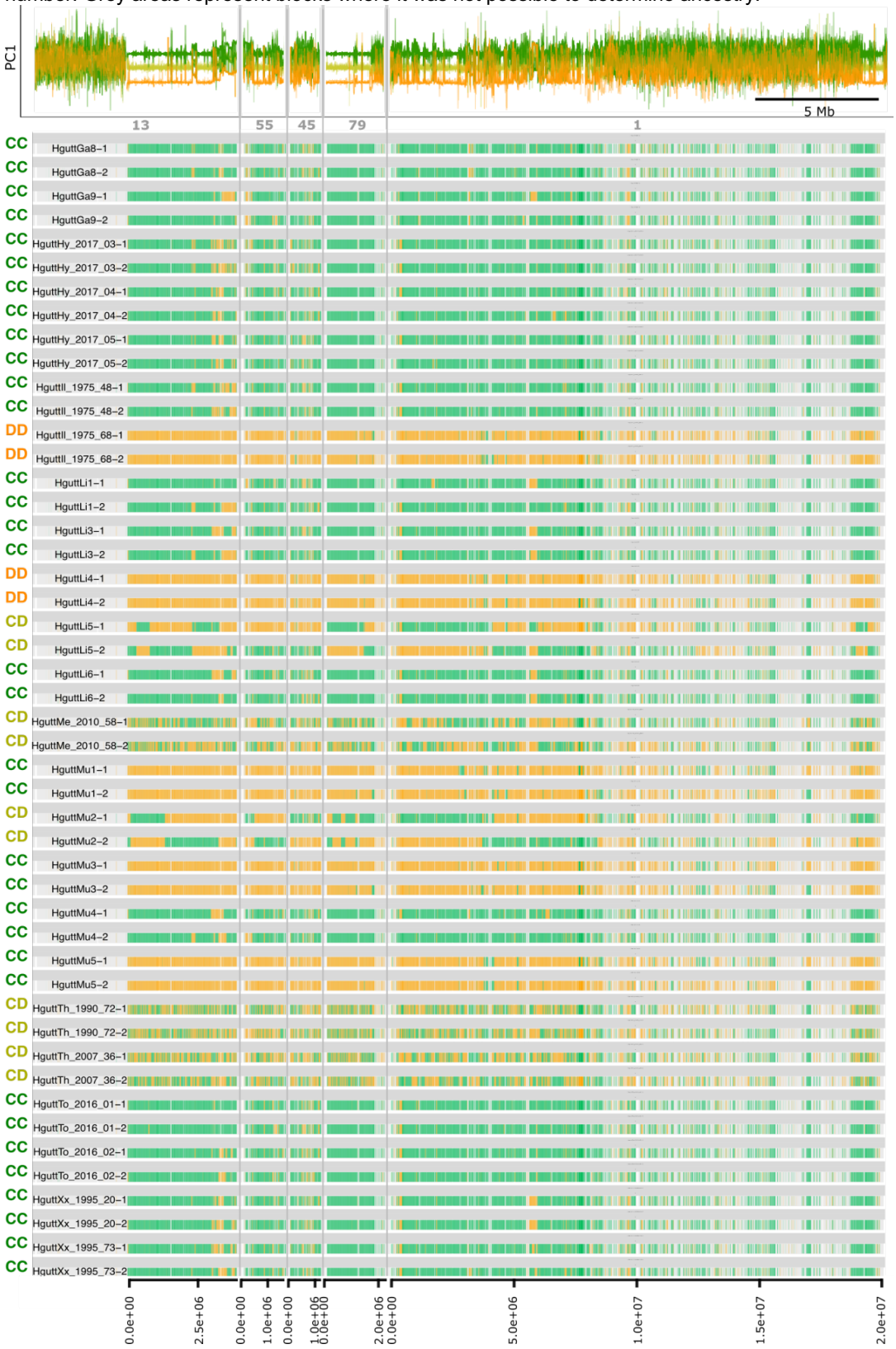
Supplementary Figure S13. F_{ST} co-plots between different Mediterranean and Atlantic lineages.



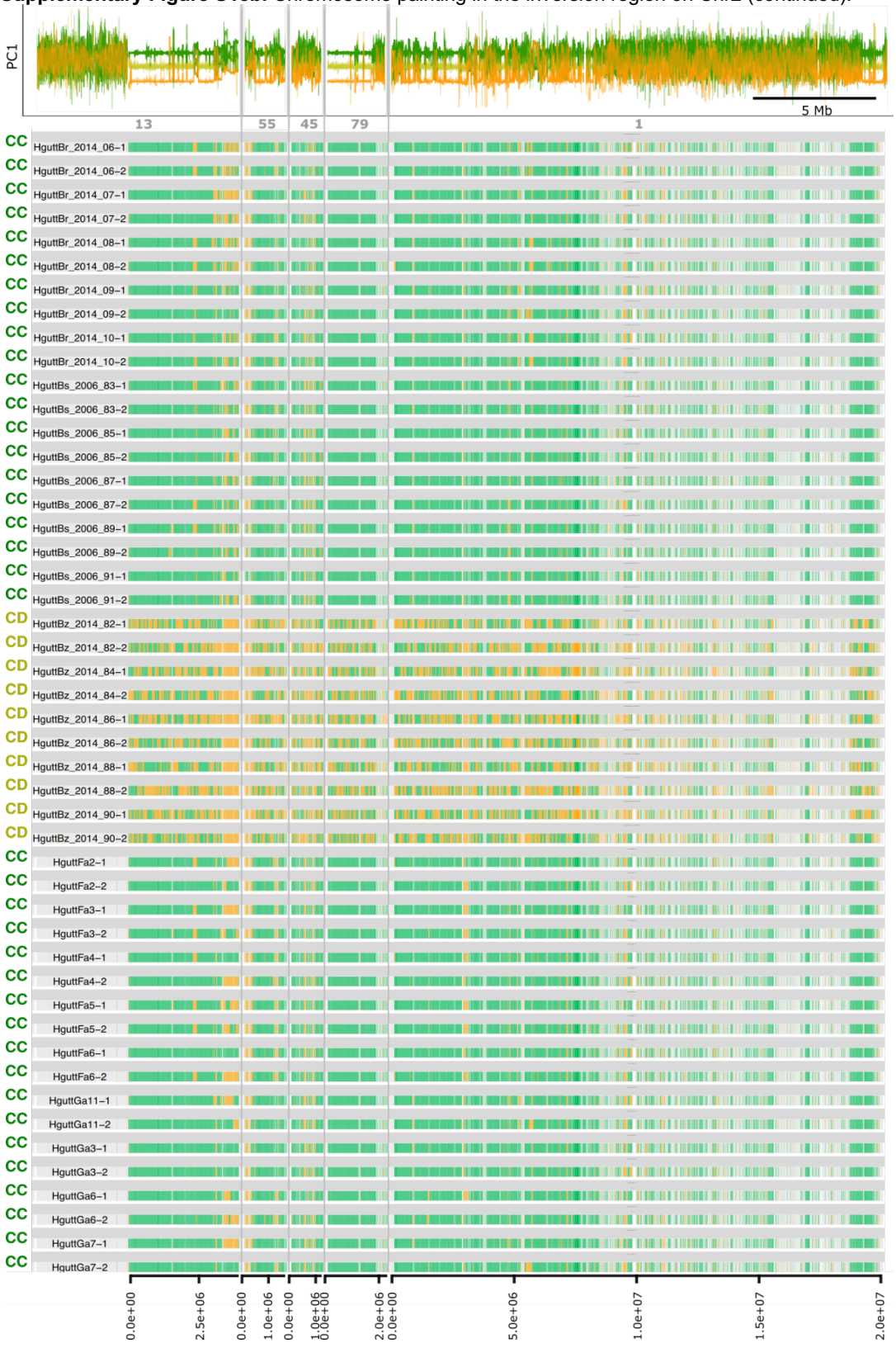
Supplementary Figure S14. Chromosome painting in the inversion region on Chr12 (scaffold 14). Codes indicate inversion genotypes and individual sample names of which the last characters ("-1" or "-2") refer to phased parental haplotypes. Inversion heterokaryotes show phasing errors in the form of haplotype switches. The top panels show the first principal component of local PCA along the inversion.



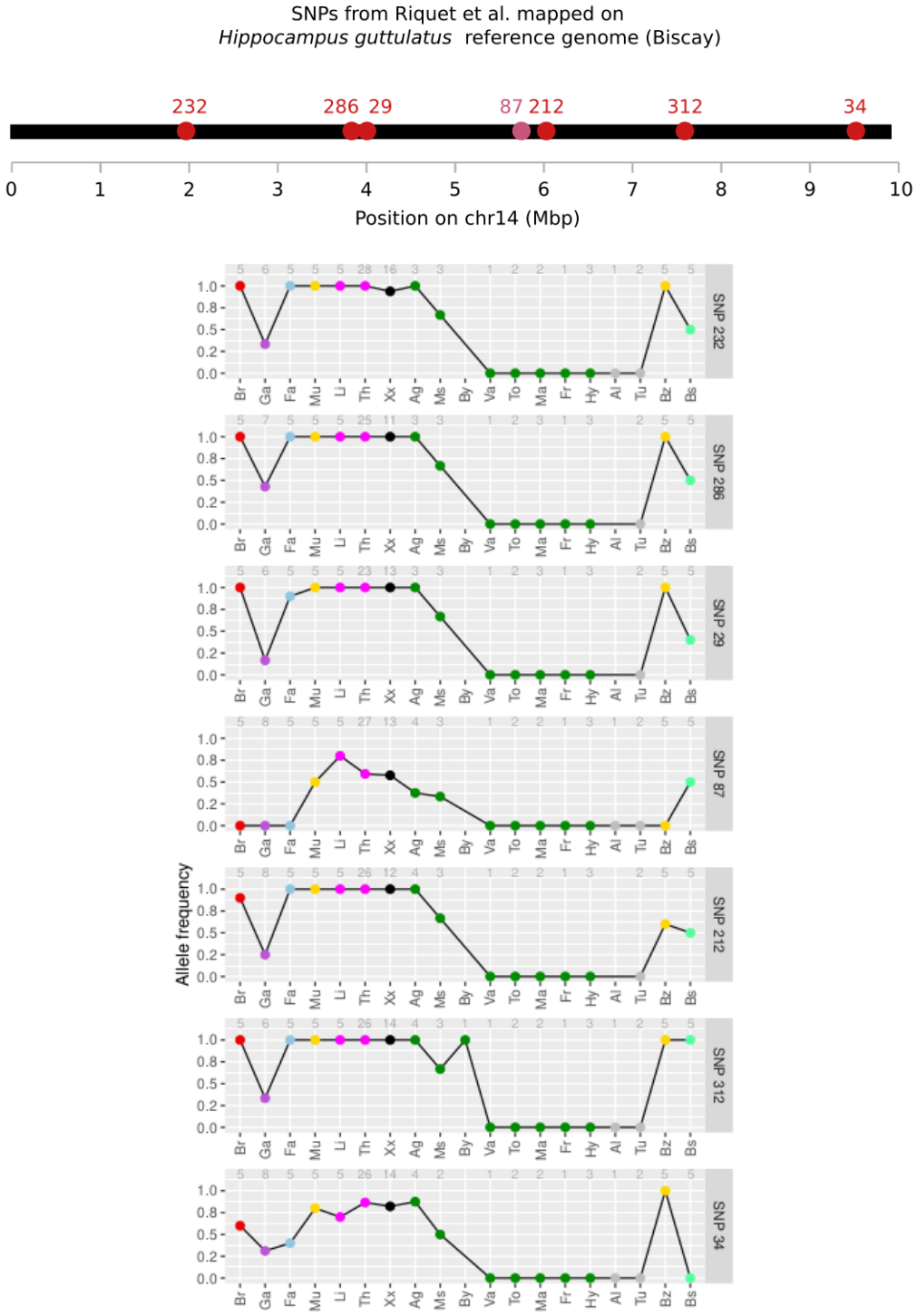
Supplementary Figure S15a. Chromosome painting in the inversion region on Chr2. Grey font indicates scaffold number. Grey areas represent blocks where it was not possible to determine ancestry.



Supplementary Figure S15b. Chromosome painting in the inversion region on Chr2 (continued).



Supplementary Figure S16. Outlier SNPs identified in Riquet et al. (2019) mapped onto scaffold 14 (*H. erectus* Chr12) of our reference genome (Hgutt_V1). Allele frequencies calculated from our whole-genome data (n=112) are shown per sampling location. Grey numbers indicate sample sizes.



Supplementary File S2 - Scripts and commands

Reference genome assembly and repeat annotation

Step	Command
nubeam-dedup	#Deduplication of 10x genomics reads nubeam-dedup -i1 Hgutt_S1_L001_R1_001.fastq.gz -i2 Hgutt_S1_L001_R2_001.fastq.gz -o1 Hgutt_DEDUP_S1_L001_R1_001.fastq.gz -o2 Hgutt_DEDUP_S1_L001_R2_001.fastq.gz -z 6 -r 0
process_10xReads	#Process raw 10x genomics reads process_10xReads.py -a -o Hgutt_DEDUP_PROC_S1 -1 Hgutt_DEDUP_S1_L001_R1_001.fastq.gz -2 Hgutt_DEDUP_S1_L001_R2_001.fastq.gz
custom R script	#Identification of rare and over-represented barcodes Rscript --vanilla Autoset_Filters.R Hgutt_DEDUP_PROC_S1_barcodes.txt barocode_list.txt
filter_10XReads	#Filters reads for status and barcodes filter_10xReads.py -m 11 -n 720 -o Hgutt_DEDUP_FILTERED_S1 -B Hgutt_DEDUP_PROC_S1_barcodes.txt -1 Hgutt_DEDUP_PROC_S1_R1_001.fastq.gz -2 Hgutt_DEDUP_PROC_S1_R2_001.fastq.gz
regen_10XReads	#Return reads to origin format regen_10xReads.py -o Hgutt_DEDUP_REGEN_S1 -1 Hgutt_DEDUP_FILTERED_S1_R1_001.fastq.gz -2 Hgutt_DEDUP_FILTERED_S1_R2_001.fastq.gz
Supernova-2.1.1	#Generate whole genome *de novo* assembly supernova run --id Hgutt_DEDUP_FILTERED_S1_all --description DEDUP_FILTERED_S1_all_reads --fastqs /DEDUP_FILTERED --maxreads=all --accept-extreme-coverage #Generate FASTA output for the assembly supernova mkoutput --style=pseudohap2 --asmdir=Hgutt_DEDUP_FILTERED_S1_all/outs/assembly --outprefix=Hippocampus_guttulatus_v1 --minsize=1000 --index --headers=short
BlobToolKit (v3.5.2) BUSCO (v5.4.4)	#Create a new BlobDir dataset blobtools create --fasta Hippocampus_guttulatus_v1.fasta --meta Hippocampus_guttulatus_v1.yaml /DATASETS/Hippocampus_guttulatus_v1 #Run BUSCO analysis busco -i ./Hippocampus_guttulatus_v1.fasta -l actinopterygii -o Hgutt_v1_BUSCO -m geno #Add results of BUSCO analysis to the BlobToolKit dataset blobtools add --busco /Hgutt_v1_BUSCO/run_actinopterygii_odb10/full_table.tsv /DATASETS/Hippocampus_guttulatus_v1 #View results blobtools view --local /DATASETS/Hippocampus_guttulatus_v1
RepeatModeler2	#Create database for RepeatModeler BuildDatabase -name Hippocampus_guttulatus_v1 Hippocampus_guttulatus_v1.fasta

	#Run RepeatModeler nohup RepeatModeler -database Hippocampus_guttulatus_v1 -threads 20 -LTRStruct >& run.out &
Tandem Repeats Finder (v4.09.1)	#Run pyTanFinder pyTanFinder.py Hippocampus_guttulatus_v1.fasta -minM 50 -maxM 2000 -minMN 2 -minA 100000 -px HguttV1_Tan -tp ./trf409.linux64
RepeatMasker (v4.0.5)	#Run RepeatMasker RepeatMasker -lib Hippocampus_guttulatus_v1-families.fa -pa 12 -a Hippocampus_guttulatus_v1.fasta #Calculate divergence for each repeat family calcDivergenceFromAlign.pl -s Hippocampus_guttulatus_v1.divsum Hippocampus_guttulatus_v1.fasta.align #Generate Repeat Landscape createRepeatLandscape.pl -div ./Hippocampus_guttulatus_v1.divsum -g 424000000 > ./Hgutt_Repeat_Landscape.html #Softmasking for Braker2 structural annotation RepeatMasker -lib Hippocampus_guttulatus_v1-families.fa -pa 12 -gff -xsmall -a Hippocampus_guttulatus_v1.fasta
trim-galore (v0.6.7-1)	#Trim reads trim_galore --paired RNAseqfile_R1.fq RNAseqfile_R2.fq
HISAT2 (v2.2.1)	#Index the softmasked reference genome hisat2-build Hippocampus_guttulatus_v1.fasta.masked /Hisat2_db/Hippocampus_guttulatus_v1_masked #Align RNA-Seq reads hisat2 -p 20 -x /Hisat2_db/Hippocampus_guttulatus_v1_masked -1 Trimmed_RNAseqfile_1.fq -2 Trimmed_RNAseqfile_2.fq -S aligned.sam
Samtools (v1.12)	#Convert sam to sorted bam samtools sort aligned.sam -o aligned_sorted.bam
Braker2 (v2.1.6)	#Run Braker2 braker.pl --species=Hippocampus_guttulatus --genome=Hippocampus_guttulatus_v1.fasta.masked --bam=aligned_1_sorted.bam,aligned_2_sorted.bam,aligned_3_sorted.bam,aligned_4_sorted.bam,aligned_5_sorted.bam,aligned_6_sorted.bam,aligned_7_sorted.bam

Whole-genome sequencing data

Some of the following commands were ran using snakemake (v7.1.1), for which snakefiles can be found at https://github.com/pierrebarry/life_tables_genetic_diversity_marine_fishes.

Step	Command
fastp (v0.23.1)	fastp -i {input.raw_R1} -I {input.raw_R2} -o {output.fastp_R1} -O {output.fastp_R2} --trim_poly_g --correction --low_complexity_filter –html {output.report_html} --json {output.report_json} –report_title {wildcards.sample} --thread $ --dont_overwrite --merge –merged_out {output.merged}

bwa mem (v0.7.17)	#paired reads bwa mem -M -t $ referencegenome_{wildcards.species} {input.fastq_R1} {input.fastq_R2} > {output.align_sam_paired}) #merged reads bwa mem -M -t $ referencegenome_{wildcards.species} {input.fastq_merged} > {output.align_sam_merged}) #combine sam files picard MergeSamFiles I={input.align_sam_paired} I={input.align_sam_merged} O={output.sam_final}
Sam to bam file (picard v2.26.8)	picard SortSam -I {input.align_sam} -O {output.align_bam_picard} -SO coordinate -CREATE_INDEX true -VALIDATION_STRINGENCY LENIENT -TMP_DIR tmp
Mark duplicates (picard v2.26.8)	picard -Xmx$g MarkDuplicates -I {input.align_bam_picard} -O {output.markdup_picard} -ASSUME_SORTED TRUE -REMOVE_DUPLICATES FALSE -CREATE_INDEX TRUE -METRICS_FILE {wildcards.sample}_duplicate_metrics.txt -VALIDATION_STRINGENCY LENIENT -TMP_DIR tmp
Add read group (picard v2.26.8)	picard AddOrReplaceReadGroups -I {input.markdup_picard} -O {output.markdup_rg_picard} -RGPL ILLUMINA -RGLB lib -RGPU genewiz -RGSM {wildcards.sample}
Get stats (samtools v1.10, htslib v1.10.2)	samtools flagstat {input.bamfile} > {output.samtools_flagstat} samtools stats -d {input.bamfile} > {output.samtools_stats}
pmdtools (v0.50)	for i in *.bam do echo pmdtools running on: $i samtools view $i | python /path/pmdtools --deamination > $i_pmdtools.txt done #Plotted with their Rscript plotPMD.R
GATK HaplotypeCaller (GATK 4.1.8.0)	gatk HaplotypeCaller -R {input.reference_genome} -I {input.markdup_rg_picard} -O {output.gvcf_first} -bamout {output.realign_bam} -ERC GVCF -G StandardAnnotation -G AS_StandardAnnotation -G StandardHCAnnotation --tmp-dir tmp
GATK GenomicsDBImport	gatk --java-options '-Xms$g -Xmx$G' GenomicsDBImport --sample-name-map name_Hgutt.txt –genomicsdb-workspace-path {output.joint_genotyping_files} --tmp-dir /home/lmeyer/tmp #name_Hgutt.txt contained sample names and the list of *_gvcf_first.g.vcf files

GATK GenotypeGVCFs	gatk --java-options '-Xms$g -Xmx$G' GenotypeGVCFs -R {input.reference_genome} -V gendb://Joint_Genotyping_{wildcards.interval} -G StandardAnnotation -O {output.joint_gvcf_first}
Merge VCFs (vcftools v0.1.16)	vcf-concat -f /path/list_vcf_files.txt | bgzip -c > {output.vcf_first}
Vcf filtering (bcftools v1.9, vcftools v0.1.16)	#Filter VCF around indels bcftools filter -g 5 --output {output.vcf_indel5bp} {input.vcf} #Keep biallelic snps vcftools --gzvcf {input.vcf_indel5bp} --remove-indels --max-alleles 2 --recode --stdout | bgzip > {output.vcf_indel5bp_snponly} #Filter missing data vcftools --gzvcf {input.vcf_indel5bp_snponly} –max-missing {params.per} --recode --stdout | bgzip > {output.vcf_indel5bp_snponly_missing} #Remove sites with extremely high depth vcftools --exclude-positions {list} –vcfgz {output.vcf_indel5bp_snponly_missing} --recode --recode-INFO-all --out {output.vcf_indel5bp_snponly_missing_depth}
ANGSD (v0.933)	#ANGSD all SNPs angsd -bam {bam_file_list} -GL 2 -trim 5 -setMaxDepth {value} -doMajorMinor 1 -minMapQ 30 -minQ 20 -doMaf 1 -SNP_pval 1e-6 -minInd 45 -doCounts 1 -minMaf 0.05 -doGlf 2 -uniqueonly 1 -remove_bads 1 -C 50 -baq 1 -doCov 1 -doIBS 2 -makeMatrix 1 -ref referencegenome_Hgutt_V1.fa -out {prefix} -P $ 2>log.txt
ANGSD (v0.933)	#ANGSD markers in linkage equilibrium angsd -bam {bam_file_list} -sites {list_sites} -rf chrs.txt -GL 2 -trim 5 -setMaxDepth {value} -doMajorMinor 1 -minMapQ 30 -minQ 20 -doMaf 1 -minInd 40 -doCounts 1 -minMaf 0.05 -doGlf 2 -uniqueonly 1 -remove_bads 1 -C 50 -baq 1 -doCov 1 -doIBS 2 -makeMatrix 1 -ref referencegenome_Hgutt_V1.fa -out {prefix} -P $ 2>log.txt

SNPRelate v1.28.0, SeqVartools v1.38.0, R v4.3.0)	showfile.gds(closeall=TRUE)
	VCF_PATH="/path/file.vcf.gz"
	
	if (file.exists("file.gds")==F){
	vcf.fn <- VCF_PATH
	seqVCF2GDS(vcf.fn, "file.gds")
	}
	
	#OPEN GDS (can start from here)
	genofile <- seqOpen("/path/file.gds")
	
	#Perform PCA
	pca <- snpgdsPCA(genofile, num.thread=5,autosome.only = T, maf=0.05)

lostruct (v0.0.0.9)	snps<-vcf_windower("/path/file.bcf",size=5000,type='bp') pcs <- eigen_windows(snps,k=2) write.table(pcs,"filename.txt")
genomics_general (python v3.9.13)	#Parse VCF and generate .geno format python /path/parseVCF.py -i file.vcf --skipIndels --minQual 30 --gtf flag=DP min=2 max=100 -o file.geno.gz # popgenWindows.py #diversity and divergence along the genome #Example for contrast between pops Mu and Ma python /path/popgenWindows.py --windType coordinate -w 25000 -m 100 -g file.geno.gz -o file.csv.gz -f phased -T $ -p Mu HguttMu1,HguttMu2,HguttMu3,HguttMu4,HguttMu5 -p Ma Hgutt_Hy_2017_03,Hgutt_Hy_2017_04,Hgutt_Hy_2017_05,Hgutt_To_2016_01,Hgutt_To_2016_02 #Scripts available at https://github.com/simonhmartin/genomics_general
Heterozygosity (vcftools v0.1.16)	vcftools --gzvcf file.vcf.gz --het --out prefix
BAMscorer (v1.4)	#select SNPs BAMscorer select_snps file.vcf.gz output_prefix --numchrom $ #Manual step to get 3 files with the individuals of each karyotype #{OUT}_AA_individuals.txt #{OUT}_BB_individuals.txt #{OUT}_db_individuals.txt #score bam files BAMscorer score_bams file.vcf.gz output_prefix path_to_bams
Twisst	#Produce a VCF with ANGSD including other spp. angsd -bam bamlist -r $ -ref referencegenome_Hgutt_V1.fa -GL 2 -doPost 1 -doGeno 1 -trim 5 -setMaxDepth {value} -doMajorMinor 1 -minMapQ 30 -minQ 20 -doMaf 1 -minInd 8 -doCounts 1 -doGlf 2 -uniqueonly 1 -remove_bads 1 -C 50 -baq 1 -doIBS 2 -doCov 1 -makeMatrix 1 -doBcf 1 --ignore-RG 0 -out {prefix} -P $ 2>log.txt #Remove lines with heterozygous genotypes bcftools view file.bcf| grep -v "0/1" > file_noHet.vcf #Remove lines where all species are "0/0" bcftools view file_noHet.vcf| grep "1/1" > tmp.vcf bcftools view -h file_noHet.vcf > VCF_header cat VCF_header tmp.vcf > file_noHet_variable.vcf #also added the header rm tmp.vcf #Make it a "phased" VCF bcftools view -H file_noHet_variable.vcf | sed 's/\/|/g' > tmp.vcf cat VCF_header tmp.vcf > file_noHet_variable_phased.vcf #also added the header rm tmp.vcf VCF_header bgzip file_noHet_variable_phased.vcf #Convert to .geno format

	python /path/parseVCF.py -i file_noHet_variable_phased.vcf.gz --skipIndels --gtf flag=DP min=2 max=100 | gzip > file_noHet_variable_phased.geno.gz #Infer the trees python path/phyml_sliding_windows.py --minPerInd 25 -T $ -g file_noHet_variable_phased.geno.gz --prefix prefix.phyml_bionj.w50 -w 50 --windType sites --model GTR --optimise n #Get topology weights python /path/twisst.py -t prefix.phyml_bionj.w50.trees.gz prefix.phyml_bionj.w50.weights.csv.gz -g A -g B -g C -g D -g E -g F -g G -g H --groupsFile groups.tsv #Scripts can be found at https://github.com/simonhmartin/twisst
SHAPEIT (v4.2.2)	#without gmap for i in $(cat scaffold_list); do shapeit4 \ --input file.bcf \ --region ${i} \ --effective-size X \ --output prefix_phased_${i}.vcf.gz \ --log phasing.log; done

tsinfer & tsdate	#See the tutorial: https://tskit.dev/tsinfer/docs/stable/tutorial.html
	
	#create samples using tsinfer_create_input_files.py
	#the function `add_diploid_sites` iterates over the variants in a `CYVCF2.VCF` object and adds them to a `tsinfer` sample data file.
	
	#run tsinfer using tsinfer_infer.py
	#The tree object is inferred with the tsinfer.infer function
	
	#run tsdate using tsdate_infer_part1.py
	#Use tsdate.build_prior_grid to specify a prior
	#Run tsdate.date to date the nodes
	
	#extract TMRCA using tsinfer_extract_info_v2.py
	#Randomly sample two leaves in each topology and extract information such as TMRCA
	
	a=1
	block_start=0
	for tree_index in range(1,len(breaks)-2):
	a=a+1
	block_start=block_start+1
	for times in range(1,3):
	inds_list = random.sample(range(96), 2)
	ind1 = inds_list[0]
	ind2 = inds_list[1]
	IND1+=[ind1]
	IND2+=[ind2]
	BLOCK_LENGTH+=[breaks[a]-breaks[block_start]]
	POSITION+=[breaks[a]]
	time=dated_ts.at_index(tree_index).tmrca(ind1,ind2)
	DIVERGENCE+=[time]
	POP1+=[corr[ind1]]
	POP2+=[corr[ind2]]

Analysis of SVs

Step	Command
EMA (v0.6.2)	#Interleave read files parallel -j20 --bar 'paste <(pigz -c -d {} | paste - - - -) <(pigz -c -d {= s:_R1_:_R2_: =} | paste - - - -) | tr "\t" "\n" | ema count -w ./barcode_list.txt -o {/.} 2>{/.}.log' ::: *_R1_*.gz #Preprocess 10x data and insert BX:Z tags paste <(pigz -c -d *_R1_*.gz | paste - - - -) <(pigz -c -d *_R2_*.gz | paste - - - -) | tr "\t" "\n" | ema preproc -w ./barcode_list.txt -b -n 500 -t 20 -o output_dir *.ema-ncnt 2>&1 | tee preproc.log #Concatenate files cat ema-bin-* > Hgutt_DEDUP_REGEN_S1_Interleaved_BXed.fastq
LRez (v2.2.4)	#Build barcode index LRez index fastq -f Hgutt_DEDUP_REGEN_S1_Interleaved_BXed.fastq -o Barcode_Index.bci -t 20
BWA (v0.7.17)	#Map linked-reads to the reference genomes Hgutt_V1 and HguttRefA bwa mem -C -t 20 Hippocampus_guttulatus_v1.fasta -p Hgutt_DEDUP_REGEN_S1_Interleaved_BXed.fastq.gz > Hgutt_10XLR_Aligned_V1.sam bwa mem -C -t 20 GCA_025802095.1_ASM2580209v1_genomic.fna -p Hgutt_DEDUP_REGEN_S1_Interleaved_BXed.fastq.gz > Hgutt_10XLR_Aligned_RefA.sam #Sort and convert to bam and make index samtools sort Hgutt_10XLR_Aligned_V1.sam -@ 4 -O bam -l 0 -m 2G -o Hgutt_10XLR_Aligned_V1.bam samtools index -b Hgutt_10XLR_Aligned_V1.bam samtools sort Hgutt_10XLR_Aligned_RefA.sam -@ 4 -O bam -l 0 -m 2G -o Hgutt_10XLR_Aligned_RefA.bam samtools index -b Hgutt_10XLR_Aligned_RefA.bam
MTG-Link	#Generate the input GFA file bed2gfa.py -bed Hgutt_14_1698000_1700000.bed -fa Hippocampus_guttulatus_v1.fasta -out Hgutt_14_1698000_1700000.gfa #Run MTG-Link mtglink.py DBG -gfa Hgutt_14_1698000_1700000.gfa -bam Hgutt_10XLR_Aligned_V1.bam -fastq Hgutt_DEDUP_REGEN_S1_Interleaved_BXed.fastq -index Barcode_Index.bci -k 61 51 41 31 21 -t 20
Minimap2 (v2.26)	# Align assembled contig on the reference genome minimap2 -a Hippocampus_guttulatus_v1.fasta Hgutt_14_1698000_1700000.gfa.14_0-1698000-L+_14_1700000-9878394-R+.g2000.flank10000.occ2.k61.a3.bxu.insertions_filtered_quality.fasta > 3kb_contig.sam
Leviathan (v1.0.2)	#Extract data mapping to Chr4.1 of HguttRefA samtools view -b Hgutt_10XLR_Aligned_RefA.bam JAOYMQ010000004.1 > Hgutt_10XLR_Aligned_RefA_4.1.bam samtools index Hgutt_10XLR_Aligned_RefA_4.1.bam #Build LRez barcode index LRez index bam -p -b Hgutt_10XLR_Aligned_RefA_4.1.bam -o Hgutt_10XLR_Aligned_RefA_4.1.bci #Run Leviathan on Chromosome 4.1

	LEVIATHAN -b Hgutt_10XLR_Aligned_RefA_4.1.bam -i Hgutt_10XLR_Aligned_RefA_4.1.bci -g GCA_025802095.1_ASM2580209v1_genomic.fna -o Hgutt_10XLR_Aligned_RefA_4.1.vcf

Résumé : De nombreuses espèces se subdivisent en formes phénotypiquement et génétiquement différenciées qui sont associées à des variations d'habitat à fine échelle. Ces écotypes peuvent représenter une étape intermédiaire dans la formation de nouvelles espèces et offrent donc des modèles utiles pour comprendre le processus de spéciation. Des questions importantes restent en suspens quant à la manière dont les adaptations locales, les contingences historiques et les composantes de l'architecture du génome interagissent dans la formation des écotypes. Cette thèse avait pour objectif d'étudier la subdivision écotypique à travers un cadre comparatif réalisé dans contexte biogéographique similaire. Nous avons ainsi étudié cinq espèces de poissons marins de l'Atlantique Nord-Est et de la Méditerranée: l'anchois européen (*Engraulis encrasicolus*), l'hippocampe moucheté (*Hippocampus guttulatus*), l'athérine (*Atherina boyeri*), le crénilabre cendré (*Symphodus cinereus*) et le syngnathe siphonostome (*Syngnathus typhle*). Ces espèces occupent une variété d'habitats différents le long du gradient écologique mer-lagune, et la comparaison de leurs histoires évolutives peut révéler des aspects importants liés à la formation des écotypes. Nous avons cherché à caractériser les rôles relatifs de l'écologie, des contingences historiques et de l'architecture génomique dans la détermination des trajectoires évolutives des paires d'écotypes chez chaque espèce. En utilisant des données de séquençage du génome entier, nous avons cherché à tester (i) si les différences génétiques sont associées aux différents types d'habitat et (ii) comment celles-ci sont maintenues en présence de flux génique. (iii) Nous avons évalué dans quelle mesure l'architecture génomique participe au maintien de la différenciation écotypique et (iv) si ces différences proviennent de nouvelles mutations, de variations génétiques pré-existantes ou de variations introgressées. Enfin, nous avons cherché (v) à caractériser le contexte historique de la divergence écotypique. Dans le chapitre I, nous étudions la structure écotypique chez *E. encrasicolus* - une espèce pélagique très mobile présentant des écotypes marins et côtiers à une large échelle géographique. Nous avons identifié de multiples variants structuraux (VSs) qui sous-tendent la différenciation écotypique et qui ont probablement été introgressés à partir d'une troisième lignée présente dans le sud de l'océan Atlantique. Dans le chapitre II, nous étudions deux VSs qui différencient les lignées géographiques et écotypiques chez *H. guttulatus*. Nos résultats montrent qu'ils correspondent à d'anciens polymorphismes intraspécifiques d'inversions chromosomiques, soumis à des dynamiques évolutives différentes et contribuant différemment à la différentiation entre écotypes. Enfin, dans le chapitre III, nous comparons les patrons éco-géographiques et les architectures génomiques associées des écotypes des cinq espèces. Nous avons constaté que la structure écotypique était généralement plus prononcée dans la Méditerranée que dans l'Atlantique, ce qui indique probablement l'influence d'une histoire biogéographique commune. De plus, la comparaison des paysages de divergence entre espèces a révélé que les grands VSs, tels que les inversions chromosomiques, sont régulièrement impliqués dans la différenciation écotypique. En raison de leur effet suppresseur sur la recombinaison, les VSs maintiennent les combinaisons alléliques impliquées dans différentes formes d'adaptations, et pourraient ainsi agir comme des barrières au flux génique entre des lignées. Bien que la présence d'un seul VS ne permette pas l'isolement reproductif, l'évolution d'un déséquilibre de liaison entre plusieurs VSs pourrait contribuer à renforcer l'isolement reproductif, même s'il n'est pas certain que cette condition soit suffisante pour achever la spéciation.

Mots-clés: Spéciation, poissons marins, écotypes, histoire évolutive, variants structuraux.

Abstract: Many species show subdivision into phenotypically and genetically differentiated forms that are associated with fine-scale habitat variation. These ecotypes may represent an intermediate stage to the formation of new species, and thus offer key models for understanding the process of speciation. Open questions remain with respect to how local adaptations, historical contingencies and components of genome architecture interact in ecotype formation. The current thesis aimed to study ecotypic subdivision in a comparative framework controlling for a similar biogeographic context. We studied five species of marine fishes from the North East Atlantic and Mediterranean Sea: the European anchovy (*Engraulis encrasicolus*), the long-snouted seahorse (*Hippocampus guttulatus*), the big-scale sand smelt (*Atherina boyeri*), the grey wrasse (*Symphodus cinereus*), and the broadnosed pipefish (*Syngnathus typhle*). These species occur in a variety of different habitats along the marine-lagoon ecological gradient, and comparing their evolutionary histories has the potential to reveal important aspects related to ecotype formation. We wished to characterise the relative roles of ecology, historical contingencies and genomic architecture in determining the evolutionary trajectories of ecotype pairs in each species. Using whole-genome sequencing data, we aimed to test (i) whether genetic differences were associated with different habitat types, and (ii) how these are maintained in the presence of gene flow. (iii) We evaluated the extent to which the genomic architecture participates in maintaining ecotypic differentiation, and (iv) whether these differences originated from new mutations, standing genetic variation, or introgressed variation. Finally, we aimed (v) to characterise the historical context of ecotypic divergence. In Chapter I, we study ecotypic structure in *E. encrasicolus* - a highly mobile pelagic species showing marine and coastal ecotypes at a wide geographic scale. We identified multiple structural variants (SVs) that underlie ecotypic differentiation and which were likely introgressed from a third lineage in the Southern Atlantic Ocean. In Chapter II, we study two SVs segregating in *H. guttulatus*, which differentiate geographical and ecotype lineages. Our results show that these correspond to large chromosomal inversions representing ancient intraspecific polymorphisms, which are subject to different evolutionary dynamics and contribute differently to ecotype formation. Finally, in Chapter III, we compare the eco-geographic patterns and associated genome architectures of ecotypes in all five species. We found that ecotype structure was generally more pronounced in the Mediterranean as compared to the Atlantic, likely indicating the influence of a shared biogeographic history. Moreover, the comparison of divergence landscapes across species revealed that large SVs, such as chromosomal inversions, are consistently involved in ecotypic differentiation. Due to their suppressive effects on recombination, SVs maintain allelic combinations and could act as barriers to gene flow between diverging lineages experiencing gene flow. Although a single SV might not be sufficient for ensuring reproductive isolation, the build-up of linkage disequilibrium among multiple SVs could help strengthen reproductive isolation, although it remains unclear whether this is a sufficient condition for speciation to complete.

Keywords: Speciation, marine fishes, ecotypes, evolutionary history, structural variants.