Intégrer des approches expérimentales et d'évolution moléculaire en génomique de la spéciation afin d'identifier les mécanismes impliqués dans la divergence entre bar atlantique et loup méditerranéen

Maud Duranton

▶ **To cite this version:**

Maud Duranton. Intégrer des approches expérimentales et d'évolution moléculaire en génomique de la spéciation afin d'identifier les mécanismes impliqués dans la divergence entre bar atlantique et loup méditerranéen. Sciences agricoles. Université Montpellier, 2019. Français. NNT : 2019MONTG031 . tel-02481233

HAL Id: tel-02481233
https://theses.hal.science/tel-02481233

Submitted on 17 Feb 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

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

En génétique et génomique

École doctorale GAIA

Unité de recherche ISEM

Intégrer des approches expérimentales et d'évolution moléculaire en génomique de la spéciation afin d'identifier les mécanismes impliqués dans la divergence entre bar atlantique et loup méditerranéen

Présentée par Maud DURANTON

Le 15 novembre 2019

Sous la direction de François BONHOMME et Pierre-Alexandre GAGNAIRE

Devant le jury composé de

Laurence DESPRES, Professeur, Université Grenoble Alpes	Rapporteur
Christophe LEMAIRE, Maître de conférences, Université d'Angers	Rapporteur
Tatiana GIRAUD, Directrice de recherche, CNRS	Examinatrice
Carole SMADJA, Directrice de recherche, CNRS	Présidente du jury
François BONHOMME, Directeur de recherche, CNRS	Directeur
Pierre-Alexandre GAGNAIRE, Chargé de recherche, CNRS	Co-directeur

UNIVERSITÉ DE MONTPELLIER

« Not only does God play dice, but... he sometimes throws them where they cannot be seen. »

Stephen Hawking

À mon (futur ?) mari ^^

Résumé:

La spéciation est un processus évolutif permettant la formation de nouvelles espèces grâce à l'établissement progressif de barrières d'isolement reproductif entre populations en cours de divergence. Comprendre de quoi sont constituées ces barrières, quelles sont les forces évolutives ayant permis leur mise en place et comment elles impactent la valeur sélective des individus hybrides sont des questions fondamentales en biologie évolutive. Cependant, répondre à ces questions en étudiant de vraies espèces qui n'interagissent plus au travers d'échanges génétiques peut s'avérer difficile tant les barrières d'isolement reproductif sont nombreuses. C'est pourquoi nous avons choisi d'étudier dans cette thèse le bar européen (*Dicentrarchus labrax*), une espèce de poisson marin subdivisée en deux lignées évolutives représentées par les populations atlantique (bar) et méditerranéenne (loup). Afin de comprendre comment la divergence s'est établie et maintenue entre ces deux lignées, nous avons combiné plusieurs approches. Dans un premier temps, une étude de génomique des populations sur des individus sauvages nous a permis de préciser le contexte démographique dans lequel la divergence a eu lieu et d'identifier les mécanismes évolutifs ayant permis à la différenciation génétique de s'établir. Nous avons alors confirmé que les variations chromosomiques du taux de recombinaison influencent la persistance dans le génome des gènes impliqués dans l'isolement reproductif et que ces régions présentaient des niveaux de divergence particulièrement élevés, que nous avons pu relier à la présence d'allèles introgressés anciens. En effet, il semblerait que des échanges génétiques ayant eu lieu par le passé entre la population atlantique et une espèce proche, le bar moucheté (*Dicentrachus punctatus*), aient facilité la mise en place de l'isolement reproductif entre les deux lignées actuelles de bar européen. Dans un deuxième temps, nous avons étudié les patrons d'évolution moléculaire des gènes qui participent à l'isolement reproductif. Nous avons alors montré que ce sont principalement des gènes sous fortes contraintes évolutive subissant de fortes pressions de sélection. Dans un troisième temps, nous avons utilisé des croisements expérimentaux afin de déterminer s'il existe une dépression d'hybridation chez les premières générations hybrides. Nous avons ainsi pu constater que la valeur sélective d'individus issus d'une première génération de rétrocroisement n'était pas moindre que celle de leurs parents méditerranéens. Il semblerait également que l'introgression d'allèles atlantiques soit favorisée chez ces hybrides aux locus où elle est contre-sélectionnée sur le long terme, révélant une dynamique complexe de sélection des haplotypes atlantiques introgressés. Ainsi, cette thèse aura influencé la vision classique du processus de spéciation qui est généralement pensée comme l'accumulation progressive de la divergence génétique entre deux populations, facilitée par l'absence de flux génique. Ici nous avons montré qu'au contraire, les échanges génétiques ayant eu lieu avec une troisième lignée ont probablement accéléré l'émergence de l'isolement reproductif entre les deux lignées de bar. De plus, l'isolement reproductif est généralement supposé passé par une forte contre sélection des hybrides de première génération. Or, nous avons montré que ce n'était probablement pas le cas chez le bar européen et que la dynamique de sélection contre l'introgression pouvait être plus complexe, le sens s'inversant au fil des générations dû aux effets de liaison entre allèles introgressés. La sélection en liaison et donc les variations locales du taux de recombinaison semblent donc jouer un rôle fondamental dans la mise en place et le maintien de l'isolement reproductif.

Mots clés : Génomique – Spéciation - Phasage haplotypique - Isolement reproductif - Contact secondaire - Evolution moléculaire.

Abstract:

Speciation is the evolutionary process allowing species formation through the progressive establishment of reproductive isolation barriers between diverging populations. Understanding what kind of loci constitute these barriers, what evolutionary forces enabled their formation and how they impact the fitness of hybrids are fundamental questions in evolutionary biology. However, studying true species that do not interact through genetic exchanges to answer these questions can be difficult as many reproductive isolation barriers exist, making the identification of the first barriers to appear very hard. That is why we decided to study the European sea bass (*Dicentrarchus labrax*), a marine fish species subdivided into two evolutionary lineages, represented by the Atlantic and Mediterranean population. In order to understand how divergence built up and maintained between these two lineages, we combined several different approaches. First, a population genomic study of wild individuals allowed us to specify the demographic context in which divergence took place and to identify the evolutionary mechanisms that allowed genetic differentiation to increase. We found that chromosomal variations of recombination rate influence the establishment of reproductive isolation. Furthermore, genomic regions involved in reproductive isolation showed particularly high levels of divergence, that we relate to the presence of old introgressed alleles. Indeed, it seems that old genetic exchanges between *D. labrax* Atlantic population and a closely related species the spotted sea bass (*Dicentrarchus punctatus*), facilitated the establishment of reproductive isolation between the two *D. labrax* lineages. Secondly, we studied molecular evolution patterns of genes involved in reproductive isolation. We showed that they are mainly genes under strong evolutionary constraint and thus undergoing strong selective pressures. Thirdly, we used experimental crossing to determine if backcrossed individuals have a reduced fitness, which could be expected if there is hybridization load. We observed that backcrossed individuals had the same fitness than their Mediterranean parents. It would also appear that the introgression of Atlantic alleles is favored for these hybrids at the same loci where they experience negative selection over the long-term, revealing a complex dynamic of selection on introgressed Atlantic haplotypes. Therefore, this work has influenced the classical view of speciation which is generally thought as the progressive accumulation of divergence between two populations, facilitated by the absence of gene flow. Here we showed that, on the contrary, genetic exchanges between *D. labrax* Atlantic population and a third lineage probably accelerated the emergence of reproductive isolation between the two lineages of European sea bass. Furthermore, reproductive isolation is generally assumed to be especially efficient if first-generation hybrids are strongly counter-selected. However, we showed that this is probably not the case for the European sea bass and that the dynamic of selection on introgression could be more complex, the direction being reversed over generation due to the effects of linkage between introgressed alleles. Linked selection and local variation of recombination rate thus seem to play a fundamental role in the establishment and maintenance of reproductive isolation.

Key words: Genomic – Haplotype phasing – Reproductive isolation – Secondary contact – Molecular evolution.

Liste des publications:

Duranton M., F. Allal, C. Fraïsse, N. Bierne, F. Bonhomme and P.-A. Gagnaire, 2018 The origin and remolding of genomic islands of differentiation in the European sea bass. Nature Communications 9: 2518. https://doi.org/10.1038/s41467-018-04963-6

Duranton M., F. Allal, S. Valière, O. Bouchez, F. Bonhomme and P.-A. Gagnaire 2019 The contribution of ancient admixture to reproductive isolation between European sea bass lineages. Evolution Lettrers (Second round of review)

Duranton M., F. Bonhomme, and P.-A. Gagnaire, 2019 The spatial scale of dispersal revealed by admixture tracts. Evolutionary Applications 0. https://doi.org/10.1111/eva.12829

Duranton M., M. Rousselle, F. Schlichta, F. Bonhomme and P.-A. Gagnaire 2019 The relation between recombination, reproductive isolation and patterns of molecular evolution in the European sea bass genome. Molecular Biology and Evolution (being submitted)

Simon A., and M. Duranton, 2018 Digest: Demographic inferences accounting for selection at linked sites†. Evolution 0. https://doi.org/10.1111/evo.13504

TABLE DES MATIÈRES

INTRODUCTION

I. La spéciation

1. Concepts et définitions

a. Sélection naturelle et dérive génétique

La spéciation est le processus évolutif permettant l'apparition de deux nouvelles espèces (ou plus) à partir d'une espèce ancestrale. Comprendre comment elle se déroule et les mécanismes évolutifs qu'elle implique est l'un des principaux objectifs de la biologie évolutive. Pour ce faire, il est nécessaire de déterminer quand (et pourquoi) elle commence et se termine et donc d'avoir une définition claire de ce qu'est une espèce. C'est ce qu'a proposé Mayr (1942) en énonçant le « concept biologique de l'espèce ». Il définit alors une espèce comme un ensemble d'individus effectivement ou potentiellement capables de se reproduire et d'engendrer une descendance viable et féconde, tout en étant reproductivement isolés d'autres groupes similaires d'individus. L'isolement reproductif devient alors un paramètre clé. Dans ce cadre-là, la spéciation s'achève quand les échanges génétiques entre populations ne sont plus possibles, c'est-à-dire quand il n'y a plus de flux génique. On peut alors décrire la spéciation comme un processus graduel au cours duquel au moins deux populations divergent génétiquement et accumulent des barrières d'isolement reproductif jusqu'à être complètement isolées reproductivement (Coyne and Orr 2004). Identifier les barrières d'isolement reproductif est alors devenu un des objectifs principaux des études de spéciation.

La question de la force évolutive principale permettant l'accumulation des différences génétiques au cours du temps a cependant longtemps été débattue au sein de la communauté scientifique. En effet, le devenir des mutations spontanées (c'est-à-dire leur fixation ou leur élimination du patrimoine génétique de la population dans laquelle elles sont apparues), qui constituent la principale source de variation génétique, est déterminé par deux forces évolutives principales : la sélection naturelle et la dérive génétique. La théorie synthétique de l'évolution, qui découle directement des travaux de Darwin, a d'abord proposé que la sélection naturelle soit la force prédominante (Huxley 1942). D'après cette théorie, les nouvelles mutations génétiques ont principalement un effet positif ou négatif sur la valeur sélective des individus, c'est-à-dire sur leur capacité à survivre et à se reproduire dans un environnement donné. L'effet d'une mutation est donc conditionné par l'environnement dans lequel elle se trouve, ce qui génère des pressions de sélection environnementales. Les mutations délétères sont négativement sélectionnées, leur fréquence dans la population diminue donc jusqu'à élimination. A l'inverse, les mutations favorables sont positivement sélectionnées et augmentent en fréquence jusqu'à fixation. La vitesse à laquelle la fréquence des mutations évolue dépend directement de l'intensité de la sélection qu'elles subissent (leur coefficient de sélection noté s) et donc de leurs effets plus ou moins importants sur la valeur sélective des individus. Au fil des générations, seules les

mutations positives sont supposées perdurer ; c'est l'adaptation. Quand les pressions de sélection sont générées directement par l'environnement, on parle d'adaptation locale. D'après la théorie synthétique de l'évolution, les variations environnementales (comme les gradients de température latitudinaux qui génèrent des variations de pressions de sélection) et l'apparition des mutations positives (qui reste un processus aléatoire) constituent donc le principal moteur de l'évolution.

Plus tard, Kimura énonça une théorie radicalement opposée : la théorie neutre de l'évolution (Kimura 1968). Il propose que la plupart des nouvelles mutations génétiques sont neutres, c'est-à-dire qu'elles n'ont pas d'effets sur la valeur sélective des individus (s=0). Elles ne subissent donc pas de pressions de sélection et leur devenir est uniquement influencé par la dérive génétique (Kimura 1983), c'est-à-dire par des processus aléatoires comme le hasard de la survie et de la reproduction des individus. Ainsi, sous ce modèle, la majorité des changements qui s'accumulent au cours du temps n'ont pas d'effet sur la valeur sélective des individus et sont conservés uniquement sous l'effet du hasard. L'intensité de la dérive génétique dépend cependant de la taille efficace de la population (N_e) qui est définie comme l'effectif d'une population idéale (*càd.* d'effectif constant, non structurée, à générations non chevauchantes et sans sélection) présentant les mêmes caractéristiques de dérive génétique que la population observée. En effet, la force de la dérive génétique étant égale à $\frac{1}{2Ne}$ (pour les organismes diploïdes), plus la taille efficace d'une population est importante et moins la dérive aura d'influence. Plus tard, Kimura et Ohta introduisent la notion de mutations quasi neutres (Kimura and Ohta 1971) c'est-à-dire avec un coefficient de sélection non nul, mais inférieur à $\frac{1}{2Ne}$ et qui évoluent donc principalement sous l'effet de la dérive génétique. Cette théorie propose donc la dérive génétique comme principale source de changements évolutifs au cours du temps, faisant de la démographie un paramètre clé de l'évolution. Ainsi, d'après la théorie neutre, il n'y a pas que les mutations positives qui se maintiennent au cours du temps, mais également des mutations neutres ou faiblement délétères qui peuvent se fixer par dérive génétique, c'est-à-dire par hasard.

Les partisans des deux théories, les « sélectionnistes » et les « neutralistes », continuent encore aujourd'hui à débattre de leur validité (Kern and Hahn 2018). Pourtant, la théorie neutre ne remet pas en cause l'existence de la sélection naturelle, mais suppose plutôt que les mutations avantageuses sont rares par rapport aux mutations neutres et délétères, et ce pour deux raisons principales. Premièrement, parce que les séquences codantes (sur lesquelles la sélection agit) ne représentent qu'une faible proportion de la totalité des génomes eucaryotes et deuxièmement, parce qu'une mutation à plus de chance de détériorer que d'améliorer une séquence protéique. En effet, les protéines subissent de fortes contraintes évolutives, leurs séquences ayant déjà été optimisées par la sélection naturelle. Ainsi, les deux théories ne sont pas incompatibles, mais évaluent différemment le

rôle relatif de la sélection naturelle et de la dérive génétique. Il existe en réalité probablement tout un gradient, des mutations délétères en passant par les mutations neutres ou quasi-neutres jusqu'aux mutations adaptatives. Pour comprendre quelle force évolutive influence réellement le devenir d'une mutation, il faut donc principalement s'intéresser à la force relative de la sélection par rapport à celle de la dérive génétique (s par rapport à $\frac{1}{2Ne}$). Il est donc nécessaire de connaître le contexte démographique dans lequel la spéciation s'est déroulée.

b. Biogéographie de la spéciation

Le contexte spatial a également son importance étant donné qu'il va influencer l'intensité des forces évolutives mises en jeu (Coyne and Orr 2004). On distingue principalement trois modes de spéciation : allopatrique, sympatrique et parapatrique.

La spéciation allopatrique

Sous ce modèle, la divergence évolue entre deux populations isolées par une barrière physique extrinsèque (ex : une chaîne de montagne, un front courantologique) empêchant les échanges génétiques. Initialement, les deux populations isolées géographiquement partagent un grand nombre d'allèles polymorphes hérités de la population ancestrale. La dérive génétique, qui agit de façon indépendante dans les deux populations, va alors faire évoluer différemment les fréquences alléliques de ces allèles. C'est le tri indépendant de ce polymorphisme ancestral entre les deux populations qui va principalement permettre à la différenciation génétique de se construire dans les premières générations de divergence. En effet, on considère généralement qu'après $10N_e$ générations de divergence le tri du polymorphisme ancestral est terminé ou quasiment. Or, en parallèle de ce tri, de nouvelles mutations apparaissent à un rythme constant indépendamment dans chaque population. Ainsi, la divergence génétique qui s'établie initialement à partir du tri du polymorphisme ancestral est progressivement nourrie par l'apparition de ces nouveaux allèles. L'établissement de la divergence commence donc grâce à des mécanismes évolutifs stochastiques, comme la dérive génétique, et est largement influencé par la démographie des populations, puis c'est la mutation qui prend le relais. Si les conditions environnementales (biotiques et abiotiques) sont différentes des deux côtés de la barrière, la sélection naturelle va pouvoir accélérer la divergence génétique et faciliter la mise en place de barrières d'isolement reproductif. Cependant, la sélection divergente n'est pas indispensable puisque l'isolement reproductif évolue ici comme une conséquence secondaire inévitable de l'accumulation (induite par l'absence de flux génique) de différences génétiques entre les deux populations.

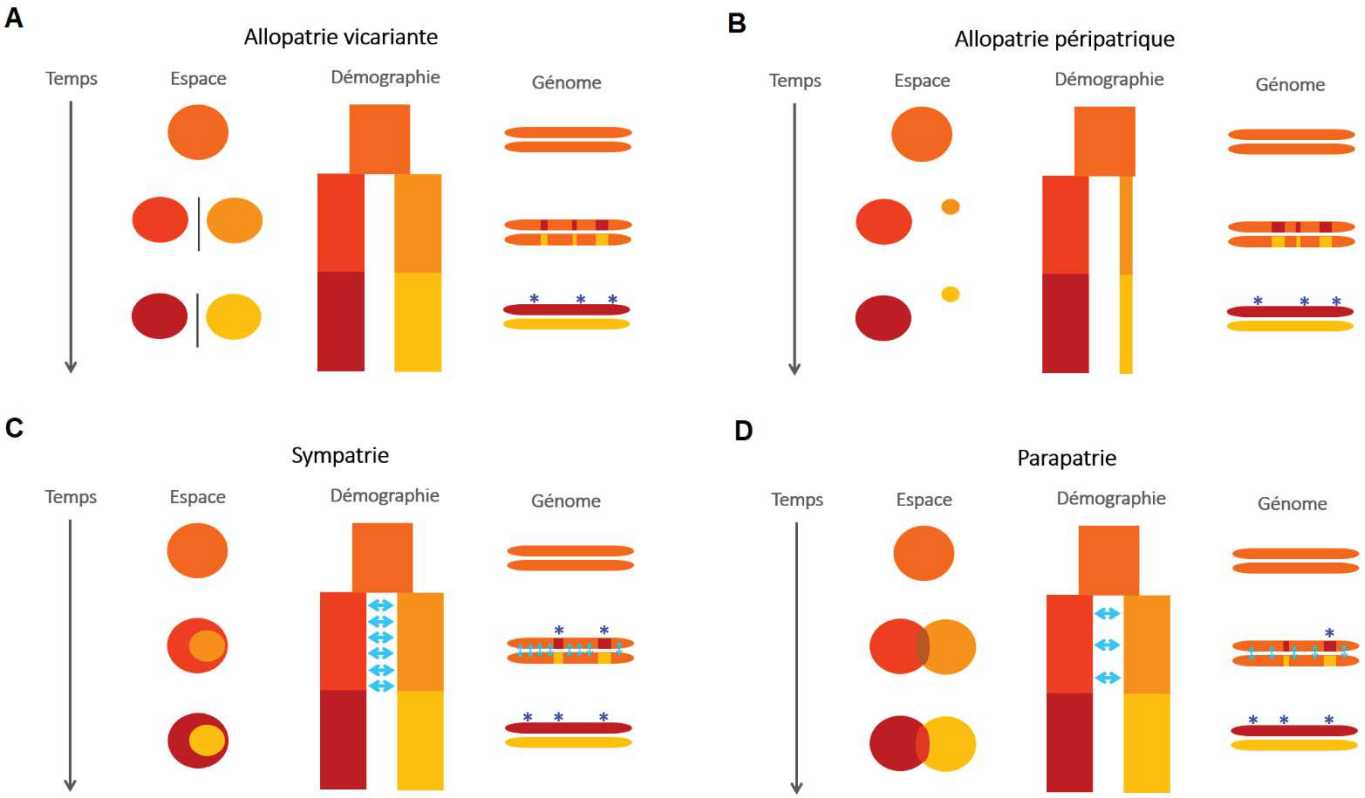


FIGURE 1 – Déroulement des différents modes de spéciation dans l'espace et le temps ainsi que leur impact sur la démographie et la différenciation génétique des populations. Une population ancestrale se divise en deux populations filles qui accumulent progressivement des différences génétiques et des locus d'isolement reproductif (étoiles bleues) jusqu'à être complétement isolées reproductivement et former deux espèces distinctes. Trois étapes sont représentées pour chaque mécanisme, avec les aires de distributions des deux populations (Espace), leur taille efficace (Démographie) avec l'intensité des échanges génétiques (flèches bleues) et le niveau de différenciation génétique (Génome) représenté le long de deux chromosomes (un chromosome par population). La spéciation peut se dérouler en allopatrie **A.** avec vicariance ou **B.** en péripatrie, **C.** en sympatrie ou **D.** en parapatrie.

On distingue parfois deux modes de spéciation allopatrique, qui se différencient principalement par la taille relative des deux populations isolées : la spéciation vicariante (Figure 1A) et péripatrique (Figure 1B). La spéciation vicariante est généralement initiée par la mise en place d'une barrière physique, qui va diviser une population ancestrale unique en deux populations de tailles similaires. La spéciation péripatrique, quant à elle, fait généralement suite au franchissement d'une barrière à la migration préexistante (montagne, fleuve, ou colonisation d'une île) par certains individus. La nouvelle population fondatrice qui se trouve de l'autre côté de la barrière a donc une taille réduite par rapport à la population originale. Cette nouvelle population peut éventuellement se retrouver dans un nouvel habitat et ainsi subir de nouvelles pressions de sélection qui induisent une sélection divergente, mais pas nécessairement. En effet, la dérive génétique peut à elle seule permettre l'établissement de la divergence étant donné qu'elle sera amplifiée par la taille réduite de la population. De plus, la nouvelle population étant établie par un petit nombre d'individus, elle subit une diminution immédiate de sa diversité génétique et un changement brutal de fréquences alléliques, ce qui augmente instantanément la différenciation génétique, c'est l'effet fondateur (Mayr 1942).

Ce mode de spéciation ne pose pas de problème théorique et est probablement un phénomène fréquent. En effet, on imagine aisément comment la formation de chaînes de montagnes, la dérive des continents ou les changements climatiques passés ont pu géographiquement isoler des populations. Les variations climatiques du Pléistocène sont notamment connues pour avoir fortement impacté les aires de répartition des espèces des régions tempérées (Hewitt 1996, 2000). De plus, on trouve dans la nature une myriade d'exemples à l'appui de la spéciation allopatrique. C'est le cas des nombreuses paires d'espèces sœurs séparées par une même barrière géographique, un argument en faveur de la spéciation vicariante. Par exemple, la fermeture de l'isthme de Panama a permis la divergence entre de nombreuses lignées d'organismes marins vivant dans les eaux du Pacifique et des Caraïbes (Knowlton Nancy and Weigt Lee A. 1998). Le fort endémisme observé sur les îles est quant à lui un argument en faveur de la spéciation péripatrique, en particulier quand une espèce sœur est retrouvée sur le contient. C'est le cas pour de nombreuses espèces endémiques de l'archipel d'Hawaï (Kaneshiro and Boake 1987; Roderick and Gillespie 1998).

La spéciation sympatrique

Le modèle de spéciation sympatrique (Figure 1C) s'oppose complètement au modèle allopatrique. Ici, la divergence évolue sans qu'il y ait de barrière aux échanges génétiques. Une forme de sélection disruptive est donc nécessaire pour initier la divergence étant donné que la dérive génétique ne joue aucun rôle (au moins au début), son effet étant contrebalancé par la migration. L'adaptation locale différentielle à deux niches écologiques distinctes pourrait par exemple induire ce genre de pressions de sélection. La spéciation sympatrique pourrait donc théoriquement se dérouler dans un

environnement spatialement hétérogène, c'est-à-dire présentant plusieurs types d'habitats. Dans ce cas-là, différents groupes d'individus pourraient se spécialiser à des habitats différents grâce à l'apparition de nouvelles mutations positivement sélectionnées dans l'un ou l'autre des environnements. Ainsi, chaque individu aura une valeur sélective élevée dans l'habitat dans lequel il s'est spécialisé et une valeur sélective réduite dans les autres habitats.

Cependant, en absence de barrières au flux génique, le brassage génétique (conséquence directe de la recombinaison génétique) induit par la reproduction avec des migrants, va casser les associations de gènes coadaptés et introduire des combinaisons alléliques défavorables. Des génotypes maladaptés seront alors produits, ce qui va s'opposer à l'adaptation locale et donc à l'établissement de la divergence (Lenormand 2002). La spéciation sympatrique ne peut donc pas se dérouler sans la mise en place de mécanismes limitant les échanges génétiques et la recombinaison, c'est-à-dire un isolement reproductif ou une forme d'association entre les gènes coadaptés (par exemple un réarrangement chromosomique comme une inversion permettant de limiter la recombinaison). Contrairement à la spéciation allopatrique où l'isolement reproductif apparait comme une conséquence de la divergence, en sympatrie, l'isolement reproductif est ici nécessaire au maintien de la divergence.

Ainsi, de nombreux modèles supposent l'apparition d'un choix d'habitat ou de partenaire sexuel parallèlement à l'adaptation à la nouvelle niche écologique (Felsenstein 1981). Une hypothèse simple permettant la spéciation sympatrique repose sur l'existence d'un gène à effet pléiotrope permettant à la fois l'adaptation à l'environnement et le choix de partenaire sexuel. On parle alors de « trait magique » (Gavrilets 2004; Servedio *et al.* 2011). Cependant, de tels gènes sont probablement rares et peu d'exemples existent dans la nature. Le plus connu reste celui des papillons mimétiques du genre *Heliconius* (Jiggins *et al.* 2005). Chez ces espèces, la couleur des ailes joue un rôle à la fois dans la reconnaissance sexuelle et dans l'adaptation locale en influençant la capacité à mimer une espèce toxique.

L'ensemble des conditions nécessaires pour permettre la spéciation sympatrique, ainsi que les nombreux paramètres influençant sa faisabilité, comme le nombre de gènes impliqués, leur niveau d'interaction, le taux de recombinaison, la force de la sélection disruptive et la stabilité de l'environnement en font un processus probablement rare (Coyne and Orr 2004). En effet, contrairement aux exemples de spéciation allopatrique, il y a peu d'exemples non controversés de spéciation sympatrique. L'exemple qui semble le plus probant est celui de la diversification rapide de nombreuses espèces de poissons de la famille des Cichlidés dans des lacs de cratères africains suite à la colonisation par un nombre restreint d'espèces. Cependant, la spéciation sympatrique a été remise

en question suite à des études de génomiques qui ont mis en évidence l'existence de flux génique secondaire après le premier évènement de colonisation (Martin *et al.* 2015). Néanmoins, le rôle de ce flux génétique secondaire dans la diversification des espèces n'ayant pas été clairement identifié (pas de traces d'introgression adaptative), l'hypothèse d'une spéciation sympatrique est encore défendue aujourd'hui (Richards *et al.* 2018).

La spéciation parapatrique

La spéciation parapatrique (Figure 1D) est un mécanisme intermédiaire entre la spéciation allopatrique et sympatrique. Sous ce modèle, la divergence entre deux populations évolue en présence de migration mais avec un flux génique limité. Ainsi, l'isolement reproductif peut se développer si l'effet homogénéisateur de la migration est inférieur à celui de la dérive génétique et de la sélection naturelle. Une forme de sélection disruptive reste donc nécessaire (bien que la dérive joue un rôle à part entière), ce qui suppose une hétérogénéité des pressions de sélection environnementale. On peut donc imaginer des populations discrètes distribuées dans des environnements différents ou une population continue distribuée le long d'un gradient environnemental. Si on envisage la sélection sympatrique comme un mécanisme plausible, les conditions moins restrictives de la spéciation parapatrique en font un mécanisme probable. Pourtant, les observations de spéciation parapatrique restent rares dans la nature. Néanmoins, ce mode de spéciation reste difficilement démontrable étant donné qu'il est difficile d'exclure l'existence d'une période d'allopatrie antérieure qui aurait permis d'initier la divergence.

Les modèles de spéciation allopatrique et sympatrique se placent donc aux deux extrêmes d'un continuum où la divergence génétique s'accumule entre les populations en absence ou en présence de flux génique et de sélection divergente. Il existe entre ces deux points tout un ensemble de modèles possibles où l'intensité des échanges génétiques varie, la spéciation parapatrique représentant un de ces intermédiaires. Cependant, ces différents modes de spéciation n'accordent pas la même importance aux rôles relatifs de la dérive génétique et de la sélection naturelle. Ainsi, on ne s'attend pas nécessairement à retrouver les mêmes mécanismes d'isolement reproductif, étant donné qu'il en existe une grande variété pouvant intervenir à différents moments du cycle de vie et étant plus ou moins dépendants de l'environnement.

2. Les différents mécanismes d'isolement reproductif

a. Isolement pré-zygotique

L'isolement pré-zygotique intervient avant la fécondation et empêche la formation d'individus hybrides. Quand la probabilité de croisement entre deux espèces est diminuée par des facteurs environnementaux, on parle d'**isolement écologique**. Deux espèces génétiquement adaptées à deux habitats différents vont vivre dans deux niches écologiques distinctes, soit parce qu'il existe un mécanisme actif de choix d'habitat, soit parce que la valeur sélective réduite des migrants dans l'autre habitat empêche leur survie (Nosil *et al.* 2005). Dans tous les cas, la probabilité de rencontre entre les deux espèces sera naturellement réduite, empêchant la reproduction interspécifique. De nombreux insectes phytophages sont par exemple spécialisés à une plante hôte sur laquelle ils vont également se reproduire (Matsubayashi *et al.* 2010). Un autre cas d'isolement écologique est dû à l'utilisation de vecteurs pour le transport des gamètes. C'est par exemple le cas des plantes angiospermes, qui ont pour la plupart recourt à des insectes pollinisateurs pour leur fécondation. Un polinisateur spécialisé à une espèce ne pourra pas polliniser la fleur d'une autre espèce même si celle-ci se trouve dans la même région géographique, réduisant les échanges génétiques. Enfin, un décalage de la période de reproduction entre deux espèces réduit également les probabilités d'hybridation, c'est l'**isolement temporel**. Ce mécanisme est assez fréquemment rencontré chez les plantes à fleurs pour qui les périodes de floraisons peuvent être décalées car déclenchées par des facteurs environnementaux différents. Ce type d'isolement reproductif évolue très probablement sous l'action de la sélection naturelle par adaptation locale et n'est donc pas attendu si les deux espèces ont évolué dans un environnement similaire.

Il existe également des mécanismes d'isolement pré-zygotique qui ne sont pas en lien avec l'environnement. De nombreuses espèces animales utilisent des mécanismes actifs de reconnaissance des partenaires sexuels qui peuvent être basés sur des signaux variés : visuels, auditifs, olfactifs ou tactiles, c'est l'**isolement comportemental**. Chez les oiseaux, le chant est notamment un critère important pour la reconnaissance entre partenaires. Ainsi, la forme du bec étant impliquée à la fois dans l'isolement reproductif (en influençant la sonorité du chant) et l'adaptation locale (en déterminant le type de ressource alimentaire accessible) peut être considéré comme un « trait magique » (Derryberry *et al.* 2012).

Plus tardivement, après la rencontre des partenaires sexuels, d'autres mécanismes d'isolement pré-zygotique peuvent également intervenir. Pour les espèces à fécondation interne il peut exister une incompatibilité morpho-anatomique des pièces génitales. De façon analogue à un mécanisme clé-serrure, l'accouplement n'est possible qu'entre individus possédant des appareils génitaux

compatibles, c'est l'**isolement mécanique**. On trouve de nombreux exemples chez les arthropodes pour qui les parties génitales mâles sont des structures parfois complexes et à évolution rapide qui peuvent être utilisées comme critère pour distinguer certaines espèces proches. Enfin le mécanisme pré-zygotique le plus tardif intervient au moment de la rencontre des gamètes, c'est l'**isolement gamétique**. Il peut être induit par des problèmes de motilité ou de viabilité des gamètes mâles hétérospécifique dans les voies génitales femelles, dans le cas d'espèces à fécondation interne (Devigili *et al.* 2018) ou à des problèmes de reconnaissance des gamètes. En effet, il faut une complémentarité moléculaire des enveloppes des gamètes pour que la fécondation ait lieu. Ce deuxième mécanisme semble particulièrement important pour les espèces à fécondation externe chez qui la rencontre entre gamètes hétérospécifique est inévitable. C'est notamment le cas de la plupart des espèces d'invertébrés marins chez qui l'isolement gamétique semble avoir joué un rôle prédominant dans le processus de spéciation (Palumbi 1992, 2009). Ces deux derniers mécanismes pourraient aussi bien évoluer sous l'action de la sélection naturelle que de la dérive génétique.

b. Isolement post-zygotique

L'isolement post-zygotique se traduit par une valeur sélective réduite des individus hybrides lié à une survie ou une stérilité réduite voire nulle. Ainsi, il peut intervenir à différents moments du cycle de vie, très tôt lorsque la fécondation a eu lieu mais que le développement embryonnaire est interrompu, ou plus tardivement si les hybrides sont stériles. De plus, Il peut se manifester dès les premières générations d'hybridation (F1) ou plus tardivement par exemple chez les hybrides de deuxième génération (F2, issus du croisement de deux individus F1) ou des individus *backcross* (issus du croisement entre un F1 et un individu de l'une des deux espèces parentales). Il arrive également que la réduction de valeur sélective ne touche qu'un seul sexe, le plus souvent le sexe hétérogamétique (càd. le sexe déterminé par des chromosomes sexuels différents, pour les espèces ayant un déterminisme du sexe chromosomique), c'est la règle de Haldane (Haldane 1922; Orr 1997).

Comme l'isolement pré-zygotique, l'isolement post-zygotique peut être dépendant de facteurs environnementaux, c'est-à-dire exogènes. Par exemple, si les individus issus de deux populations différentes sont adaptés à deux environnements différents, on peut imaginer que les hybrides vont posséder des caractères intermédiaires et donc avoir une valeur sélective réduite dans les deux habitats. Ce genre de mécanisme ne peut évoluer que si les deux espèces ont divergé dans des environnements différents. Un mécanisme génétique simple qui peut être à l'origine de ce type de dépression d'hybridation est la sous-dominance. On parle de sous-dominance quand le génotype hétérozygote a une valeur sélective inférieure à celle des deux génotypes homozygotes. Cependant, au-delà de l'adaptation à l'environnement, il existe également au sein des génomes de nombreuses

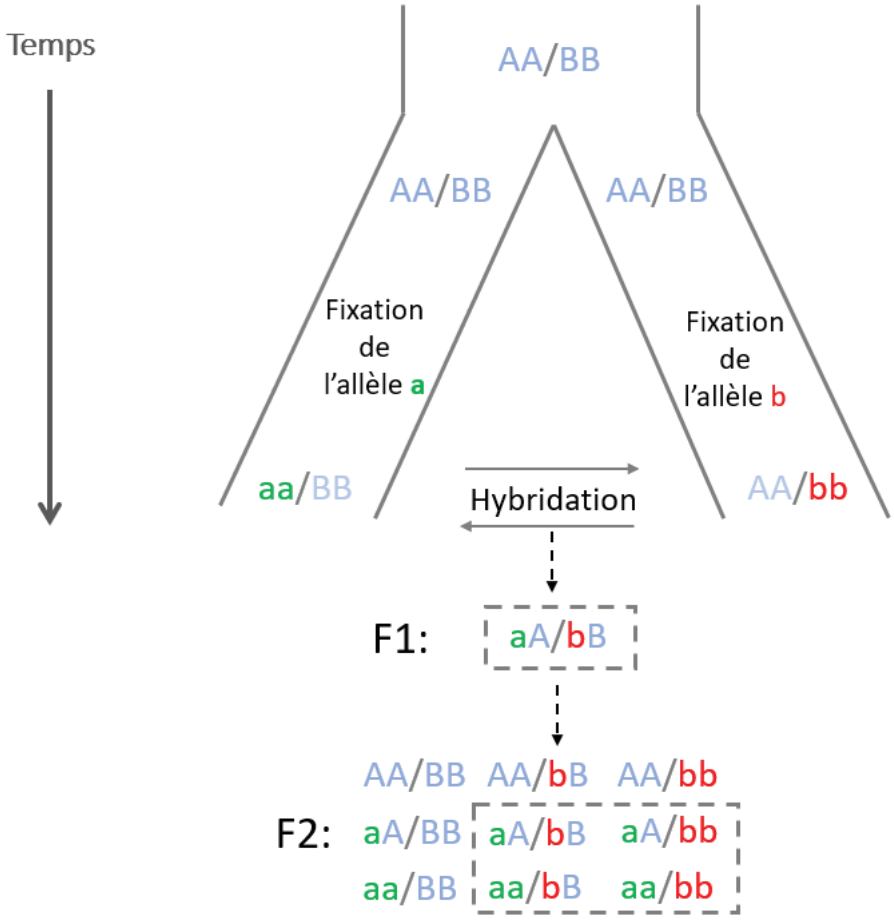


FIGURE 2 - Modèle bi-locus d'une incompatibilité de Dobzhansky-Muller. Deux lignées divergent indépendamment l'une de l'autre à partir d'une lignée ancestrale. Chaque population fixe un nouvel allèle à un locus différent. La formation de génotypes hybrides entre ces deux lignées crée de nouvelles combinaisons alléliques qui n'ont jamais été testées par la sélection naturelle et se révèlent incompatibles. Les génotypes incompatibles en cas de codominance sont encadrés.

coadaptations entre gènes. En effet, chaque gène s'exprime dans un environnement génomique donné dans lequel il est en interaction avec d'autres gènes au travers d'interactions dites épistatiques, qui peuvent être positives ou négatives et qui ne sont pas nécessairement en rapport avec l'environnement. Ainsi, la sélection naturelle va pouvoir, au fil des générations, optimiser les interactions entre les différents allèles d'un même pool génique. La sélection est alors générée par des facteurs dits endogènes.

Dobzhansky (Dobzhansky 1937) et Muller (Muller 1942) ont proposé un modèle permettant d'expliquer l'apparition d'un isolement reproductif au travers des interactions épistatiques (Figure 2). Le cas le plus simple correspond à l'interaction de deux locus bi-alléliques. Deux populations fixent indépendamment deux nouveaux allèles à deux locus différents (a et b) (Figure 2). Ces deux allèles sont neutres ou avantageux dans leur fond génétique d'origine (c'est-à-dire respectivement associé à B et A), mais leur association n'ayant jamais été testée par la sélection naturelle, elle peut se révéler délétère une fois crée dans les génotypes hybrides. La fixation de mutations faiblement délétères par dérive génétique dans l'une des populations, suivie par la fixation de mutations compensatoires (Kimura 1985) peut également être à l'origine d'incompatibilités Dobzhansky-Muller (DMI) (parfois également appelées BDMI pour Bateson-Dobzhansky-Muller *incompatibilities* (Bateson 1909)). Finalement, si les interactions entre allèles sont plus compliquées avec des relations de dominance des allèles ancestraux, les effets délétères ne seront révélés que si les allèles dérivés sont présents à l'état homozygote. Ce modèle a donc, contrairement au modèle de sous-dominance, l'avantage d'expliquer pourquoi l'isolement reproductif peut parfois apparaître dans les générations d'hybridation plus tardives (post F1). Un exemple bien connu d'interactions épistatiques négatives entre gènes d'isolement reproductif est celui souvent observé entre les gènes mitochondriaux et nucléaires impliqués dans des fonctions mitochondriales, à l'origine d'incompatibilités Dobzhansky-Muller (Burton and Barreto 2012). Quel que soit le type d'isolement reproductif impliqué, la formation d'hybrides est possible tant que l'isolement n'est pas complet, ce qui nécessite généralement la combinaison de plusieurs mécanismes. Chez les vraies espèces qui n'échangent plus génétiquement, il peut donc être difficile d'identifier parmi les nombreuses barrières d'isolement reproductif les premières à s'être mises en place et à avoir initié la spéciation (Via 2009).

3. De la génétique à la génomique des populations

a. Des allozymes aux génomes phasés

Pour comprendre comment la divergence s'établit au niveau génomique au cours du processus de spéciation et quelles sont les bases génétiques de l'isolement reproductif, il est nécessaire de décrire la variation génétique intra- et inter-populationnelle. Pour ce faire, différents types de marqueurs moléculaires ont été développés au fil du temps, permettant d'obtenir différents niveaux de précision (Sunnucks 2000). Le séquençage direct de la molécule d'ADN n'étant à l'origine pas possible, les premières études se sont concentrées sur les variations de propriétés électrophorétiques de protéines. En effet, sachant que les gènes codent pour la synthèse des protéines, comparer les propriétés d'une même protéine produite chez différents individus permet de déterminer s'il existe entre ceux-ci des différences dans la séquence du gène qui code pour cette protéine (Charlesworth and Charlesworth 2017). C'est la mobilité des protéines dans un champ électrique qui est généralement comparée, ce qui a permis de révéler l'existence d'allozymes, c'est-à-dire d'enzymes codées par différents allèles (*càd.* différentes versions) d'un même gène. L'utilisation de cette technique dans de nombreux groupes d'organismes a été à l'origine de découvertes majeures, notamment en démontrant que le polymorphisme génétique n'est pas rare contrairement à ce qui était précédemment envisagé (Lewontin and Hubby 1966) mais également en révélant l'existence d'espèces cryptiques (Chilton *et al.* 1992), c'est-à-dire morphologiquement identiques mais génétiquement différenciées. Cependant, cette technique présente certaines limitations car elle ne permet d'étudier que les séquences codantes et ne révèle que les changements nucléotidiques entraînant des changements d'acides aminés qui modifient les propriétés de la protéine.

L'utilisation des allozymes a progressivement été remplacée par des marqueurs électrophorétiques à ADN, souvent révélés via des enzymes de restriction (ex : RFLP). Dans les années 80, des avancées techniques comme l'introduction de la PCR (*Polymerase Chain Reaction*) et le développement de machines de séquençage Sanger automatisées ont rendu possible l'étude de régions spécifiques de l'ADN, permettant d'accéder directement à la séquence nucléotidique de gènes. Cependant, les coûts élevés du séquençage ont longtemps limité son utilisation à large échelle aux organismes modèles comme l'homme. Beaucoup d'études se sont alors concentrées sur l'ADN des mitochondries qui sont présentes dans la plupart des cellules eucaryotes. En effet, la transmission uniparentale de l'ADN mitochondrial ainsi que son absence de recombinaison en font un marqueur particulier qui a tendance à accumuler plus rapidement les mutations que le génome nucléaire. Les changements dans L'ADN mitochondrial peuvent donc rapidement permettre de diagnostiquer les espèces, c'est pourquoi il est largement utilisé pour les études phylogénétiques, c'est-à-dire pour l'étude des relations de parentés entre organismes (Boore and Brown 1998) et phylogéographiques. Plus récemment, l'ADN

mitochondrial a été très utilisé dans les études d'ADN environnemental qui permettent de détecter la présence d'organismes vivants dans un milieu donné (eau, sédiments, terre) grâce à l'ADN qu'ils y laissent (Taberlet *et al.* 2012). L'ADN mitochondrial est généralement utilisé comme « code barre » car il présente l'avantage d'être présent en plus grande quantité dans les cellules et permet facilement d'identifier à quel organisme il appartient grâce aux importantes bases de données disponibles. L'ADN mitochondrial environnemental est notamment utilisé en écologie pour offrir une meilleure description de la diversité spécifique de certains milieux (Rees *et al.* 2014). Cependant, l'absence de recombinaison de l'ADN mitochondrial fait qu'il se comporte comme un locus unique rendant difficile l'étude de la diversité génétique intra-espèce (Galtier *et al.* 2009). Ainsi, l'utilisation de séquences mitochondriales en génétique des populations est souvent couplée à celle de marqueurs nucléaires, notamment les marqueurs microsatellites qui correspondent à des séquences d'un a six nucléotides répétés en tandem. Ces marqueurs, largement utilisés dans les années 2000, présentent l'avantage d'être nombreux dans les génomes, d'être très polymorphes (parfois plusieurs dizaines d'allèles de longueur différente par locus) et de suivre les règles de transmission mendélienne (Li *et al.* 2002). Ils se sont donc révélés très utiles pour étudier la structure génétique et d'apparentement des populations naturelles. Néanmoins, il peut être plus difficile de les utiliser pour reconstruire l'histoire évolutive des espèces. En effet, les différences de longueurs entre allèles peuvent difficilement être reliées à leur histoire mutationnelle (en raison des risques d'homoplasie) et il peut également exister des différences importantes de taux de mutations entre les différents allèles d'un même locus (Zhang and Hewitt 2003)

Aujourd'hui, les marqueurs les plus communément utilisés sont les SNPs (*Single Nucleotide Polymorphism*) qui correspondent au changement d'une seule base dans une séquence d'ADN. Bien qu'en théorie il puisse exister pour chaque nucléotide le long du génome quatre états possibles (A, T, C ou G) les SNPs sont principalement bi-alléliques. En effet, le taux de mutation (μ) étant très faible, la probabilité d'avoir deux changements indépendants à la même position génomique est très faible. Ainsi les SNPs peuvent paraître moins informatif que les autres types de marqueurs précédemment cités qui sont tous multi-alléliques (Vignal *et al.* 2002). Malgré tout, ils présentent l'avantage de permettre d'étudier un nombre quasiment illimité de locus le long du génome. En effet, avec le développement des nouvelles technologies de séquençage à haut débit, il est désormais possible de séquencer à relativement faible coût des génomes entiers pour des organismes non modèles (Therkildsen and Palumbi 2017). De nombreuses méthodes ont donc été développées pour étudier, à partir des fréquences alléliques des différents SNPs, la structure génétique, la démographie et la connectivité des populations (Morin *et al.* 2004; Gutenkunst *et al.* 2009; Pickrell and Pritchard 2012). Cependant, toutes ces méthodes font l'hypothèse que les différents marqueurs évoluent

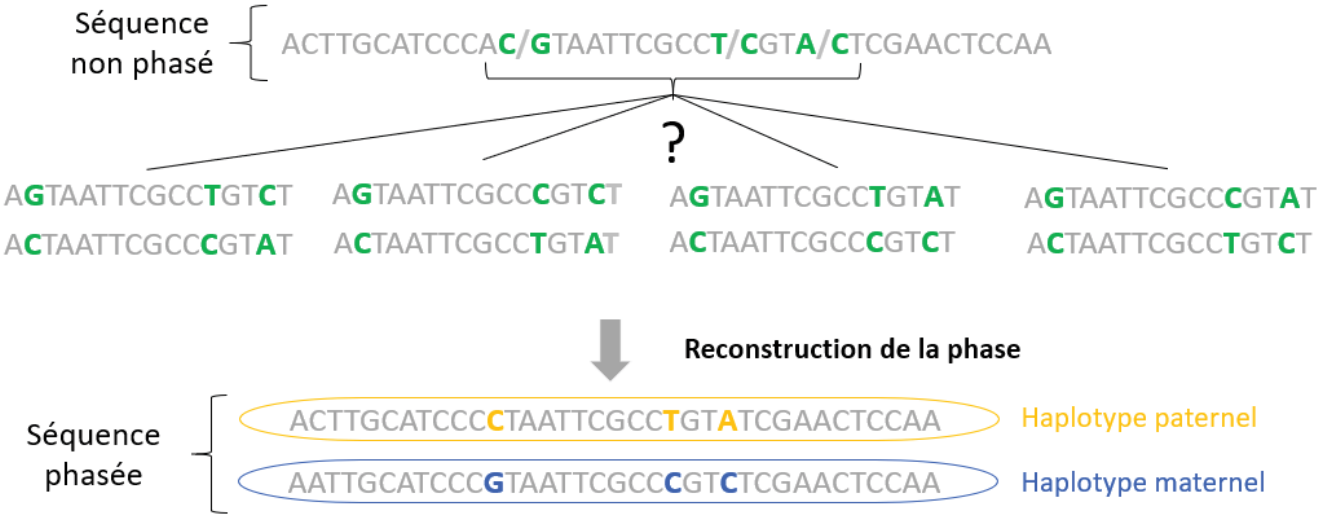


FIGURE 3 – Information gagnée par reconstruction de la phase. Une séquence est représentée avec les positions non variables en gris et 3 SNPs en vert ainsi que les différentes associations possibles entre ces SNPs. La reconstruction de la phase permet d'identifier la bonne association et de reconstruire l'haplotype paternel en bleu et maternel en jaune.

indépendamment les uns des autres. Lorsque le nombre de marqueurs utilisés devient très grand, la liaison physique le long des séquences chromosomiques devient en réalité non négligeable. Cette information sur la combinaison particulière des allèles liés, formant ce qu'on appelle des haplotypes, est aujourd'hui de plus en plus accessible pour les études de génomique des populations. Chez les espèces diploïdes, chaque individu possède une copie autosomale héritée de sa mère et une de son père, chacune composée de variants génétiques différents associés de façon particulière. Or, la plupart des techniques de séquençage nécessitent de découper la molécule d'ADN en petits fragments, ce qui a pour conséquence d'effacer l'information de liaison des variants le long des haplotypes parentaux. Les différences entre les deux copies parentales apparaissent alors comme des sites hétérozygotes chez le descendant mais on ne sait pas comment les variants sont réellement associés le long des chromosomes paternel et maternel (Figure 3). Différentes méthodes se basant sur différents types d'informations, ont été développées pour reconstruire cette information le long du génome, permettant d'obtenir des génomes dits phasés.

Les méthodes indirectes vont inférer la phase à partir de données classiques de séquençage obtenues pour plusieurs individus (Browning and Browning 2011; Rhee *et al.* 2016). Si les individus sont apparentés, alors les règles de transmission mendéliennes peuvent être utilisées, ce qui va réduire le nombre d'individus nécessaires. Par exemple, dans le cas d'un trio père-mère-enfant, si la mère et l'enfant sont hétérozygotes (C/T) et le père homozygote (T/T) alors le nucléotide C se trouve nécessairement sur l'haplotype maternel. En appliquant cette logique à l'ensemble des SNPs présents le long du génome, on peut entièrement reconstituer les haplotypes parentaux. Les seuls cas de figure où la phase ne pourra pas être reconstituée correspond aux locus pour lesquels les trois individus sont hétérozygotes. Si les relations de parenté entre individus ne sont pas connues mais que les génomes d'un grand nombre d'individus issus d'une même population sont disponibles, la phase pourra être reconstruite statistiquement, c'est-à-dire en estimant la probabilité d'observer chaque haplotype en fonction de la fréquence des différents SNPs observée dans la population. Plus récemment, des méthodes dites directes ont été développées. Elles reposent sur l'utilisation de nouvelles technologies de séquençage qui permettent de conserver l'information de liaison pendant le séquençage, par exemple en évitant de découper la molécule d'ADN et en séquençant directement de longs fragments (Snyder 2016).

Récupérer l'information de la phase permet d'obtenir une couche d'information supplémentaire par rapport à l'utilisation de SNPs individuels indépendants. En effet, avec les SNPs ce sont principalement les données de fréquences génotypiques ou alléliques qui sont utilisées, alors qu'avec les données phasées, l'information de déséquilibre de liaison s'ajoute à celle des fréquences. De plus, il existe des attendus théoriques sur la structure haplotypique d'une population naturelle évoluant neutralement

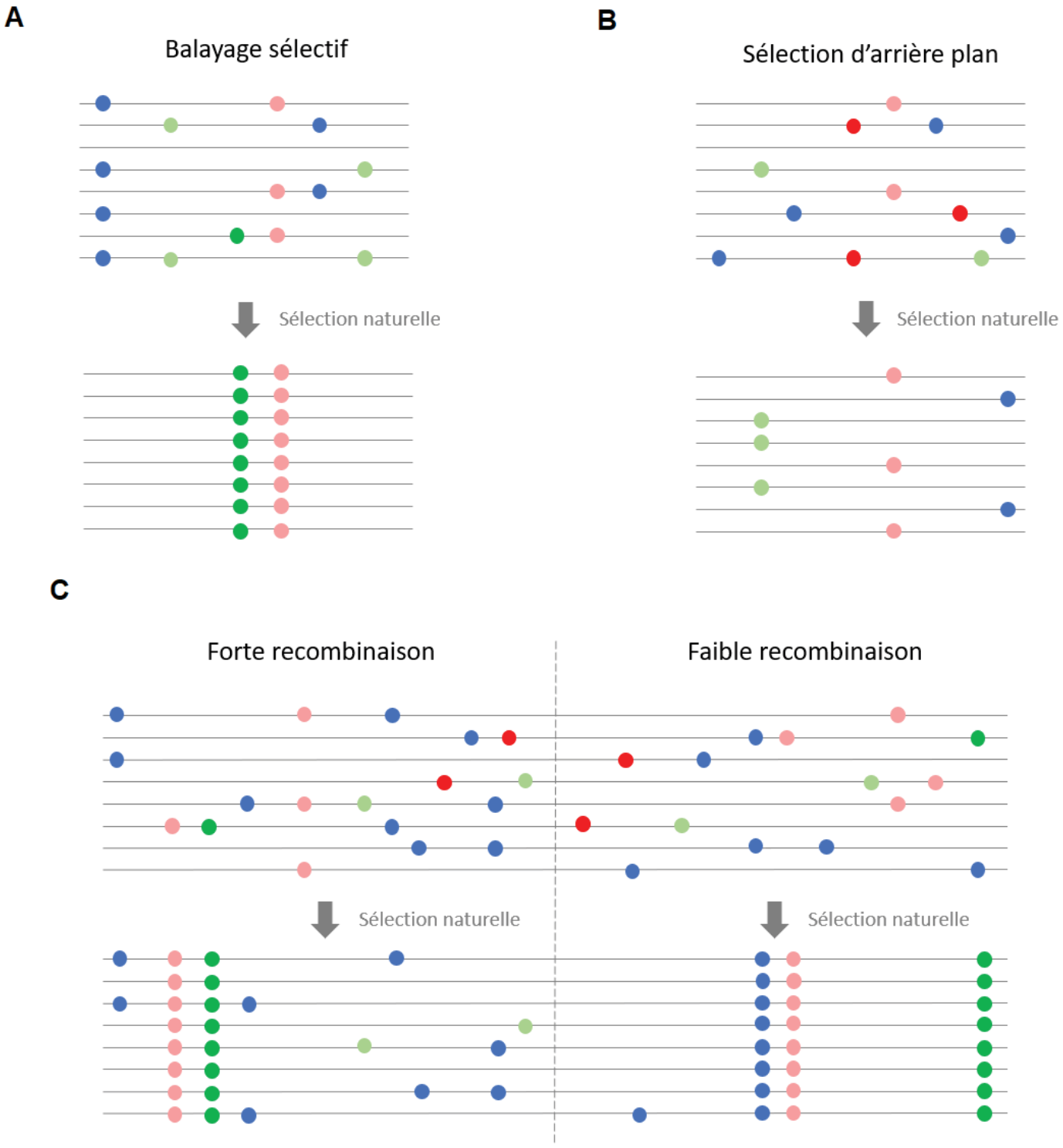


FIGURE 4 – Effet de la sélection en liaison sur la diversité nucléotidique. Impact **A.** d'un balayage sélectif, **B.** de la sélection d'arrière-plan et **C.** de l'effet répété de la sélection d'arrière-plan et de balayages sélectifs dans deux régions génomiques ayant un taux de recombinaison différent (faible ou fort). Les ronds représentent des mutations neutres (bleu), délétères (fortement en rouge et faiblement en rose) et positivement sélectionnées (fortement en vert foncé et faiblement en vert clair) disposées le long de 8 haplotypes.

(Hill and Robertson 1968). Ainsi, comparer ces attendus neutres aux données réelles peut permettre de détecter des processus évolutifs qui influencent la structure haplotypique, comme la sélection naturelle et la démographie (Nordborg and Tavaré 2002). Enfin, l'analyse de génomes entiers permet de prendre en compte l'effet de la recombinaison, processus naturel qui influence notamment la force et l'efficacité de la sélection (Stapley *et al.* 2017).

b. L'architecture génomique et le rôle de la recombinaison

La recombinaison génétique est le processus permettant l'échange de segments d'ADN entre les paires de chromosomes homologues durant la méiose. Elle permet de créer de nouvelles associations alléliques en mélangeant les variants génétiques d'origine paternelle et maternelle le long des haplotypes transmis à la génération suivante. Elle peut donc avoir des effets positifs, en créant de nouvelles associations avantageuses et ainsi favoriser l'adaptation, ou des effets négatifs en cassant des associations favorables (Stapley *et al.* 2017). La probabilité d'avoir un évènement de recombinaison entre deux locus lors d'une méiose définit le niveau de liaison génétique qui existe entre ces marqueurs. Quand le taux de recombinaison (c'est-à-dire la probabilité) est faible, la liaison génétique entre marqueurs est forte, ils seront donc très probablement transmis ensemble sur le même haplotype à la génération suivante. Pour comprendre l'évolution des génomes, il est utile de considérer la liaison génétique qui existe entre les marqueurs, étant donné qu'elle peut moduler la force non seulement de la sélection naturelle mais également de la dérive génétique.

En effet, la sélection qui s'applique sur un locus donné se répercute indirectement sur ceux qui lui sont liés, qui subissent les effets plus ou moins atténués de la sélection en liaison. Par exemple, quand un locus est positivement sélectionné, ce n'est en réalité pas uniquement sa fréquence qui va augmenter, mais celle de l'haplotype sur lequel il est situé (Maynard Smith and Haigh 1974). La fréquence de tous les locus localisés sur cet haplotype augmente donc également, qu'ils aient un effet négatif, positif ou neutre sur la valeur sélective des individus, c'est l'auto-stop génétique (*genetic hitchhiking* en anglais). Si l'haplotype arrive à fixation dans la population, on parle de balayage sélectif (Figure 4A). Il en va de même quand un locus est négativement sélectionné, la fréquence de tous les allèles qui lui sont liés va diminuer quel que soit leur effet, c'est la sélection d'arrière-plan (Charlesworth *et al.* 1993) (Figure 4B). Plus la recombinaison est faible et donc la liaison génétique forte, plus l'impact de la sélection en liaison s'étendra sur une grande région génomique. Etant donné qu'il existe des variations de taux de recombinaison le long du génome, l'intensité de la sélection en liaison varie donc également (Figure 4C).

Une des conséquences de la sélection en liaison est une diminution de la diversité génétique aux locus neutres liés aux locus sélectionnés. En effet, au moment où un haplotype atteint la fixation par

balayage sélectif, toute la diversité génétique de la région qu'il couvre disparait. De même, de façon moins intense, la sélection d'arrière-plan qui élimine les haplotypes contenant des mutations délétères et donc y compris des variants neutres en liaison, diminue la diversité génétique. Dans les régions à faible taux de recombinaison, ces effets se transmettent sur de plus grandes distances physiques et se traduisent par une diminution locale de la taille efficace. Or, une taille efficace réduite implique une sélection moins efficace et une dérive génétique plus forte. Toutes les régions génomiques n'ont donc pas les mêmes propriétés et n'évoluent pas sous l'action des mêmes forces évolutives. Les régions à faible taux de recombinaison ont en effet, plus de risque d'accumuler des mutations faiblement délétères par dérive génétique, car elles ont un coefficient de sélection trop faible pour être efficacement contre-sélectionnées dans ces régions. De plus, si les patrons de recombinaison sont stables sur le long terme évolutif, ils vont participer au modelage des patrons de diversité génétique le long du génome, ce qui peut créer des corrélations de profils génomiques de diversité entre espèces (Burri 2017). Ainsi, la recombinaison pouvant agir comme un facteur confondant, il est nécessaire de la prendre en compte afin d'identifier clairement les forces évolutives mises en jeu au cours du processus de spéciation.

II. Spéciation et hybridation

1. Les zones hybrides

a. La zone grise du processus de spéciation

Le concept biologique de l'espèce fournit une définition simple et facilement applicable aux lignées évolutives suffisamment éloignées de ce qu'est une espèce. Cependant, elle est difficilement applicable aux lignées proches qui continuent à échanger des gènes par hybridation. En effet, un contact entre deux populations partiellement isolées permet la formation d'hybrides dans une zone géographique restreinte appelée zone hybride ou zone de contact, de chaque côté de laquelle se trouvent les populations parentales (Barton and Hewitt 1985). Or, au cours des dernières années, de nombreuses études de génétique des populations ont révélé que l'hybridation est un phénomène fréquent dans la nature (Payseur and Rieseberg 2016). En effet, le processus de spéciation étant continu et les barrières d'isolement reproductif s'accumulant progressivement, la formation de vraies espèces biologiques, qui sont des entités discrètes prend du temps. Tant que l'isolement reproductif reste incomplet, les lignées divergentes continuent à s'échanger des gènes et se situent donc dans une zone intermédiaire entre des populations et des vraies espèces, appelée la zone grise du processus de spéciation (Roux *et al.* 2016). Les zones hybrides ont donc longtemps été considérées comme des laboratoires naturels pour étudier la spéciation (Endler 1977; Hewitt 1988; Harrison and Larson 2016). En effet, il peut être difficile d'identifier les forces évolutives ayant permis aux premières barrières

d'isolement reproductif de s'installer et d'initier la divergence chez de vraies espèces qui n'interagissent plus génétiquement tellement les barrières d'isolement sont nombreuses.

L'origine des zones d'hybridation pose cependant question. En effet, elles peuvent résulter d'une divergence primaire (*càd.* d'une spéciation sympatrique en cours) ou secondaire, c'est-à-dire de la remise en contact de deux populations ayant commencé leur divergence en allopatrie (*càd.* un contact secondaire) (Endler 1977). En effet, les modifications temporelles des aires de répartition des espèces peuvent conduire à un isolement allopatrique des populations d'une même espèce. Les variations climatiques du Pléistocène sont notamment connues pour avoir profondément impacté les aires de répartition des espèces des régions tempérées. Pendant les périodes glaciaires, les populations ont dû migrer dans des zones plus favorables à leur survie, appelées refuges glaciaires. Les populations isolées dans des refuges différents ont alors commencé à diverger génétiquement, puis la recolonisation post-glaciaire a permis leur remise en contact (Hewitt 1996, 2000).

Face à une zone hybride il est donc difficile d'exclure la possibilité qu'il y ait eu antérieurement une période d'allopatrie qui aurait permis d'initier la divergence. Répondre à cette question est pourtant essentiel si on veut pouvoir identifier quelles ont été les forces évolutives impliquées dans la mise en place des barrières d'isolement reproductif. En effet, la divergence primaire implique nécessairement une forme de sélection directionnelle comme par exemple de l'adaptation locale. Au contraire, lors d'un contact secondaire l'action de la dérive génétique a pu, à elle seule, suffire à faire évoluer une forme d'isolement reproductif. Néanmoins, quel que soit le mécanisme ayant permis la formation de la zone d'hybridation, à la fois la dérive génétique et la sélection naturelle peuvent avoir agi, c'est leur importance relative qui diffère.

b. Etude des zones hybrides : les clines de fréquence allélique

D'un point de vue génétique, les zones hybrides sont caractérisées par des gradients spatiaux de fréquence allélique appelés clines, au niveau desquels la fréquence des allèles diagnostiques (permettant de discriminer les populations parentales) passe rapidement de 0 à 1 (Barton and Hewitt 1985) (Figure 5A). La présence de clines indique que ces allèles ne s'échangent pas librement entre les deux populations, ils sont donc délétères dans la population receveuse ce qui crée un isolement reproductif partiel. Les clines sont maintenus par deux forces antagonistes. La migration, qui a un effet homogénéisateur et tend à les aplatir et la sélection qui agit contre les individus hybrides et les migrants et les maintient abruptes (Barton 1979) (Figure 5A). La forme du cline, notamment sa largeur et sa pente, peuvent permettre d'estimer la force de la sélection qui agit sur l'allèle étudié. Cette contre sélection peut être induite par des facteurs exogènes, c'est-à-dire en lien avec l'environnement, ou endogènes, c'est à dire liés aux interactions entre gènes. Il est important d'identifier quels

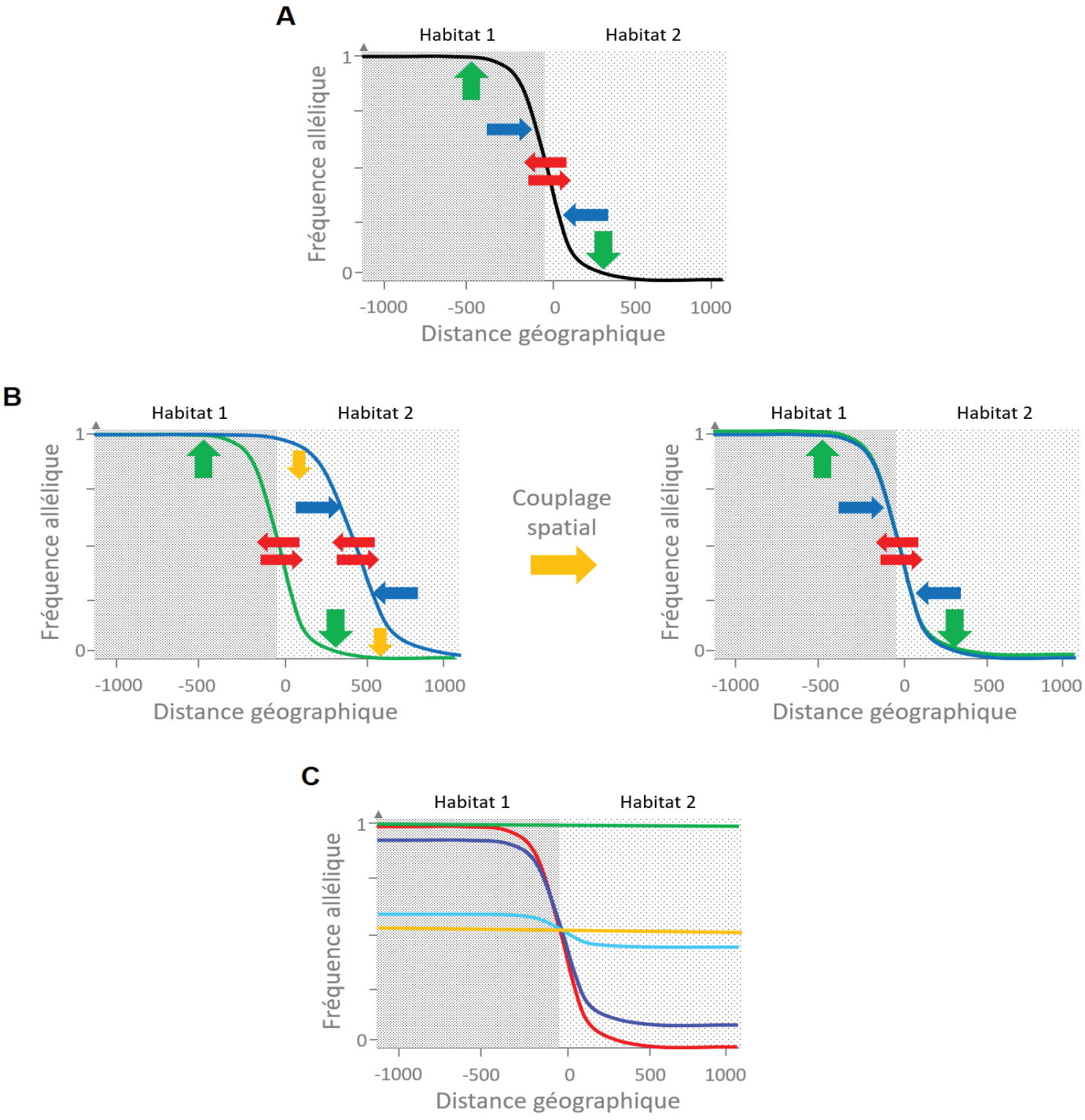


FIGURE 5 – Maintien et positionnement des clines de fréquence allélique au niveau d'une zone hybride. A. Cline de fréquence allélique maintenu par la sélection contre les migrants maladaptés à leur nouvel habitat (flèches vertes), la contre sélection contre les hybrides (flèches bleues) et la migration (flèches rouges). **B.** Un cline endogène (en bleu) chevauchant partiellement un cline exogène (en vert) se déplace sous l'effet du couplage (flèches jaunes) et vient se superposer au cline d'adaptation locale. **C.** Clines de fréquence allélique pour différents allèles : impliqué dans l'isolement reproductif (rouge), avantageux dans les deux populations (vert), neutre (jaune), neutre mais faiblement (bleu clair) ou fortement (bleu foncé) lié à un locus barrière.

mécanismes maintiennent les clines afin de déterminer quelles forces évolutives ont permis la mise en place de l'isolement reproductif.

Cependant, déterminer si une zone hybride est maintenue par des facteurs endogènes et exogènes peut s'avérer compliqué étant donné que la nature de la sélection affecte peu la forme des clines et que les différents types de clines interagissent entre eux. En effet, les locus impliqués dans l'isolement reproductif peuvent entrer en interaction au travers d'un mécanisme de couplage spatial des clines (Bierne *et al.* 2011a) (Figure 5B). La position des clines endogènes n'étant pas contrainte spatialement, ils peuvent se faire attirer par les clines d'adaptation locale qui eux sont fixés géographiquement par les facteurs environnementaux qui les génèrent. De plus, les barrières environnementales et les zones à faible densité de population où la migration est réduite ont tendance à bloquer les clines (Barton 1979). Un exemple de ce type de zone hybride existe dans le Nord-Est des Etats-Unis entre deux espèces de grillon : *Gryllus firmus* et *Gryllus pennsylvanicus*. Une partie des barrières d'isolement reproductif qui existent entre ces deux espèces semblent être liées à l'existence de clines d'adaptation locale, les deux espèces étant adaptées à des types de sols différents (Larson *et al.* 2014). Cependant, des épisodes d'allopatrie anciens auraient également pu permettre l'évolution d'incompatibilités endogènes. Il existe également un isolement reproductif temporel des deux grillons, *G. firmus* atteignant la maturité plus tardivement que *G. pennsylvanicus* (Harrison 1985). Les zones hybrides associées à des limites d'habitats peuvent donc être issues d'une histoire complexe et être maintenues par des mécanismes multiples.

L'isolement reproductif n'étant que partiel au niveau des zones hybrides, tous les allèles n'ont pas le même comportement. En effet, seuls les allèles impliqués dans l'isolement reproductif présentent des clines de fréquence allélique abrupts et bloquent les échanges génétiques (Figure 5C courbe rouge). Les allèles avantageux peuvent facilement traverser la barrière et ne présentent donc pas de clines (Figure 5C courbe verte). Les allèles complètement neutres peuvent, quant à eux, s'échanger librement entre les deux lignées et leur fréquence tend donc à s'équilibrer autour de la même valeur dans les deux populations (Figure 5C courbe jaune). Toutes les régions génomiques ne sont donc pas affectées de la même façon par le flux génique, certaines étant plus perméables que d'autres, c'est pourquoi on parle de barrière semi-perméable (Harrison 1993). On peut donc retrouver loin de la zone hybride des individus issus d'un grand nombre de générations d'hybridation chez qui la sélection négative à agit pendant de nombreuses générations pour éliminer les allèles délétères. Si la sélection a été suffisamment efficace, ces individus ne possèdent que les allèles issus de leur lignée aux locus d'isolement reproductif mais les allèles des deux populations aux locus neutres, ils sont alors qualifiés d'individus introgressés. Il en va de même pour les populations desquelles sont issus ces individus qui ne sont plus vues comme des populations hybrides mais introgressées. En effet, contrairement aux

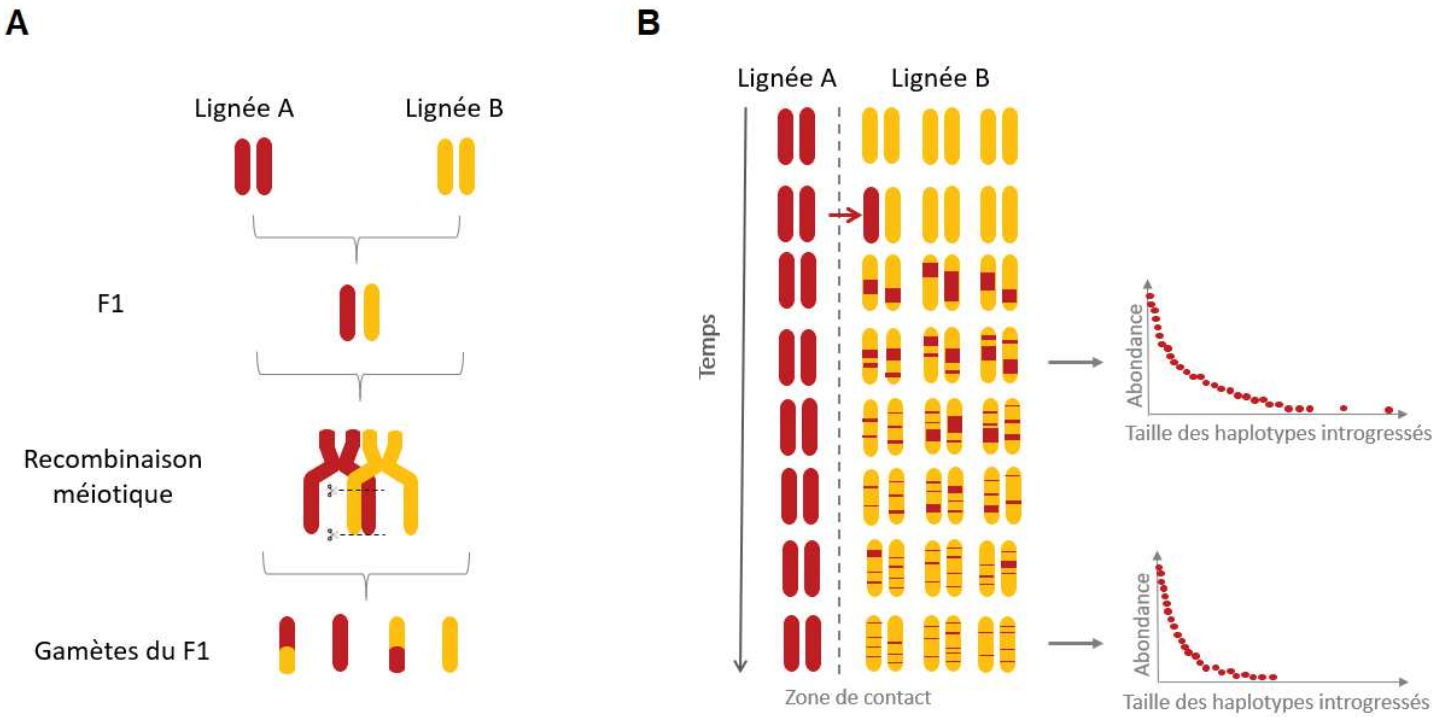


FIGURE 6 – Hybridation et recombinaison. Représentation schématique de l'influence de la recombinaison sur le mélange de deux fonds génétiques issus de la lignée A (rouge) et B (jaune). Deux chromosomes sont représentés par individu. **A.** Le croisement de deux individus issus des populations parentales permet la formation d'hybrides de première génération (F1). La recombinaison qui a lieu au moment de la méiose va permettre de casser les associations entre allèles issus du même fond génétique et donc de raccourcir les haplotypes. **B.** Rencontre au niveau d'une zone de contact à un instant donné entre la lignée A et B pour qui respectivement un et trois individus sont représentés à chaque pas de temps, permettant l'échange de matériel génétique de la lignée A vers B (flèche rouge). Une fois entrés dans le fond génétique de la lignée B les haplotypes originaires de la lignée A sont progressivement raccourcis par la recombinaison à chaque génération. La forme de la distribution de taille des haplotypes originaires de A introgressés dans B change donc avec le temps.

populations hybrides chez qui tous les allèles des deux lignées sont présents en fréquences intermédiaires, les populations introgressées sont supposées fixées aux locus d'isolement et polymorphes aux locus neutres (Bierne *et al.* 2003).

Cependant, les allèles n'étant pas indépendant mais liés génétiquement les uns aux autres, le comportement des allèles neutres est en réalité principalement déterminé par leur niveau de liaison aux locus sous sélection. Plus ils sont en fort déséquilibre de liaison avec des locus barrières, plus leurs clines seront abrupts et moins ils circuleront librement entre les deux populations (Barton and Bengtsson 1986) (Figure 5C courbes bleues). De plus, il peut également y avoir de la liaison génétique entre gènes barrières, ce qui leurs permet de cumuler plus facilement leur effets et donc de réduire plus efficacement le flux génique (Yeaman *et al.* 2016). Enfin, le niveau de couplage existant, y compris entre des locus d'isolement situés sur des groupes de liaison différents, module également la force de la barrière (Barton 1983; Kruuk *et al.* 1999). Le niveau de liaison génétique entre allèles a un impact sur l'intensité des échanges génétiques, il est donc nécessaire de prendre en compte la liaison génétique afin d'identifier clairement les régions génomiques impliquées dans l'isolement reproductif.

a. Etude des zones hybrides : les haplotypes introgressés

L'échange de matériel génétique entre populations au travers d'une zone de contact se fait *via* l'intermédiaire d'hybrides de première génération qui vont eux-mêmes se reproduire avec des individus issus des lignées parentales. Ainsi, quand les allèles entrent dans un nouveau fond génétique ils n'arrivent pas indépendamment, mais fortement liés les uns aux autres sous la forme de blocs haplotypiques, également appelés haplotypes introgressés ou haplotypes migrants. Chez les hybrides de première génération, les haplotypes des deux lignées s'étendent tout le long des chromosomes. La recombinaison qui a lieu à chaque génération va ensuite casser les associations alléliques et progressivement raccourcir les haplotypes introgressés (Pool and Nielsen 2009; Liang and Nielsen 2014) (Figure 6A). Ainsi, plus le nombre de générations depuis le contact est grand et plus la taille des haplotypes introgressés présents dans la population est réduite (Figure 6B). Le génome d'individus issus de populations introgressées peut alors être vu comme une mosaïque de fragments chromosomiques originaires de différentes lignées (Tang *et al.* 2006).

Les haplotypes introgressés ne sont cependant pas directement observables dans les génomes, c'est pourquoi de nombreuses méthodes ont été développées pour permettre de les identifier (Yuan *et al.* 2017; Geza *et al.* 2018). En effet, une fois révélés, ils donnent accès à une couche d'information supplémentaire pour estimer à la fois la date et l'intensité des échanges génétiques qui se sont produits entre deux lignées (Sousa and Hey 2013). Etant donné que les haplotypes migrants sont

progressivement raccourcis à chaque génération, leur longueur est inversement proportionnelle au temps qui s'est écoulé depuis le contact.

On s'attend donc à ce que les longs haplotypes introgressés soient rentrés dans la population plus récemment que les courts. Une façon de les étudier est de regarder leur distribution de taille au sein de la population. En effet, la forme de cette distribution est notamment influencée par le temps écoulé depuis le début du contact (Figure 6B). Il existe une formule analytique permettant de relier la taille moyenne des haplotypes introgressés ($\bar{L}$), le nombre de générations écoulées depuis le contact (t), le taux local de recombinaison (r) et l'intensité des échanges génétiques (m) : $\bar{L} = [(1-m)r(t-1)]^{-1}$ (Racimo *et al.* 2015). Ainsi, la recombinaison peut ici servir d'horloge pour dater le début des échanges génétiques entre les deux lignées. Des travaux ont également été développés non pas sur la taille des haplotypes introgressés mais sur le nombre de jonctions entre blocs chromosomiques d'origine différentes (Janzen *et al.* 2018; Hvala *et al.* 2018). L'idée sous-jacente est également d'utiliser la recombinaison comme une horloge, le nombre de jonction étant proportionnel au nombre d'évènements de recombinaisons qui se sont produit depuis le contact et donc au nombre de générations écoulées depuis l'hybridation initiale.

Avoir accès aux haplotypes introgressés peut donc permettre d'avoir une image plus précise de l'histoire des échanges génétiques entre lignées divergentes et donc sur le moment de la mise en place des zones hybrides. De plus, ils permettent d'avoir accès à une mesure directe de l'intensité du flux génique ce qui peut permettre de révéler les effets que ces haplotypes ont sur la valeur sélective des individus.

2. Effets sélectifs de l'introgression et recombinaison

a. Effets positifs

Quand deux lignées s'hybrident, il peut arriver que l'une d'entre elles possède des allèles avantageux dans le fond génétique et/ou l'environnement de l'autre lignée. Dans ce cas-là, ces allèles vont facilement traverser la zone hybride et être positivement sélectionnés dans le nouveau fond génétique jusqu'à atteindre la fixation, c'est l'introgression adaptative (Racimo *et al.* 2017). De nombreuses études ont montré que l'acquisition de phénotypes adaptatifs pouvait se faire par hybridation. En effet, l'introgression adaptative, en permettant le transfert de variants alléliques en liaison qui ont déjà été testés par la sélection naturelle dans leur environnement d'origine, facilite l'adaptation locale (Martin and Jiggins 2017). Un des exemples les plus connus concerne notre propre espèce. Quand la forme anatomique moderne de l'homme (*Homo sapiens*) est sortie d'Afrique, les populations

eurasiatiques ont rencontré de nouvelles conditions environnementales qui ont généré de nouvelles pressions de sélection, comme par exemple la haute altitude où la captation d'oxygène est plus difficile. Il a été montré que les Tibétains possèdent une forme particulière du gène *EPAS1* présentant une signature de sélection positive. Il a été montré que cet allèle particulier, qui en modulant la concentration en hémoglobine permet l'adaptation à l'hypoxie, a été obtenu par hybridation avec une lignée humaine aujourd'hui éteinte, l'homme de Denisova (Huerta-Sánchez *et al.* 2014). En effet, en sortant d'Afrique les populations d'*Homo sapiens* ont rencontré d'autres lignées humaines aujourd'hui éteintes avec lesquelles elles se sont hybridées. L'introgression adaptative ne se limite pas aux humains et a également été décrite chez d'autre vertébrés, comme le lièvre d'Amérique (*Lepus americanus*) chez qui elle a permis l'acquisition d'un camouflage saisonnier (Jones *et al.* 2018), mais également des invertébrés, comme les papillons du genre *Heliconius* chez qui elle a permis le transfert de gènes de mimétisme entre espèces (Dasmahapatra *et al.* 2012).

Il est également possible que les individus hybrides aient une meilleure valeur sélective que leurs parents, on parle alors de vigueur hybride ou d'hétérosis. Il y a hétérosis quand les individus hétérozygotes ont une meilleure valeur sélective que les homozygotes. Deux mécanismes génétiques peuvent en être à l'origine. Le premier est la superdominance, qui n'implique qu'un seul locus. Dans ce cas, c'est l'interaction au sein des génotypes hybrides des deux allèles issus des deux fonds génétiques divergents qui permet d'obtenir un phénotype avec une meilleure valeur sélective que ceux des parents. Ce mécanisme a notamment été décrit chez la tomate (*Solanum lycopersicum*) chez qui les individus hétérozygotes à certains locus, présentent une meilleure croissance et produisent plus de fruits (Lippman and Zamir 2007). Le deuxième mécanisme est la superdominance associative (Ohta and Kimura 1970). Dans ce cas-là, c'est le masquage des mutations récessives faiblement délétères qui ségrégent à faible fréquence dans le fond génétique de la population qui permet d'augmenter la valeur sélective des individus hétérozygotes. Plus le nombre de mutations faiblement délétères masquées est grand plus la valeur sélective des individus hybrides sera améliorée.

Dans le cas d'une population introgressée, l'hétérosis peut donc être générée par le masquage des mutations faiblement délétères récessives privées à la population receveuse, par les haplotypes introgressés (Figure 7A). L'effet de superdominance associative locale étant proportionnel au nombre de mutations masquées, il sera d'autant plus fort que les haplotypes introgressés sont longs. L'effet est donc maximal chez les F1 et diminue progressivement au fur et à mesure des générations, avec le raccourcissement des haplotypes introgressés par la recombinaison. L'effet sera également plus fort pour les populations chez qui ségrégent un grand nombre de mutations faiblement délétères. Dans ce cas-là, l'hétérosis peut même favoriser l'entrée d'haplotypes introgressés, quel que soit le niveau de fardeau génétique (ensemble des mutations délétères ségrégant dans une population) de la

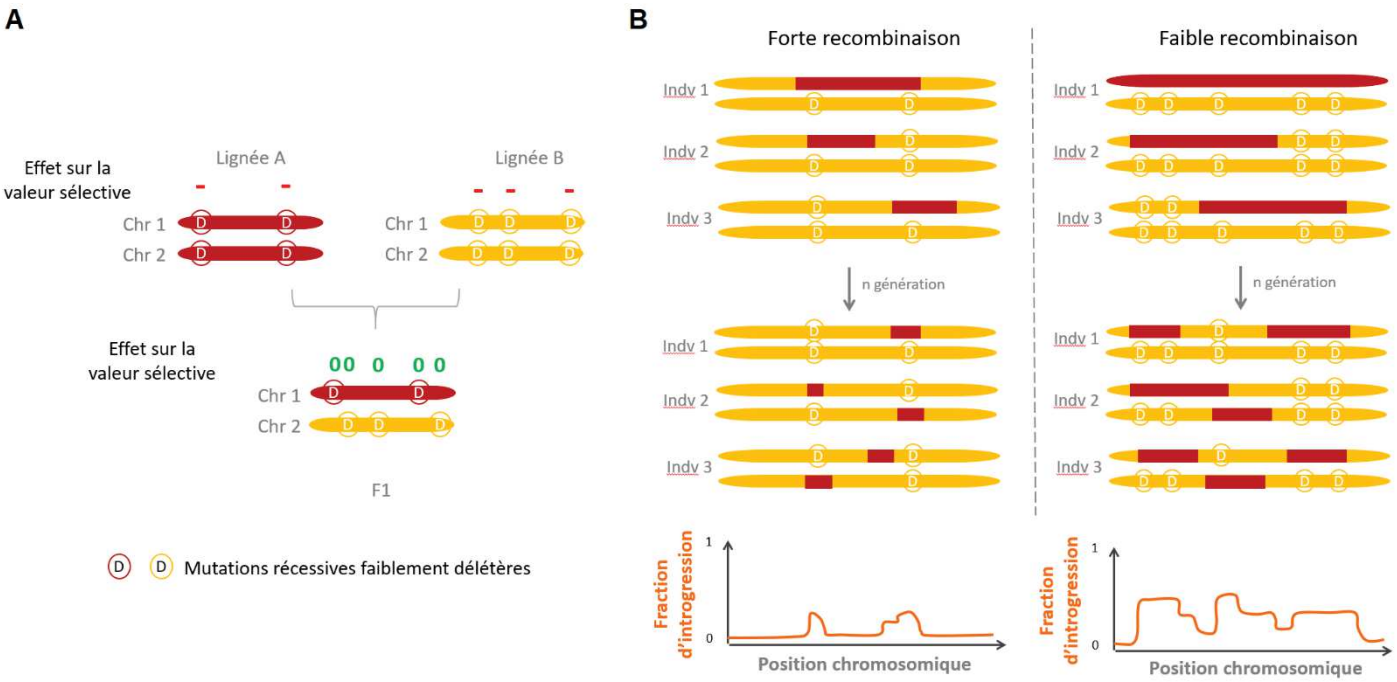


FIGURE 7 – Hétérosis et recombinaison. Les couleurs représentent les deux fonds génétiques desquels peuvent être issues les fragments chromosomiques, celui de la lignée A (rouge) et B (jaune), deux copies chromosomiques sont représentées par individu. **A.** Des mutations récessives faiblement délétères différentes ségrégent dans le fond génétique de la lignée A et B. Elles expriment leur effet délétère (moins rouges) chez les individus à l'état homozygote. Un hybride de première génération (F1) étant complétement hétérozygote, toutes les mutations faiblement délétères privées aux lignées A et B sont masquées (zéros verts), il a donc une meilleure valeur sélective que ses parents. **B.** L'effet d'hétérosis local généré par les fragments chromosomiques introgressés est plus marqué dans les régions à faible taux de recombinaison car ces fragments ont tendance à être plus longs et les mutations faiblement délétères y sont également plus nombreuses. Ainsi, l'hétérosis peut transitoirement favoriser l'introgression et augmenter localement la fraction d'introgression des régions à faible taux de recombinaison par rapport aux régions à fort taux.

population donneuse (Kim *et al.* 2018). L'hybridation entre *Homo sapiens* et l'homme de Neandertal a probablement généré de la superdominance associative favorisant l'introgression d'haplotypes néandertaliens dans les génomes eurasiatiques. En effet, la population néandertalienne aurait accumulé un grand nombre de mutations faiblement délétères suite à un goulot d'étranglement, c'est-à-dire une réduction brutale de sa taille efficace qui a fortement diminué sa diversité génétique, augmentant la dérive génétique et diminuant l'efficacité de la sélection (Harris and Nielsen 2013).

Les variations du taux local de recombinaison en modulant la force de la superdominance associative peuvent également influencer les patrons locaux d'introgression. En effet, la superdominance associative étant générée par le déséquilibre de liaison, elle tend à être plus forte dans les régions où la recombinaison est réduite et ce pour deux raisons. Premièrement, les haplotypes introgressés tendent à y être plus long, ce qui augmente leur effet de masquage. Deuxièmement, la taille efficace y étant réduite à cause de la sélection en liaison, c'est également là que ségrégent le plus de mutations faiblement délétères. Ainsi, l'hétérosis peut localement augmenter la fraction d'introgression dans les régions à faible taux de recombinaison par rapport aux régions à fort taux (Figure 7B).

b. Effets négatifs

Les allèles introgressés peuvent avoir un effet délétère sur la valeur sélective des hybrides, on parle de dépression d'hybridation (Figure 8A). Si cet effet délétère est suffisamment fort alors il peut conduire à un isolement reproductif partiel entre les lignées qui s'hybrident. Que la sélection soit générée par des facteurs endogènes ou exogènes les allèles introgressés délétères seront contre-sélectionnés et éliminés sur le long terme. Par exemple, on retrouve dans les génomes des Européens des régions complètement dépourvues d'allèles introgressés provenant de Neandertal et de Denisova, fait potentiellement dû à la sélection contre l'introgression qui était probablement délétère (Sankararaman *et al.* 2014, 2016). Cependant, la force de la sélection étant modulée par la recombinaison, on ne s'attend pas à retrouver les mêmes patrons d'introgression dans les régions à faible et fort taux de recombinaison. En effet, la liaison génétique étant forte entre les allèles introgressés délétères dans les régions à faible taux de recombinaison, les haplotypes introgressés y seront retirés plus efficacement, qu'ils soient impliqués dans l'isolement reproductif ou non. De plus, étant donné que les haplotypes sont plus longs, de nombreux allèles neutres seront également retirés en même temps. Dans les régions à forte recombinaison au contraire, les allèles neutres pourront plus facilement recombiner et se séparer des allèles délétères, leur permettant de se maintenir dans la population. Ainsi, si l'introgression est délétère, on s'attend à retrouver plus d'introgression dans les régions à fort taux de recombinaison (Figure 8B). C'est effectivement ce qui a été observé chez l'homme où les niveaux d'introgression d'allèles hérités de l'homme de Néandertal et Denisova sont plus élevés dans les régions à fort taux de recombinaison (Schumer *et al.* 2018). Cette corrélation

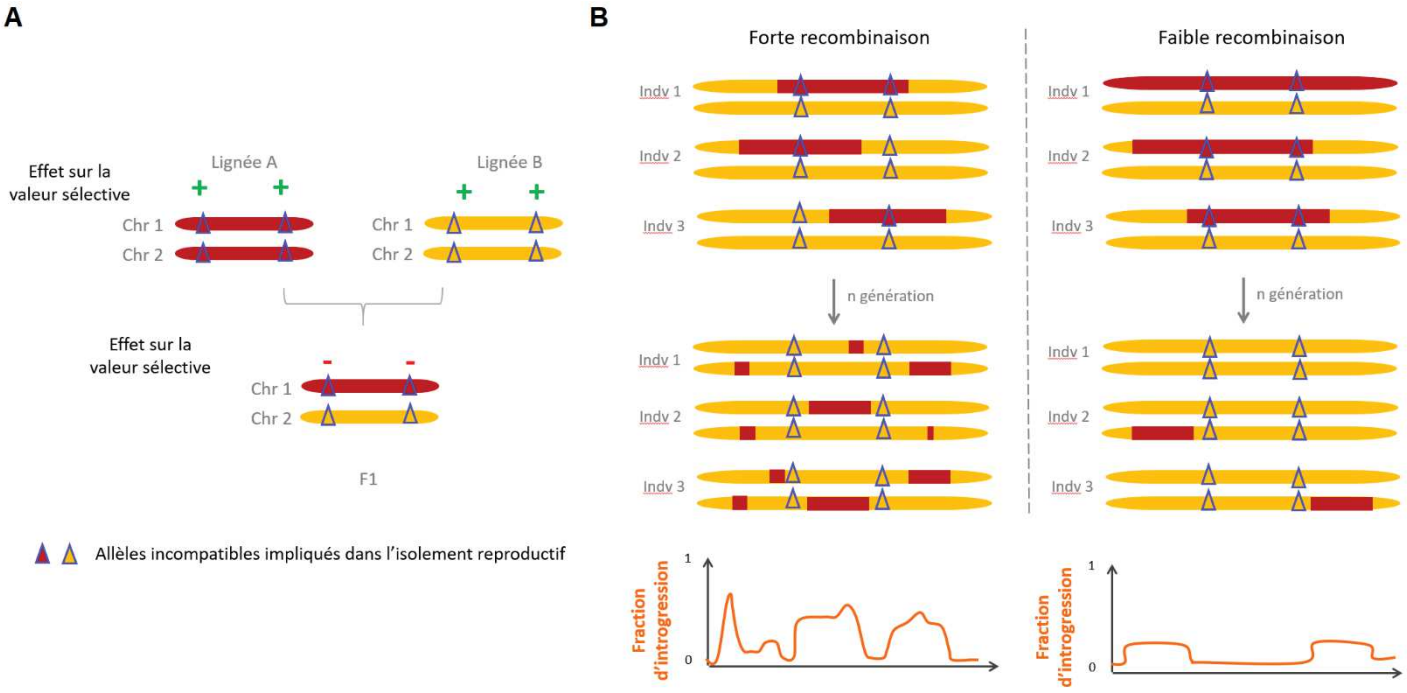


FIGURE 8 – Locus incompatibles entre lignées et recombinaison. Les couleurs représentent les deux fonds génétiques desquels peuvent être issus les fragments chromosomiques, celui de la lignée A (rouge) et B (jaune), deux copies chromosomiques sont représentées par individu. **A.** Des mutations avantageuses dans leur fond génétique d'origine (plus verts) se révèlent incompatibles quand elles sont associées. Un hybride de première génération (F1) étant complétement hétérozygote, a une valeur sélective réduite, c'est la dépression d'hybridation **B.** Dans les régions à fort taux de recombinaison les allèles introgressés neutres parviennent à se dissocier des allèles délétères et donc à se maintenir dans la population. A l'inverse, dans les régions à faible taux de recombinaison ils sont retirés en même temps que les délétères. La fraction d'introgression tend donc à être plus élevée dans les régions à fort taux de recombinaison.

positive entre recombinaison et introgression a également été observée chez les poissons du genre *Cichlida* (Gante *et al.* 2016), les poissons porte-épée du *Xiphophorus* (Schumer *et al.* 2018) et démontrée à l'aide de simulations (Martin and Jiggins 2017). Pour conclure, l'introgression pouvant avoir à la fois des effets positifs et négatifs, c'est la sélection qui va modeler les patrons génomiques d'introgression. Cependant, la force de la sélection étant modulée par la recombinaison, c'est en réalité de l'interaction entre la sélection et le taux local de recombinaison que résultent les patrons chromosomiques d'introgression. Cependant, la sélection positive tend à générer une corrélation négative entre introgression et taux de recombinaison alors que la sélection négative quant à elle, tend à générer une corrélation positive (Figure 7B et 8).

3. Hybridation et spéciation des processus opposés ?

a. Le renforcement

Alors que la spéciation permet l'accumulation de différences entre les populations qui divergent, l'hybridation tend à homogénéiser les fonds génétiques des lignées qui s'hybrident, c'est pourquoi les deux mécanismes peuvent paraître opposés. L'hypothèse de spéciation par renforcement (Blair 1955) propose au contraire l'hybridation comme un mécanisme permettant de renforcer l'isolement reproductif et donc de faciliter le maintien des différences génétiques. Le renforcement se produit au niveau des zones hybrides, quand les lignées ont commencé à accumuler des barrières d'isolement post-zygotique mais qu'il n'existe pas encore d'isolement pré-zygotique. Au niveau de la zone de contact, les individus issus des deux populations vont se reproduire aléatoirement entre eux, générant des hybrides. Or, étant donné qu'il existe des barrières d'isolement post-zygotique, les individus hybrides vont subir une dépression d'hybridation et auront donc une valeur sélective amoindrie. La production d'hybrides représente donc un coût pour les parents qui vont investir des ressources pour des descendants ayant une valeur sélective réduite. La sélection naturelle peut donc agir pour sélectionner tout trait permettant d'éviter la formation d'individus hybrides.

Ainsi, les individus se reproduisant aléatoirement seront contre-sélectionnés et au contraire ceux préférant se reproduire avec des individus issus de la même lignée (*càd.* ayant une préférence homogame) seront avantagés car ils produisent des descendants viables et fertiles. Une préférence des femelles pour les mâles de leur propre lignée permet, par exemple, de générer de l'isolement pré-zygotique. Il y aura donc au sein de chaque population de plus en plus d'individus homogames ce qui va générer un déséquilibre de liaison entre les gènes responsables des incompatibilités hybrides et de l'homogamie, renforçant l'isolement reproductif. On parle ici de renforcement car il faut que des barrières post-zygotiques existent pour que le renforcement puisse opérer. Dans les autres modèles,

l'isolement pré-zygotique évolue comme une conséquence de l'accumulation de divergences génétiques entre deux lignées sous l'action de la sélection naturelle ou de la dérive génétique. Sous l'hypothèse du renforcement, au contraire, l'isolement pré-zygotique est une adaptation en soi car c'est ce caractère qui est directement sous sélection.

L'idée que le renforcement puisse conduire à une spéciation complète pose cependant certains problèmes théoriques. En effet, le flux génétique qui a lieu au niveau de la zone hybride va tendre à casser les associations entre les gènes codant pour l'homogamie et les incompatibilités hybrides ce qui peut empêcher le renforcement. De plus, l'isolement pré-zygotique n'évoluant que dans la zone où les hybrides sont formés, cela limite l'efficacité de la sélection agissant sur un nombre restreint d'individus. Enfin, plus l'isolement pré-zygotique est fort moins il y a d'hybrides formés et donc plus les pressions de sélection contre l'hybridation diminuent. Ainsi, le renforcement ne peut émerger que dans des conditions particulières (Lemmon and Kirkpatrick 2006). Il existe cependant des exemples empiriques en faveur du renforcement. La comparaison des patrons de spéciation entre plusieurs espèces de drosophiles a par exemple montré qu'entre espèces d'âge comparable, l'isolement pré-zygotique est plus fort pour les espèces sympatriques qu'allopatriques, ce qui est attendu sous l'hypothèse du renforcement (Coyne and Orr 2004). Le renforcement semble également avoir participé à la mise en place de l'isolement pré-zygotique entre deux sous-espèces de souris du complexe *Mus musculus* présentent sur le continent européen. Il existe une zone hybride qui s'étend de la Norvège à la Mer Noire entre *M. m. domesticus* à l'ouest et *M. m. musculus* à l'est, qui est issue d'un contact secondaire. Des études ont montré que pour les deux sous-espèces, à la fois les mâles et les femelles issus des populations situées au niveau de la zone hybride du Danemark, se reproduisent préférentiellement avec des individus de la même lignée (Smadja and Ganem 2002, 2005; Smadja *et al.* 2004). Au contraire, les populations situées loin de la zone hybrides n'expriment pas de préférence homogame (Smadja and Ganem 2005; Smadja *et al.* 2015).

b. La spéciation hybride

Le modèle de spéciation hybride propose que de nouvelles espèces puissent se former grâce à l'hybridation de deux lignées en cours de divergence. Pour cela, il faut que les individus hybrides évoluent une forme d'isolement reproductif avec les lignées parentales dont ils sont issus. Ce mode de spéciation semble être facilement réalisable chez les plantes grâce à la polyploïdisation, c'est-à-dire l'augmentation du nombre de copies chromosomiques (Mallet 2007). Une duplication du génome des hybrides les rendant tétraploïdes crée immédiatement un isolement reproductif complet avec les lignées parentales. En effet, quand les individus hybrides polyploïdes se croisent avec des individus diploïdes issus de l'une des lignées parentales, des descendants avec un nombre anormal de copies chromosomiques sont produits. Mêmes si ces descendants sont viables, ils sont généralement stériles

car ils produisent des gamètes avec un nombre déséquilibré de chromosomes. La plupart des plantes pouvant se reproduire de façon sexuée comme asexuée et étant capable de s'auto-féconder, l'impossibilité de se reproduire avec les lignées parentales n'empêche pas la reproduction des hybrides. C'est pourquoi ce mode de spéciation est envisageable chez les plantes (Wood *et al.* 2009) mais difficilement réalisable chez les animaux, où la spéciation hybride est envisagée comme étant principalement homoploïde, c'est-à-dire sans changement dans le nombre de copies chromosomiques.

La spéciation hybride peut donc être définie comme un processus de spéciation au cours duquel l'hybridation a joué un rôle crucial dans la mise en place de l'isolement reproductif (Gross and Rieseberg 2005; Mallet 2007; Schumer *et al.* 2014). Plusieurs mécanismes différents peuvent être envisagés pour expliquer comment l'hybridation peut permette la spéciation. Le plus souvent, il est proposé que l'hybridation introduise des gènes qui, grâce aux nouvelles combinaison alléliques générées par la recombinaison, génèrent de nouveaux phénotypes permettant aux hybrides de coloniser de nouvelles niches écologiques (Gross and Rieseberg 2005). En effet, les hybrides présentent généralement une diversité phénotypique plus large que celle de leurs parents, c'est ce que l'on appelle la transgression phénotypique (Rieseberg *et al.* 1999). Il a également été proposé que l'hybridation pourrait permettre de réassembler d'anciens allèles ayant émerger il y a longtemps et ainsi créer de nouvelles associations facilitant la spéciation (Marques *et al.* 2019). Récemment, il a été avancé que la spéciation hybride homoploïde serait en réalité plus fréquente que ce qui était précédemment envisagé (Mallet 2005; Mavárez and Linares 2008; Seehausen 2013). Cependant, comme l'ont argumenté Schumer *et al.* dans leur article de 2014, la plupart des études ne valident pas certains critères essentiels pour prouver l'existence d'une spéciation hybride. En effet, il y a spéciation hybride quand (*i*) le fond génétique de la nouvelle espèce porte des traces nettes d'hybridation, (*ii*) il existe un isolement reproductif clair entre la population hybride et les lignées parentales et (*iii*) l'isolement reproductif est une conséquence direct de l'hybridation (Schumer *et al.* 2014).

Ainsi, bien que la plupart des études démontrent clairement l'existence de populations hybrides ou admixées, elles n'établissent pas de lien direct entre l'hybridation et la mise en place de l'isolement reproductif (Mavárez and Linares 2008; Hermansen *et al.* 2011). Un exemple convainquant est toutefois celui des papillons du genre *Heliconius*, chez qui les gènes ayant permis la spéciation dérivent de l'hybridation. En effet, ces gènes qui influencent les patrons de colorations des ailes, ont permis le mimétisme et sont également impliqués dans l'isolement reproductif, la coloration des ailes permettant la reconnaissance entre partenaires sexuels (Salazar *et al.* 2010). Dans tous les cas, bien que l'hybridation permette souvent de générer de nouveaux traits, il est difficile d'envisager que ces traits soient systématiquement plus adaptés (écologiquement ou intrinsèquement) que les

parentaux.Ainsi, la spéciation hybride par sélection positive des génotypes hybrides est probablement un mécanisme assez rare.

Un autre mécanisme par lequel l'hybridation peut générer une forme d'isolement reproductif est au travers des conflits génétiques qu'elle génère. En effet, il a été montré que l'accumulation d'incompatibilités de type Dobzhansky-Muller (DMI) est fréquente quand deux lignées divergent (Fishman and Willis 2001; Presgraves 2010). Si l'effet des DMI n'est que faiblement délétère, elles peuvent ségréger dans la population hybride et diminuer la valeur sélective des individus hybrides, la sélection va donc agir afin d'éliminer les conflits induits par ces incompatibilités. Un conflit génétique induit par une DMI à deux locus peut être résolu en fixant l'un des deux allèles parentaux à chaque locus. Si la population hybride a été générée par un mélange équivalent des deux populations parentales, il y a 50% de chance de fixer l'un ou l'autre des deux allèles (Schumer *et al.* 2015). Ainsi, la résolution de DMI multiples indépendamment vers l'un ou l'autre des allèles parentaux va déplacer les incompatibilités entre la population hybride et les deux lignées parentales générant ainsi un isolement reproductif avec ces deux dernières. L'accumulation de DMI lors de la divergence étant inévitable, ce mécanisme permet d'expliquer de façon relativement simple comment l'hybridation peut générer de l'isolement reproductif.

La différence principale entre ces deux mécanismes permettant la spéciation hybride est la nature des forces évolutives qui génèrent les barrières d'isolement reproductif. Le premier a besoin qu'une forme de sélection positive agisse sur les hybrides, alors que dans le deuxième, les barrières sont des DMI qui peuvent évoluer simplement sous l'action de la dérive génétique. Or, les patrons générés par ces deux mécanismes peuvent être très similaires (Schumer *et al.* 2015). Ainsi, même si la spéciation hybride est établie, la nature des forces évolutives ayant permis la formation des espèces peut être compliquée à déterminer.

III. La génomique de la spéciation : des patrons génomiques aux processus évolutifs

1. Comprendre le contexte démographique de la divergence

a. Les modèles démographiques

Identifier les locus impliqués dans l'isolement reproductif et comprendre quelles sont les forces évolutives ayant permis leur mise en place est l'un des objectifs principaux en génomique de la spéciation. Cependant, l'histoire démographique des espèces peut laisser dans les génomes des traces qui peuvent être confondues avec celles de la sélection. En effet, les populations peuvent subir des alternances de périodes avec et sans flux génique, ce qui va laisser une empreinte sur le polymorphisme des génomes (Abbott *et al.* 2013). Un contact secondaire peut, par exemple, générer des clines d'introgression qui sont corrélés avec des variables environnementales, ce qui peut être interprété comme le signe d'une adaptation locale à un gradient environnemental. Déterminer si les patrons observés sont issus d'un contact secondaire ou d'une divergence primaire est donc une première étape essentielle afin d'identifier les processus évolutifs mis en jeu lors de la divergence.

Les données de polymorphisme recueillies chez un grand nombre d'individus issus de deux lignées évolutives distinctes, contiennent des informations à la fois sur les aspects temporels et démographiques de leur divergence. Ainsi, en comparant ces données réelles à des modèles constituant une représentation simplifiée du processus de divergence, il est possible d'évaluer statistiquement la vraisemblance des différents scénarios de spéciation sous lesquels les patrons observés ont pu évoluer. Les premiers modèles développés avaient pour objectif de distinguer des scénarios simples de spéciation allopatrique et sympatrique, respectivement grâce au modèle de Strict Isolement sans flux génique (SI) et d'Isolement avec Migration (IM) (Figure 9A-B). Des méthodes permettent ensuite de comparer des données simulées par coalescence sous ces modèles aux données réelles et d'ajuster les différents paramètres des modèles (Temps de divergence, taille efficace des populations, intensité des échanges génétiques, etc…) par maximisation de la vraisemblance (Nielsen and Wakeley 2001). Ce type de modèle a permis de distinguer l'effet de la migration de celui du tri incomplet du polymorphisme ancestral, une question fondamentale en génomique de la spéciation. En effet, deux lignées peuvent partager des polymorphismes soit parce qu'elles les ont hérités d'un ancêtre commun, soit parce qu'elles se les sont échangés par hybridation. Distinguer ces deux processus est donc essentiel pour comprendre dans quel contexte s'est déroulée la spéciation.

Comparer le modèle IM et SI permet de déterminer s'il y a eu des échanges génétiques au cours de la divergence, mais n'indique pas comment ces échanges se sont réparti au cours du temps. En effet, le modèle IM englobe à la fois divergence primaire et contact secondaire (Roux *et al.* 2014). Des modèles

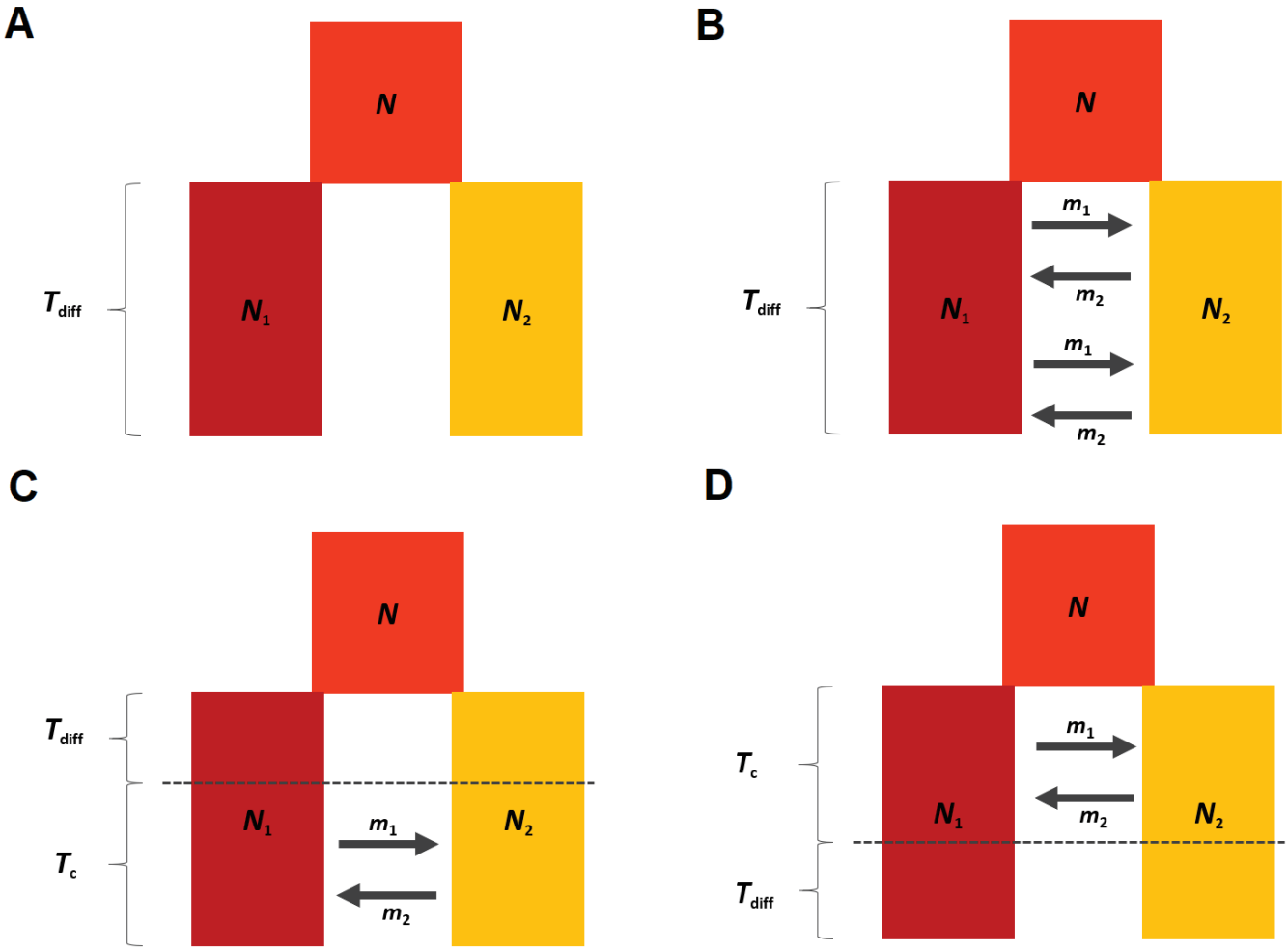


FIGURE 9 – Représentation schématique de quatre modèles de spéciation différents. Une population ancestrale de taille N se divise en deux populations filles de taille N_1 et N_2 qui divergent pendant T_{diff} générations **A.** sans flux génique (SI) ou **B.** avec un flux génique d'intensité m_1 de la population 1 vers 2 et m_2 de la population 2 vers 1 (IM). La période de divergence allopatrique peut également être **C.** suivie (SC) ou D. précédée d'une période de flux génique (AM) d'une durée de T_c générations.

plus complexes ont donc été développés par la suite, afin de considérer des variations temporelles dans les échanges génétiques ; le modèle de contact secondaire (SC) et de migration ancienne (AM) (Figure 9C-D). Cependant, le calcul analytique de la vraisemblance n'étant pas possible pour des modèles plus complexes, de nouvelles méthodes basées sur des statistiques qui résument les données ont été développées (Beaumont *et al.* 2002). Ces méthodes de calcul bayésien approximé appelées ABC « *Approximate Bayesian Computation* » comparent des statistiques mesurées sur les données réelles à celles calculées sur des données simulées sous différents scénarios de divergence, afin d'identifier le scénario ayant le plus probablement engendré les données observées.

Ces modèles restent cependant des simplifications de la réalité et négligent notamment l'hétérogénéité des processus à l'œuvre le long des génomes. En effet, les barrières aux échanges génétiques entre espèces étant généralement semi-perméables (Barton and Hewitt 1989), le flux génique n'est pas homogène le long du génome mais varie à cause des effets de sélection contre l'introgression. Ainsi, proche des gènes barrières les niveaux de flux génique vont être particulièrement réduits, alors qu'ils peuvent être très élevés en cas d'introgression adaptative. Plusieurs études ont alors proposé de prendre en compte ces variations en supposant l'existence de groupes de locus n'ayant pas les mêmes valeurs de paramètres de migration (Sousa *et al.* 2013; Roux *et al.* 2013; Tine *et al.* 2014). De même, les variations du taux local de recombinaison font varier l'intensité de la sélection en liaison et donc la taille efficace, ce qui récemment a été pris en compte par certaines études (Roux *et al.* 2016; Rougeux *et al.* 2017; Rougemont and Bernatchez 2018).

b. L'information de la liaison génétique

La plupart des méthodes d'inférence démographique développées utilisent l'information des fréquences alléliques pour inférer la durée et l'intensité des échanges génétiques entre populations (Gutenkunst *et al.* 2009; Csilléry *et al.* 2010; Pickrell and Pritchard 2012). Seulement, ces méthodes ne sont pas adaptées à l'étude de génomes entiers car elles négligent la liaison génétique qui existe entre les marqueurs. En effet, un faible nombre de locus échantillonnés suffisamment loin les uns des autres dans le génome peuvent être considérés comme indépendants. Cependant, il a été montré que négliger l'effet de la recombinaison quand un grand nombre de locus sont étudiés peut biaiser les estimations de taille de populations et de temps de divergence (Schierup and Hein 2000; Strasburg and Rieseberg 2010).

L'information haplotypique peut cependant se révéler très utile pour inférer l'histoire démographique des espèces. En effet, on peut décrire au sein d'une population des fragments chromosomiques identiques par ascendance (*Identical by descent* IBD), c'est-à-dire deux haplotypes présent chez deux individus et identiques car hérités d'un ancêtre commun (Browning and Browning 2011). Plus ces

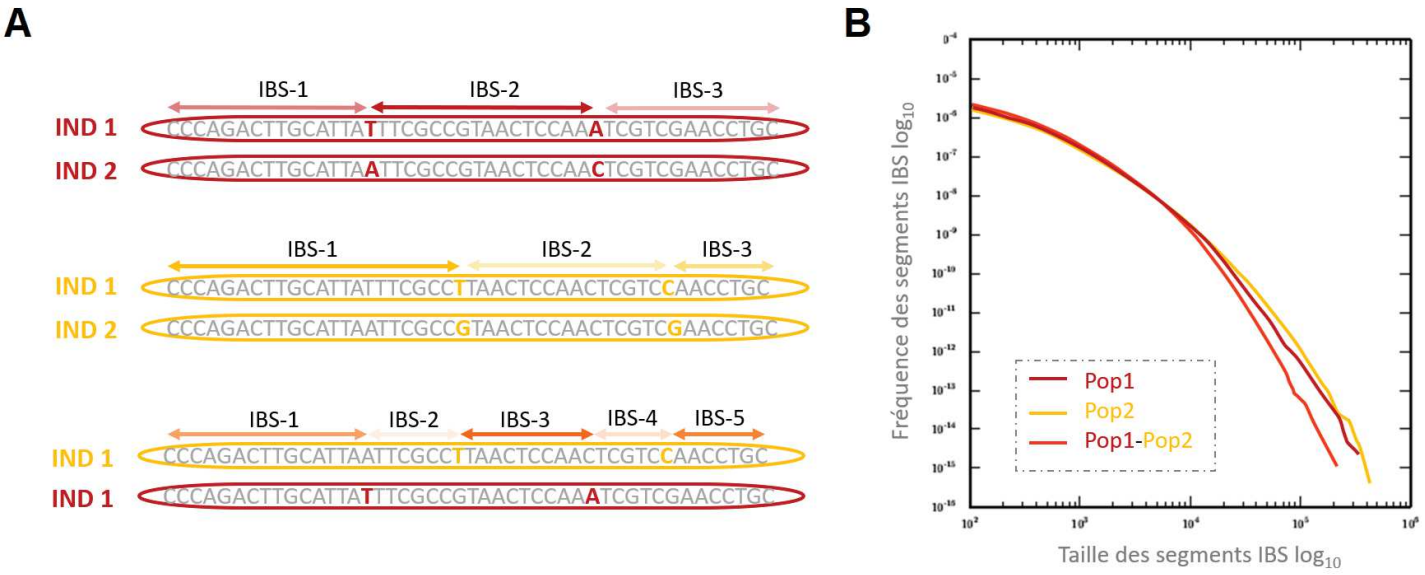


FIGURE 10 – Les segments identiques par état (IBS). A. Identification des segments IBS entre individus issu de la population 1 (rouge), de la population 2 (jaune) et entre les deux. Distribution de longueur des haplotypes identiques partagés entre et au sein de la population 1 et 2.

haplotypes sont longs, plus l'ancêtre commun qui les a transmis aux deux individus étudiés est récent, la taille des fragments IBD est donc informative des paramètres démographiques de la population. Les changements démographiques vont donc affecter la taille des fragments IBD. Plus une population aura une taille efficace large, plus elle sera diversifiée génétiquement et donc plus les fragments IBD seront courts. On peut également chercher des segments IBD entre individus issus de deux populations différentes, ce qui équivaut à s'intéresser aux haplotypes migrants ou introgressés. La recombinaison agissant également dans ce cas-là comme une horloge, la distribution de taille des fragments IBD contient de l'information sur la démographie récente de la population.

Plutôt que d'étudier les segments IBD dont il faut pouvoir retracer la généalogie, Harris et Nielsen (2013) ont proposé de s'intéresser aux segments identiques par état (*Identical by state* IBS). Un segment IBS de taille *L* est définit comme *L* paires de bases contigües encadrées par deux SNPs (Figure 10A) qui sont observés entre deux individus ou au sein d'un même individu (pour les diploïdes). Pour étudier l'histoire de la divergence et des échanges génétiques entre deux lignées, il faut s'intéresser à la distribution de taille des segments IBS au sein de chacune des deux populations mais également à celle mesurée entre les populations (Figure 10B). En effet, les distributions intra-populationnelles portent le signal des changements démographiques propres à chaque population alors que la distribution inter-populationnelle porte la trace des échanges génétiques (Harris and Nielsen 2013). Un flux génétique récent entre les deux populations va par exemple créer un excès d'haplotypes longs partagés entre les deux populations. Ces trois distributions peuvent être prédites sous différents modèles d'histoire démographique et la vraisemblance de chaque modèle est ensuite maximisée en comparant les distributions observées et théoriques.

De nombreuses méthodes ont été développées pour étudier l'histoire démographique des espèces en prenant en compte l'information de la liaison génétique et en considérant des scénarios de plus en plus complexes (Gravel 2012; Palamara *et al.* 2012; Pugach *et al.* 2016; Browning *et al.* 2018; Ni *et al.* 2018). Utiliser ce genre d'approches est une première étape essentielle pour comprendre comment la divergence entre les lignées a évolué et pouvoir par la suite identifier les régions impliquées dans l'isolement reproductif et les processus évolutifs ayant permis leur mise en place.

2. Identifier les régions génomiques impliquées dans l'isolement reproductif

Une fois le contexte de la divergence étudié, l'objectif est d'identifier les régions génomiques, voire même plus précisément les locus, impliqués dans l'isolement reproductif. Ceci inclut les locus sous

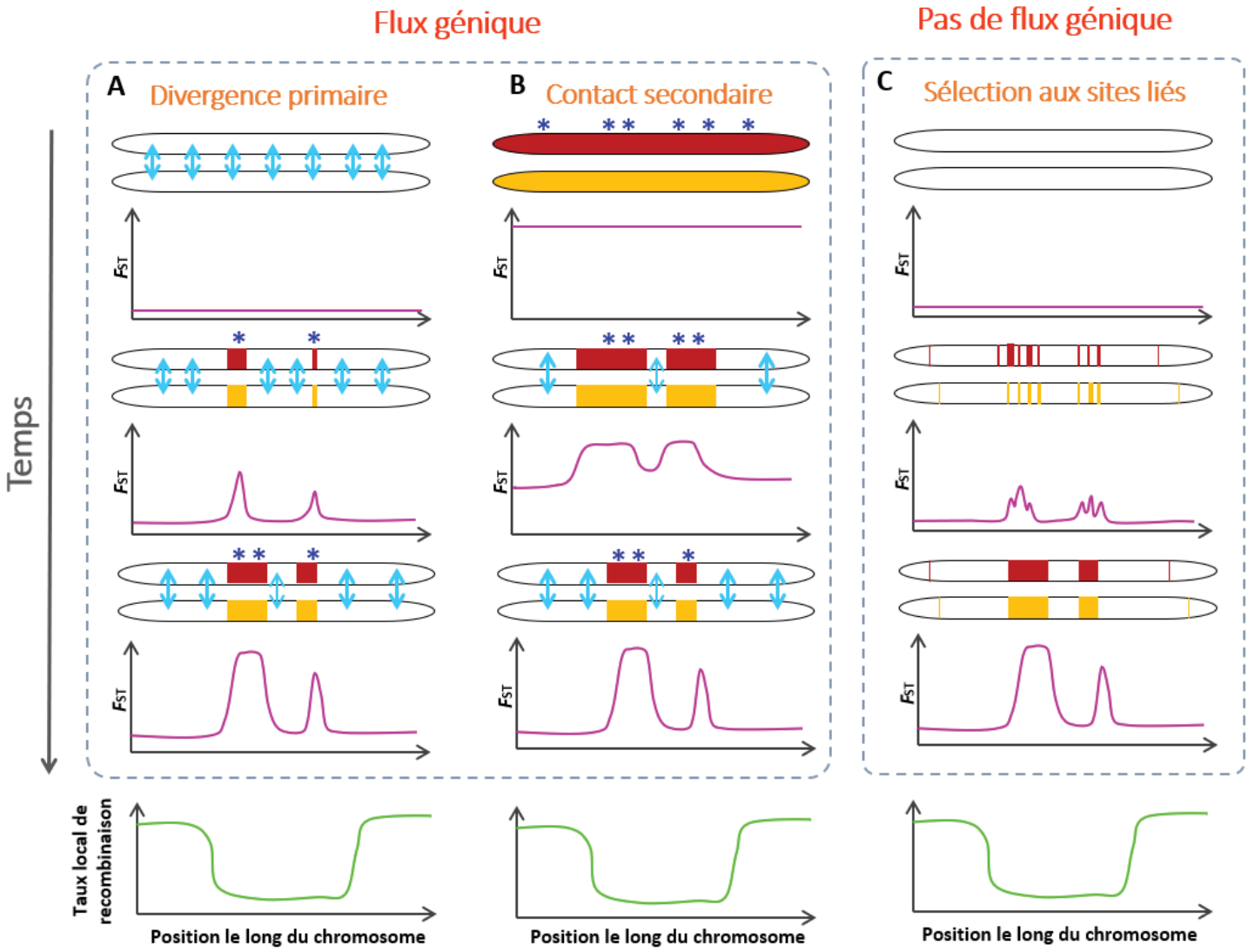


FIGURE 11 – Représentation schématique des différents mécanismes pouvant expliquer la formation des îlots génomiques de différenciation entre deux lignées en cours de divergence. Pour chaque modèle trois étapes sont représentées avec pour chacune un chromosome par population. Les couleurs (rouge et jaune) indiquent les régions chromosomiques où la différenciation génétique (F_{ST}) est marquée entre les deux lignées. Les flèches bleues représentent le flux génique entre les deux populations et les étoiles bleues la présence de locus impliqués dans l'isolement reproductif que ce soit dû à des incompatibilités génétiques ou à des gènes d'adaptation locale. Le paysage de F_{ST} est représenté à chacune des trois étapes afin de montrer son évolution. **A.** Divergence primaire initiée par une réduction du flux génique autour des gènes d'adaptation locale et progressivement étendue à d'autres régions génomiques par établissement de nouvelles mutations localement avantageuses. **B.** Contact secondaire entre populations génétiquement différenciées induisant une érosion progressive de la différenciation préexistante autour des gènes impliqués dans l'isolement reproductif. **C.** Différenciation hétérogène induite par des variations d'intensité des effets de la sélection aux sites liés en absence de flux génique. Les variations chromosomiques du taux de recombinaison affectent les trois mécanismes en modulant : (*i*) l'intensité de la sélection en liaison, (ii) le cumul de l'effet barrière des locus impliqués dans l'isolement reproductif et (iii) le maintien des combinaisons adaptatives face au flux génique. Qu'ils agissent de façon combinée ou séparément ces effets tendent à concentrer les îlots génomiques de différenciation dans les régions à faible taux de recombinaison quel que soit le mécanisme de divergence.

sélection écologique divergente, impliqués dans le choix de partenaires ou dans des incompatibilités post-zygotiques intrinsèques qui peuvent être neutres au sein de la population dans laquelle ils ont évolué (Ravinet *et al.* 2017). Ces locus vont avoir pour effet de réduire le taux local de flux génique et ainsi permettre à la différenciation génétique, c'est-à-dire à la différence de fréquence allélique, de s'accumuler entre les lignées. De nombreuses études se sont donc concentrées sur les patrons de différenciation génétique, la mesure la plus classiquement utilisée étant le F_{ST} de Wright (1949), qui permet de mesurer le niveaux de différenciation entre sous-populations prédéfinies. Il peut être approximé comme $F_{ST} = (\pi_{total} - \pi_{intra\text{-}pop}) / \pi_{total}$ où π correspond à la diversité génétique mesurée au sein de chaque population ($\pi_{intra\text{-}pop}$) ou en considérant tous les individus (π_{total}) et est donc une mesure relative de la différenciation génétique.

Le développement des nouvelles technologies de séquençage à haut débit a permis d'obtenir une vision pan-génomique de la différenciation. De nombreuses études réalisées chez des paires d'espèces partiellement isolées ont mis en évidence l'existence de patrons de différenciation génétique hétérogènes le long des génomes (Turner *et al.* 2005; Harr 2006; Nosil *et al.* 2009; Nadeau *et al.* 2012; Ellegren *et al.* 2012; Gagnaire *et al.* 2013). Ces patrons se caractérisent par l'existence de régions génomiques faiblement différenciées qui alternent avec des régions ou la différenciation est forte, appelées îlots génomique de différenciation (Harr 2006). Ces îlots ont tout d'abord été vu comme des marqueurs de la présence de locus d'isolement reproductif et ont donc été appelés « îlots de spéciation » (Turner *et al.* 2005). Différents scénarios de divergence mettant l'accent sur les effets relatifs de la sélection et du flux génique ont alors été élaborés pour expliquer leur formation.

Le premier est un modèle de divergence primaire au cours duquel les îlots se construisent progressivement (Figure 11A). Sous ce modèle deux populations divergent en sympatrie en s'adaptant à deux environnements différents. Lors de la première phase de ce continuum, la sélection divergente agissant sur un petit nombre de gènes impliqués dans l'adaptation locale augmente le niveau de différenciation génétique sur ces locus, ainsi qu'aux locus neutres liés (Feder *et al.* 2012). Ces régions étant impliquées dans l'adaptation locale, elles sont protégées du flux génique, les migrants étant contre-sélectionnés en dehors de leur habitat d'origine. Cette protection faciliterait l'établissement de nouvelles mutations localement avantageuses ce qui fait progressivement augmenter la taille des îlots de différenciation (Smadja *et al.* 2008; Via and West 2008; Via 2009). Ce mécanisme appelé « *divergence hitchhiking* » pourarit ainsi favoriser l'accumulation de nouveaux gènes d'adaptation au cours du temps, jusqu'à entrainer une réduction du flux génique à l'échelle du génome lors de la phase de « *genome hitchhiking* » (Feder *et al.* 2012).

Le deuxième mécanisme proposé pour expliquer la formation des îlots est celui d'une érosion de la différenciation lors d'un contact secondaire entre populations divergentes (Figure 11B). Lors d'une première phase de divergence en allopatrie, les populations peuvent fixer des allèles conférant des adaptations à leurs environnements respectifs ou des incompatibilités génétiques (Bierne *et al.* 2011b). En effet, la fixation d'incompatibilités de type Dobzhansky-Muller est quasiment inévitable durant la divergence allopatrique (Presgraves 2010). La reprise des échanges génétiques lors du contact va alors venir éroder la différenciation précédemment accumulée pendant la phase d'allopatrie, sauf au niveau des locus impliqués dans l'isolement reproductif qui agissent comme des barrières à l'introgression (Barton and Hewitt 1985; Barton and Bengtsson 1986; Payseur 2010). Ici, les îlots sont donc formés par l'érosion hétérogène d'un état prédifférencié dans une dynamique temporelle inversée par rapport à celle du mécanisme précédent.

Sous ces deux mécanismes, les îlots de différenciation sont principalement attendus dans les régions où la recombinaison génétique est faible. En effet, dans le cas de la divergence primaire, l'effet d'auto-stop qui permet aux îlots de s'étendre est d'autant plus fort que la recombinaison est faible. Dans le cas du contact secondaire, les locus d'isolement reproductifs combinent plus facilement leurs effets et donc résistent mieux à l'introgression dans les régions où ils sont en forte liaison génétique (Yeaman *et al.* 2016). Les patrons génomiques générés par ces deux mécanismes sont donc identiques et il peut être difficile de les distinguer si le contexte géographique de la divergence n'est pas connu (Bierne *et al.* 2013). Bien que sous ces deux modèles les îlots contiennent des locus impliqués dans l'isolement reproductif, les forces évolutives ayant permis leur mise en place sont différentes. En effet, une forme de sélection disruptive est nécessaire pour que les îlots se mettent en place dans le modèle de divergence primaire, alors que pour le modèle de contact secondaire, l'action seule de la dérive génétique peut permettre la mise en place des locus d'isolement reproductif. C'est pourquoi il est nécessaire d'identifier au préalable dans quel contexte démographique et géographique a eu lieu la divergence afin de déterminer quelles sont les forces évolutives qui ont permis sa mise en place.

Un troisième modèle n'impliquant pas de flux génétique hétérogène a également été proposé (Nachman and Payseur 2012; Cruickshank and Hahn 2014) (Figure 11C). Il a été montré que les îlots de F_{ST} peuvent également se former pendant une période de divergence allopatrique en l'absence de locus impliqués dans l'isolement reproductif. En effet, dans les régions à faible taux de recombinaison, la sélection en liaison qu'elle soit positive (balayage sélectif (Maynard Smith and Haigh 1974)) ou négative (sélection d'arrière-plan (Charlesworth *et al.* 1993)) réduit localement la diversité génétique. Or, le F_{ST} est une valeur relative de la différenciation génétique entre populations qui dépend du niveau de diversité génétique intra-populationnel. En effet, étant donné qu'il est négativement corrélé à la diversité intra-populationnelle, des valeurs fortes de F_{ST} peuvent être localement générées par des

réductions de la diversité intra-populationnelle. Les variations locales de taux de recombinaison peuvent donc générer des variations de diversité génétique et donc localement créer des pics de F_{ST}. En effet, bien que les balayages sélectifs soient assez rares dans les génomes, l'effet répété de la sélection d'arrière-plan au niveau des séquences fonctionnelles peut être suffisant pour faire varier la diversité génétique.

Plusieurs mécanismes impliquant ou pas la présence de locus d'isolement reproductif peuvent donc générer des îlots de différenciation génétique (Harrison and Larson 2016; Wolf and Ellegren 2016; Ravinet *et al.* 2017). Il a donc été proposé d'utiliser d'autres statistiques en plus du F_{ST} pour pouvoir distinguer les îlots impliqués dans l'isolement reproductif de ceux qui ne le sont pas. Le d_{XY} (Nei 1987) qui mesure la divergence brute entre populations comme le nombre moyen de substitutions nucléotidiques observées entre les populations a par exemple été proposé pour distinguer les îlots impliqués dans l'isolement reproductif des autres (Cruickshank and Hahn 2014). En effet, le d_{XY} peut être vu comme une mesure absolue de la divergence étant donné qu'il n'est pas confondu par des processus intra-populationnels. Ainsi, si une région génomique présente à la fois des valeurs élevées de d_{XY} et de F_{ST} il y a de grandes chances qu'elle soit impliquée dans l'isolement reproductif. Avoir une mesure directe du flux génique reste néanmoins le moyen le plus direct pour révéler la présence de barrières d'isolement reproductif. Dans tous les cas les patrons observés dans les génomes peuvent être générés par différentes formes d'interaction entre sélection et démographie et ne représentent qu'une image instantanée d'un processus dynamique. C'est pourquoi l'étude du processus de spéciation est principalement une démarche rétrospective.

3. Détecter la sélection : les approches d'évolution moléculaire

Une fois les régions impliquées dans l'isolement reproductif identifiées, une question perdure sur les gènes qui les composent et les forces évolutives ayant permis leur mise en place. En effet, on peut se demander si l'isolement reproductif est lié à des gènes subissant une contrainte évolutive particulière en lien avec leur fonction mais également quelles pressions de sélection (positive ou négative) ont permis l'apparition des nouveaux allèles. En effet, on peut aussi bien imaginer que l'isolement reproductif soit généré par des allèles impliqués dans l'adaptation et ayant évolué sous sélection positive, que par des allèles faiblement neutres (ou faiblement délétères) dans leur fond génétique d'origine ayant fixés par dérive génétique et impliqué dans des interaction épistatiques complexes. De même, l'isolement pourrait être associé à des séquences à évolution rapide, simplement en raison de leur vitesse d'évolution ou à des gènes fortement conservés phylogénétiquement, un changement étant plus susceptible d'induire des effets importants du fait de leur importance fonctionnelle. Etudier les signatures d'évolution moléculaire des gènes au cœurs des îlots impliqués dans l'isolement

reproductif est donc une étape essentielle afin de comprendre comment les barrières d'isolement reproductif se mettent en place.

a. Le ratio dN/dS

L'évolution moléculaire est une discipline qui combine des concepts de biologie moléculaire et de biologie évolutive pour étudier l'évolution des molécules biologiques. Elle a notamment pour objectif de déterminer si les changements observés dans les molécules, et par extension les génomes, sont dus à des processus neutres ou sélectifs. Différentes méthodes ont été développées pour détecter les locus sous sélection en se basant sur l'étude des régions codantes des génomes. Le principe de certains tests est de comparer le nombre de changements synonymes et non-synonymes (Nei and Gojobori 1986). Quand une mutation apparait dans une séquence codante d'ADN il y a deux possibilités : elle peut entrainer ou pas un changement dans la séquence protéique. En effet, le code génétique reliant les séquences d'ADN des gènes à leur protéine est déterminé par des triplets de trois nucléotides (appelés codons). L'ADN étant constitué de quatre bases nucléotidiques différentes (A, T, G et C), il existe 64 combinaisons de trois nucléotides possibles, qui ne codent que pour 22 acides aminées (qui composent les protéines). Certains codons codent donc pour le même acide aminé et le changement d'un nucléotide dans ces codons peut ne pas entrainer de changements d'acide aminé. Par exemple, l'isoleucine est codée par 3 codons différents ATT, ATC ou ATA. C'est pourquoi le code génétique est dit redondant ou dégénéré. Si le changement d'acide nucléique n'entraine pas de changement d'acide aminé on parle de mutation synonyme, dans le cas contraire on parle de mutation non-synonyme.

Les changements synonymes sont considérés comme neutres et servent donc de contrôle reflétant le taux de mutation de base du génome. Les changements non-synonymes quant à eux peuvent être avantageux, neutres ou délétères et sont donc sous sélection positive ou négative. Ainsi, pour déterminer si un gène donné a subi des pressions de sélection au cours du temps on va comparer la séquence de ce gène entre deux espèces suffisamment éloignées phylogénétiquement et compter le nombre de substitutions synonymes (dS) et non-synonyme (dN). Le ratio de dN/dS permet ensuite d'estimer s'il y a eu plus ou moins de changements non-synonymes par rapport aux synonymes. Si aucune pression de sélection n'agit sur le gène alors on s'attend à voir autant de changements non-synonymes que synonymes et dN/dS = 1. Les mutations non-synonymes qui apparaissent dans un gène qui a déjà été optimisé par la sélection sont majoritairement délétères, donc éliminées par la sélection naturelle et n'arrivent pas à fixation. Le ratio du dN/dS est donc inférieur à 1 ce qui indique que ce gène est sous sélection purificatrice. Enfin, si les mutations non-synonymes sont conservées par la sélection naturelle c'est qu'elles sont avantageuses, un ratio du dN/dS supérieur à 1 est donc le signe d'une sélection positive.

Cependant l'interprétation du ratio dN/dS reste très conservative. En effet, un ratio inférieur à 1 ne veut pas forcément dire qu'aucune des substitutions non-synonymes observées n'a été avantageuse étant donné que les pressions de sélections qui s'exercent sur un gène peuvent varier d'un site à l'autre. Si la majorité des sites sont sous sélection purificatrice mais que certains sont sous sélection positive, ce qui est généralement le cas du site actif de la protéine, alors le ratio dN/dS sera inférieur à 1 bien que certains sites évoluent sous sélection positive. La sélection positive reste donc particulièrement difficile à détecter.

b. Le ratio π_n/π_s et le test McDonald-Kreitman

La comparaison du nombre de changements synonymes et non-synonymes peut également se faire en étudiant le polymorphisme d'une espèce et non pas la divergence entre deux lignées. Dans ce cas-là on regarde au sein d'un échantillon d'individus issus d'une même population le nombre moyen de polymorphismes synonymes (π_s) et non-synonymes (π_n). La différence principale entre le ratio dN/dS et π_n/π_s vient du fait que la dynamique de maintien des mutations positives et délétères est différente dans la divergence et le polymorphisme. En effet, les mutations faiblement délétères peuvent échapper à la sélection quand elles sont à faible fréquence et donc se maintenir dans le polymorphisme sans pour autant jamais atteindre la fixation. Ainsi, on peut s'attendre à voir des valeurs élevées de π_n à cause de la présence de mutations faiblement délétères et non pas à la présence de mutations positives. Cet effet sera d'autant plus marqué que la taille efficace est réduite car la sélection purificatrice sera moins efficace à purger les mutations faiblement délétères. De plus, les mutations non-synonymes avantageuses vont le plus souvent fixer rapidement au sein de la population et donc ne contribuer que temporairement au π_n et ainsi être plus visible dans la divergence (dN) que le polymorphisme.

McDonald et Kreitman (1991) ont alors proposé un test basé sur la comparaison des ratio dN/dS et π_n/π_s avec l'idée que si toutes les mutations sont neutres alors dN/dS = π_n/π_s. Les mutations adaptatives ne ségrégant pas dans le polymorphisme, elles sont principalement dans la divergence, un dN/dS > π_n/π_s est alors vu comme le signe d'une sélection positive. A l'inverse les mutations délétères étant majoritairement présentes dans le polymorphisme sans pour autant atteindre la fixation, un π_n/π_s > dN/dS est vu comme le signe de l'accumulation de mutations faiblement délétères. On peut alors définir l'index de neutralité (*NI*) comme *NI* = $(\pi_n/\pi_s)/(dN/dS)$ qui va indiquer le sens et l'écart par rapport à la neutralité, c'est-à-dire *NI*=1. Plus tard une extension de ce test a été proposée afin d'obtenir la proportion de substitutions non-synonymes avantageuses α où $\alpha = 1 - NI$ c'est-à-dire $\alpha = 1 - (dS * \pi_n / dN * \pi_s)$ (Smith and Eyre-Walker 2002). Ce qui permet également d'estimer le taux de substitution adaptatives $\omega_a = \alpha * (dN/dS)$ et le taux de substitutions non-adaptatives $\omega_{na} = (1 - \alpha) * (dN/dS)$.

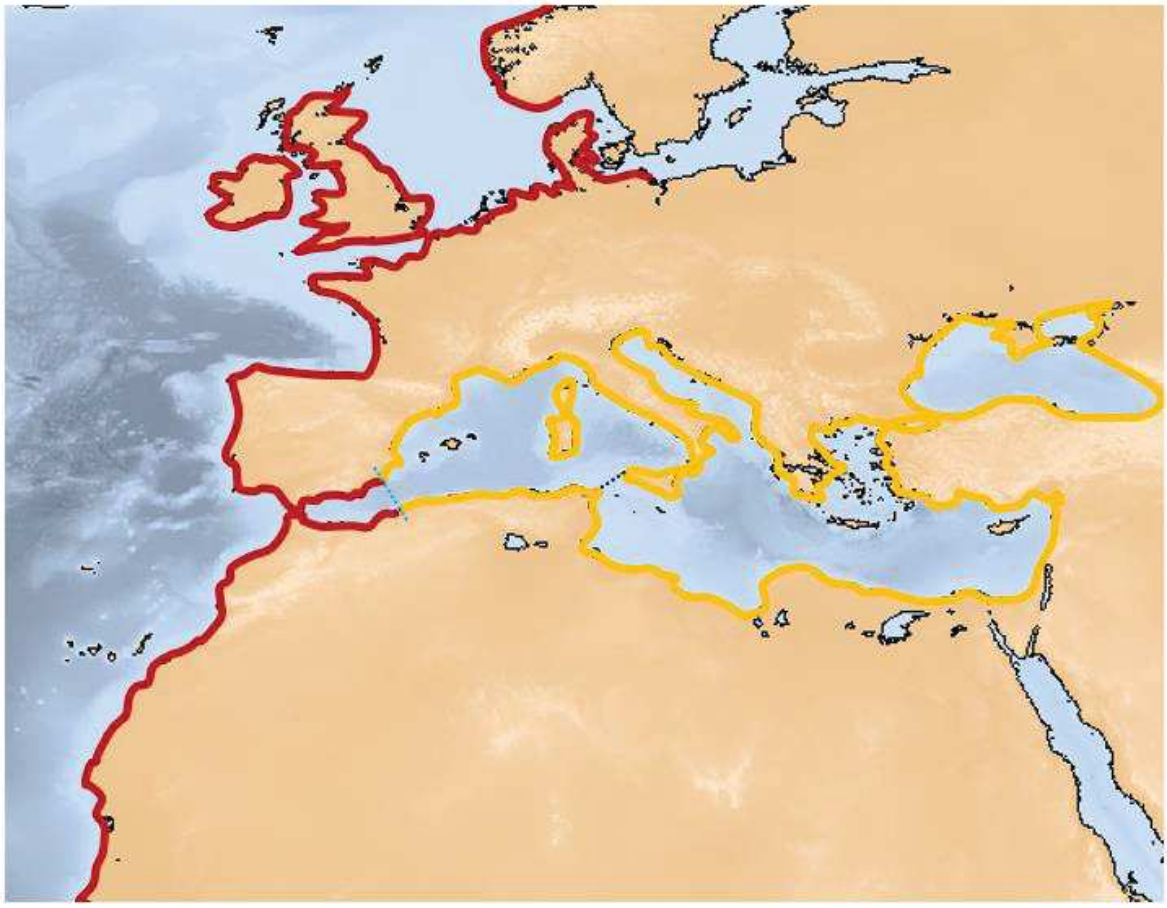


FIGURE 12 – Aire de répartition du bar européen. La couleur rouge représente la répartition des individus issus de la population atlantique et la couleur jaune celle des individus méditerranéens.

IV. Modèle d'étude et objectifs de la thèse

1. Le bar européen, *Dicentrarchus labrax*

Le bar Européen (*Dicentrarchus labrax*) est un poisson téléostéen de la famille des moronidés (*Moronidae*) qui comprend six espèces ; *Dicentrarchus labrax* et *Dicentrarchus punctatus* sur les côtes européennes, *Morone saxatilis*, *Morone americana*, *Morone mississippiensis* et *Morone chrysops* sur les côtes ou dans les rivières nord-est américaines. Son aire de répartition va de l'Atlantique Nord-Est (des côtes du Maroc au sud au sud de la Norvège au nord) et s'étend sur toute la Mer Méditerranée ainsi que la mer Noire (Figure 12). On le retrouve dans les eaux côtières, jusqu'à une centaine de mètres de fond et environ 80 km des côtes mais également dans les zones estuariennes et les lagunes côtières particulièrement en Méditerranée (Kelley 1988; Dufour *et al.* 2009). C'est un poisson euryhalin et eurytherme supportant une large gamme de salinités et de températures ce qui explique qu'on le retrouve aussi bien dans les eaux froides de la Mer du Nord que dans les eaux chaudes des lagunes tunisiennes. C'est une espèce démersale qui nage près du fond pour chasser ses proies majoritairement des poissons (Pickett and Pawson 1994).

La reproduction a lieu une fois par an et est légèrement décalée entre l'Atlantique et la Méditerranée. En effet, la ponte a lieu de décembre à mars en Méditerranée et jusqu'en juin en Atlantique en fonction de la température de l'eau (Pawson *et al.* 2000). Durant la phase larvaire qui est relativement longue chez cette espèces (8 à 12 semaines) les larves se déplacent avec les courants et entrent généralement dans les estuaires où les juvéniles vont passer l'été et ressortir à l'automne (Pickett and Pawson 1994; Dufour *et al.* 2009). Cependant, il semblerait qu'il existe deux stratégies d'exploitation de l'habitat pour les juvéniles. En effet, certains restent en mer ce qui leur donne accès à une ressource alimentaire plus réduite mais garanti une stabilité des paramètres environnementaux alors que ceux préférant l'habitat lagunaire ont un accès plus facile à la nourriture mais subissent des conditions environnementales plus changeantes.

La structure génétique de l'espèce a été étudiée depuis longtemps, tout d'abord à l'aide d'allozymes (Allegrucci *et al.* 1997), puis de marqueurs microsatellites (Naciri *et al.* 1999; Bahri-Sfar *et al.* 2000), mitochondriaux (Lemaire *et al.* 2005), SNPs (Souche 2009) et plus récemment grâce au séquençage du génome du bar (Tine *et al.* 2014). Toutes ces études ont confirmé l'existence de deux lignées évolutivesdifférentes représentées par la population atlantique et méditerranéenne (où le bar est appelé loup), les deux lignées possédant en plus des différences dans leur génome nucléaire des haplotypes mitochondriaux différents (Lemaire *et al.* 2005). La zone de transition entre les deux fonds génétiques se situe au niveau du front océanique Almeria-Oran (Naciri *et al.* 1999), comme c'est le cas pour de nombreuses espèces marines ayant une distribution atlantico-méditerranéenne (Patarnello *et*

al. 2007). Il a également été montré que les deux lignées s'hybrident au niveau de la mer d'Alboran (Lemaire *et al.* 2005).

Certaines études se sont également focalisées sur la description de la structure génétique au sein de chacune des deux lignées. La population atlantique apparait très faiblement structurée mais Il existe néanmoins une légère différenciation de la population du sud du Portugal par rapport au reste de l'Atlantique, probablement liée à l'influx de matériel génétique depuis la Méditerranée (Castilho and McAndrew 1998; Souche 2009). La population de Méditerranée est quant à elle beaucoup plus structurée. En effet, il existe une sub-subdivision entre la Méditerranée Ouest et Est, la transition se faisant au niveau du détroit siculo-tunisien qui agit comme une barrière à la dispersion (Bahri-Sfar *et al.* 2000; Souche 2009).

L'étude la plus récente ayant étudié la différenciation génétique entre les deux lignées à l'échelle du génome a révélé qu'elles ont divergé suite à l'isolement géographique de deux populations il y a environ 270 000 ans (Tine *et al.* 2014). Ces deux lignées se sont par la suite remises en contact au début de la période interglaciaire actuelle il y a environ 12 000 ans, illustrant l'influence des cycles glaciaires du Pléistocène sur la structuration génétique des populations (Hewitt 1996, 2000). Les deux lignées ont donc probablement divergé en allopatrie dans deux refuges glaciaires différents. Ce contact secondaire a permis la reprise des échanges génétique de façon asymétrique, avec une introgression majoritaire d'allèles atlantiques dans le fond génétique méditerranéen. L'analyse du paysage chromosomique de différenciation entre lignées a également révélé l'existence d'îlots génomiques de différenciation sur la quasi-totalité des chromosomes, le plus souvent dans les régions à faible taux de recombinaison (Tine *et al.* 2014). En effet, les taux de recombinaison sont assez variables le long du génome, les régions centro-chromosomique présentant des taux de recombinaison faible contrastant avec les régions péri-chromosomique à forte recombinaison et ce pour quasiment tous les chromosomes (Tine *et al.* 2014).

2. Objectifs de la thèse

Le bar européen constitue un bon modèle d'étude pour comprendre quels sont les forces évolutives impliquées dans le processus de spéciation et plus particulièrement la mise en place de l'isolement reproductif. En effet, étudier les patrons génomiques de différenciation afin d'identifier les locus impliqués dans l'isolement reproductif peut être compliqué si tous les mécanismes confondants et le contexte géographique de la divergence ne sont pas pris en compte. Or, une précédente étude a déjà démontré qu'un contact secondaire entre la lignée atlantique et méditerranéenne de bar européen a permis la reprise des échanges génétiques (Tine *et al.* 2014), qui sont essentiels pour révéler la présence de locus impliqués dans l'isolement reproductif. De plus, la plupart des espèces de poissons

marins présentent des caractéristiques idéales pour les études de génomique des populations. En effet, elles ont généralement de grandes tailles de population (N_e) permettant à la sélection d'être plus efficace ainsi qu'une phase larvaire planctonique permettant une migration (m) à longue distance et donc la possibilité à un fort flux génique (N_e*m). Cette espèce représente donc un modèle idéal pour comprendre les interactions existantes entre migration et sélection contre l'introgression et ainsi identifier les régions impliquées dans l'isolement reproductif. En effet, le fait que la différenciation génétique se soit maintenue entre les deux lignées malgré de nombreuses générations d'hybridation suggère fortement que l'introgression à un effet délétère sur le long terme. Cependant, il a également été montré que les hybrides atlantiques-méditerranéens de première génération ont une survie supérieure à celle de leur parents (Guinand *et al.* 2017). Il semblerait donc que l'introgression ait des effets différents à court et long terme. Comprendre clairement quelles forces évolutives ont permis à la divergence de se mettre en place et de se maintenir nécessite donc de combiner plusieurs approches afin d'identifier les locus impliqués dans l'isolement reproductif, comprendre quelles forces évolutives ont permis leur mise en place et quels sont leurs effets sur l'introgression à court et long terme. C'est ce que propose de faire cette thèse au travers de quatre chapitres principaux.

L'objectif principale du **Chapitre 1** est de clarifier dans quel contexte la divergence entre la lignée atlantique et méditerranéenne a eu lieu afin d'identifier quel mécanisme est à l'origine des îlots de différenciation génétique observés entre le bar et le loup. En effet, bien que l'étude de Tine *et al.* (2014) ai mis en évidence l'existence d'un contact secondaire entre les deux lignées de bar européen, les variations climatiques du Pléistocène laissent penser que plusieurs périodes d'isolement-contact ont pu se répéter dans le temps. L'utilisation de nouveaux génomes entièrement séquencés et phasés a permis d'utiliser une nouvelle méthode d'inférence démographique basée sur l'information de la liaison génétique. La détection des haplotypes atlantiques introgressés en méditerranée a ensuite permis d'avoir une mesure directe du flux génique et de comprendre comment les îlots génomiques se sont formés. Ceci a permis de distinguer les îlots génomiques impliqués dans l'isolement reproductif de ceux qui ne le sont pas.

Une fois les régions impliquées dans l'isolement reproductif clairement identifiées et délimitées, nous avons constaté qu'elles présentaient un niveau de divergence (d_{XY}) beaucoup plus élevé que ce qui pouvait être attendu sous un processus classique de divergence allopatrique entre la lignée atlantique et méditerranéenne. L'objectif du **Chapitre 2** a donc été de déterminer si cet excès de divergence est dû à la présence d'allèles anciens maintenus polymorphes dans la population ancestrale et différentiellement triés entre les deux lignées ou à un évènement d'introgression ancien avec une troisième lignée. Pour cela, différentes méthodes pour détecter l'introgression ancienne ont été utilisées.

L'objectif du **chapitre 3** a été de comprendre quelles sont les forces évolutives ayant permis la mise en place des barrières d'isolement reproductif identifiées entre la lignée atlantique et méditerranéenne. Pour cela, des méthodes d'évolution moléculaire ont été utilisées afin de comparer les signatures d'évolution moléculaire des gènes localisés dans les régions impliquées dans l'isolement reproductif par rapport aux autres. Etant donné que la recombinaison a un impact sur l'efficacité de la sélection en modulant la taille efficace le long du génome, le taux de recombinaison populationnel a également été réestimé. Ceci a permis d'établir un lien entre taux local de recombinaison, force de la sélection et mise en place de l'isolement reproductif.

Sachant que l'isolement reproductif passe principalement par la contre sélection des individus hybrides, la présence de locus impliqués dans l'isolement reproductif entre la lignée atlantique et méditerranéenne semble difficile à relier à l'existence d'une vigueur hybride chez les F1 (Guinand *et al.* 2017). Il est donc nécessaire de comprendre comment se comportent les locus d'isolement reproductif dans les premières générations d'hybridation. Or ces individus sont trop rares pour être facilement observables en nature. Le **Chapitre 4** porte donc sur l'analyse de croisement expérimentaux d'individus mâles F1 avec des femelles méditerranéennes afin de produire des *backcross*-méditerranéens. L'objectif étant de déterminer s'il existe des différences de valeurs sélectives entre les méditerranéens et les hybrides (en termes de survie et de condition) et de voir comment se comportent les locus d'isolement reproductif dans les premières générations d'hybridation.

Enfin, les données haplotypiques utilisées dans le chapitre 1 ont permis de répondre à une question plus appliquée. En effet, l'objectif du **chapitre 5** a été de comparer la distribution de taille des haplotypes introgressées dans la population Ouest- et Est-méditerranéenne afin d'estimer le temps de migration entre les deux populations méditerranéennes. Les haplotypes entrant en premier dans la population Ouest ils sont en moyenne plus longs que ceux atteignant la population Est. La différence de taille entre les deux représente l'action de la recombinaison pendant le temps nécessaire aux haplotypes pour traverser la Méditerranée. Cette approche originale nous a permis de montrer que la longueur des haplotypes introgressés permet d'estimer l'échelle spatiale de la dispersion pour améliorer les mesures de gestion et de conservation de l 'espèce.

V. Références

Abbott R., D. Albach, S. Ansell, J. W. Arntzen, S. J. E. Baird, *et al.*, 2013 Hybridization and speciation. J. Evol. Biol. 26: 229–246. https://doi.org/10.1111/j.1420-9101.2012.02599.x

Allegrucci G., C. Fortunato, and V. Sbordoni, 1997 Genetic structure and allozyme variation of sea bass (Dicentrarchus labrax and D. punctatus) in the Mediterranean Sea. Marine Biology 128: 347–358. https://doi.org/10.1007/s002270050100

Bahri-Sfar L., C. Lemaire, O. K. B. Hassine, and F. Bonhomme, 2000 Fragmentation of sea bass populations in the western and eastern Mediterranean as revealed by microsatellite polymorphism. Proceedings of the Royal Society of London B: Biological Sciences 267: 929–935. https://doi.org/10.1098/rspb.2000.1092

Barton N., 1979 Gene flow past a cline. Heredity 43: 333–339.

Barton N. H., 1983 Multilocus Clines. Evolution 37: 454–471. https://doi.org/10.1111/j.1558-5646.1983.tb05563.x

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

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

Barton N. H., and G. M. Hewitt, 1989 Adaptation, speciation and hybrid zones. Nature 341: 497. https://doi.org/10.1038/341497a0

Bateson W., 1909 Heredity and variation in modern lights. Darwin and modern science 85–101.

Beaumont M. A., W. Zhang, and D. J. Balding, 2002 Approximate Bayesian Computation in Population Genetics. Genetics 162: 2025–2035.

Bierne N., P. Borsa, C. Daguin, D. Jollivet, F. Viard, *et al.*, 2003 Introgression patterns in the mosaic hybrid zone between Mytilus edulis and M. galloprovincialis. Molecular Ecology 12: 447–461. https://doi.org/10.1046/j.1365-294X.2003.01730.x

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

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

Bierne N., P.-A. Gagnaire, and P. David, 2013 the geography of introgression in a patchy environment and the thorn in the side of ecological speciation. Current Zoology 59: 72–86.

Blair W. F., 1955 Mating Call and Stage of Speciation in the Microhyla Olivacea—M. Carolinensis Complex. Evolution 9: 469–480. https://doi.org/10.1111/j.1558-5646.1955.tb01556.x

Boore J. L., and W. M. Brown, 1998 Big trees from little genomes: mitochondrial gene order as a phylogenetic tool. Current Opinion in Genetics & Development 8: 668–674. https://doi.org/10.1016/S0959-437X(98)80035-X

Browning S. R., and B. L. Browning, 2011 Haplotype phasing: existing methods and new developments. Nat Rev Genet 12: 703–714. https://doi.org/10.1038/nrg3054

Browning S. R., B. L. Browning, M. L. Daviglus, R. A. Durazo-Arvizu, N. Schneiderman, *et al.*, 2018 Ancestry-specific recent effective population size in the Americas. PLOS Genetics 14: e1007385. https://doi.org/10.1371/journal.pgen.1007385

Burri R., 2017 Interpreting differentiation landscapes in the light of long-term linked selection. Evolution Letters. https://doi.org/10.1002/evl3.14

Burton R. S., and F. S. Barreto, 2012 A disproportionate role for mtDNA in Dobzhansky–Muller incompatibilities? Mol Ecol 21: 4942–4957. https://doi.org/10.1111/mec.12006

Castilho R., and B. McAndrew, 1998 Two polymorphic microsatellite markers in the European seabass, Dicentrarchus labrax (L.). Animal Genetics.

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

Charlesworth B., and D. Charlesworth, 2017 Population genetics from 1966 to 2016. Heredity 118: 2–9. https://doi.org/10.1038/hdy.2016.55

Chilton N. B., I. Beveridge, and R. H. Andrews, 1992 Detection by allozyme electrophoresis of cryptic species of Hypodontus macropi (Nematoda: Strongyloidea) from macropodid marsupials. International Journal for Parasitology 22: 271–279. https://doi.org/10.1016/S0020-7519(05)80004-9

Coyne J. A., and A. H. Orr, 2004 *Speciation*. Sunderland MA, massachusetts U.S.A.

Cruickshank T. E., and M. W. Hahn, 2014 Reanalysis suggests that genomic islands of speciation are due to reduced diversity, not reduced gene flow. Mol Ecol 23: 3133–3157. https://doi.org/10.1111/mec.12796

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

Dasmahapatra K. K., J. R. Walters, A. D. Briscoe, J. W. Davey, A. Whibley, *et al.*, 2012 Butterfly genome reveals promiscuous exchange of mimicry adaptations among species. Nature 487: 94–98. https://doi.org/10.1038/nature11041

Derryberry E. P., N. Seddon, S. Claramunt, J. A. Tobias, A. Baker, *et al.*, 2012 Correlated Evolution of Beak Morphology and Song in the Neotropical Woodcreeper Radiation. Evolution 66: 2784–2797. https://doi.org/10.1111/j.1558-5646.2012.01642.x

Devigili A., J. L. Fitzpatrick, C. Gasparini, I. W. Ramnarine, A. Pilastro, *et al.*, 2018 Possible glimpses into early speciation: the effect of ovarian fluid on sperm velocity accords with post-copulatory isolation between two guppy populations. Journal of Evolutionary Biology 31: 66–74. https://doi.org/10.1111/jeb.13194

Dobzhansky T. grigorovitch, 1937 *Genetics and the origin of species*.

Dufour V., M. Cantou, and F. Lecomte, 2009 Identification of sea bass (Dicentrarchus labrax) nursery areas in the north-western Mediterranean Sea. Journal of the Marine Biological Association of the United Kingdom 89: 1367–1374. https://doi.org/10.1017/S0025315409000368

Ellegren H., L. Smeds, R. Burri, P. I. Olason, N. Backström, *et al.*, 2012 The genomic landscape of species divergence in Ficedula flycatchers. Nature 491: 756–760. https://doi.org/10.1038/nature11584

Endler J. A., 1977 *Geographic Variation, Speciation, and Clines*. Princeton University Press.

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

Felsenstein J., 1981 Skepticism Towards Santa Rosalia, or Why Are There so Few Kinds of Animals? Evolution 35: 124–138. https://doi.org/10.1111/j.1558-5646.1981.tb04864.x

Fishman L., and J. H. Willis, 2001 Evidence for Dobzhansky-Muller Incompatibilites Contributing to the Sterility of Hybrids Between Mimulus Guttatus and M. Nasutus. Evolution 55: 1932–1942. https://doi.org/10.1111/j.0014-3820.2001.tb01311.x

Gagnaire P.-A., S. A. Pavey, E. Normandeau, and L. Bernatchez, 2013 The Genetic Architecture of Reproductive Isolation During Speciation-with-Gene-Flow in Lake Whitefish Species Pairs Assessed by Rad Sequencing. Evolution 67: 2483–2497. https://doi.org/10.1111/evo.12075

Galtier N., B. Nabholz, S. Glémin, and G. D. D. Hurst, 2009 Mitochondrial DNA as a marker of molecular diversity: a reappraisal. Molecular Ecology 18: 4541–4550. https://doi.org/10.1111/j.1365-294X.2009.04380.x

Gante H. F., M. Matschiner, M. Malmstrøm, K. S. Jakobsen, S. Jentoft, *et al.*, 2016 Genomics of speciation and introgression in Princess cichlid fishes from Lake Tanganyika. Molecular Ecology 25: 6143–6161. https://doi.org/10.1111/mec.13767

Gavrilets S., 2004 *Fitness landscapes and the origin of species (MPB-41)*. Princeton University Press.

Geza E., J. Mugo, N. J. Mulder, A. Wonkam, E. R. Chimusa, *et al.*, 2018 A comprehensive survey of models for dissecting local ancestry deconvolution in human genome. Brief Bioinform. https://doi.org/10.1093/bib/bby044

Gravel S., 2012 Population Genetics Models of Local Ancestry. Genetics 191: 607–619. https://doi.org/10.1534/genetics.112.139808

Gross B. L., and L. H. Rieseberg, 2005 The Ecological Genetics of Homoploid Hybrid Speciation. J Hered 96: 241–252. https://doi.org/10.1093/jhered/esi026

Guinand B., M. Vandeputte, M. Dupont-Nivet, A. Vergnet, P. Haffray, *et al.*, 2017 Metapopulation patterns of additive and nonadditive genetic variance in the sea bass (Dicentrarchus labrax). Ecol Evol 7: 2777–2790. https://doi.org/10.1002/ece3.2832

Gutenkunst R. N., R. D. Hernandez, S. H. Williamson, and C. D. Bustamante, 2009 Inferring the Joint Demographic History of Multiple Populations from Multidimensional SNP Frequency Data. PLOS Genetics 5: e1000695. https://doi.org/10.1371/journal.pgen.1000695

Haldane J. BS., 1922 Sex ratio and unisexual sterility in hybrid animals. Journal of genetics 12: 101–109.

Harr B., 2006 Genomic islands of differentiation between house mouse subspecies. Genome Res. 16: 730–737. https://doi.org/10.1101/gr.5045006

Harris K., and R. Nielsen, 2013 Inferring Demographic History from a Spectrum of Shared Haplotype Lengths. PLOS Genet 9. https://doi.org/10.1371/journal.pgen.1003521

Harrison R. G., 1985 Barriers to Gene Exchange Between Closely Related Cricket Species. Ii. Life Cycle Variation and Temporal Isolation. Evolution 39: 244–259. https://doi.org/10.1111/j.1558-5646.1985.tb05664.x

Harrison R. G., and E. L. Larson, 2016 Heterogeneous genome divergence, differential introgression, and the origin and structure of hybrid zones. Mol Ecol 25: 2454–2466. https://doi.org/10.1111/mec.13582

Hermansen J. S., S. A. Sæther, T. O. Elgvin, T. Borge, E. Hjelle, *et al.*, 2011 Hybrid speciation in sparrows I: phenotypic intermediacy, genetic admixture and barriers to gene flow. Molecular Ecology 20: 3812–3822. https://doi.org/10.1111/j.1365-294X.2011.05183.x

Hewitt G. M., 1988 Hybrid zones-natural laboratories for evolutionary studies. Trends in Ecology & Evolution 3: 158–167. https://doi.org/10.1016/0169-5347(88)90033-X

Hewitt G. M., 1996 Some genetic consequences of ice ages, and their role in divergence and speciation. Biological Journal of the Linnean Society 58: 247–276. https://doi.org/10.1111/j.1095-8312.1996.tb01434.x

Hewitt G., 2000 The genetic legacy of the Quaternary ice ages. Nature 405: 907–913. https://doi.org/10.1038/35016000

Hill W. G., and A. Robertson, 1968 Linkage disequilibrium in finite populations. Theoret. Appl. Genetics 38: 226–231. https://doi.org/10.1007/BF01245622

Huerta-Sánchez E., X. Jin, Asan, Z. Bianba, B. M. Peter, *et al.*, 2014 Altitude adaptation in Tibetans caused by introgression of Denisovan-like DNA. Nature 512: 194–197. https://doi.org/10.1038/nature13408

Huxley J., 1942 *Evolution the modern synthesis.* George Allen and Unwin.

Hvala J. A., M. E. Frayer, and B. A. Payseur, 2018 Signatures of hybridization and speciation in genomic patterns of ancestry*. Evolution 72: 1540–1552. https://doi.org/10.1111/evo.13509

Janzen T., A. W. Nolte, and A. Traulsen, 2018 The breakdown of genomic ancestry blocks in hybrid lineages given a finite number of recombination sites. Evolution 72: 735–750. https://doi.org/10.1111/evo.13436

Jiggins C. D., I. Emelianov, and J. Mallet, 2005 Assortative mating and speciation as pleiotropic effects of ecological adaptation: examples in moths and butterflies. SYMPOSIUM-ROYAL ENTOMOLOGICAL SOCIETY OF LONDON 22: 451.

Jones M. R., L. S. Mills, P. C. Alves, C. M. Callahan, J. M. Alves, *et al.*, 2018 Adaptive introgression underlies polymorphic seasonal camouflage in snowshoe hares. Science 360: 1355–1358. https://doi.org/10.1126/science.aar5273

Kaneshiro K. Y., and C. R. B. Boake, 1987 Sexual selection and speciation: Issues raised by Hawaiian Drosophila. Trends in Ecology & Evolution 2: 207–212. https://doi.org/10.1016/0169-5347(87)90022-X

Kelley D. F., 1988 The importance of estuaries for sea-bass, Dicentrarchus labrax (L.). Journal of Fish Biology 33: 25–33. https://doi.org/10.1111/j.1095-8649.1988.tb05555.x

Kern A. D., and M. W. Hahn, 2018 The Neutral Theory in Light of Natural Selection. Mol Biol Evol. https://doi.org/10.1093/molbev/msy092

Kim B. Y., C. D. Huber, and K. E. Lohmueller, 2018 Deleterious variation shapes the genomic landscape of introgression. PLOS Genetics 14: e1007741. https://doi.org/10.1371/journal.pgen.1007741

Kimura M., 1968 Evolutionary rate at the molecular level. Nature 217: 624–626.

Kimura M., and T. Ohta, 1971 On the rate of molecular evolution. J Mol Evol 1: 1–17. https://doi.org/10.1007/BF01659390

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

Kimura M., 1985 The role of compensatory neutral mutations in molecular evolution. J. Genet. 64: 7. https://doi.org/10.1007/BF02923549

Knowlton Nancy, and Weigt Lee A., 1998 New dates and new rates for divergence across the Isthmus of Panama. Proceedings of the Royal Society of London. Series B: Biological Sciences 265: 2257–2263. https://doi.org/10.1098/rspb.1998.0568

Kruuk L. E. B., S. J. E. Baird, K. S. Gale, and N. H. Barton, 1999 A Comparison of Multilocus Clines Maintained by Environmental Adaptation or by Selection Against Hybrids. Genetics 153: 1959–1971.

Larson E. L., T. A. White, C. L. Ross, and R. G. Harrison, 2014 Gene flow and the maintenance of species boundaries. Molecular Ecology 23: 1668–1678. https://doi.org/10.1111/mec.12601

Lemaire C., J.-J. Versini, and F. Bonhomme, 2005 Maintenance of genetic differentiation across a transition zone in the sea: discordance between nuclear and cytoplasmic markers. Journal of Evolutionary Biology 18: 70–80. https://doi.org/10.1111/j.1420-9101.2004.00828.x

Lemmon A. R., and M. Kirkpatrick, 2006 Reinforcement and the Genetics of Hybrid Incompatibilities. Genetics 173: 1145–1155. https://doi.org/10.1534/genetics.105.048199

Lenormand T., 2002 Gene flow and the limits to natural selection. Trends in Ecology & Evolution 17: 183–189. https://doi.org/10.1016/S0169-5347(02)02497-7

Lewontin R. C., and J. L. Hubby, 1966 A Molecular Approach to the Study of Genic Heterozygosity in Natural Populations. II. Amount of Variation and Degree of Heterozygosity in Natural Populations of DROSOPHILA PSEUDOOBSCURA. Genetics 54: 595–609.

Li Y.-C., A. B. Korol, T. Fahima, A. Beiles, and E. Nevo, 2002 Microsatellites: genomic distribution, putative functions and mutational mechanisms: a review. Molecular Ecology 11: 2453–2465. https://doi.org/10.1046/j.1365-294X.2002.01643.x

Liang M., and R. Nielsen, 2014 The Lengths of Admixture Tracts. Genetics 197: 953–967. https://doi.org/10.1534/genetics.114.162362

Lippman Z. B., and D. Zamir, 2007 Heterosis: revisiting the magic. Trends Genet. 23: 60–66. https://doi.org/10.1016/j.tig.2006.12.006

Mallet J., 2005 Hybridization as an invasion of the genome. Trends in Ecology & Evolution 20: 229–237. https://doi.org/10.1016/j.tree.2005.02.010

Mallet J., 2007 Hybrid speciation. Nature 446: 279–283. https://doi.org/10.1038/nature05706

Marques D. A., J. I. Meier, and O. Seehausen, 2019 A Combinatorial View on Speciation and Adaptive Radiation. Trends in Ecology & Evolution. https://doi.org/10.1016/j.tree.2019.02.008

Martin C. H., J. S. Cutler, J. P. Friel, C. D. Touokong, G. Coop, *et al.*, 2015 Complex histories of repeated gene flow in Cameroon crater lake cichlids cast doubt on one of the clearest examples of sympatric speciation. Evolution 69: 1406–1422. https://doi.org/10.1111/evo.12674

Martin S. H., and C. D. Jiggins, 2017 Interpreting the genomic landscape of introgression. Current Opinion in Genetics & Development 47: 69–74. https://doi.org/10.1016/j.gde.2017.08.007

Matsubayashi K. W., I. Ohshima, and P. Nosil, 2010 Ecological speciation in phytophagous insects. Entomologia Experimentalis et Applicata 134: 1–27. https://doi.org/10.1111/j.1570-7458.2009.00916.x

Mavárez J., and M. Linares, 2008 Homoploid hybrid speciation in animals. Molecular Ecology 17: 4181–4185. https://doi.org/10.1111/j.1365-294X.2008.03898.x

Maynard Smith J. M., and J. Haigh, 1974 The hitch-hiking effect of a favourable gene. Genetics Research 23: 23–35. https://doi.org/10.1017/S0016672300014634

Mayr E., 1942 *Systematics and the origin of species, from the viewpoint of a zoologist.* Harvard University Press.

McDonald J. H., and M. Kreitman, 1991 Adaptative protein evolution at the Adh locus in Drosophila. Nature 351: 652.

Morin P. A., G. Luikart, R. K. Wayne, and the SNP workshop group, 2004 SNPs in ecology, evolution and conservation. Trends in Ecology & Evolution 19: 208–216. https://doi.org/10.1016/j.tree.2004.01.009

Muller H., 1942 Isolating mechanisms, evolution, and temperature. Biol. Symp. 6: 71–125.

Nachman M. W., and B. A. Payseur, 2012 Recombination rate variation and speciation: theoretical predictions and empirical results from rabbits and mice. Phil. Trans. R. Soc. B 367: 409–421. https://doi.org/10.1098/rstb.2011.0249

Naciri M., C. Lemaire, P. Borsa, and F. Bonhomme, 1999 Genetic study of the Atlantic/Mediterranean transition in sea bass (Dicentrarchus labrax). J Hered 90: 591–596. https://doi.org/10.1093/jhered/90.6.591

Nadeau N. J., A. Whibley, R. T. Jones, J. W. Davey, K. K. Dasmahapatra, *et al.*, 2012 Genomic islands of divergence in hybridizing Heliconius butterflies identified by large-scale targeted sequencing. Phil. Trans. R. Soc. B 367: 343–353. https://doi.org/10.1098/rstb.2011.0198

Nei M., and T. Gojobori, 1986 Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol Biol Evol 3: 418–426. https://doi.org/10.1093/oxfordjournals.molbev.a040410

Nei M., 1987 *Molecular evolutionary genetics*. Columbia university press.

Ni X., K. Yuan, C. Liu, Q. Feng, L. Tian, *et al.*, 2018 MultiWaver 2.0 : modeling discrete and continuous gene flow to reconstruct complex population admixtures. European Journal of Human Genetics 1. https://doi.org/10.1038/s41431-018-0259-3

Nielsen R., and J. Wakeley, 2001 Distinguishing Migration From Isolation: A Markov Chain Monte Carlo Approach. Genetics 158: 885–896.

Nordborg M., and S. Tavaré, 2002 Linkage disequilibrium: what history has to tell us. Trends in Genetics 18: 83–90. https://doi.org/10.1016/S0168-9525(02)02557-X

Nosil P., T. H. Vines, and D. J. Funk, 2005 Reproductive isolation caused by natural selection against immigrants from divergent habitats. Evolution 59: 705–719.

Nosil P., D. J. Funk, and D. Ortiz-Barrientos, 2009 Divergent selection and heterogeneous genomic divergence. Molecular Ecology 18: 375–402. https://doi.org/10.1111/j.1365-294X.2008.03946.x

Ohta T., and M. Kimura, 1970 Development of associative overdominance through linkage disequilibrium in finite populations*. Genetics Research 16: 165–177. https://doi.org/10.1017/S0016672300002391

Orr H. A., 1997 Haldane's Rule. Annual Review of Ecology and Systematics 28: 195–218. https://doi.org/10.1146/annurev.ecolsys.28.1.195

Palamara P. F., T. Lencz, A. Darvasi, and I. Pe'er, 2012 Length Distributions of Identity by Descent Reveal Fine-Scale Demographic History. Am J Hum Genet 91: 809–822. https://doi.org/10.1016/j.ajhg.2012.08.030

Palumbi S. R., 1992 Marine speciation on a small planet. Trends in Ecology & Evolution 7: 114–118. https://doi.org/10.1016/0169-5347(92)90144-Z

Palumbi S. R., 2009 Speciation and the evolution of gamete recognition genes: pattern and process. Heredity 102: 66–76. https://doi.org/10.1038/hdy.2008.104

Patarnello T., F. a. M. J. Volckaert, and R. Castilho, 2007 Pillars of Hercules: is the Atlantic–Mediterranean transition a phylogeographical break? Molecular Ecology 16: 4426–4444. https://doi.org/10.1111/j.1365-294X.2007.03477.x

Pawson M. G., G. D. Pickett, and P. R. Witthames, 2000 The influence of temperature on the onset of first maturity in sea bass. Journal of Fish Biology 56: 319–327. https://doi.org/10.1111/j.1095-8649.2000.tb02109.x

Payseur B. A., 2010 Using differential introgression in hybrid zones to identify genomic regions involved in speciation. Molecular Ecology Resources 10: 806–820. https://doi.org/10.1111/j.1755-0998.2010.02883.x

Payseur B. A., and L. H. Rieseberg, 2016 A genomic perspective on hybridization and speciation. Mol Ecol 25: 2337–2360. https://doi.org/10.1111/mec.13557

Pickett G. D., and M. G. Pawson, 1994 *Sea Bass: Biology*. Springer Science & Business Media.

Pickrell J. K., and J. K. Pritchard, 2012 Inference of Population Splits and Mixtures from Genome-Wide Allele Frequency Data. PLoS Genet 8. https://doi.org/10.1371/journal.pgen.1002967

Pool J. E., and R. Nielsen, 2009 Inference of Historical Changes in Migration Rate From the Lengths of Migrant Tracts. Genetics 181: 711–719. https://doi.org/10.1534/genetics.108.098095

Presgraves D. C., 2010 The molecular evolutionary basis of species formation. Nature Reviews Genetics 11: 175–180. https://doi.org/10.1038/nrg2718

Pugach I., R. Matveev, V. Spitsyn, S. Makarov, I. Novgorodov, *et al.*, 2016 The Complex Admixture History and Recent Southern Origins of Siberian Populations. Mol Biol Evol 33: 1777–1795. https://doi.org/10.1093/molbev/msw055

Racimo F., S. Sankararaman, R. Nielsen, and E. Huerta-Sánchez, 2015 Evidence for archaic adaptive introgression in humans. Nat Rev Genet 16: 359–371. https://doi.org/10.1038/nrg3936

Racimo F., D. Marnetto, and E. Huerta-Sánchez, 2017 Signatures of Archaic Adaptive Introgression in Present-Day Human Populations. Mol Biol Evol 34: 296–317. https://doi.org/10.1093/molbev/msw216

Ravinet M., R. Faria, R. K. Butlin, J. Galindo, N. Bierne, *et al.*, 2017 Interpreting the genomic landscape of speciation: a road map for finding barriers to gene flow. J. Evol. Biol. 30: 1450–1477. https://doi.org/10.1111/jeb.13047

Rees H. C., B. C. Maddison, D. J. Middleditch, J. R. M. Patmore, and K. C. Gough, 2014 REVIEW: The detection of aquatic animal species using environmental DNA – a review of eDNA as a survey tool in ecology. Journal of Applied Ecology 51: 1450–1459. https://doi.org/10.1111/1365-2664.12306

Rhee J.-K., H. Li, J.-G. Joung, K.-B. Hwang, B.-T. Zhang, *et al.*, 2016 Survey of computational haplotype determination methods for single individual. Genes Genom 38: 1–12. https://doi.org/10.1007/s13258-015-0342-x

Richards E. J., J. W. Poelstra, and C. H. Martin, 2018 Don't throw out the sympatric speciation with the crater lake water: fine-scale investigation of introgression provides equivocal support for causal role of secondary gene flow in one of the clearest examples of sympatric speciation. Evolution Letters 2: 524–540. https://doi.org/10.1002/evl3.78

Rieseberg L. H., M. A. Archer, and R. K. Wayne, 1999 Transgressive segregation, adaptation and speciation. Heredity 83: 363–372. https://doi.org/10.1046/j.1365-2540.1999.00617.x

Roderick G. K., and R. G. Gillespie, 1998 Speciation and phylogeography of Hawaiian terrestrial arthropods. Molecular Ecology 7: 519–531. https://doi.org/10.1046/j.1365-294x.1998.00309.x

Rougemont Q., and L. Bernatchez, 2018 The demographic history of Atlantic salmon (Salmo salar) across its distribution range reconstructed from approximate Bayesian computations*. Evolution 72: 1261–1277. https://doi.org/10.1111/evo.13486

Rougeux C., L. Bernatchez, and P.-A. Gagnaire, 2017 Modeling the Multiple Facets of Speciation-with-Gene-Flow toward Inferring the Divergence History of Lake Whitefish Species Pairs (Coregonus clupeaformis). Genome Biol Evol 9: 2057–2074. https://doi.org/10.1093/gbe/evx150

Roux C., G. Tsagkogeorga, N. Bierne, and N. Galtier, 2013 Crossing the Species Barrier: Genomic Hotspots of Introgression between Two Highly Divergent Ciona intestinalis Species. Mol Biol Evol 30: 1574–1587. https://doi.org/10.1093/molbev/mst066

Roux C., C. Fraïsse, V. Castric, X. Vekemans, G. H. Pogson, *et al.*, 2014 Can we continue to neglect genomic variation in introgression rates when inferring the history of speciation? A case study in a *Mytilus* hybrid zone. J. Evol. Biol. 27: 1662–1675. https://doi.org/10.1111/jeb.12425

Roux C., C. Fraïsse, J. Romiguier, Y. Anciaux, N. Galtier, *et al.*, 2016 Shedding Light on the Grey Zone of Speciation along a Continuum of Genomic Divergence. PLOS Biology 14. https://doi.org/10.1371/journal.pbio.2000234

Salazar C., S. W. Baxter, C. Pardo-Diaz, G. Wu, A. Surridge, *et al.*, 2010 Genetic Evidence for Hybrid Trait Speciation in Heliconius Butterflies. PLOS Genetics 6: e1000930. https://doi.org/10.1371/journal.pgen.1000930

Sankararaman S., S. Mallick, M. Dannemann, K. Prüfer, J. Kelso, *et al.*, 2014 The genomic landscape of Neanderthal ancestry in present-day humans. Nature 507: 354–357. https://doi.org/10.1038/nature12961

Sankararaman S., S. Mallick, N. Patterson, and D. Reich, 2016 The Combined Landscape of Denisovan and Neanderthal Ancestry in Present-Day Humans. Current Biology 26: 1241–1247. https://doi.org/10.1016/j.cub.2016.03.037

Schierup M. H., and J. Hein, 2000 Consequences of Recombination on Traditional Phylogenetic Analysis. Genetics 156: 879–891.

Schumer M., G. G. Rosenthal, and P. Andolfatto, 2014 How Common Is Homoploid Hybrid Speciation? Evolution 68: 1553–1560. https://doi.org/10.1111/evo.12399

Schumer M., R. Cui, G. G. Rosenthal, and P. Andolfatto, 2015 Reproductive Isolation of Hybrid Populations Driven by Genetic Incompatibilities. PLOS Genetics 11: e1005041. https://doi.org/10.1371/journal.pgen.1005041

Schumer M., C. Xu, D. L. Powell, A. Durvasula, L. Skov, *et al.*, 2018 Natural selection interacts with recombination to shape the evolution of hybrid genomes. Science eaar3684. https://doi.org/10.1126/science.aar3684

Seehausen O., 2013 Conditions when hybridization might predispose populations for adaptive radiation. Journal of Evolutionary Biology 26: 279–281. https://doi.org/10.1111/jeb.12026

Servedio M. R., G. S. V. Doorn, M. Kopp, A. M. Frame, and P. Nosil, 2011 Magic traits in speciation: 'magic' but not rare? Trends in Ecology & Evolution 26: 389–397. https://doi.org/10.1016/j.tree.2011.04.005

Smadja C., and G. Ganem, 2002 Subspecies recognition in the house mouse: a study of two populations from the border of a hybrid zone. Behav Ecol 13: 312–320. https://doi.org/10.1093/beheco/13.3.312

Smadja C., J. Catalan, and G. Ganem, 2004 Strong premating divergence in a unimodal hybrid zone between two subspecies of the house mouse. Journal of Evolutionary Biology 17: 165–176. https://doi.org/10.1046/j.1420-9101.2003.00647.x

Smadja C., and G. Ganem, 2005 Asymmetrical reproductive character displacement in the house mouse. Journal of Evolutionary Biology 18: 1485–1493. https://doi.org/10.1111/j.1420-9101.2005.00944.x

Smadja C., J. Galindo, and R. Butlin, 2008 Hitching a lift on the road to speciation. Molecular Ecology 17: 4177–4180. https://doi.org/10.1111/j.1365-294X.2008.03917.x

Smadja C. M., E. Loire, P. Caminade, M. Thoma, Y. Latour, *et al.*, 2015 Seeking signatures of reinforcement at the genetic level: a hitchhiking mapping and candidate gene approach in the house mouse. Molecular Ecology 24: 4222–4237. https://doi.org/10.1111/mec.13301

Smith N. G. C., and A. Eyre-Walker, 2002 Adaptive protein evolution in Drosophila. Nature 415: 1022. https://doi.org/10.1038/4151022a

Snyder C. W., 2016 Evolution of global temperature over the past two million years. Nature 538: 226–228. https://doi.org/10.1038/nature19798

Souche E., 2009 Genomic variation in European Sea bass: from SNP discovery within ESTs to genome scan

Sousa V. C., M. Carneiro, N. Ferrand, and J. Hey, 2013 Identifying Loci Under Selection Against Gene Flow in Isolation-with-Migration Models. Genetics 194: 211–233. https://doi.org/10.1534/genetics.113.149211

Sousa V., and J. Hey, 2013 Understanding the origin of species with genome-scale data: modelling gene flow. Nature Reviews Genetics 14: 404–414. https://doi.org/10.1038/nrg3446

Stapley J., P. G. D. Feulner, S. E. Johnston, A. W. Santure, and C. M. Smadja, 2017 Recombination: the good, the bad and the variable. Phil. Trans. R. Soc. B 372: 20170279. https://doi.org/10.1098/rstb.2017.0279

Strasburg J. L., and L. H. Rieseberg, 2010 How Robust Are "Isolation with Migration" Analyses to Violations of the IM Model? A Simulation Study. Mol Biol Evol 27: 297–310. https://doi.org/10.1093/molbev/msp233

Sunnucks P., 2000 Efficient genetic markers for population biology. Trends in Ecology & Evolution 15: 199–203. https://doi.org/10.1016/S0169-5347(00)01825-5

Taberlet P., E. Coissac, M. Hajibabaei, and L. H. Rieseberg, 2012 Environmental DNA. Molecular Ecology 21: 1789–1793. https://doi.org/10.1111/j.1365-294X.2012.05542.x

Tang H., M. Coram, P. Wang, X. Zhu, and N. Risch, 2006 Reconstructing Genetic Ancestry Blocks in Admixed Individuals. The American Journal of Human Genetics 79: 1–12. https://doi.org/10.1086/504302

Therkildsen N. O., and S. R. Palumbi, 2017 Practical low-coverage genomewide sequencing of hundreds of individually barcoded samples for population and evolutionary genomics in nonmodel species. Molecular Ecology Resources 17: 194–208. https://doi.org/10.1111/1755-0998.12593

Tine M., H. Kuhl, P.-A. Gagnaire, B. Louro, E. Desmarais, *et al.*, 2014 European sea bass genome and its variation provide insights into adaptation to euryhalinity and speciation. Nature Communications 5: 5770.

Turner T. L., M. W. Hahn, and S. V. Nuzhdin, 2005 Genomic Islands of Speciation in Anopheles gambiae. PLOS Biol 3. https://doi.org/10.1371/journal.pbio.0030285

Via S., and J. West, 2008 The genetic mosaic suggests a new role for hitchhiking in ecological speciation. Molecular Ecology 17: 4334–4345. https://doi.org/10.1111/j.1365-294X.2008.03921.x

Via S., 2009 Natural selection in action during speciation. PNAS 106: 9939–9946. https://doi.org/10.1073/pnas.0901397106

Vignal A., D. Milan, M. SanCristobal, and A. Eggen, 2002 A review on SNP and other types of molecular markers and their use in animal genetics. Genetics Selection Evolution 34: 275. https://doi.org/10.1186/1297-9686-34-3-275

Wolf J. B. W., and H. Ellegren, 2016 Making sense of genomic islands of differentiation in light of speciation. Nat Rev Genet 18: 87–100. https://doi.org/10.1038/nrg.2016.133

Wood T. E., N. Takebayashi, M. S. Barker, I. Mayrose, P. B. Greenspoon, *et al.*, 2009 The frequency of polyploid speciation in vascular plants. PNAS 106: 13875–13879. https://doi.org/10.1073/pnas.0811575106

Wright S., 1949 The Genetical Structure of Populations. Annals of Eugenics 15: 323–354. https://doi.org/10.1111/j.1469-1809.1949.tb02451.x

Yeaman S., S. Aeschbacher, and R. Bürger, 2016 The evolution of genomic islands by increased establishment probability of linked alleles. Mol Ecol 25: 2542–2558. https://doi.org/10.1111/mec.13611

Yuan K., Y. Zhou, X. Ni, Y. Wang, C. Liu, *et al.*, 2017 Models, methods and tools for ancestry inference and admixture analysis. Quant Biol 5: 236–250. https://doi.org/10.1007/s40484-017-0117-2

Zhang D.-X., and G. M. Hewitt, 2003 Nuclear DNA analyses in genetic studies of populations: practice, problems and prospects. Molecular Ecology 12: 563–584. https://doi.org/10.1046/j.1365-294X.2003.01773.x

CHAPITRE 1 :

Inférence de l'histoire de la divergence et identification des mécanismes à l'origine des îlots génomiques de différenciation

Les résultats de ce chapitre sont exposés en détails dans l'article joint publié dans *Nature communications*.

Ce chapitre a deux objectifs principaux. Le premier est d'affiner notre connaissance de l'histoire de la divergence entre le bar et le loup et le deuxième est d'identifier les mécanismes évolutifs ayant permis la mise en place des îlots génomiques de différenciation observés entre les deux lignées (Tine *et al.* 2014). Pour ce faire, nous avons décidé d'utiliser une couche d'information supplémentaire encore jamais utilisée chez cette espèce non modèle, celle de la liaison génétique. Seize individus sauvages dont quatre mâles atlantiques, huit femelles ouest-méditerranéennes et quatre mâles est-méditerranéens ont été croisés pour générer huit familles. Les génomes des parents de chaque famille ainsi que celui d'un descendant choisi aléatoirement ont été entièrement séquencés, ce qui nous a permis de reconstruire leur phase (i.e. reconstituer les associations entre allèles transmis par chacun des deux parents) grâce aux règles de transmission mendéliennes (McKenna *et al.* 2010). Nous avons ensuite utilisé une méthode d'inférence d'ascendance locale (Lawson *et al.* 2012) afin d'identifier les fragments chromosomiques atlantiques introgressés en Méditerranée et inversement (Figure 1).

Inférences démographiques

L'existence d'un contact secondaire entre la lignée atlantique et méditerranéenne avait déjà été mise en évidence, plaçant le début de la divergence il y a 300 000 ans (Tine *et al.* 2014). Or, les cycles glaciaires du Pléistocène ayant une durée d'environ 100 000 ans (Snyder 2016), nous avons voulu déterminer s'il n'y a pas eu plusieurs périodes d'isolement/contact entre le bar et le loup. Pour ce faire j'ai utilisé la méthode d'inférence démographique présentée précédemment qui se base sur la distribution des fragments IBS (Harris and Nielsen 2013). J'ai ainsi développé un modèle d'histoire démographique flexible, permettant de modéliser avec un nombre réduit de paramètres trois scénarios de divergence : une période de divergence avec flux génique continu, un contact secondaire ou des contacts secondaires répétés (Figure 3). Contrairement à notre hypothèse de travail, je n'ai pas pu mettre en évidence l'existence d'une connectivité cyclique entre les deux lignées. Cependant, la méthode utilisée ici ne prend pas en compte les variations de taille efficace que peuvent avoir subi les populations pendant les périodes glaciaires (goulots d'étranglement ou expansions), ni la contre-sélection qui agit sur les haplotypes introgressés. Or ces deux phénomènes impactent la forme des distributions de longueur des fragments IBS. De plus, on peut se demander si la trace des évènements de contacts anciens n'est pas effacée par le contact le plus récent, la recombinaison et le temps ayant le même impact sur la taille des segments IBS. Néanmoins, le modèle de contact secondaire est celui présentant la meilleure vraisemblance avec des valeurs de paramètres estimées très proche des précédentes (Tine *et al.* 2014). Nous avons donc confirmé à l'aide de simulations que les distributions de longueur des fragments introgressés étaient très bien reproduites par ce modèle de contact

secondaire, confirmant l'existence d'une phase de divergence allopatrique suivie d'une phase de flux génique entre ces deux lignées (Figure 2). Néanmoins, nos résultats ne permettent pas d'exclure l'existence de contacts antérieur qui aurait pu ne pas être détecté dû au manque de signal dans les données, chaque nouveau contact ayant tendance à effacer la trace du précédent.

Les îlots génomiques de différenciation

Les îlots de différenciation génétique observés entre la lignée atlantique et méditerranéenne de bar européen ont pu être généré par deux mécanismes différents. Premièrement, par l'action de la sélection en liaison qui étant plus forte dans les régions à faible taux de recombinaison, réduit la taille efficace en gonflant artificiellement les valeurs de différenciation. Deuxièmement, par l'existence de locus d'isolement reproductif qui permettent localement à la différenciation de se maintenir en présence de flux génique. En combinant l'utilisation de différentes statistiques permettant de mesurer la différenciation (F_{ST}) et la divergence (d_{XY}) génétique ainsi que les niveaux d'introgression le long du génome (Fraction d'introgression, longueur des fragments introgressés et $RND_{min} = d_{XY\text{-}min} / d_{XY\text{-}outgroup}$), nous avons pu distinguer les îlots générés par ces deux mécanismes. Ceux caractérisés par de fortes valeurs de différenciation et de divergence ainsi que de faibles niveaux d'introgression, qui sont donc capables de se maintenir en présence de flux génique et sont donc très probablement impliqués dans l'isolement reproductif entre les deux lignées de *D. labrax*. Ceux présentant de forts niveaux d'introgression, qui ont dû être générés par les effets de la sélection en liaison et sont encore visible aujourd'hui car dans une phase transitoire d'un processus dynamique d'érosion qui a débuté mais n'est pas terminé (Figure 4).

Cependant, les îlots étant tous majoritairement présents dans les régions à faible taux de recombinaison, nous avons voulu confirmer que la sélection en liaison seule ne pouvait pas générer les patrons observés. A l'aide de simulations, nous avons montré que les valeurs des statistiques observées pour les îlots supposés impliqués dans l'isolement reproductif ne pouvaient être obtenues qu'en présence de sélection en liaison et d'isolement reproductif (Figure 6). Il semblerait donc que les deux mécanismes aient agit chez cette espèce et que ce soit leur combinaison qui ait permis la formation de certains des îlots observés. En effet, pendant les périodes d'allopatrie, la sélection en liaison réduit la taille efficace des régions à faible recombinaison par rapport au reste du génome. Le polymorphisme intra-populationnel y est donc plus faible, ce qui a pour effet d'augmenter le F_{ST}. Des allèles d'isolement reproductif peuvent également évoluer dans ces régions suite à la fixation de mutations faiblement délétères (impliquées dans des DMI) ou avantageuses (adaptation local), qui vont plus facilement résister à l'introgression en cumulant leurs effets dans les régions à faible recombinaison. A la fin de la période d'allopatrie, la différenciation est donc plus forte dans les régions à faible taux de recombinaison, qu'elles contiennent ou non des locus impliqués dans l'isolement

reproductif. Pendant la période de flux génique, la différenciation est ensuite progressivement érodée dans les régions qui ne contiennent pas de locus impliqués dans l'isolement reproductif, accentuant l'hétérogénéité du paysage de différenciation. La différenciation étant plus élevées au départ dans les régions à faible recombinaison, leur érosion prend plus de temps, c'est pourquoi certains îlots sont encore détectables aujourd'hui. Les îlots génomiques de différenciation observés chez *D. labrax* sont donc générés par deux mécanismes qui agissent à différents moments du processus de divergence mais peuvent combiner leurs effets.

Nos résultats nous permettent donc de distinguer les îlots impliqués dans l'isolement reproductif des autres, mais nous ne savons toujours pas comment ont évolué les allèles responsables de l'isolement reproductif. En effet, plusieurs mécanismes en lien avec des facteurs endogènes ou exogènes peuvent permettre à l'isolement reproductif de se mettre en place. C'est pourquoi l'étude des signatures d'évolution moléculaire des gènes aux cœurs des îlots impliqués dans l'isolement reproductif reste une étape nécessaire pour comprendre quels mécanismes empêchent les échanges génétiques. De plus, nous avons constaté que les îlots impliqués dans l'isolement productif présentaient des valeurs de divergence particulièrement élevé. Plusieurs mécanismes pouvant générer ces patrons particuliers, il nous faut dans un premier temps les distinguer afin d'identifier l'origine des allèles impliqués dans l'isolement reproductif et pouvoir étudier leurs patrons d'évolution moléculaire.

Références

Duranton M., F. Allal, C. Fraïsse, N. Bierne, F. Bonhomme, *et al.*, 2018 The origin and remolding of genomic islands of differentiation in the European sea bass. Nature Communications 9: 2518. https://doi.org/10.1038/s41467-018-04963-6

Harris K., and R. Nielsen, 2013 Inferring Demographic History from a Spectrum of Shared Haplotype Lengths. PLOS Genet 9. https://doi.org/10.1371/journal.pgen.1003521

Lawson D. J., G. Hellenthal, S. Myers, and D. Falush, 2012 Inference of Population Structure using Dense Haplotype Data. PLOS Genet 8. https://doi.org/10.1371/journal.pgen.1002453

McKenna A., M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, *et al.*, 2010 The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 20: 1297–1303. https://doi.org/10.1101/gr.107524.110

Snyder C. W., 2016 Evolution of global temperature over the past two million years. Nature 538: 226–228. https://doi.org/10.1038/nature19798

Tine M., H. Kuhl, P.-A. Gagnaire, B. Louro, E. Desmarais, *et al.*, 2014 European sea bass genome and its variation provide insights into adaptation to euryhalinity and speciation. Nature Communications 5: 5770.

nature COMMUNICATIONS

Corrected: Publisher correction

ARTICLE

DOI: 10.1038/s41467-018-04963-6 OPEN

The origin and remolding of genomic islands of differentiation in the European sea bass

Maud Duranton[1,2], François Allal[2,3], Christelle Fraïsse[1,2], Nicolas Bierne[1,2], François Bonhomme[1,2] & Pierre-Alexandre Gagnaire[1,2]

Speciation is a complex process that leads to the progressive establishment of reproductive isolation barriers between diverging populations. Genome-wide comparisons between closely related species have revealed the existence of heterogeneous divergence patterns, dominated by genomic islands of increased divergence supposed to contain reproductive isolation loci. However, this divergence landscape only provides a static picture of the dynamic process of speciation, during which confounding mechanisms unrelated to speciation can interfere. Here we use haplotype-resolved whole-genome sequences to identify the mechanisms responsible for the formation of genomic islands between Atlantic and Mediterranean sea bass lineages. Local ancestry patterns show that genomic islands first emerged in allopatry through linked selection acting on a heterogeneous recombination landscape. Then, upon secondary contact, preexisting islands were strongly remolded by differential introgression, revealing variable fitness effects among regions involved in reproductive isolation. Interestingly, we find that divergent regions containing ancient polymorphisms conferred the strongest resistance to introgression.

[1] Institut des Sciences de l'Evolution de Montpellier - UMR5554 UM-CNRS-IRD-EPHE, Place Eugène Bataillon, 34095 Montpellier, France. [2] Université de Montpellier, Place Eugène Bataillon, 34095 Montpellier, France. [3] MARBEC, Université de Montpellier, Ifremer-CNRS-IRD-UM, 34250 Palavas-les-Flots, France. Correspondence and requests for materials should be addressed to M.D. (email: maud.duranton@umontpellier.fr)

Understanding how genetic differences accumulate between populations over time to eventually form new species is one of the main objectives in evolutionary biology[1,2]. Speciation is generally thought as a gradual mechanism that proceeds through intermediate stages whereby gene flow is not completely interrupted and genomes remain permeable to genetic exchanges[3–6]. As long as species can still hybridize, studying gene exchange provides access to variety of evolutionary mechanisms involved at different stages of the speciation process[7,8]. Advanced sequencing technologies now provide a genome-wide view of divergence between closely related species, improving our understanding of how speciation unfolds at the molecular level[9]. A growing number of speciation genomics studies have demonstrated the existence of heterogeneous genomic divergence patterns between entities at different stages of speciation[10–20]. However, difficulties to relate empirical divergence patterns to the underlying mechanisms involved in their formation limit the potential of speciation genomics approaches[21,22].

Heterogeneous genome divergence between taxa can have several possible causes that need to be individually assessed for understanding the underlying mechanisms generating regions of increased divergence[23,24], the so-called genomic islands[10,11,20]. Among them, accelerated rates of lineage sorting within populations due to recurrent events of either selective sweeps[25] or background selection (BGS)[26] can generate incidental islands of relative divergence that are not necessarily related to reproductive isolation (RI)[27]. An important objective of speciation research is therefore to identify and understand the origin of genomic islands associated with barrier loci responsible for gene flow reduction between diverging populations[22]. Such islands may be themselves explained by different mechanisms depending on the intensity and timing of gene flow and the genomic architecture of RI[24]. Elucidating the typical conditions under which each mechanism is at play is central to understanding the roles of selection and gene flow in the speciation process.

The identification of genomic regions that are truly resistant to introgression remains a challenging task, especially because the aforementioned mechanisms are influenced by the recombination landscape and therefore tend to affect similar regions of the genome[27–29]. To disentangle the role of these confounding factors, substantial levels of gene flow may be needed to properly reveal the genomic regions involved in RI. Moreover, the analysis of gene flow and selection may be facilitated by the direct detection of introgressed chromosomal segments, combined with an explicit consideration of the demographic history[30]. Here we developed this type of approach in a high gene flow marine species.

The European sea bass (*Dicentrarchus labrax*) provides an interesting model to understand the evolution of genomic islands[31]. The species is subdivided into an Atlantic and a Mediterranean lineage that hybridize in the Alboran sea[32]. Historical demographic inferences revealed that the two lineages have started to diverge in allopatry around 300,000 years before present (BP) and then experienced a post-glacial secondary contact generating varying rates of introgression across the genome[31]. This evolutionary history mirrors the distributional range shifts that occurred across many taxa during glacial periods especially in the Atlantic–Mediterranean region[33], which are recognized as an important source of species diversification[34,35]. Admittedly, however, the divergence history of sea bass lineages may involve a more complex succession of divergence and contact periods potentially paced by the quasi-100,000-year glacial cycles during the Pleistocene[36]. Here we characterize local ancestry patterns using haplotype-resolved whole-genome sequences within a geographic context to (i) infer the divergence history of sea bass lineages from the length spectrum of introgressed tracts and shared haplotypes and (ii) identify the different mechanisms involved in the formation and remolding of genomic islands of differentiation. Our results show that genomic regions experiencing stronger linked selection due to low recombination not only diverge faster during allopatric episodes but also better resist introgression in the presence of gene flow. These findings support that multiple loci affect RI in sea bass, with the most divergent having the strongest effects.

Results

Spatial population structure and admixture. The genetic relationships of the newly sequenced genomes with respect to the range-wide population structure of the European sea bass were evaluated with a Principal Component (PC) Analysis including 112 additional individuals genotyped at 13,094 common single nucleotide polymorphisms (SNPs; Fig. 1a, b and Supplementary Note 2). The main component of genetic variation (axis 1, 43.54% of explained variance) clearly distinguished Atlantic from Mediterranean populations, while the second axis (2.59%) revealed a subtle genetic differentiation between eastern and western Mediterranean basins (E-MED and W-MED, respectively). Genetic admixture was found to occur along the Algerian coast, which is the principal zone where Atlantic alleles enter the Mediterranean sea. The resulting inflow of Atlantic alleles within the Mediterranean generates a longitudinal gradient of introgression illustrated by a shift between W-MED and E-MED samples along the first PC axis (Fig. 1b).

Migrant tracts' identification. Spatial introgression patterns were also detected at the chromosome level using local ancestry inference based on 2,628,725 phased SNPs. The proportion of the genome occupied by migrant tracts was twice higher in the W-MED (31%) compared to the E-MED population (13%), which also displayed shorter migrant tracts (Fig. 1c, d). The longest introgressed haplotype detected in the W-MED (2.03 Mb) was twice as long as the longest one found in the E-MED population (0.98 Mb). More generally, the genome-wide distribution of migrant tract length showed a reduced abundance of tracts over all length classes in the E-MED compared to the W-MED population. This shift is consistent with the action of recombination that progressively erodes recently introgressed tracts as they diffuse by migration from the entrance to the bottom of the Mediterranean sea. Consistently, this effect was not apparent for the shortest migrant tracts (i.e., <50 kb) that probably reside in the Mediterranean sea for a much longer time than the time needed to diffuse from west to east. The Atlantic population was the least introgressed, with <5% of its genome occupied by tracts of Mediterranean ancestry. Migrant tracts were also shorter (maximum length 0.62 Mb) and less abundant over the whole-length spectrum (Fig. 1c, d). This is consistent with a reduced amount of gene flow from the Mediterranean to the Atlantic population[31]. Finally, our method for reconstructing ancestral Mediterranean genomes effectively removed migrant tracts (Supplementary Note 4), leaving very small residual amounts of Atlantic haplotypes in the reconstructed Mediterranean genomes (Fig. 1c), except for genomic regions that were strongly introgressed by Atlantic alleles (Supplementary Figs. 5 and 6).

Analysis of migrant tract length distribution. An 85% fraction of the 100 kb windows located in low-recombining regions of the Mediterranean genomes present introgressed Atlantic tracts that are on average >50 kb (Supplementary Fig. 7). Using a recombination clock, we found that this observation is consistent with introgression occurring during the past 17,000 years. Since the most part of the introgressed tracts length distribution is

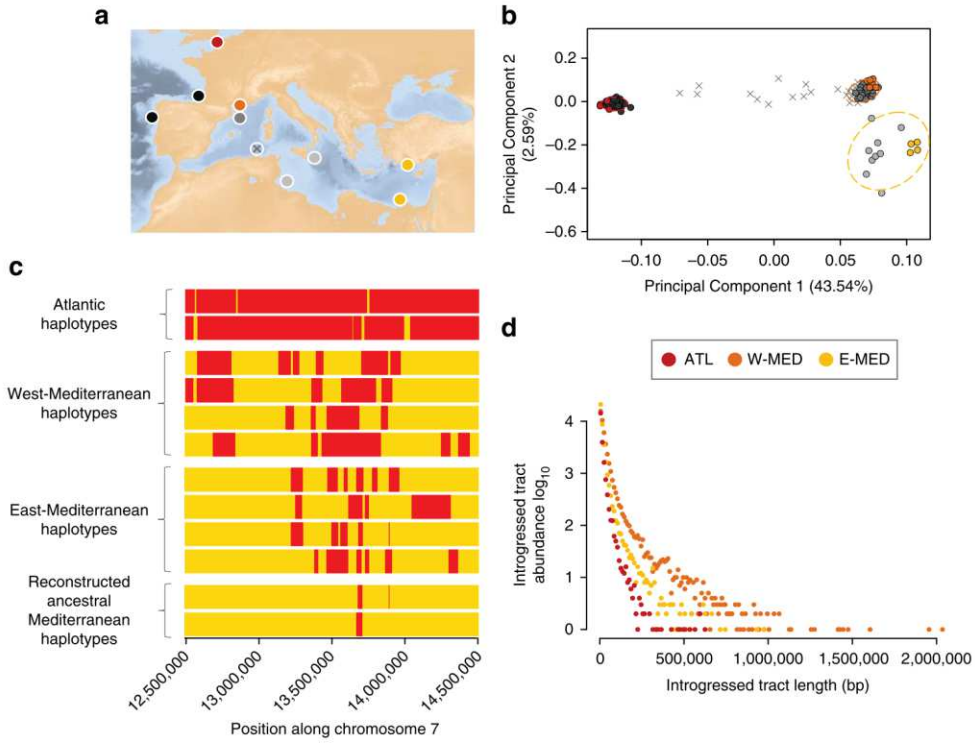


Fig. 1 Spatial population structure and local ancestry patterns. **a** Geographical location of samples, including the newly sequenced genomes (colored circles) and additional reference samples from the Atlantic (dark gray), W-MED (gray), E-MED (light gray), and the Algerian admixture zone (gray crosses). **b** Principal Component Analysis of newly sequenced genomes combined with 112 individuals genotyped at 13,094 common SNPs (MAF > 0.1). The first PCA axis distinguishes Atlantic and Mediterranean populations while the second axis reveals a subtle population structure between W-MED and E-MED. Some individuals from the Algerian coast represent admixed genotypes between Atlantic and Mediterranean populations. **c** Schematic representation of a 2 Mb region within chromosome 7, showing the mosaic of ancestry blocks derived from Atlantic (red) and Mediterranean (yellow) populations. For simplification, we only display two individual haplotypes from Atlantic samples, four from W-MED, four from E-MED, and two reconstructed ancestral Mediterranean haplotypes from E-MED samples. **d** Migrant tract length distribution obtained for the Atlantic (red, showing tracts of Mediterranean origin), W-MED (orange), and E-MED (yellow) populations (showing tracts of Atlantic origin) using four individuals per population and a total of 2,628,725 phased SNPs

consistent with a recent secondary contact, we evaluated the goodness-of-fit of the previously inferred post-glacial secondary contact model[31], which places the onset of gene flow 11,500 years ago. We performed coalescent simulations with variable recombination rates under this model to generate whole-genome data using the recombination structure of sea bass chromosomes. The length distribution of migrant tracts obtained from these simulations reproduced well the observed distribution for both the Atlantic and Mediterranean populations (Fig. 2). Therefore, the secondary contact model inferred from the joint site frequency spectrum without using linkage information[31] has a high predictive power regarding the length distribution of introgressed tracts when accounting for recombination rate variation.

Testing waves of historical gene flow. Initial divergence between sea bass lineages has been dated around 270,000 years BP[31], corresponding to three glacial cycles[36] during which possible genetic interactions may have occurred when interglacial conditions were similar to present. In order to address whether short migrant tracts found within low-recombining regions of the genome could result from waves of historical gene flow, we developed a flexible model of divergence in which the history of admixture can take different forms (Fig. 3a). The modeling scenarios were subdivided into three categories according to the distribution of admixture pulses over time: (i) continuous migration, (ii) secondary contact, and (iii) periodic pulses (Fig. 3b). Our demographic inferences showed an increase in

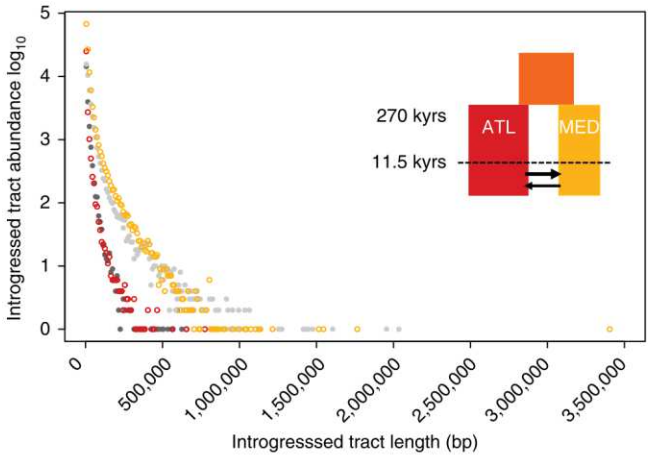


Fig. 2 Observed and simulated migrant tract length distributions in the Atlantic and Mediterranean populations. Observed distributions (gray) of migrant tract length are compared with simulated distributions (colored) under the post-glacial secondary contact scenario illustrated in the top-right corner[31]. The abundance of introgressed tracts as a function of their length is represented for observed vs. simulated data in the Atlantic (dark gray vs. red circles, showing tracts of Mediterranean origin) and Mediterranean populations (light gray vs. yellow circles, showing tracts of Atlantic origin)

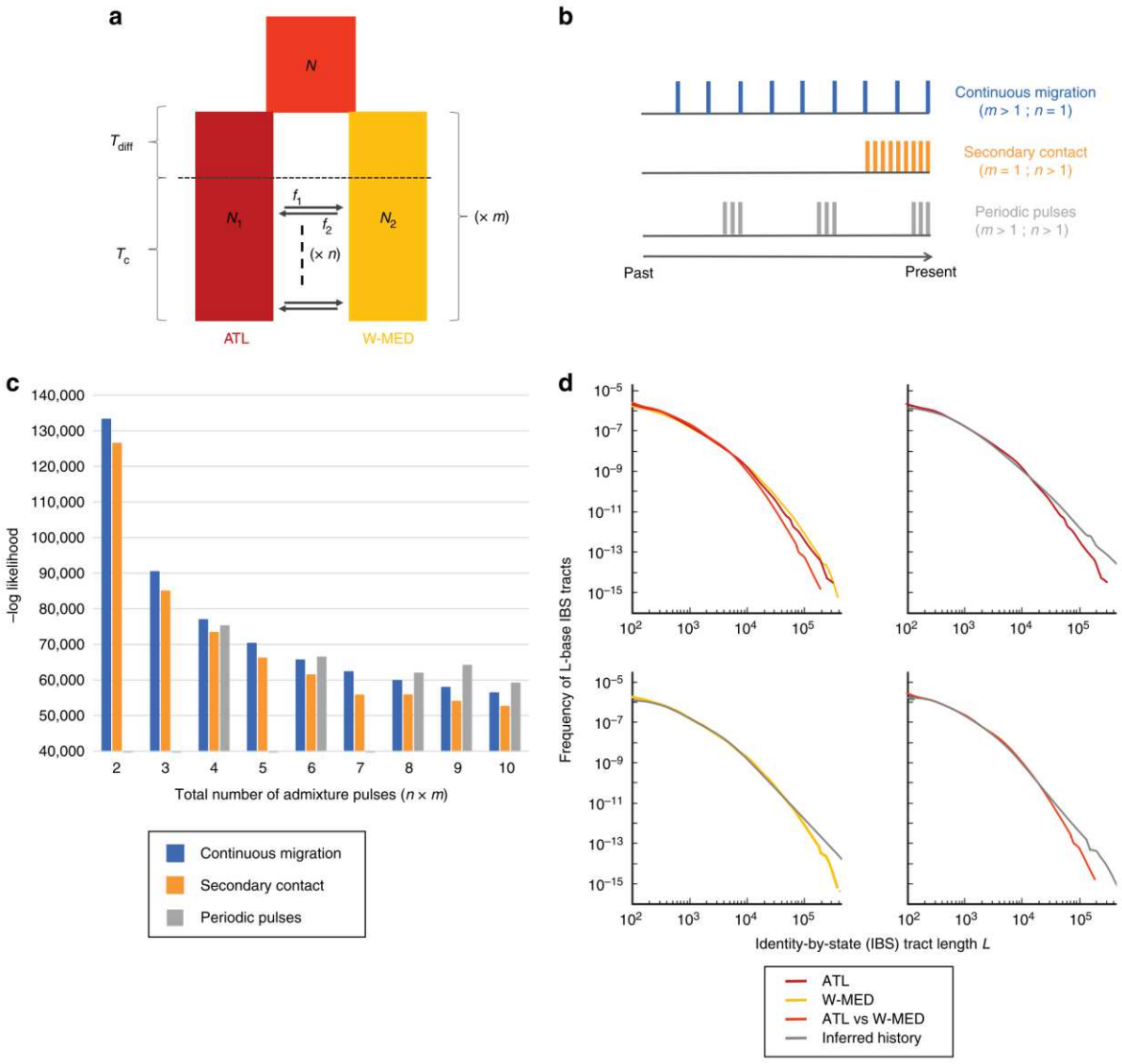


Fig. 3 Demographic history inferred from the length distribution of IBS tracts. **a** Flexible demographic model accounting for multiple equal-length episodes of divergence and gene flow between Atlantic and Mediterranean sea bass populations. An ancestral population of size N splits into two populations of size N_1 and N_2, experiencing one to several (m) cycles of interrupted gene flow during T_{diff} generations followed by migration during T_c generations. Each contact episode contains one to several (n) pulses of admixture (black arrows), replacing Mediterranean and Atlantic populations by a proportion f_1 and f_2 of migrants, respectively. The most recent admixture pulse occurs at time $T_c/100$ before the end of each contact episode, and preceding pulses are homogeneously distributed every $(T_c - T_c/100)/n$ time interval. **b** Modeling scenarios fall into three different categories according to the distribution of admixture pulses over time: continuous migration, secondary contact, and periodic pulses. The illustrated example shows these three categories with nine pulses represented by vertical bars along the timeline. **c** The log-likelihood values obtained for each of the three modeling scenarios from two to ten admixture pulses. Two possible configurations of the periodic pulses scenario exist for models including a total of six, eight, and ten pulses (Supplementary Table 2), only the best of which is represented here. The periodic pulses scenario is not defined for two, three, five, and seven pulses. **d** Goodness-of-fit of the best model ($m = 1$, $n = 10$), showing the length distributions of IBS tracts from observed data (colored lines) compared to model prediction (gray lines). Upper-left: the three distributions observed within ATL and W-MED and between ATL and W-MED populations. Upper-right: model fit within ATL. Lower-left: model fit within W-MED. Lower-right: model fit between ATL and W-MED

likelihood with increasing numbers of pulses in each scenario (Fig. 3c). This is because the total number of admixture pulses contained in each model (i.e., the product $m \times n$) acts as a hidden nuisance parameter, even though the three different scenarios were built using the same number of model parameters. Therefore, we only compared likelihood values among scenarios with identical number of admixture pulses, considering up to 10 pulses in total (the likelihood tended to flatten out beyond this value). We found the secondary contact scenario to be the best-supported model across the entire range of admixture pulse number (Fig. 3c and Supplementary Table 2) and therefore found no support for a periodic pulse model with separate waves of gene flow. The best-fit secondary contact model ($n = 10$ pulses, Fig. 3d) provided clear support for asymmetric introgression, with a more than six-fold higher introgression rate from the Atlantic into the Mediterranean than in the opposite direction, which is consistent with previous findings[31]. The duration of allopatric divergence relative to the secondary contact period was found shorter than previously reported (Supplementary Table 3). Nevertheless, the splitting times estimated between Atlantic and

Mediterranean sea bass lineages were highly consistent across methods (ca. 300,000 years BP for the best identity-by-state (IBS) tract model vs. 270,000 years BP from the best JAFS model[31]).

Linking genomic islands to modes of selection. We investigated chromosomal patterns of genetic differentiation (F_{ST}) and absolute sequence divergence (d_{XY}) between Atlantic and Mediterranean populations. We found highly varying levels of relative and absolute divergence across the genome, with Mb-scale regions of elevated divergence preferentially mapping to low-recombining regions (Fig. 4a–c and Supplementary Fig. 9). Consistent with predictions from the linked selection hypothesis[27,37], F_{ST} and d_{XY} were, respectively, negatively and positively related to the population-scaled recombination rate (ρ

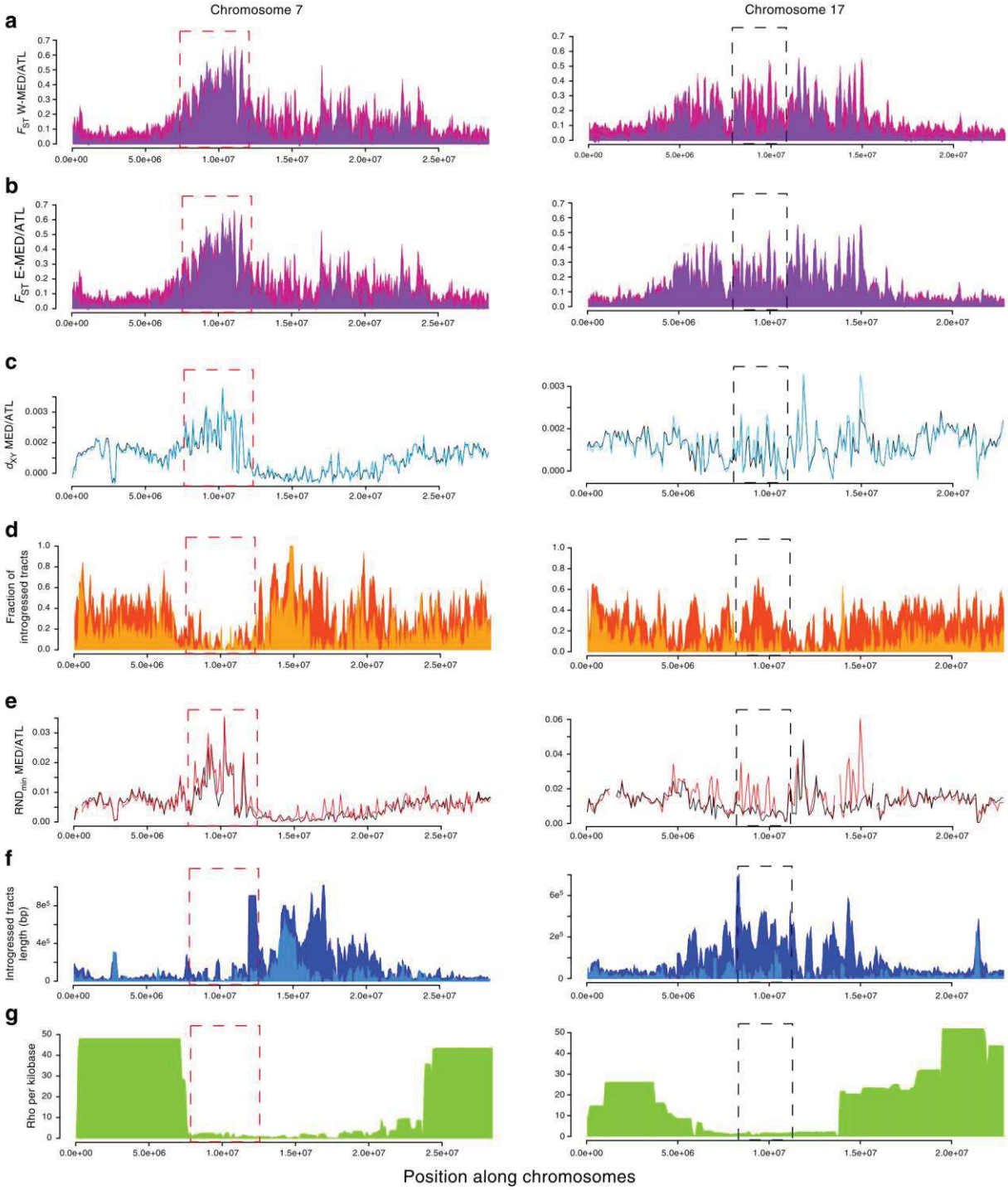


Fig. 4 Population genetic statistics calculated in non-overlapping 100 kb windows along chromosomes 7 and 17. **a** F_{ST} measured between the Atlantic and contemporary (purple) or ancestral reconstructed (mauve) W-MED or E-MED (**b**) population. **c** d_{XY} calculated between the Atlantic and the W-MED (black) or E-MED (blue) population. **d** Fraction of introgressed tracts in the W-MED (orange) or E-MED (yellow) population. **e** RND_{min} measured between the Atlantic and W-MED (black) or E-MED (red) population. **f** Average length of introgressed tracts in the W-MED (dark blue) or E-MED (light blue) population. **g** Population-scaled recombination rate ($\rho = 4N_e r$) averaged between Atlantic and Mediterranean population[31]

$= 4N_e r$) (Fig. 5a–d, Supplementary Figs. 10 and 11). However, the highest F_{ST} values tended to be associated with high values of d_{XY} mapping preferentially to low-recombining regions (Fig. 5), which is not expected under the single action of linked selection[38]. Specifically, we found that the most divergent windows in terms of both F_{ST} and d_{XY} were also associated with high RND_{min} values (Fig. 5e, f and Supplementary Fig. 12) or low frequencies of introgressed tracts (Fig. 5g, h and Supplementary Fig. 13), indicating increased resistance to gene flow. This is consistent with the similar levels of contemporary and ancestral F_{ST} values reconstructed in these regions (Fig. 4a, b, mauve vs. purple F_{ST} plots in the red dotted box).

In order to evaluate the extent to which these observations support the existence of genomic islands resistant to gene flow, we simulated the secondary contact scenario under different modes of selection. Our simulations show that only the BGS+RI model can reproduce the combinations of statistics observed within genomic islands characterized by high values of both F_{ST}, d_{XY}, and RND_{min} (Fig. 6a). By contrast, the Neutral model cannot generate high F_{ST} values due to migration, whereas, although the BGS model could reach such values provided sufficient strength of background selection (BGS), it would on the other hand generate low d_{XY} values. The comparison between observed and simulated data revealed that the BGS+RI model outperformed the Neutral and the BGS model for 16.7% of observed genomic windows, which are also characterized by high values of d_{XY}, F_{ST}, and RND_{min} (Fig. 6b). Therefore, barriers to gene flow between Atlantic and Mediterranean sea bass tend to involve regions of low recombination where both relative and absolute divergence are higher than expected under the sole effect of linked selection.

Variable degrees of resistance to gene flow among genomic regions were also detected using spatial comparisons of divergence patterns. Despite the strong correlation observed between ATL/W-MED and ATL/E-MED F_{ST} patterns (Supplementary Fig. 5), some peaks of ancestral differentiation almost completely vanished in ATL/W-MED contemporary patterns but remained remarkably unchanged in the ATL/E-MED comparison (Fig. 4, black dotted box). By contrast, some peaks seem to have resisted to introgression in both ATL/W-MED and ATL/E-MED comparisons (Fig. 4, red dotted box). These differences are consistent with variable strength of the barrier effect among genomic regions involved in RI and possibly to local introgression swamping (i.e., adaptive introgression in the W-MED or genetic incompatibilities that escaped from coupling).

Despite selection against introgression in some genomic regions, the length of introgressed tracts at the genome scale was largely determined by the local recombination rate (Fig. 4f±g). However, we found that the mean length of introgressed tracts was reduced (up to four-fold) in the close vicinity (<200 kb) of the strongest barriers to gene flow identified by outlier RND_{min} values (Supplementary Fig. 14), consistent with theoretical predictions[39]. When combined across all chromosomes, such regions only represented 4% of the genome.

Finally, we found a negative correlation between the local fraction of introgressed tracts and both ancestral F_{ST} and RND_{max} (Supplementary Fig. 15). This indicates that the regions with the highest level of precontact differentiation and maximal divergence were the less likely to introgress during the recent secondary contact episode. Moreover, observed values of d_{XY} and RND_{min} within genomic islands resistant to gene flow were higher than those obtained using simulations under the BGS+RI model. These results thus support that the haplotypes located in the genomic regions containing RI loci are older than the average age of alleles across the genome.

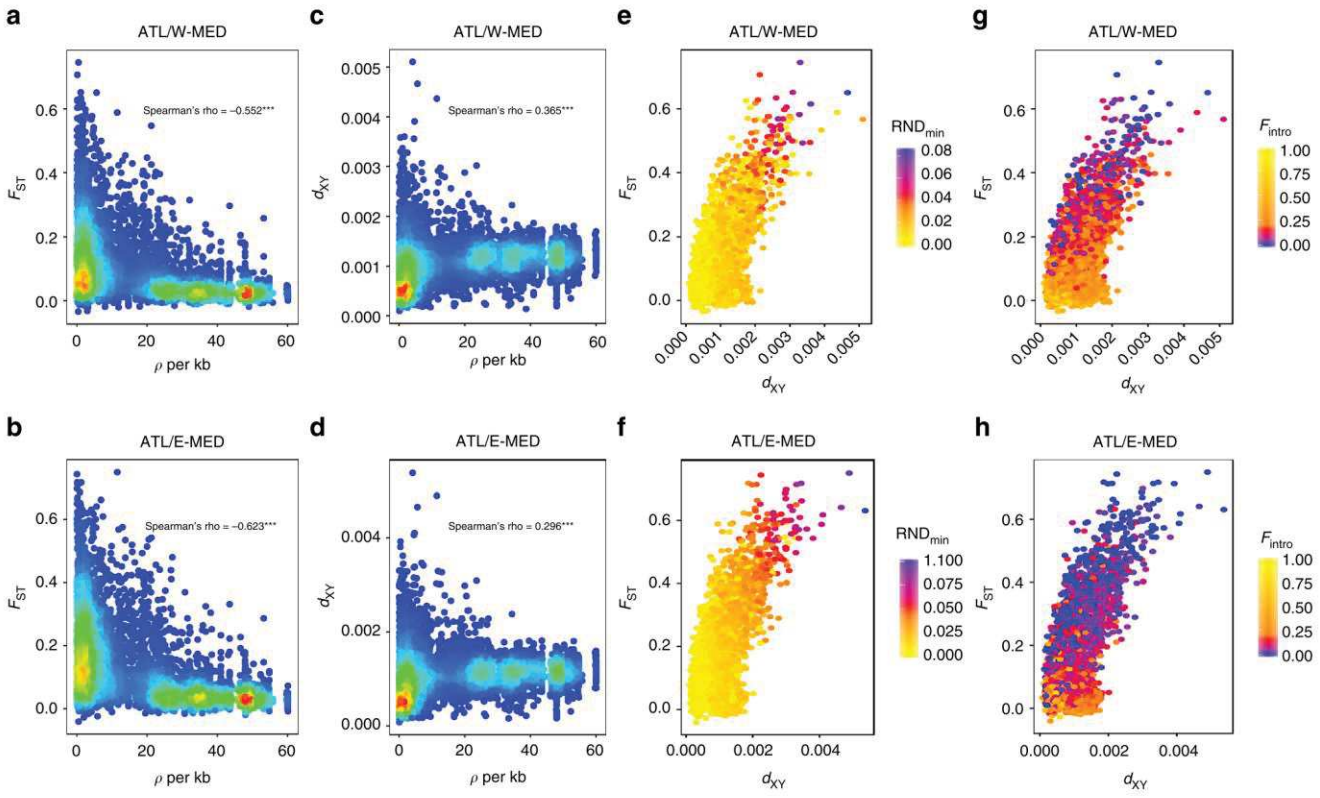


Fig. 5 Relationships between divergence (F_{ST} and d_{XY}), the population-scaled recombination rate ($\rho = 4N_e r$), and introgression statistics (RND_{min} and F_{intro}). **a**–**d** The density of points appears in color scale from low (blue) to high (red) densities. **e**–**h** The color scale indicates the value of RND_{min} (**e**, **f**) or the frequency of introgression (**g**, **h**) in the corresponding window from low (blue) to high (yellow) introgression rate values

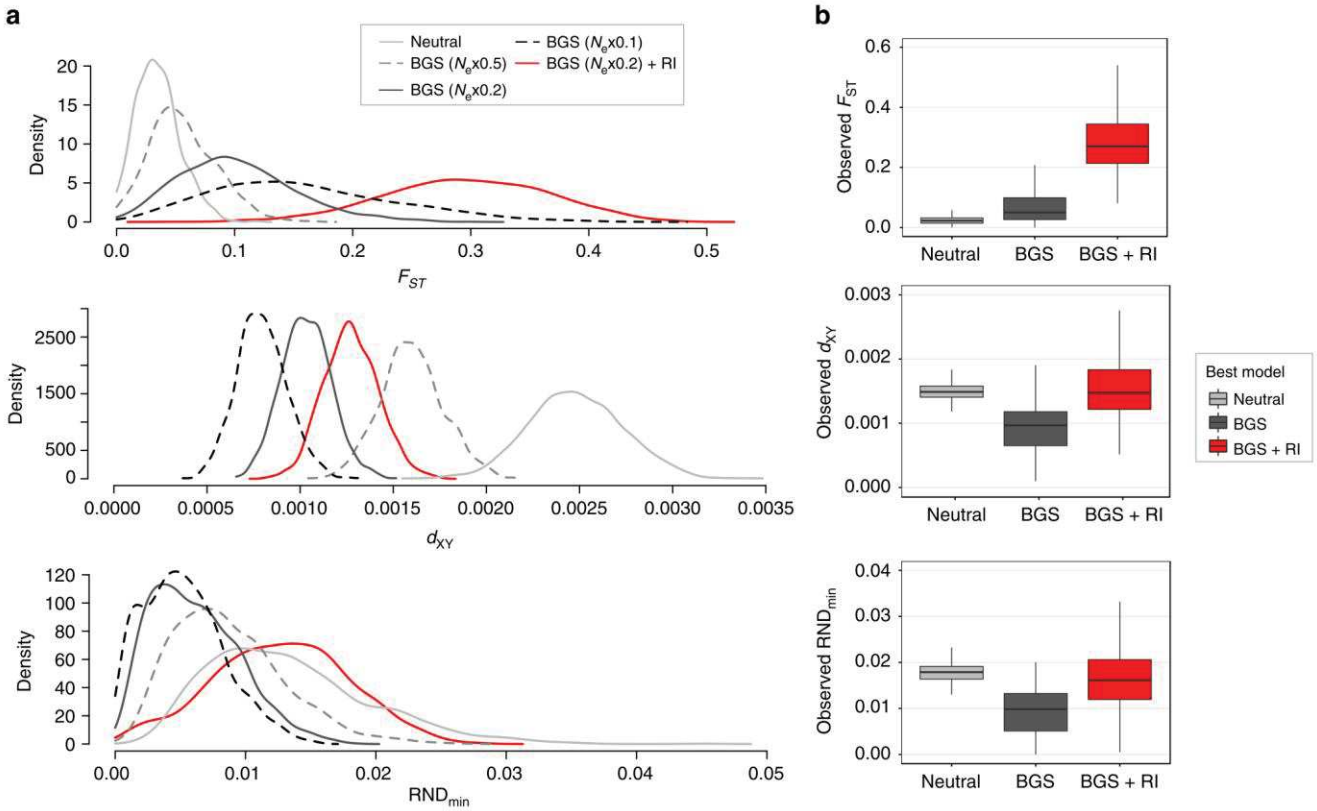


Fig. 6 Simulations under different modes of selection to understand the mechanisms underlying genomic islands. **a** Comparison among simulated distributions of 100 kb window-averaged F_{ST} (top), d_{XY} (middle), and RND_{min} (bottom) obtained under different versions of the secondary contact scenario, including neutral divergence and introgression (Neutral), varying strengths of background selection (BGS, from 0.5 to 0.1 × N_e), and BGS with reproductive isolation (BGS+RI). **b** Comparison among observed distributions of F_{ST} (top), d_{XY} (middle), and RND_{min} (bottom) for real genomic windows of 100 kb that were either assigned to the Neutral model (7% of windows), the BGS model (76% of windows), or the BGS+RI model (17% of windows) based on Euclidean distances to simulated data

Discussion

We used two different approaches based on haplotype information to resolve the history of divergence and gene flow between Atlantic and Mediterranean sea bass lineages. First, our coalescent simulations with recombination showed that the observed length distribution of introgressed tracts can be well reproduced by the post-glacial secondary contact model inferred in a previous study[31]. We then evaluated whether the presence of short introgressed tracts in low-recombining regions could indicate the existence of older admixture events. Although there is a possibility that, during the whole divergence history, the quasi-100,000-year glacial cycles[36] have promoted discrete waves of gene flow, we did not find evidence for a cyclic connectivity model representing glacial oscillations. Instead, our demographic inferences based on the IBS tract spectrum also supported a scenario of secondary contact and confirmed that Atlantic and Mediterranean sea bass lineages have started to diverge around 300,000 years BP. We nonetheless detected an older time of contact that might explain the shortest introgressed tracts found in low-recombining regions.

Admittedly, the power of our inferences may be limited by the confounding effects of time and recombination on the length of introgressed tracts[40–42]. Therefore, haplotype-based methods may be sensitive to local inaccuracies in the estimation of recombination rate along the genome. Furthermore, although the magnitude of divergence was captured by the between-lineages spectrum of shared IBS tracts, the two within-lineages spectrums were highly similar, possibly leading to a lack of signal to precisely estimate the duration of secondary contact with the IBS tract method.

The two approaches implemented here neglect the effects of temporal changes in effective population size and selection against introgressed tracts. Populations surviving in glacial refugia possibly experienced bottlenecks[34], which could have impacted the IBS tract spectrum[43]. Furthermore, the removal of long blocks of foreign ancestry by selection can reduce the average length of introgressed tracts, although we showed that this only happens in a minor fraction (4%) of the sea bass genome. Therefore, selection against migrant tracts is unlikely to cause significant violations to our neutral modeling approach, which should not interfere with its capacity to discriminate among alternative divergence scenarios. Nevertheless, unaccounted selection may explain the over-predicted abundance of introgressed tracts <500 kb in the Mediterranean (Fig. 2) and of IBS tracts >100 kb (Fig. 3). Future works will have to integrate the effect of selection against introgression within demographic models, as previously done for demographic inference from unphased data[6,44,45].

The role of gene flow in generating genomic islands is a long standing debate[3,22,27,37,46,47]. In particular, whether gene flow simply remodels evolving or pre-existing divergence patterns or constrains divergence to evolve principally in low-recombining regions remains an open question[24]. Our results demonstrate that linked selection[48] has increased the rate of lineage sorting in low-recombining regions, generating heterogeneous genome divergence between sea bass lineages during their geographic isolation.

This was supported both by the genome-wide correlations between divergence and recombination and the reconstructed ancestral landscape of divergence. The reduction of recombination in the center of chromosomes relative to their peripheries is a common feature of fish genomes[49,50], which is possibly due to crossover interference[51] and male heterochiasmy[52]. Therefore, linked selection generating heterogeneous differentiation across the genome can be seen as a null expectation for allopatric divergence in sea bass, as in other teleost fish.

Gene flow after secondary contact has the potential to remodel heterogeneous divergence landscapes by eroding neutral differentiation[28], sometimes rapidly[24]. However, how RI loci affect the dynamics of erosion of pre-existing islands of differentiation during secondary contact remains poorly understood. Our direct detection of introgressed tracts revealed broad variation in the rate of introgression among regions displaying similar levels of divergence. In some regions where ancestral peaks of F_{ST} have been found, the amount of introgression was high and the peaks almost vanished after secondary contact. This suggests that these incidental islands do not contain RI loci[27], or if they contained any, have managed to escape coupling with other such genes. On the other hand, reduced introgression in many other regions confirms the view that introgressed alleles generally have negative fitness effects in the foreign genetic background[30,53] (although we did not specifically address the extent of adaptive introgression in that study). The spatial comparison of introgression patterns between recipient populations (W-MED and E-MED) at variable distances from the source population (ATL) provides indirect cues about the distribution of fitness effects of introgressed tracts. Some genomic islands were resistant to introgression in both Mediterranean populations, indicating strong selection against introgressed tracts. In most genomic islands, however, introgression was more reduced in the E-MED compared to the W-MED. This either suggests that stronger migration overwhelms the effect of selection in the W-MED or that selection takes more time to remove weakly selected migrant tracts as they diffuse from the western to the eastern part of the Mediterranean sea.

An important result stemming from the analysis of introgression is that the degree of resistance to gene flow for a given region was positively related to the past level of differentiation and to the absolute divergence between haplotypes, independently of the local mutation rate. Therefore, the strength of selection against introgressed tracts at a given genomic location seems to be at least partly explained by the coalescence time between haplotypes. The expected amount of absolute divergence in low-recombining regions of the sea bass genome can be determined by summing values of ancestral diversity ($\theta_{anc} \approx 0.001$, estimated from the mean diversity of contemporary populations) and sequence divergence due to the accumulation of mutations in both lineages after split ($2\mu T \approx 0.001$, determined using the divergence time estimated from demographic models). This amounts to $E(d_{XY}) = \theta_{anc} + 2\mu T = 0.002$, a value twice higher than the observed genome-wide average, which is decreased by introgression. By contrast, the strongest barriers to gene flow between Atlantic and Mediterranean sea bass involve regions with higher d_{XY} values ranging between 0.002 and 0.005. Our simulations under the BGS +RI model confirmed that this excess of coalescence time is not expected under a scenario whereby differential introgression reshapes a heterogeneous divergence landscape previously established by linked selection alone[38]. To explain these results, we thus need to consider the sorting of ancient polymorphisms during the divergence period, subsequently acting as barrier loci upon secondary contact.

The origin of such alleles that started to diverge before the average coalescent time expected from historical reconstructions remains uncertain. Recent studies have emphasized the role of introgression from a distantly related lineage as a source of new adaptations[54–56], which can possibly play a role in RI[57]. Although we previously found no evidence for contemporary gene flow between *D. labrax* and its closest relative *D. punctatus*[31], we cannot rule out a possible past admixture with *D. punctatus* or an extinct lineage or the existence of an ancestral population structure[42]. Another possible explanation involves the existence of balanced polymorphisms of various types (e.g., frequency-dependent selection or local adaptation) maintained for a long time in the ancestral population, followed by the fixation of alternative alleles in the derived populations[58].

Whatever the origin of the alleles that contribute to RI, our results clearly show that barrier loci tend to map preferentially to low-recombining regions. Such pattern may arise through different but non-mutually exclusive mechanisms. First, during isolation, weakly deleterious mutations are more likely to become fixed by drift or due to hitchhiking with a positively selected allele if recombination is reduced[59]. This may subsequently trigger the fixation of compensatory mutations independently in each population, which could become genetic incompatibilities upon contact[60]. Another effect of linked selection is to accelerate lineage sorting[27], increasing the chance to fix alternative alleles in low-recombining regions during short isolation periods (i.e., <10 N_e generations). Finally, during secondary contacts, the retention of divergence is facilitated when multiple incompatibility loci combine their effects through linkage, especially if some of these loci are involved in local adaptation[61]. Therefore, the density of selected sites determines, in interaction with recombination, the strength of selection against introgressed tracts and the tendency for increased neutral introgression near chromosome extremities[24,28,30,53,62]. These different effects are likely to be amplified if ancestral variation has been fueled by foreign alleles coming from a distant lineage during the divergence history.

To conclude, our results shed new light on the origin and remolding of genomic islands during the divergence history of European sea bass lineages. Thanks to the use of haplotypic information, we provide a more mechanistic understanding of the complex interplay between linked selection, allele age, and resistance to introgression. The recombination landscape appears to be an essential driver of the observed genomic patterns, influencing both lineage sorting and introgression. Our findings also support that the genomic islands generated by linked selection tend to be disproportionately involved in RI during allopatric speciation, although some of them are purely incidental and are currently being eroded by gene flow. Finally, the probability of introgression in a particular genomic region is negatively related to the level of divergence between alleles, a result possibly indicating that either past admixture or long-term balancing selection has participated to the evolution of reproductive barriers in sea bass.

Methods

Whole-genome resequencing and haplotyping. Haplotype-resolved whole genomes were obtained using a phasing-by-transmission approach[63]. Eight parent–offspring trios were generated using 16 wild European sea bass, including 4 males from the Atlantic Ocean (English Channel, $♂_{ATL}$), 4 males from the eastern Mediterranean sea (2 from Turkey and 2 from Egypt, $♂_{E\text{-}MED}$), and 8 females from the western Mediterranean sea (Gulf of Lion, $♀_{W\text{-}MED}$). This trio design is well adapted to recover haplotype information and use local ancestry for inferring the history of divergence, which does not require large sample sizes. Wild parents were crossed in the laboratory to produce 4 families of $♂_{ATL} \times ♀_{W\text{-}MED}$ and 4 families of $♂_{E\text{-}MED} \times ♀_{W\text{-}MED}$ (Supplementary Fig. 1, Supplementary Table 1). Artificial mating and fish rearing were performed in normal conditions at the Ifremer aquaculture facility under experimental agreement C 34-192-6 and in agreement with the French decree no. 2013 -118 1 February 2013 NOR:AGRG1231951D. For each family, the two parents and one randomly selected descendant were submitted to whole-genome resequencing.

Individual genomes were sequenced to an average depth of 15.5× (Supplementary Table 1, Supplementary Fig. 2) using Illumina paired-end reads of

100 bp. Reference alignment to the sea bass genome[31] was performed using BWA-mem v0.7.5a[64]. After filtering duplicate reads, the mean coverage depth per family ranged between 11.3× and 17.6×. We followed the Genome Analysis Toolkit (GATK v3.3-0) best practice pipeline[65,66], applying both base-quality score recalibration and variant-quality score recalibration before calling variants (Supplementary Note 1). Phasing-by-transmission was performed using default parameters. For all downstream analyses, we only used SNPs from parental genomes that were successfully phased using the information contained in the genome of their offspring. SNPs located on the mitochondrial chromosome and the ungrouped fraction of the genome were excluded. Finally, we applied a stringent filter on individual genotype quality (>30) and only retained variants without any missing or excluded genotype. Our final dataset consisted of 2,628,725 SNPs phased into chromosome-wide haplotypes from 14 individuals.

Detection of introgressed haplotypes. We used Chromopainter v0.04[67] to perform local ancestry inference and identify migrant tracts resulting from introgression between Atlantic and Mediterranean lineages. The program uses a hidden Markov Model to estimate the probability of Atlantic and Mediterranean ancestry at each variable position along the genome using patterns of haplotype similarity. Each individual was separately considered as a "recipient" and compared to "donor" individuals present in reference Atlantic and Mediterranean populations. A recipient chromosome was reconstructed as a combination of DNA chunks from donor individuals, the donor of each chunk being identified as the most similar haplotype from the reference populations. Therefore, the local ancestry profile of a given recipient chromosome was made of a mosaic of DNA chunks for which the probability to be inherited from either lineages was inferred by Chromopainter.

We developed a method to determine the starting and ending positions of each migrant tract (Supplementary Fig. 3). Atlantic migrant tracts within Mediterranean genomes were delimited by analyzing the probability of Atlantic ancestry inferred by Chromopainter. A given haplotype was considered as truly Atlantic when this probability was >0.95 and as Mediterranean when it was <0.05. Positions with intermediate probabilities lying between 0.05 and 0.95 corresponded to ambiguous regions that were assigned the same ancestry as non-ambiguous neighboring sites. Therefore, an Atlantic tract was defined by a starting position where the inferred probability of Atlantic ancestry reached 0.95 and an ending position where it dropped below 0.05. To avoid bias in the estimated length of introgressed tracts depending on the reading direction of ancestry profiles, we shifted the beginning and ending point of each tract to the closest position with an ancestry probability of 0.5. The presence of Mediterranean migrant tracts within Atlantic genomes were identified in the same way, by analyzing the probability profile of Mediterranean ancestry.

Chromopainter requires non-introgressed reference individuals from every potential source population in order to detect introgressed haplotypes in focal samples. Although the Atlantic lineage is only slightly introgressed by Mediterranean alleles, the Mediterranean lineage is by contrast heavily impacted by gene flow. Therefore, the presence of Atlantic haplotypes in both Atlantic and Mediterranean references populations can confound the accurate prediction of local ancestry by Chromopainter. We thus developed a procedure to reconstruct non-introgressed Mediterranean genomes by removing migrant tracts in a stepwise manner (Supplementary Note 3, Supplementary Fig. 4). Introgressed segments of Atlantic origin were identified by exploiting the different levels of introgression existing between W-MED and E-MED populations (31% in the W-MED compared to 13% in the E-MED population, see Results). This strategy enabled us to reconstruct the ancestral genetic diversity of the Mediterranean populations before secondary gene flow from the Atlantic, which significantly improved the detection of introgressed tracts (Supplementary Note 4, Supplementary Figs 5 and 6).

Analysis of migrant tract length distribution. The most abundant class of tracts found in low-recombining regions of the sea bass genome have an average length of 50 kb (Supplementary Fig. 7). Using an analytical expectation for the average length of migrant tracts following an admixture pulse[42], we estimated that these tracts have introgressed the W-MED population approximately 17,000 years BP (Supplementary Note 5). Given that 85% of the Atlantic introgressed tracts found in low-recombining regions are on average >50 kb, the length distribution of migrant tracts is consistent with a post-glacial secondary contact. To assess the goodness-of-fit of the secondary contact model previously inferred from the joint allele frequency spectrum[31], we compared the length distributions of migrant tracts observed in Atlantic and Mediterranean populations to simulated distributions. We used the coalescent simulator msprime v0.4.0[68] to generate genome-scale haplotype data with variable recombination rates under the previously inferred secondary contact model[31] (Supplementary Fig. 8), using the same parameter values. We then analyzed simulated data with Chromopainter to get the genome-wide distribution of migrant tract length in each population and compared it with the observed distributions.

Testing waves of historical gene flow. The demographic history of Atlantic and Mediterranean sea bass populations was inferred from the length distribution of tracts of IBS using a composite likelihood method[43]. We extended this approach to test for successive waves of gene flow during divergence (Supplementary Note 6). We developed a simple and flexible model that can account for multiple equal-length episodes of divergence and gene flow between two populations (Fig. 3a). Using only nine parameters, the model can represent a large range of demographic scenarios falling into three categories. (i) Continuous migration (an approximation of the Isolation-with-Migration model) is modeled using several contacts ($m > 1$) each containing a single pulse ($n = 1$), with no isolation period separating contact episodes ($T_{\mathrm{diff}} \approx 0$). In this scenario, the m pulses are therefore continuously distributed along the whole divergence history, with T_c generations separating two consecutive pulses. (ii) Secondary contact with a single period of isolation and gene flow ($m = 1$), including a long enough interruption of gene flow to allow divergence before contact $\left(T_{\mathrm{diff}} > \frac{T_c}{n}\right)$. (iii) Periodic pulses with $m > 1$ and a long enough interruption of gene flow to initiate divergence between two successive contacts ($T_{\mathrm{diff}} > 0.1 \times N_e$ and $T_{\mathrm{diff}} > \frac{T_c}{n}$). In secondary contact and periodic pulses scenarios, continuous gene flow within contact episodes can be approximated with several admixture pulses sufficiently close in time ($\frac{T_c}{n} \ll N_e$). Even though every scenario can be modeled using the same number of parameters, the total number of admixture pulses in a given scenario (i.e., the product $m \times n$) acts as a nuisance parameter. Therefore, we only compared estimated composite likelihoods among scenarios having the same total number of admixture pulses (Fig. 3c and Supplementary Table 2).

The information about the timing of introgression events is expected to be better preserved within low-recombining regions[40,42]. Therefore, we only used sequence information from the low-recombining fraction of the sea bass genome (using $\rho \leq 10$, the population-scaled recombination rate estimated by ref. [31]) to infer the demographic divergence history from the length distribution of IBS tracts. Inferences were performed using a grid search on the total number of pulses, making the product $m \times n$ vary from 1 to 10. For every combination of m and n values, we used 20 independent runs in which we let the 7 other parameters being freely estimated during composite likelihood optimization (Supplementary Table 3).

Whole-genome alignment with an outgroup species. We used the 35,012 scaffolds from the *Morone saxatilis* genome assembly (www.ncbi.nlm.nih.gov/assembly/GCA_001663605.1/; sequence length = 585.2 Mb; N50 = 30 kb) to perform alignments against the reference genome of *D. labrax*. The Mauve Contig Mover tool[69] from the Mauve software v2.4.0[70] was used to match every scaffold from *M. saxatilis* to the *D. labrax* genome and generate a list of scaffolds matching to each *D. labrax* chromosome. Matching scaffolds were properly ordered and assembled along each chromosome to generate 24 pseudo-chromosomes using the software Abacas version 1.3.1[71]. We then aligned *M. saxatilis* pseudo-chromosomes to the *D. labrax* genome using the algorithm progressiveMauve in order to identify insertions in the *M. saxatilis* genome that were removed to only conserve homologous sites present in *D. labrax*. After whole-genome alignment, mapped scaffolds were evenly distributed among the 24 sea bass chromosomes, providing outgroup information for 52% of the *D. labrax* genome. The genome-wide average nucleotide divergence between *D. labrax* and *M. saxatilis* was 4.35% (s.d. = 2.36%).

Population genomics statistics. We used several complementary approaches to assess the level of divergence and introgression between lineages. First, the extent of genetic differentiation between Atlantic and Mediterranean populations of *D. labrax* was evaluated using both relative (F_{ST}[72]) and absolute (d_{XY}[73]) measures of divergence. We used Vcftools v0.1.11[74] and MVFTools v3.0[75] to calculate the average F_{ST} and d_{XY} in non-overlapping 100 kb windows. Second, we used Chromopainter outputs to directly measure the percentage of positions occupied by migrant tracts and calculate their average length in 100 kb windows. Finally, we calculated the minimum and maximum pairwise distance between haplotypes sampled from Atlantic and Mediterranean populations (d_{XYmin} and d_{XYmax}) and divided these values by the divergence between *D. labrax* and *M. saxatilis* (d_{out}) in each window to compute RND_{min}[76] and RND_{max} statistics. The RND_{min} ratio is more sensitive to low-frequency migrant tracts than F_{ST} and d_{XY}. When introgression occurs at a genome-wide scale, the highest RND_{min} values indicate regions of reduced introgression, whereas RND_{max} rather reflects the maximal absolute divergence among haplotypes irrespective to introgression. Both statistics are robust to mutation rate variation across the genome.

Testing for RI. In order to test whether genomic islands are actively reshaped by migration and selection, we used simulations to generate predictions under the secondary contact scenario assuming different selective effects. More specifically, we evaluated the influence of BGS and RI on the distribution of F_{ST}, d_{XY}, and RND_{min} statistics compared to neutral expectations. The effect of RI and BGS in low-recombining regions were approximated by respectively reducing the effective population size (N_e)[77] and the effective migration rate (m_e)[28] compared to neutral loci. We used msprime v.0.0.4[68] to simulate a post-glacial secondary contact model under three different selection scenarios. (i) A neutral model parameterized using the values inferred from the joint allele–frequency spectrum[31] (i.e., similar to Fig. 2). (ii) A BGS model with an increased rate of lineage sorting and neutral introgression. Three possible levels of BGS were applied to both ancestral and derived populations (i.e., $0.5 \times N_e$, $0.2 \times N_e$, and $0.1 \times N_e$). The middle range value was empirically estimated to correspond to the reduction in nucleotide diversity

observed in low compared to highly recombining regions due to linked selection[31]. Finally, (iii) A BGS+RI model with an increased rate of lineage sorting and a decreased rate of introgression. The effective population size was reduced to $0.2 \times N_e$ in both ancestral and derived populations, and the effective migration rate was reduced to $\sim 0.2 \times m$ compared to neutral regions, as previously inferred for genomic island loci[31]. For each model, we simulated a 150 Mb chromosome with an average recombination rate corresponding to the mean value calculated across the sea bass genome (6.85 cM/Mb). Divergence and introgression statistics were calculated in non-overlapping 100 kb windows using Vcftools[74] for F_{ST} and MVFtools[75] for d_{XY} and RND_{min}. To test which simulated model best explains our data, we compared the joint distributions of d_{XY}, F_{ST}, and RND_{min} obtained by simulations under each model to the data. For each 100 kb window, we calculated the mean Euclidean distance between the Z-scored values of observed and simulated statistics under each of the three models (i.e., Neutral, BGS, and BGS+RI). Models were then ranked by their Euclidean distance to observed data, to determine the best model for each 100 kb window.

Data availability. Sequence reads have been deposited in the GenBank Sequence Read Archive under the accession code BioProject PRJNA472842.

Received: 4 December 2017 Accepted: 23 May 2018
Published online: 28 June 2018

References

1. Mayr, E. *Systematics and the Origin of Species, from the Viewpoint of a Zoologist* (Harvard University Press, Cambridge, 1942).
2. Coyne, J. A. & Orr, A. H. *Speciation* (Sinauer Associates, Sunderland, MA, 2004).
3. Wu, C.-I. The genic view of the process of speciation. *J. Evol. Biol.* **14**, 851–865 (2001).
4. Feder, J. L., Egan, S. P. & Nosil, P. The genomics of speciation-with-gene-flow. *Trends Genet.* **28**, 342–350 (2012).
5. Harrison, R. G. & Larson, E. L. Hybridization, introgression, and the nature of species boundaries. *J. Hered.* **105**, 795–809 (2014).
6. Roux, C. et al. Shedding light on the grey zone of speciation along a continuum of genomic divergence. *PLoS Biol.* **14**, e2000234 (2016).
7. Barton, N. Gene flow past a cline. *Heredity* **43**, 333–339 (1979).
8. Hewitt, G. M. Hybrid zones-natural laboratories for evolutionary studies. *Trends Ecol. Evol.* **3**, 158–167 (1988).
9. Seehausen, O. et al. Genomics and the origin of species. *Nat. Rev. Genet.* **15**, 176–192 (2014).
10. Turner, T. L., Hahn, M. W. & Nuzhdin, S. V. Genomic islands of speciation in *Anopheles gambiae*. *PLoS Biol.* **3**, e285 (2005).
11. Harr, B. Genomic islands of differentiation between house mouse subspecies. *Genome Res.* **16**, 730–737 (2006).
12. Ellegren, H. et al. The genomic landscape of species divergence in *Ficedula* flycatchers. *Nature* **491**, 756–760 (2012).
13. Jones, F. C. et al. The genomic basis of adaptive evolution in threespine sticklebacks. *Nature* **484**, 55–61 (2012).
14. Martin, S. H. et al. Genome-wide evidence for speciation with gene flow in *Heliconius* butterflies. *Genome Res.* **23**, 1817–1828 (2013).
15. Gagnaire, P.-A., Pavey, S. A., Normandeau, E. & Bernatchez, L. The genetic architecture of reproductive isolation during speciation-with-gene-flow in lake whitefish species pairs assessed by RAD sequencing. *Evolution* **67**, 2483–2497 (2013).
16. Renaut, S., Owens, G. L. & Rieseberg, L. H. Shared selective pressure and local genomic landscape lead to repeatable patterns of genomic divergence in sunflowers. *Mol. Ecol.* **23**, 311–324 (2014).
17. Soria-Carrasco, V. et al. Stick insect genomes reveal natural selection's role in parallel speciation. *Science* **344**, 738–742 (2014).
18. Marques, D. A. et al. Genomics of rapid incipient speciation in sympatric threespine stickleback. *PLoS Genet.* **12**, e1005887 (2016).
19. Nadeau, N. J. et al. Genomic islands of divergence in hybridizing *Heliconius* butterflies identified by large-scale targeted sequencing. *Philos. Trans. R. Soc. B* **367**, 343–353 (2012).
20. Nosil, P., Funk, D. J. & Ortiz-Barrientos, D. Divergent selection and heterogeneous genomic divergence. *Mol. Ecol.* **18**, 375–402 (2009).
21. Wolf, J. B. W. & Ellegren, H. Making sense of genomic islands of differentiation in light of speciation. *Nat. Rev. Genet.* **18**, 87–100 (2016).
22. Ravinet, M. et al. Interpreting the genomic landscape of speciation: a road map for finding barriers to gene flow. *J. Evol. Biol.* **30**, 1450–1477 (2017).
23. Harrison, R. G. & Larson, E. L. Heterogeneous genome divergence, differential introgression, and the origin and structure of hybrid zones. *Mol. Ecol.* **25**, 2454–2466 (2016).
24. Yeaman, S., Aeschbacher, S. & Bürger, R. The evolution of genomic islands by increased establishment probability of linked alleles. *Mol. Ecol.* **25**, 2542–2558 (2016).
25. Maynard Smith, J. M. & Haigh, J. The hitch-hiking effect of a favourable gene. *Genet. Res.* **23**, 23–35 (1974).
26. Charlesworth, B., Morgan, M. T. & Charlesworth, D. The effect of deleterious mutations on neutral molecular variation. *Genetics* **134**, 1289–1303 (1993).
27. Cruickshank, T. E. & Hahn, M. W. Reanalysis suggests that genomic islands of speciation are due to reduced diversity, not reduced gene flow. *Mol. Ecol.* **23**, 3133–3157 (2014).
28. Barton, N. & Bengtsson, B. O. The barrier to genetic exchange between hybridising populations. *Heredity* **57**, 357–376 (1986).
29. Nachman, M. W. & Payseur, B. A. Recombination rate variation and speciation: theoretical predictions and empirical results from rabbits and mice. *Philos. Trans. R. Soc. B* **367**, 409–421 (2012).
30. Martin, S. H. & Jiggins, C. D. Interpreting the genomic landscape of introgression. *Curr. Opin. Genet. Dev.* **47**, 69–74 (2017).
31. Tine, M. et al. European sea bass genome and its variation provide insights into adaptation to euryhalinity and speciation. *Nat. Commun.* **5**, 5770 (2014).
32. Lemaire, C., Versini, J.-J. & Bonhomme, F. Maintenance of genetic differentiation across a transition zone in the sea: discordance between nuclear and cytoplasmic markers. *J. Evol. Biol.* **18**, 70–80 (2005).
33. Patarnello, T., Volckaert, Fa. M. J. & Castilho, R. Pillars of Hercules: is the Atlantic–Mediterranean transition a phylogeographical break? *Mol. Ecol.* **16**, 4426–4444 (2007).
34. Hewitt, G. The genetic legacy of the Quaternary ice ages. *Nature* **405**, 907–913 (2000).
35. Hewitt, G. M. Some genetic consequences of ice ages, and their role in divergence and speciation. *Biol. J. Linn. Soc.* **58**, 247–276 (1996).
36. Snyder, C. W. Evolution of global temperature over the past two million years. *Nature* **538**, 226–228 (2016).
37. Burri, R. Interpreting differentiation landscapes in the light of long-term linked selection. *Evol. Lett.* **1**, 118–131 (2017).
38. Burri, R. Dissecting differentiation landscapes: a linked selection's perspective. *J. Evol. Biol.* **30**, 1501–1505 (2017).
39. Sedghifar, A., Brandvain, Y. & Ralph, P. Beyond clines: lineages and haplotype blocks in hybrid zones. *Mol. Ecol.* **25**, 2559–2576 (2016).
40. Pool, J. E. & Nielsen, R. Inference of historical changes in migration rate from the lengths of migrant tracts. *Genetics* **181**, 711–719 (2009).
41. Liang, M. & Nielsen, R. The lengths of admixture tracts. *Genetics* **197**, 953–967 (2014).
42. Racimo, F., Sankararaman, S., Nielsen, R. & Huerta-Sánchez, E. Evidence for archaic adaptive introgression in humans. *Nat. Rev. Genet.* **16**, 359–371 (2015).
43. Harris, K. & Nielsen, R. Inferring demographic history from a spectrum of shared haplotype lengths. *PLoS Genet.* **9**, e1003521 (2013).
44. Sousa, V. C., Carneiro, M., Ferrand, N. & Hey, J. Identifying loci under selection against gene flow in isolation-with-migration models. *Genetics* **194**, 211–233 (2013).
45. Rougeux, C., Bernatchez, L. & Gagnaire, P.-A. Modeling the multiple facets of speciation-with-gene-flow toward inferring the divergence history of lake whitefish species pairs (*Coregonus clupeaformis*). *Genome Biol. Evol.* **9**, 2057–2074 (2017).
46. Charlesworth, B., Nordborg, M. & Charlesworth, D. The effects of local selection, balanced polymorphism and background selection on equilibrium patterns of genetic diversity in subdivided populations. *Genet. Res.* **70**, 155–174 (1997).
47. Noor, Ma. F. & Bennett, S. M. Islands of speciation or mirages in the desert? Examining the role of restricted recombination in maintaining species. *Heredity* **103**, 439–444 (2009).
48. Cutter, A. D. & Payseur, B. A. Genomic signatures of selection at linked sites: unifying the disparity among species. *Nat. Rev. Genet.* **14**, 262–274 (2013).
49. Bradley, K. M. et al. An SNP-based linkage map for zebrafish reveals sex determination loci. *G3 Genes Genomes Genet.* **1**, 3–9 (2011).
50. Roesti, M., Hendry, A. P., Salzburger, W. & Berner, D. Genome divergence during evolutionary diversification as revealed in replicate lake–stream stickleback population pairs. *Mol. Ecol.* **21**, 2852–2862 (2012).
51. Allendorf, F. W. et al. Effects of crossovers between homeologs on inheritance and population genomics in polyploid-derived salmonid fishes. *J. Hered.* **106**, 217–227 (2015).
52. Lien, S. et al. A dense SNP-based linkage map for Atlantic salmon (*Salmo salar*) reveals extended chromosome homeologies and striking differences in sex-specific recombination patterns. *BMC Genomics* **12**, 615 (2011).
53. Schumer, M. et al. Natural selection interacts with recombination to shape the evolution of hybrid genomes. *Science* **360**, 656–660 (2018).
54. Dasmahapatra, K. K. et al. Butterfly genome reveals promiscuous exchange of mimicry adaptations among species. *Nature* **487**, 94–98 (2012).

55. Huerta-Sánchez, E. et al. Altitude adaptation in Tibetans caused by introgression of Denisovan-like DNA. *Nature* **512**, 194–197 (2014).
56. Lamichhaney, S. et al. Evolution of Darwin's finches and their beaks revealed by genome sequencing. *Nature* **518**, 371–375 (2015).
57. Abbott, R. et al. Hybridization and speciation. *J. Evol. Biol.* **26**, 229–246 (2013).
58. Guerrero, R. F. & Hahn, M. W. Speciation as a sieve for ancestral polymorphism. *Mol. Ecol.* **26**, 5362–5368 (2017).
59. Birky, C. W. & Walsh, J. B. Effects of linkage on rates of molecular evolution. *PNAS* **85**, 6414–6418 (1988).
60. Lindtke, D. & Buerkle, C. A. The genetic architecture of hybrid incompatibilities and their effect on barriers to introgression in secondary contact. *Evolution* **69**, 1987–2004 (2015).
61. Bank, C., Bürger, R. & Hermisson, J. The limits to parapatric speciation: Dobzhansky–Muller incompatibilities in a continent–island model. *Genetics* **191**, 845–863 (2012).
62. Aeschbacher, S., Selby, J. P., Willis, J. H. & Coop, G. Population-genomic inference of the strength and timing of selection against gene flow. *PNAS* **114**, 7061–7066 (2017).
63. Browning, S. R. & Browning, B. L. Haplotype phasing: existing methods and new developments. *Nat. Rev. Genet.* **12**, 703–714 (2011).
64. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. Preprint at https://arxiv.org/abs/1303.3997 (2013).
65. McKenna, A. et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
66. Van der Auwera, G. A. et al. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr. Protoc. Bioinformatics* **11**, 11.10.1–11.10.33 (2013).
67. Lawson, D. J., Hellenthal, G., Myers, S. & Falush, D. Inference of population structure using dense haplotype data. *PLoS Genet.* **8**, e1002453 (2012).
68. Kelleher, J., Etheridge, A. M. & McVean, G. Efficient coalescent simulation and genealogical analysis for large sample sizes. *PLoS Comput. Biol.* **12**, e1004842 (2016).
69. Rissman, A. I. et al. Reordering contigs of draft genomes using the Mauve Aligner. *Bioinformatics* **25**, 2071–2073 (2009).
70. Darling, A. C. E., Mau, B., Blattner, F. R. & Perna, N. T. Mauve: multiple alignment of conserved genomic sequence with rearrangements. *Genome Res.* **14**, 1394–1403 (2004).
71. Assefa, S., Keane, T. M., Otto, T. D., Newbold, C. & Berriman, M. ABACAS: algorithm-based automatic contiguation of assembled sequences. *Bioinformatics* **25**, 1968–1969 (2009).
72. Weir, B. S. & Cockerham, C. C. Estimating F-Statistics for the analysis of population structure. *Evolution* **38**, 1358–1370 (1984).
73. Nei, M. *Molecular Evolutionary Genetics* (Columbia University Press, New York, 1987).
74. Danecek, P. et al. The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158 (2011).
75. Pease, J. & Rosenzweig, B. Encoding data using biological principles: the multisample variant format for phylogenomics and population genomics. *IEEE/ACM Trans. Comput. Biol. Bioinform.* https://doi.org/10.1109/TCBB.2015.2509997 (2015).
76. Rosenzweig, B. K., Pease, J. B., Besansky, N. J. & Hahn, M. W. Powerful methods for detecting introgressed regions from population genomic data. *Mol. Ecol.* **25**, 2387–2397 (2016).
77. Charlesworth, B. Effective population size and patterns of molecular evolution and variation. *Nat. Rev. Genet.* **10**, 195–205 (2009).

Acknowledgements

This work was supported by the ANR grants LABRAD-SEQ 11-PDOC-009-01 and CoGeDiv ANR-17-CE02-0006-01 to P.-A.G. We thank Véronique Dhennin and Julien Derop from the sequencing platform of UMR 8199 Génomique Intégrative et Modélisation des Maladies Métaboliques, as well as Alain Vergnet from the experimental aquaculture facilities of Ifremer Palavas for their precious help. We are also grateful to Lamya Chaoui, Hichem Kara, Lilia Bahri-Sfar, and Filipe Martinho for providing samples from Algeria, Tunisia, and Portugal.

Author contributions

M.D. and P.-A.G. wrote the manuscript with inputs from F.A., C.F., N.B., and F.B. Experimental crosses were managed by F.A., genome alignment and SNP calling was performed by C.F. and P.-A.G., and M.D performed all population genomic analyses. P.-A.G. conceived the project and managed financial support and genome sequencing.

Additional information

Supplementary Information accompanies this paper at https://doi.org/10.1038/s41467-018-04963-6.

Competing interests: The authors declare no competing interests.

Reprints and permission information is available online at http://npg.nature.com/reprintsandpermissions/

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2018

CHAPITRE 2 :

Déterminer l'origine des allèles impliqués dans l'isolement reproductif

Les résultats de ce chapitre sont exposés en détails dans l'article joint en révisions à *Evolution Letters*.

L'objectif principal de ce chapitre a été d'identifier l'origine des allèles localisés dans les îlots génomiques de différenciation impliqués dans l'isolement reproductif. En effet, ces régions présentent des niveaux de divergence plus élevés qu'attendu. Dans un modèle neutre, $d_{XY} = \theta_{anc} + 2\mu t$ où θ_{anc} correspond à la diversité génétique de la population ancestrale, μ au taux de mutation de l'espèce et t au temps de divergence entre les deux lignées. On peut donc estimer le niveau de divergence attendu entre la lignée atlantique et méditerranéenne de bar européen. Sachant que $\mu \approx 1e^{-8}$, $t \approx 60000$ générations et que θ_{anc} peut être approximé par la diversité nucléotidique de la population actuelle $\pi \approx 0{,}001$ on s'attend à $d_{XY} \approx 0{,}002$. Or on observe dans les régions impliquées dans l'isolement reproductif des valeurs jusqu'à 0,004 (cf chapitre 1). Ces valeurs particulièrement élevées indiquent que la divergence entre les allèles présents en Atlantique et en Méditerranée a commencé avant que les lignées débutent leur divergence allopatrique. Plusieurs hypothèses peuvent expliquer l'origine de ces vieux allèles, l'une d'entre elle étant leur introgression lors d'un contact avec une troisième lignée (Green *et al.* 2010; Meyer *et al.* 2012).

Nous avons voulu tester cette hypothèse en cherchant des traces de flux génique entre *D. labrax* et *D. punctatus*, la seule espèce phylogénétiquement proche vivant en sympatrie partielle avec le bar européen. Pour cela nous avons utilisé trois méthodes différentes. La première est un test appelé ABBA-BABA (Martin *et al.* 2013) qui se base sur la topologie d'arbres phylogénétiques réalisés le long de l'alignement des génomes de 4 taxons. Nous avons donc utilisé ici les deux lignées de bar européen, *D. punctatus* et le bar rayé (*Morone saxatilis*) qui sert ici de groupe externe. Le test permet de rechercher des excès d'allèles dérivés présent chez *D. punctatus* (noté B) et partagés avec la lignée atlantique ou méditerranéenne, le groupe externe possédant toujours l'allèle ancestral (noté A). En absence de flux génique, le polymorphisme étant trié aléatoirement entre l'Atlantique et la Méditerranée, on s'attend à retrouver autant de polymorphismes dérivés partagés entre *D. punctaus* et les deux lignées de bar européen, soit autant de généalogies de type ABBA que BABA. L'excès d'un type de généalogie sera alors interprété comme la trace d'échanges génétiques entre *D. punctatus* et une des deux lignées de bar européen (en fonction du type de généalogie en excès).

La deuxième méthode permet d'identifier des segments archaïques introgressés sans utiliser de génome de référence, ce qui contrairement à la méthode précédente permet de ne pas avoir d'a priori sur la lignée donneuse (Skov *et al.* 2018). L'idée derrière étant que, si parmi deux populations d'une même espèce seule une a eu des échanges génétiques avec une autre lignée, alors la présence d'allèles introgressés va localement augmenter le polymorphisme de cette population par rapport à l'autre. Le test recherche donc des régions génomiques présentant des excès de polymorphismes privés révélant

la présence d'haplotypes introgressés. Enfin, nous avons utilisé la statistique $RND_{min} = d_{XY\text{-}min} / d_{XY\text{-}outgroup}$ (Rosenzweig *et al.* 2016) en mesurant le long du génome la divergence minimale entre la lignée atlantique ou méditerranéenne de *D. labrax* et *D. punctatus* divisée par la divergence moyenne entre ces deux espèces et *M. saxatilis*. Dans un contexte où les échanges génétiques sont fréquents, des valeurs particulièrement élevées de RND_{min} indique que ces régions ont résisté au flux génique. L'avantage de cette méthode est donc que contrairement aux deux autres, ce n'est pas une méthode relative qui se base sur la comparaison des niveaux d'introgression de deux populations. Ainsi, elle va nous permettre de détecter des traces d'introgression même si les échanges ont eu lieu entre *D. punctatus* et les deux lignées de *D. labrax*.

Nous avons ainsi pu détecter des régions génomiques présentant des excès de généalogies ABBA par rapport à BABA et présentant en Atlantique de fortes valeurs de RNDmin où des fragments archaïques introgressés ont été identifiés, signe qu'il y a eu des échanges génétiques entre *D. punctatus* et la lignée atlantique de *D. labrax* (Figure 2). Sachant que la taille des fragments introgressés est notamment déterminée par leur date d'introgression, nous avons pu utiliser la distribution de taille des fragments archaïques introgressés pour dater les échanges génétiques entre ces deux lignées (Figure 3). Le flux génique aurait donc eu lieu il y a environ 80 000 ans, quand la lignée atlantique et méditerranéenne de bar européen étaient isolées dans des refuges glaciaires différents. Les allèles de *D. punctatus,* après être entrés en Atlantique, ont ensuite pu introgresser la Méditerranée quand les échanges génétiques entre les deux lignées de *D. labrax* ont repris. Il existe donc un fort différentiel d'introgression uniquement au niveau des régions impliquées dans l'isolement reproductif entre la lignée atlantique et méditerranéenne. En effet, nous avons pu délimiter ces régions en utilisant un modèle de chaine de Markov caché qui analyse le ratio du F_{ST} sur la fraction d'introgression (cf chapitre 1) le long des génomes. Nous avons alors pu montrer qu'en Atlantique les allèles originaires de *D. punctatus* y sont présent à forte fréquence (voir fixés) alors qu'ils sont présents à faible fréquence dans les autres régions génomiques. Il semblerait donc que l'introgression d'allèles de *D. punctatus* dans les génomes atlantiques ait participé à la mise en place des barrières d'isolement reproductif entre la lignée atlantique et méditerranéenne de *D. labrax*.

Plusieurs mécanismes auraient pu conduire à la fixation de ces allèles originaires de *D. punctatus* dans la lignée atlantique de *D. labrax,* tout en contribuant à renforcer la barrière à l'introgression avec la lignée méditerranéenne. Le premier serait une introgression adaptative, si les allèles de *D. punctatus* sont avantageux dans l'environnement atlantique mais désavantageux en Méditerranée (Martin and Jiggins 2017). Dans ce cas, les allèles se fixent rapidement dans le fond génétique atlantique mais ne peuvent pas entrer en Méditerranée. Le deuxième est la fixation d'allèles de *D. punctatus* délétères par dérive génétique, qui est particulièrement forte dans les régions à faible recombinaison (Whitlock

et al. 2000). Cette fixation d'allèles délétères aurait pu être facilitée par les goulots d'étranglement qu'ont pu subir les populations pendant les périodes glaciaires (Hewitt 2000) ou l'effet d'hétérosis que peut générer l'introgression dans le fond génétique de *D. labrax* atlantique (Kim *et al.* 2018). Une fois introgressés en Méditerranée, ces allèles révèlent leurs effets délétères et génèrent de la dépression d'hybridation. Le troisième mécanisme, qui est probablement le plus parcimonieux, est celui de la résolution de conflits liés à la présence d'incompatibilités génétiques entre *D. labrax* et *D. punctatus* (Schumer *et al.* 2015). En effet, la fixation de DMI est quasiment inévitable quand deux populations divergent en allopatrie (Presgraves 2010). Ainsi, le contact entre *D. punctatus* et *D. labrax* a pu révéler un grand nombre de ces incompatibilités qui peuvent être résolues en fixant alternativement l'allèle de *D. punctatus* ou de *D. labrax*. Or si l'allèle fixé en Atlantique est celui de *D. punctatus*, alors l'incompatibilité n'existe plus entre *D. punctatus* et *D. labrax* mais entre la lignée atlantique et méditerranéenne du bar européen. Ce mécanisme peut être vu comme le transfert d'incompatibilités génétiques entre lignées évolutives, le bar atlantique ayant hérité d'une partie des allèles de *D. punctatus* que ne possède pas le loup méditerranéen recréant les DMI lors de leurs échanges génétiques.

Notre étude s'ajoute donc aux précédentes ayant mis en évidence le rôle de l'hybridation dans la formation de barrières d'isolement reproductif (Runemark *et al.* 2018; Schumer *et al.* 2018; Eberlein *et al.* 2019). Cependant, la question du devenir des allèles atlantiques introgressés en Méditerranéen reste entière. En effet, les DMI s'étant résolue entre *D. punctatus* et la lignée atlantique de *D. labrax* on pourrait s'attendre à ce que le même processus se produise entre la lignée atlantique et méditerranéenne de bar européen, qui représente au niveau de ces îlots une version ancestrale de ce que pouvait être le génome de *D. labrax*. Ceci supprimerait alors l'effet d'isolement reproductif et pourrait permettre une réhomogénéisation totale des deux fonds génétiques. Le contact n'a peut-être simplement pas duré suffisamment longtemps pour permettre la résolution complète de toutes les incompatibilités. Cependant, les conditions démo-géographiques du contact entre les deux lignées de *D. labrax* sont peut-être très éloignées de celles du contact entre *D. punctatus* et la lignée atlantique de *D. labrax*. Dans ces conditions, la résolution des incompatibilités n'est peut-être pas possible et l'isolement reproductif pourrait alors se maintenir voir se renforcer entre les deux lignées. Des analyses complémentaires sont donc nécessaires pour comprendre quel mécanisme a permis la fixation des allèles de *D. punctatus* en Atlantique. L'étude détaillée des patrons d'évolution moléculaire des gènes aux cœurs des îlots génomiques résistants à l'introgression pourrait permettre de répondre en partie à cette question (cf chapitre 3)

Références

Eberlein C., M. Hénault, A. Fijarczyk, G. Charron, M. Bouvier, *et al.*, 2019 Hybridization is a recurrent evolutionary stimulus in wild yeast speciation. Nature Communications 10: 923. https://doi.org/10.1038/s41467-019-08809-7

Green R. E., J. Krause, A. W. Briggs, T. Maricic, U. Stenzel, *et al.*, 2010 A Draft Sequence of the Neandertal Genome. Science 328: 710–722. https://doi.org/10.1126/science.1188021

Hewitt G., 2000 The genetic legacy of the Quaternary ice ages. Nature 405: 907–913. https://doi.org/10.1038/35016000

Kim B. Y., C. D. Huber, and K. E. Lohmueller, 2018 Deleterious variation shapes the genomic landscape of introgression. PLOS Genetics 14: e1007741. https://doi.org/10.1371/journal.pgen.1007741

Martin S. H., K. K. Dasmahapatra, N. J. Nadeau, C. Salazar, J. R. Walters, *et al.*, 2013 Genome-wide evidence for speciation with gene flow in Heliconius butterflies. Genome Res. 23: 1817–1828. https://doi.org/10.1101/gr.159426.113

Martin S. H., and C. D. Jiggins, 2017 Interpreting the genomic landscape of introgression. Current Opinion in Genetics & Development 47: 69–74. https://doi.org/10.1016/j.gde.2017.08.007

Meyer M., M. Kircher, M.-T. Gansauge, H. Li, F. Racimo, *et al.*, 2012 A High-Coverage Genome Sequence from an Archaic Denisovan Individual. Science 338: 222–226. https://doi.org/10.1126/science.1224344

Presgraves D. C., 2010 The molecular evolutionary basis of species formation. Nature Reviews Genetics 11: 175–180. https://doi.org/10.1038/nrg2718

Rosenzweig B. K., J. B. Pease, N. J. Besansky, and M. W. Hahn, 2016 Powerful methods for detecting introgressed regions from population genomic data. Mol Ecol 2387–2397. https://doi.org/10.1111/mec.13610

Runemark A., C. N. Trier, F. Eroukhmanoff, J. S. Hermansen, M. Matschiner, *et al.*, 2018 Variation and constraints in hybrid genome formation. Nature Ecology & Evolution 2: 549. https://doi.org/10.1038/s41559-017-0437-7

Schumer M., R. Cui, G. G. Rosenthal, and P. Andolfatto, 2015 Reproductive Isolation of Hybrid Populations Driven by Genetic Incompatibilities. PLOS Genetics 11: e1005041. https://doi.org/10.1371/journal.pgen.1005041

Schumer M., C. Xu, D. L. Powell, A. Durvasula, L. Skov, *et al.*, 2018 Natural selection interacts with recombination to shape the evolution of hybrid genomes. Science eaar3684. https://doi.org/10.1126/science.aar3684

Skov L., R. Hui, V. Shchur, A. Hobolth, A. Scally, *et al.*, 2018 Detecting archaic introgression using an unadmixed outgroup. PLOS Genetics 14: e1007641. https://doi.org/10.1371/journal.pgen.1007641

Whitlock M. C., Pä. K. Ingvarsson, and T. Hatfield, 2000 Local drift load and the heterosis of interconnected populations. Heredity 84: 452–457. https://doi.org/10.1046/j.1365-2540.2000.00693.x

bioRxiv preprint first posted online May. 17, 2019; doi: http://dx.doi.org/10.1101/641829. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

The contribution of ancient admixture to reproductive isolation between European sea bass lineages

Short Title: *Ancient admixture in sea bass speciation*

Maud Duranton[1*], François Allal[2], Sophie Valière[3], Olivier Bouchez[3], François Bonhomme[1] and Pierre-Alexandre Gagnaire[1]

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

[2] MARBEC, Université de Montpellier, Ifremer-CNRS-IRD-UM, Palavas-les-Flots, France

[3]INRA, US 1426, GeT-PlaGe, Genotoul, Castanet-Tolosan, France.

[*]Corresponding author

bioRxiv preprint first posted online May. 17, 2019; doi: http://dx.doi.org/10.1101/641829. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

Abstract:

Understanding how new species arise through the progressive establishment of reproductive isolation barriers between diverging populations is a major goal in Evolutionary Biology. One important result of speciation genomics studies is that the genomic regions involved in reproductive isolation frequently harbor anciently diverged haplotypes that predate the reconstructed history of species divergence. The possible origins of these old alleles remain highly debated, since they relate to contrasted mechanisms of speciation that are not fully understood yet. In the European sea bass (*Dicentrarchus labrax*), the genomic regions involved in reproductive isolation between Atlantic and Mediterranean lineages are enriched for anciently diverged alleles of unknown origin. Here, we used haplotype-resolved whole-genome sequences to test whether divergent haplotypes could have originated from a closely related species, the spotted sea bass (*Dicentrarchus punctatus*). We found that an ancient admixture event between *D. labrax* and *D. punctatus* is responsible for the presence of shared derived alleles that segregate at low frequencies in both lineages of *D. labrax*. An exception to this was found within regions involved in reproductive isolation between the two *D. labrax* lineages. In those regions, archaic tracts originating from *D. punctatus* locally reached high frequencies or even fixation in Atlantic genomes but were almost absent in the Mediterranean. We showed that the ancient admixture event most likely occurred between *D. punctatus* and the *D. labrax* Atlantic lineage, while Atlantic and Mediterranean *D. labrax* lineages were experiencing allopatric isolation. Our results suggest that local adaptive introgression and/or the resolution of genomic conflicts provoked by ancient admixture have probably participated to the establishment of reproductive isolation between the two *D. labrax* lineages.

Author summary

Speciation is often viewed as a progressive accumulation of reproductive isolation barriers between two diverging lineages through the time. When initiated, the speciation process may however take different routes, sometimes leading to the erosion of an established species barrier or to the acquisition of new speciation genes transferred from another species boundary. Here, we describe such a case in the European sea bass. This marine fish species has split 300,000 years ago into an Atlantic and a Mediterranean lineage, which remained partially reproductively isolated after experiencing postglacial secondary contact. For unknown reasons, genomic regions involved in reproductive isolation between lineages have started to diverge well before the split. We here show that diverged alleles were acquired by the Atlantic lineage from an ancient event of admixture with a parapatric sister species about 80,000 years ago. Introgressed foreign alleles that were locally driven

bioRxiv preprint first posted online May. 17, 2019; doi: http://dx.doi.org/10.1101/641829. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

35 to high frequencies in the Atlantic have subsequently resisted to introgression within the
36 Mediterranean during the postglacial secondary contact, thus contributing to increased reproductive
37 isolation between two sea bass lineages. These results support the view that reproductive isolation
38 barriers can evolve via reticulate gene flow across multiple species boundaries.

39

40 **Introduction**

41

42 Speciation is the evolutionary process that leads to the emergence of new species through the
43 progressive establishment of Reproductive Isolation (RI) barriers between diverging populations (1).
44 Identifying those barriers and understanding the eco-evolutionary context in which they evolved has
45 been at the core of the speciation genetics research program (2,3). Over the last decade, progresses
46 in sequencing technologies have allowed to gain important insights into the genetic basis of
47 reproductive isolation barriers through the study of genome-wide differentiation/divergence patterns
48 between closely related species (4–8). An important result of speciation genomics studies was that the
49 age of the alleles located within genomic regions involved in RI is often much older than the average
50 coalescent time computed across the whole genome. This finding indicates that the regions involved
51 in RI tend to be enriched for anciently diverged haplotypes. An example of this comes from the fixed
52 chromosomal inversions involved in RI between *Drosophila pseudoobscura* and *D. persimilis*, which
53 show higher divergence than collinear regions of the genome (9). Another case is provided by the large
54 genomic regions of ancient ancestry that have been found across the threespine stickleback's genome,
55 which are involved in RI between marine and freshwater populations (10,11). A third example, among
56 others (see Marques *et al.* (2019) for a review), was described in Darwin's finches, whereby genomic
57 regions showing increased divergence in several species pairs also display anciently diverged
58 haplogroups that originated before the species splits (13).

59 Different hypotheses can explain the origin and the maintenance of these highly divergent
60 haplotypes. First, polymorphism has possibly been maintained over the long term in the ancestral
61 population before being differentially sorted between the descendant lineages (14). This hypothesis
62 has been proposed to explain the excess of haplotype divergence in the aforementioned examples
63 (9,10,13). One mechanism that may explain the long-term maintenance of polymorphism is ancestral
64 population structure, that is, subdivision owing to barriers to gene flow in the ancestral population
65 (15). In addition to demography, balancing selection due to either frequency-dependent selection,
66 heterozygote advantage (overdominance) or heterogeneous selection in space or time (16) can also

bioRxiv preprint first posted online May. 17, 2019; doi: http://dx.doi.org/10.1101/641829. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

promote the maintenance of ancient polymorphisms. For instance, in Darwin's finches, balancing selection has been proposed to explain the maintenance of divergent haplogroups associated with beak shape, due to the selective advantage of rare beak morphologies, or changing environmental conditions inducing heterogeneous selection (13). An alternative explanation to the presence of anciently diverged alleles is admixture with a divergent lineage. Contemporary hybridization has long been recognized as a common phenomenon in plants and animals (17,18), and cases of ancient admixture are increasingly detected by genomic studies. One emblematic example is past admixture between modern humans and two extinct archaic hominin lineages, Neanderthal and Denisova (19–21). More recently, ancient introgression from the extinct cave bear has also been detected in the genomes of living brown bears (22). Therefore, past admixture is increasingly recognized as a source of anciently diverged alleles in contemporary genomes.

Understanding why and how divergent haplogroups tend to disproportionately contribute to the buildup of RI between nascent species remains, however, highly challenging. First, because retention of ancestral polymorphism and past admixture are notoriously difficult to distinguish and not mutually exclusive hypotheses to explain the presence of anciently diverged alleles (23–27). Furthermore, identifying the genomic regions that resist introgression is still a major obstacle to the detection of RI loci (28). These tasks are now facilitated by the direct assessment of local ancestry along individual genome sequences (29,30), thus paving the way for assessing the role of ancient admixture in speciation. Here, we use new haplotype-resolved whole-genome sequences to delineate the regions involved in RI between European sea bass lineages and understand the origin of the divergent haplogroups they contain.

The European sea bass (*Dicentrarchus labrax*) is a marine fish subdivided into two glacial lineages, which currently correspond to Atlantic and Mediterranean populations (31). These two lineages have diverged in allopatry for *c.a.* 300,000 years before experiencing a secondary contact since the last glacial retreat (32). Postglacial gene flow between the two lineages is strongly asymmetrical, mostly occurring from the Atlantic to the Mediterranean genetic background (32). This resulted in a spatial introgression gradient within the Mediterranean Sea, illustrated by a more than twofold higher Atlantic ancestry in the western (31%) compared to the eastern (13%) Mediterranean population (30). A detailed analysis of local ancestry tracts across Mediterranean and Atlantic sea bass genomes has provided direct evidence for highly heterogeneous rates of gene flow along most chromosomes (Duranton *et al.* 2018). This mosaic introgression pattern was attributed to the effect of multiple small effect RI loci mainly located in low-recombining regions that present particularly high values of nucleotidic divergence (d_{XY}). It is generally assumed that increased d_{XY} indicates the presence of haplotypes that started to diverge earlier than the rest of the genome. However, regions of

bioRxiv preprint first posted online May. 17, 2019; doi: http://dx.doi.org/10.1101/641829. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

increased divergence may simply have resisted gene flow during secondary contact, while haplotypes in the remainder of the genome got rejuvenated due to recombination. This later hypothesis, however has been rejected in the European sea bass using simulations accounting for both background selection and selection against introgressed tracts (30). Therefore, anciently diverged alleles are unlikely to have evolved within the 300,000 years divergence history inferred from genome-wide polymorphism data and are thus older. In the present study, we use new haplotype-resolved whole-genome sequences to accurately delineate regions involved in RI and investigate the mechanisms underlying their excess of divergence. We specifically test for past admixture with a closely related species using a new genome sequence from the parapatrically distributed spotted sea bass (*Dicentrarchus punctatus*). Our results show that gene flow occurred between *D. punctatus* and the Atlantic lineage of *D. labrax* about 80,000 years ago, resulting in a low background ancestry from *D. punctatus* in contemporary *D. labrax* genomes. By contrast, genomic regions involved in RI between the two *D. labrax* lineages generally display high frequencies of haplotypes derived from *D. punctatus* in the Atlantic, while these archaic tracts remain rare in the Mediterranean. This suggests that ancient admixture has played an important role in the evolution of RI between Atlantic and Mediterranean sea bass lineages, consistently with predictions from models of local adaptive introgression and selection against genetic incompatibilities.

Results

Phylogenomic analysis

We reconstructed the genetic relationships among the three Moronid species used in our study: the striped bass (*Morone saxatilis*), the spotted sea bass (*Dicentrarchus punctatus*) and the European sea bass (*Dicentrarchus labrax*), which is further subdivided into two partially reproductively isolated populations: the Atlantic and Mediterranean sea bass lineages. All of the 3,329 maximum-likelihood phylogenetic trees generated in non-overlapping 50kb windows showed the same topology, corresponding to the expected species tree (Figure 1A). However, when similar reconstructions were performed in 2 kb windows, 4.6% of conflicting genealogies were found with an excess of trees in which *D. punctatus* grouped with the Atlantic (2.87%) versus with the Mediterranean (1.68%) *D. labrax* lineage (Supplementary Figure 3). The relative branch lengths of the species tree largely reflected the mean nucleotide divergence (d_{XY}) measured between each pair of four species/lineages (Figure 1B). We found 4.5% of absolute sequence divergence between the outgroup *M. saxatilis* and the two *Dicentrarchus* species. Divergence between *D. labrax* and *D. punctatus* (0.55%) was more than five times higher than divergence between Atlantic and Mediterranean *D. labrax* lineages (0.1%), consistently with previous estimates (30,32). We found a slightly higher divergence between *D. punctatus* and the eastern Mediterranean (0.56%) compared to the Atlantic *D. labrax* lineage (0.53%).

bioRxiv preprint first posted online May. 17, 2019; doi: http://dx.doi.org/10.1101/641829. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

Within *D. labrax*, divergence to the Atlantic population was higher for the eastern (0.1%) compared to the western Mediterranean population (0.09%) (consistent with the PCA, Supplementary Figure 2), as expected due to gene flow between Atlantic and Mediterranean *D. labrax* lineages (30,32).

***Test for foreign introgression within* D. labrax**

Chromosomal patterns of absolute sequence divergence (d_{XY}) between the Atlantic and Mediterranean lineages of *D. labrax* (Fig 2A and Supplementary Figure 4A) showed highly heterogeneous divergence along the genome, as reported in previous studies (30,32). To determine if local excesses of d_{XY} can be explained by past admixture with another lineage, we first looked for gene flow between *D. labrax* lineages and *D. punctatus* using the ABBA-BABA test. Some genomic regions showed particularly high values of the f_D statistics, thus reflecting locally elevated ancestry from *D. punctatus* within the *D. labrax* Atlantic lineage (Figure 2B and Supplementary Figure 4B red curve). By contrast, when the f_D statistics was used to measure local *D. punctatus* ancestry within *D. labrax* Mediterranean populations, low and relatively homogeneous introgression patterns were found across the entire genome (Figure 2B and Supplementary Figure 4B blue and green curves). This finding thus indicates highly heterogeneous introgression of spotted sea bass alleles within the Atlantic *D. labrax* lineage, and comparatively lower introgression within the Mediterranean lineage.

We also searched for the presence of archaic introgressed tracts in *D. labrax* genomes. A relatively low fraction of archaic tracts ($F_{archaic}$) was found along the genome in both Atlantic (4.85% in non-RI islands) and Mediterranean (2.73% in non-RI islands) *D. labrax* individuals (Figure 2C and Supplementary Figure 4C). In some regions, however, $F_{archaic}$ was particularly high in the Atlantic (i.e. >30%) compared to the Mediterranean lineage. Interestingly, those regions also presented the highest f_D values (Figure 2B red curve), and there was a highly significant positive correlation between f_D and $F_{archaic}$ in Atlantic *D. labrax* genomes (Spearman's rho = 0.281***). These results thus support the hypothesis that the detected archaic segments that locally reach high frequencies in some regions of Atlantic *D. labrax* genomes have been inherited from *D. punctatus* at some time in the past. Furthermore, regions of particularly increased *D. punctatus* ancestry also showed the highest absolute divergence values between Atlantic and Mediterranean *D. labrax* lineages, with positive genome-wide correlations being found with d_{XY} for both f_D (Spearman's rho = 0.281***) and $F_{archaic}$ (Spearman's rho = 0.531***). Lastly, we used the RND_{min} statistics to detect chromosomal variations in ancient introgression. Values of RND_{min} measured between *D. punctatus* and the Atlantic *D. labrax* lineage were low and relatively constant along chromosomes (Figure 2D and 4D red curves), indicating widespread (although locally rare) introgression across the genome. By contrast, RND_{min} was higher

bioRxiv preprint first posted online May. 17, 2019; doi: http://dx.doi.org/10.1101/641829. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

and highly variable when measured with the Mediterranean *D. labrax* populations (Figure 2D and 4D blue and green curves), indicating that introgression from *D. punctatus* is absent or nearly absent in some genomic regions of the Mediterranean lineage. These regions, that seem resistant to *D. punctatus* introgression in Mediterranean *D. labrax* genomes, also showed elevated values of $F_{archaic}$ (genome-wide Spearman's rho = 0.472***) and f_D (genome wide spearman's rho = 0.223***) in Atlantic genomes, along with increased d_{XY} between Atlantic and Mediterranean *D. labrax* lineages (genome-wide Spearman's rho = 0.717***). These results thus indicate the existence of outlying patterns of *D. punctatus* ancestry in the most divergent genomic regions between *D. labrax* lineages, due to respectively increased and decreased frequencies of anciently introgressed tracts in the Atlantic and Mediterranean lineages, compared to the background level.

Finally, our HMM approach allowed categorizing 70,738 SNP that are likely associated with RI islands between the two *D. labrax* lineages (Figure 2E and Supplementary Figure 4E). We found a good concordance between the positions of RI islands identified with the SNP and window-based methods, although the former allowed us to detect narrower RI-associated regions with a higher resolution (Supplementary Figure 5C and F). As expected, all these regions displayed increased levels of ancient *D. punctatus* introgression in the Atlantic but decreased *D. punctatus* ancestry in the Mediterranean (Figure 2), thus strengthening the association of RI-islands to differential rates of archaic ancestry.

***Estimation of the time since introgression between* D. punctatus *and* D. labrax**

We estimated the timing of past gene flow between *D. punctatus* and *D. labrax* by first comparing the length distribution of *D. punctatus* tracts introgressed into Atlantic *D. labrax* genomes to that of Atlantic *D. labrax* tracts introgressed into western Mediterranean *D. labrax* genomes (Figure 3A). The two distributions showed similar shapes although *D. punctatus* tracts were on average almost ten-time shorter ($\bar{L}_{punctatus}$ = 5,513 kb) than Atlantic *D. labrax* tracts ($\bar{L}_{labrax}$ = 52,026 kb). *D. punctatus* tracts were also less abundant in almost all length classes except for the shortest tracts (Figure 3A). We estimated the average time since introgression for both distributions as $t_{\text{labrax - punctatus}} = \frac{1}{((1-0.096)\cdot 3.693e^{-8}\cdot 5513)} + 1$ and $t_{\text{Atlantic - Mediterranean}} = \frac{1}{((1-0.341)\cdot 3.23e^{-8}\cdot 52026)} + 1$, which placed the contact between *D. punctatus* and *D. labrax* approximately 6 times earlier than the one between the two *D. labrax* lineages. Using the age of secondary contact previously estimated between Atlantic and Mediterranean sea bass lineages (i.e. 11,500 years, Tine et al. 2014; Duranton et al. 2018) as a calibration time-point, ancient gene flow between the two species was dated to *ca.* 70,000 years ago. Secondly, we converted the estimated values of the transition parameter (p) of the HMM model used to detect archaic introgressed tracts to estimate one value of T_{admix} for each chromosome

bioRxiv preprint first posted online May. 17, 2019; doi: http://dx.doi.org/10.1101/641829. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

(Supplementary Table 2). From the obtained time distribution (Figure 3B), we estimated the most probable time of ancient admixture to T_{admix} = 91,149 ($CI_{90\%}$ = [85,831 ; 110,645]) years.

The frequency of D. punctatus *derived mutations in* D. labrax

We used the conditioned site frequency spectrum between Atlantic and Mediterranean lineages as a way to represent how derived *D. punctatus* alleles segregate in *D. labrax*. For SNPs that were not associated to RI-islands by the HMM approach, the one-dimensional distribution of allele frequencies (*CSFS*) was highly similar between Atlantic and Mediterranean *D. labrax* lineages, showing a bimodal shape with few intermediate frequency variants (Figure 4A). Most *D. punctatus* derived alleles were present at either low or high frequencies, with ancestral mutations almost fixed in both *D. labrax* lineages being about 100 times more abundant than *D. punctatus* derived mutations almost fixed in both *D. labrax* lineages in the *CJSFS* (Figure 4C). This result showed that the combined effects of incomplete lineage sorting and introgression during species divergence has resulted in very similar amounts of *D. punctatus* derived mutations between Atlantic and Mediterranean *D. labrax* lineages. By contrast, SNPs found to be associated with RI islands showed a large excess of *D. punctatus* derived alleles that were fixed or almost fixed in the Atlantic population, while segregating at low frequencies in the Mediterranean populations (Figure 4B and D). This remained true whatever the Mediterranean population (east, west or both) considered in the analysis (Supplementary Figure 7). The excess of high-frequency *D. punctatus* derived mutations in the Atlantic sea bass lineage was also clearly visible in the reversal of the *CSFS* in RI islands compared to non-RI regions (Figure 4A and B). Therefore, differential introgression of *D. punctatus* derived mutations in RI islands is most likely due to their direct role in reproductive isolation, rather than a delayed post-glacial rehomogenization due to already-existing genetic barriers between *D. labrax* lineages in these regions.

Discussion

Recent speciation genomics studies have revealed that genomic regions involved in RI often contain anciently diverged alleles (e.g. Meier *et al.* 2017; Han *et al.* 2017; Nelson and Cresko 2018). One of the competing hypotheses to explain their origin is ancient admixture with an already diverged lineage. Our main objective here was to determine if such a scenario could explain the excess of divergence observed in RI regions between Atlantic and Mediterranean *D. labrax* lineages (30). To achieve this goal, we used different complementary approaches that collectively provided strong support for ancient introgression from the sister species *Dicentrarchus punctatus*. Despite low divergence (d_{XY} =

bioRxiv preprint first posted online May. 17, 2019; doi: http://dx.doi.org/10.1101/641829. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

0.55%), partially overlapping range distributions and interfertility in artificial crosses (48), contemporary hybridization has not been observed in the wild between *D. labrax* and *D. punctatus* (Tine et al. 2014). We here show that interspecies admixture has likely happened earlier in the past, bringing new key elements to understand the complex evolutionary history of unachieved speciation between Atlantic and Mediterranean sea bass lineages.

Extent of ancient admixture

Overall, the average fraction of contemporary genomes derived from ancient admixture was lower than 6% (i.e. 5.39% in the Atlantic and 2.82% in the Mediterranean lineage), which is only slightly higher than the estimated persistence of archaic ancestry in humans and brown bears (22,49). Whether these low background levels reflect a relatively limited contribution of genetic material from *D. punctatus* during admixture, or the impact of long-term selection against admixed foreign ancestry (29,50,51) was out of the scope of this study. Instead, we focused on understanding the marked excess of shared derived mutations found between *D. punctatus* and the Atlantic compared to the Mediterranean *D. labrax* lineage in RI-associated regions. This finding was strengthened by the locally increased frequency of archaic introgressed tracts found in Atlantic genomes within regions associated to RI with the Mediterranean lineage. Such locally elevated differences in the frequency of *D. punctatus* derived alleles explain the increased sequence divergence previously observed in RI islands between Atlantic and Mediterranean lineages (Duranton et al. 2018). Below, we consider potential limitations to the detection of archaic introgression from contemporary genomes, and the related challenge of dating ancient admixture. We then discuss how the genomic mosaicism of species ancestry may relate to different mechanisms potentially involved in European sea bass speciation.

Separating ancient introgression from shared ancestral variation

Distinguishing past introgression and shared ancestral variation from ABBA-BABA and f_D statistics can be difficult, especially in regions of reduced divergence (39). Therefore, the positive correlations observed between f_D, the inferred frequency of archaic segments, and d_{XY} provided good support that regions of high *D. punctatus* ancestry in the Atlantic are responsible for increased divergence between *D. labrax* lineages. Admittedly, past gene flow may also have occurred with another now extinct species rather than with *D. punctatus*, as it has been shown for other species (20,22,52). However, since *D. labrax* harbors shared derived alleles with *D. punctatus*, any alternative ghost donor lineage must have shared a long common history with the spotted sea bass.

Another potential issue with the tests performed to detect ancient admixture is that they often rely on differential introgression patterns between two candidate recipient populations (39,40). Therefore, these tests only enabled us to detect regions where the level of archaic introgression differs

bioRxiv preprint first posted online May. 17, 2019; doi: http://dx.doi.org/10.1101/641829. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

between Atlantic and Mediterranean *D. labrax* lineages. This problem could be particularly acute outside RI regions, where post-glacial gene flow between *D. labrax* lineages has almost completely rehomogenized allele frequencies (Tine et al. 2014; Duranton et al. 2018). To determine whether *D. punctatus* ancestry was simply absent or present but at similar levels in both lineages, we used the RND_{min} statistics that does not rely on the comparison of two populations (Rosenzweig et al. 2016). Low and nearly constant RND_{min} values indicated a widespread presence (although most of the time at low frequencies) of anciently introgressed tracts along Atlantic *D. labrax* genomes. By contrast, regions of elevated RND_{min} that coincided with the location of RI-islands revealed local resistance to introgression in Mediterranean *D. labrax* genomes. Therefore, both lineages contain *D. punctatus* introgressed tracts at relatively similar levels outside RI islands, which contrasys with strong archaic ancestry differences found within RI-islands.

Timing of ancient introgression

To understand why *D. punctatus* alleles were rare within RI genomic regions in the Mediterranean, we reconstructed the history of ancient admixture by estimating the time of contact between *D. punctatus* and *D. labrax*. The two different methods respectively inferred a contact taking place approximately 70,000 and 90,000 years ago. Although these two estimates slightly differ, they both place ancient admixture during the last glacial period (53), when Atlantic and Mediterranean *D. labrax* lineages were inferred to be geographically isolated (30,32). The current distribution range of *D. punctatus* partially overlaps with the southern part of the *D. labrax* distributional area in both the Atlantic (i.e. from southern Biscay to Morocco) and southern Mediterranean Sea (i.e. North African shores). It is thus likely that the latitudinal range shifts that occurred during quaternary ice ages (54) have favored hybridization by further increasing the range overlap between the two species, as they were coexisting in the Iberian or the north-western African Atlantic refugium (55). Once the two *D. labrax* lineages came into secondary contact after the last glacial maximum, the *D. punctatus* alleles already introgressed within Atlantic genomes could have readily introgressed Mediterranean genomes. This hypothesis was supported by the observed gradient of decreasing *D. punctatus* ancestry from the Atlantic to the eastern Mediterranean lineage, which mirrored the gradient in Atlantic ancestry generated by the post-glacial secondary contact (30). The fact that *D. punctatus* tracts have most probably introgressed the Mediterranean lineage secondarily, indicates that ancient hybridization has only occurred in the Atlantic during the last glacial period. A possible explanation is the absence of sympatry between *D. punctatus* and *D. labrax* within the Mediterranean during the last glacial period. A missing piece of the reconstructed historical scenario remains with respect to the role of *D. punctatus* alleles in RI.

bioRxiv preprint first posted online May. 17, 2019; doi: http://dx.doi.org/10.1101/641829. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

Causative role of high-frequency *D. punctatus* alleles in RI-islands

If most of the currently observed RI-islands between *D. labrax* lineages were already existing before ancient admixture with *D. punctatus*, such genetic barriers would have impeded the introgression of *D. punctatus* alleles within the Mediterranean lineage (30). However, they would not account for increased frequencies of *D. punctatus* derived alleles within RI-islands in the Atlantic lineage. The fact that, in Atlantic *D. labrax*, regions associated to RI exhibited closely fixed *D. punctatus* derived alleles that comparatively occurred at low frequencies elsewhere in the genome strongly supports their direct role in the establishment of RI. This finding thus indicates that *D. punctatus* alleles have been first locally driven to high frequencies in the Atlantic *D. labrax* lineage, while being secondarily prevented from introgression within the Mediterranean lineage.

Why anciently introgressed alleles contribute to RI?

Locally adaptive introgression

Understanding the underlying evolutionary mechanisms through which admixture has contributed to the buildup of reproductive isolation remains highly challenging (56,57). One evolutionary force that can drive an allele to fixation is local positive selection. *D. punctatus* alleles may have fixed in the Atlantic *D. labrax* lineage following admixture because they provided a selective advantage in the Atlantic environment compared to ancestral *D. labrax* alleles, a process called adaptive introgression (58). Several studies have revealed that the acquisition of adaptive phenotypes can be done through hybridization, such as altitude adaptation in humans (59), mimicry in *Heliconius* butterflies (60) or among others, seasonal camouflage in the snowshoe hares (61). Indeed, adaptive introgression allows the rapid transfer of linked variants that have already been tested by natural selection in their original environment, thus facilitating local adaptation (62). Therefore, it is theoretically possible that the Atlantic *D. labrax* lineage has received from *D. punctatus* advantageous alleles in the Atlantic environment that revealed to be deleterious in the Mediterranean Sea. Nevertheless, adaptive introgression is usually difficult to prove since it can be confounded with other processes such as uncoupling of an incompatibility from a multilocus genetic barrier (Fraïsse *et al.* 2014). Furthermore, it has been argued that adaptive introgression cannot play an important role in reproductive isolation, because unconditionally favorable alleles spread easily between diverging lineages until RI is nearly complete (65).

Fixation-compensation of deleterious mutations

bioRxiv preprint first posted online May. 17, 2019; doi: http://dx.doi.org/10.1101/641829. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

Another evolutionary force that may have driven *D. punctatus* derived alleles to fixation is genetic drift, which can induce the fixation of deleterious mutations and thus increase mutation load (66). When gene flow occurred between *D. punctatus* and *D. labrax* during the last glacial period, populations of each species were probably experiencing bottlenecks (54), which decreased the efficiency of selection and enhanced the probability to fix deleterious mutations by drift. Weakly deleterious *D. punctatus* alleles may therefore have introgressed and fixed within the *D. labrax* Atlantic population. Another related mechanism that may have influenced the outcome of hybridization is associative overdominance, due to the masking of recessive deleterious mutations in admixed genotypes (Whitlock et al. 2000; Bierne *et al.* 2002). Heterosis can locally increase the introgression rate of foreign alleles, even if interbreeding populations have similar amounts of deleterious variation (68). Therefore, heterosis may have favored the introgression of weakly deleterious *D. punctatus* variants in a bottlenecked Atlantic *D. labrax* lineage. Subsequently, when Atlantic and Mediterranean *D. labrax* lineages reconnected following postglacial recolonizations, expanding populations would have been sufficiently large to reveal the deleterious effects of the introgressed alleles, generating hybrid depression and hybridization load (29,69). Furthermore, the Atlantic population may have had enough time to evolve compensatory mutations (70), which could have become substrate for increased RI. The fact that most genomic regions involved in RI between *D. labrax* lineages exhibit low recombination rates (Tine *et al.* 2014; Duranton et al. 2018) could indicate a role of slightly deleterious alleles in RI, since selection is less efficient when linkage is strong.

Reciprocal sorting of DMIs

Reproductive isolation may also have evolved through the resolution of genetic conflicts resulting from the contact between two diverged populations (71,72). Because each population has almost inevitably fixed new adaptive or nearly neutral variants that reveal incompatible when combined in hybrid genomes (73), Bateson-Dobzhansky-Muller incompatibilities (BDMIs) are recognized as a common substrate for speciation (2). A genomic conflict induced by a two-locus BDMI can be resolved by fixing one of either parental alleles. In a hybrid population generated by an equal mixture of individuals from both parental populations, there is a 50% chance of fixing either parental combination (71). Therefore, the resolution of multiple BDMIs in an admixed population offers ample opportunity to reciprocally resolve independent BDMIs with respect to the origin of the parental allelic combination, which results in RI from both parental populations. Even in the presence of skewed initial admixture proportions, fixation of the minor parent combination can still happen with a sufficient number of BDMIs (71). Therefore, the resolution of genetic conflicts between *D. punctatus* and *D. labrax* alleles in the Atlantic lineage may have induced the fixation of *D. punctatus* alleles at some incompatibility loci. Upon contact between Atlantic and Mediterranean *D. labrax* lineages, fixed *D. punctatus* alleles may have recreated

bioRxiv preprint first posted online May. 17, 2019; doi: http://dx.doi.org/10.1101/641829. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

the BDMIs, thus contributing to RI. This non-adaptive speciation model due to selection against genetic incompatibilities has the advantage to explain both the fixation of *D. punctatus* alleles within the *D. labrax* Atlantic population, and their incompatibility with the Mediterranean lineage. Verbally, it can be seen as a case whereby speciation reversal between lineages A and B contributes to strengthen RI between lineages B and C through the transfer of incompatibilities between two porous species boundaries.

Conclusion

To conclude, our results show that divergent haplotypes that were introgressed from *D. punctatus* about 80,000 year ago have contributed to the strengthening of nascent RI between Atlantic and Mediterranean *D. labrax* lineages. The resulting genomic architecture of RI between contemporary *D. labrax* lineages is thus constituted by a mosaic of fixed blocks of different ancestries, that is, a mixture of genetic barriers inherited from the own *D. labrax* divergence history and the contribution of ancient admixture. Although additional analyses will be needed to fully understand which process has driven the fixation of *D. punctatus* alleles within Atlantic genomes, the resolution of genetic conflicts between *D. punctatus* and *D. labrax* seems the most parsimonious hypothesis (Schumer *et al.* 2015; Blanckaert and Bank 2018). This speciation mechanism can be thought of as a transfer of incompatibilities between two species boundaries, from the strongest to the weakest barrier, which is eventually strengthened by the displacement of genetic conflicts inherited from an ancient episode of admixture. Our finding adds to previous reports showing that postglacial and recent hybridization events have played a role in the buildup of RI between admixed and parental lineages by generating similar genomic mosaics of ancestries (29,74,75). The contribution of ancient admixture in European sea bass speciation suggests that significantly older admixture events, which may have left cryptic signatures in contemporary genomes, can be involved in seemingly recent speciation histories.

Material and methods

Whole-genome resequencing and haplotyping

We sequenced the whole genome of one *Dicentrarchus punctatus* individual from the Atlantic Ocean (Gulf of Cadiz, PUN) and 59 new *Dicentrarchus labrax* individual genomes. Fifty-two of them were wild individuals captured from the Atlantic Ocean (English Channel, 10 males $♂_{AT}$), the western Mediterranean Sea (Gulf of Lion, 14 females $♀_{WME}$ and 9 males $♂_{WME}$) and the eastern Mediterranean Sea (Turkey, 10 males $♂_{NEM}$ and Egypt, 9 males $♂_{SEM}$). Some of these specimens were involved in

bioRxiv preprint first posted online May. 17, 2019; doi: http://dx.doi.org/10.1101/641829. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

406 experimental crosses to generate first generation hybrids. Seven F1 hybrids obtained from 7 different
407 biparental families (pedigree $\male_{AT}$ x $\female_{WME}$) were also submitted to whole-genome sequencing. All captive
408 breeding procedures were performed at Ifremer's experimental aquaculture facility (agreement for
409 experiments with animals: C 34-192-6), where fish were reared in normal aquaculture conditions in
410 agreement with the French decree no. 2013-118 (1 February 2013 NOR:AGRG1231951D).
411 Whole genome sequencing libraries were prepared separately for each individual using either
412 the Illumina TruSeq DNA PCR-Free (40 individuals) or the TruSeq DNA Nano protocol (20 individuals),
413 depending on DNA concentration (Supplementary Table 1). Pools of 5 individually barcoded libraries
414 were then sequenced on 12 separate lanes of an Illumina HiSeq3000 using 2x150bp PE reads at the
415 GeT-PlaGe Genomics platform (Toulouse, France). Thirty-three individuals were sequenced twice due
416 to insufficient amounts of sequence reads obtained in the first run (Supplementary Table 1). For each
417 individual, the alignment of PE reads to the sea bass reference genome (32) was performed using BWA-
418 mem v0.7.5a (33) with default parameters. Duplicate reads were marked using Picard version 1.112
419 before being removed, producing a mean coverage depth of 33.8X per individual (Supplementary
420 Figure 1). We then followed GATK's (version 3.3-0-g37228af) best practice pipeline for individual
421 variant calling (using HaplotypeCaller), to joint genotyping, genotype refinement and variant filtering
422 (using Filter Expression: QD<10; MQ<50; FS>7; MQRankSum<-1.5; ReadPosRankSum<-1.5). We used
423 the BQSR algorithm to recalibrate base quality scores using a set of high-quality variants identified in
424 a previous study (30), and to perform variant quality score recalibration using the VQSR algorithm.
425 Hard filtering was then applied to exclude low-quality genotypes with a GQ score < 30. For the 7
426 mother-father-offspring trios, we used family-based priors for genotype refinement. We obtained a
427 total of 14,579,961 SNPs after filtering for indels, missing data (using --max-missing-count 8) and
428 removing the mitochondrial and ungrouped scaffolds (chromosome UN) in VCFtools v0.1.11 (34).
429 We performed haplotype phasing in *D. labrax* after removing the *D. punctatus* individual and
430 merging the 59 newly sequenced genomes with the 16 genomes already obtained in Duranton *et al.*
431 (2018). Fifteen individuals that were involved in family crosses (i.e. newly sequenced or not already
432 phased in the previous study) were submitted to phasing-by-transmission using the
433 PhaseByTransmission algorithm in GATK with default parameters and a mutation rate prior of 10^{-8} for
434 *de novo* mutations. For all individuals, variants located on a same read pair were directly phased using
435 physical phasing information. Non-related *D. labrax* individuals were then statistically phased using
436 the reference-based phasing algorithm implemented in Eagle2 (version 2.4) (35). The 22 parents
437 phased with the phasing-by-transmission approach were used to build a European sea bass reference
438 haplotype library (F1 genomes were excluded since their haplotype information was redundant with
439 that of their parents), which was used in Eagle2 to improve statistical phasing. We finally filtered out

bioRxiv preprint first posted online May. 17, 2019; doi: http://dx.doi.org/10.1101/641829. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

SNPs that were not phased or not genotyped over all individuals (using --max-missing-count 0 and –phased in VCFtools), to generate a dataset of haplotype-resolved whole-genome sequences from 68 unrelated *D. labrax* individuals (14 AT, 31 WME, 11 SEM and 12 NEM), containing 5,074,249 phased SNPs without missing data. The genetic relationships of the newly sequenced genomes with respect to the 16 already available was evaluated with a Principal Component Analysis (Supplementary Figure 2). Although we detected a slight genetic differentiation between North and South eastern Mediterranean samples on the PCA (Supplementary Figure 2), we later determined that they present similar genome-wide average levels of Atlantic ancestry and introgressed tract length (18,148 bp for the South and 17,769 bp for the North, Supplementary Figure 5). Therefore, we regrouped these samples together within a single eastern Mediterranean population, similarly to Duranton et al. (2018).

Phylogenomic analyses

We used RAxML v.8.2.12 (36) to generate maximum-likelihood trees of Moronids genomes in non-overlapping 50 kb windows (including *Morone saxatilis*, *D. punctatus* and the Atlantic and Mediterranean *D. labrax* lineages). Ancient admixture is expected to generate discordant trees among genomic windows. However, if admixture is ancient, introgressed tracts may be too short to influence the phylogenetic signal in 50 kb windows. Therefore, we did the same analyses using a 2 kb window size to increase the resolution of local genealogies while keeping enough informative sites. In order to account for disparities among species' genome sequence datasets, we used only one individual haplome for each species/lineage for this analysis. The alignment of these four haplomes spanned 52% of the *D. labrax* genome, a fact largely due to the fragmentation of the *M. saxatilis* (SAX) genome that produced discontinuous local alignments to the *D. labrax* reference genome (30). In order to account for this fragmentation, we only analyzed windows with less than 10% missing data in local alignments. We obtained 3,329 and 155,155 trees under the GTRGAMMA model for analyses based on 50 and 2 kb windows, respectively. Trees generated in windows of similar size were then superposed using DensiTree v2.2.5 (37) for visualization. In order to provide indications for genome-wide average absolute sequence divergence between all pairs of species and lineages, we calculated d_{XY} with the same individual haplomes used for the RAxML analysis using MVFTools v5.1.2 (38) and averaged distance values calculated in non-overlapping 50kb windows.

***Tests for foreign introgression within* D. labrax**

We tested for admixture between *D. labrax* and another species using three different methods that capture complementary aspects of the data. Since *D. punctatus* is the only closely related species parapatrically distributed with *D. labrax*, we first tested for historical gene flow between these two species. To do so, we used the ABBA-BABA test (19,23) with *M. saxatilis* as the outgroup (O), *D.*

bioRxiv preprint first posted online May. 17, 2019; doi: http://dx.doi.org/10.1101/641829. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

475 *punctatus* as the potential donor species (P3) and the two *D. labrax* lineages as potential recipient
476 populations (P1 and P2). We used the dataset containing 14,579,961 SNPs from *D. punctatus* and
477 unphased *D. labrax* samples, and only kept sites that were available for both *M. saxatilis* and *D.*
478 *punctatus* in genome alignments, representing a total of 9,606,462 SNPs. This allowed testing for
479 different amounts of gene flow between P3 and P2, and P3 and P1, by comparing the number of
480 genealogies of type ((P1,(P2,P3),O) (*i.e.* ABBA genealogies) and ((P2,(P1,P3),O) (*i.e.* BABA genealogies).
481 An excess of shared derived alleles between the donor and one of the two recipient populations (i.e.
482 excess of ABBA over BABA genealogies, or vice versa) indicates gene flow from *D. punctatus* to *D.*
483 *labrax* population P2 or P1, respectively. Although the ABBA-BABA test is adequate to detect
484 introgression, the Patterson's *D* statistic that measures the imbalance between the two types of
485 genealogies $\left(D = \frac{sum(ABBA)-sum(BABA)}{sum(ABBA)+sum(BABA)}\right)$ is not appropriate to quantify introgression over small
486 genomic windows (39). Therefore, we used the f_D statistics ($f_D = \frac{s(P1,P2,P3,O)}{s(P1,PD,PD,O)}$) to estimate admixture
487 proportion between P2 et P3, were *s* = sum(ABBA-BABA) and *PD* corresponds to the most likely donor
488 population (i.e. the population with the higher frequency of the derived allele). In order to test for
489 admixture between *D. punctatus* and different populations of *D. labrax*, we made different tests using
490 successively the Atlantic (AT), eastern (EME combining SEM and NEM individuals), western (WME) or
491 the whole Mediterranean (MED) populations of *D. labrax* populations as P2 or P1. We used scripts
492 from Martin *et al.* (2015) to estimate the number of ABBA and BABA genealogies and the f_D statistics
493 in non-overlapping 50 kb windows along the genome, keeping only windows containing at least 500
494 SNPs.

495 Secondly, we used a method that allows identifying archaic introgressed tracts without using
496 an archaic reference genome for the donor species (40). The main advantage of this method is that it
497 makes no assumption on the identity of the donor species. Basically, it looks for local excesses of
498 private variants in a candidate recipient population by comparison to another non-admixed population
499 (40). In order to test for archaic introgression within the Atlantic *D. labrax* lineage, we identified
500 variants that were not shared with the eastern Mediterranean population, and conversely to test for
501 introgression in the Mediterranean *D. labrax* lineage. We only analyzed the eastern Mediterranean
502 population because the western Mediterranean is more strongly impacted by gene flow from the
503 Atlantic (30,32). We used the phased genomes dataset containing 5,074,249 SNPs, assuming a
504 constant mutation and call rate to run the model in 1000 bp windows along each chromosome. For
505 each window, the probability that an individual haplotype contains an archaic introgressed fragment
506 was estimated to identify introgressed windows with a posterior probability superior to 0.8 (40). We
507 then combined individual profiles of introgressed windows to estimate the fraction of introgressed
508 archaic tracts in each population ($F_{archaic}$), as the fraction of haplotypes for which a window was

bioRxiv preprint first posted online May. 17, 2019; doi: http://dx.doi.org/10.1101/641829. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

identified as introgressed. The inferred fraction of introgressed archaic tracts was finally averaged in non-overlapping 50 kb windows along the genome.

Finally, we used the RND_{min} statistics, which is sensitive to rare introgression while being robust to mutation rate variation across the genome (41). The main advantage of this statistic is that, unlike the two former methods, it does not rely on the comparison of two recipient populations that differ in their level of introgression. The RND_{min} corresponds to the ratio of the minimal pairwise distance between haplotypes from the potential donor and recipient populations (d_{min}) over the average divergence of those populations to an outgroup species (d_{out}). If gene flow has occurred genome-wide, then locally elevated RND_{min} values indicate regions where introgression has been limited or absent. We used MVFTools v5.1.2 (38) to measure d_{min} between *D. punctatus* and different population of *D. labrax* (AT, EME, WME, MED). For this analysis, we used 4,943,488 SNPs polymorphic sites that were phased within *D. labrax* and non-missing in *D. punctatus*. On one hand, this dataset excludes a large number of variants that are differentially fixed between *D. punctatus* and *D. labrax*, and therefore underestimates the real level of divergence between *D. punctatus* and *D. labrax.* On the other hand, excluding diagnostic SNPs rendered the test more sensitive to the detection of ancient introgression, since the accumulation of divergence after introgression only adds noise to chromosomal variations in RND_{min}. We estimated d_{out} by averaging the divergence measures between *M. saxatilis* and the two *Dicentrarchus* species. All values were averaged in non-overlapping 50 kb windows along the genome.

Detection of introgressed tracts between Atlantic and Mediterranean* D. labrax *lineages

In order to test whether ancient introgression has influenced genomic patterns of post-glacial gene flow between Atlantic and Mediterranean *D. labrax* lineages, we mapped Atlantic tracts introgressed into Mediterranean genomes and conversely. Local ancestry inference was performed with Chromopainter v0.04 (42), an HMM-based program that estimates the probability of Atlantic and Mediterranean ancestry for each variable position along each haplome. To do so, it compares a focal haplotype to reference populations composed of non-introgressed Atlantic and Mediterranean haplotypes. Since Mediterranean individuals are introgressed to various extents by Atlantic alleles, we used a pure Mediterranean reference population reconstituted by Duranton *et al.* (2018) with the same model parameters. We then identified the starting and ending position of each introgressed tract within both Atlantic and Mediterranean genetic backgrounds by analyzing the ancestry probability profiles inferred by Chromopainter, following the same methodology as in Duranton *et al.* (2018). Identified tracts in each *D. labrax* population (Supplementary Figure 5) were then combined to estimate the fraction of introgressed tracts (F_{intro}) for each position along the genome, which was finally averaged in non-overlapping 50 kb windows.

bioRxiv preprint first posted online May. 17, 2019; doi: http://dx.doi.org/10.1101/641829. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

Delineation of RI regions between Atlantic and Mediterranean* D. labrax *lineages

We adapted the HMM approach developed by Hofer *et al.* (2012) to precisely delineate genomic regions involved in RI between Atlantic and Mediterranean *D. labrax* lineages. Genomic regions involved in RI between European sea bass lineages are characterized by elevated genetic differentiation and increased resistance to gene flow (30,32). Therefore, we combined both measures of F_{ST} (43,44) and resistance to introgression measured as the inverse of F_{intro} (28,30). To identify true RI islands in our HMM strategy, we thus used the ratio of F_{ST} over F_{intro} (i.e. the frequency of Atlantic tracts within western Mediterranean *D. labrax* genomes). Our rationale was that these regions should be associated with both high F_{ST} (Supplementary Figure 6A and D) and low F_{intro} values (Supplementary Figure 6B and E) (30), hence elevated F_{ST}/F_{intro} ratio values (Supplementary Figure 6C and F). We used the HMM approach to map RI at two different scales, a SNP-by-SNP (Supplementary Figure 6A-C) and a 50kb window scale, which was more suitable to delineate regions (Supplementary Figure 6D-F). We used VCFtools v0.1.15 (34) to estimate F_{ST} between the Atlantic and the western Mediterranean *D. labrax* lineage for each SNPs and every non-overlapping 50kb window along the genome. The HMM was designed with three different states corresponding to low (i.e. neutral genomic regions), intermediate (i.e. regions experiencing linked selection) and high F_{ST}/F_{intro} ratio values (i.e. regions involved in RI). The most likely state of each SNP/window was inferred by running the HMM algorithm chromosome by chromosome. Finally, we controlled for false discovery rate and retained only SNPs/windows with an FDR-corrected p-value inferior to 0.001 (43).

***Estimation of the time since foreign introgression within* D. labrax**

We used two different approaches to estimate the time since foreign introgression within *D. labrax*. First, we relied on the fact that the length of introgressed tracts is informative of the time elapsed since introgression. Recombination progressively shortens migrant tracts across generations following introgression into a new genetic background (45,46). Since we found a good correspondence between the inferred fraction of introgressed archaic tracts ($F_{intro\text{-}archaic}$) and f_D values (see results) using *D. punctatus* as a donor species, we used the length of archaic haplotypes that were identified with the method of Skov *et al.* (2018). Only archaic tracts found in Atlantic genomes within windows involved in RI between Atlantic and Mediterranean *D. labrax* lineages were considered, corresponding to 1310 windows of 50 kb. This choice was justified because archaic tract detection relies on a signal of differential introgression between two populations. Therefore, archaic tracts can be correctly identified and delimited only if they are present in one lineage (e.g. the Atlantic) but absent in the other (e.g. the Mediterranean), which was only the case in RI islands between Atlantic and Mediterranean *D. labrax* lineages (see results).

bioRxiv preprint first posted online May. 17, 2019; doi: http://dx.doi.org/10.1101/641829. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

Under simple assumptions, there is an analytical expectation for the average length of introgressed tracts ($\bar{L}$) as a function of the number of generations since introgression (t), the local recombination rate (r in Morgans per bp) and the proportion of admixture (f), which takes the form $\bar{L} = [(1-f)\, r\, (t-1)]^{-1}$ (26). We used this equation to estimate the age of admixture between *D. punctatus* and the Atlantic lineage of *D. labrax* ($t_{\text{labrax – punctatus}}$), as well as between the two lineages of *D. labrax* ($t_{\text{Atlantic – Mediterranean}}$). For each estimation, we used the average value of the retained windows. Hence, for $t_{\text{labrax – punctatus}}$: f = 0.096, r = 3.693e^{-8} M/bp (32) and $\bar{L}$ = 5,513 bp, and for $t_{\text{Atlantic – Mediterranean}}$: f = 0.341, r = 3.23e^{-8} M/bp, and $\bar{L}$ = 52,026 bp. Since we only considered a relatively small fraction of the genome to call archaic tracts, we could not obtain precise estimations of those parameters. Therefore, we estimated the age of contact between *D. punctatus* and *D. labrax* by reference to the age of the post-glacial secondary contact between Atlantic and Mediterranean lineages of *D. labrax*, which has been more precisely estimated to 2,300 generations using a larger fraction of the genome (30,32).

Secondly, we transformed the estimated transition parameter values of the HMM model used to detect archaic introgressed tracts using $p \approx T_{admix} \cdot 2 \cdot r \cdot L \cdot a$ (40). In this equation, p is the probability of transition from the *D. labrax* to the archaic ancestry state, T_{admix} represents the admixture time in generations, r the recombination rate in Morgan per bp, a the admixture proportion and L the size of the window (here L = 1000 bp). Parameter p was estimated separately for each chromosome by averaging over the values estimated per individual haplome. We finally estimated T_{admix} chromosome by chromosome using the average recombination rate and the fraction of archaic introgressed tracts of each chromosome (Supplementary Table 2). The time in generations was converted into years using a generation time of 5 years (32). We then obtained a distribution for T_{admix} across the 24 chromosomes, from which we identified the maximum and its 90% confidence interval by bootstrapping the distribution 10,000 times.

***Characterizing foreign ancestry tracts within* D. labrax**

We used Spearman's correlation test to evaluate relationships among $F_{\text{intro-archaic}}$, f_D, d_{XY} and RND_{min} statistics that relate to a series of predictive hypotheses. More specifically, if *D. punctatus* has anciently contributed to *D. labrax* in the Atlantic, the local abundance of archaic tracts inferred within Atlantic *D. labrax* genomes (F_{archaic}) should be positively correlated to f_D. Moreover, if the abundance of archaic tracts within Atlantic *D. labrax* explains the presence of anciently diverged alleles between Atlantic and Mediterranean sea bass lineages, d_{XY} should increase with the amount of ancient admixture. Finally, if regions involved in RI between Atlantic and Mediterranean sea bass lineages harbor reduced frequencies of *D. punctatus* derived tracts in the Mediterranean, a positive correlation is expected between RND_{min} measured between *D. punctatus* and Mediterranean *D. labrax* and ancient admixture from *D. punctatus* within the Atlantic.

bioRxiv preprint first posted online May. 17, 2019; doi: http://dx.doi.org/10.1101/641829. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

We then focused on SNP-level statistics to specifically address the frequency distributions of derived mutations from *D. punctatus* within *D. labrax* genomes, separately in the Atlantic and Mediterranean lineages. Since anciently introgressed alleles most likely originated from *D. punctatus* (see results), we used *D. labrax* polymorphic sites for which *M. saxatilis* harbors the ancestral and *D. punctatus* the derived state (*i.e.* ABBA-BABA informative sites) to characterize ancient introgression. For each of these SNPs, we measured the frequency of the *D. punctatus* derived allele separately in the Atlantic and Mediterranean *D. labrax* populations using VCFtools. We then separated SNPs associated to RI islands from those that were not associated to RI islands in the SNP-based HMM analysis to represent the site frequency spectrum of each *D. labrax* lineage, conditioned on *D. punctatus* being derived (*CSFS*). Finally, two conditioned joint site frequency spectra (*CJSFS*) were generated (i.e. for RI and non-RI SNPs) to represent the bi-dimensional SFS between Atlantic and Mediterranean *D. labrax* lineages, conditioning on sites that have the derived allele in *D. punctatus*. These analyses aimed at distinguishing two hypotheses with respect to the mechanisms underlying differential introgression of *D. punctatus* derived mutations in RI islands between Atlantic and Mediterranean *D. labrax*. Our first hypothesis was that anciently introgressed alleles are not directly involved in RI but simply maintained at different frequencies because genetic barriers between *D. labrax* lineages (i.e. unrelated to the history of ancient admixture) have impeded their post-glacial rehomogenization. In this case, we expected no excess of high-frequency *D. punctatus* derived mutations in RI islands compared to non-RI regions. Alternatively, under the hypothesis that anciently introgressed alleles are associated with reproductive isolation in sea bass, an excess of *D. punctatus* derived mutations almost fixed within RI islands in the Atlantic but nearly absent in the Mediterranean was expected compared to the alternate configuration (i.e. almost fixed in the Mediterranean and nearly absent in the Atlantic).

Ackowledgement

This work was co-founded by the GeneSea project (n° R FEA 4700 16 FA 100 0005) by the French Government and the European Union (EMFF, European Maritime and Fisheries Fund) at the "Appels à projets Innovants" managed by the FranceAgrimer Office and the ANR grant CoGeDiv (ANR-17-CE02-0006-01 to P.-A.G). This work was performed in collaboration with the GeT core facility, Toulouse, France (http://get.genotoul.fr), and was supported by France Génomique National infrastructure, funded as part of “Investissement d’avenir” program managed by Agence Nationale pour la Recherche (contract ANR-10-INBS-09). We also thank Ifremer's experimental aquaculture platform for the breeding and the rearing of the hybrid populations

bioRxiv preprint first posted online May. 17, 2019; doi: http://dx.doi.org/10.1101/641829. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

References

1. Coyne JA, Orr AH. Speciation. Sinauer Associates. massachusetts U.S.A.: Sunderland MA; 2004.

2. Presgraves DC. The molecular evolutionary basis of species formation. Nature Reviews Genetics. mars 2010;11(3):175-80.

3. Wolf Jochen B. W., Lindell Johan, Backström Niclas. Speciation genetics: current status and evolving approaches. Philosophical Transactions of the Royal Society B: Biological Sciences. 12 juin 2010;365(1547):1717-33.

4. Feder JL, Egan SP, Nosil P. The genomics of speciation-with-gene-flow. Trends in Genetics. 1 juill 2012;28(7):342-50.

5. Seehausen O, Butlin RK, Keller I, Wagner CE, Boughman JW, Hohenlohe PA, et al. Genomics and the origin of species. Nat Rev Genet. mars 2014;15(3):176-92.

6. Harrison RG, Larson EL. Heterogeneous genome divergence, differential introgression, and the origin and structure of hybrid zones. Mol Ecol. 1 juin 2016;25(11):2454-66.

7. Wolf JBW, Ellegren H. Making sense of genomic islands of differentiation in light of speciation. Nat Rev Genet. 14 nov 2016;18:87-100.

8. Ravinet M, Faria R, Butlin RK, Galindo J, Bierne N, Rafajlović M, et al. Interpreting the genomic landscape of speciation: a road map for finding barriers to gene flow. J Evol Biol. 1 août 2017;30(8):1450-77.

9. Fuller ZL, Leonard CJ, Young RE, Schaeffer SW, Phadnis N. Ancestral polymorphisms explain the role of chromosomal inversions in speciation. PLOS Genetics. 30 juill 2018;14(7):e1007526.

10. Colosimo PF, Hosemann KE, Balabhadra S, Villarreal G, Dickson M, Grimwood J, et al. Widespread Parallel Evolution in Sticklebacks by Repeated Fixation of Ectodysplasin Alleles. Science. 25 mars 2005;307(5717):1928-33.

11. Nelson TC, Cresko WA. Ancient genomic variation underlies repeated ecological adaptation in young stickleback populations. Evolution Letters. 2018;2(1):9-21.

12. Marques DA, Meier JI, Seehausen O. A Combinatorial View on Speciation and Adaptive Radiation. Trends in Ecology & Evolution [Internet]. 15 mars 2019 [cité 10 avr 2019]; Disponible sur: http://www.sciencedirect.com/science/article/pii/S0169534719300552

13. Han F, Lamichhaney S, Grant BR, Grant PR, Andersson L, Webster MT. Gene flow, ancient polymorphism, and ecological adaptation shape the genomic landscape of divergence among Darwin's finches. Genome Res [Internet]. 25 avr 2017 [cité 10 janv 2019]; Disponible sur: http://genome.cshlp.org/content/early/2017/04/25/gr.212522.116

14. Guerrero RF, Hahn MW. Speciation as a Sieve for Ancestral Polymorphism. Mol Ecol [Internet]. 2017 [cité 11 août 2017]; Disponible sur: http://onlinelibrary.wiley.com/doi/10.1111/mec.14290/abstract

bioRxiv preprint first posted online May. 17, 2019; doi: http://dx.doi.org/10.1101/641829. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

15. Slatkin M, Pollack JL. Subdivision in an Ancestral Species Creates Asymmetry in Gene Trees. Mol Biol Evol. 1 oct 2008;25(10):2241-6.

16. Charlesworth D. Balancing Selection and Its Effects on Sequences in Nearby Genome Regions. PLOS Genetics. 28 avr 2006;2(4):e64.

17. Abbott R, Albach D, Ansell S, Arntzen JW, Baird SJE, Bierne N, et al. Hybridization and speciation. J Evol Biol. 1 févr 2013;26(2):229-46.

18. Payseur BA, Rieseberg LH. A genomic perspective on hybridization and speciation. Mol Ecol. 1 juin 2016;25(11):2337-60.

19. Green RE, Krause J, Briggs AW, Maricic T, Stenzel U, Kircher M, et al. A Draft Sequence of the Neandertal Genome. Science. 7 mai 2010;328(5979):710-22.

20. Meyer M, Kircher M, Gansauge M-T, Li H, Racimo F, Mallick S, et al. A High-Coverage Genome Sequence from an Archaic Denisovan Individual. Science. 12 oct 2012;338(6104):222-6.

21. Pääbo S. The diverse origins of the human gene pool. Nature Reviews Genetics. 18 mai 2015;16:313-4.

22. Barlow A, Cahill JA, Hartmann S, Theunert C, Xenikoudakis G, Fortes GG, et al. Partial genomic survival of cave bears in living brown bears. Nature Ecology & Evolution. oct 2018;2(10):1563-70.

23. Durand EY, Patterson N, Reich D, Slatkin M. Testing for Ancient Admixture between Closely Related Populations. Mol Biol Evol. 1 août 2011;28(8):2239-52.

24. Eriksson A, Manica A. Effect of ancient population structure on the degree of polymorphism shared between modern human populations and ancient hominins. Proc Natl Acad Sci U S A. 28 août 2012;109(35):13956-60.

25. Welch JJ, Jiggins CD. Standing and flowing: the complex origins of adaptive variation. Molecular Ecology. 2014;23(16):3935-7.

26. Racimo F, Sankararaman S, Nielsen R, Huerta-Sánchez E. Evidence for archaic adaptive introgression in humans. Nat Rev Genet. juin 2015;16(6):359-71.

27. Theunert C, Slatkin M. Distinguishing Recent Admixture from Ancestral Population Structure. Genome Biol Evol. 1 mars 2017;9(3):427-37.

28. Cruickshank TE, Hahn MW. Reanalysis suggests that genomic islands of speciation are due to reduced diversity, not reduced gene flow. Mol Ecol. 1 juill 2014;23(13):3133-57.

29. Schumer M, Xu C, Powell DL, Durvasula A, Skov L, Holland C, et al. Natural selection interacts with recombination to shape the evolution of hybrid genomes. Science. 19 avr 2018;eaar3684.

30. Duranton M, Allal F, Fraïsse C, Bierne N, Bonhomme F, Gagnaire P-A. The origin and remolding of genomic islands of differentiation in the European sea bass. Nature Communications. 28 juin 2018;9(1):2518.

bioRxiv preprint first posted online May. 17, 2019; doi: http://dx.doi.org/10.1101/641829. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

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

32. Tine M, Kuhl H, Gagnaire P-A, Louro B, Desmarais E, Martins RS, et al. European sea bass genome and its variation provide insights into adaptation to euryhalinity and speciation. Nature Communications. 2014;5:5770.

33. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv: preprint [Internet]. 16 mars 2013 [cité 27 avr 2017]; Disponible sur: http://arxiv.org/abs/1303.3997

34. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call format and VCFtools. Bioinformatics. 8 janv 2011;27(15):2156-8.

35. Loh P-R, Danecek P, Palamara PF, Fuchsberger C, Reshef YA, Finucane HK, et al. Reference-based phasing using the Haplotype Reference Consortium panel. Nature Genetics. nov 2016;48(11):1443-8.

36. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 1 mai 2014;30(9):1312-3.

37. Bouckaert R, Heled J. DensiTree 2: Seeing Trees Through the Forest. bioRxiv. 8 déc 2014;012401.

38. Pease J, Rosenzweig B. Encoding Data Using Biological Principles: the Multisample Variant Format for Phylogenomics and Population Genomics. IEEE/ACM Transactions on Computational Biology and Bioinformatics. 2015;PP(99):1-1.

39. Martin SH, Davey JW, Jiggins CD. Evaluating the Use of ABBA–BABA Statistics to Locate Introgressed Loci. Mol Biol Evol. 1 janv 2015;32(1):244-57.

40. Skov L, Hui R, Shchur V, Hobolth A, Scally A, Schierup MH, et al. Detecting archaic introgression using an unadmixed outgroup. PLOS Genetics. 18 sept 2018;14(9):e1007641.

41. Rosenzweig BK, Pease JB, Besansky NJ, Hahn MW. Powerful methods for detecting introgressed regions from population genomic data. Mol Ecol. 1 mars 2016;2387-97.

42. Lawson DJ, Hellenthal G, Myers S, Falush D. Inference of Population Structure using Dense Haplotype Data. PLOS Genet [Internet]. 26 janv 2012 [cité 4 avr 2016];8(1). Disponible sur: http://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1002453

43. Hofer T, Foll M, Excoffier L. Evolutionary forces shaping genomic islands of population differentiation in humans. BMC Genomics. 22 mars 2012;13:107.

44. Marques DA, Lucek K, Meier JI, Mwaiko S, Wagner CE, Excoffier L, et al. Genomics of Rapid Incipient Speciation in Sympatric Threespine Stickleback. PLOS Genetics [Internet]. 29 févr 2016 [cité 11 mai 2017];12(2). Disponible sur: http://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1005887

45. Pool JE, Nielsen R. Inference of Historical Changes in Migration Rate From the Lengths of Migrant Tracts. Genetics. 1 févr 2009;181(2):711-9.

bioRxiv preprint first posted online May. 17, 2019; doi: http://dx.doi.org/10.1101/641829. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

46. Liang M, Nielsen R. The Lengths of Admixture Tracts. Genetics. 1 juill 2014;197(3):953-67.

47. Meier JI, Marques DA, Mwaiko S, Wagner CE, Excoffier L, Seehausen O. Ancient hybridization fuels rapid cichlid fish adaptive radiations. Nature Communications. 10 févr 2017;8:14363.

48. Ky C-L, Vergnet A, Molinari N, Fauvel C, Bonhomme F. Fitness of early life stages in F1 interspecific hybrids between Dicentrarchus labrax and D. punctatus. Aquat Living Resour. 1 janv 2012;25(1):67-75.

49. Sankararaman S, Mallick S, Patterson N, Reich D. The Combined Landscape of Denisovan and Neanderthal Ancestry in Present-Day Humans. Current Biology. 9 mai 2016;26(9):1241-7.

50. Harris K, Nielsen R. The Genetic Cost of Neanderthal Introgression. Genetics. 1 juin 2016;203(2):881-91.

51. Juric I, Aeschbacher S, Coop G. The Strength of Selection against Neanderthal Introgression. PLOS Genetics. 8 nov 2016;12(11):e1006340.

52. Gopalakrishnan S, Sinding M-HS, Ramos-Madrigal J, Niemann J, Castruita JAS, Vieira FG, et al. Interspecific Gene Flow Shaped the Evolution of the Genus Canis. Current Biology [Internet]. 18 oct 2018 [cité 24 oct 2018];0(0). Disponible sur: https://www.cell.com/current-biology/abstract/S0960-9822(18)31125-4

53. Snyder CW. Evolution of global temperature over the past two million years. Nature. 13 oct 2016;538(7624):226-8.

54. Hewitt G. The genetic legacy of the Quaternary ice ages. Nature. 22 juin 2000;405(6789):907-13.

55. Maggs CA, Castilho R, Foltz D, Henzler C, Jolly MT, Kelly J, et al. Evaluating Signatures of Glacial Refugia for North Atlantic Benthic Marine Taxa. Ecology. 2008;89(sp11):S108-22.

56. Schumer M, Rosenthal GG, Andolfatto P. How Common Is Homoploid Hybrid Speciation? Evolution. 2014;68(6):1553-60.

57. Schumer M, Rosenthal GG, Andolfatto P. What do we mean when we talk about hybrid speciation? Heredity. avr 2018;120(4):379.

58. Racimo F, Marnetto D, Huerta-Sánchez E. Signatures of Archaic Adaptive Introgression in Present-Day Human Populations. Mol Biol Evol. 1 févr 2017;34(2):296-317.

59. Huerta-Sánchez E, Jin X, Asan, Bianba Z, Peter BM, Vinckenbosch N, et al. Altitude adaptation in Tibetans caused by introgression of Denisovan-like DNA. Nature. 14 août 2014;512(7513):194-7.

60. Dasmahapatra KK, Walters JR, Briscoe AD, Davey JW, Whibley A, Nadeau NJ, et al. Butterfly genome reveals promiscuous exchange of mimicry adaptations among species. Nature. 5 juill 2012;487(7405):94-8.

61. Jones MR, Mills LS, Alves PC, Callahan CM, Alves JM, Lafferty DJR, et al. Adaptive introgression underlies polymorphic seasonal camouflage in snowshoe hares. Science. 22 juin 2018;360(6395):1355-8.

bioRxiv preprint first posted online May. 17, 2019; doi: http://dx.doi.org/10.1101/641829. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

62. Martin SH, Jiggins CD. Interpreting the genomic landscape of introgression. Current Opinion in Genetics & Development. déc 2017;47:69-74.

63. Hedrick PW. Adaptive introgression in animals: examples and comparison to new mutation and standing variation as sources of adaptive variation. Molecular Ecology. 2013;22(18):4606-18.

64. Fraïsse C, Roux C, Welch JJ, Bierne N. Gene-Flow in a Mosaic Hybrid Zone: Is Local Introgression Adaptive? Genetics. 1 juill 2014;197(3):939-51.

65. Barton NH. Does hybridization influence speciation? Journal of Evolutionary Biology. 2013;26(2):267-9.

66. Whitlock MC, Ingvarsson PäK, Hatfield T. Local drift load and the heterosis of interconnected populations. Heredity. 2000;84(4):452-7.

67. Bierne N, Lenormand T, Bonhomme F, David P. Deleterious mutations in a hybrid zone: can mutational load decrease the barrier to gene flow? Genetics Research. déc 2002;80(3):197-204.

68. Kim BY, Huber CD, Lohmueller KE. Deleterious variation shapes the genomic landscape of introgression. PLOS Genetics. 22 oct 2018;14(10):e1007741.

69. Shpak M. The Role of Deleterious Mutations in Allopatric Speciation. Evolution. 2005;59(7):1389-99.

70. Kimura M. The role of compensatory neutral mutations in molecular evolution. J Genet. 1 juill 1985;64(1):7.

71. Schumer M, Cui R, Rosenthal GG, Andolfatto P. Reproductive Isolation of Hybrid Populations Driven by Genetic Incompatibilities. PLOS Genetics. 13 mars 2015;11(3):e1005041.

72. Blanckaert A, Bank C. In search of the Goldilocks zone for hybrid speciation. PLOS Genetics. 7 sept 2018;14(9):e1007613.

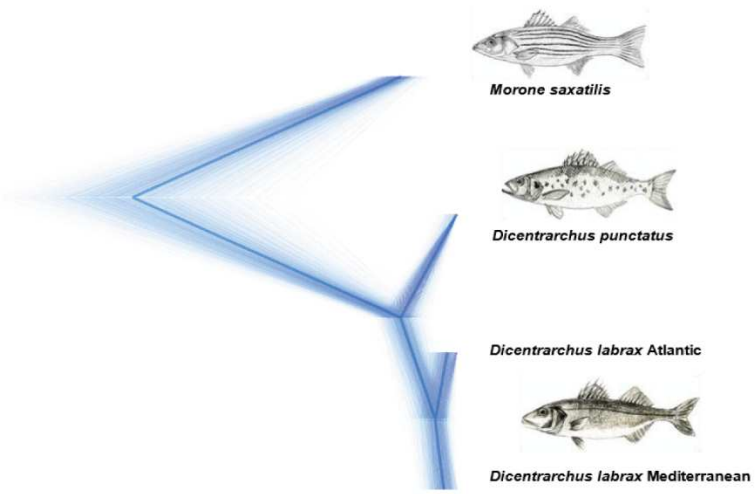
73. Dobzhansky T grigorovitch. Genetics and the origin of species. Columbia university press. 1937.

74. Runemark A, Trier CN, Eroukhmanoff F, Hermansen JS, Matschiner M, Ravinet M, et al. Variation and constraints in hybrid genome formation. Nature Ecology & Evolution. mars 2018;2(3):549.

75. Eberlein C, Hénault M, Fijarczyk A, Charron G, Bouvier M, Kohn LM, et al. Hybridization is a recurrent evolutionary stimulus in wild yeast speciation. Nature Communications. 25 févr 2019;10(1):923.

bioRxiv preprint first posted online May. 17, 2019; doi: http://dx.doi.org/10.1101/641829. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

A



B

	East-Mediterranean *D. labrax*	West-Mediterranean *D. labrax*	Atlantic *D. labrax*	*D. punctatus*
East-Mediterranean *D. labrax*				
West-Mediterranean *D. labrax*	0.0007			
Atlantic *D. labrax*	0.0010	0.0009		
D. punctatus	0.0056	0.0055	0.0053	
M. saxatilis	0.0454	0.0453	0.0451	0.0461

Figure 1 - **Phylogenetic relationships between *D. labrax* lineages, *D. punctatus* and the outgroup species *M. saxatilis*. A.** Concordance of 3,329 maximum-likelihood trees reconstructed in non-overlapping 50 kb windows along the genome (thick lines) and superimposed to the consensus species tree (bold line) using DensiTree v2.2.5 (bouckaert and heled 2014). Only one haplome from each species/lineage was used for tree reconstruction. Discordant trees that disproportionately grouped the Atlantic *D. labrax* lineage with *D. punctatus* were observed at a more local scale using 2 kb windows (Supplementary Figure 3). **B.** Genome-wide average pairwise sequence divergence between species/lineages measured by d_{xy} using the same individual haplotypes as for the phylogenetic relationships.

bioRxiv preprint first posted online May. 17, 2019; doi: http://dx.doi.org/10.1101/641829. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

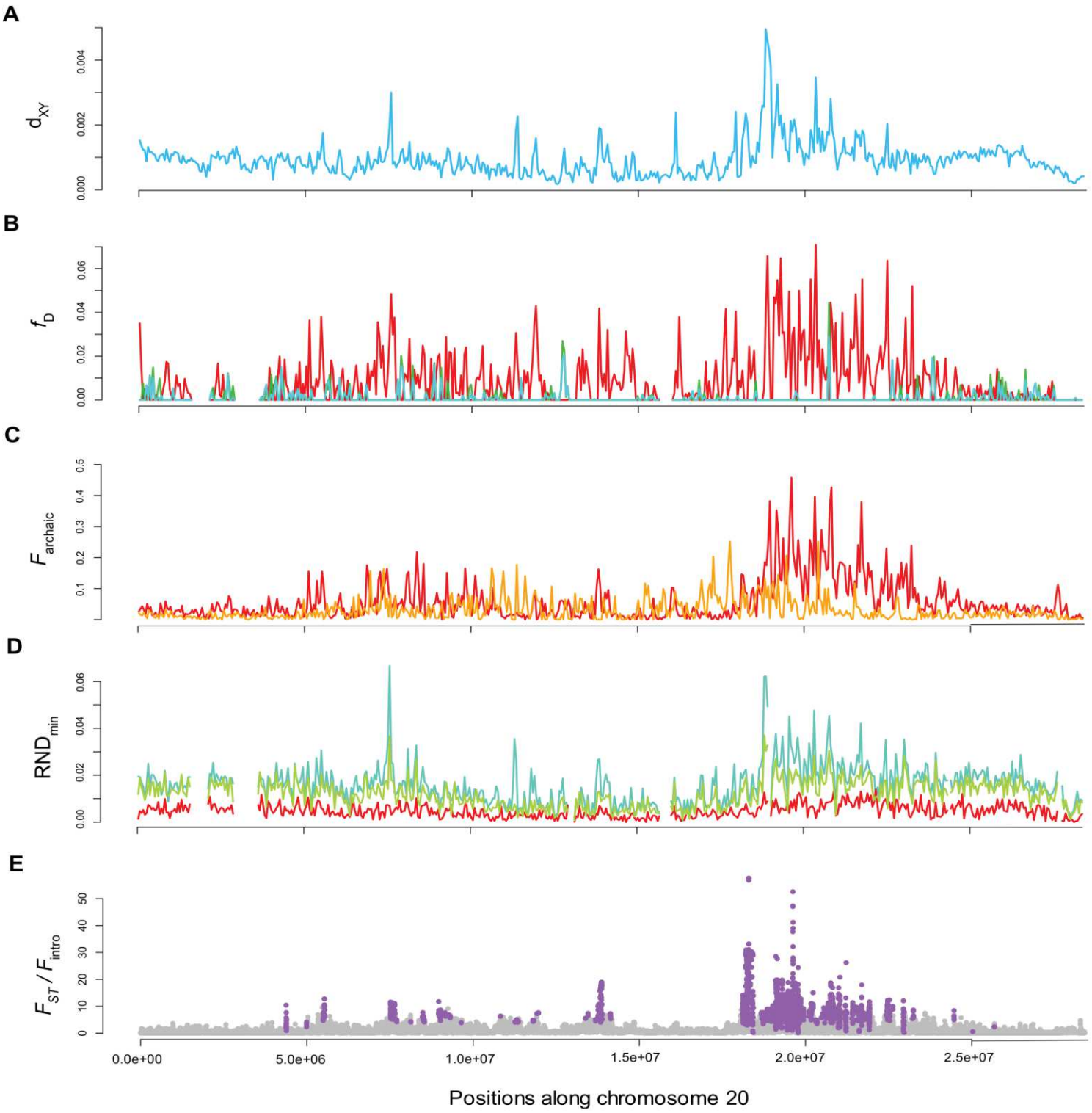


Figure 2 – **Divergence and introgression statistics measured in non-overlapping 50 kb windows along chromosome 20. A.** d_{XY} measured between the Atlantic and Mediterranean (including eastern and western population) *D. labrax* lineages. **B.** f_D statistics measured using (((MED, AT), PUN), SAX) in red, (((AT, WEM), PUN), SAX) in green, and (((ATL, SEM), PUN), SAX) in blue. **C.** Fraction of archaic introgressed tracts ($F_{archaic}$) inferred in the eastern Mediterranean and Atlantic populations of *D. labrax*. **D.** RND_{min} measured between *D. punctatus* and *D. labrax* Atlantic (red), western (green) and eastern (blue) Mediterranean populations. **E.** F_{ST} between the Atlantic and western Mediterranean population of *D. labrax* divided by the fraction of Atlantic tracts introgressed into the western Mediterranean genomes for each SNP along the chromosome. Purple points show SNPs with significant associations to reproductive isolation after applying FDR correction to the probabilities determined with the HMM approach.

bioRxiv preprint first posted online May. 17, 2019; doi: http://dx.doi.org/10.1101/641829. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

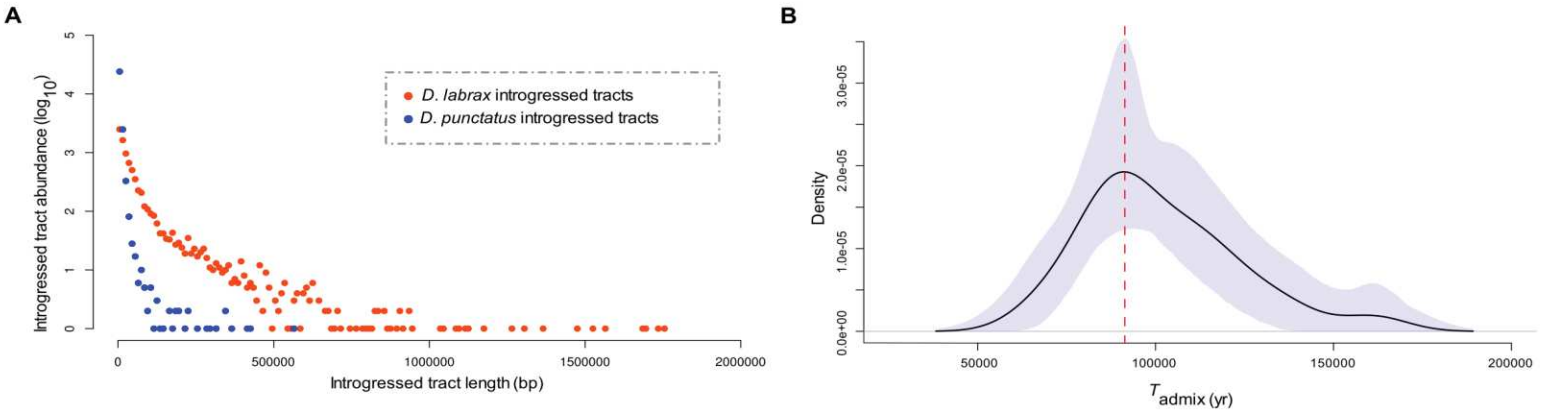


Figure 3 – **Estimation of the time since admixture between *D. punctatus* and Atlantic *D. labrax*. A.** Length distributions of *D. punctatus* tracts introgressed into Atlantic *D. labrax* genomes (blue) and Atlantic *D. labrax* tracts introgressed into western Mediterranean *D. labrax* genomes (orange). Both distributions were generated using similar sequence lengths (totalizing 65.6 Mb) along the genomes of 14 Atlantic and 14 Mediterranean individuals, so that tract abundances can be compared. **B.** Distribution of estimated time since admixture between *D. punctatus* and *D. labrax* (T_{admix}) obtained from estimated transition parameter values of the HMM model over the 24 chromosomes. The maximum of the distribution is represented by the vertical red dashed line and the blue shape represents the 95% credibility envelope of the distribution obtained using 10,000 bootstrap resampling.

bioRxiv preprint first posted online May. 17, 2019; doi: http://dx.doi.org/10.1101/641829. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

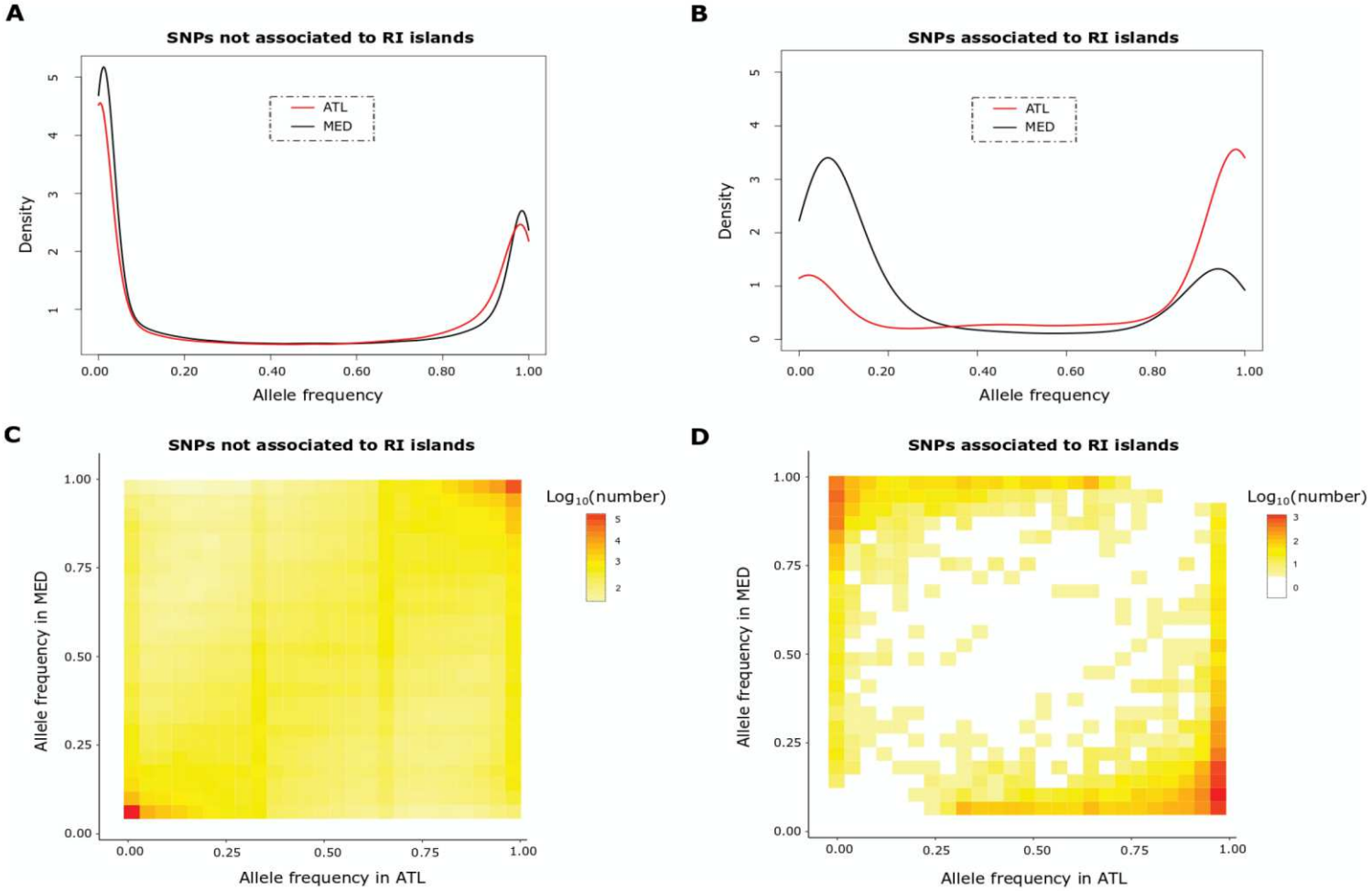


Figure 4 – **One and two-dimensional Site Frequency Spectra of *D. punctatus* derived alleles segregating in *D. labrax*. A.** Conditional Site Frequency Spectra (*CSFS*) of *D. punctatus* derived alleles in AT (red) and MED (black) *D. labrax* lineages for categories of SNPs that are either not associated, or (**B**) associated to RI islands identified between the two *D. labrax* lineages. **C.** Conditional Joint Site Frequency Spectra (*CJSFS*) of derived *D. punctatus* alleles between MED (54 individuals) and AT (14 individuals) lineages based of 618,842 SNPs not involved in RI, and (**D**) 7,372 SNPs involved in RI.

bioRxiv preprint first posted online May. 17, 2019; doi: http://dx.doi.org/10.1101/641829. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

Supporting Information Legends

Supplementary Table 1 – Summary statistics of sequencing and mapping data for each individual. Individuals whose name is in bold are those involved in crossing.

Supplementary Figure 1 - **Depth of coverage per individual**. Median (dark gray), first (light gray) and third (black) quartile of the depth of coverage for the 10 Atlantic males (AT), the 23 individuals from the western Mediterranean sea (14 females (F) and 9 males (WM/WME)), the 19 individuals from the eastern Mediterranean sea (9 males from the south (SEM) and 9 males from the north (NEM)), the 7 hybrids (F1) and the D. punctatus individual (Punc).

Supplementary Figure 2 - Principal Component Analysis of the European sea bass population genetic structure. The analysis was performed on the 52 newly sequenced genomes (colored circles) and the 16 from a previous data set (1) (gray circles with colorful outline). We used the R package adegenet (2) on 91,073 SNPs with a minor allele frequency > 0.4. Individuals originated from four different geographic locations the Atlantic ocean (red, AT), the west (orange, WME), the north-east (dark yellow, NEM) and the south east (light yellow, SEM) of the Mediterranean sea. The first PCA axis explains 39.76% of the total inertia and distinguish the Atlantic and Mediterranean populations while the second PCA axis explains 6.55% of the total inertia and reveals a structure within the Mediterranean population.

Supplementary Figure 3 –Consensus trees of the 155,155 Maximum-likelihood trees inferred in 2kb windows along the genome between M. saxatilis, D. punctatus and Atlantic and Mediterranean D. labrax lineages. There were four different topologies, the most frequent representing the species tree; 94.5% (blue), the second one grouping the Atlantic lineage with D. punctatus; 2.87 % (orange), a third one grouping D. punctatus with the Mediterranean lineage; 1.68 % (green) and a last one with unresolved relationship between D. labrax lineages and D. punctatus; 0.05% (not showed).

Supplementary Figure 4 – Statistics measured in non-overlapping 50 kb windows along the genome. A. dXY measured between the Atlantic and Mediterranean (including eastern and western population) D. labrax lineage B. fD statistic measured using in red (((MED, AT), PUN), SAX), in green (((AT, WEST), PUN), SAX) and in blue (((AT, EAST), PUN), SAX). C. Fraction of archaic introgressed tracts (Farchaic) in the eastern Mediterranean and Atlantic population of D. labrax. D. RNDmin measured between D. punctatus and D. labrax Atlantic (red), western (green) and eastern (blue) Mediterranean populations. E. Ratio of FST and Fintro used to run the HMM approaches on 50 kb windows that rely on 3 states 1 (light grey), 2 (medium grey) and 3 (dark grey). Red points passed the control for false discovery. We defined island of reproductive isolation as continuous regions containing only red and dark grey points (red boxes). F. Ratio of FST and Fintro used to run the HMM approaches on SNPs, purple points are SNPs identified as involved in reproductive isolation that passed the control for false discovery.

Supplementary Figure 5 – Introgressed tract length distributions. Length distributions of Mediterranean tracts introgressed into the Atlantic population (red) and of Atlantic tracts introgressed into the western (orange), north-eastern (dark yellow) and south-eastern (light yellow) Mediterranean populations. Distributions were generated over the whole genome using 11 individuals per population.

bioRxiv preprint first posted online May. 17, 2019; doi: http://dx.doi.org/10.1101/641829. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

Supplementary Figure 6 – Data and results for the SNPs and 50kb window based HMM approach to identify regions involved in reproductive isolation between the two lineage of D. labrax along chromosome 7. A. FST measured between the Atlantic and western Mediterranean population of D. labrax for each SNPs and in every non-overlapping 50 kb windows (D). B. Fraction of Atlantic tracts introgressed in western Mediterranean genomes (Fintro) for each SNPs and in every non-overlapping 50 kb windows (E). C. Statistic analyzed by the HMM approaches (FST divided by Fintro) for each SNPs and every 50 kb non-overlapping window (F). Ratio of FST and Fintro used to run the HMM approaches that rely on 3 states that identify; neutral genomic regions (state 1, light grey), genomic regions under linked selection (state 2, medium grey) and genomic regions involved in reproductive isolation (state3, dark grey). Red points are those that passed the control for false discovery. For the window approach we defined island of RI as continuous regions containing only red and dark grey points (red boxes).

Supplementary Figure 7 – Distributions and joint allele-frequency spectrums of derived D. punctatus alleles present in D. labrax. Distribution of D. punctatus derived alleles frequency in AT (red) and WEST (black) (A) or East (D) D. labrax individuals for loci involved (solid line) or not (dashed lines) in reproductive isolation between the two D. labrax lineages. B. Joint allele-frequency spectrum of derived D. punctatus allele for the WEST (31 individuals) and AT (14 individuals) populations for 594,797 SNPs not involved and 7,372 SNPs involved (C) in reproductive isolation. E. Joint allele-frequency spectrum of derived D. punctatus allele for the EAST (23 individuals) and AT (14 individuals) populations for 594,454 SNPs not involved and 7,366 SNPs involved (C) in reproductive isolation.

Supplementary Table 2 – values used to estimate Tadmix for each chromosome.

CHAPITRE 3 :

Etude des patrons d'évolution moléculaire des gènes aux cœurs des îlots génomiques résistants à l'introgression

Les résultats de ce chapitre sont exposés en détails dans l'article joint.

Maintenant que les régions génomiques impliquées dans l'isolement reproductif entre le bar et le loup ont été clairement délimitées (voir Chapitre 2), nous avons voulu comprendre quels mécanismes évolutifs ont permis leur mise en place. Pour cela nous nous sommes focalisés sur l'étude des patrons d'évolution moléculaire des gènes localisés au cœur des îlots génomiques de spéciation. Cependant, nous avons également montré que ces îlots sont préférentiellement localisés dans les régions à faible taux de recombinaison (voir Chapitre 1). En effet, quand la recombinaison est faible, les allèles d'isolement reproductif sont fortement liés les uns aux autres ce qui leur permet de cumuler leurs effets et de résister plus facilement à l'introgression (Nachman and Payseur 2012; Ortiz-Barrientos *et al.* 2016). Or, la recombinaison est connue pour largement influencer les patrons d'évolution moléculaire (Charlesworth 2009; Campos *et al.* 2014). En effet, la sélection en liaison (positive par balayage sélectif (Smith and Haigh 1974) ou négative par sélection d'arrière-plan (Charlesworth *et al.* 1993)) tends à réduire la diversité génétique des régions à faible taux de recombinaison et donc leur taille efficace. Or, l'efficacité de la sélection étant directement liée à la taille efficace, elle est plus efficace dans les régions à fort taux comparé aux régions à faible taux de recombinaison, où plus de mutations faiblement délétères vont ségréger et moins de mutations avantageuses vont se fixer (Campos *et al.* 2014).

Il est donc nécessaire d'étudier et de prendre en compte les variations du taux de recombinaison pour comparer les patrons d'évolution moléculaire des gènes impliqués et non-impliqués dans l'isolement reproductif. Nous avons donc dans un premier temps utilisé une méthode basée sur l'analyse du déséquilibre de liaison (LDHelmet (Chan *et al.* 2012)) pour estimer le taux de recombinaison populationnel des populations atlantique, ouest- et est-méditerranéennes. Nous avons pu montrer que les variations inférées à large échelle sont très similaires entre les trois populations, mais pas celles à faible échelle. En effet, les positions des points chauds de recombinaison détectés au sein des trois populations ne se recoupent quasiment pas, suggérant une évolution rapide de ces positions dans la lignée de *D. labrax*. Nous avons également montré que les gènes impliqués dans l'isolement reproductif sont moins associés aux points chauds que ce qu'on aurait pu attendre au vu de leur distribution génomique. Étant donné que nous n'avons pas établi de lien entre isolement reproductif et points chauds de recombinaison, nous nous sommes, par la suite, concentrés sur les variations de recombinaison à large échelle.

Nous avons donc voulu vérifier que le lien attendu entre recombinaison et patrons d'évolution moléculaire est également présent dans le génome du bar européen. Pour cela, nous nous sommes concentrés sur les gènes non impliqués dans l'isolement reproductif et les avons classés en six

catégories en fonction de leur taux de recombinaison. Nous avons ensuite estimé pour chaque catégorie le ratio du nombre de substitutions (calculées par rapport au bar américain *Morone saxatilis*) non-synonyme sur synonyme (dN/dS), le ratio du nombre de polymorphismes non-synonymes sur synonyme ($\pi N/\pi S$) et la proportion de substitutions adaptatives (α). Nous avons ainsi pu montrer qu'il existe une forte corrélation entre ces valeurs et la recombinaison (Figure 1), qui s'explique par l'action de processus évolutifs neutres (effet mutagène de la recombinaison), quasi-neutres (sélection en liaison) et sélectifs (efficacité de la sélection purificatrice). Il est donc bien nécessaire de considérer l'effet de la recombinaison pour comparer les patrons d'évolution moléculaire des gènes associés et non-associés à l'isolement reproductif.

Afin de contrôler l'effet de la recombinaison, nous avons sous-échantillonné les gènes non impliqués dans l'isolement reproductif afin que leur distribution de taux de recombinaison soit similaire à celle des gènes impliqués dans l'isolement reproductif et avons comparé leurs patrons d'évolution moléculaire (Figure 3). Nous avons ainsi pu montrer que les gènes impliqués dans l'isolement reproductif subissent de plus fortes pressions de sélection purificatrice (dN/dS plus faible), évoluent en fixant plus de changements adaptatifs (α plus fort) et tendent à être plus conservés à l'échelle phylogénétique (scores de conservation plus élevés) (Figure 4) que les gènes non impliqués dans l'isolement reproductif. Cette contradiction apparente entre évolution adaptative et conservation élevée à l'échelle phylogénétique pourrait indiquer que les changements adaptatifs détectés ont essentiellement eu lieu dans la lignée *labrax*, sur des gènes par ailleurs plutôt soumis à des contraintes sélectives fortes. Enfin, nous nous sommes intéressés aux fonctions des gènes impliqués dans l'isolement reproductif en cherchant si une ou plusieurs fonctions étaient surreprésentées dans ce groupe de gènes par rapport au reste du génome. Nous avons ainsi pu monter que des fonctions telles que le transport transmembranaire d'ions et des processus cognitifs (liées au comportement et à la mémoire) étaient surreprésentées parmi les gènes impliqués dans l'isolement reproductif (Figure 5). Le bar européen étant une espèce euryhaline, le fait que les gènes impliqués dans le transport d'ions soient également impliqués dans l'isolement reproductif semble suggérer que l'isolement reproductif entre le bar et le loup est au moins en partie dû à des allèles sous sélection divergente impliqués dans des adaptations locales. Cependant, l'isolement reproductif entre ces deux lignées étant fortement polygénique (Duranton *et al.* 2018), il reste difficile de quantifier le rôle relatif de l'adaptation locale par rapport à d'autres processus que nous ne pouvions pas détecter ici sans réaliser d'études fonctionnelles spécifiques, telles que les incompatibilités génétiques (DMI).

Ainsi, notre étude a permis de montrer l'existence d'un lien entre recombinaison et patrons d'évolution moléculaire que nous expliquons par l'action de processus à la fois neutres, quasi-neutres et sélectifs. Nous avons également montré que ce sont des éléments conservés du génome, évoluant

généralement sous forte sélection purificatrice, qui participent principalement à l'isolement reproductif chez le bar. On peut donc imaginer que le peu de changements qui touchent ces gènes, bien que principalement adaptatifs, puissent induire des effets importants, facilitant l'établissement de l'isolement reproductif.

Références

Campos J. L., D. L. Halligan, P. R. Haddrill, and B. Charlesworth, 2014 The Relation between Recombination Rate and Patterns of Molecular Evolution and Variation in Drosophila melanogaster. Mol Biol Evol 31: 1010–1028. https://doi.org/10.1093/molbev/msu056

Chan A. H., P. A. Jenkins, and Y. S. Song, 2012 Genome-Wide Fine-Scale Recombination Rate Variation in Drosophila melanogaster. PLOS Genetics 8: e1003090. https://doi.org/10.1371/journal.pgen.1003090

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

Charlesworth B., 2009 Effective population size and patterns of molecular evolution and variation. Nat Rev Genet 10: 195–205. https://doi.org/10.1038/nrg2526

Duranton M., F. Allal, C. Fraïsse, N. Bierne, F. Bonhomme, *et al.*, 2018 The origin and remolding of genomic islands of differentiation in the European sea bass. Nature Communications 9: 2518. https://doi.org/10.1038/s41467-018-04963-6

Nachman M. W., and B. A. Payseur, 2012 Recombination rate variation and speciation: theoretical predictions and empirical results from rabbits and mice. Phil. Trans. R. Soc. B 367: 409–421. https://doi.org/10.1098/rstb.2011.0249

Ortiz-Barrientos D., J. Engelstädter, and L. H. Rieseberg, 2016 Recombination Rate Evolution and the Origin of Species. Trends in Ecology & Evolution 31: 226–236. https://doi.org/10.1016/j.tree.2015.12.016

Smith J. M., and J. Haigh, 1974 The hitch-hiking effect of a favourable gene. Genetics Research 23: 23–35. https://doi.org/10.1017/S0016672300014634

Article: Discoveries

The relation between recombination, reproductive isolation and patterns of molecular evolution in the European sea bass genome

Maud Duranton[1*], Marjolaine Rousselle [1], Flavià Schlichta[1], François Bonhomme[1] and Pierre-Alexandre Gagnaire[1]

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

[*]Corresponding author: durantonmaud@gmail.com

Abstract:

Identifying the evolutionary mechanisms involved in the establishment of reproductive isolation has been the focus of many studies in speciation genomics. Some of them focused on how reduced recombination can hamper the maintenance of reproductive isolation loci in the face of gene flow through particular genomic architecture. However, many questions remain unresolved as to the nature of genes involved in reproductive isolation, the evolutionary forces under which they usually evolve and how recombination influences their evolution. In this study, we used fully phased genome sequences to understand how reproductive isolation has evolved between the Atlantic and Mediterranean lineages of European sea bass. We compared molecular evolution patterns of genes involved in reproductive isolation to those of other genes. Since recombination is known to influence these patterns in a wide range of organisms, we specifically controlled for variation in local recombination rate between the two sets of genes. Therefore, we first precisely described both large and small-scale variation in recombination rate across the genome. We show that patterns of molecular evolution are highly correlated to recombination rate variation through both neutral, nearly neutral and selective processes. We also identified recombination hotspots showing a rapid location turnover between sea bass lineages. Finally, after controlling for recombination, we showed that genes involved in reproductive isolation tend to be on average more evolutionarily conserved, while accumulating more adaptive changes than the rest of the genome in the European sea bass.

Introduction

Under the biological species concept framework, understanding speciation necessarily involves identifying the mechanisms underlying the establishment of reproductive isolation (RI). Loci involved in RI can be defined as positions in the genome contributing to a local reduction of the effective migration rate compared to the genomic background rate (Ravinet *et al.* 2017). Thus, many recent studies have focused on uncovering aspects of the genetic architecture that help to more effectively reduce gene flow (Gompert *et al.* 2012; Lindtke and Buerkle 2015; Martin and Jiggins 2017). On the other hand, different kinds of loci may generate RI, such as those involved in hybrid incompatibilities, mating preferences/differences or local adaptation. Which of these mechanisms is the main driver of speciation is still unknown for most models in speciation research and thus many questions persist as to the nature of genes involved in speciation (Butlin *et al.* 2012). For example, do the same genes tend to be involved in different speciation events due to their function? Are these genes usually under strong or weak evolutionary constraints at the phylogenetic level? Do they show particular patterns of molecular evolution? Solving those questions might be a first step to resolve a long-standing debate in evolutionary biology about the relative role in the speciation process of natural selection and genetic drift or adaptation and genomic conflicts (Nosil and Schluter 2011).

According to several theoretical models of speciation, in the presence of gene flow loci involved in RI are more likely to be found in genomic regions experiencing reduced recombination (Nachman and Payseur 2012; Ortiz-Barrientos *et al.* 2016). Indeed, tight linkage allows cumulating their effects and generating a more efficient barrier to migration (through a higher impact on the fitness of individuals carrying them in heterozygous condition). Thus, structures that efficiently reduce recombination such as inversions are often enriched in RI loci (Rieseberg 2001; Noor *et al.* 2001; Faria *et al.* 2019). The link between RI and recombination is often made to understand how limited recombination helps RI loci to be maintained in the face of gene flow (Martin and Jiggins 2017), but little is known about how recombination influences the molecular evolution of RI alleles. Indeed, regions of reduced recombination present particular patterns of molecular evolution (Charlesworth 2009; Campos *et al.* 2014), as evolution at a given site is influenced by selection acting on linked sites, a process known as Hill–Robertson effect (Hill and Robertson 1966; Felsenstein 1974). Selection at linked sites can be positive through selective sweeps, when the fixation of a favorable mutation drives linked neutral or weakly deleterious mutations to fixation (Smith and Haigh 1974). It can also be negative through the action of background selection, which by removing deleterious mutations also removes linked variation (Charlesworth *et al.* 1993). All these processes therefore induce a reduction of the effective population size which is more pronounced in low- compared to highly-recombining genomic regions. Thus, selection tends to be less efficient in low-recombining regions, which in turn causes them to accumulate more slightly deleterious mutations and fix less advantageous ones (Campos *et al.* 2014).

Therefore, to understand which evolutionary mechanisms can lead to RI through the study of molecular evolution patterns, it is first important to disentangle the relationships between recombination, molecular evolution and RI. Here, we propose to implement this strategy in the European sea bass (*Dicentrarchus labrax*), a marine fish species subdivided into two glacial lineages that are currently represented by the Atlantic and Mediterranean populations (Lemaire *et al.* 2005). These two lineages have started to diverge in allopatry *c. a.* 300,000 years ago and later came into secondary contact *c. a.* 11,500 years ago, allowing gene flow mainly from the Atlantic to the Mediterranean genetic background (Tine *et al.* 2014). This species represents an ideal case study for three main reasons. First, the existence of RI between the two lineages has been demonstrated using the variable rates of introgression of Atlantic haplotypes within Mediterranean genomes (Duranton *et al.* 2018) and the genomic regions involved were precisely delineated (Duranton *et al.* 2019). Secondly, there is a strong variation in recombination rate along the genome, with centro-chromosomic regions recombining on average ten times less than peri-chromosomic regions (Tine *et al.* 2014). Finally, it has been shown that RI loci tend to be preferentially located in low-recombining regions (Duranton *et al.* 2018).

69

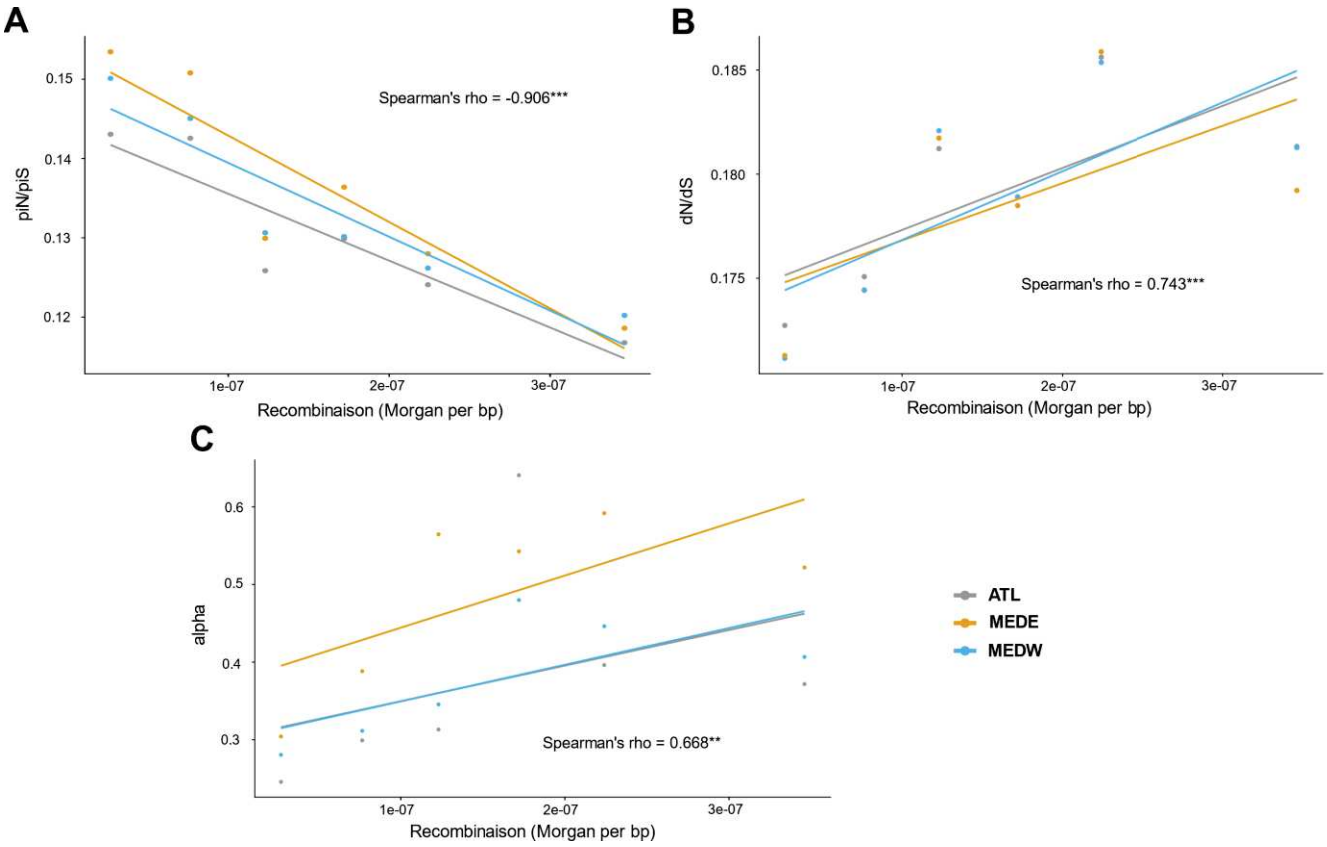


Figure 1 – **Correlation between recombination rate and patterns of molecular evolution.** Measures were computed independently for the Atlantic (ATL; grey), western (MEDW; blue) and eastern-Mediterranean (MEDE; orange) populations. Stars indicate the level of significance of the p-value (** when p-value < 0.01 and *** when p-value < 0.001) for the linear correlation measured including all populations. **A.** piN/piS **B.** dN/dS and **C.** alpha.

Here we used recently published new individual genomes (Duranton *et al.* 2019) to firstly, precisely describe variations in recombination rate along the genome of the European sea bass. We used an approach relying on linkage-disequilibrium, to study large-scale variation and looked for fine-scale variation by detecting recombination hotspots. The gene responsible for the rapid evolution of recombination hotspots in many organisms, PRDM9, (McVean *et al.* 2004; Baudat and de Massy 2007; Myers *et al.* 2010; Billings *et al.* 2013; Baker *et al.* 2015, 2017) is not functional in the European sea bass (Baker *et al.* 2017). However, it has been shown that rapidly evolving hotspots can be generated through other mechanisms, as it is the case for the threespine stickleback fish (Shanfelter *et al.* 2019). Secondly, we looked for relationships between recombination rate variation, patterns of molecular evolution and RI. Finally, we focused on genes involved in RI and studied their patterns of molecular evolution (taking into account their level of recombination), their level of evolutionary constraints and looked for functional enrichment associated with RI genes. Once variation in recombination rate was taken into account, our results indicate that genes involved in RI tend to be on average more conserved at the phylogenetic scale, under stronger purifying selection and to evolve more through adaptative changes than other genes in the European sea bass.

Results

Relation between large-scale variation in recombination rate and patterns of molecular evolution

We first tested if large-scale variations in recombination rate across the genome of *D. labrax* influence patterns of molecular evolution. We found highly significant positive correlations between both the nucleotidic polymorphism measured across synonymous sites (πS) and non-synonymous ones (πN) and the local recombination rate of genes (Supplementary Figure 3A-B). Since the level of polymorphism is tightly linked to the effective population size, a positive correlation is indeed expected between π and recombination rate due to the effect of linked selection on local N_e (Corbett-Detig *et al.* 2015). This correlation is not influenced by the direct effect of selection of the studied sites for synonymous mutations that are assumed to be neutral. However, direct selection against weakly deleterious mutations that are expected to represent a large proportion of the non-synonymous fraction of polymorphism affects the correlation between πN and recombination. Since direct selection against negative mutations is more efficient in highly recombining regions, more weakly deleterious mutations should be removed by direct selection in regions of high compared to low recombination rates. Consistent with this expectation, the slope of the positive correlation between π and the recombination rate was weaker for non-synonymous compared to synonymous mutations (π slope = $1.865e^{-4}$, πN slope = $2.355e^{-5}$). Therefore, the πN/πS ratio was negatively correlated with the recombination rate (Figure 1A). Likewise, variation in effective population size also explained

differences among the regressions performed separately in the three populations. The eastern-Mediterranean population, which has the smallest N_e displayed the lowest values of π, whereas the Atlantic population, which has the largest N_e consistently showed the highest nucleotide diversity and the lowest πN/πS ratios (Figure 1A and Supplementary Figure 3A-B).

We also found highly significant positive correlations between the divergence parameters dS, dN and their ratio dN/dS with the recombination rate of genes for each of the three populations (Figure 1B and Supplementary Figure 3C-D). These positive correlations are not expected under the neutral theory (Birky and Walsh 1988). Indeed, the common ancestor of *D. labrax* and *M. saxatilis* is ancient enough for divergence between these two species to be mainly generated by the fixation of new mutations independently in each species rather than the sorting of ancestral polymorphism. Therefore, the positive correlation observed between πS and recombination rate, and attributed to the effect of linked selection on neutral diversity, is not expected with dS. Indeed, the neutral theory predicts that the accumulation of neutral mutations is independent from the effective population size. Thus, synonymous substitutions should accumulate at the same pace in low and high recombining regions, generating no correlation between dS and recombination rate. Mildly deleterious mutations segregating in the polymorphism are supposed to contribute relatively less to divergence if selection is efficient enough to prevent their fixation. In this case, negative correlations are expected between both dN and dN/dS and the recombination rate (Charlesworth 1994). By opposition, the positive correlations found here between these parameters and the local recombination rate may thus be indicative of relaxed Hill-Robertson interference among positively selected sites in highly recombining regions (Hill and Robertson 1966; Peck 1994). Differences observed between the three *D. labrax* populations reflect recent evolutionary histories as they are the result of processes that happened after the three populations split. Indeed, the eastern Mediterranean population shows the highest values of dN and dS because it has the smallest effective population size and thus experience higher genetic drift increasing the divergence with the outgroup.

Finally, we also found a significant positive correlation between both ωa (the adaptative substitution rate) and α (the proportion of adaptative amino-acid substitutions) and the recombination rate, whereas a negative correlation was found between ωna (the non-adaptative substitution rate) and recombination (Figure 1C and Supplementary Figure 3E-F). Since selection is less efficient in low-recombining regions due to their reduced effective population size, the fixation of adaptive mutations could be more difficult in low compared to highly-recombining regions. Indeed, advantageous mutations are expecteéd to more easily recombine from deleterious linked mutations and fix in highly recombining regions, generating a positive correlation between ωa and recombination. Likewise, deleterious mutations maybe less effectively removed from low-recombining regions, generating the

observed negative correlation between ωna and recombination. Lastly, while western-Mediterranean and Atlantic populations present very similar values of ωna, ωa and α, the eastern-Mediterranean population shows higher values of ωa and α and lower values of ωna across the range of recombination rate values.

Fine scale variation in recombination rate and the mapping of recombination hotspots

We first wanted to determine if we could identify fine scale variation in recombination rate in the form of hotspots of recombination in the different populations of *D. labrax*. We found 2379 recombination hotspots in the Atlantic population, 2015 in the western-Mediterranean population and 993 in the eastern-Mediterranean population. Since hotspots are more easily detected in the presence of high rates of polymorphism, they should be found more easily in large populations, which likely explains the differences in hotspot numbers between the three populations. Since large-scale variations in recombination rate are highly correlated among *D. labrax* populations (Spearman's rho estimated between population-scaled recombination rate values measured within 50kb window between ATL-MEDW: 0,743***, ATL-MEDE: 0,719*** and MEDE-MEDW: 0,772***) with a reduced recombination rate in centro-chromosomic regions, we wanted to determine if recombination hotspots positions were conserved as well. We found 252 hotspots in common between the Atlantic and the western-Mediterranean population (representing respectively 10.59% and 12.51% of Atlantic and western-Mediterranean hotspots), 107 between Atlantic and eastern-Mediterranean population (representing respectively 4.50% and 10.77% of Atlantic and eastern-Mediterranean hotspots) and 147 between eastern and western-Mediterranean population (representing respectively 7.29% and 14.80% of western and eastern-Mediterranean hotspots). Only 33 hotspots were common to all three populations.

These results support that hotspots location largely differs between the different populations of *D. labrax*, possibly indicating a rapid turnover of hotspots position. We thus investigated whether patterns of molecular evolution associated to recombination hotspots were representative of the local recombination rate of the hotspot itself or the genomic region where the hotspot sits. In line with the negative correlation found between $\pi N/\pi S$ and recombination rate (Figure 1A), we found that hotspots present the lowest values of $\pi N/\pi S$ compared to low- and highly-recombining regions in all three populations (Figure 2A). This was associated to respectively high and moderately high values of πS and πN in hotspots (Supplementary Figure 4A-B). This finding confirms that identified recombination hotspots display the typical diversity patterns expected for regions with high recombination rates. However, contrary to what was expected from the positive correlations between divergence parameters (dN, dS, and their ratio dN/dS) and recombination (Figure 1B, Supplementary Figure3C-D), hotspots generally displayed rather low values of dN, dS and dN/dS ratio, even if there

172

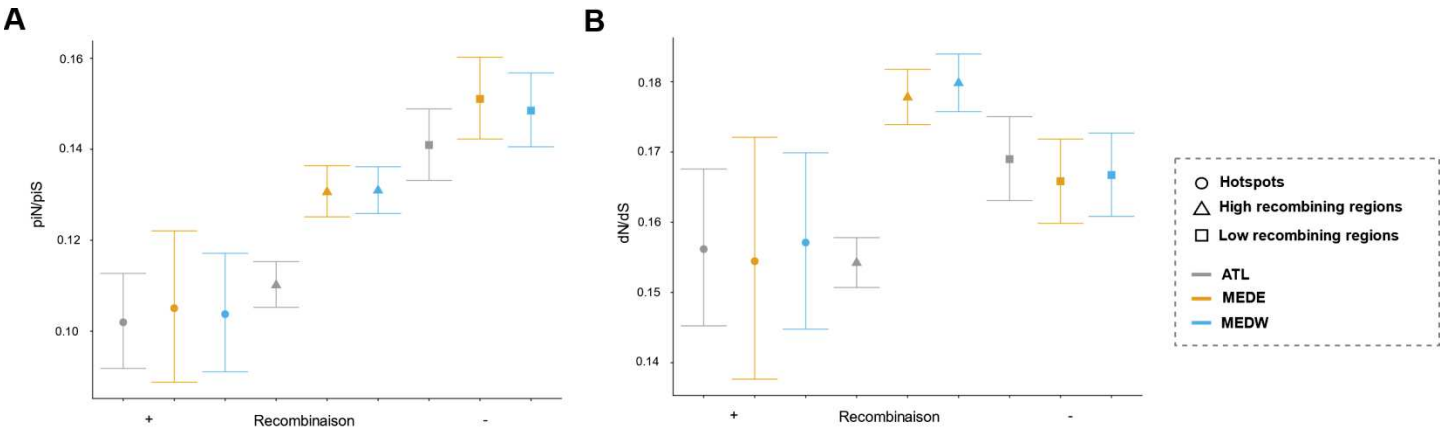


Figure 2 – **Patterns of molecular evolution in recombination hotspots compared to the rest of the genome.** Measures were computed for recombination hotspots (circles), regions of high (triangles) and low (squares) recombination, independently for the Atlantic (ATL; grey), western (MEDW; blue) and eastern-Mediterranean (MEDE; orange) populations. Verticals bars represent the 95% confidence intervals. **A.** piN/piS and **B.** dN/dS.

was some variation among populations (Figure 2B and Supplementary Figure 4C-D). Since patterns of divergence reveal a more ancient history than patterns of polymorphism, the fact that hotspots do not present the patterns of divergence expected in high-recombining regions could indicate that their location has changed through time. Finally, we wanted to determine if hotspots tend not to be associated with RI loci as predicted from the inverse relationship between recombination and the accumulation of RIs. We thus used a chi-squared test to compare the proportion of RI-associated SNPs located in recombination hotspots over the total number of RI SNPs to the proportion of 2kb windows that were identified as hotspots over the entire genome. We found that for all three populations, there are significantly less SNPs involved in RI located in hotspots than expected based on the genome-wide distribution of hotspots ($p\text{-value}_{ATL}=2.2e^{-16}$, $p\text{-value}_{MEDW}=2.2e^{-16}$ and $p\text{-value}_{MEDE}=3.81e^{-10}$).

Molecular evolution patterns of RI-associated genes

One of our main objectives was to determine whether genes involved in RI present particular patterns of molecular evolution by comparison with genes that are not involved in RI. However, RI regions tend to be disproportionally located in low recombining regions (Duranton *et al.* 2018), as illustrated here by contrasting distributions of recombination rate values between genes involved/not involved in RI (Supplementary Figure 1). Furthermore, we showed that recombination influences patterns of molecular evolution at multiple levels in the *D. labrax* genome (Figure 1). Thus, in order to control for recombination rate effects while testing for differences in molecular evolution patterns between different sets of genes, we subsampled genes that are not involved in RI in order to reproduce the distribution of recombination rates observed for genes involved in RI (Supplementary Figure 2).

First, we looked at polymorphism data and showed that like previously (Figure 1A), differences in π among populations mainly reflect differences in effective population size (Supplementary Figure 5A-B). Likewise, π values tend to be lower for genes involved in RI compared to those not involved and this across all populations. Part of this finding might be attributed to the absence of gene flow between Atlantic and Mediterranean sea bass lineages within regions involved in RI, which may decrease the effective population size and thus polymorphism. For the $\pi N/\pi S$ ratio, there was a slight tendency for lower values at genes involved in RI in Mediterranean populations, while the opposite was observed in Atlantic (Figure 3A). However, confidence intervals were largely overlapping between RI-associated and non-RI-associated sets of genes.

Secondly, we compared divergence patterns between candidate RI-associated genes and non-RI-associated sets of genes. For genes not involved in RI, values of dN, dS and especially dN/dS were very similar among the three populations (Supplementary Figure 5C-D and Figure 3B). This may be expected as gene flow is ongoing in those regions and genes have therefore shared *labrax*-outgroup divergent

206

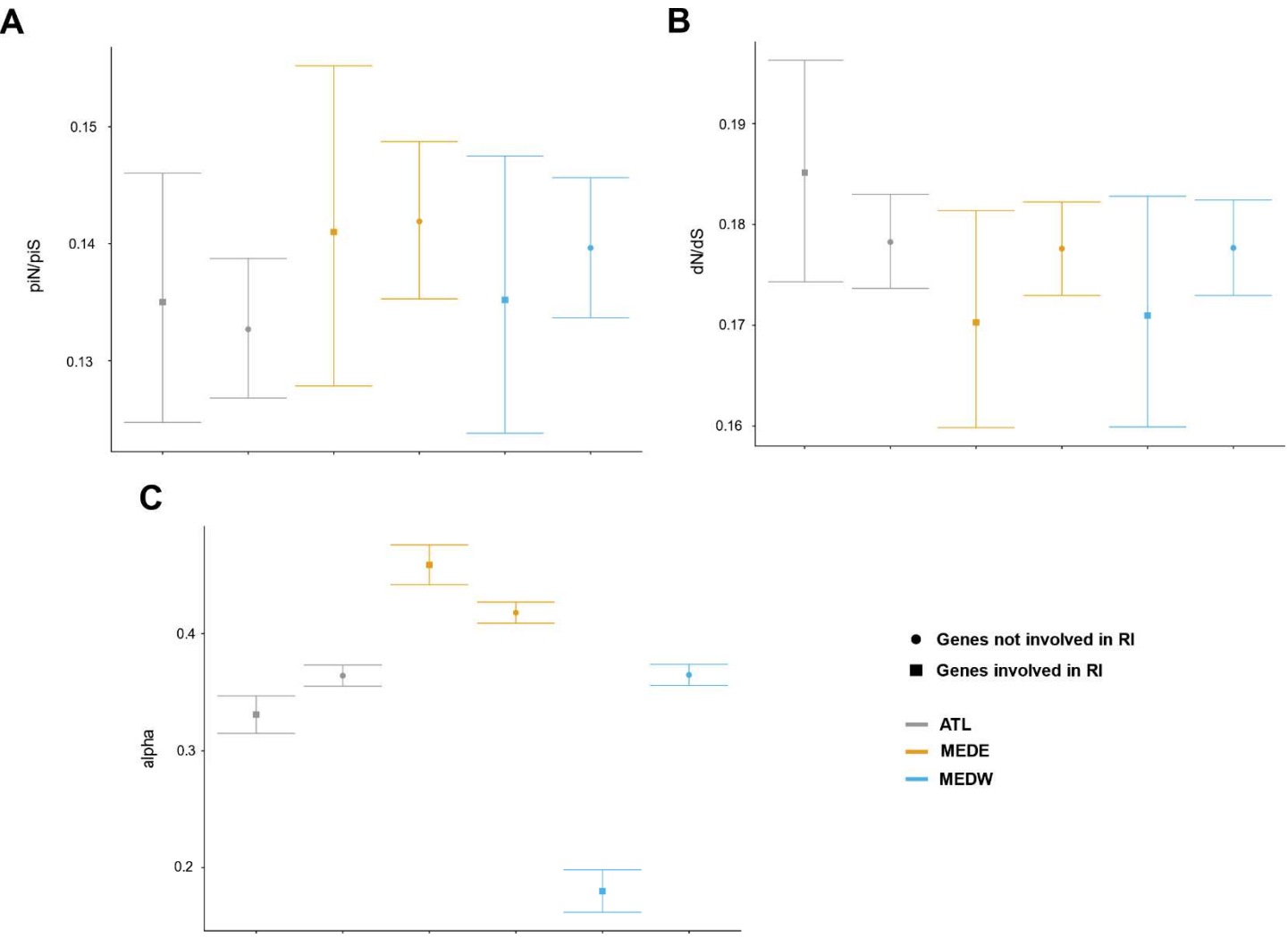


Figure 3 – **Comparison of molecular evolution patterns.** Measures were computed independently for the Atlantic (ATL; grey), western (MEDW; blue) and eastern-Mediterranean (MEDE; orange) populations for genes involved (squares) or not (circles) in RI presenting similar recombination rates. Verticals bars represent the 95% confidence intervals. **A.** piN/piS **B.** dN/dS and **C.** alpha.

sites among populations. By contrast, genes involved in RI showed different divergence patterns depending on which of the Atlantic or Mediterranean population was used as a *D. labrax* ingroup. Indeed, for both Mediterranean populations, genes involved in RI tended to have lower values of dN/dS compared to genes not involved in RI, while we found the opposite for the Atlantic population (Figure 3B). This might indicate that the genes involved in RI tend to be under stronger selective pressure than genes not involved in RI in Mediterranean population, while this tendency would be opposite in Atlantic. Indeed, for Mediterranean populations, dN values tended to be higher for non-RI genes (Supplementary Figure5 D), as expected when selection is relaxed. Finally, we compared patterns of non-synonymous amino acid substitutions due to adaptation. For the eastern-Mediterranean populations values of both ωa and α tend to be higher and ωna lower for genes involved in RI compared to genes not-involved in RI (Figure 3C and Supplementary Figure 5E-F). Therefore, it seems that changes that occurred in genes involved in RI were on average more positive than those in non-RI genes. However, we observed the exact opposite patterns for the Atlantic and western-Mediterranean populations (Figure 3C and Supplementary Figure 5E-F).

Comparison of phylogenetic conservation scores between genes involved or not in RI

As the method used to calculate conservation scores is based on branch length, it could be influenced by variation in recombination rate correlated to genetic diversity. Since, we know that RI regions tend to have lower recombination rate than non-RI regions, differences between genes involved or not in RI could reflect differences in recombination level. To determine if our estimations of conservation scores were influenced by recombination, we compared non-genic regions involved or not in RI. We found that whatever the level of stringency used to detect conserved elements (5%, 25% and 50%) and for both phylogenetic analysis levels (family level or Perciformes level) there are no differences between non-genic regions involved or not in RI (Figure 4, green compared to purple). This indicates that our estimations are not influenced by variation in recombination rate. We also found as expected that all conservation scores tend to be higher in genes compared to non-genic regions (Figure 4 pink compared to green and blue to purple). By contrast, conservation scores tend to be higher for genes involved compared to those not involved in RI (Figure 4 pink compared to blue).

Functional enrichment analysis

We found that some molecular functions are over-represented in the subset of genes involved in RI (Figure 5 and Supplementary Table 1 and 2). The most significant functional enrichment appears for genes coding for transporter activity especially ion transmembrane transport (GO:0034220, PV = 10^{-8}). However, it seems to also be the case for genes involved in cognitive processes such as learning and memory (GO:0007611, PV = 10^{-5}) and behavior and cell organization.

241

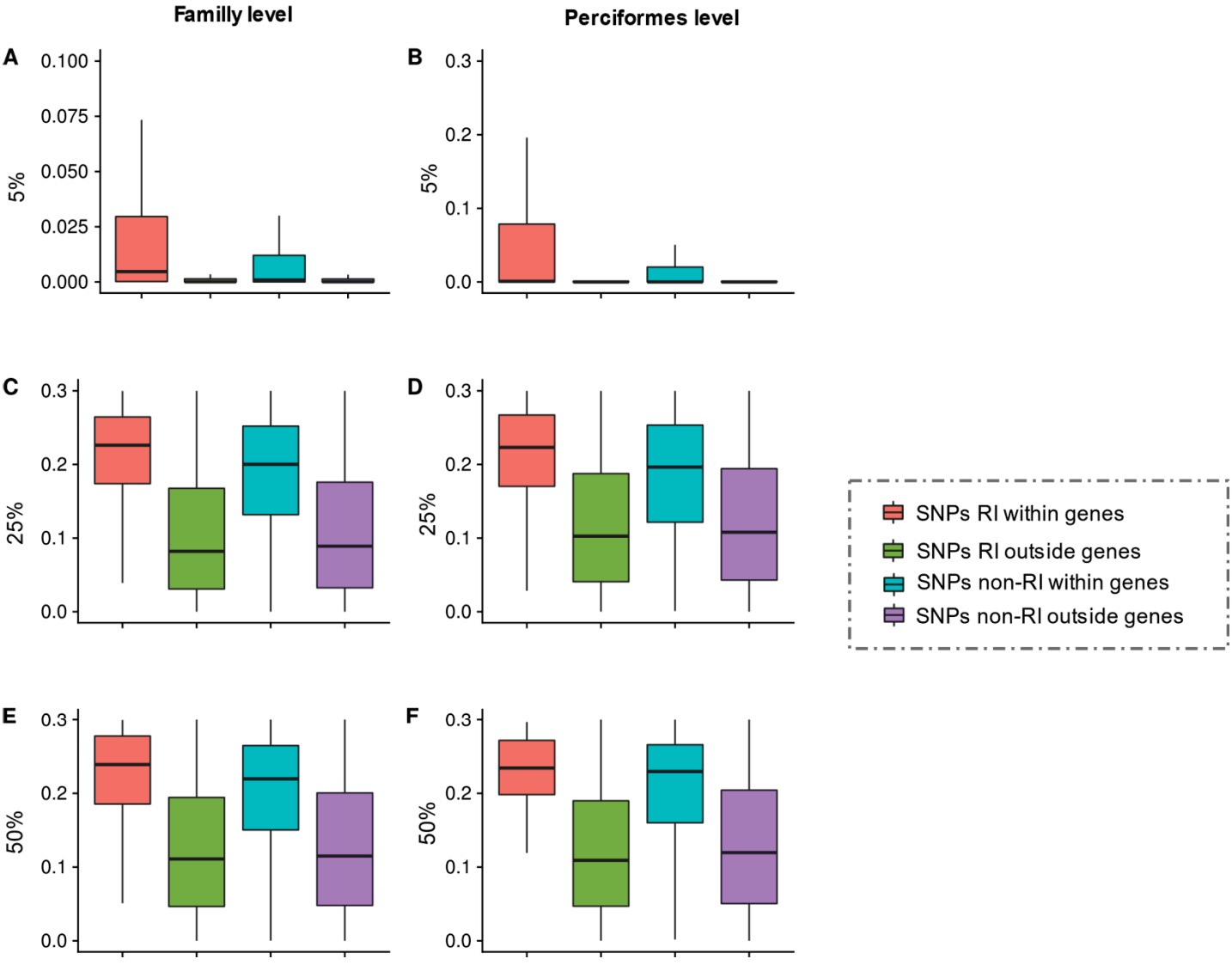


Figure 4 – **Conservation scores.** Measures were computed at the family (A,C and E) and the Perciformes level (B,D and F) for SNPs involved in RI within (pink) and outside (green) genes and SNPs not involved in RI within (blue) and outside (purple) genes. Horizontal barres represent the median and vertical bars the first and third quartile of the distribution. Conservation scores estimated at the family level for the **A.** 5%, **C.** 25% and **E.** 50% most conserved regions of the genome and at the Perciformes level for the **B.** 5%, **D.** 25% and **F.** 50% most conserved regions of the genome.

Discussion

Linking large-scale variation in recombination rate to molecular evolution patterns

The patterns of molecular evolution in the European sea bass genome are influenced by the action of both direct and indirect selection. The positive correlation observed between polymorphism and recombination (Supplementary Figure 3A-B) is expected under nearly neutral evolution. Indeed, the indirect effect of (positive and negative) selection tend to remove more variants in low recombining regions where linkage extends over larger stretches of the genome, generating a positive correlation between genetic diversity and recombination (Campos *et al.* 2014). The negative correlation found here between πN/πS ratio and recombination (Figure 1A) was due to the attenuation of the positive correlation between πN and recombination compared to correlation for πS, which is expected due to the direct effect of selection. Indeed, selection being more efficient in highly recombining regions, fewer weakly deleterious (and hence non-synonymous) mutations segregate in the polymorphism in those regions. Likewise, the positive correlation found between α and recombination (Figure 1C) is expected under adaptive evolution and has been found in other organisms such as *Drosophila melanogaster* (Presgraves 2005; Campos *et al.* 2014). Consistently with an easier fixation of advantageous mutations and removal of deleterious ones in highly recombining regions, we also found a positive correlation between ωa and recombination and a negative correlation for ωna (Supplementary Figure 3E-F). In addition, some patterns such as the observed differences in polymorphism levels between the three populations (Supplementary Figure 3A-B) rather reflect purely neutral processes. Indeed, they are generated by the differences in genetic drift intensity that populations undergo due to their differences in effective population size (Charlesworth 2009). Consistently with previous knowledge on effective population sizes in *D. labax*, the Atlantic population showed the lowest πN/πS ratios across all recombination classes.

The positive correlations observed between divergence and recombination (Figure 1B and Figure 3C-D) are however not expected under the action of indirect or direct selection. Indeed, indirect selection is assumed to mainly affect rates of polymorphism and not divergence, especially when the outgroup used to measure divergence is phylogenetically distant, as it is the case here with *M. saxatilis*. Thus, although linked selection can explain the positive correlation between πS and recombination, it cannot account for the positive correlation between dS and recombination. Likewise, direct selection on deleterious mutations is expected to generate a negative correlation between dN and recombination, as the removal of negative mutations is more efficient in highly recombining regions. Therefore, the mutagenic effect of recombination; which has been described in human (Hellmann *et al.* 2003), could be at play in the European sea bass genome as well and generate the positive correlation observed between divergence and recombination. Another neutral process related to recombination is GC-

276

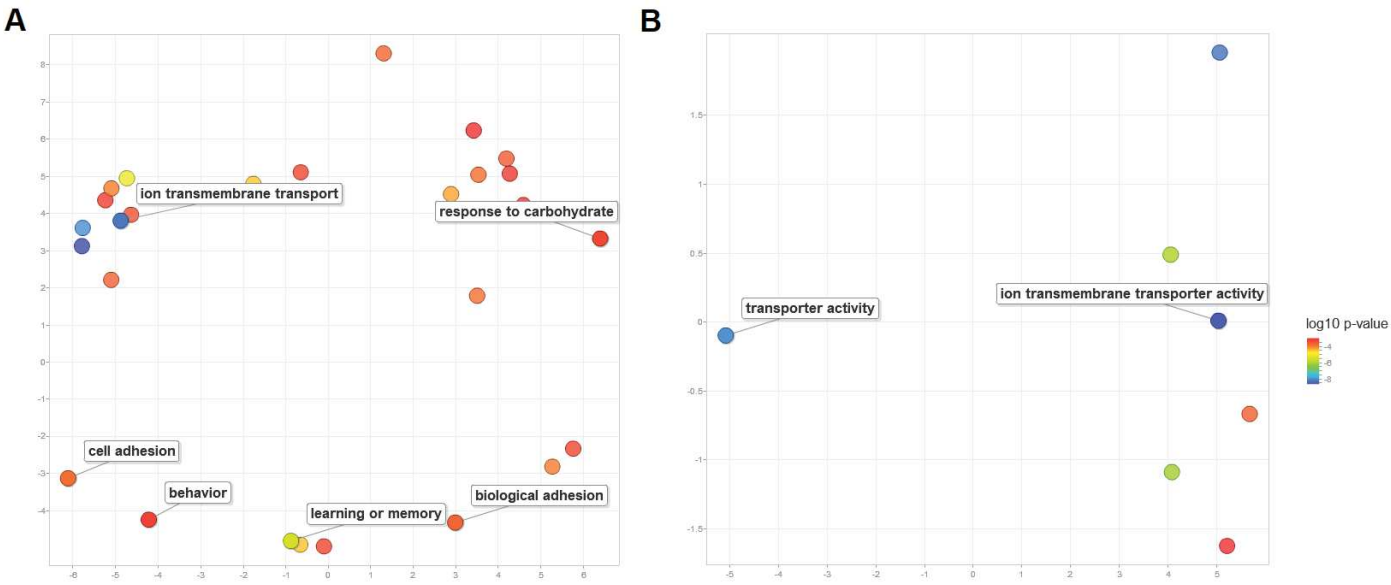


Figure 5 – ***Functional enrichment analysis of gene involved in RI compared to the rest of the genome.*** ***A.*** *At the level of biological process and* ***B.*** *molecular function. Sets of genes were group according to their function and are projected in a space representing the similarity between GO terms. The color represents the p-value of the enrichment analysis.*

biased gene conversion, which intensity is positively correlated to recombination, and is known to leave footprints that can mimic positive selection (Galtier and Duret 2007). Finally, differences in divergence values observed between the three *D. labrax* populations and *M. saxatilis* might reflect differences in branch lengths between *D. labrax* populations. Even if these mutations represent only a small fraction of the total substitutions separating *D. labrax* from *M. saxatilis*, they probably contribute to the observed patterns, indicating that the contribution of the recent history of divergence between the two *D. labrax* lineages cannot be neglected to interpret the present results.

Our results thus clearly demonstrate that patterns of molecular evolution in the European sea bass genome are largely influenced by large-scale variation of recombination rate, whatever them being due to neutral, nearly neutral or selective processes (Figure 1). We also confirmed that genes involved in RI tend to map to regions with lower recombination rates compared to non-RI genes (Supplementary Figure 1). Therefore, it is mandatory to control for recombination to compare molecular evolution patterns between these two sets of genes. One question remains however, concerning fine-scale variation in recombination rate. Indeed, large-scale variation have been well described in fish genomes, and can be partly attributed to the effect of cross-over interferences acting on chromosomes of different lengths (review Otto and Payseur 2019). This is thus presumably highly conserved over large phylogenetic scale due to the stability of fish karyotype (Mank and Avise 2006), but little is known about fine-scale variations

The mapping of recombination hotspots

In mammalian genomes, recombination hotspots are largely associated with the long zinc-finger recombination protein PRDM9 (Parvanov *et al.* 2010) that is partially or completely lost in many fish species (Baker *et al.* 2017). However, recombination hotspots have also been identified outside mammalians, as for example in the swordtail fish, where they are associated to promoter-like features, due to either particular binding motif or a greater chromatin accessibility (Baker *et al.* 2017). The rapid evolution of PRDM9 sequence is assumed to be responsible for the rapid evolution of hotspots location between species (Myers *et al.* 2010), thus for species without PRDM9 hotspots location are thought to be usually conserved over longer evolutionary time scales (Baker *et al.* 2017). However, a recent study showed that in the threespine stickleback, recombination hotspots are not associated to PRDM9 motif (RDM9 being non-functional), but differ between closely related populations, suggesting that there may be another mechanism for targeting recombination hotspots location and its turnover in such species (Shanfelter *et al.* 2019).

Here we show that recombination hotspots can be detected in the European sea bass genome and that their location largely differ between populations (with less than 5% of hotspots found in common

to the three populations) suggesting a rapid turnover of hotspots location in this species. As expected, the levels of polymorphism of genes intersecting hotspots location was similar or higher to those of highly recombining regions (Figure 2A and Supplementary Figure 4A-B). However, divergence values were particularly low (Figure 2B and Supplementary Figure 4C-D) contrary to our expectations based on the positive correlation found between divergence and recombination (Figure 1B and Supplementary Figure 3C-D). Since divergence values reflect a more ancient history than polymorphism data, this tends to confirm that the location of hotspots is not conserved over long evolutionary time scales in sea bass. Determining whether the location of hotspots is associated to promoter-like features, particular binding motifs or simply related to the level of chromatin accessibility, is behind the scope of this paper. Nevertheless, our finding that supports a rapid evolution of hotspot location suggests that it is probably not associated to the location of promoter features, since synteny tends to be highly conserved among fish genomes. Similarly to what was found for the threespine stickleback, there might be a mechanism that allows double strand break formation (Shanfelter *et al.* 2019), possibly targeting particular alleles of epigenetic changes which frequency changes among populations. Finally, we found that SNPs involved in RI tend to be less associated to recombination hotspots than expected based on the genome-wide distribution of hotspots. Therefore, considering large-scale variation in recombination rate alone seems sufficient to study molecular evolution patterns of RI associated genes.

Evolution of RI-associated genes

To understand under which kind of selective pressure RI associated genes evolve, we compared patterns of molecular evolution (controlling for recombination) and conservation scores between genes associated and not associated to RI. No clear tendency emerged from the comparisons of $\pi N/\pi S$ ratios as confidence intervals were largely overlapping between the two categories of genes (Figure 3A). For divergence patterns, Mediterranean populations tended to present slightly lower values of dN/dS for RI genes (Figure 3B, blue and orange), which could indicate stronger purifying selection pressures. In the contrary, we found the opposite tendency for the Atlantic population (Figure 3B grey). However, higher value of dN/dS observed for RI-associated genes in Atlantic was due to lower dS and not higher dN values (Supplementary Figure 5C-D, grey) and thus was probably not indicating relaxed purifying selection on these genes. Since differences in dN/dS calculated using different *D. labrax* populations only reflect changes that occurred after the split between the two *D. labrax* lineages (*c.a.* 300,00 years ago), the shift of tendency between the Atlantic and the Mediterranean populations could be the consequence of recent evolutionary changes. One major recent event which has impacted the divergence between Atlantic and Mediterranean sea bass lineages is the genetics exchanges that

occurred 80,000 years ago between the spotted sea bass (*Dicentrarchus punctatus*) and the Atlantic lineage (Duranton *et al.* 2019). Indeed, we showed in a previous study that this past introgression led to the fixation of *D. punctatus* alleles within Atlantic genomes, which facilitated the establishment of RI between the two *D. labrax* lineages (Duranton *et al.* 2019). One hypothesis proposed to explain why *D. punctatus* alleles may have fixed within Atlantic genomes, while contributing to RI with the Mediterranean lineage is through the resolution of genetic conflicts (Schumer *et al.* 2015; Blanckaert and Bank 2018; Duranton *et al.* 2019). One way to resolve a conflict caused by a two-locus Bateson-Dobzhansky-Muller incompatibilities (BDMIs) (Bateson 1909; Dobzhansky 1937; Muller 1942) is to fix one of either parental haplotypes. The probability of fixing one of the two parental haplotype and generate an incompatibility, strongly depends on admixture proportion, selection coefficient, linkage architecture and dominance relation (Schumer *et al.* 2015; Blanckaert and Bank 2018). Nevertheless, some particular conditions can favor the fixation of the ancestral haplotype (Blanckaert and Bank 2018). If the resolution of genetic conflict between *D. labrax* and *D. punctatus* resulted in the preferential fixation of ancestral haplotypes at incompatibility loci, then the divergence between *M. saxatilis* and the Atlantic *D. labrax* lineage would be lower in these genomic regions, which could explain the observed reduction in dS values at RI loci. Therefore, the dN/dS ration of RI-genes was probably influenced by ancient introgression in the Atlantic *D. labrax*, but not in the Mediterranean lineage where it seems that RI genes tend to be under stronger purifying selection than non-RI genes.

We then wanted to determine whether changes occurring in RI genes are either more or less adaptative than those occurring in other genes. For the eastern Mediterranean population, genes involved in RI clearly tend to evolve through more adaptative changes, as they show higher values of both α and ωa along with lower values of ωna, compared to non-RI genes (Figure 3C and Supplementary Figure 5E-F orange). However, the opposite patterns were observed for the Atlantic and western-Mediterranean populations (Figure 3C and Supplementary Figure 5E-F grey and blue). One thing that is common to these two populations is the presence of anciently introgressed alleles in regions involved in RI. Indeed, Atlantic genomes contain *D. punctatus* alleles (Duranton *et al.* 2019) and even if these genomic regions are involved in RI, Atlantic haplotypes can still introgressed at low rates within western-Mediterranean genomes (Duranton *et al.* 2018). However, introgressed alleles are fixed in the Atlantic population while they are polymorphic in the wester-Mediterranean population, thus deeply impacting the shape of the synonymous and non-synonymous site frequency spectrum (SFS). Since estimation of α are made with a correction based on these SFS, they might be biased by the effect of introgression in the western-Mediterranean populations. However, we are more confident on our estimations for the Atlantic population. Therefore, If the shift of tendency between the Atlantic and eastern-Mediterranean population is due to the presence of introgressed *D.*

punctatus alleles in the Atlantic population, then our results tend to indicate that this introgression was not adaptative for the Atlantic lineage. Patterns of molecular evolution thus tend to indicate that RI genes are usually under stronger purifying selection pressures (lower values of dN/dS for both Mediterranean populations) while they fixed more adaptive changes during the recent Mediterranean population history (higher values of α for the eastern-Mediterranean population) than non-RI genes. In order, to determine if this is also true on a deeper evolutionary scale, we compared conservation scores between SNPs located within RI and non-RI genes. We found that whatever the level of stringency used to detect conserved elements (5%, 25% and 50%) and for both phylogenetic analysis levels (family level or Perciformes level), RI-genes tend to be more constrained that non-RI genes (Figure 4, pink compared to blue). Our analysis on a deeper phylogenetic scale thus confirms our previous results indicating that RI-genes tend to evolve under stronger purifying selection than non-RI genes.

Function of RI genes

Finally, we wanted to determine in which molecular and cellular functions the genes associated to RI were involved. One function that was found clearly overrepresented in this set of genes compared to the rest of the genome, was the ion transmembrane transport. The European sea bass is an euryhaline fish found in environments with a wide variety of salinity levels. It has previously been shown that gene families involved in ion regulation harbored higher numbers of gene copies compared to other teleost fish genomes, providing a genetic basis for adaptation to euryhalinity (Tine *et al.* 2014). Therefore, our findings might indicate that RI between Atlantic and Mediterranean *D. labrax* lineages has been partly caused by divergent selection acting on genes involved in osmoregulation that contribute to differential adaptations between Atlantic and Mediterranean environment. Nevertheless, we cannot exclude that thoese genes are involved in RI due to genetics incompatibilities. Indeed, RI in the European sea bass seems to be highly polygenic (Duranton *et al.* 2018), and other mechanisms such as those related to past admixture with the spotted sea bass are probably also at play, among others. Thus, even if local adaptation seems to play a role in RI, we cannot say from these results that it is the main driver of RI. Indeed, other processes such as BDMIs that are a good substrate for speciation (Presgraves 2010) are not expected to target a particular gene family and would thus be unnoticed with this kind of analyses. Interestingly and contrary to what we could expect, we did not find an enrichment for gene linked to mitochondrial function. Indeed, mitochondria does not flow freely between the two lineages (Lemaire *et al.* 2005), indicating a possible role for coevolved mitonuclear interactions which are known to cause genetic incompatibilities in other systems(Sloan *et al.* 2017; Morales *et al.* 2018).

Conclusion

From a broad perspective, our study confirms that a link exists between large-scale variation in recombination rate and molecular evolution patterns, which is due to both neutral, quasi neutral and selective processes. It also underlines that recombination hotspots with a rapidly evolving location can exist in the absence of a functional PRDM9 gene, suggesting the existence of another mechanism regulating double strand break formation and subsequent repair. From a speciation perspective, our study suggests that genes involved in RI tend to be more evolutionarily conserved over the long term, while accumulating more adaptive changes than the rest of the genome during the recent divergence history of sea bass. It also suggests that local adaptation participates in RI between the Atlantic and Mediterranean *D. labrax* lineages, through differential adaptations to salinity. However, we could not quantify the relative contribution of local adaptation compared to other processes that were not specifically targeted here. Therefore, it seems that conserved elements play a more important role in sea bass speciation than fast evolving gene, even if fast evolving features could have appeared as attractive candidates for the buildup of RI. One possible explanation for this finding is that conserved elements, due to their functional importance, could generate larger effects on RI even after accumulating small changes in DNA sequences.

Material and Methods

Whole genome resequencing and phasing

In this study, we used haplotype-resolved whole genome sequences published in (Duranton *et al.* 2019) and (Duranton *et al.* 2018). Our data set includes genomes of 68 wild *Dicentrarchus labrax* individuals originating from the Atlantic Ocean (English Channel; 14 males, $♂_{AT}$), the western (Gulf of Lion; 22 females $♀_{WME}$ and 9 males $♂_{WME}$) and eastern Mediterranean Sea (Turkey; 12 males $♂_{NEM}$ and Egypt; 11 males $♂_{SEM}$). Among these, 22 individuals were involved in experimental crossing to generate 15 different families (2 $♂_{SEM}$ x $♀_{WME}$; 2 $♂_{NEM}$ x $♀_{WME}$ and 11 $♂_{AT}$ x $♀_{WME}$). In addition to these parental genomes, we also sequenced the whole genome of one descend per family to generate parent-offspring trios in order to phase the two parental genomes using a phasing-by-transmission approach (Browning and Browning 2011). The descendants were then removed from the final data set as they only provide redundant information compared to their parents. Therefore, our dataset contains 22 genomes from unrelated individuals that were phased using phasing-by-transmission. The genomes of the 46 remaining individuals, which were not involved in crosses, were phased using the reference-based phasing algorithm Eagle2 (version 2.4) (Loh *et al.* 2016), with the 22 genomes phased-by-transmission as references. Our final dataset only contained fully phased variants (excluding indels)

with no missing data, representing a total of 5,074,249 SNPs. As an outgroup, we used the diploid genome of the *Dicentrarchus punctatus* individual originating from the Atlantic Ocean (Golf of Cadiz) that was previously sequenced (Duranton *et al.* 2019) and the haploid reference draft genome of *Morone saxatilis* that was partially aligned to the reference genome of *D. labrax*, covering 52% of the assembly (Duranton *et al.* 2018).

Estimation of local population-scaled recombination rates

We used LDHelmet v1.10 (Chan *et al.* 2012) to estimate the population-scaled recombination rate ρ ($\rho = 4N_e r$ with N_e being the effective population size and r the recombination rate) along each chromosome for each of our three *D. labrax* populations, separately (*i.e.* Atlantic and Mediterranean east and west). The same number of individuals (*i.e.* 14, which is the total number of individual available for the Atlantic population) was used in each population to avoid potential biases due to different sample sizes. LDHelmet estimates the population-sclaed recombination rate locally by analyzing patterns of linkage disequilibrium along phased genomes using a coalescent-based reversible-jump Markov chain Monte Carlo (rjMCMC) simulation approach. To improve the accuracy of the estimations, the program can use for each marker a prior probability on which nucleotide represents the ancestral state. We thus defined the most likely ancestral state of every marker by looking at the information contained in the *M. saxatilis* outgroup genome (52% of the positions were available). When the outgroup information was missing, we referred to the *D. punctatus* genome and only considered homozygous sites, since the ancestral state could not be defined if there is shared polymorphism segregating in the two species. We assigned a prior probability of 0.91 to the variant identified as ancestral and 0.03 for the three remaining alternative nucleotides to allow for uncertainty in variant orientation (Shanfelter *et al.* 2019). In cases where the ancestral state could not be identified due to missing outgroup information or shared polymorphism, we used the frequency of each nucleotide in the considered *D. labrax* population as a prior probability (Shanfelter *et al.* 2019). For that purpose, we used VCFtools v 0.1.15 (Danecek *et al.* 2011) to determine the frequency of each nucleotide in every *D. labrax* population.

LDHelmet analyses were run chromosome by chromosome within windows of 50 adjacent SNPs (-w 50). To create the likelihood tables, we used a personalized grid of recombination rates (-r 0.0 0.01 1.0 0.1 10.0) since the recombination rate of the European sea bass seems to be lower than that of *Drosophila melanogaster* (Chan *et al.* 2012; Tine *et al.* 2014). Padé files were then generated using 11 coefficients (-x 11) as recommended. Finally, we ran five independent rjMCMC simulations each with 1,000,000 iterations after 100,000 burn-in iterations and a block penalty of 10, following recommendations (-b 10 -burn_in 100000 -n 1000000). Since we did not have any empirical substitution matrix for the European sea bass, we used the Jukes-Cantor mutation matrix. For each

population, the local recombination rate was estimated as the average per-site rate calculated over the five rjMCMC runs, which was then averaged in non-overlapping 2kb and 50kb windows along the genome.

Recombination hotspots were identified as 2kb windows with a recombination rate 10 times higher than the mean recombination rate of each surrounding 50kb windows (Chan *et al.* 2012; Shanfelter *et al.* 2019). If several consecutive hotspots were found within the same 10kb region, we only kept the 2kb window with the highest recombination rate (Shanfelter *et al.* 2019). Finally, to avoid possible artefacts created by the rjMCMC simulation, we only retained hotspots that were detected in all five independent runs. Hotspots were not called within 2kb windows containing less than 5 SNPs. Finally, hotspots were considered to be shared between two *D. labrax* populations if they were located within the same 6kb region.

Identification of genes involved in RI between D. labrax lineages

SNPs involved in RI were previously identified using a HMM approach analyzing the ratio between F_{ST} (calculated between ATL and WMED populations) and the fraction of Atlantic tracts introgressed within the western Mediterranean population (F_{intro}) (see Duranton *et al.* 2019). The rationale behind this test is that only the regions involved in RI are characterized by both high levels of genetic differentiation and reduced levels of introgression (Duranton *et al.* 2018). We used gene annotations from (Tine *et al.* 2014) to identify protein-coding genes involved in RI between the Atlantic and Mediterranean *D. labrax* lineages as genes containing at least one SNP previously detected in the HMM test. We also attributed to each gene a recombination rate corresponding to that of the 50kb window in which the gene seats. Given that we estimated population-scaled recombination rates ($\rho = 4N_e r$), differences in effective population sizes among the three populations were accounted for to rescale ρ values with reference to the effective population size of the Atlantic population. To do so, we first used the effective population size estimated for the Atlantic population ($N_{e\,ATL} = 100{,}372$) by Tine *et al.* (2014) to estimate the genome-wide average r value from ρ values estimated in the Atlantic *D. labrax* population. We then used the genome-wide average value of r to estimate the effective population sizes of western and eastern Mediterranean populations and calculate N_e ratios between the Atlantic and each Mediterranean population. The effective size of the Atlantic population was found to be 2.14 times larger than that of the western Mediterranean and 6.49 times larger than that of the eastern Mediterranean population. These ratios were used to rescale Mediterranean ρ values with reference to the effective population size of the Atlantic population, that is, to obtain a separate local estimate of $\rho = 4N_{e\,ATL} r$ in each of the three populations. We then averaged these rescaled ρ values across the three populations within each 50kb window to assign a recombination rate to each gene. In order to determine if there is a link between recombination rate and other molecular evolution parameters,

we defined 6 categories of recombination rate range values ([0 - $5e^{-8}$],]$5e^{-8}$ - $1e^{-7}$],]$1e^{-7}$ - $1.5e^{-7}$],]$1.5e^{-7}$ - $2e^{-7}$],]$2e^{-7}$ - $2.5e^{-7}$] and]$2.5e^{-7}$ - $1.2e^{-6}$]), each containing a fairly close number of genes not identified as being involved in RI (3156, 3268, 3870, 2387, 1947, 2055). We verified that RI-associated genes are disproportionally located in low-recombining regions (Duranton *et al.* 2018) by comparing the recombination rate distributions of genes involved and not involved in RI (Supplementary Figure 1). In order to compare patterns of molecular evolution between categories of genes involved and not involved in RI without the confounding effect of recombination, we selected a subset of genes not involved in RI presenting a similar distribution of recombination rate values to that of RI-associated genes. This was done chromosome by chromosome by subsampling genes not involved in RI that best reproduce the observed distribution of RI-associated genes (Supplementary Figure 2).

Estimation of πN/πS, dN/dS and α ratios

We computed the ratio of non-synonymous (πN) to synonymous (πS) nucleotide diversity (πN/πS) within populations, as well as the ratio of non-synonymous (dN) to synonymous (dS) substitutions (dN/dS) between species using the software dNdSpiNpiS_1.0 developed within the PopPhyl project (https://kimura.univ-montp2.fr/PopPhyl/index.php?section=tools). We set the transition/transversion ratio (κ) to 2.7, corresponding to the median value of κ for genes located on chromosome LG1A, and estimated via bppml (version 2.4) (Guéguen *et al.* 2013). SNPs for which the outgroup was polymorphic or displayed a mutational state not found in any of the focal *D. labrax* populations were discarded. Non-synonymous to synonymous ratios were first computed for all genes not involved in RI that were sorted within 6 recombination rate classes, in order to determine if molecular evolution parameters are influenced by recombination. We then controlled for recombination rate variation to test for differences in selective constraints between RI-associated and non-RI-associated genes. We therefore computed these same ratios for all genes involved RI and compared them to a subset of genes not involved in RI but presenting similar recombination rate distributions to RI-associated genes. All ratio values were computed independently for the three *D. labrax* populations.

We also estimated the proportion of adaptive amino-acid substitutions (α), the adaptive substitution rate (ωa) and the non-adaptive substitution rate (ωna) for the same categories of genes using the approach of (Eyre-Walker and Keightley 2009) as implemented in (Galtier 2016) (program Grapes v.1.0). The program models the deleterious effect of non-synonymous mutations using polymorphism data by fitting a Gamma distribution of the fitness effects (DFE) of mutations to the synonymous and non-synonymous site frequency spectra (SFS), computed for each set of genes. The DFE estimated from the distortion of the non-synonymous compared to the synonymous SFS is then used to predict

548 the expected dN/dS ratio under near-neutrality. The difference between observed and expected dN/dS
549 provides an estimate of the proportion of adaptive non-synonymous substitutions, α. The rate of
550 adaptive and non-adaptive amino-acid substitution were then obtained as following: ωa = α(dN/dS)
551 and ωna = (1-α)(dN/dS). We chose to model the negative component of the DFE as a negative gamma
552 distribution and the positive component of the DFE as an exponential (i.e. model "GammaExpo" in
553 (Galtier 2016)). When estimating DFE model parameters, we accounted for recent demographic effects
554 by using nuisance parameters, which correct each class of frequency of the synonymous and non-
555 synonymous SFS relative to the neutral expectation in an equilibrium Wright–Fisher population (Eyre-
556 Walker *et al.* 2006).

557 ***Inference of nucleotide conservation scores***

558 To identify evolutionarily conserved sequence elements and estimate conservation scores across
559 multispecies genome alignments, we ran a phylogenetic comparative analysis using the program
560 *phastCons* (Siepel *et al.* 2005; Hubisz *et al.* 2011). The analyses were run at two different phylogenetic
561 levels, first the *Moronidae* family level, using the aligned genomes of *D. punctatus*, *D. labrax* and *M.*
562 *saxatilis* (Duranton *et al.* 2018, 2019). Secondly, the Perciformes order level, aligning the stickleback
563 (*Gasterosteus aculeatus*) genome to the *D. labrax* reference, following the pipeline developed by the
564 UCSC for the analysis of conservation scores (genomewiki.ucsc.edu/index.php/Whole_genome
565 _alignment_howto). The program *phastCons* is based on a phylogenetic hidden markovian model
566 (phylo-HMM) implemented in the *R* package *RPHAST* (Siepel *et al.* 2005; Hubisz *et al.* 2011) in which a
567 regional alignment is the result of either a conserved or a non-conserved model of evolution. The only
568 parameter that differs between both models is the branch lengths of the locally inferred phylogenetic
569 tree that are scaled down by a factor *ρ* in the conserved model. A multi-species alignment containing
570 only four-fold degenerated sites was built to construct the phylogenetic tree under the neutral model
571 to be used in the HMM. To run the phylo-HMM model, the expected length of the predicted conserved
572 elements (ω) and the percentage of the genome that is covered by conserved elements (γ) were fixed
573 to specific values. In order to have enough phylogenetic information to produce reliable conservation
574 scores, we fixed ω to 100 bp at the family level and 50 bp for the Perciformes analysis level due to
575 larger phylogenetic distance. The target coverage parameter γ was fixed to either 5%, 25% and 50% of
576 the genome in order to analyze the most conserved elements with different levels of stringency. We
577 then divided SNPs into four categories depending on whether they were identified as being involved
578 in RI and if they were located inside a protein-coding gene before comparing their conservation scores.

579 ***Functional enrichment analysis***

580 We finally tested if the genes present in genomic regions associated to RI were involved in a particular
581 function. We thus used Gorilla tool (Eden *et al.* 2007, 2009) to determine overrepresented Gene

582 Ontology (GO) categories in the pool of genes identified as involved in RI (n=1594 genes) compared to
583 the rest of the genome (n=14371). We used the gene annotations from Tine *et al.* (2014) and compared
584 to the database for *Homo sapiens* in Gorilla. Over the 15965 genes, 14283 were recognized, 3509 were
585 discarded because they were duplicated, living us with 10583 genes that could be associated to a GO
586 term. Results were visualized using the REViGO online tool (Supek *et al.* 2011) (http://revigo.irb.hr/) that
587 remove redundant and similar GO terms.

588

589 **Ackowledgement**

590 This work was founded by the ANR grant CoGeDiv (ANR-17-CE02-0006-01 to P.-A.G). We are grateful
591 to Thibault Leroy for his help on the bioinformatic analyses and to Pierre-Louis Stenger for his help on
592 the functional enrichment analyses.

Baker C. L., S. Kajita, M. Walker, R. L. Saxl, N. Raghupathy, *et al.*, 2015 PRDM9 Drives Evolutionary Erosion of Hotspots in Mus musculus through Haplotype-Specific Initiation of Meiotic Recombination. PLOS Genetics 11: e1004916. https://doi.org/10.1371/journal.pgen.1004916

Baker Z., M. Schumer, Y. Haba, L. Bashkirova, C. Holland, *et al.*, 2017 Repeated losses of PRDM9-directed recombination despite the conservation of PRDM9 across vertebrates. eLife 6. https://doi.org/10.7554/eLife.24133

Bateson W., 1909 Heredity and variation in modern lights. Darwin and modern science 85–101.

Baudat F., and B. de Massy, 2007 Regulating double-stranded DNA break repair towards crossover or non-crossover during mammalian meiosis. Chromosome Res 15: 565–577. https://doi.org/10.1007/s10577-007-1140-3

Billings T., E. D. Parvanov, C. L. Baker, M. Walker, K. Paigen, *et al.*, 2013 DNA binding specificities of the long zinc-finger recombination protein PRDM9. Genome Biology 14: R35. https://doi.org/10.1186/gb-2013-14-4-r35

Birky C. W., and J. B. Walsh, 1988 Effects of linkage on rates of molecular evolution. PNAS 85: 6414–6418.

Blanckaert A., and C. Bank, 2018 In search of the Goldilocks zone for hybrid speciation. PLOS Genetics 14: e1007613. https://doi.org/10.1371/journal.pgen.1007613

Browning S. R., and B. L. Browning, 2011 Haplotype phasing: existing methods and new developments. Nat Rev Genet 12: 703–714. https://doi.org/10.1038/nrg3054

Butlin R., A. Debelle, C. Kerth, R. R. Snook, L. W. Beukeboom, *et al.*, 2012 What do we need to know about speciation? Trends Ecol Evol 27: 27–39. https://doi.org/10.1016/j.tree.2011.09.002

Campos J. L., D. L. Halligan, P. R. Haddrill, and B. Charlesworth, 2014 The Relation between Recombination Rate and Patterns of Molecular Evolution and Variation in Drosophila melanogaster. Mol Biol Evol 31: 1010–1028. https://doi.org/10.1093/molbev/msu056

Chan A. H., P. A. Jenkins, and Y. S. Song, 2012 Genome-Wide Fine-Scale Recombination Rate Variation in Drosophila melanogaster. PLOS Genetics 8: e1003090. https://doi.org/10.1371/journal.pgen.1003090

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

Charlesworth B., 1994 The effect of background selection against deleterious mutations on weakly selected, linked variants. Genetics Research 63: 213–227. https://doi.org/10.1017/S0016672300032365

Charlesworth B., 2009 Effective population size and patterns of molecular evolution and variation. Nat Rev Genet 10: 195–205. https://doi.org/10.1038/nrg2526

Corbett-Detig R. B., D. L. Hartl, and T. B. Sackton, 2015 Natural Selection Constrains Neutral Diversity across A Wide Range of Species. PLOS Biology 13: e1002112. https://doi.org/10.1371/journal.pbio.1002112

Danecek P., A. Auton, G. Abecasis, C. A. Albers, E. Banks, *et al.*, 2011 The variant call format and VCFtools. Bioinformatics 27: 2156–2158. https://doi.org/10.1093/bioinformatics/btr330

Dobzhansky T. grigorovitch, 1937 *Genetics and the origin of species*.

Duranton M., F. Allal, C. Fraïsse, N. Bierne, F. Bonhomme, *et al.*, 2018 The origin and remolding of genomic islands of differentiation in the European sea bass. Nature Communications 9: 2518. https://doi.org/10.1038/s41467-018-04963-6

Duranton M., F. Allal, S. Valière, O. Bouchez, F. Bonhomme, *et al.*, 2019 The contribution of ancient admixture to reproductive isolation between European sea bass lineages. bioRxiv 641829. https://doi.org/10.1101/641829

Eden E., D. Lipson, S. Yogev, and Z. Yakhini, 2007 Discovering Motifs in Ranked Lists of DNA Sequences. PLOS Computational Biology 3: e39. https://doi.org/10.1371/journal.pcbi.0030039

Eden E., R. Navon, I. Steinfeld, D. Lipson, and Z. Yakhini, 2009 GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. BMC Bioinformatics 10: 48. https://doi.org/10.1186/1471-2105-10-48

Eyre-Walker A., M. Woolfit, and T. Phelps, 2006 The Distribution of Fitness Effects of New Deleterious Amino Acid Mutations in Humans. Genetics 173: 891–900. https://doi.org/10.1534/genetics.106.057570

Eyre-Walker A., and P. D. Keightley, 2009 Estimating the Rate of Adaptive Molecular Evolution in the Presence of Slightly Deleterious Mutations and Population Size Change. Mol Biol Evol 26: 2097–2108. https://doi.org/10.1093/molbev/msp119

Faria R., K. Johannesson, R. K. Butlin, and A. M. Westram, 2019 Evolving Inversions. Trends in Ecology & Evolution. https://doi.org/10.1016/j.tree.2018.12.005

Felsenstein J., 1974 The Evolutionary Advantage of Recombination. Genetics 78: 737–756.

Galtier N., and L. Duret, 2007 Adaptation or biased gene conversion? Extending the null hypothesis of molecular evolution. Trends in Genetics 23: 273–277. https://doi.org/10.1016/j.tig.2007.03.011

Galtier N., 2016 Adaptive Protein Evolution in Animals and the Effective Population Size Hypothesis. PLOS Genetics 12: e1005774. https://doi.org/10.1371/journal.pgen.1005774

Gompert Z., T. L. Parchman, and C. A. Buerkle, 2012 Genomics of isolation in hybrids. Philosophical Transactions of the Royal Society B: Biological Sciences 367: 439–450. https://doi.org/10.1098/rstb.2011.0196

Guéguen L., S. Gaillard, B. Boussau, M. Gouy, M. Groussin, *et al.*, 2013 Bio++: Efficient Extensible Libraries and Tools for Computational Molecular Evolution. Mol Biol Evol 30: 1745–1750. https://doi.org/10.1093/molbev/mst097

Hellmann I., I. Ebersberger, S. E. Ptak, S. Pääbo, and M. Przeworski, 2003 A Neutral Explanation for the Correlation of Diversity with Recombination Rates in Humans. The American Journal of Human Genetics 72: 1527–1535. https://doi.org/10.1086/375657

Hill W. G., and A. Robertson, 1966 The effect of linkage on limits to artificial selection. Genetics Research 8: 269–294. https://doi.org/10.1017/S0016672300010156

Hubisz M. J., K. S. Pollard, and A. Siepel, 2011 PHAST and RPHAST: phylogenetic analysis with space/time models. Brief Bioinform 12: 41–51. https://doi.org/10.1093/bib/bbq072

Lemaire C., J.-J. Versini, and F. Bonhomme, 2005 Maintenance of genetic differentiation across a transition zone in the sea: discordance between nuclear and cytoplasmic markers. Journal of Evolutionary Biology 18: 70–80. https://doi.org/10.1111/j.1420-9101.2004.00828.x

Lindtke D., and C. A. Buerkle, 2015 The genetic architecture of hybrid incompatibilities and their effect on barriers to introgression in secondary contact. Evolution 69: 1987–2004. https://doi.org/10.1111/evo.12725

Loh P.-R., P. Danecek, P. F. Palamara, C. Fuchsberger, Y. A. Reshef, *et al.*, 2016 Reference-based phasing using the Haplotype Reference Consortium panel. Nature Genetics 48: 1443–1448. https://doi.org/10.1038/ng.3679

Mank J. E., and J. C. Avise, 2006 Phylogenetic conservation of chromosome numbers in Actinopterygiian fishes. Genetica 127: 321–327. https://doi.org/10.1007/s10709-005-5248-0

Martin S. H., and C. D. Jiggins, 2017 Interpreting the genomic landscape of introgression. Current Opinion in Genetics & Development 47: 69–74. https://doi.org/10.1016/j.gde.2017.08.007

McVean G. A. T., S. R. Myers, S. Hunt, P. Deloukas, D. R. Bentley, *et al.*, 2004 The Fine-Scale Structure of Recombination Rate Variation in the Human Genome. Science 304: 581–584. https://doi.org/10.1126/science.1092500

Morales H. E., A. Pavlova, N. Amos, R. Major, A. Kilian, *et al.*, 2018 Concordant divergence of mitogenomes and a mitonuclear gene cluster in bird lineages inhabiting different climates. Nat Ecol Evol 2: 1258–1267. https://doi.org/10.1038/s41559-018-0606-3

Muller H., 1942 Isolating mechanisms, evolution, and temperature. Biol. Symp. 6: 71–125.

Myers S., R. Bowden, A. Tumian, R. E. Bontrop, C. Freeman, *et al.*, 2010 Drive Against Hotspot Motifs in Primates Implicates the PRDM9 Gene in Meiotic Recombination. Science 327: 876–879. https://doi.org/10.1126/science.1182363

Nachman M. W., and B. A. Payseur, 2012 Recombination rate variation and speciation: theoretical predictions and empirical results from rabbits and mice. Phil. Trans. R. Soc. B 367: 409–421. https://doi.org/10.1098/rstb.2011.0249

Noor M. A. F., K. L. Grams, L. A. Bertucci, and J. Reiland, 2001 Chromosomal inversions and the reproductive isolation of species. PNAS 98: 12084–12088. https://doi.org/10.1073/pnas.221274498

Nosil P., and D. Schluter, 2011 The genes underlying the process of speciation. Trends in Ecology & Evolution 26: 160–167. https://doi.org/10.1016/j.tree.2011.01.001

Ortiz-Barrientos D., J. Engelstädter, and L. H. Rieseberg, 2016 Recombination Rate Evolution and the Origin of Species. Trends in Ecology & Evolution 31: 226–236. https://doi.org/10.1016/j.tree.2015.12.016

Otto S. P., and B. A. Payseur, 2019 Crossover Interference: Shedding Light on the Evolution of Recombination. Annual Review of Genetics 53: null. https://doi.org/10.1146/annurev-genet-040119-093957

Parvanov E. D., P. M. Petkov, and K. Paigen, 2010 Prdm9 Controls Activation of Mammalian Recombination Hotspots. Science 327: 835–835. https://doi.org/10.1126/science.1181495

Peck J. R., 1994 A ruby in the rubbish: beneficial mutations, deleterious mutations and the evolution of sex. Genetics 137: 597–606.

Presgraves D. C., 2005 Recombination Enhances Protein Adaptation in Drosophila melanogaster. Current Biology 15: 1651–1656. https://doi.org/10.1016/j.cub.2005.07.065

Presgraves D. C., 2010 The molecular evolutionary basis of species formation. Nature Reviews Genetics 11: 175–180. https://doi.org/10.1038/nrg2718

Ravinet M., R. Faria, R. K. Butlin, J. Galindo, N. Bierne, *et al.*, 2017 Interpreting the genomic landscape of speciation: a road map for finding barriers to gene flow. J. Evol. Biol. 30: 1450–1477. https://doi.org/10.1111/jeb.13047

Rieseberg L. H., 2001 Chromosomal rearrangements and speciation. Trends in Ecology & Evolution 16: 351–358. https://doi.org/10.1016/S0169-5347(01)02187-5

Schumer M., R. Cui, G. G. Rosenthal, and P. Andolfatto, 2015 Reproductive Isolation of Hybrid Populations Driven by Genetic Incompatibilities. PLOS Genetics 11: e1005041. https://doi.org/10.1371/journal.pgen.1005041

Shanfelter A. F., S. L. Archambeault, and M. A. White, 2019 Divergent Fine-Scale Recombination Landscapes between a Freshwater and Marine Population of Threespine Stickleback Fish. Genome Biol Evol 11: 1552–1572. https://doi.org/10.1093/gbe/evz090

Siepel A., G. Bejerano, J. S. Pedersen, A. S. Hinrichs, M. Hou, *et al.*, 2005 Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res. 15: 1034–1050. https://doi.org/10.1101/gr.3715005

Sloan D. B., J. C. Havird, and J. Sharbrough, 2017 The on-again, off-again relationship between mitochondrial genomes and species boundaries. Mol Ecol n/a-n/a. https://doi.org/10.1111/mec.13959

Smith J. M., and J. Haigh, 1974 The hitch-hiking effect of a favourable gene. Genetics Research 23: 23–35. https://doi.org/10.1017/S0016672300014634

Supek F., M. Bošnjak, N. Škunca, and T. Šmuc, 2011 REVIGO Summarizes and Visualizes Long Lists of Gene Ontology Terms. PLOS ONE 6: e21800. https://doi.org/10.1371/journal.pone.0021800

Tine M., H. Kuhl, P.-A. Gagnaire, B. Louro, E. Desmarais, *et al.*, 2014 European sea bass genome and its variation provide insights into adaptation to euryhalinity and speciation. Nature Communications 5: 5770.

Supplementary Figure 1 – Distribution of recombination rate for gene involved (red) and not involved (black) in RI. The vertical dotted line represents the threshold we fixed to sort regions of low and high recombination.

Supplementary Figure 2 – **Example for the chromosome LG1B of the subsampling of gene not involved in RI to reproduce the distribution of recombination rate of genes involved in RI**. Distribution of recombination rate for gene involved (black) and not involved in RI before (red) and after the subsampling (blue).

Supplementary Figure 3 – **Correlation between recombination rate and patterns of molecular evolution.** Measures were computed independently for the Atlantic (ATL; grey), western (MEDW; blue) and eastern-Mediterranean (MEDE; orange) populations. Stars indicate the level of significance of the p-value (** when p-value < 0.01 and *** when p-value < 0.001) for the linear correlation measured including all populations. **A.** piS, **B.** piN, **C.** dS, **D.** dN, **E.** ωa and **F.** ωna.

Supplementary Figure 4 – **Patterns of molecular evolution in recombination hotspots compared to the rest of the genome.** Measures were computed for recombination hotspots (circles), regions of high (triangles) and low (squares) recombination, independently for the Atlantic (ATL; grey), western (MEDW; blue) and eastern-Mediterranean (MEDE; orange) populations. Verticals bars represent the 95% confidence intervals. **A.** piS, **B.** piN, **C.** dS and **D.** dN.

Supplementary Figure 5 – **Comparison of molecular evolution patterns.** Measures were computed independently for the Atlantic (ATL; grey), western (MEDW; blue) and eastern-Mediterranean (MEDE; orange) populations for genes involved (squares) or not (circles) in RI presenting similar recombination rates. Verticals bars represent the 95% confidence intervals. **A.** piS, **B.** piN, **C.** dS, **D.** dN, **E.** ωa and **F.** ωna.

Supplementary Table 1 – **Enrichment analysis at the level of biological process between genes involved or not in RI.**

Supplementary Table 2 – **Enrichment analysis at the level of molecular functions between genes involved or not in RI.**

CHAPITRE 4 :

Les hybrides ont-ils une valeur sélective plus faible que leurs parents ? Analyse de croisements expérimentaux

Ce chapitre a fait l'objet de l'encadrement du stage de master de 1 de Manon Leroy

Introduction

La spéciation est un processus évolutif graduel qui permet aux barrières d'isolement reproductif (IR) d'apparaître progressivement entre deux populations qui divergent génétiquement, jusqu'à mener à un isolement complet (Coyne and Orr 2004). Comprendre quelles sont les premières barrières d'isolement à se mettre en place ainsi que leur impact sur la valeur sélective des individus est l'un des objectifs principaux en biologie évolutive. En effet, ces barrières peuvent intervenir à différents stades du cycle de vie, on parle d'isolement pré-zygotique avant la fécondation et d'isolement post-zygotique après. De plus, elles peuvent être plus ou moins en lien avec des variables environnementales, certaines étant complètement dépendantes de l'environnement et d'autres pas du tout. Par exemple, deux lignées évolutives adaptées à deux habitats différents auront une probabilité réduite de se reproduire entre elles car les migrants sont mal adaptés dans l'autre habitat (Nosil *et al.* 2005). Ici, c'est l'environnement qui va réduire la probabilité de rencontre entre les individus des deux lignées, on parle donc d'isolement écologique. Complètement à l'opposé, certaines barrières sont totalement indépendantes de l'environnement, ce qui peut être le cas de celles qui résultent d'incompatibilités Dobzhansky Muller (DMI) (Dobzhansky 1937; Muller 1942). Ce type d'incompatibilité peut apparaître quand deux populations qui divergent fixent indépendamment deux nouveaux allèles à deux locus différents (DMI bilocus). Ces nouveaux allèles sont neutres ou avantageux dans leur fond génétique d'origine mais peuvent se révéler incompatibles quand combinés au sein de génotypes hybrides. Ici, ce sont donc les interactions épistatiques entre gènes qui vont générer l'IR (DMI multilocus). Si les interactions entre allèles sont plus complexes avec des relations de dominance, les effets délétères ne seront révélés que si les allèles sont à l'état homozygote ce qui ne se produit pas à la première génération d'hybridation. Dans tous les cas, quel que soit le type de barrières d'IR, c'est souvent le fait que les hybrides aient une valeur sélective plus faible que celle de leurs parents qui crée l'IR. Plus la dépression d'hybridation apparait tôt dans les générations d'hybrides, plus la barrière au flux génique est efficace (Barton and Hewitt 1985)

Identifier les régions génomiques impliquées dans l'IR afin de comprendre quelles en sont les bases génétiques est donc l'un des objectifs des études de génomique de la spéciation. Cependant, cette démarche peut être difficile chez de vraies espèces qui n'interagissent plus génétiquement tellement les barrières d'IR sont nombreuses. Etudier des lignées évolutives entre lesquelles le flux génique n'est que partiellement interrompu peut au contraire permettre d'identifier les régions génomiques imperméables aux échanges génétiques et donc impliquées dans l'IR par opposition à celles qui ne le sont pas (Harrison and Larson 2016; Ravinet *et al.* 2017). C'est pourquoi le bar européen (*Dicentrarchus labrax*) constitue un bon modèle d'étude. En effet, cette espèce est subdivisée en deux lignées

évolutives génétiquement distinctes, le bar atlantique et le loup méditerranéen qui s'hybrident naturellement au niveau de la mer d'Alboran (Lemaire *et al.* 2005). Une étude précédente de génétique des populations a montré que ces deux lignées ont commencé à diverger en allopatrie il y a environ 300 000 ans, puis se sont remises en contact il y a 11 500 ans à la fin de la dernière période glaciaire (Tine *et al.* 2014). Ce contact a permis la reprise des échanges génétiques mais de façon très asymétrique avec une introgression majoritaire d'allèles atlantiques dans le fond génétique méditerranéen. De plus, il semblerait qu'il existe une contre sélection des allèles atlantiques introgressés en Méditerranée. En effet, les niveaux de flux génique sont très hétérogènes le long du génome, les haplotypes atlantiques étant quasiment absents de certaines régions génomiques chez toutes les populations Méditerranéennes (Duranton *et al.* 2018). De plus, l'introgression est beaucoup plus fréquente dans la population de Méditerranée ouest (31%) que dans celle de l'est (13%) ce qui semble indiquer que les haplotypes atlantiques introgressés sont progressivement contre-sélectionnés pendant leur diffusion en Méditerranée (Duranton *et al.* 2018). En effet, bien que ce déséquilibre puisse être attendu, la population ouest étant plus proche de l'atlantique, les échanges génétiques durent depuis suffisamment longtemps pour que les fréquences alléliques se soient rééquilibrées entre les différentes populations en l'absence de contre- sélection. Il existe donc nécessairement une forme d'IR entre les lignées atlantique et méditerranéenne de bar européen.

Identifier les effets sélectifs des allèles introgressés peut cependant être difficile dans les populations naturelles. En effet, il est souvent très compliqué d'avoir accès aux premières générations d'hybridation chez qui la contre-sélection des combinaisons alléliques délétères est la plus visible. En raison du tri sélectif des recombinaisons génétiques générées lors de l'hybridation, les allèles incompatibles introgressent peu et sont donc rarement observables en population naturelle. Ce sont donc principalement les effets de la sélection sur le long terme qui sont identifiables dans les populations introgressées. De plus, l'IR passant principalement par la contre-sélection des individus hybrides, la mise en évidence d'un isolement reproductif entre les deux lignées de bar européen semble difficile à relier à l'existence d'une vigueur hybride chez les hybrides de première génération (Guinand *et al.* 2017). Il est donc nécessaire d'avoir recours à des croisements expérimentaux afin d'avoir accès aux premières générations d'hybridation et ainsi identifier les locus d'IR et comprendre la dynamique de contre-sélection des allèles atlantiques.

Nous avons donc cherché à tester s'il existe une dépression d'hybridation chez des individus issus de rétrocroisements méditerranéens (*backcross*-MED) afin d'étudier le sens majoritaire de l'introgression en nature. Nous avons donc comparé la valeur sélective d'individus ouest-méditerranéens (MED) par rapport à celle des *backcross*-MED à différents stades du cycle de vie. Nous avons, pour cela, comparé les taux de fécondation, de mortalité et de croissance ainsi que l'indice de condition des individus

backcross-MED et MED à différents temps. Cette étude néglige cependant une part importante de la valeur sélective qui est déterminé par la capacité des individus à se reproduire. En effet, la maturité sexuelle étant atteinte à l'âge de trois ans chez le bar européen, nous n'avons pas pu mesurer de paramètres en lien avec la fertilité. Enfin, 380 individus *backcross*-MED ont été génotypés à l'aide d'une puce 57000 SNPs afin de comparer les patrons d'introgression des allèles atlantiques sur le court et long terme.

Matériels et Méthodes

1. Croisements expérimentaux

Deux types de croisements expérimentaux ont été réalisés. Des croisements intra-lignées entre individus ouest-méditerranéens pour générer des individus méditerranéens qui servent de contrôle. Des rétrocroisements entre des femelles ouest-méditerranéennes et des mâles hybrides F1 (issus du croisement entre une femelle ouest-méditerranéenne et un mâle atlantique) pour générer des *backcross*-MED. Ces croisements ont été réalisés à la station IFREMER de Palavas-les-Flots à partir de 10 ♀ $_{MED}$ sauvages non apparentées et de 14 ♂ parmi lesquels 7 ♂ $_{MED}$ et 7 ♂ $_{F1}$ issus de 4 familles grands-maternelles (Figure 1). Ces quatorze mâles ont été élevés dans les mêmes conditions et ont le même âge. 72 familles ont été produites, 36 méditerranéennes et 36 *backcross*-MED. Pour chaque famille, 10 ml d'œufs ont été fécondés artificiellement avec une quantité identique de sperme de chaque mâle. Les œufs ont été répartis dans 20 incubateurs différents, deux par femelle, un pour les œufs fécondés par un mâle méditerranéen et un pour les œufs fécondés par un mâle F1.

Pour chaque femelle une petite quantité du mélange d'œufs fécondés par les mâles méditerranéens et hybrides a été récupérée afin d'estimer un taux de fécondation pour chaque femelle en fonction de l'origine des mâles. Pour cela, trois personnes ont compté indépendamment le nombre d'œufs effectivement fécondés sur un échantillon de 100. Quarante-huit heures après la fécondation, les œufs vivants ont été triés des œufs non-fécondés et morts afin d'estimer un taux de mortalité embryonnaire pour chaque femelle en tenant compte de l'origine des pères. Pour chaque femelle la même quantité d'œuf fécondés par des méditerranéens et des hybrides a été récupérée afin d'obtenir un mélange d'œufs 50% MED 50% *backcross*-MED. La même quantité d'œufs n'a pas pu être récupérée pour toutes les femelles au vu des différences de taux de fécondation et de survie mais la contribution relative de chaque femelle au pool de départ a été prise en compte pour la suite des analyses. Les larves, juvéniles et adultes ont ensuite été élevées en environnement commun dans les conditions standardisées d'aquaculture.

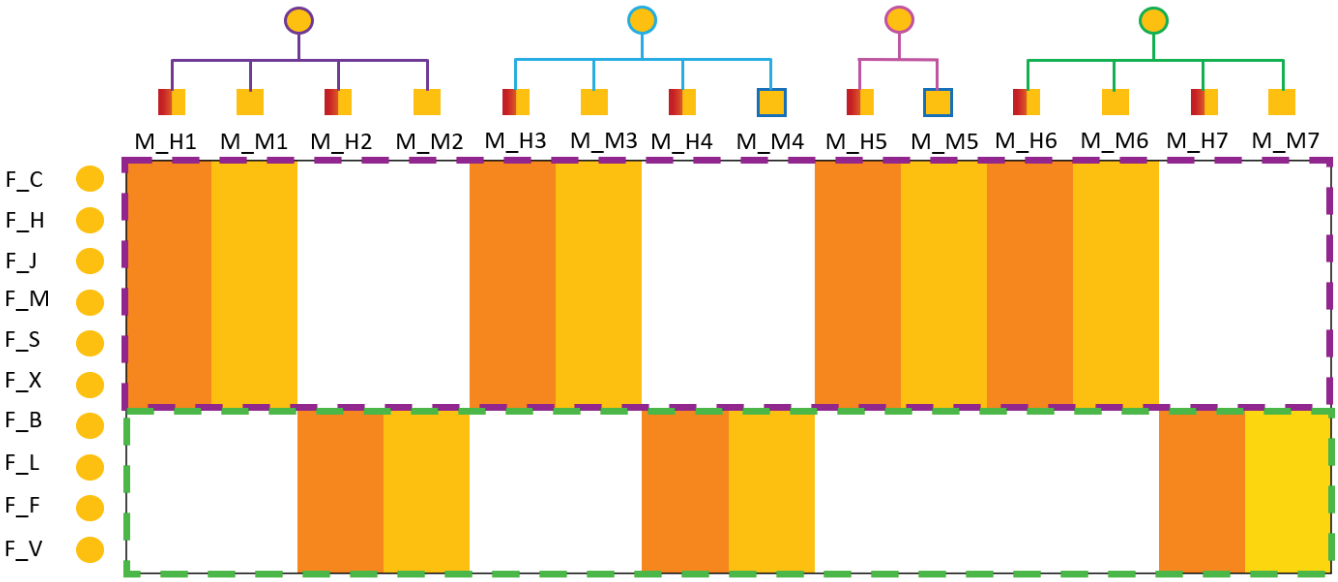


FIGURE 1 – Plan de croisement. Les ronds symbolisent les individus femelles (F) et les carrés les mâles (M). Les individus méditerranéens sont représentés en jaune, les *backcross*-MED en orange et les hybrides F1 issus du croisement entre une femelle méditerranéenne et un mâle atlantique en rouge/jaune. Les relations de parentés entre les mâles sont représentées au-dessus de leurs noms. Ils sont issus de 4 femelles différentes et ont tous des pères différents sauf deux mâles (carrés jaunes aux contours bleus). Les cases colorées dans le tableau indiquent les individus qui ont été effectivement croisés entre eux. Il y a deux blocs de croisements, le premier composé de 6 femelles croisées avec les 8 mêmes mâles (pointillés violets) et un de 4 femelles croisées avec les 6 mêmes mâles (pointillés verts).

1. Biométrie

La première biométrie a eu lieu 317 jours après l'éclosion. 1536 juvéniles ont été anesthésiés avec de la benzocaïne avant d'être soumis à une chaîne de biométrie composée de cinq étapes. Premièrement, une puce RFID sous-cutanée leur a été implantée au niveau de la nageoire pectorale droite afin d'identifier individuellement chaque juvénile. Ils ont ensuite été pesés et mesurés et une photo de chaque poisson sur sa face latérale droite a été prise à l'aide d'un appareil photo numérique fixé sur un pied. Les malformations classiquement observées en élevage aquacole ont été renseignées afin de ne pas biaiser les analyses morpho-anatomiques. Enfin un morceau de la nageoire caudale a été découpé et conservé dans de l'éthanol afin de réaliser des extractions d'ADN pour pouvoir génotyper les individus et déterminer leur origine. La deuxième biométrie des individus survivants, au cours de laquelle seul le poids et la taille des individus ont été remesurés a eu lieu environ 1 an plus tard (soit 681 jours après l'éclosion).

2. Génotypage et assignations parentales

Les 1536 individus ayant été inclus dans la biométrie ont été génotypés sur 48 SNPs choisi pour être présents en fréquence intermédiaire dans les populations atlantiques et méditerranéennes de bar européen (fréquence de l'allèle mineur > 0,25) et répartis sur l'ensemble des 24 chromosomes pour ne pas être physiquement liés. Le but ici étant d'identifier les parents des descendants et non pas de distinguer les atlantiques des méditerranéens, nous n'avons pas choisi des SNPs diagnostiques mais des SNPs en fréquence intermédiaire pour avoir une plus grande probabilité d'observer du polymorphisme chez les individus et augmenter la puissance d'assignation parentale. Les extractions d'ADN ainsi que le génotypage ont été réalisés par la plateforme de génotypage GENTYANE à Clermont-Ferrand. Les résultats de génotypage consistent en des signaux d'intensité de fluorescence pour chaque SNP et chaque individu qui sont analysés grâce à une méthode de classification non-hiérarchique qui permet de classer les individus en trois groupes (les deux homozygotes et l'hétérozygote). Pour assigner un individu à un groupe donné, la méthode regarde la distance de chaque individu par rapport au barycentre des trois groupes précédemment définis. Nous avons sélectionné un seuil de distance de 80%, plus strict que celui de 65% par défaut afin de ne conserver que les individus pour lesquels le génotypage est sûr sans éliminer trop de données. Toutefois, il arrive que l'assignation ne se fasse pas correctement (Annexe 4 Figure 1A). C'est pourquoi les données ont été traitées manuellement afin de réaliser des corrections d'affectation en se basant sur la corrélation entre les fréquences alléliques des parents et des descendants, qui est supposée être forte si les génotypes sont corrects. Nous avons ainsi pu identifier les individus qui s'écartaient de cette corrélation dû à des erreurs de génotypage (Annexe 4 Figure 1B) et corriger leur génotype pour les

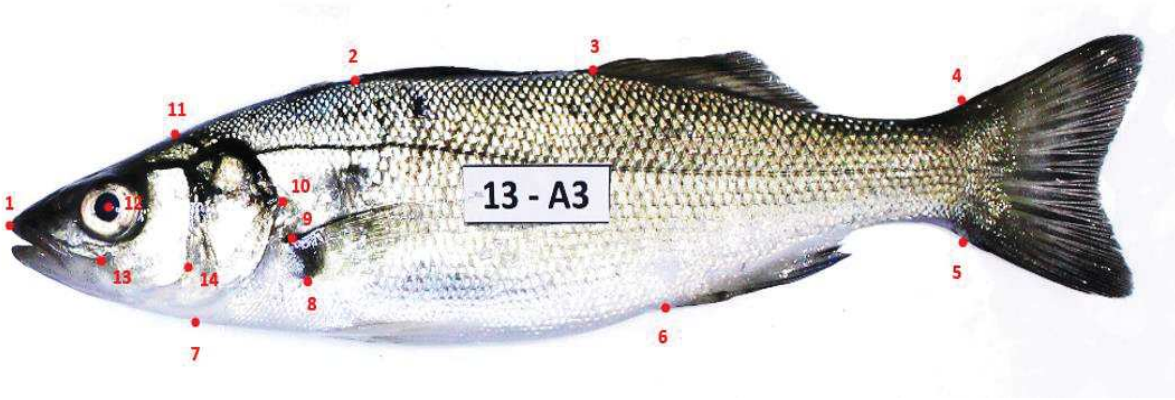


FIGURE 2 – Positionnement des points morpho-anatomiques utilisés pour cette étude. **1**: extrémité de la mâchoire supérieure, **2**: insertion antérieure de la première nageoire dorsale, **3**: insertion antérieure de la seconde nageoire dorsale, **4**: insertion dorsale de la nageoire caudale, **5**: insertion ventrale de la nageoire caudale, **6**: insertion antérieure de la nageoire anale, **7**: insertion ventrale de l'opercule, **8**: insertion ventrale de la nageoire pectorale, **9**: insertion dorsale de la nageoire pectorale, **10**: extrémité postérieure de l'opercule, **11**: extrémité dorso-postérieure de la tête et début du filet dorsal, **12**: centre de l'orbite oculaire, **13**: extrémité postérieure de la lèvre postérieure, **14**: extrémité ventro-postérieure de la joue

locus concernés (Annexe 4 Figure 1C) en veillant à ce que le coefficient de corrélation des fréquences alléliques entre parents et descendants reste inchangé. Afin de réaliser les assignations parentales, nous n'avons conservé que les individus clairement génotypés sur au moins 24 SNPs soit 1343 individus. Les assignations parentales ont été réalisées par maximum de vraisemblance avec le programme COLONY2 (Jones and Wang 2010) qui utilise les génotypes multi-locus pour inférer les relations de parenté et de fratrie. Les assignations ont été réalisées en paramétrant pour une haute précision de recherche du maximum de vraisemblance et un taux d'erreur de génotypage de 2,5% qui a été ajusté pour les locus présentant des problèmes de génotypage. Par la suite nous n'avons conservé que les individus assignés avec une probabilité supérieure à 0,95 soit 995 individus afin d'estimer les taux de survie. Pour vérifier la puissance de COLONY2 sur notre jeu de données, nous avons fait des simulations. A partir des génotypes parentaux connus, 30 descendants ont été simulés par famille (soit 2160 individus) en faisant pour chacun des 48 locus un tirage aléatoire parmi les allèles parentaux. Nous avons ainsi pu évaluer le taux d'erreur d'assignation de COLONY2 en fonction du seuil de probabilité d'assignation utilisé.

Parmi les individus *backcross-MED* précédemment identifiés, 380 répartis entre les différentes familles ont été sélectionnés pour être génotypés à l'aide d'une puce à ADN de 57000 SNPs. Les SNPs sélectionnés sur la puce sont répartis de façon homogène le long du génome en fonction des variations locales du taux de recombinaison, les régions à plus fort taux de recombinaison contenant plus de SNPs que celles à faible taux. Les données ont ensuite été filtrées afin d'éliminer les SNPs présentant des erreurs de génotypage (des homozygotes des deux types mais pas d'hétérozygotes ce qui n'est pas possible au vu du plan de croisement) ce qui nous a permis d'obtenir un total de 50517 SNPs pour les 380 individus.

3. Analyses morpho-anatomiques

Nous avons, dans un premier temps, utilisé la taille et le poids mesurés des individus afin de déterminer s'il existe des différences de valeur sélective entre les MED et les *backcross*-MED. Nous avons calculé le facteur de condition de Le Cren (K) comme le rapport du poids observé au poids attendu par la régression linéaire poids-taille. Si K>1 le poisson est en bonne condition, si K<1 alors il est en mauvaise condition. La croissance (G) à quant à elle été calculée comme le ratio de la taille (mm) sur l'âge des individus (jours). Les valeurs de G et K ont ensuite été centrées-réduites. Nous avons ensuite utilisé les photos afin de déterminer s'il existe des différences morphologiques entre les MED et les *backcross*-MED. Pour cela, 14 points morpho-anatomiques durs (Figure 2) ont été positionnés manuellement à l'aide du logiciel TPS Dig 2 (ROHLF 2006). Ces points anatomiques sont couramment utilisés pour les études morphologiques chez *D. labrax* (Vandeputte *et al.* 2017).

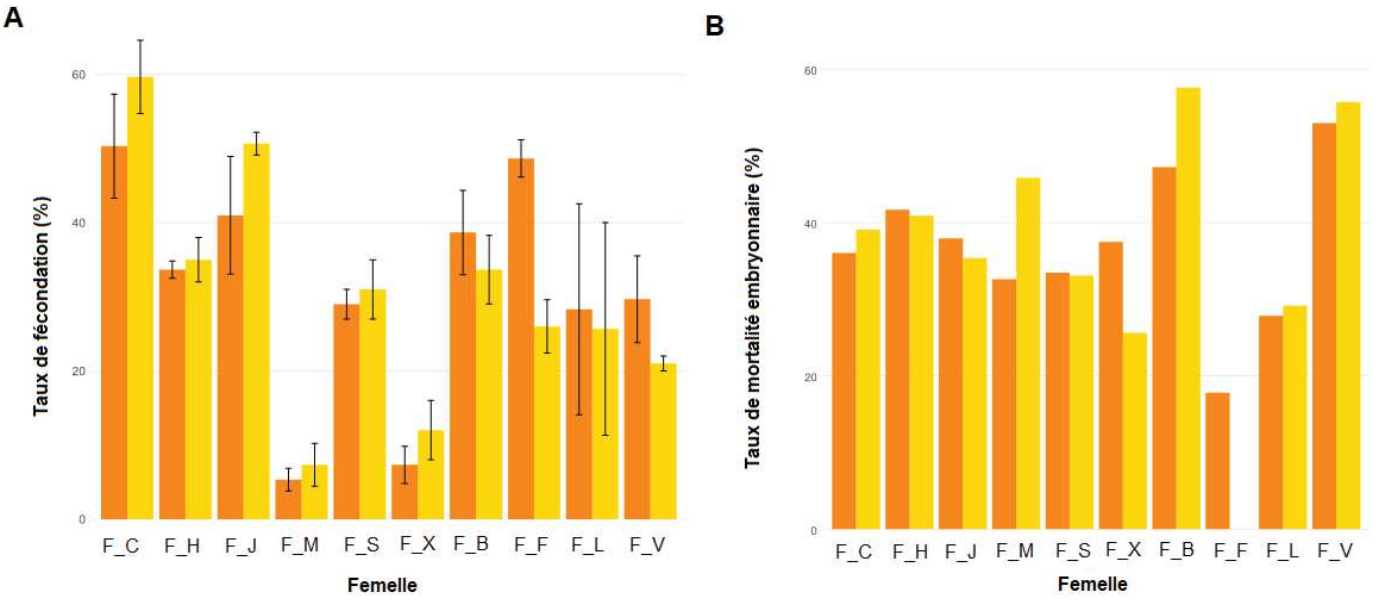


FIGURE 3 – Taux de fécondation et de mortalité embryonnaire des descendants de chaque femelle. La couleur représente l'origine des pères, orange pour les hybrides et jaune pour les méditerranéens. **A.** taux de fécondation, les barres d'erreurs représentent l'intervalle de confiance, les mesures ayant été faites par 3 personnes. **B.** Taux de mortalité embryonnaire.

Variable réponse	Vraisemblance	Effets fixes		P-value		Test de Wald
		Backcross-MED	MED	*Backcross*-MED	MED	Comparaison des deux effets estimés
Taux de fécondation	-33	-1,25	-1,29	*	*	0,93
Taux de mortalité embryonnaire	-12	-0,55	-0,56	0,40	0,39	-

TABLEAU 1 – Effets fixes estimés pour la meilleure vraisemblance des GLMM. Les étoiles symbolisent la P-value (* : < 0,05 ; ** : < 0,01 ; *** : < 0,001). La P-value indique si les effets fixes estimés sont significativement différents de 0 et le test de Wald s'ils sont significativement différents pour les MED et les *backcross*-MED.

Nous avons dans un premier temps réalisé une analyse Procruste qui permet d'homogénéiser la morphologie des individus indépendamment de leur taille relative à l'aide du package R Geomorph (Adams and Otárola-Castillo 2013). Nous avons ensuite utilisé une Analyse en Composante Principale (ACP) à partir des coordonnées Procruste et extrait les coordonnées de chaque individu sur les différents axes. A l'aide de régression linéaire nous avons ensuite déterminé quels axes capturaient le mieux la variance de croissance et de condition. Afin de mieux visualiser les changements morphologiques portés par chaque axe de l'ACP des grilles de déformation ont été générées à l'aide du package R Geomorph et du logiciel Morpho J (Klingenberg 2011).

4. Analyses statistiques

Afin de déterminer s'il existe des différences de taux de fécondation, de mortalité, de survie et morpho-anatomiques entre les MED et les *backcross*-MED nous avons utilisé des modèles linéaires mixtes généralisés (GLMM). Le plan de croisement permet de prendre en compte les facteurs confondants comme les effets parentaux qui ont donc toujours été estimés comme des effets aléatoires. La forme générale d'un GLMM est $Y_i = \beta T_i + bZ_i + \varepsilon_i$ où Y_i correspond à l'observation associée au descendant *i* parmi les *n* pour la variable réponse analysée, T_i est le vecteur reliant la variable explicative (ici le type de descendant *backcross*-MED ou MED) au vecteur d'effets fixes β, Z_i est le vecteur contenant les effets *b* des r variables aléatoires associées à l'observation *i* et ε_i correspond à la valeur résiduelle de cette observation. Deux approches ont été utilisées (quand c'était possible en fonction de la forme de la distribution des données) pour analyser les modèles, une par maximum de vraisemblance (MV) avec la package R glmm et une autre par approche bayésienne (B) avec le package MCMCglmm. Afin de déterminer si les différences entre les β observés étaient significatives, le test de WALD (Wald and Wolfowitz 1948) a été utilisé pour l'approche par MV et les intervalles de crédibilité à 95% ont été comparés pour l'approche B.

5. Lien génotype-phénotype chez les individus *backcross*-MED

Afin de comprendre comment les allèles atlantiques et méditerranéens interagissent dans les génomes des *backcross*-MED nous avons regardé la corrélation entre la fréquence de l'allèle méditerranéen chez les parents et les *backcross*-MED. De fortes distorsions dans la corrélation ne sont pas attendues même en présence d'effets sélectifs mais sont plus probablement dues à des erreurs de génotypage. Nous avons donc éliminé du jeu de données tous les SNPs qui s'écartaient significativement de la corrélation, ce qui nous a laissé 49 993 SNPs. Parmi tous ces SNPs nous nous sommes ensuite intéressés plus particulièrement à ceux précédemment identifiés à l'aide d'un modèle de Markov caché (cf Chapitre 2) comme impliqués dans l'IR entre la lignée atlantique et méditerranéenne (762). Nous avons également regardé les SNPs impliqués dans l'IR et différentiellement fixés entre les populations

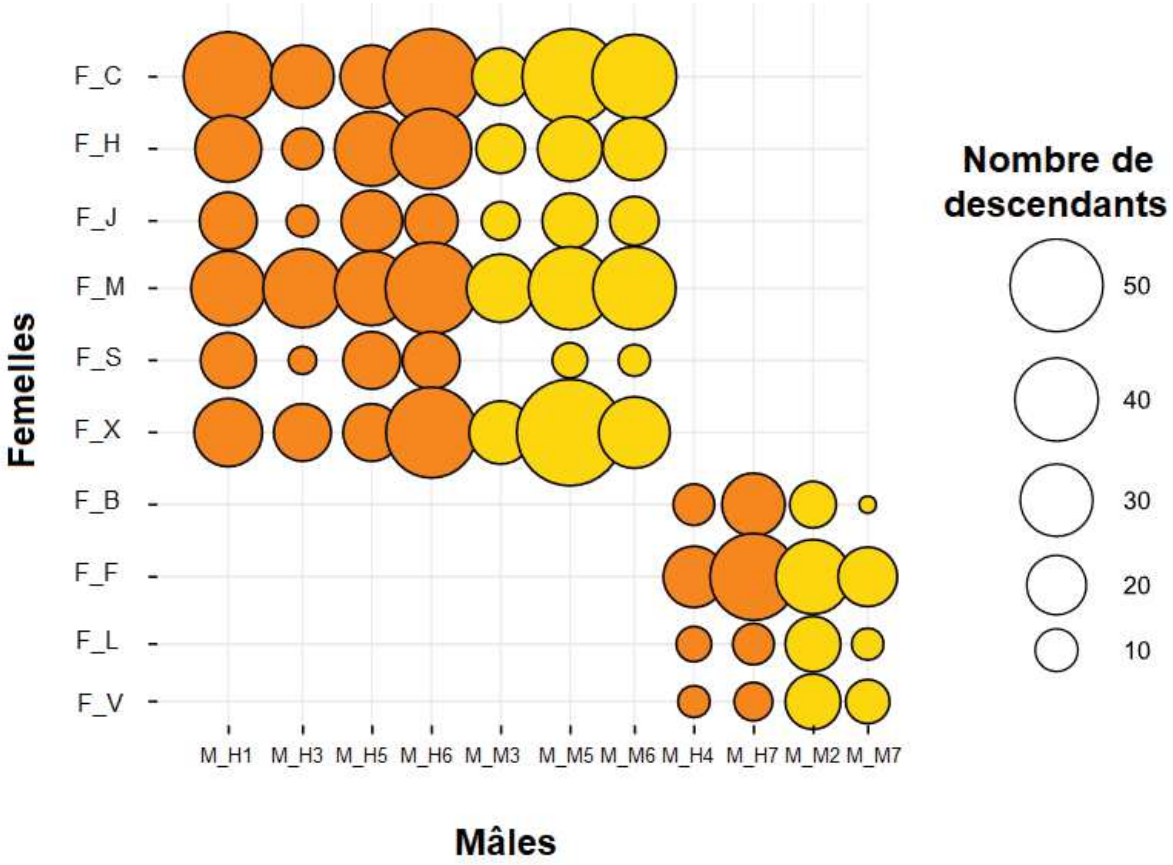


FIGURE 4 – Nombre de descendants des 58 familles considérées dans l'étude. La taille des cercles est proportionnelle au nombre de descendants et la couleur représente le type de croisement méditerranéen (jaune) ou *backcross*-MED (orange). Le nom des parents est le même qu'en figure 1.

Variable réponse	Vraisemblance	Effets fixes estimés		P-value		Comparaison des effets fixes	
		Backcross-MED	MED	*Backcross*-MED	MED	*Backcross*-MED	MED
Survie à 1 an (MV)	1960	2,65	2,56	***	***	Test de Wald : 0,32	
Survie à 1 an (B)	-	2,63	2,52	***	***	$IC_{95\%}$[2,10 ; 3,17]	$IC_{95\%}$[1,97 ; 3,08]

TABLEAU 2 – Effets fixes estimés avec un GLMM pour la survie à 1 an par l'approche du maximum de vraisemblance et bayésienne. Les étoiles symbolisent la P-value (* : < 0,05 ; ** : < 0,01 ; *** : < 0,001). La P-value indique si les effets fixes estimés sont significativement différents de 0 et le test de Wald s'ils sont significativement différents pour les MED et les *backcross*-MED pour l'approche du maximum de vraisemblance. Pour l'approche bayésienne la comparaison des intervalles de confiance permet de dire si les effets sont significativement différents.

naturelles (127) ou différentiellement fixés entre les géniteurs (156). Afin de déterminer s'il existe un lien entre le génotype des individus hybrides et leur phénotype, nous nous sommes concentrés sur les SNPs différentiellement fixés entre les géniteurs et impliqués dans l'IR. Nous avons représenté chaque individu comme un point dans un triangle qui indique leur indice hybride (c'est-à-dire la proportion d'allèles qui sont originaire de Méditerranée, les individus atlantiques ont donc un indice hybride de 0 et les méditerranéens un indice hybride de 1) et leur niveau d'hétérozygotie (les individus atlantiques et méditerranéens étant complètement homozygotes et les F1 complètement hétérozygotes). Nous avons ensuite cherché à voir si la position des individus dans ce triangle pouvait expliquer leur condition ou leur croissance.

Résultats

1. Taux de fécondation et de mortalité embryonnaire

Le taux de fécondation et de mortalité embryonnaire des descendants de chaque femelle a été comparé en prenant en compte l'origine des pères (Figure 3). Nous avons utilisé des GLMM afin de déterminer s'il existe un effet significatif de l'origine des pères en intégrant en effet aléatoire les effets parentaux (Tableau 1). Chaque modèle a été ajusté 5 fois indépendamment par maximum de vraisemblance afin de vérifier la bonne convergence des valeurs de paramètres estimés. Pour le taux de fécondation les effets fixes sont significativement différents de 0 pour les deux types de croisements ($p < 0,005$ dans les deux cas) mais ne sont pas significativement différents l'un de l'autre (Test de Wald $p = 0,93$). Au contraire on ne détecte aucun effet significatif de l'origine des pères sur le taux de mortalité embryonnaire. L'origine des pères n'a donc pas d'influence sur le taux de fécondation ni sur le taux de mortalité embryonnaire.

2. Assignations parentales et mesure de la survie post-embryonnaire

Sur les 14 mâles ayant participé aux croisements, 3 mâles dont deux méditerranéens et un hybride n'ont été assignés à aucun descendant, le programme ayant inféré 3 nouveaux génotypes. En effet, COLONY2 peut supposer que tous les génotypes parentaux ne sont pas connus et ainsi inférer des génotypes parentaux probables. Ce problème semble être couramment rencontré quand le programme est utilisé avec des données SNPs (sachant qu'il peut aussi utiliser des données microsatellites). Nous avons donc décidé d'éliminer les individus assignés à ces trois mâles fictifs afin de ne conserver que les parents réels. Nous n'avons conservé que les individus assignés avec une probabilité d'assignation supérieure à 0,95 soit 995 individus assignés à leurs deux parents, répartis entre 32 familles MED et 26 familles *backcross*-MED. Le nombre de descendants par famille a été standardisé par la quantité d'œufs viables récupérés 48h après la fécondation pour éliminer la variance de survie embryonnaire et ne voir que la survie post-éclosion (Figure 4). Nos simulations nous ont

Variable réponse	Effets fixes estimés		P-value		$IC_{95\%}$	
	Backcross-MED	MED	*Backcross*-MED	MED	*Backcross*-MED	MED
G à 1 an	0,43	0,43	***	***	[0,41 ; 0,44]	[0,41 ; 0,44]
K à 1 an	1,01	1,00	***	***	[1,00 ; 1,02]	[0,98 ; 1,01]

TABLEAU 3 – Effets fixes estimés avec des GLMM pour la croissance et la condition mesurée à 1 an par l'approche bayésienne. Les étoiles symbolisent la P-value (* : < 0,05 ; ** : < 0,01 ; *** : < 0,001). La P-value indique si les effets fixes estimés sont significativement différents de 0 et la comparaison des intervalles de confiance permet de dire s'ils sont significativement différents pour les MED et les *backcross*-MED.

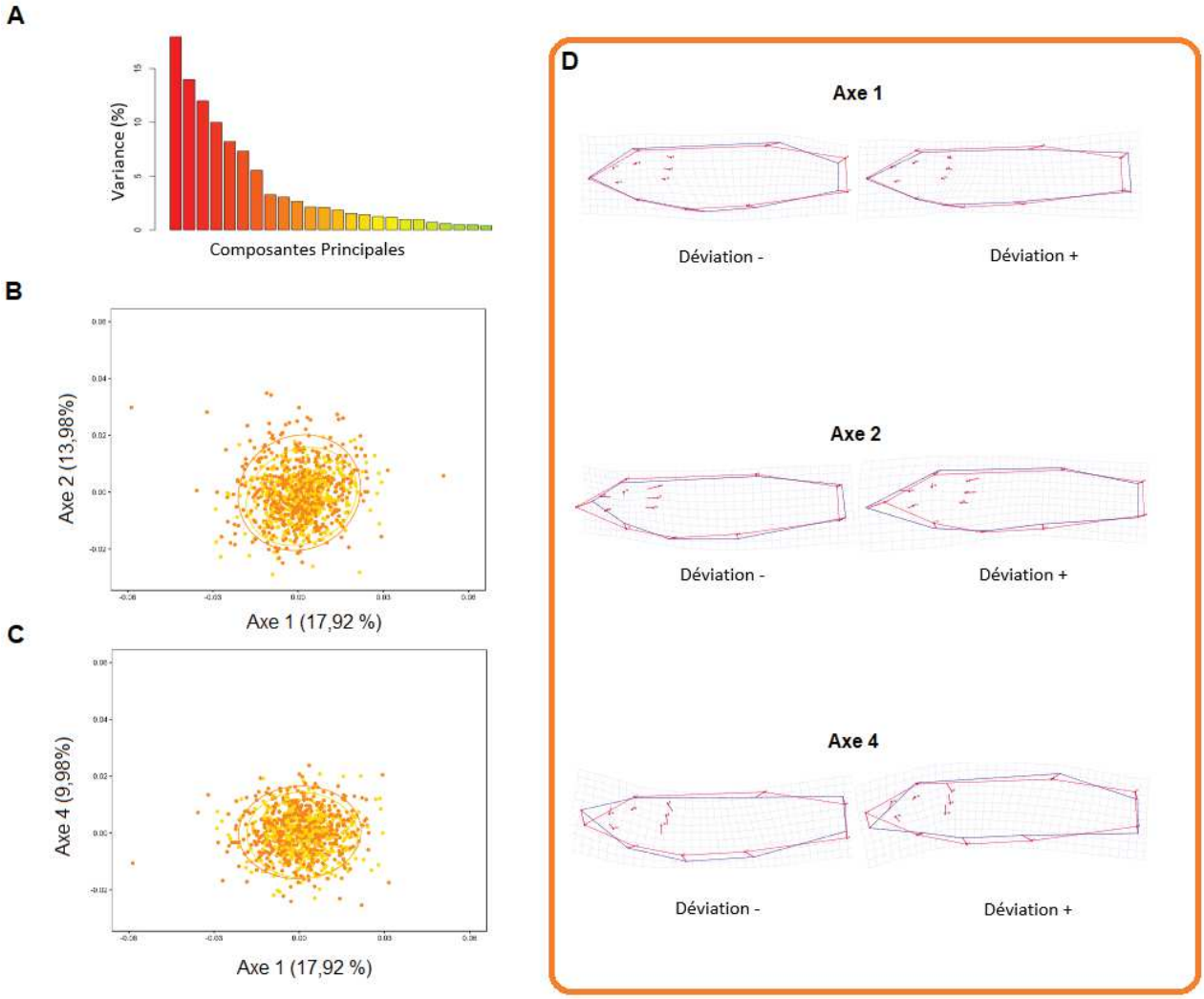


FIGURE 5 – ACP sur les points morpho-anatomiques des 995 individus assignés à deux parents. **A.** Histogramme de la variance portée par chaque axe. Positions des individus sur **B.** les axes 1 et 2 et **C.** les axes 1 et 4. Les individus *backcross*-MED sont représentés en orange et les MED en jaune. **D.** Grilles de déformations associées aux axes 1, 2 et 4. La forme consensus des individus est représentée en rouge et la forme bleue montre la déformation le long de l'axe en positif et négatif.

également permis de montrer que COLONY2 ne fait pas d'erreurs d'assignation quel que soit le seuil de probabilité utilisé (Annexe 4 Tableau 1). Cependant, plus la valeur de probabilité d'assignation utilisée est élevée, plus le nombre d'individus assignés (au deux ou à un seul parent) est faible, ce qui démontre un manque de puissance du programme. Néanmoins, ces simulations nous permettent de dire que bien que nous n'ayons pas pu identifier les parents d'un nombre non négligeable d'individus, nous pouvons avoir confiance dans les assignations faites par COLONY2.

Nous avons ensuite cherché à déterminer si l'origine des mâles influence le nombre d'individus survivants en utilisant un GLMM (Tableau 2) qui a été ajusté à la fois par approche bayésienne et maximum de vraisemblance (ajustée 5 fois indépendamment afin de s'assurer de la bonne convergence du modèle). Avec les deux approches les effets estimés sont très similaires et significativement différents de zéro ($p < 0,001$). Cependant ils ne sont pas significativement différents l'un de l'autre étant donné que le test de Wald n'est pas significatif ($p = 0,32$) et que les intervalles de confiance à 95% se chevauchent. De plus les distributions des effets fixes estimés pour les *backcross*-MED et les MED par l'approche bayésienne sont très similaires (Figure Supplémentaire 2). Il n'y a donc pas d'effets du type de croisement sur la survie à 1 an.

3. Analyses morpho-anatomiques des juvéniles

Nous avons ensuite voulu voir si la croissance et la condition des individus mesurée lors de la première biométrie pouvaient être expliquées par l'origine des pères. Pour cela, nous avons ajusté par approche bayésienne deux GLMM (Tableau 3). Bien que les effets fixes soient significativement différents de zéro pour les deux types de croisements à la fois pour la croissance et la condition ($p < 0,001$) ils ne sont pas significativement différents pour les MED et les *backcross*-MED étant donné que les intervalles de confiance se chevauchent. En effet, pour la croissance et la condition, les distributions des effets fixes sont chevauchantes (Figure Supplémentaire 3). La croissance et la condition des individus à l'âge d'un an environ ne dépendent donc pas de l'origine de leur père.

Nous avons ensuite utilisé les coordonnées des 14 points morpho-anatomique numérisées sur les 995 descendants assignés à leurs deux parents et corrigées par l'analyse Procruste pour réaliser une ACP (Figure 5). Nous nous sommes focalisés sur les 7 premiers axes étant donné que ce sont eux qui portent la plus grande part de variance, 74,9 % cumulés (Figure 5A). A l'aide de GLMM nous avons cherché à savoir quels axes portaient l'information du type de croisement (Tableau Supplémentaire 2). Aucun des 7 premiers axes ne capture de façon significative l'information du type de croisement. Nous avons ensuite utilisé des régressions linéaires pour voir quels axes portaient l'information de la variance associée à la croissance et la condition (Tableau Supplémentaire 2). L'axe 1 capture de façon significative, quoique faible, l'information de la croissance ($R^2 = 0,18$ et $p = 0,29e^{-37}$) et l'axe 4 la

Variable réponse	Vraisemblance	Effets fixes estimés		P-value		Comparaison des effets fixes	
		Backcross-MED	MED	*Backcross*-MED	MED	*Backcross*-MED	MED
Taux de mortalité de 1 à 2 ans (B)	-31	-0,87	-1,70	*	**	Test de Wald : 0,69	
Survie à 2 an (B)	-	2,39	2,39	***	***	$IC_{95\%}$[1,88 ; 2,94]	$IC_{95\%}$[1,82 ; 2,93]
G à 2 ans (B)	-	-3,24	4,27	-	-	$IC_{95\%}$[-12,66 ; 6,38]	$IC_{95\%}$[-6,15 ; 14,36]
K à 2 ans (B)	-	5,76	-2,75	-	-	$IC_{95\%}$[-3,88 ; 15,47]	$IC_{95\%}$[-13,01 ; 7,54]

TABLEAU 4 – Effets fixes estimés avec un GLMM pour la mortalité entre l'âge d'1 et 2 ans et la survie, croissance et condition à 2 ans. Les étoiles symbolisent la P-value (* : < 0,05 ; ** : < 0,01 ; *** : < 0,001). La P-value indique si les effets fixes estimés sont significativement différents de 0 et le test de Wald s'ils sont significativement différents pour les MED et les *backcross*-MED pour l'approche du maximum de vraisemblance. Pour l'approche bayésienne la comparaison des intervalles de confiance

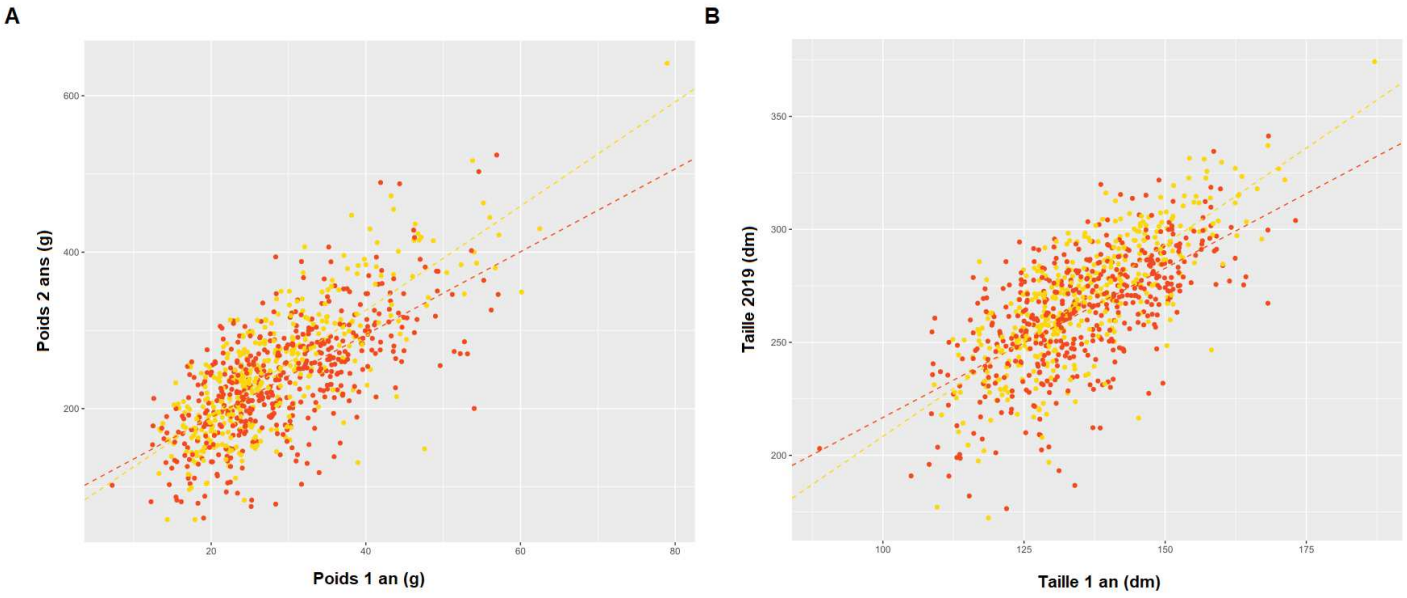


FIGURE 6 – Corrélation entre deux variables phénotypiques mesurées à 1 an d'écart. Les points jaunes correspondent aux individus MED et les oranges aux *backcross*-MED. Les droites en pointillées représentent les régressions linéaires pour les deux groupes d'individus. Corrélation entre **A.** le poids et **B.** la taille mesurée à 1 et 2 ans.

condition (R^2= 0,24 et p = 0,50e^{-94}). De plus, l'axe 2 explique une part de la variance associée à la condition (R^2= 0,10 et p = 0,28e^{-37}) et de la croissance (R^2= 0,06 et p = 0,17e^{-23}). Les grilles de déformation des trois axes de l'ACP considérés (Figure 5D) correspondent bien aux analyses de corrélation effectuées, les poissons se distinguant sur un axe longitudinal lié aux différences de croissance le long de l'axe 1 de l'ACP et sur un axe dorso-ventrale le long de l'axe 4 de l'ACP correspondant à des différences de condition. Les changements portés par l'axe 2 se font principalement au niveau de la tête. Cependant, les critères morphologiques utilisés ne permettent pas de distinguer les *backcross*-MED des MED, et la variance de croissance et de condition qui existe entre les individus n'est pas expliquée par l'origine des pères.

4. Survie, croissance et condition des poissons à deux ans

Nous avons ensuite analysé les données de la seconde biométrie afin de déterminer si des différences de survie ou de morphologie commençaient à apparaitre entre les *backcross*-MED et les MED plus tardivement dans le développement des poissons. Sur les 995 poissons étudiés la première année 819 étaient toujours vivants l'année suivante. Nous avons testé à l'aide de modèles GLMM s'il existent des différences de survie à deux ans et des différences de mortalité entre 1 et 2 ans entre les *backcross*-MED et les MED (Tableau 3). Il n'y a pas de différences de taux de mortalité entre les *backcross*-MED et les MED entre la première et la deuxième année le test de Wald n'étant pas significatif. Nous avons également montré qu'il n'y a pas de différence de survie entre les deux types d'individus étant donné que les intervalles de confiance des effets fixes estimés et leurs distributions se chevauchent (Figure Supplémentaire 4). Ici le modèle a été analysé uniquement avec l'approche bayésienne étant donné que nous avions précédemment montré que les deux approches (maximum de vraisemblance et bayésienne) donnent des résultats très similaires (Tableau 2). De plus, nous n'avons pas pu mettre en évidence l'existence d'un effet du type de croisement sur la croissance et la condition des individus mesurées à deux ans. Nous nous sommes ensuite intéressés aux corrélations entre le poids et la taille mesurés à 1 et à 2 ans (Figure 6). Nous avons alors pu montrer que bien qu'il n'existe pas de différence moyenne entre le poids et la taille des *backcross*-MED et des MED les pentes de régression sont différentes. En effet, à la fois pour le poids et la taille la régression est plus pentue pour les individus MED que *backcross*-MED (Figure 6). A l'aide d'un modèle linéaire nous avons montré qu'il y a un effet significatif du type de croisement à la fois sur la taille et le poids (p-value sur poids = 1,75e^{-5} et p-value pour la taille =1,75e^{-4}), ce qui indique que les pentes des régressions sont significativement différentes. Il semblerait donc que les individus MED aient grandi plus et pris plus de poids que les *backcross*-MED pendant leur deuxième année de vie.

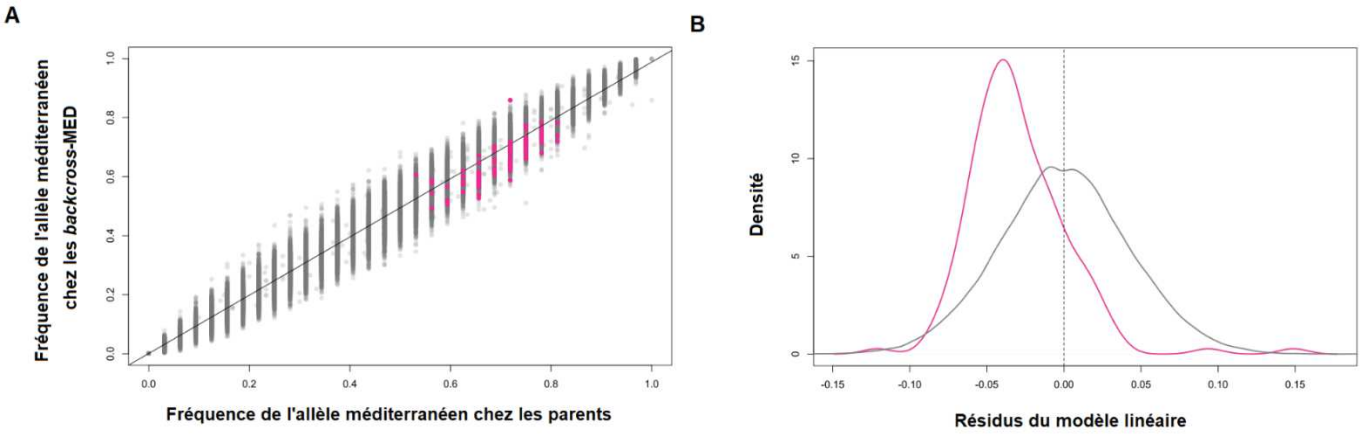


FIGURE 7 – Corrélation linéaire des fréquences alléliques des parents et des *backcross*-MED sur 49993 SNPs dont 156 impliqués dans l'isolement reproductif et différentiellement fixés entre les grands-parents atlantiques et ouest-méditerranéens des *backcross*-MED. A. Chaque point représente un SNPs non impliqué (gris) ou impliqué dans l'IR et différentiellement fixé entre les grands-parents (rose). La droite rouge représente la régression linéaire entre les fréquences alléliques. **B.** Distribution des écarts au modèle pour les SNPs non impliqués (gris) ou impliqué dans l'IR et différentiellement fixé entre les grands-parents (rose).

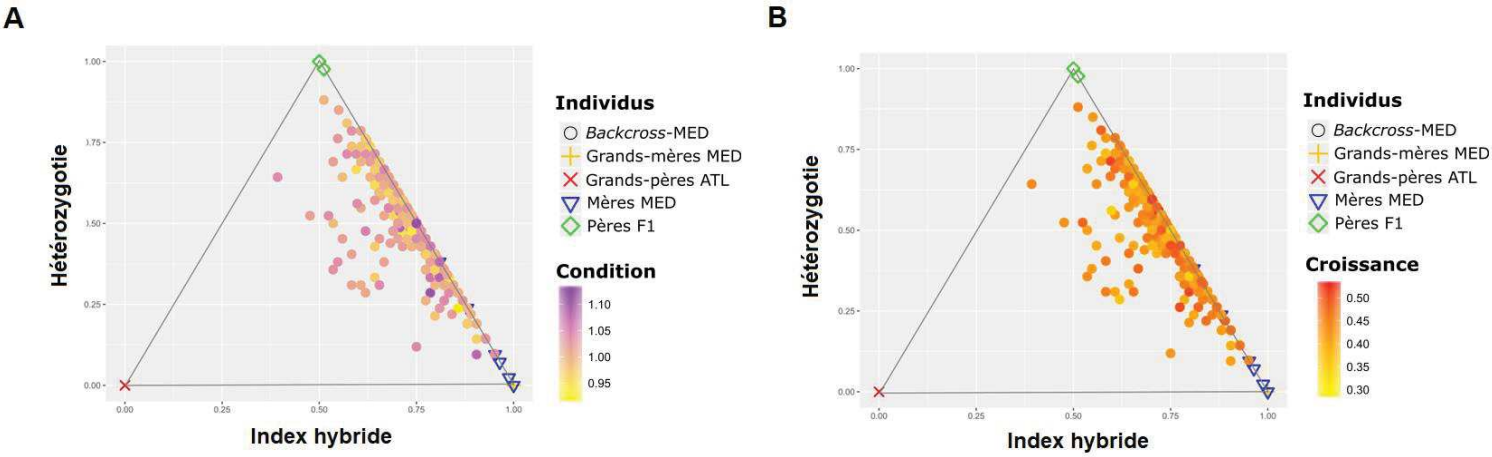


FIGURE 8 – Etude du lien génotype-phénotype chez 380 *backcross*-MED sur 42 SNPs impliqués dans l'isolement reproductif et présentant des fortes valeurs de F_{ST} entre les géniteurs atlantiques et méditerranéens. Les *backcross*-MED (ronds) ainsi que leurs ascendants, leurs grand-mères et mères d'origine méditerranéenne (respectivement croix jaunes et triangles bleus), leurs grands-pères d'origine atlantique (croix rouges) et leurs pères hybrides de première génération (losanges verts) sont représentés dans un espace triangulaire indiquant en abscisse leur indice hybride (c'est-à-dire la proportion d'allèles d'origine méditerranéenne) et en ordonnés leur niveau d'hétérozygotie. La couleur des ronds indique **A.** la condition et **B.** la croissance des *backcross*-MED à 1 ans.

5. Lien génotype-phénotype chez les individus *backcross*-MED

Nous avons voulu voir si les locus précédemment identifiés (cf Chapitre 2) comme impliqués dans l'IR au sein des populations naturelles de *D. labrax* (soit 762 SNPs sur 49 993 analysés avec la puce à ADN utilisée pour génotyper les *backcross*-MED) avaient un comportement particulier chez les *backcross*-MED. Pour cela nous avons étudié la corrélation des fréquences alléliques chez les parents et les descendants (R^2= 0,96 et $p < 2,2e^{-16}$) et regardé le comportement de ces locus en particulier afin de voir s'ils s'écartent de la corrélation (Figure Supplémentaire 5). On constate que pour les SNPs qui ne sont pas impliqués dans l'IR, la distribution des écarts au modèle est centrée sur zéro alors que pour ceux impliqués dans l'IR elle est légèrement décalée vers les valeurs négatives (Figure Supplémentaire 5B). Il semblerait donc que les individus *backcross*-MED tendent à porter moins souvent l'allèle méditerranéen qu'attendu. Parmi les SNPs impliqués dans l'IR nous nous sommes ensuite focalisés sur ceux différentiellement fixés entre les populations naturelles atlantiques et est-méditerranéennes, soit 127 SNPs (Figure Supplémentaire 6) et différentiellement fixés entre les individus atlantiques et ouest-méditerranéens ayant servi de géniteurs pour les croisements expérimentaux (correspondant aux grands-parents des individus étudiés ici) soit 156 SNPs (Figure 7). Dans les deux cas, la distribution des résidus du modèle linéaire est très nettement décalée vers les valeurs négatives, indiquant qu'à ces locus particuliers les *backcross*-MED portent moins souvent l'allèle méditerranéen qu'attendu. Ainsi, il semblerait que les allèles atlantiques soient favorisés chez les *backcross*-MED à ces locus particuliers.

Nous avons ensuite voulu savoir s'il existait un lien entre le génotype des individus sur ces locus et leur valeur sélective, évaluée par leur croissance et leur condition à 1 ans (Figure Supplémentaire 7). Les locus utilisés ici ont été choisis car différentiellement fixés entre les grands-parents atlantiques et méditerranéen. Les individus ouest-méditerranéens étant fortement introgressés, les mères ne sont pas positionnées dans l'angle inférieur droit du triangle comme elles devraient l'être, ce qui décale les individus *backcross*-MED de leur position attendue dans le triangle (Figure Supplémentaire 7). Nous n'avons donc conservé que les SNPs ayant un F_{ST} mesuré entre les mères et les grands-pères atlantiques supérieur à 0.9, soit 42 SNPs (Figure 8). Une des mères étant fortement introgressée, il existe une grande variance au niveau des génotypes des *backcross*-MED. Cependant, nous n'avons pas pu mettre en évidence l'existence d'un lien entre le génotype des individus et leur condition et croissance à 1 an (Figure 8) et à 2 ans (Figure Supplémentaire 8). De même nous avons voulu voir si les individus *backcross*-MED morts entre la première et deuxième biométrie présentaient des génotypes particuliers mais il semble que ce ne soit pas le cas étant donné que ces individus se répartissent de façon homogène dans l'espace des génotypes des *backcross*-MED (Figure Supplémentaire 9). Cependant, les individus qui sont morts durant la deuxième année avaient une croissance significativement plus faible que les survivants à l'âge d'un an (Figure Supplémentaire 10).

Discussion

L'objectif principal de cette étude était de déterminer s'il existe une différence de valeur sélective entre les individus *backcross*-MED et MED qui pourrait expliquer l'existence d'une barrière au flux génique observée dans la nature entre la lignée atlantique et méditerranéenne de *D. labrax*. Les individus ayant été élevés dans des conditions standardisées d'aquaculture, notre étude ne permet pas de tester l'existence de barrières extrinsèques au flux génique qui pourraient être générées par une forme d'isolement comportemental ou une adaptation différentielle des bars et des loups aux environnements atlantique et méditerranéen. En effet, la période de ponte qui est décalée d'un mois entre les deux lignée ainsi que la mise en évidence d'un comportement philopatrique du bar en atlantique (de Pontual *et al.* 2019) génèrent probablement un certain niveau d'isolement pré-zygotique. Cependant, même si ces barrières existent, elles n'empêchent pas l'introgression d'allèles atlantiques en Méditerranée et n'expliquent pas l'hétérogénéité du flux génique le long du génome et son différentiel entre la population ouest- et est-méditerranéenne (Tine *et al.* 2014; Duranton *et al.* 2018). Il existe donc nécessairement une forme d'isolement post-zygotique entre les deux lignées. Notre étude pourrait donc permettre de révéler l'existence de barrières intrinsèques intervenant à différents stades du cycle de vie, de la fécondation (barrières pré-zygotiques tardive), à l'éclosion jusqu'à l'âge d'un et deux ans. Néanmoins, la valeur sélective d'un individu étant en grande partie déterminée par sa capacité à se reproduire, une part importante de la valeur sélective n'est pas évaluée ici. En effet, la maturité sexuelle étant atteinte à l'âge de trois ans chez le bar européen, nous n'avons pas pu faire de mesures sur les traits en lien avec la fertilité.

La mise en évidence d'une différence de valeur sélective entre deux groupes d'individus nécessitant qu'ils aient grandi dans les mêmes conditions, il nous a fallu distinguer à l'aide de marqueurs génétiques les *backcross*-MED des MED. Pour cela nous avons utilisé le logiciel COLONY2 afin de réaliser des assignations parentales. Nous avons pu montrer à l'aide de simulations que pour cette étude, le taux d'erreur est nul quand le seuil de probabilité d'assignation est fixé à 0,95 (Tableau Supplémentaire 1). Ainsi, bien qu'un certain pourcentage d'individus n'ait pas pu être assignés, nous sommes confiants quant à notre capacité à distinguer les *backcross*-Med des MED. Malgré tout, nous n'avons pas mis en évidence l'existence d'une différence de valeur sélective entre les deux groupes d'individus, que ce soit au niveau du taux de fécondation, de mortalité embryonnaire, de la survie, la croissance et la condition mesurées à 1 et 2 ans. Il semblerait cependant qu'entre l'âge d'un et deux ans le poids et la taille des individus MED aient augmenté plus rapidement que celui des *backcross*-MED (Figure 6).

Les conditions standardisées d'aquaculture dans lesquelles ont grandi les individus génèrent des pressions de sélection moins fortes que l'environnement naturel. Ainsi, certains génotypes avec une valeur sélective plus faible ont peut-être pu survivre, alors qu'ils auraient été éliminés dans la nature, ce qui aurait pu masquer des différences potentielles de valeur sélective entre les groupes. Cependant, une étude précédente a déjà mis en évidence des différences de valeurs sélectives entre différentes populations de bar européen dans des conditions similaires (Guinand *et al.* 2017). En effet il a été montré que les F1 atlantique/ouest-méditerranéen avait une meilleur survie que les individus issus des populations parentales, les atlantiques survivant mieux que les est-méditerranéens, eux-mêmes survivant mieux que les ouest-méditerranéens (Guinand *et al.* 2017). On peut donc penser que si des différences de valeur sélectives existaient, les conditions expérimentales auraient permis de les révéler.

Cependant, le fait que les deux groupes d'individus aient la même valeur sélective apparente ne veut pas nécessairement dire que les *backcross*-MED ne souffrent pas de dépression d'hybridation. En effet, l'existence d'hétérosis chez les hybrides de première génération a été démontrée par une étude précédente (Guinand *et al.* 2017). Or, un mécanisme permettant de générer de l'hétérosis est le masquage (au sein des génomes hybrides) des mutations faiblement délétères récessives qui ségrégent dans les fonds génétiques des deux populations parentales (Charlesworth and Willis 2009). L'hétérozygotie étant diminuée de moitié à chaque génération de rétrocroisement, l'hétérosis tend à diminuer progressivement au fil des générations. On peut donc s'attendre à ce qu'il y ait de l'hétérosis résiduelle chez les *backcross*-MED, ce qui pourrait masquer l'effet de la dépression d'hybridation. En effet, l'étude de Guinand *et al.* (2017) a également montré que le croisement intraspécifique avec la valeur sélective la plus faible est celui faisant intervenir deux individus ouest-méditerranéens. Sachant que la population ouest-méditerranéenne est celle présentant les plus fort niveaux d'introgression et qu'il a été mis en évidence l'existence d'une contre-sélection des allèles atlantiques en Méditerranée (Duranton *et al.* 2018), on peut se demander si la population ouest-méditerranéenne ne subit pas encore une faible dépression d'hybridation qui expliquerait la mauvaise survie des individus ouest-méditerranéens précédemment observée (Guinand *et al.* 2017). Chez les *backcross*-MED l'hétérosis résiduelle pourrait donc compenser les effets de la dépression d'hybridation.

Nous nous sommes donc intéressés chez les *backcross*-MED au comportement des allèles atlantiques aux locus identifiés comme impliqués dans l'IR entre les populations naturelles. Etant donné que les allèles atlantiques sont contre-sélectionnés en Méditerranée sur le long terme, on s'attend à ce que les allèles méditerranéens soient favorisés à ces locus chez les *backcross*-MED. Or, on observe l'inverse, les allèles méditerranéens étant moins présents qu'attendu (Figure 7). Il semblerait donc que la présence d'allèles atlantiques à ces locus soit avantageuse, bien que nous n'ayons pas pu mettre en

évidence de lien avec la croissance et la condition des individus (Figure 8 et Annexe 4 Figure 9). Cependant, l'effet sur la valeur sélective des individus peut être lié à d'autres traits phénotypiques que nous n'avons pas pu mesurer dans cette étude, notamment pour tout ce qui touche à la reproduction, une autre composante très importante de la valeur sélective. Il semblerait donc que la sélection sur les allèles atlantiques introgressés soit différente sur le court et long terme, les allèles atlantiques étant favorisés chez les *backcross*-MED aux locus où ils sont éliminés sur le long terme dans les populations naturelles.

L'hétérosis pourrait expliquer pourquoi le sens de la sélection sur les allèles atlantiques introgressés change au cours du temps. En effet, l'hétérosis étant générée par le masquage des mutations faiblement délétères récessives, plus les haplotypes introgressés sont longs, plus le masquage est important et donc plus l'effet d'hétérosis est fort, c'est la superdominance associative (Ohta and Kimura 1970). Ainsi, l'hétérosis est maximale à la première génération d'hybridation puis diminue progressivement à mesure que les fragments introgressés sont raccourcis par la recombinaison à chaque génération de rétrocroisement. L'introgression d'allèles atlantiques lors des premières générations d'hybridation pourrait donc être favorisée pour leurs effet d'hétérosis (Kim *et al.* 2018). En effet, la population méditerranéenne ayant une taille efficace plus faible que l'atlantique il est probable qu'elle ait un fardeau génétique plus élevé ce qui pourrait favoriser l'introgression dans ce sens. De plus, l'effet d'hétérosis étant lié aux mutations faiblement délétères qui sont fixées dans le fond génétique méditerranéen, on l'attend particulièrement dans les régions où la recombinaison est faible. En effet, la taille efficace étant réduite dans ces régions, la sélection est moins efficace et un plus grand nombre de mutations faiblement délétères y ségrégent. Or c'est principalement dans les régions à faible taux de recombinaison que se trouvent les îlots résistants à l'introgression. On peut donc imaginer qu'un certain nombre de mutations faiblement délétères récessives présentent dans les génomes méditerranéens soient liées à des locus impliqués dans l'IR. D'autant plus que dans les régions qui ne résistent pas à l'introgression elles auraient pu être remplacées par des allèles atlantiques plus avantageux. C'est pourquoi on s'attend à ce que les mêmes régions génomiques soient impliquées dans l'IR et l'hétérosis.

Il semblerait donc que dans les premières générations d'hybridation, les haplotypes introgressés étant longs, c'est principalement leur effet de masquage des mutations faiblement délétères récessives qui est visible, leur introgression pourrait alors être favorisé par l'hétérosis (Kim *et al.* 2018). Plus, tard la recombinaison ayant raccourci les haplotypes, l'effet d'hétérosis diminue et les effets délétères commencent à se révéler et les allèles introgressés sont alors contre-sélectionnés. Ce processus a déjà été envisagé chez l'homme, chez qui l'introgression d'allèles néanderthaliens aurait été facilité par la superdominance associative, qui aurait diminuée autour de la vingtième génération de

rétrocroisement révélant les effets délétères de l'introgression (Harris and Nielsen 2016; Juric *et al.* 2016). Lors des premières générations d'introgression c'est donc principalement l'effet bloc des haplotypes introgressés qui est visible alors que dans les générations d'introgression plus tardives, c'est l'effet individuel des locus qui est révélé. C'est pourquoi la sélection peut changer au cours du temps rendant la dynamique d'introgression complexe.

Conclusion

En conclusion, la sélection sur les allèles atlantiques introgressés dans les génomes méditerranéens semble s'exercer de manière différente sur le court et le long terme. Dans les premières générations d'hybridation, les allèles atlantiques étant assis sur de long haplotypes, ils sont favorisés pour l'hétérosis qu'ils génèrent. Plus tard, la recombinaison diminuant l'hétérosis, les effets délétères sont révélés et ils sont alors contre-sélectionnés. L'effet de l'introgression semble donc étroitement lié à la taille des fragments introgressés et donc la dynamique d'introgression directement en lien avec la recombinaison. Il est donc nécessaire d'avoir une vision allant au-delà des premières générations d'hybridation pour comprendre comment agissent les barrières d'IR, étant donné que contrairement à ce qui était précédemment envisagé, elles peuvent ne pas se révéler chez les premières générations d'hybridation mais doivent apparaître plus tardivement.

Références

Adams D. C., and E. Otárola-Castillo, 2013 geomorph: an r package for the collection and analysis of geometric morphometric shape data. Methods in Ecology and Evolution 4: 393–399. https://doi.org/10.1111/2041-210X.12035

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

Charlesworth D., and J. H. Willis, 2009 The genetics of inbreeding depression. Nature Reviews Genetics 10: 783–796. https://doi.org/10.1038/nrg2664

Coyne J. A., and A. H. Orr, 2004 *Speciation*. Sunderland MA, massachusetts U.S.A.

Dobzhansky T. grigorovitch, 1937 *Genetics and the origin of species*.

Duranton M., F. Allal, C. Fraïsse, N. Bierne, F. Bonhomme, *et al.*, 2018 The origin and remolding of genomic islands of differentiation in the European sea bass. Nature Communications 9: 2518. https://doi.org/10.1038/s41467-018-04963-6

Guinand B., M. Vandeputte, M. Dupont-Nivet, A. Vergnet, P. Haffray, *et al.*, 2017 Metapopulation patterns of additive and nonadditive genetic variance in the sea bass (Dicentrarchus labrax). Ecol Evol 7: 2777–2790. https://doi.org/10.1002/ece3.2832

Harris K., and R. Nielsen, 2016 The Genetic Cost of Neanderthal Introgression. Genetics 203: 881–891. https://doi.org/10.1534/genetics.116.186890

Harrison R. G., and E. L. Larson, 2016 Heterogeneous genome divergence, differential introgression, and the origin and structure of hybrid zones. Mol Ecol 25: 2454–2466. https://doi.org/10.1111/mec.13582

Jones O. R., and J. Wang, 2010 COLONY: a program for parentage and sibship inference from multilocus genotype data. Molecular Ecology Resources 10: 551–555. https://doi.org/10.1111/j.1755-0998.2009.02787.x

Juric I., S. Aeschbacher, and G. Coop, 2016 The Strength of Selection against Neanderthal Introgression. PLOS Genetics 12: e1006340. https://doi.org/10.1371/journal.pgen.1006340

Kim B. Y., C. D. Huber, and K. E. Lohmueller, 2018 Deleterious variation shapes the genomic landscape of introgression. PLOS Genetics 14: e1007741. https://doi.org/10.1371/journal.pgen.1007741

Klingenberg C. P., 2011 MorphoJ: an integrated software package for geometric morphometrics. Molecular Ecology Resources 11: 353–357. https://doi.org/10.1111/j.1755-0998.2010.02924.x

Lemaire C., J.-J. Versini, and F. Bonhomme, 2005 Maintenance of genetic differentiation across a transition zone in the sea: discordance between nuclear and cytoplasmic markers. Journal of Evolutionary Biology 18: 70–80. https://doi.org/10.1111/j.1420-9101.2004.00828.x

Muller H., 1942 Isolating mechanisms, evolution, and temperature. Biol. Symp. 6: 71–125.

Nosil P., T. H. Vines, and D. J. Funk, 2005 Reproductive isolation caused by natural selection against immigrants from divergent habitats. Evolution 59: 705–719.

Ohta T., and M. Kimura, 1970 Development of associative overdominance through linkage disequilibrium in finite populations*. Genetics Research 16: 165–177. https://doi.org/10.1017/S0016672300002391

Pontual H. de, M. Lalire, R. Fablet, C. Laspougeas, F. Garren, *et al.*, 2019 New insights into behavioural ecology of European seabass off the West Coast of France: implications at local and population scales. ICES J Mar Sci. https://doi.org/10.1093/icesjms/fsy086

Ravinet M., R. Faria, R. K. Butlin, J. Galindo, N. Bierne, *et al.*, 2017 Interpreting the genomic landscape of speciation: a road map for finding barriers to gene flow. J. Evol. Biol. 30: 1450–1477. https://doi.org/10.1111/jeb.13047

ROHLF F. J., 2006 tpsDig, version 2.10. http://life.bio.sunysb.edu/morph/index.html.

Tine M., H. Kuhl, P.-A. Gagnaire, B. Louro, E. Desmarais, *et al.*, 2014 European sea bass genome and its variation provide insights into adaptation to euryhalinity and speciation. Nature Communications 5: 5770.

Vandeputte M., A. Puledda, A. S. Tyran, A. Bestin, C. Coulombet, *et al.*, 2017 Investigation of morphological predictors of fillet and carcass yield in European sea bass (Dicentrarchus labrax) for application in selective breeding. Aquaculture 470: 40–49. https://doi.org/10.1016/j.aquaculture.2016.12.014

Wald A., and J. Wolfowitz, 1948 Optimum Character of the Sequential Probability Ratio Test. Ann. Math. Statist. 19: 326–339. https://doi.org/10.1214/aoms/1177730197

CHAPITRE 5 :

Estimation de la distance de dispersion du bar européen en Méditerranée

Les résultats de ce chapitre sont exposés en détails dans l'article joint publié dans *Evolutionary Applications.*

Bien que les questions posées dans cette thèse soient majoritairement théoriques, les données obtenues peuvent également permettre de répondre à des questions beaucoup plus appliquées. Par exemple, pour avoir une gestion adaptée des populations naturelles il est essentiel de connaître le niveau de connectivité qui existe entre elles. En effet, la migration peut assurer la stabilité de certaines populations ayant un taux de natalité trop faible par rapport au taux de mortalité (Runge *et al.* 2006; Furrer and Pasinelli 2016). De plus, le bar européen est un poisson d'intérêt économique important, notamment pour la France qui est le principal pays à l'exploiter aussi bien pour la pêche professionnelle (les débarquements de bar français représentant les deux tiers des débarquements internationaux) que récréative. Il est donc essentiel d'avoir une bonne connaissance de la dynamique des populations naturelles afin d'avoir une gestion adéquate des stocks de pêche. Par ailleurs, la connectivité réelle réalisée dans la nature, en particulier dans le milieu marin, est un paramètre-clé pour comprendre les phénomènes de sélection contre-sélection et adaptation locale.

Etudier la connectivité des populations marines grâce à des approches directes de type capture-marquage-recapture peut cependant se révéler compliqué, étant donné que la dispersion a majoritairement lieu pendant la phase larvaire (Selkoe *et al.* 2016). C'est pourquoi des approches basées sur l'utilisation de marqueurs génétiques neutres ou sous sélection ont été proposées (Gagnaire *et al.* 2015; Selkoe *et al.* 2016). Ici, nous avons utilisé la distribution de tailles des fragments atlantiques introgressés en Méditerranée-Ouest et Est afin d'estimer l'échelle spatiale de la dispersion du bar européen au sein de la Méditerranée. En effet, les haplotypes introgressés d'origine atlantique entrant en premier dans la population ouest, il leur faut un certain temps pour atteindre la population est, temps pendant lequel ils se font progressivement raccourcir par la recombinaison qui agit à chaque génération (Pool and Nielsen 2009; Liang and Nielsen 2014; Racimo *et al.* 2015). La différence de taille des haplotypes introgressés entre l'Ouest et l'Est représente donc l'action de la recombinaison pendant le temps nécessaire pour traverser la Méditerranée.

Nous avons donc utilisé la formule permettant de relier la taille des haplotypes introgressés à leur date d'introgression ($\bar{L} = [(1-m)r(t-1)]^{-1}$ (Racimo *et al.* 2015)) indépendamment dans les deux populations méditerranéennes et comparé les deux estimations pour estimer le temps nécessaire pour qu'un haplotype diffuse d'une population à l'autre. Etant donné que cette formule suppose que l'introgression est neutre et que nous savons que les allèles atlantiques sont contre-sélectionnés en Méditerranée, nous avons éliminé des analyses les régions potentiellement impliquées dans l'isolement reproductif. Dans un premier temps nous avons appliqué la formule sur les haplotypes

présents dans des fenêtres de 100 kb non chevauchantes afin d'obtenir une distribution des temps d'introgression qui prenne en compte les variations locales du taux de recombinaison. En effet, le génome du bar européen peut être divisé en deux catégories en fonction de la recombinaison, les régions centro-chrosomiques où la recombinaison est faible et les régions péri-chromosomiques où la recombinaison est en moyenne dix fois plus forte. Nous avons ainsi pu estimer un taux de dispersion d'environ 15 km par génération (Figure 2). Dans un deuxième temps, nous avons divisé le génome en 11 catégories de taux de recombinaison pour lesquelles nous avons estimé $\overline{L}$ puis *t* à partir de la pente de la distribution log-transformée de la longueur des fragments introgressés. Ceci permet d'avoir une estimation plus précise de la taille des fragments introgressés mais néglige les variations fines du taux de recombinaison. Nous avons alors obtenu un taux de dispersion d'environ 7 km par génération (Figure 3). Enfin, sachant que la formule que nous utilisons suppose un évènement d'admixture instantané alors que le flux génique entre les lignées atlantique et méditerranéenne est continu depuis environ 11 000 ans, nous avons voulu valider notre approche à l'aide de simulations. Nous avons ainsi pu montrer que non seulement cette méthode donne des résultats cohérents en présence de flux génique continu mais qu'elle est également pertinente pour une large gamme de durée de contact, d'intensité de flux génétique et de nombre d'individus échantillonnés (Figure 4).

Bien que nos deux estimations soient légèrement différentes entre les deux approches, nos résultats sont cohérents avec les études de génétique des populations ayant précédemment mis en évidence l'existence d'une structure entre la Méditerranée ouest et est (Allegrucci *et al.* 1997; Bahri-Sfar *et al.* 2000; Souche *et al.* 2015). De plus, cette valeur relativement faible de dispersion compte tenu de la phase larvaire longue (8 à 12 semaines) et de la forte mobilité des juvéniles et adultes va également dans le sens d'un comportement philopatrique soupçonné chez le bar européen (Castilho and Ciftci 2005; de Pontual *et al.* 2019). Cependant, comme nous n'avons utilisé que deux points d'échantillonnage, cette distance représente une moyenne et peut masquer la présence de barrières de dispersion qui réduisent localement le flux génique. La circulation particulière des eaux dans le détroit siculo-tunisien est notamment connue pour réduire le flux génique entre les populations occidentales et orientales de Méditerranée et ce chez plusieurs espèces de poissons (Quéré *et al.* 2012; Pascual *et al.* 2017). Néanmoins, nos estimations fournissent déjà des informations pertinentes pour la conservation et la gestion des populations, et ouvrent la voie vers de nouveaux développements méthodologiques plus sophistiqués.

Références

Allegrucci G., C. Fortunato, and V. Sbordoni, 1997 Genetic structure and allozyme variation of sea bass (Dicentrarchus labrax and D. punctatus) in the Mediterranean Sea. Marine Biology 128: 347–358. https://doi.org/10.1007/s002270050100

Bahri-Sfar L., C. Lemaire, O. K. B. Hassine, and F. Bonhomme, 2000 Fragmentation of sea bass populations in the western and eastern Mediterranean as revealed by microsatellite polymorphism. Proceedings of the Royal Society of London B: Biological Sciences 267: 929–935. https://doi.org/10.1098/rspb.2000.1092

Castilho R., and Y. Ciftci, 2005 Genetic differentiation between close eastern Mediterranean Dicentrarchus labrax (L.) populations. Journal of Fish Biology 67: 1746–1752. https://doi.org/10.1111/j.1095-8649.2005.00869.x

Furrer R. D., and G. Pasinelli, 2016 Empirical evidence for source–sink populations: a review on occurrence, assessments and implications. Biol Rev 91: 782–795. https://doi.org/10.1111/brv.12195

Gagnaire P.-A., T. Broquet, D. Aurelle, F. Viard, A. Souissi, *et al.*, 2015 Using neutral, selected, and hitchhiker loci to assess connectivity of marine populations in the genomic era. Evol Appl 8: 769–786. https://doi.org/10.1111/eva.12288

Liang M., and R. Nielsen, 2014 The Lengths of Admixture Tracts. Genetics 197: 953–967. https://doi.org/10.1534/genetics.114.162362

Pascual M., B. Rives, C. Schunter, and E. Macpherson, 2017 Impact of life history traits on gene flow: A multispecies systematic review across oceanographic barriers in the Mediterranean Sea. PLOS ONE 12: e0176419. https://doi.org/10.1371/journal.pone.0176419

Pontual H. de, M. Lalire, R. Fablet, C. Laspougeas, F. Garren, *et al.*, 2019 New insights into behavioural ecology of European seabass off the West Coast of France: implications at local and population scales. ICES J Mar Sci. https://doi.org/10.1093/icesjms/fsy086

Pool J. E., and R. Nielsen, 2009 Inference of Historical Changes in Migration Rate From the Lengths of Migrant Tracts. Genetics 181: 711–719. https://doi.org/10.1534/genetics.108.098095

Quéré N., E. Desmarais, C. S. Tsigenopoulos, K. Belkhir, F. Bonhomme, *et al.*, 2012 Gene flow at major transitional areas in sea bass (Dicentrarchus labrax) and the possible emergence of a hybrid swarm. Ecol Evol 2: 3061–3078. https://doi.org/10.1002/ece3.406

Racimo F., S. Sankararaman, R. Nielsen, and E. Huerta-Sánchez, 2015 Evidence for archaic adaptive introgression in humans. Nat Rev Genet 16: 359–371. https://doi.org/10.1038/nrg3936

Runge J. P., M. C. Runge, and J. D. Nichols, 2006 The Role of Local Populations within a Landscape Context: Defining and Classifying Sources and Sinks. The American Naturalist 167: 925–938. https://doi.org/10.1086/503531

Selkoe K. A., C. C. D' Aloia, E. D. Crandall, M. Iacchei, L. Liggins, *et al.*, 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

Souche E. L., B. Hellemans, M. Babbucci, E. MacAoidh, B. Guinand, *et al.*, 2015 Range-wide population structure of European sea bass Dicentrarchus labrax. Biol J Linn Soc 116: 86–105. https://doi.org/10.1111/bij.12572

Received: 5 October 2018 | Accepted: 28 May 2019

DOI: 10.1111/eva.12829

ORIGINAL ARTICLE

Evolutionary Applications Open Access WILEY

The spatial scale of dispersal revealed by admixture tracts

Maud Duranton | François Bonhomme | Pierre-Alexandre Gagnaire

ISEM, Univ Montpellier, CNRS, EPHE, IRD, Montpellier, France

Correspondence
Maud Duranton, ISEM, Univ Montpellier, CNRS, EPHE, IRD, Montpellier, France.
Email: durantonmaud@gmail.com

Funding information
Agence Nationale de la Recherche, Grant/Award Number: ANR-17-CE02-0006-01 and LABRAD-SEQ 11-PDOC-009-01

Abstract

Evaluating species dispersal across the landscape is essential to design appropriate management and conservation actions. However, technical difficulties often preclude direct measures of individual movement, while indirect genetic approaches rely on assumptions that sometimes limit their application. Here, we show that the temporal decay of admixture tracts lengths can be used to assess genetic connectivity within a population introgressed by foreign haplotypes. We present a proof-of-concept approach based on local ancestry inference in a high gene flow marine fish species, the European sea bass (*Dicentrarchus labrax*). Genetic admixture in the contact zone between Atlantic and Mediterranean sea bass lineages allows the introgression of Atlantic haplotype tracts within the Mediterranean Sea. Once introgressed, blocks of foreign ancestry are progressively eroded by recombination as they diffuse from the western to the eastern Mediterranean basin, providing a means to estimate dispersal. By comparing the length distributions of Atlantic tracts between two Mediterranean populations located at different distances from the contact zone, we estimated the average per-generation dispersal distance within the Mediterranean lineage to less than 50 km. Using simulations, we showed that this approach is robust to a range of demographic histories and sample sizes. Our results thus support that the length of admixture tracts can be used together with a recombination clock to estimate genetic connectivity in species for which the neutral migration-drift balance is not informative or simply does not exist.

KEYWORDS
admixture tracts, connectivity, dispersal, introgression, spatial genetics

1 | INTRODUCTION

Demographic connectivity among populations plays an important role in the dynamics and resilience of species. First, it ensures the stability of local populations whose growth rate or persistence depends on immigration due to low local birth rates or high mortality rates (Furrer & Pasinelli, 2016; Pulliam, 1988; Runge, Runge, & Nichols, 2006). Demographic connectivity also contributes to the overall stability of metapopulations, by increasing the colonization potential of empty patches (Hanski, 1998). Therefore, improving empirical knowledge of species dispersal capabilities is of prime importance for understanding the ecoevolutionary dynamics of natural populations and provides helpful information for designing management and conservation actions (Lowe & Allendorf, 2010).

Quasi-direct measures of individual movement between populations can be obtained using methods such as capture–mark–recapture field experiments, parentage analyses, or assignment tests. These approaches enable evaluating how net immigration

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.
© 2019 The Authors. *Evolutionary Applications* published by John Wiley & Sons Ltd

contributes to population growth relative to local recruitment (Lowe, 2003), that is, demographic connectivity. However, they are often extremely difficult to implement (Broquet & Petit, 2009), especially for marine species in which dispersal usually takes place during a larval stage (Selkoe et al., 2016).

Indirect genetic approaches provide less demanding alternatives to evaluate average dispersal rates and distances (Broquet & Petit, 2009), although they remain uninformative regarding the contribution of dispersal to population demography, and hence stability (Lowe & Allendorf, 2010). Estimation of dispersal scales from isolation-by-distance (IBD) patterns (Rousset, 1997) has been used with success in several marine species (Palumbi, 2003; Pinsky, Montes Jr., & Palumbi, 2010; Pinsky et al., 2017; Puebla, Bermingham, & McMillan, 2012), sometimes providing consistent estimates of single-generation dispersal distances compared to direct parentage assignment methods (Pinsky et al., 2017). Nevertheless, these methods are associated with a number of assumptions which potentially limit their range of application, such as equilibrium conditions between migration and drift. Equilibrium can take a long time to establish in species with large effective population sizes, which is commonly the case in marine species. Moreover, an independent assessment of the effective density of reproducing individuals (i.e., capturing drift effects) is required for estimating the standard deviation of dispersal distances from IBD patterns (Rousset, 1997). Therefore, such approaches are not always applicable even though IBD patterns are often observed in marine species (Selkoe et al., 2016).

The recent availability of genomewide polymorphism data in nonmodel organisms has opened new research avenues for assessing connectivity, especially with the information contained in selected and hitchhiker loci (Gagnaire et al., 2015). On the other hand, a renewed interest in neutral inferences has been shown thanks to the availability of haplotype data, which have a high potential to shed light on dispersal (Cayuela et al., 2018; Gagnaire et al., 2015; Pool & Nielsen, 2009). For instance, long identical-by-descent (IBD) blocks shared between individuals have been used to infer recent demography (Palamara & Pe'er, 2013; Ringbauer, Coop, & Barton, 2017). Similarly, the distribution of migrant tracts has also proved useful for inferring the timing of recent admixture events (Gravel, 2012; Pool & Nielsen, 2009). This second type of approach relies on the fact that gene flow between divergent gene pools (e.g., populations, lineages, subspecies, ecotypes) allows migrant chromosomes to enter a new genetic background with which they recombine. As migrant chromosomes diffuse through the landscape within the introgressed population, they are progressively shortened by recombination at each generation (Liang & Nielsen, 2014; Pool & Nielsen, 2009). Therefore, the length of migrant tracts (also called admixture tracts) is informative of the time elapsed since introgression, while being relatively robust to the effect of effective population size (Racimo, Sankararaman, Nielsen, & Huerta-Sánchez, 2015). Analyzing the migrant tract length distribution in a spatial context should therefore enable to estimate the speed at which migrant tracts diffuse within an introgressed lineage and to ultimately estimate single-generation dispersal distances on conservation-relevant timescales. This method only requires gene flow between genetically differentiated populations and mapped SNPs to detect and measure the length of introgressed tracts using local ancestry inference (LAI). LAI methods have been developed that work even without the need to use phased markers (Baran et al., 2012; Guan, 2014). Nevertheless, a large variety of direct (Snyder, Adey, Kitzman, & Shendure, 2015) and indirect (Browning & Browning, 2011; Rhee et al., 2016) phasing methods can be used to facilitate the delineation of migrant tracts. Since admixture between diverging lineages is relatively common in nature (Payseur & Rieseberg, 2016), introgressed tracts have the potential to bring new information on dispersal in many species that remain difficult to study with direct tagging or classical indirect genetic approaches, which is the case for many marine species.

To illustrate this approach, we apply this framework to estimate the spatial scale of dispersal in a highly exploited marine fish, the European sea bass (*Dicentrarchus labrax*). This species is subdivided into an Atlantic and a Mediterranean glacial lineage, which started to diverge in allopatry about 300,000 years BP and remain currently partially reproductively isolated (Tine et al., 2014). Since the end of the last glacial period, asymmetrical gene flow allows the entry of Atlantic migrant tracts within the western Mediterranean population. In a recent study, we showed that these migrant tracts are on average shorter in the eastern compared to the western Mediterranean population, consistent with the action of recombination during the diffusion of Atlantic haplotypes across the Mediterranean Sea (Duranton et al., 2018). Here, we estimate the spatial scale of dispersal within the Mediterranean sea bass lineage by comparing the length distribution of introgressed Atlantic tracts between two different populations located at different distances from the contact zone with the Atlantic lineage. Furthermore, we use simulations to evaluate the robustness and the generality of this strategy to different admixture histories and sample sizes. With the development of new LAI methods to estimate the length distribution of admixture tracts (Corbett-Detig & Nielsen, 2017; Medina, Thornlow, Nielsen, & Corbett-Detig, 2018), we expect that quantitative assessments of dispersal will be obtained in several other species, which may help to resolve a long-standing issue in conservation biology.

2 | MATERIALS AND METHODS

2.1 | Whole-genome resequencing, phasing, and local ancestry inference

Our analysis relies on the use of haplotype-resolved whole-genome sequences already published in Duranton et al. (2018). Briefly, we sequenced different mother–father–child trios obtained in experimental crossings to perform chromosome-wide phasing-by-transmission (Browning & Browning, 2011). Females from the western Mediterranean Sea (Gulf of Lion, n = 8, $♀_{W\text{-}MED}$) were crossed with males from either the Atlantic Ocean (English Channel, n = 4, $♂_{ATL}$) or the eastern Mediterranean Sea (Turkey n = 2 and Egypt n = 2,

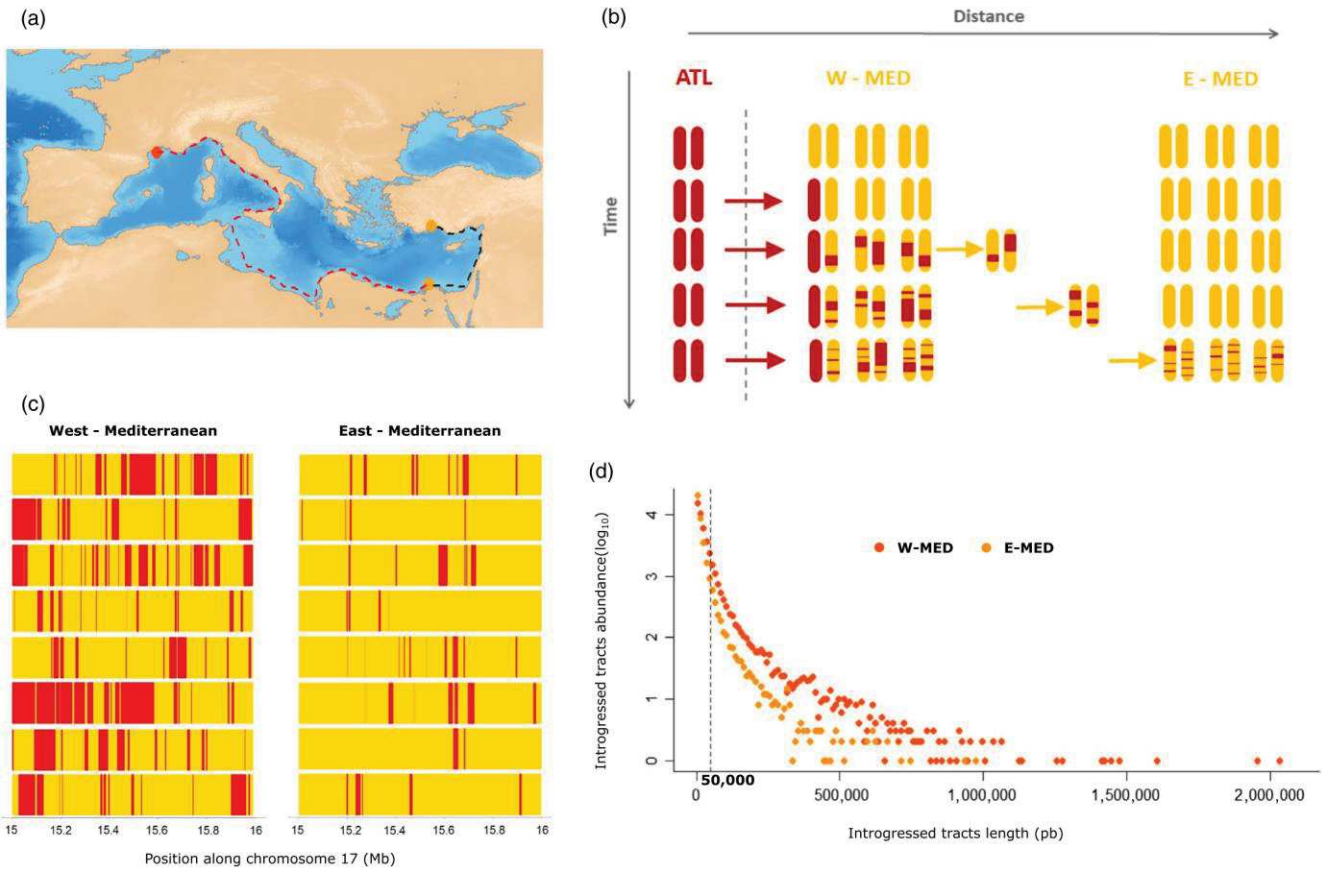


FIGURE 1 Introgression and diffusion of Atlantic tracts within the Mediterranean genetic background. (a) Geographical map showing the least cost dispersal distance of European sea bass (dotted lines) in continental waters of less than 200 m deep, between the western and the eastern Mediterranean Sea. Colored circles represent the geographical locations of western (orange) and eastern (yellow) Mediterranean samples. (b) Schematic representation of the diffusion-recombination process of Atlantic introgressed tracts over time and space in the Mediterranean Sea. At each time step, 2 chromosomes are represented for the Atlantic population and 6 for the western and eastern Mediterranean populations. Blocks of Atlantic ancestry are colored in red, and Mediterranean tracts are in yellow. Red arrows represent the transfer of Atlantic blocks into the Mediterranean background through hybridization, and yellow arrows represent the diffusion of Atlantic blocks due to gene flow among populations within the Mediterranean Sea. (c) Schematic representation of the mosaic of ancestry tracts from the Atlantic (red) and Mediterranean (yellow) populations in a 1 Mb region of chromosome 17 for 8 haplomes from the western and eastern Mediterranean populations. (d) Genomewide distribution of introgressed tract length for the western (orange, W-MED) and eastern (yellow, E-MED) Mediterranean populations, using four individuals for each population. The vertical dotted line represents the threshold length value (50 kb) below which we do not consider introgressed tracts to estimate the difference in introgression times between sampling locations

$\delta_{E\text{-}MED}$) to generate 8 families: 4 $\delta_{ATL} \times \varphi_{W\text{-}MED}$ and 4 $\delta_{E\text{-}MED} \times \varphi_{W\text{-}MED}$ (Figure 1a). This sampling design allowed generating phased whole-genome sequences from Mediterranean populations located at different distances from the contact zone with the Atlantic lineage (either near: W-MED, or far: E-MED). Since no genetic differentiation has been found between samples from Egypt and Turkey (Duranton et al., 2018), all E-MED individuals were grouped together within a single population. Low-quality and unphased genotypes were filtered to only retain sites with unambiguous transmission patterns and no missing data.

The filtered dataset consisting of 2,628,725 SNPs fully phased into chromosome-wide haplotypes was used to perform LAI (Duranton et al., 2018). Blocks of Atlantic origin introgressed into the Mediterranean genetic background were identified along each individual chromosome haplotype with Chromopainter (Lawson, Hellenthal, Myers, & Falush, 2012). We then refined the delineation of tract junctions to generate the length distribution of Atlantic migrant tracts separately for the western and eastern Mediterranean populations. The limited sampling size for each population in our trio design was largely compensated by the amount of haplotype information per sample, since each individual genome is composed of a mosaic of hundreds of Atlantic and Mediterranean tracts. Therefore, only a small number of phased whole-genome sequences provided sufficient information to obtain a clear picture of the admixture tract length distribution for the W-MED and E-MED populations (Duranton et al., 2018). This important aspect of the implemented methodology was also assessed using simulations (see below).

2.2 | Estimation of introgression time from migrant tract length

Once introgressed within a divergent genetic background, migrant tracts are progressively shortened by recombination across generations (Liang & Nielsen, 2014; Pool & Nielsen, 2009). Therefore, long migrant tracts are expected to have introgressed on average more recently than short migrant tracts. In the European sea bass, blocks of Atlantic ancestry must enter the Mediterranean Sea from its western side near the Gibraltar strait before they diffuse eastward (Figure 1a). This diffusion across the Mediterranean seascape takes a certain number of generations during which migrant tracts are eroded by recombination (Figure 1b). Therefore, the shift between the migrant tracts length distributions of two sampling locations at different distances from the contact zone directly reflects the time it takes for an Atlantic haplotype to diffuse from the nearest to the farthest location with respect to the contact zone. Here, we estimated the time of entrance for Atlantic tracts found in western and eastern Mediterranean populations and calculated the difference between these two estimates to evaluate the average time for a migrant tract to cross the Mediterranean Sea. To estimate the time of entrance in each location, we focused on neutral genomic regions and used an analytical expectation for the mean length of migrant tracts ($\bar{L}$) as a function of the time since initial admixture (t, expressed in generations), the admixture proportion of the population considered (f, the fraction of Atlantic ancestry), and the local recombination rate (r, in Morgan per base pair) (Racimo et al., 2015).

$$\bar{L} = [(1-f)r(t-1)]^{-1} \quad (1)$$

2.3 | Data filtering

The length distribution of migrant tracts is influenced by the temporal dynamics of introgression. Given that gene flow has been introducing Atlantic alleles within the Mediterranean since the end of the last glacial period (Tine et al., 2014), haplotypes of variable ages (and therefore variable lengths) are expected to be found at any Mediterranean location. The shortest tracts that reside in the Mediterranean for a much longer time than it takes to diffuse from west to east have very similar lengths among locations. Therefore, the shift in the length of Atlantic tracts between western and eastern Mediterranean population is all the more important that the tracts have introgressed recently and therefore remain long (Figure 1c,d). For that reason, we only considered blocks of Atlantic ancestry longer than 50 kb, since shorter blocks are less informative for estimating recent introgression. Applying a length threshold to remove short tracts has been also used to control for technical limitations to measure short introgressed tracts (Ni et al., 2016). Moreover, because migrant tracts length is more difficult to estimate in highly recombining regions of the genome, we excluded such regions from the analysis. In the sea bass genome, local recombination rates tend to be markedly reduced in central chromosomal regions ($\rho = 4N_e r$ usually <5 per kb) compared to chromosome extremities (ρ usually >40 per kb) (Tine et al., 2014). We thus applied a population-scaled recombination rate threshold of $\rho = 10$ per kb to keep only low-recombining regions.

Our analysis of migrant tract length relies on a neutral theory. In order to avoid potentially confounding effects of selection against Atlantic alleles, we filtered genomic regions that probably contain barrier loci that locally reduce gene flow between Atlantic and Mediterranean lineages. These regions, which represent ~4% of the genome (Duranton et al., 2018), were identified using the RND_{min} statistics (Rosenzweig, Pease, Besansky, & Hahn, 2016) calculated in 100 kb windows. This minimum relative node depth statistics corresponds to the ratio of the minimal value of d_{XY} calculated between all Atlantic and eastern Mediterranean individuals (d_{min}) to the mean value of d_{XY} measured between *D. labrax* and the outgroup species *Morone saxatilis* (d_{out}). An empirical upper threshold RND_{min} value of 0.03 (i.e., corresponding to the 95th percentile of the distribution) was used to conservatively exclude genomic regions influenced by selection from our dataset, according to previous results (Duranton et al., 2018).

2.4 | Estimation of the average tract length ($\bar{L}$)

We used two different approaches to estimate the average introgressed tract length for each of the two Mediterranean populations. Our first method specifically addresses the direct influence of local recombination rate (r) variations on the length distribution of introgressed tracts. Broadscale variation in recombination rate along chromosomes is commonly observed in eukaryotes (Haenel, Laurentino, Roesti, & Berner, 2018), including teleost fishes (Bradley et al., 2011; Roesti, Hendry, Salzburger, & Berner, 2012) and among them the European sea bass (Tine et al., 2014). As a result, the length of migrant tracts is expected to decrease at variable rates across the sea bass genome, even though we excluded the most highly recombining regions from our analysis. To account for these variations, we calculated the average length of introgressed tracts locally within nonoverlapping 100 kb windows, using all introgressed tracts that were either fully contained within, or simply overlapping each focal window. The average time since introgression was then estimated separately for each window using equation (1) with the average length of introgressed tracts and the local recombination rate value estimated for that window (Tine et al., 2014). We also used the observed genomewide admixture proportions for the western ($f = 0.31$) and eastern ($f = 0.13$) Mediterranean populations (Duranton et al., 2018). After removing windows showing no introgressed tracts, we retained a total of 2,092 and 1,065 windows for the western and eastern Mediterranean populations, respectively. We then merged time estimates across windows to generate the distribution of introgression times separately in each Mediterranean population. Finally, we bootstrapped both distributions 10,000 times and identified the maximum value of each bootstrap replicate. We then used the 0.025 and 0.975 quantile values of bootstrapped maxima to estimate a

95% confidence interval for the maximum value of the distribution in each population. The difference between the two maxima represents the average number of generations taken for a migrant tract to cross the Mediterranean Sea.

The second method builds on the fact that the abundance of introgressed tracts as a function of their length follows an exponential distribution (Gravel, 2012; Pool & Nielsen, 2009; Racimo et al., 2015) with a mean $\bar{L}=\frac{1}{\lambda}$, where λ corresponds to the rate parameter of the exponential distribution. To estimate λ, we log-transformed the tract abundance values from the introgressed tract length distribution to obtain a linear distribution which slope equals $-\lambda$. In order to estimate this slope using data from similar recombination rate regions, genomic windows of 100 kb were grouped into eleven recombination rate categories, which were designed to receive an equal number of windows (i.e., 209 windows in each category). For each category, we then separated the tracts into twenty bins of tract length and used only bins with at least five tracts to fit the linear regression. Finally, we plotted the values of $\lambda=\frac{1}{\bar{L}}$ estimated for every recombination rate category as a function of the average recombination rate of the 209 windows used in the corresponding category. We fitted a linear regression to this distribution forcing the intercept to equal zero and determined its slope $a = (1 - f)(t - 1)$, (where f corresponds to the admixture proportion) which allowed us to estimate the time since introgression as $t=\frac{a}{1-f}+1$ separately for the eastern and western Mediterranean populations. The difference between the two estimates corresponds to the number of generations necessary for a track to diffuse from west to east.

2.5 | Estimate of the least coast distance between the Mediterranean populations

We used the R package marmap (Pante & Simon-Bouhet, 2013) to estimate the least cost distances between western and eastern Mediterranean sampling locations. Since the European sea bass is a neritic benthopelagic species occupying shallow continental waters (Pickett & Pawson, 1994), we considered that dispersal only occurs through areas where the maximum depth is less than 200 m. We estimated the distance between the western and both north and southeastern Mediterranean locations as $Dist_{\text{west-south\_east}}$ = 4,891 km and $Dist_{\text{west-north\_east}}$ = 6,005 km (Figure 1a). Since we analyzed all eastern individuals together without separating northern from southern samples, we used the average distance $Dist_{\text{west-east}}$ = 5,448 km between the western and the two eastern populations to calculate the per-generation dispersal distance.

2.6 | Validation of the methodology by simulations

We used neutral simulations to test whether the length distribution of introgressed tracts can be used to reliably estimate the diffusion time of Atlantic haplotypes between western and eastern Mediterranean populations. To do so, we used the coalescent simulator msprime v0.6.2 (Kelleher, Etheridge, & McVean, 2016) to simulate the length distribution of Atlantic tracts introgressed within the western Mediterranean population under a secondary contact model (see legend of Figure 4a). Demographic and temporal simulation parameters were set to values that were previously shown to accurately reproduce this distribution (Duranton et al., 2018). We then aimed at modeling the diffusion of introgressed tracts from the western toward the eastern Mediterranean population using the same simulation framework. This diffusion is characterized by the decay of introgressed haplotype length due to the recombination events that occur every generation during the time it takes to cross the Mediterranean Sea, and it is therefore not directly influenced by the Atlantic population. To model this diffusion, we thus considered a third episode lasting for T_{diff} generations, in which gene flow between the Atlantic and western Mediterranean population stops. During these T_{diff} generations, the erosion of introgressed haplotypes that are not renewed anymore by gene flow from the Atlantic mimics what happens when haplotypes diffuse away from the western to reach the eastern Mediterranean population. In this way, the spatial diffusion process of introgressed tracts is simply modeled by reproducing the temporal erosion of introgressed tracts after they enter the western Mediterranean, without making spatially explicit simulations. To identify and measure the exact length of introgressed tracts within simulated Mediterranean genomes, we used functions implemented in Skov et al., (2018), which track the local ancestry of recombining genomic segments during the course of the coalescent simulations. We simulated a single chromosome of 25 Mb (i.e., the average length of *D. labrax* chromosomes) with constant recombination and mutation rates of $6.84e^{-8}$ (i.e., the mean recombination rate calculated across the sea bass genome) and $1e^{-8}$, respectively.

To estimate the introgression time from simulated tracts sampled at times Date_T_{diff} and Date_T_0 within the Mediterranean population, we used Equation (1) in which $\bar{L}$ was calculated as the average length of introgressed tracts and f as the genomic fraction occupied by introgressed tracts. The difference between the estimated time since introgression at Date_T_0 and Date_T_{diff} represents the diffusion time of Atlantic haplotypes (i.e., parameter T_{diff} in our model), as they move from the western to the eastern Mediterranean population.

Our first objective was to determine whether the proposed approach allows to reliably estimate the diffusion time parameter T_{diff}. We therefore parameterized our historical demographic model using the same parameter values as in Duranton et al., (2018) (i.e., T_{div} = 54,000 generations, T_c = 2,300 generations and $N_{\text{MED}}{}^{*}m_1$ = 7 migrants per generation) and explored a range of T_{diff} values spanning those estimated between the two Mediterranean populations (i.e., 20, 50, 150, 350, 550, 800, 1,200, and 1,600 generations). We performed 10 replicate simulations for each T_{diff} value and sampled 4 individuals per population at each sampling time, which corresponds to the number of individuals used in this study.

Our second objective was to determine the robustness of our method to different sample sizes and demographic histories. In order to assess the effect of the amount of data, we used the parameter values of the European sea bass (i.e., T_c = 2,300 and T_{diff} = 550 generations, $N_{\text{MED}}{}^{*}m_1$ = 7 migrants per generation) and sampled either 1, 2, 3, 4, 5, 6, or 7 individuals at each time point.

Simulations were run 10 times for each value, and T_{diff} was estimated for every replicate. We then fixed the number of sampled individuals to 4 and made the secondary contact duration parameter (T_c) vary around the estimated number of generations of admixture in the sea bass (i.e., using T_c = 250, 500, 1,000, 1,500, 2,300, 3,500, and 5,000). We then did the same for the number of migrants entering the Mediterranean population per generation (using $N_{MED}{}^{*}m_1$ = 1, 3, 5, 7, 10, 15, and 20). These explored values allowed us to consider a wide range of gene flow durations and intensities, which partly covers the diversity of settings found in other species. We did not make the divergence time (T_{div}) vary since this parameter mostly influences the accuracy of the detection of introgressed tracts, which were called directly in our simulations instead of being inferred with a LAI method as we did from real data.

3 | RESULTS

3.1 | Dating introgression in nonoverlapping 100 kb windows

Our first method, which estimates the time since introgression using the average length of admixture tracts in nonoverlapping 100 kb window, specifically accounted for local recombination rate variation among windows. By combining the estimated times across all retained windows, we obtained a distribution of introgression time for the western and eastern Mediterranean populations, separately (Figure 2). The two distributions showed similar shapes (Figure 2a), except that more recently introgressed Atlantic tracts were observed in the western compared to the eastern Mediterranean population. Furthermore, the eastern distribution was slightly shifted toward longer introgression times, indicating that introgressed tracts are on average older in this population (Figure 2b). We estimated the maximum of each distribution and its 95 percent confidence interval as T_{west_max} = 831.01 (CI95% = [721.55; 967.81]) and t_{east_max} = 1,186.15 (CI95% = [1,071.48; 1,329.47]) generations for the western and eastern Mediterranean, respectively. Therefore, there was a difference in diffusion time of t_{diff_1} = 1,186.15–831.01 = 355.14 (CI95% = [103.67; 607.92]) generations between locations. This difference corresponds to a per-generation dispersal distance of $d_{west\text{-}east\_1}$ = 5,448/355.14 = 15.34 km ($CI_{95\%}$ = [8.96; 52.55]).

3.2 | Dating introgression using the log-transformed distribution of tracts length

Our second method that modeled the log-transformed distribution of admixture tracts length to estimate the mean tract length ($\bar{L}$) relied on the analysis of 2,299 windows grouped into eleven recombination rate categories (Table 1). Although we delimited the range of each recombination rate category to evenly distribute windows across categories, the total amount of information slightly differed among categories due to varying amounts of admixture tracts per window. For that reason, the slope of the regression of the log-transformed distribution of admixture tracts length was only marginally significant for some recombination rate categories with limited amount of data (i.e., five categories in the eastern population, Table 1). As expected, the estimated average length of introgressed tracts was shorter in the eastern compared to the western population for every category. We then plotted the eleven estimated values of $\frac{1}{\bar{L}}$ as a function of their corresponding recombination rate, separately for the western and eastern Mediterranean populations (Figure 3). We estimated the slope of the linear regression to a_{west} = 323.49 (*SE* = 53.23; *R*-squared = 0.77 and *p*-value = $1.19e^{-4}$) for the western and to a_{east} = 1,100.2 (*SE* = 151.67; *R*-squared = 0.82 and *p*-value = $2.75e^{-5}$) for the eastern population. Using these values, we estimated t_{west} to 469.83 (CI95% = [315.54; 624.12]) and t_{east} to 1,265.60 (CI95% = [916.92; 1,265.60]) generations. Thus, we estimated a diffusion time t_{diff_2} = 1,265.60–469.83 = 795.77 generations (CI95% = [292.8; 950.06]) and a per-generation dispersal distance of $d_{west\text{-}east\_2}$ = 5,448/795.77 = 6.85 km per generation (CI95% = [5.73; 18.61]).

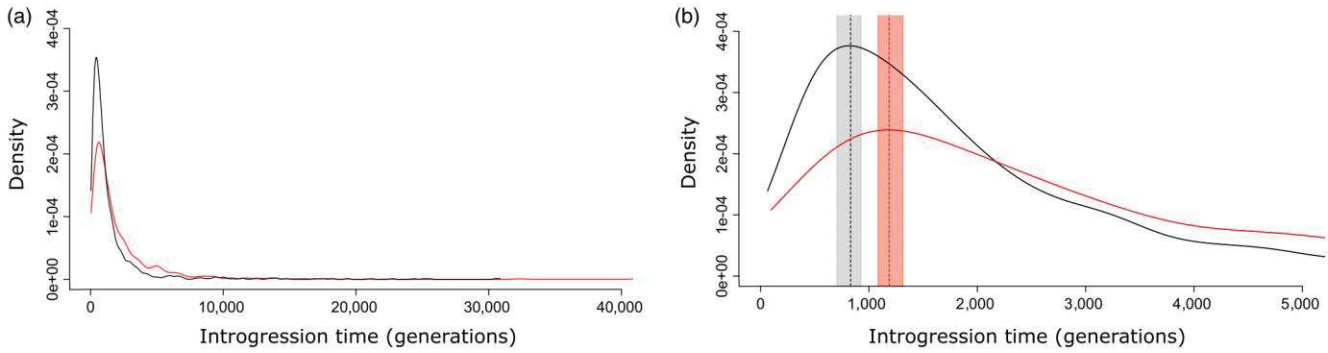


FIGURE 2 Distributions of estimated time since introgression. Times were estimated in non-overlapping 100 kb windows, separately for the western (black) and eastern (red) Mediterranean populations. (b) Dotted vertical bars show the maximum of each distribution, and shaded rectangles represent their 95 percent confidence intervals

TABLE 1 Summary statistics of the linear correlations between the log-transformed number of admixture tracts and their length in eleven categories of local recombination rate

Recombination rate (M per bp)	Number of windows	Population	Number of tracts	$\lambda = \frac{1}{L}$	*SE*	*R*-squared	*p*-value
$6.10e^{-10}$	209	W-MED	689	$2.06e^{-6}$	$1.66e^{-7}$	0.93	**$2.14e^{-7}$**
		E-MED	306	$3.74e^{-6}$	$1.10e^{-6}$	0.57	0.012
$1.59e^{-9}$	209	W-MED	831	$2.17e^{-6}$	$2.35e^{-7}$	0.86	**$4.38e^{-7}$**
		E-MED	294	$1.12e^{-5}$	$1.44e^{-6}$	0.90	**$5.55e^{-4}$**
$2.32e^{-9}$	209	W-MED	856	$2.54e^{-6}$	$3.95e^{-7}$	0.76	**$3.21e^{-7}$**
		E-MED	283	$7.06e^{-6}$	$1.46e^{-6}$	0.67	**$6.73e^{-4}$**
$3.21e^{-9}$	209	W-MED	823	$4.50e^{-6}$	$5.42e^{-7}$	0.88	**$3.32e^{-5}$**
		E-MED	174	$1.11e^{-5}$	$2.63e^{-6}$	0.67	**0.004**
$4.09e^{-9}$	209	W-MED	931	$5.52e^{-6}$	$4.27e^{-7}$	0.95	**$3.88e^{-6}$**
		E-MED	219	$1.11e^{-5}$	$2.11e^{-6}$	0.75	**$7.75e^{-4}$**
$5.26e^{-9}$	209	W-MED	736	$2.29e^{-6}$	$4.77e^{-7}$	0.65	**$5.52e^{-4}$**
		E-MED	175	$1.05e^{-5}$	$2.03e^{-6}$	0.74	**$8.63e^{-4}$**
$7.064e^{-9}$	209	W-MED	874	$4.56e^{-6}$	$5.76e^{-7}$	0.86	**$2.41e^{-5}$**
		E-MED	218	$1.56e^{-5}$	$7.81e^{-7}$	0.99	**$2.72e^{-4}$**
$9.040e^{-9}$	209	W-MED	713	$3.72e^{-6}$	$5.65e^{-7}$	0.71	**$6.26e^{-6}$**
		E-MED	144	$1.69e^{-5}$	$4.10e^{-6}$	0.76	0.014
$1.278e^{-8}$	209	W-MED	657	$7.15e^{-6}$	$6.70e^{-7}$	0.90	**$3.80e^{-7}$**
		E-MED	116	$6.66e^{-6}$	$2.52e^{-6}$	0.40	0.029
$1.786e^{-8}$	209	W-MED	652	$4.38e^{-6}$	$8.73e^{-7}$	0.67	**$3.90e^{-4}$**
		E-MED	112	$1.06e^{-5}$	$3.55e^{-6}$	0.57	0.031
$3.807e^{-8}$	209	W-MED	556	$1.03e^{-5}$	$1.60e^{-6}$	0.83	**$3.57e^{-4}$**
		E-MED	76	$4.28e^{-5}$	$1.06e^{-5}$	0.83	0.057

Note: Model summaries are presented separately for the western and eastern Mediterranean populations. The *p*-values of significant regressions after applying Bonferroni correction for multiple tests ($p < 0.0045$) appear in bold.

3.3 | Validation of the methodology using simulations

We used simulated data to test whether the average length of introgressed tracts measured at two time points after entering a recipient population can be used to reliably estimate their diffusion time. We showed that there is a strong correlation between the simulated and estimated values of diffusion times (T_{diff}) (Figure 4b). This indicates that measuring the difference in time since admixture between two populations connected by gene flow allows to accurately estimate the number of generations that it takes to connect them through dispersal. Although we did not explicitly consider the spatial diffusion process here, this should not strongly affect the realism of our simulations since the erosion of tracts mostly depends on recombination and time. Indeed, for neutral regions, the length distribution of introgressed tracts is not influenced by the effective population size. Furthermore, the equation used here to estimate the time since admixture assumes a unique pulse of historical admixture (Gravel, 2012; Racimo et al., 2015). Because we simulated a continuous period of gene flow, our results also indicate that the used formula remains applicable in a context of continuous migration.

We further tested if our estimations of diffusion time were robust to a range of sample sizes and demographic histories. Results were relatively stable across a wide range of sample sizes, even when only one individual was used (Figure 4c). To account for varying demographic histories, we modeled different durations and intensities of gene flow and showed that none of these parameters strongly impacts the estimation of diffusion time (Figure 4d and e). The variance in estimated values among replicates was high only in cases of low introgression levels, which were either due to a low migration rate or a short period of gene flow. This indicates that in such cases, introgressed tracts were not abundant enough to precisely estimate their average length with only four individuals sampled at each time point. Increasing the sample size in cases of limited introgression would thus probably improve the reliability of time estimates. It is also important to consider that we only simulated a single 25 Mb chromosome per replicate run due to computational limitations. However, real datasets such as the one used here in the European sea bass cover the whole genome (i.e., 24 chromosomes) and therefore contain larger amounts of data. Therefore, the proposed methodology is likely to give consistent results even with only small numbers of genomes sequenced in each population.

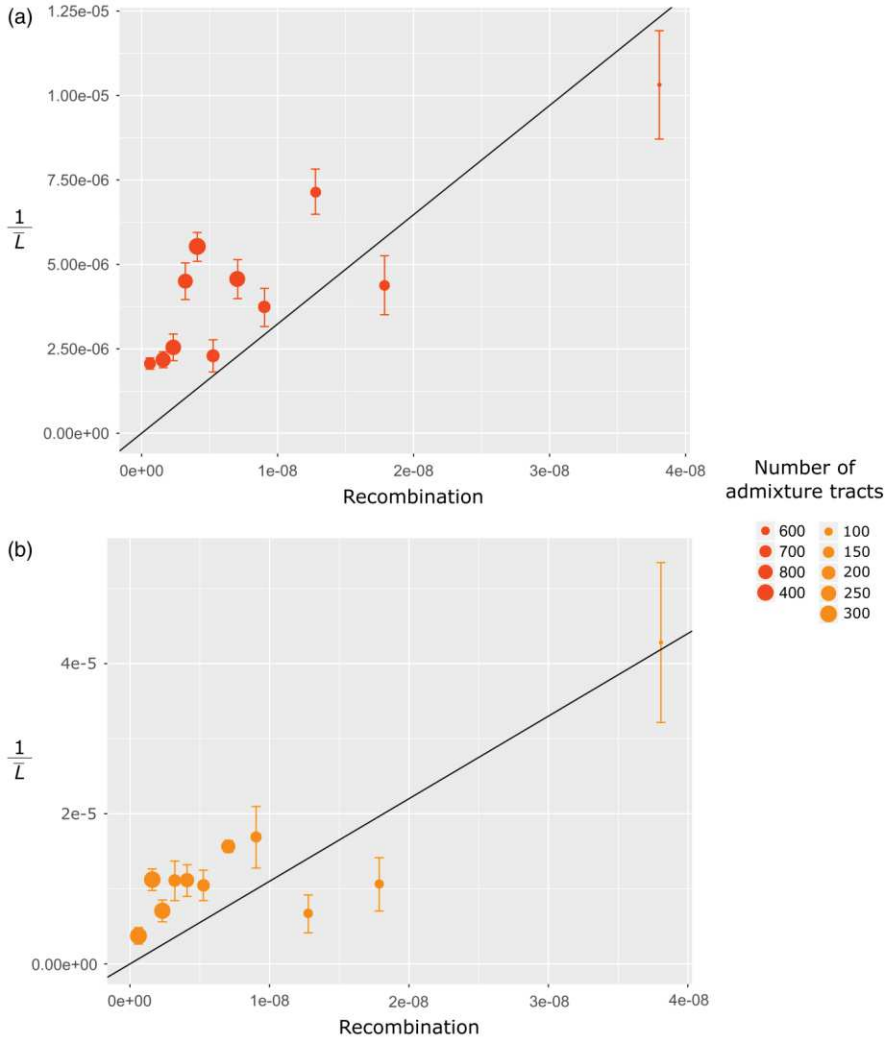


FIGURE 3 Correlations between $\frac{1}{L}$ and the average recombination rate estimated for each of the eleven recombination rate categories (each containing 209 genomic windows) for the eastern (a) and western (b) Mediterranean population. The size of the points indicates the number of admixture tracts contained in the 209 windows of each recombination rate category, and the vertical bars indicate the standard error of the estimated statistics. The black line shows the linear regression fitted for each population

4 | DISCUSSION

We used the information contained in the length of admixture tracts as a means to estimate the spatial scale of dispersal within a population receiving genetic material from a distinct lineage. Introgressed tracts entering a recipient population are progressively shortened every generation by recombination, providing a clock that keeps track of the history of introgression (Liang & Nielsen, 2014; Pool & Nielsen, 2009). Here, the proposed methodology relies on the fact that introgressed tracts get on average shorter when they reach locations farther away from the contact zone, a process considered to be relatively independent from the effective population size (Gravel, 2012; Racimo et al., 2015). The difference in tract length between two locations at different distances from the contact zone thus represents the action of recombination during the time needed to connect these two locations through multigenerational dispersal (Figure 1b) (Duranton et al., 2018). Using empirical data and coalescent simulations, we show that this approach provides a means to estimate the spatial scale of dispersal in populations that are not at migration–drift equilibrium.

Despite our efforts to filter the European sea bass data prior to the estimation of time parameters, a few confounding factors may have had undesirable effects on our analyses. The first one is selection against introgressed Atlantic tracts in some regions of the Mediterranean genomes. To control for that effect, we conservatively removed genomic windows that showed signs of selection against introgression, in order to consider only regions where Atlantic tracts have likely introgressed neutrally. A second possible source of bias comes from technical limitations of LAI methods to estimate the length of short introgressed tracts (Ni et al., 2016). This category of tracts is, however, less informative for the type of analysis presented here, since their whole history of recombination within the Mediterranean is likely much longer than the time needed to diffuse across the Mediterranean Sea. On the contrary, long migrant tracts are more likely to display contrasted lengths between remote locations. For these reasons, we did not consider highly recombining genomic windows ($4N_e r > 10$) where the differential in tract length is rapidly lost, as well as introgressed fragments shorter than 50 kb in the remaining windows.

Although removing short tracts allows extracting the most informative fraction of the data, it may have biased the estimation of the time since introgression obtained from our first methodology. This filtering step should indeed slightly increase the average tract

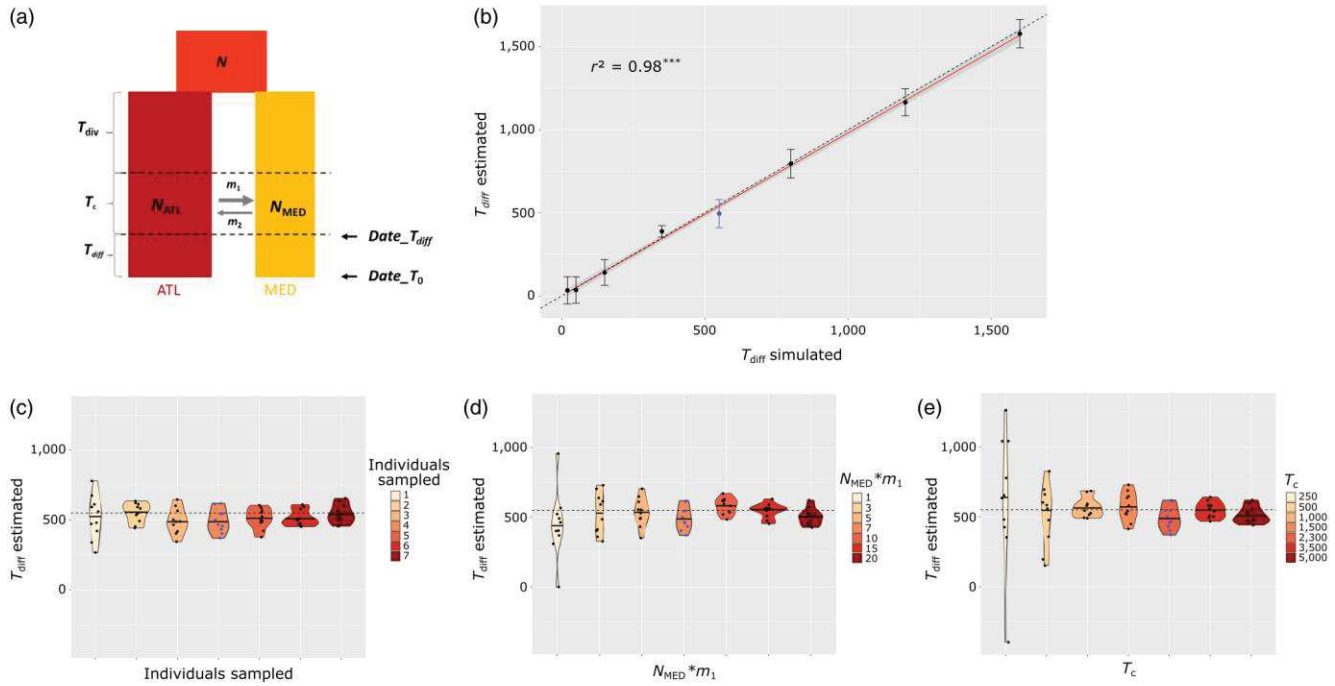


FIGURE 4 Reliability evaluation of the approach to estimate the diffusion time separating two sampling time points. (a) Schematic representation of the model used in simulations. An ancestral population of size N splits into two populations ATL and MED of size N_{ATL} and N_{MED}, which diverge during T_{div} generations and then exchange genes for T_c generations and then follows a third episode without gene flow lasting for T_{diff} generations, which represents the diffusion period during which the length of introgressed tracts is only influenced by the erosion process in the recipient population. The Mediterranean population is sampled two times, the first one at the beginning of the period without gene flow (Date_T_{diff}) and the second one at the end (date_T_0). The migration rates from the ATL to the MED population and in the opposite direction are represented by parameters m_1 and m_2, respectively. (b) Correlation between the estimated and simulated values of T_{diff} using 4 individuals per sampling time point, T_c = 2,300 and m_1*N_{MED} = 7. For each value of T_{diff}, simulations were run 10 times, each point represents the average estimated value over replicates and the vertical bar the standard error. The dashed line represents the equation y = x, and the red line is the linear regression with its confidence interval in grey shade. Estimated values of T_{diff} using model parameters estimated for the European sea bass: T_c = 2,300, m_1*N_{MED} = 7 and T_{diff} = 550 (represented by the dashed line) using (c) different sample sizes, (d) N_{MED}*m_1 values, and (e) durations of contact (T_C). Points represent estimated T_{diff} values of the 10 replicate simulations for each tested value, and horizontal bars their median value. Blue points represent simulations that were modeled using the same model parameter values as in the European sea bass

length within windows, which in turn should affect the distribution of the time since introgression. Therefore, the estimated time may be underestimated for both populations, but possibly more so for the eastern population that contains a relatively higher fraction of short tracts. Our filtering of short tracts may thus lead to an underestimation of the diffusion time between locations, that is, an overestimation of the per-generation dispersal distance.

To overcome this potential difficulty, we used a second methodology that models the mean length of introgressed tracts from their log-transformed distribution. As such, this approach is largely independent of the distribution tail and thus insensitive to removing short tracts. This filtering step only reduces the amount of data and therefore the power of the regression approach, but without modifying the regression slope. On the downside, this method needs to group windows with similar recombination rate values so that each category has enough introgressed tracts to perform powerful regressions. The average recombination rate value used for each of the eleven categories may cause some loss of precision regarding fine-scale recombination rate variation, as compared to our first approach. However, windows within a given category displayed a small variance in recombination rate. Therefore, we speculate that averaging recombination rate values among windows within categories did not strongly affect our inferences. A result supporting this conjecture was the reasonably high amounts of variance in mean tracts length explained by the linear regression models fitted for each recombination rate category (Table 1).

Overall, our two methodologies can be seen as complementary. The first one allows to consider more fine-scaled variations in recombination rate along the genome but might be sensitive to the removal of short tracts in the presence of historical admixture. The second approach is probably robust to the removal of short tracts but at the expense of summarizing variation in recombination rate within bins. Despite these differences, the two values of the time since introgression estimated for the eastern population (t_{east_max} = 1,186.15 ($CI_{95\%}$ = [1,071.48; 1,329.47]) and t_{east} = 1,265.60 (CI95% = [916.92; 1,265.60])) were very similar, and the second method provided an older estimate, as expected. The two values estimated for the western population (T_{west_max} = 831.01 (CI95% = [721.55; 967.81]) and

t_{west} = 469.83 (CI95% = [315.54; 624.12])) were, however, more different, with a younger estimate obtained with the second method, which is contrary to our expectation. A possible explanation could be that since the western population is closer from the contact zone, it has a higher variance in tract length that reduces the precision of estimated time. Nevertheless, our two quantitative estimates of the per-generation dispersal distance were very close to each other and displayed largely overlapping confidence intervals (first method: $d_{west\text{-}east\_1}$ = 15.34 km ($CI_{95\%}$ = [8.96; 52.55]) and second method: $d_{west\text{-}east\_2}$ = 6.85 km (CI95% = 5.73; 18.61])). As expected, the second method which is less prone to overestimate dispersal due to the removal of short tracts provided a smaller estimate the per-generation dispersal distance. Since our simulation study also supported the validity of the implemented approach (Figure 4b), we are confident that our empirical numerical estimates provide reliable indications of the spatial scale of dispersal in the European sea bass.

One possible limitation of this study could be that the analytical expectation we used assumes a unique pulse of admixture (Gravel, 2012; Racimo et al., 2015), while gene flow between the two sea bass lineages has been ongoing since the last glacial retreat (Tine et al., 2014). Methods accounting for continuous gene flow (Gravel, 2012; Ni et al., 2016) provide more realistic modeling of the migration history but are more suitable for inferring recent admixture (i.e., around 100 generations). Therefore, such methods would have little power with cases of postglacial gene flow, as it is the case in the European sea bass. Furthermore, our simulations showed that using the theoretical expectation for a single pulse of admixture provides rather accurate results (Figure 4b). Therefore, our choice of methodology offers a good compromise considering the methods currently available. We anticipate, however, that the increasing accessibility to local ancestry information in population genomic studies will foster the development of new methods that better account for continuous historical gene flow.

Our numerical estimations seem consistent with previous population genetics studies demonstrating the existence of a genetic structure between western and eastern Mediterranean populations using allozymes (Allegrucci, Fortunato, & Sbordoni, 1997), microsatellite markers (Bahri-Sfar, Lemaire, Hassine, & Bonhomme, 2000; Quéré et al., 2012), and SNPs (Souche et al., 2015). These results are also in line with the suspected philopatric behavior of the European sea bass (Bahri-Sfar et al., 2000; Castilho & Ciftci, 2005; de Pontual et al., 2019). Nevertheless, since we only used two sampling locations for this proof-of-concept study, our estimated dispersal distances should be interpreted with caution. Indeed, they represent an average dispersal distance calculated between two distant populations. Therefore, inferred distances may be decreased by local dispersal barriers that reduce gene flow somewhere in between the two sampling sites. For instance, water circulation features in the Siculo-Tunisian strait are known to reduce gene flow between the western and eastern Mediterranean populations in several Mediterranean fish species (but see Pascual, Rives, Schunter, & Macpherson, 2017) including sea bass (Bahri-Sfar et al., 2000; Quéré et al., 2012). A finer-scale sampling of the Mediterranean Sea could thus allow refining local estimates of dispersal distance to test and quantify the effect of such barriers on dispersal. Nonetheless, our estimates already provide relevant ecological information on the spatial scale of dispersal in Mediterranean sea bass for conservation and management purposes. Even accounting for uncertainty, our results indicate relatively short effective dispersal distances (<50 km per generation) considering the potential offered by the larval phase of 8 to 12 weeks and the mobility of juvenile and adult stages (Bahri-Sfar et al., 2000). This illustrates the complex relationships existing between pelagic larval duration and gene flow in marine species (Nanninga & Manica, 2018; Selkoe & Toonen, 2011) and raises the question of the long-distance benefits of marine reserves in terms of demographic connectivity (Manel et al., 2019).

Here, we used the European sea bass as a case study to illustrate the potential of admixture tracts for estimating dispersal in non-equilibrium populations. However, our main message is that similar approaches could be applied to a wide range of species, especially marine organisms in which estimating dispersal distances remains a challenging issue (Gagnaire et al., 2015). Despite the apparent lack of strong physical barriers to dispersal in the marine realm, genetic studies have shown that many species are subdivided into genetically interacting cryptic lineages, ecotypes, or partially reproductively isolated populations (see Bierne, Welch, Loire, Bonhomme, & David, 2011 for a review). For instance, even at the regional scale of the North Atlantic Ocean, such subdivisions have been illustrated by a number of studies in several commercial fish species including the Atlantic herring (*Clupea harengus*) (Guo, Li, & Merilä, 2016; Lamichhaney et al., 2012; Limborg et al., 2012; Martinez Barrio et al., 2016), the Atlantic cod (*Gadus morhua*) (Berg et al., 2016; Bradbury et al., 2014; Hemmer-Hansen et al., 2013; Karlsen et al., 2013; Sodeland et al., 2016), the turbot (*Scophthalmus maximus*) (Nielsen, Nielsen, Meldrup, & Hansen, 2004; Vandamme et al., 2014), the European hake (*Merluccius merluccius*) (Milano et al., 2014; Nielsen et al., 2012), and the European anchovy (*Engraulis encrasicolus*) (Le Moan, Gagnaire, & Bonhomme, 2016). The lack of reliable measures of dispersal within and among populations of such species can generate mismatches between the delineation of fisheries management units and conservation objectives (Reiss, Hoarau, Dickey-Collas, & Wolff, 2009). Moreover, spatial patterns of admixture can be confounded with evidence for either strong isolation-by-distance or local adaptation, since introgression gradients tend to increase genetic differentiation between populations located at different distances from a contact zone (Gagnaire et al., 2015). If not accounted for, such gradients may lead to an underestimation of dispersal distances from isolation-by-distance patterns. Furthermore, when introgression and environmental gradients are spatially overlapped, as it is commonly the case, it is very difficult to distinguish loci under local selection from introgression clines at neutral loci. Therefore, the risk of misinterpreting dispersal and local adaption patterns is particularly high when admixture occurs between genetically differentiated groups. A careful analysis of genetic variation including demographic inferences to reconstruct the evolutionary history of the studied populations thus appears to be a first important step in genetic connectivity studies.

An important prerequisite for applying the methodology developed in our study is to accurately identify and measure introgressed tracts. Although this may require phased haplotype data to perform LAI, haplotype phasing approaches are making this task increasingly feasible (Browning & Browning, 2011; Rhee et al., 2016). Having access to a chromosome-level reference genome assembly will no longer remain necessary with haplotype-resolved genome sequencing methods based on long read sequencing technologies (Browning & Browning, 2011; Snyder et al., 2015). In parallel to ongoing progress in sequencing, a wide variety of methods for LAI have been developed (Geza et al., 2018; Yuan et al., 2017). Some of them allow to correct phasing errors while identifying admixture tracts (Dias-Alves, Mairal, & Blum, 2018; Maples, Gravel, Kenny, & Bustamante, 2013; Salter-Townshend & Myers, 2018), making LAI accessible with data that are phased with a large variety of methods, including statistical phasing. Phased sequence data are even not necessary with some methods that only require the physical (Baran et al., 2012) or genetic (Guan, 2014) position of variants to identify ancestry blocks (Baran et al., 2012; Guan, 2014). Recently, a new method has also been developed to perform LAI directly from reads pileup data in population samples with arbitrary ploidy (Corbett-Detig & Nielsen, 2017). Therefore, there is a good potential for using the length of tracts to date admixture, including with reduced-representation sequencing data that still represent the most common type of data used in population genomic studies. An example of this kind of approach has been successfully applied with ddRAD SNPs for identifying introgressed tracts in supplemented populations of wild brown trout (Leitwein, Gagnaire, Desmarais, Berrebi, & Guinand, 2018).

The ability to correctly estimate the length of introgressed tracts admittedly depends on the density of markers. However, the minimal marker density required depends on the average length of introgressed tracts and therefore on the time since the beginning of admixture. The more recent the introgression is, the less the density of genetic markers is necessary. In all cases, the precise delineation of migrant tracts will be facilitated by a stronger divergence between admixing lineages (Gravel, 2012). Therefore, reduced-representation sequencing approaches such as RAD-sequencing, which can generate from 10 to 1,000 loci per Mb (Andrews, Good, Miller, Luikart, & Hohenlohe, 2016), can offer the flexibility suitable to date both recent admixture between young lineages and more ancient introgression between divergent lineages.

Finally, our simulation study showed that the proposed methodology can give accurate results for a wide range of sample sizes and demographic histories, and this for inferring a large range of diffusion times. For instance, assuming a demographic history similar to that of the European sea bass, we showed that a single individual may be sufficient to provide reliable results. More generally, the method can be applied either with a small number of whole-genome sequences for cases of historical gene flow or using larger sample sizes with reduced-representation sequencing data for studying recent admixture. Finally, because the method seems to be accurate over a large range of gene flow intensities and diffusion times, it can be used to measure dispersal at a more refined spatial scale than the one considered here.

5 | CONCLUSION

Our study illustrates the potential of admixture tracts for estimating the spatial scale of dispersal in nonequilibrium populations, which is an essential parameter to design appropriate management and conservation actions. Although methodological improvements will be needed to better account for ancient migrations, the proposed approach provides a roadmap to generate valuable information on a conservation-relevant timescale and is already well suited for species with relatively recent admixture histories. The development of new methods that simultaneously estimate local ancestry and the time since admixture (Corbett-Detig & Nielsen, 2017) should further accelerate the interest for this kind of approach. This is especially true for species in which direct measures of dispersal are not applicable and the neutral migration–drift balance is not informative or simply does not exist, which is the case for many marine species.

ACKNOWLEDGEMENTS

This work was supported by the ANR grants LABRAD-SEQ 11-PDOC-009-01 and CoGeDiv ANR-17-CE02-0006-01 to P.-A.G. We thank the International Marine Connectivity Network (GDRI iMarCo) for insightful discussions. The authors are also grateful to the Associate Editor Luciano Beheregaray and two anonymous reviewers for their constructive comments.

CONFLICT OF INTEREST

None declared.

DATA ACCESSIBILITY

Sequence reads are available on GenBank under the accession code PRJNA472842 (Duranton et al., 2018).

ORCID

Maud Duranton https://orcid.org/0000-0003-3943-8061
François Bonhomme https://orcid.org/0000-0002-8792-9239
Pierre-Alexandre Gagnaire https://orcid.org/0000-0002-1908-3235

REFERENCES

Allegrucci, G., Fortunato, C., & Sbordoni, V. (1997). Genetic structure and allozyme variation of sea bass (*Dicentrarchus labrax* and *D. punctatus*) in the Mediterranean Sea. *Marine Biology*, *128*, 347–358. https://doi.org/10.1007/s002270050100

Andrews, K. R., Good, J. M., Miller, M. R., Luikart, G., & Hohenlohe, P. A. (2016). Harnessing the power of RADseq for ecological and evolutionary genomics. *Nature Reviews Genetics*, *17*, 81–92. https://doi.org/10.1038/nrg.2015.28

Bahri-Sfar, L., Lemaire, C., Hassine, O. K. B., & Bonhomme, F. (2000). Fragmentation of sea bass populations in the western and eastern Mediterranean as revealed by microsatellite polymorphism. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, *267*(1446), 929–935. https://doi.org/10.1098/rspb.2000.1092.

Baran, Y., Pasaniuc, B., Sankararaman, S., Torgerson, D. G., Gignoux, C., Eng, C., … Halperin, E. (2012). Fast and accurate inference of local ancestry in Latino populations. *Bioinformatics*, *28*, 1359–1367. https://doi.org/10.1093/bioinformatics/bts144

Berg, P. R., Star, B., Pampoulie, C., Sodeland, M., Barth, J. M. I., Knutsen, H., … Jentoft, S. (2016). Three chromosomal rearrangements promote genomic divergence between migratory and stationary ecotypes of Atlantic cod. *Scientific Reports*, *6*, 23246. 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*, 2044–2072. https://doi.org/10.1111/j.1365-294X.2011.05080.x

Bradbury, I. R., Bowman, S., Borza, T., Snelgrove, P. V. R., Hutchings, J. A., Berg, P. R., … Bentzen, P. (2014). Long distance linkage disequilibrium and limited hybridization suggest cryptic speciation in Atlantic Cod. *PLoS ONE*, *9*, e106380. https://doi.org/10.1371/journal.pone.0106380

Bradley, K. M., Breyer, J. P., Melville, D. B., Broman, K. W., Knapik, E. W., & Smith, J. R. (2011). An SNP-based linkage map for Zebrafish reveals sex determination Loci. *G3: Genes, Genomes, Genetics*, *1*, 3–9. https://doi.org/10.1534/g3.111.000190

Broquet, T., & Petit, E. J. (2009). Molecular estimation of dispersal for ecology and population genetics. *Annual Review of Ecology, Evolution, and Systematics*, *40*, 193–216. https://doi.org/10.1146/annurev.ecolsys.110308.120324

Browning, S. R., & Browning, B. L. (2011). Haplotype phasing: Existing methods and new developments. *Nature Reviews Genetics*, *12*, 703–714. https://doi.org/10.1038/nrg3054

Castilho, R., & Ciftci, Y. (2005). Genetic differentiation between close eastern Mediterranean *Dicentrarchus labrax* (L.) populations. *Journal of Fish Biology*, *67*, 1746–1752. https://doi.org/10.1111/j.1095-8649.2005.00869.x

Cayuela, H., Rougemont, Q., Prunier, J. G., Moore, J.-S., Clobert, J., Besnard, A., & Bernatchez, L. (2018). Demographic and genetic approaches to study dispersal in wild animal populations: A methodological review. *Molecular Ecology*, *27*(20), 3976–4010. https://doi.org/10.1111/mec.14848

Corbett-Detig, R., & Nielsen, R. (2017). A hidden Markov model approach for simultaneously estimating local ancestry and admixture time using next generation sequence data in samples of arbitrary ploidy. *PLOS Genetics*, *13*, e1006529. https://doi.org/10.1371/journal.pgen.1006529

de Pontual, H., Lalire, M., Fablet, R., Laspougeas, C., Garren, F., Martin, S., … Woillez, M. (2019). New insights into behavioural ecology of European seabass off the West Coast of France: Implications at local and population scales. *ICES Journal of Marine Science*, *76*(2), 501–515. https://doi.org/10.1093/icesjms/fsy086

Dias-Alves, T., Mairal, J., & Blum, M. G. B. (2018). Loter: A software package to infer local ancestry for a wide range of species. *Molecular Biology and Evolution*, *35*, 2318–2326. https://doi.org/10.1093/molbev/msy126

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*, 2518. https://doi.org/10.1038/s41467-018-04963-6

Furrer, R. D., & Pasinelli, G. (2016). Empirical evidence for source–sink populations: A review on occurrence, assessments and implications. *Biological Reviews*, *91*, 782–795. https://doi.org/10.1111/brv.12195

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

Geza, E., Mugo, J., Mulder, N. J., Wonkam, A., Chimusa, E. R., & Mazandu, G. K. (2018). A comprehensive survey of models for dissecting local ancestry deconvolution in human genome. *Briefings in Bioinformatics*, *6*(29). https://doi.org/10.1093/bib/bby044

Gravel, S. (2012). Population Genetics Models of Local Ancestry. *Genetics*, *191*, 607–619. https://doi.org/10.1534/genetics.112.139808

Guan, Y. (2014). Detecting Structure of Haplotypes and Local Ancestry. *Genetics*, *196*, 625. https://doi.org/10.1534/genetics.113.160697

Guo, B., Li, Z., & Merilä, J. (2016). Population genomic evidence for adaptive differentiation in the Baltic Sea herring. *Molecular Ecology*, *25*, 2833–2852. https://doi.org/10.1111/mec.13657

Haenel, Q., Laurentino, T. G., Roesti, M., & Berner, D. (2018). Meta-analysis of chromosome-scale crossover rate variation in eukaryotes and its significance to evolutionary genomics. *Molecular Ecology*, *27*, 2477–2497. https://doi.org/10.1111/mec.14699

Hanski, I. (1998). Metapopulation dynamics. *Nature*, *396*, 41–49. https://doi.org/10.1038/23876

Hemmer-Hansen, J., Nielsen, E. E., Therkildsen, N. O., Taylor, M. I., Ogden, R., Geffen, A. J., … Carvalho, G. R. (2013). A genomic island linked to ecotype divergence in Atlantic cod. *Molecular Ecology*, *22*, 2653–2667. https://doi.org/10.1111/mec.12284

Karlsen, B. O., Klingan, K., Emblem, Å., Jørgensen, T. E., Jueterbock, A., Furmanek, T., … Moum, T. (2013). Genomic divergence between the migratory and stationary ecotypes of Atlantic cod. *Molecular Ecology*, *22*, 5098–5111. https://doi.org/10.1111/mec.12454

Kelleher, J., Etheridge, A. M., & McVean, G. (2016). Efficient Coalescent Simulation and Genealogical Analysis for Large Sample Sizes. *PLoS Computational Biology*, *12*, e1004842. https://doi.org/10.1371/journal.pcbi.1004842

Lamichhaney, S., Barrio, A. M., Rafati, N., Sundstrom, G., Rubin, C.-J., Gilbert, E. R., … Andersson, L. (2012). Population-scale sequencing reveals genetic differentiation due to local adaptation in Atlantic herring. *PNAS*, *109*, 19345–19350. https://doi.org/10.1073/pnas.1216128109

Lawson, D. J., Hellenthal, G., Myers, S., & Falush, D. (2012). Inference of Population Structure using Dense Haplotype Data. *PLoS Genetics*, *8*, https://doi.org/10.1371/journal.pgen.1002453

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*, 3187–3202. https://doi.org/10.1111/mec.13627

Leitwein, M., Gagnaire, P.-A., Desmarais, E., Berrebi, P., & Guinand, B. (2018). Genomic consequences of a recent three-way admixture in supplemented wild brown trout populations revealed by local ancestry tracts. *Molecular Ecology*, *27*, 3466–3483. https://doi.org/10.1111/mec.14816

Liang, M., & Nielsen, R. (2014). The Lengths of Admixture Tracts. *Genetics*, *197*, 953–967. https://doi.org/10.1534/genetics.114.162362

Limborg, M. T., Helyar, S. J., DeBruyn, M., Taylor, M. I., Nielsen, E. E., Ogden, R., … Bekkevold, D. (2012). Environmental selection on transcriptome-derived SNPs in a high gene flow marine fish, the Atlantic herring (*Clupea harengus*). *Molecular Ecology*, *21*, 3686–3703. https://doi.org/10.1111/j.1365-294X.2012.05639.x

Lowe, W. H. (2003). Linking Dispersal to Local Population Dynamics: A Case Study Using a Headwater Salamander System. *Ecology*, *84*, 2145–2154. https://doi.org/10.1890/0012-9658(2003)084[2145:LDTLPD]2.0.CO;2

Lowe, W. H., & Allendorf, F. W. (2010). What can genetics tell us about population connectivity? *Molecular Ecology*, *19*, 3038–3051. https://doi.org/10.1111/j.1365-294X.2010.04688.x

Manel, S., Loiseau, N., Andrello, M., Fietz, K., Goñi, R., Forcada, A., … Mouillot, D. (2019). Long-Distance Benefits of Marine Reserves: Myth or Reality? *Trends in Ecology & Evolution*, *34*, 342–354. https://doi.org/10.1016/j.tree.2019.01.002

Maples, B. K., Gravel, S., Kenny, E. E., & Bustamante, C. D. (2013). RFMix: A Discriminative Modeling Approach for Rapid and Robust Local-Ancestry Inference. *The American Journal of Human Genetics*, *93*, 278–288. https://doi.org/10.1016/j.ajhg.2013.06.020

Martinez, B. A., Lamichhaney, S., Fan, G., Rafati, N., Pettersson, M., Zhang, H. … Andersson, L. (2016). The genetic basis for ecological adaptation of the Atlantic herring revealed by genome sequencing. *eLife*, *5*, p. e12081. https://doi.org/10.7554/eLife.12081.

Medina, P., Thornlow, B., Nielsen, R., & Corbett-Detig, R. (2018) *Estimating the timing of multiple admixture pulses during local ancestry inference*. bioRxiv 314617. https://doi.org/10.1101/314617

Milano, I., Babbucci, M., Cariani, A., Atanassova, M., Bekkevold, D., Carvalho, G. R., … Bargelloni, L. (2014). Outlier SNP markers reveal fine-scale genetic structuring across European hake populations (*Merluccius merluccius*). *Molecular Ecology*, *23*, 118–135. https://doi.org/10.1111/mec.12568

Nanninga, G. B., & Manica, A. (2018). Larval swimming capacities affect genetic differentiation and range size in demersal marine fishes. *Marine Ecology Progress Series*, *589*, 1–12. https://doi.org/10.3354/meps12515

Ni, X., Yang, X., Guo, W., Yuan, K., Zhou, Y., Ma, Z., & Xu, S. (2016). Length distribution of Ancestral tracks under a general admixture model and its applications in population history inference. *Scientific Reports*, *6*, 20048. https://doi.org/10.1038/srep20048

Nielsen, E. E., Cariani, A., Aoidh, E. M., Maes, G. E., Milano, I., Ogden, R., … Carvalho, G. R. (2012). Gene-associated markers provide tools for tackling illegal fishing and false eco-certification. *Nature Communications*, *3*, 851. https://doi.org/10.1038/ncomms1845

Nielsen, E. E., Nielsen, P. H., Meldrup, D., & Hansen, M. M. (2004). Genetic population structure of turbot (*Scophthalmus maximus* L.) supports the presence of multiple hybrid zones for marine fishes in the transition zone between the Baltic Sea and the North Sea. *Molecular Ecology*, *13*, 585–595. https://doi.org/10.1046/j.1365-294X.2004.02097.x

Palamara, P. F., & Pe'er, I. (2013). Inference of historical migration rates via haplotype sharing. *Bioinformatics*, *29*, i180–i188. https://doi.org/10.1093/bioinformatics/btt239

Palumbi, S. R. (2003). Population genetics, demographic connectivity, and the design of marine reserves. *Ecological Applications*, *13*, S146–S158.

Pante, E., & Simon-Bouhet, B. (2013). marmap: A package for importing, plotting and analyzing bathymetric and topographic data in R. *PLoS ONE*, *8*, e73051. https://doi.org/10.1371/journal.pone.0073051

Pascual, M., Rives, B., Schunter, C., & Macpherson, E. (2017). Impact of life history traits on gene flow: A multispecies systematic review across oceanographic barriers in the Mediterranean Sea. *PLoS ONE*, *12*, e0176419. https://doi.org/10.1371/journal.pone.0176419

Payseur, B. A., & Rieseberg, L. H. (2016). A genomic perspective on hybridization and speciation. *Molecular Ecology*, *25*, 2337–2360. https://doi.org/10.1111/mec.13557

Pickett, G. D., & Pawson, M. G. (1994). *Sea Bass: Biology*. London: Springer Science & Business Media.

Pinsky, M. L., Montes, H. R. Jr., & Palumbi, S. R. (2010). Using Isolation by Distance and Effective Density to Estimate Dispersal Scales in Anemonefish. *Evolution*, *64*, 2688–2700. https://doi.org/10.1111/j.1558-5646.2010.01003.x

Pinsky, M. L., Saenz-Agudelo, P., Salles, O. C., Almany, G. R., Bode, M., Berumen, M. L., … Planes, S. (2017). Marine dispersal scales are congruent over evolutionary and ecological time. *Current Biology*, *27*, 149–154. https://doi.org/10.1016/j.cub.2016.10.053

Pool, J. E., & Nielsen, R. (2009). Inference of historical changes in migration rate from the lengths of migrant tracts. *Genetics*, *181*, 711–719. https://doi.org/10.1534/genetics.108.098095

Puebla, O., Bermingham, E., & McMillan, W. O. (2012). On the spatial scale of dispersal in coral reef fishes. *Molecular Ecology*, *21*, 5675–5688. https://doi.org/10.1111/j.1365-294X.2012.05734.x

Pulliam, H. R. (1988). Sources, sinks, and population regulation. *The American Naturalist*, *132*, 652–661. https://doi.org/10.1086/284880

Quéré, N., Desmarais, E., Tsigenopoulos, C. S., Belkhir, K., Bonhomme, F.,Guinand, B. (2012). Gene flow at major transitional areas in sea bass (*Dicentrarchus labrax*) and the possible emergence of a hybrid swarm. *Ecology and Evolution*, *2*, 3061–3078. https://doi.org/10.1002/ece3.406

Racimo, F., Sankararaman, S., Nielsen, R., & Huerta-Sánchez, E. (2015). Evidence for archaic adaptive introgression in humans. *Nature Reviews Genetics*, *16*, 359–371. https://doi.org/10.1038/nrg3936

Reiss, H., Hoarau, G., Dickey-Collas, M., & Wolff, W. J. (2009). Genetic population structure of marine fish: Mismatch between biological and fisheries management units. *Fish and Fisheries*, *10*, 361–395. https://doi.org/10.1111/j.1467-2979.2008.00324.x

Rhee, J.-K., Li, H., Joung, J.-G., Hwang, K.-B., Zhang, B.-T., & Shin, S.-Y. (2016). Survey of computational haplotype determination methods for single individual. *Genes & Genomics*, *38*, 1–12. https://doi.org/10.1007/s13258-015-0342-x

Ringbauer, H., Coop, G., & Barton, N. H. (2017). Inferring recent demography from isolation by distance of long shared sequence blocks. *Genetics*, *205*(3), 1335–1351. https://doi.org/10.1534/genetics.116.196220

Roesti, M., Hendry, A. P., Salzburger, W., & Berner, D. (2012). Genome divergence during evolutionary diversification as revealed in replicate lake–stream stickleback population pairs. *Molecular Ecology*, *21*, 2852–2862. https://doi.org/10.1111/j.1365-294X.2012.05509.x

Rosenzweig, B. K., Pease, J. B., Besansky, N. J., & Hahn, M. W. (2016). Powerful methods for detecting introgressed regions from population genomic data. *Molecular Ecology*, *25*, 2387–2397. https://doi.org/10.1111/mec.13610

Rousset, F. (1997). Genetic differentiation and estimation of gene flow from F-statistics under isolation by distance. *Genetics*, *145*, 1219–1228.

Runge, J. P., Runge, M. C., & Nichols, J. D. (2006). The role of local populations within a landscape context: Defining and classifying sources and sinks. *The American Naturalist*, *167*, 925–938. https://doi.org/10.1086/503531

Salter-Townshend, M., & Myers, S. (2018) *Fine-scale Inference of Ancestry Segments without Prior Knowledge of Admixing Groups*. bioRxiv 376137. https://doi.org/10.1101/376137

Selkoe, K. A., D'Aloia, C. C., Crandall, E. D., Iacchei, M., Liggins, L., Puritz, J. B., … Toonen, R. J. (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

Selkoe, K. A., & Toonen, R. J. (2011). Marine connectivity: A new look at pelagic larval duration and genetic metrics of dispersal. *Marine Ecology Progress Series*, *436*, 291–305. https://doi.org/10.3354/meps09238

Skov, L., Hui, R., Shchur, V., Hobolth, A., Scally, A., Schierup, M. H., & Durbin, R. (2018). Detecting archaic introgression using an unadmixed outgroup. *PLOS Genetics*, *14*, e1007641. https://doi.org/10.1371/journal.pgen.1007641

Snyder, M. W., Adey, A., Kitzman, J. O., & Shendure, J. (2015). Haplotype-resolved genome sequencing: Experimental methods and applications. *Nature Reviews Genetics*, *16*, 344–358. https://doi.org/10.1038/nrg3903

Sodeland, M., Jorde, P. E., Lien, S., Jentoft, S., Berg, P. R., Grove, H., … Knutsen, H. (2016). "Islands of Divergence" in the Atlantic Cod genome represent polymorphic chromosomal rearrangements. *Genome Biology and Evolution*, *8*, 1012–1022. https://doi.org/10.1093/gbe/evw057

Souche, E. L., Hellemans, B., Babbucci, M., MacAoidh, E., Guinand, B.,Bargelloni, L., & Volckaert, F. (2015). Range-wide population structure of European sea bass *Dicentrarchus labrax*. *Biological Journal of the Linnean Society*, *116*, 86–105. https://doi.org/10.1111/bij.12572

Tine, M., Kuhl, H., Gagnaire, P.-A., Louro, B., Desmarais, E., Martins, R. S. T., … Reinhardt, R. (2014). European sea bass genome and its variation provide insights into adaptation to euryhalinity and speciation. *Nature Communications*, *5*, 5770. https://doi.org/10.1038/ncomms6770

Vandamme, S. G., Maes, G. E., Raeymaekers, J. A., Cottenie, K., Imsland, A. K., Hellemans, B., … Volckaert, F. A. (2014). Regional environmental pressure influences population differentiation in turbot (*Scophthalmus maximus*). *Molecular Ecology*, *23*, 618–636. https://doi.org/10.1111/mec.12628

Yuan, K., Zhou, Y., Ni, X., Wang, Y., Liu, C., & Xu, S. (2017). Models, methods and tools for ancestry inference and admixture analysis. *Quant Biol*, *5*, 236–250. https://doi.org/10.1007/s40484-017-0117-2

How to cite this article: Duranton M, Bonhomme F, Gagnaire P-A. The spatial scale of dispersal revealed by admixture tracts. *Evol Appl*. 2019;00:1–14. https://doi.org/10.1111/eva.12829

DISCUSSION

I. Bilan des résultats

Au cours de cette thèse, j'ai cherché à comprendre quels mécanismes évolutifs permettent la mise en place de l'isolement reproductif et donc l'aboutissement de la spéciation. Pour cela, j'ai étudié le bar européen (*Dicentrarchus labrax*), une espèce de poisson marin subdivisée en deux lignées évolutives représentées par la population atlantique (bar) et méditerranéenne (loup) (Lemaire *et al.* 2005). Au commencement de ma thèse, l'avancée des travaux sur cette espèce avait mis en lumière l'existence d'un contact secondaire entre ces deux lignées, en lien avec les variations climatiques du Pléistocène (Tine *et al.* 2014). L'existence d'îlots génomiques de différenciation avait également été mise en évidence, probablement due à l'existence d'une barrière semi-perméable aux échanges génétiques entre les deux populations (Tine *et al.* 2014). Mon premier objectif a donc été d'identifier les mécanismes ayant permis la formation de ces îlots, afin de distinguer ceux liés à l'isolement reproductif de ceux qui ne le seraient pas (**Chapitre 1**). L'utilisation de génomes entièrement séquencés et phasés m'a alors permis d'accéder à une couche d'information supplémentaire à celle des fréquences alléliques, celle de la liaison génétique. J'ai ainsi pu identifier dans les génomes méditerranéens les haplotypes atlantiques introgressés, ce qui m'a permis d'avoir une mesure directe du flux génique. J'ai ainsi montré qu'il existe chez le bar européen deux types d'îlots génomiques de différenciation, ceux impliqués dans l'isolement reproductif (également appelés îlots de spéciation (Turner *et al.* 2005)) et ceux qui ne le sont pas (« *incidental islands* » en anglais (Cruickshank and Hahn 2014)). Ces îlots ont été créés par deux mécanismes distincts ayant agi à deux périodes différentes de la divergence, mais facilitant tous les deux la mise en place des îlots dans les régions à faible taux de recombinaison. Pendant la période d'allopatrie, l'action de la sélection en liaison réduit plus fortement la diversité génétique intra-populationnelle des régions à faible taux de recombinaison en y accélérant le tri du polymorphisme ancestral, faisant ainsi artificiellement monter le niveau de différenciation génétique mesuré à l'aide de statistiques de divergence relatives (i.e. sensibles au niveau de polymorphisme) comme le F_{ST} (Noor and Bennett 2009; Cruickshank and Hahn 2014). Les populations peuvent également, lors de la phase allopatrique, fixer des allèles impliqués dans des incompatibilités génétiques ou dans l'adaptation locale, qui vont générer de l'isolement reproductif. Au moment du contact, le flux génique va venir éroder la différenciation génétique précédemment accumulée, sauf au niveau des locus d'isolement reproductif qui agissent comme des barrières à l'introgression (Bierne *et al.* 2013), accentuant l'hétérogénéité du paysage de différenciation déjà modelé par la différenciation la sélection en liaison pendant en allopatrie. Les allèles d'isolement reproductif cumulant plus facilement leurs effets dans les régions à faible taux de recombinaison, c'est dans ces régions qu'ils résistent le mieux à l'introgression (Yeaman *et al.* 2016). On retrouve donc bien deux types d'îlots, ceux n'étant pas impliqués dans l'isolement reproductif qui sont en cours d'érosion et

ceux impliqués dans l'isolement qui se maintiennent face au flux génique. Ces résultats soulignent donc l'importance du rôle du paysage génomique de recombinaison dans la mise en place de l'isolement reproductif. De plus, la présence de nombreux îlots génomiques résistants à l'introgression suggère que l'isolement reproductif entre le bar et le loup à des bases polygéniques. Il existe en outre une gradation dans le niveau de résistance à l'introgression de ces différents îlots génomiques de spéciation, reflétant possiblement des effets individuels plus ou moins important sur l'isolement reproductif.

Un autre résultat marquant de ce premier chapitre de thèse a été l'observation dans les îlots génomiques de spéciation d'un niveau de divergence nucléotidique plus élevé qu'attendu sous un modèle neutre. Cet excès de divergence entre les séquences atlantiques et méditerranéennes s'explique par la présence d'allèles anciens, ayant initié leur divergence avant que les deux lignées de bar européen n'aient commencé leur phase de divergence allopatrique. L'objectif principale du **Chapitre 2** a donc été d'identifier l'origine de ces allèles anciens. Une hypothèse pouvant expliquer leur origine est qu'ils aient introgressés lors d'un contact avec une troisième lignée (Green *et al.* 2010; Meyer *et al.* 2012). J'ai pu tester spécifiquement cette hypothèse à l'aide du séquençage de nouveaux génomes de bar européen et de celui du bar moucheté (*Dicentrarchus punctatus*), seule espèce phylogénétiquement proche vivant en sympatrie partielle avec le bar européen. Dans un premier temps, j'ai délimité précisément les régions impliquées dans l'isolement reproductif entre le bar et le loup en utilisant l'information de la différenciation génétique et de l'intensité du flux génique. J'ai ensuite utilisé trois méthodes différentes pour détecter des traces d'échanges génétiques entre les deux espèces de bar. J'ai ainsi pu montrer que des échanges ont eu lieu entre la lignée atlantique de bar européen et le bar moucheté. Ces échanges ont permis l'introgression d'allèles d'origine *D. punctatus* tout le long du génome de *D. labrax*. Cependant, ces allèles sont globalement présents à faible fréquence (<5%) alors qu'ils sont fixés dans les régions impliquées dans l'isolement reproductif. Il semblerait donc que l'introgression d'allèles de *D. punctatus* (datée à il y a environ 80 000 ans à partir de la taille des haplotypes introgressés) ait permis d'accélérer la mise en place de l'isolement reproductif entre le bar et le loup. Une hypothèse permettant d'expliquer comment ces allèles ont pu se fixer en Atlantique tout en générant de l'isolement reproductif avec la Méditerranée est celle de la résolution d'incompatibilités génétiques entre *D. punctatus* et *D. labrax* (Schumer *et al.* 2015; Blanckaert and Bank 2018). On peut en effet supposer que des incompatibilités génétiques de type Dobzhanky-Mueller (DMI) existaient entre le bar européen et le bar moucheté au moment de leur contact (ceux-ci étant entre eux plus divergents que les lignées de bar européen entre elles), la fixation de telles DMI étant quasi inévitable lors de la divergence de deux lignées en allopatrie (Presgraves 2010). Or, la résolution dans une population admixée d'une incompatibilité génétique entre allèles

dérivés à deux locus passe nécessairement par la fixation d'un allèle parental à l'un des deux locus. En fonction des conditions (e.g. proportions lors du mélange, relations de dominance et coefficients de sélection des allèles, …), la probabilité de fixation des allèles parentaux originaires de chacune des deux lignées peut être fortement différente. Par exemple, dans une population d'ascendance majoritaire *D. labrax* (e.g. 95%) et minoritaire *D. punctatus* (e.g. 5%), on peut s'attendre à ce que la majorité des conflits soient résolu par la fixation de l'allèle d'origine *D. labrax*. Toutefois, on peut dans tous les cas s'attendre à ce que certaines résolutions d'incompatibilités entre paires de locus se soient produites par fixation de l'allèle issu de *D. punctatus*. Dans ces cas-là, si l'incompatibilité résolue entre *D. punctatus* et la lignée atlantique de *D. labrax* n'existe plus, elle se retrouve en revanche transférée entre les lignées atlantique et méditerranéenne de *D. labrax*. Ainsi, ces résultats tendent à montrer que l'introgression (ici un épisode ancien d'admixture) avec une espèce proche a favorisé la mise en place de l'isolement reproductif, comme cela a déjà été proposé chez d'autres espèces (Runemark *et al.* 2018; Schumer *et al.* 2018; Eberlein *et al.* 2019).

Maintenant que l'origine des allèles impliqués dans l'isolement reproductif est plus claire, au moins pour une partie d'entre eux, des questions perdurent notamment sur la nature des gènes impliqués et les pressions de sélection sous lesquelles ils évoluent. J'ai donc cherché à répondre à ces questions dans le **Chapitre 3,** en étudiant les signatures d'évolution moléculaire des gènes au cœur des îlots de spéciation. Cependant, comme nous l'avons vu dans le **Chapitre 1**, la plupart des gènes impliqués dans l'isolement reproductif sont localisés dans les régions à faible taux de recombinaison. Or, il est connu que les patrons d'évolution des gènes sont largement influencés par la recombinaison, notamment à cause des effets de sélection en liaison (Campos *et al.* 2014). J'ai donc dans un premier temps, réestimé le taux de recombinaison populationnel le long du génome du bar européen afin de montrer qu'il existe effectivement une corrélation entre le niveau de recombinaison et les différents paramètres d'évolution moléculaire estimés (dN, dS, πN, πS, ωa, ωna). Afin que la comparaison des patrons d'évolution moléculaire des gènes impliqués et non-impliqués dans l'isolement reproductif reflète des différences de pression de sélection et non des différences de taux de recombinaison, j'ai étudié un sous-ensemble de gènes non-impliqués dans l'isolement reproductif ayant une distribution de taux de recombinaison similaire à celle des gènes impliqués dans l'IR. J'ai ainsi pu montrer que les gènes impliqués dans l'isolement reproductif subissent de plus fortes pressions de sélection purificatrice, sont phylogénétiquement plus conservés mais ont accumulé plus de changements adaptatifs récents dans la lignée *D. labrax*. Ces résultats suggèrent que des changements d'acides nucléiques sur ces gènes pourraient avoir de plus grandes conséquences phénotypiques, expliquant possiblement leur implication dans l'IR. Enfin, j'ai cherché si les gènes impliqués dans l'isolement reproductif présentaient une surreprésentation fonctionnelle particulière et j'ai ainsi pu montrer que des fonctions, telles que

le transport transmembranaire d'ions et des processus cognitifs (liés au comportement et à la mémoire), étaient surreprésentées parmi les gènes d'IR. Le bar européen étant une espèce euryhaline, le fait que les gènes impliqués dans l'isolement reproductif soient également impliqués dans le transport d'ions pourrait suggérer que ces allèles ont évolué sous sélection divergente en réponse à des pressions sélectives locales. Cependant, nous ne pouvons exclure l'hypothèse que les gènes responsables du transport d'ions génèrent de l'isolement reproductif parce qu'ils sont plus facilement impliqués que d'autres dans des DMI. De plus, l'isolement reproductif entre le bar et le loup étant polygénique, il est difficile de quantifier le rôle relatif potentiel de l'adaptation locale par rapport à d'autres processus que nous ne pouvions pas détecter ici sans études de surreprésentation fonctionnelle, telles que les incompatibilités génétiques.

Après avoir délimité les régions impliquées dans l'isolement reproductif et identifié les pressions de sélection auxquelles elles sont contraintes, je me suis intéressée au comportement de ces allèles dans les génomes hybrides et leur impact sur leur valeur sélective. En effet, ces allèles étant contre-sélectionnés sur le long terme, on peut s'attendre à ce qu'ils génèrent de la dépression d'hybridation. Dans le **Chapitre 4** j'ai donc comparé la valeur sélective d'individus issus de rétrocroisements entre des mâles hybrides de première génération avec des femelles ouest-méditerranéennes (*backcross-med*) et des individus ouest-méditerranéens. Pour cela des individus *backcross-med* et ouest-méditerranéens partageant les mêmes mères ont été élevés dans les mêmes conditions (en jardin commun) jusqu'à l'âge de deux ans. Je n'ai pas pu mettre en évidence l'existence de différences de taux de fécondation, de taux de survie, de croissance ou de facteur de condition entres ces deux groupes d'individus, suggérant l'absence de différences de valeur sélective. Cependant, une large composante de la valeur sélective est liée au succès de reproduction. Or, la maturité sexuelle étant atteinte à l'âge de trois ans chez le bar, il se peut que des différences de valeur sélective entre les deux groupes existent mais n'aient pas pu être mise en évidence par cette étude. Les génomes de 380 *backcross-med* ont ensuite été séquencés à l'aide d'une puce à ADN de 57000 SNPs. Je me suis alors intéressée plus particulièrement aux SNPs identifiés comme impliqués dans l'isolement reproductif et différentiellement fixés entre les géniteurs atlantiques et méditerranéens des croisements. J'ai alors pu montrer que les allèles atlantiques étaient présents en plus forte fréquence qu'attendue dans les génomes des *backcross-med*. Il semblerait donc que les allèles atlantiques soient favorisés dans les premières générations d'hybridation aux mêmes locus où ils sont contre-sélectionnés sur le long terme. Ceci pourrait s'expliquer par une dynamique temporelle complexe entre dépression d'hybridation et hétérosis. En effet, dans les premières générations d'hybridation, les allèles atlantiques étant associés entre eux sur de longs haplotypes, peuvent facilement générer de l'hétérosis locale en masquant les mutations récessives faiblement délétères présentes en Méditerranée, c'est la

superdominance associative locale (Ohta and Kimura 1970; Pamilo and Pálsson 1998). Ce mécanisme expliquerait la meilleure valeur sélective des hybrides de première génération (F1) par rapport à celle de leurs parents précédemment décrite entre le bar et le loup (Guinand *et al.* 2017). Dans les générations plus tardives, la recombinaison ayant cassé les associations alléliques (les haplotypes atlantiques deviennent plus courts), l'effet d'hétérosis diminue rapidement permettant de révéler l'effet délétère des allèles introgressés. Ceci pourrait également expliquer pourquoi la survie des individus ouest-méditerranéen, qui sont en moyenne introgressés à 30% (cf **Chapitre 1**), est plus faible que celle des individus issus des autres populations moins fortement introgressées (Guinand *et al.* 2017). Ainsi, la dynamique d'introgression est probablement plus complexe que ce qui était précédemment envisagé au démarrage de cette thèse et en lien étroit avec la recombinaison, l'effet des haplotypes introgressés semblant dépendre directement de leur longueur.

Finalement, le **Chapitre 5** m'a permis d'utiliser les données générées lors de cette thèse pour répondre à une question plus appliquée, celle de l'estimation de la distance moyenne de dispersion intergénérationnelle du bar européen en Méditerranée. En effet, la connectivité est un paramètre démographique important qui entre en jeu dans la stabilité des populations. Il est donc nécessaire de connaitre ce paramètre pour avoir une bonne gestion des populations naturelles (Runge *et al.* 2006; Furrer and Pasinelli 2016). J'ai donc utilisé les données phasées générées dans le **Chapitre 1** pour estimer la distance de dispersion parents-enfants du bar en Méditerranée. En effet, l'information haplotypique qui n'est encore que très peu utilisée dans le domaine de la conservation, peut permettre d'aborder d'anciennes questions sous un nouvel angle. Par exemple, la distribution de tailles des haplotypes introgressés n'est quasiment jamais utilisée pour répondre à des questions appliquées alors qu'elle contient énormément d'informations sur la démographie des populations, en particulier sur la migration. En effet, la longueur des haplotypes introgressés est directement liée à leur date d'introgression et au taux de recombinaison de la région dans laquelle ils se trouvent (Pool and Nielsen 2009; Racimo *et al.* 2015). Ainsi, en comparant la taille moyenne des haplotypes atlantiques introgressés dans les populations ouest- et est-méditerranéenne, j'ai pu estimer le temps nécessaire aux haplotypes pour traverser la Méditerranée et donc la distance de dispersion intergénérationnelle de l'espèce en Méditerranée. J'ai également vérifié que cette approche était valide à l'aide de simulations. Etonnamment, au vu des capacités de mouvements des individus adultes et de la durée importante de la phase larvaire chez cette espèce, la distance de dispersion estimée est relativement faible (5 et 15 km par génération). Ces résultats sont en accord avec de précédentes études ayant mis en évidence l'existence d'une structure au sein de la Méditerranée (Bahri-Sfar *et al.* 2000; Castilho and Ciftci 2005). Cette étude confirme également le potentiel des données haplotypiques pour répondre à des questions plus appliquées au domaine de la conservation.

II. Perspectives

Les principaux résultats de cette thèse auront largement contribué à modifier ma vision du processus de spéciation et plus particulièrement des mécanismes évolutifs impliqués et de leur importance.

1. L'hétérogénéité de la recombinaison

L'intensité locale de la recombinaison apparaît désormais comme un paramètre clé à considérer pour comprendre l'évolution des espèces et ce pour deux raisons principales. Premièrement, la recombinaison influence la spéciation en influant sur le positionnement des locus d'isolement reproductif dans les génomes. En effet, en présence de flux génique, les locus d'isolement reproductif localisés dans les régions à faible taux de recombinaison cumulent leurs effets et peuvent plus facilement résister à l'introgression (Nachman and Payseur 2012; Ortiz-Barrientos *et al.* 2016). C'est pourquoi des structures réduisant efficacement la recombinaison comme les inversions, sont souvent impliquées dans l'isolement reproductif (Rieseberg 2001; Noor *et al.* 2001). C'est effectivement ce qui a été observé chez le bar européen, les îlots génomiques résistant à l'introgression étant principalement localisés dans les régions à faible recombinaison. Deuxièmement, la recombinaison en modulant la force de la sélection en liaison, joue un rôle prédominant dans le modelage des patrons de différenciation génétique. En effet, si la recombinaison est hétérogène le long du génome, les régions génomiques ayant un fort taux de recombinaison subiront moins les effets de la sélection indirecte par rapport à celles avec un faible taux de recombinaison. Ainsi, la recombinaison peut en interaction avec la sélection participer à générer des paysages de différenciation génétique hétérogènes, comme c'est le cas chez le bar européen. Or, une différenciation hétérogène étant généralement interprétée comme le signe de l'action de la sélection directe rendant le flux génique hétérogène le long du génome, ne pas prendre en compte les variations du taux de recombinaison peut conduire à une mauvaise interprétation des processus évolutifs mis en jeu (Noor and Bennett 2009; Cruickshank and Hahn 2014). De plus, les variations du taux de recombinaison à large échelle étant généralement conservées entre espèces proches, la sélection en liaison peut agir sur le long terme évolutif et générer des corrélations entre les patrons de diversité et de différenciation observés chez différentes espèces (Burri *et al.* 2015; Burri 2017). La recombinaison influence également les patrons génomiques d'introgression. En effet, si l'introgression est globalement délétère alors une corrélation positive est attendue entre recombinaison et introgression (Schumer *et al.* 2018). C'est pourquoi, dans les régions où la recombinaison est forte, les mutations neutres introgressées peuvent plus facilement se séparer des délétères et se maintenir dans le nouveau fond génétique. Un domaine en particulier où l'hétérogénéité de la recombinaison peut compliquer l'interprétation des patrons est l'inférence de l'histoire démographique des espèces (voir Annexe 5). En effet, la plupart des méthodes d'inférence supposent que les génomes évoluent complètement neutralement alors que la sélection

en liaison est omniprésente (Kern and Hahn 2018). Seule quelques études ont essayé d'intégrer l'effet de la sélection en liaison dans les inférences démographiques en implémentant des variations de taille efficace le long du génome (Roux *et al.* 2016; Rougeux *et al.* 2017; Rougemont and Bernatchez 2018). Il parait donc essentiel de considérer l'impact de la recombinaison pour identifier les processus évolutifs impliqués dans la spéciation

2. Rôle de l'hybridation dans la spéciation

Comme pour la lignée atlantique de bar européen et le bar moucheté, plusieurs études ont mis en évidence l'existence d'échanges génétiques anciens entre lignées en cours de divergence (Green *et al.* 2010; Meyer *et al.* 2012; Barlow *et al.* 2018). Ces études montrent bien que la spéciation étant un processus long et graduel, il existe de nombreuses possibilités d'hybridation entre lignées avant que l'isolement reproductif ne soit complet et les échanges génétiques impossibles. La spéciation apparait donc comme un processus réticulé pouvant faire intervenir (au moins au début de la divergence) deux ou plusieurs lignées proches. La spéciation pourrait alors se dérouler avec une alternance de périodes avec et sans flux génique entre ces différentes lignées. Ainsi, se focaliser sur deux lignées divergentes pour comprendre comment la spéciation se déroule sans considérer l'existence possibles d'interactions avec une troisième lignée (actuelle voire éteinte), pourrait conduire à une mauvaise identification des processus évolutifs impliqués. De plus en plus de méthodes sont développées pour inférer l'histoire démographique incluant des échanges génétiques entre plus que deux populations (Gravel *et al.* 2013; Hellenthal *et al.* 2014; Pugach *et al.* 2016; Ni *et al.* 2018a; b). Utiliser ce genre d'approches chez le bar européen en incluant le bar moucheté pourrait permettre d'avoir une meilleure image de l'histoire de la divergence entre ces lignées. Il semblerait également que ces anciens échanges génétiques aient favorisé la mise en place de l'isolement reproductif entre la lignée atlantique et méditerranéenne de bar européen. D'autres études ont également mis en avant le rôle de l'hybridation dans l'isolement reproductif (Runemark *et al.* 2018; Schumer *et al.* 2018; Eberlein *et al.* 2019). Or, le flux génique est généralement vu comme un processus permettant d'homogénéiser les fonds génétiques et donc s'opposant à la spéciation. En réalité, l'hybridation pourrait permettre le transfert d'incompatibilités génétiques entre lignées évolutives (Schumer *et al.* 2015; Blanckaert and Bank 2018) et ainsi accélérer la mise en place de l'isolement reproductif entre lignées en cours de divergence.

3. Qu'est-ce qu'un locus d'isolement reproductif ?

La spéciation est généralement définie comme un mécanisme permettant l'accumulation de locus d'isolement reproductif entre deux lignées en cours de divergence. En effet, pour qu'il y ait spéciation il faut que les fonds génétiques de deux lignées se différencient et que cette différenciation puisse se maintenir face à un éventuel flux génique. Une façon simple d'empêcher la ré-homogénéisation des

fonds génétiques est d'empêcher la reproduction entre individus issus des deux lignées. Or il n'y a pas de problèmes de reproduction entre le bar et le loup étant donné que les hybrides de première génération ainsi que les individus issus de rétrocroisement survivent bien. Pourtant, la différenciation génétique arrive à se maintenir puisqu'elle est encore visible aujourd'hui malgré plus de 10 000 ans d'échanges génétiques. La dynamique de sélection sur les haplotypes introgressés semble en réalité plus complexe. Dans les premières générations, les haplotypes introgressés sont favorisés car étant longs ils génèrent plus facilement de l'hétérosis par superdominance associative locale (Ohta and Kimura 1970). Dans les générations plus tardives, la recombinaison ayant cassé les associations alléliques, la contre-sélection des allèles migrants peut agir et leur introgression est bloquée en raison de leur effet délétère qui prend progressivement le dessus sur l'hétérosis atténuée qu'ils peuvent générer. Ainsi, la sélection contre les allèles migrants n'agissant que dans les générations d'hybridation plus tardives, peut-on vraiment parler d'isolement reproductif ?

Des questions se posent alors sur la nature des mutations capables de générer ce type de patrons. Par exemple, des mutations inconditionnellement délétères fixées en Atlantique ne devraient pas pouvoir introgresser en Méditerranée et devraient par conséquent générer des patrons similaires à ceux observés. Ces mutations diminuent effectivement la valeur sélective des hybrides en générant un fardeau d'hybridation mais peuvent-elles se maintenir durablement en Atlantique face au flux génique permettant l'entrée des « bons allèles » Méditerranéens ? Si les mutations impliquées dans le fardeau d'hybridation n'ont un effet barrière que dans une seule direction, peut-on dire qu'elles participent réellement à la spéciation ? Cette caractéristique est pourtant également rencontrée chez les mutations conditionnellement neutres impliquées dans l'adaptation locale et les DMIs. Cependant, chez ces dernières, l'effet contexte-dépendant (de l'environnement pour les premières et du fond génomique pour les secondes) en présence de nombreux locus générant des barrières asymétriques de direction variables peut générer une barrière efficace au flux génique. Qu'en est-il des mutations délétères dont l'effet est inconditionnel ? Si leur niveau de liaison génétique avec des mutations résistantes à l'introgression est suffisant, elles sont protégées de l'introgression par ces dernières et doivent pouvoir se maintenir dans la population. Dans ce cas-là, elles peuvent contribuer effectivement à renforcer la barrière au flux génique générée par des incompatibilités génétiques. Par exemple, on peut considérer le cas d'un allèle atlantique incompatible avec le fond génétique méditerranéen, qui étant lié à plusieurs mutations inconditionnellement délétères verrait son niveau de contre-sélection accru une fois introgressé en Méditerranée. Est-ce suffisant pour dire que ces mutations responsables du fardeau d'hybridation contribuent à la spéciation ? Ceci revient à se demander plus généralement quelle est la proportion relative de gènes résistant à l'introgression car contribuant au fardeau d'hybridation par rapport à ceux participant directement à la barrière au flux

génique (DMIs et adaptations locales) ? Cette proportion génère-t-elle une barrière au flux génique suffisamment forte pour permettre à la différenciation de continuer à s'accumuler et ainsi avancer vers une spéciation entre les deux lignées, ou les fonds génétiques vont-ils finir par s'homogénéiser ? Néanmoins, la question qui se pose plus largement est : quel type de locus peut-on considérer comme étant réellement impliqués dans la spéciation ?

Références

Bahri-Sfar L., C. Lemaire, O. K. B. Hassine, and F. Bonhomme, 2000 Fragmentation of sea bass populations in the western and eastern Mediterranean as revealed by microsatellite polymorphism. Proceedings of the Royal Society of London B: Biological Sciences 267: 929–935. https://doi.org/10.1098/rspb.2000.1092

Barlow A., J. A. Cahill, S. Hartmann, C. Theunert, G. Xenikoudakis, *et al.*, 2018 Partial genomic survival of cave bears in living brown bears. Nature Ecology & Evolution 2: 1563–1570. https://doi.org/10.1038/s41559-018-0654-8

Bierne N., P.-A. Gagnaire, and P. David, 2013 the geography of introgression in a patchy environment and the thorn in the side of ecological speciation. Current Zoology 59: 72–86.

Blanckaert A., and C. Bank, 2018 In search of the Goldilocks zone for hybrid speciation. PLOS Genetics 14: e1007613. https://doi.org/10.1371/journal.pgen.1007613

Burri R., A. Nater, T. Kawakami, C. F. Mugal, P. I. Olason, *et al.*, 2015 Linked selection and recombination rate variation drive the evolution of the genomic landscape of differentiation across the speciation continuum of Ficedula flycatchers. Genome Res. 25: 1656–1665. https://doi.org/10.1101/gr.196485.115

Burri R., 2017 Interpreting differentiation landscapes in the light of long-term linked selection. Evolution Letters. https://doi.org/10.1002/evl3.14

Campos J. L., D. L. Halligan, P. R. Haddrill, and B. Charlesworth, 2014 The Relation between Recombination Rate and Patterns of Molecular Evolution and Variation in Drosophila melanogaster. Mol Biol Evol 31: 1010–1028. https://doi.org/10.1093/molbev/msu056

Castilho R., and Y. Ciftci, 2005 Genetic differentiation between close eastern Mediterranean Dicentrarchus labrax (L.) populations. Journal of Fish Biology 67: 1746–1752. https://doi.org/10.1111/j.1095-8649.2005.00869.x

Cruickshank T. E., and M. W. Hahn, 2014 Reanalysis suggests that genomic islands of speciation are due to reduced diversity, not reduced gene flow. Mol Ecol 23: 3133–3157. https://doi.org/10.1111/mec.12796

Eberlein C., M. Hénault, A. Fijarczyk, G. Charron, M. Bouvier, *et al.*, 2019 Hybridization is a recurrent evolutionary stimulus in wild yeast speciation. Nature Communications 10: 923. https://doi.org/10.1038/s41467-019-08809-7

Furrer R. D., and G. Pasinelli, 2016 Empirical evidence for source–sink populations: a review on occurrence, assessments and implications. Biol Rev 91: 782–795. https://doi.org/10.1111/brv.12195

Gravel S., F. Zakharia, A. Moreno-Estrada, J. K. Byrnes, M. Muzzio, *et al.*, 2013 Reconstructing Native American Migrations from Whole-Genome and Whole-Exome Data. PLOS Genetics 9: e1004023. https://doi.org/10.1371/journal.pgen.1004023

Green R. E., J. Krause, A. W. Briggs, T. Maricic, U. Stenzel, *et al.*, 2010 A Draft Sequence of the Neandertal Genome. Science 328: 710–722. https://doi.org/10.1126/science.1188021

Guinand B., M. Vandeputte, M. Dupont-Nivet, A. Vergnet, P. Haffray, *et al.*, 2017 Metapopulation patterns of additive and nonadditive genetic variance in the sea bass (Dicentrarchus labrax). Ecol Evol 7: 2777–2790. https://doi.org/10.1002/ece3.2832

Hellenthal G., G. B. J. Busby, G. Band, J. F. Wilson, C. Capelli, *et al.*, 2014 A Genetic Atlas of Human Admixture History. Science 343: 747–751. https://doi.org/10.1126/science.1243518

Kern A. D., and M. W. Hahn, 2018 The Neutral Theory in Light of Natural Selection. Mol Biol Evol. https://doi.org/10.1093/molbev/msy092

Lemaire C., J.-J. Versini, and F. Bonhomme, 2005 Maintenance of genetic differentiation across a transition zone in the sea: discordance between nuclear and cytoplasmic markers. Journal of Evolutionary Biology 18: 70–80. https://doi.org/10.1111/j.1420-9101.2004.00828.x

Meyer M., M. Kircher, M.-T. Gansauge, H. Li, F. Racimo, *et al.*, 2012 A High-Coverage Genome Sequence from an Archaic Denisovan Individual. Science 338: 222–226. https://doi.org/10.1126/science.1224344

Nachman M. W., and B. A. Payseur, 2012 Recombination rate variation and speciation: theoretical predictions and empirical results from rabbits and mice. Phil. Trans. R. Soc. B 367: 409–421. https://doi.org/10.1098/rstb.2011.0249

Ni X., K. Yuan, X. Yang, Q. Feng, W. Guo, *et al.*, 2018a Inference of multiple-wave admixtures by length distribution of ancestral tracks. Heredity 121: 52. https://doi.org/10.1038/s41437-017-0041-2

Ni X., K. Yuan, C. Liu, Q. Feng, L. Tian, *et al.*, 2018b MultiWaver 2.0 : modeling discrete and continuous gene flow to reconstruct complex population admixtures. European Journal of Human Genetics 1. https://doi.org/10.1038/s41431-018-0259-3

Noor M. A. F., K. L. Grams, L. A. Bertucci, and J. Reiland, 2001 Chromosomal inversions and the reproductive isolation of species. PNAS 98: 12084–12088. https://doi.org/10.1073/pnas.221274498

Noor M. a. F., and S. M. Bennett, 2009 Islands of speciation or mirages in the desert? Examining the role of restricted recombination in maintaining species. Heredity 103: 439–444. https://doi.org/10.1038/hdy.2009.151

Ohta T., and M. Kimura, 1970 Development of associative overdominance through linkage disequilibrium in finite populations*. Genetics Research 16: 165–177. https://doi.org/10.1017/S0016672300002391

Ortiz-Barrientos D., J. Engelstädter, and L. H. Rieseberg, 2016 Recombination Rate Evolution and the Origin of Species. Trends in Ecology & Evolution 31: 226–236. https://doi.org/10.1016/j.tree.2015.12.016

Pamilo P., and S. Pálsson, 1998 Associative overdominance, heterozygosity and fitness. Heredity 81: 381–389. https://doi.org/10.1046/j.1365-2540.1998.00395.x

Pool J. E., and R. Nielsen, 2009 Inference of Historical Changes in Migration Rate From the Lengths of Migrant Tracts. Genetics 181: 711–719. https://doi.org/10.1534/genetics.108.098095

Presgraves D. C., 2010 The molecular evolutionary basis of species formation. Nature Reviews Genetics 11: 175–180. https://doi.org/10.1038/nrg2718

Pugach I., R. Matveev, V. Spitsyn, S. Makarov, I. Novgorodov, *et al.*, 2016 The Complex Admixture History and Recent Southern Origins of Siberian Populations. Mol Biol Evol 33: 1777–1795. https://doi.org/10.1093/molbev/msw055

Racimo F., S. Sankararaman, R. Nielsen, and E. Huerta-Sánchez, 2015 Evidence for archaic adaptive introgression in humans. Nat Rev Genet 16: 359–371. https://doi.org/10.1038/nrg3936

Rieseberg L. H., 2001 Chromosomal rearrangements and speciation. Trends in Ecology & Evolution 16: 351–358. https://doi.org/10.1016/S0169-5347(01)02187-5

Rougemont Q., and L. Bernatchez, 2018 The demographic history of Atlantic salmon (Salmo salar) across its distribution range reconstructed from approximate Bayesian computations*. Evolution 72: 1261–1277. https://doi.org/10.1111/evo.13486

Rougeux C., L. Bernatchez, and P.-A. Gagnaire, 2017 Modeling the Multiple Facets of Speciation-with-Gene-Flow toward Inferring the Divergence History of Lake Whitefish Species Pairs (Coregonus clupeaformis). Genome Biol Evol 9: 2057–2074. https://doi.org/10.1093/gbe/evx150

Roux C., C. Fraïsse, J. Romiguier, Y. Anciaux, N. Galtier, *et al.*, 2016 Shedding Light on the Grey Zone of Speciation along a Continuum of Genomic Divergence. PLOS Biology 14. https://doi.org/10.1371/journal.pbio.2000234

Runemark A., C. N. Trier, F. Eroukhmanoff, J. S. Hermansen, M. Matschiner, *et al.*, 2018 Variation and constraints in hybrid genome formation. Nature Ecology & Evolution 2: 549. https://doi.org/10.1038/s41559-017-0437-7

Runge J. P., M. C. Runge, and J. D. Nichols, 2006 The Role of Local Populations within a Landscape Context: Defining and Classifying Sources and Sinks. The American Naturalist 167: 925–938. https://doi.org/10.1086/503531

Schumer M., R. Cui, G. G. Rosenthal, and P. Andolfatto, 2015 Reproductive Isolation of Hybrid Populations Driven by Genetic Incompatibilities. PLOS Genetics 11: e1005041. https://doi.org/10.1371/journal.pgen.1005041

Schumer M., C. Xu, D. L. Powell, A. Durvasula, L. Skov, *et al.*, 2018 Natural selection interacts with recombination to shape the evolution of hybrid genomes. Science eaar3684. https://doi.org/10.1126/science.aar3684

Tine M., H. Kuhl, P.-A. Gagnaire, B. Louro, E. Desmarais, *et al.*, 2014 European sea bass genome and its variation provide insights into adaptation to euryhalinity and speciation. Nature Communications 5: 5770.

Turner T. L., M. W. Hahn, and S. V. Nuzhdin, 2005 Genomic Islands of Speciation in Anopheles gambiae. PLOS Biol 3. https://doi.org/10.1371/journal.pbio.0030285

Yeaman S., S. Aeschbacher, and R. Bürger, 2016 The evolution of genomic islands by increased establishment probability of linked alleles. Mol Ecol 25: 2542–2558. https://doi.org/10.1111/mec.13611

ANNEXE 1: Matériel supplémentaire de l'article:

The origin and remolding of genomic islands of differentiation in the European sea bass.

Supplementary Information

Duranton et al.

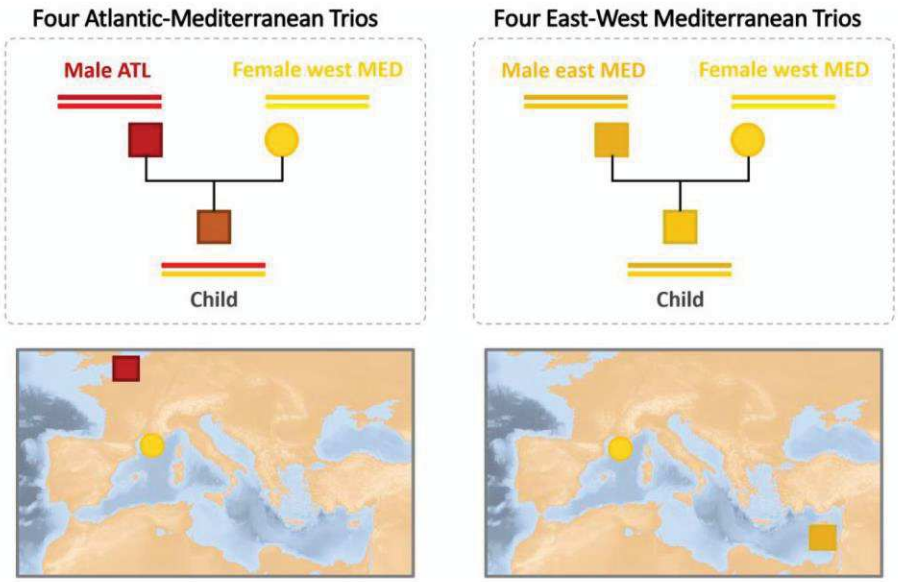
The origin and remolding of genomic islands of differentiation in the European sea bass

Supplementary Note 1: Whole genome resequencing and haplotyping

Haplotype-resolved whole genomes were obtained using a phasing-by-transmission approach. Experimental crosses were produced between wild sea bass to generate parent-offspring trios and phase parental genomes using their offspring (Supplementary Fig. 1). All experimental procedures were conducted at Ifremer's experimental aquaculture infrastructure (agreement for experiments with animals: C 34-192-6) in agreement with the French Decree n° 2013 –118 1st February 2013 NOR: AGRG1231951D (which transposes Directive 2010-63-EU into research practice for the Care and Use of Laboratory Animals). No specific agreement was required according to Directive 2010-63-EU, article 1.5 (i.e. practices not likely to cause pain, suffering, distress or lasting harm equivalent to, or higher than, that caused by the introduction of a needle in accordance with good veterinary practice are excluded from the Directive), since the fish were reared in normal conditions.

Whole genome sequencing libraries were prepared separately for each of the 24 individuals using the SPRIworks Library Preparation System (Beckman Coulter) to select 350-450bp DNA fragments. The three individual libraries from each trio were pooled and sequenced on a separate lane of an Illumina Hi-Seq 2500 platform using 2×100pb PE reads at the LIGAN-PM Genomics platform (Lille, France) (Supplementary Table 1). Variant discovery and haplotype calling was performed following the GATK version 3.3-0-g37228af best practice pipeline. Raw reads were first aligned to the sea bass reference genome[1] using BWA-mem version 0.7.5a[2] and duplicates were marked using Picard version 1.112 (http://broadinstitute.github.io/picard/index.html). The following steps were performed using GATK, starting with local realignment around indels, individual variants calling using the HaplotypeCaller, joint genotyping, genotype refinement using family priors and hard-filtering of variants to retain the most confident SNPs and indels (Filter Expression: QD<10; MQ<50; FS>7; MQRankSum<-1.5; ReadPosRankSum<-1.5). This subset of high-quality variants was then used to recalibrate base quality scores using the BQSR algorithm. The HaplotypeCaller was ran again on recalibrated sequence data to call individual variants, followed by joint genotyping, variant quality score recalibration using the VQSR algorithm with the previously identified subset of high-quality variants, and genotype refinement using family-based priors. Variants were then filtered to exclude low-quality genotypes with a GQ score <30. All trios were finally phased given parents and child genotype likelihoods and a mutation prior of 10^{-8} for *de novo* mutations using the PhaseByTransmission algorithm with default parameters. Only sites where parent/child transmission could be determined unambiguously were phased (excluding Mendelian violations and uninformative sites). In addition to phasing based on transmission information within trios, we also scored physical phasing information with the HaplotypeCaller algorithm when available (*i.e.* for closely linked variable sites located on the same read pair). For all downstream analyses, we only used parental genomes and only retained SNPs that were successfully phased using the information contained in their children's genome. Two female genomes were excluded prior to using family-based priors in the GATK pipeline since they appeared to be misidentified mothers in their trios

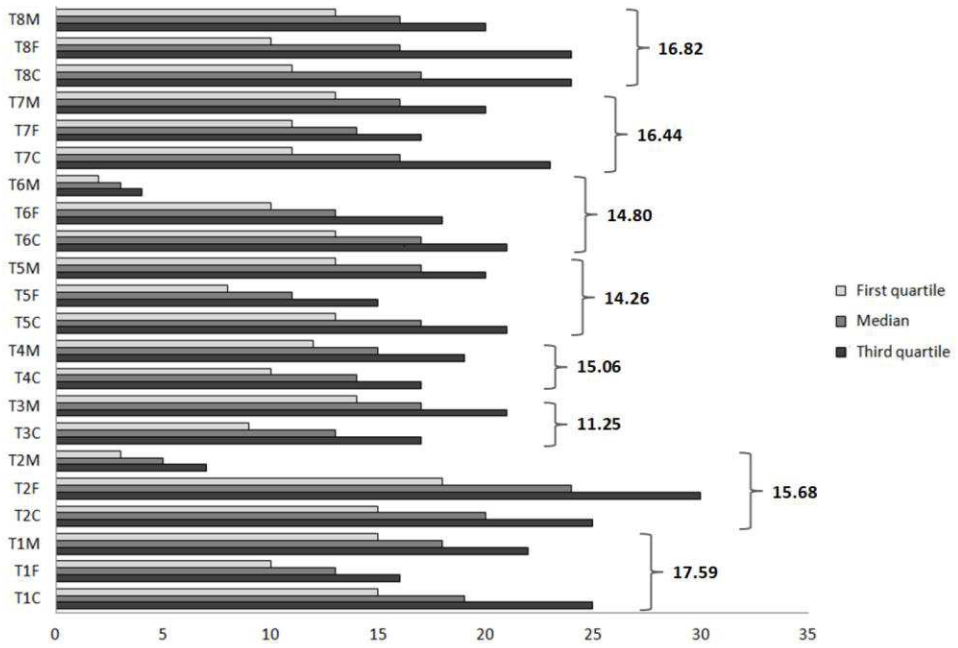
(T3F and T4F, Supplementary Table 1), which were therefore treated as father-child duos for phasing. Finally, we used VCFtools version 0.1.11[3] to only retain SNP variants (indels were removed) without missing genotype in the 14 parental genomes. Variants located on the ungrouped fraction of the genome (i.e. non-assembled scaffolds) as well as on the mitochondrial chromosome were excluded. Our final dataset consisted of 2,628,725 SNPs phased into chromosome-wide haplotypes from 4 males of Atlantic origin, 6 females from the western Mediterranean Sea, and 4 males from the eastern Mediterranean Sea. Mapping summary statistics and depth of coverage for each individual included in this study are detailed in Supplementary Table 1 and Supplementary Fig. 2.



Supplementary Figure 1 – **Schematic representation of the two different types of crossings generated to produce sea bass trios**. Four trios were obtained by crossing parents from Atlantic and Mediterranean sea bass lineages (left panel), and four trios by crossing parents from eastern and western Mediterranean populations (right panel). The geographic area of parents' origin is indicated for each type of crossing. Male parents are represented by the squares and females by circles. The red color symbolizes the Atlantic lineage, and the two Mediterranean populations are colored in light yellow (western Mediterranean) and dark yellow (eastern Mediterranean).

Supplementary Table 1 – **Summary statistics of sequencing and mapping data for each individual.**

Individual	Unpaired reads examined	Reads pairs examined	Unmapped reads	Unpaired read duplicates	Read pair duplicates	Percent duplication	Estimated library size
T1C	1081927	66809192	2616613	247413	884126	0.014964	2540848981
T1F	2668171	47343728	4479891	1249028	5158971	0.118812	201992483
T1M	1139190	61374922	2718960	166362	730774	0.01314	2611630891
T2C	3018619	75084510	5055081	940730	6420707	0.089969	414227864
T2F	1468390	84761271	3688068	352752	2419315	0.030361	1466128093
T2M	6397183	13596684	7005255	1458522	1688617	0.143962	50139501
T3C	802017	43185540	1938499	138166	378074	0.010259	2502021318
T3F	1369036	55513323	3683430	381458	1792017	0.035282	848979317
T3M	1097071	57112388	3451939	241435	758592	0.01525	2177220972
T4C	1338977	45811338	2851255	202624	921213	0.021999	1147167323
T4F	2198974	66545468	7145418	984486	2123317	0.038666	1036654039
T4M	1339876	51424532	2990880	197138	1253063	0.025946	1051212241
T5C	1338511	59270268	3003509	310932	1460799	0.026965	1197082582
T5F	1980060	45252463	3513888	528965	7825578	0.174949	115496763
T5M	1062988	56612039	2660466	168302	1131827	0.021279	1415430537
T6C	2578026	63399201	4439602	717966	6377110	0.104132	294754616
T6F	1931413	51031022	3829965	389923	1962146	0.041485	659999628
T6M	3583979	7260050	3926337	1129388	1351914	0.211732	17009615
T7C	758612	60554046	2344194	234454	927538	0.017146	1982688539
T7F	1725368	47251888	3494428	303753	2317839	0.05133	471050437
T7M	814605	53756361	2280031	119211	744762	0.014851	1966354587
T8C	1398328	63282016	4178788	395006	1133963	0.02081	1775204766
T8F	1272884	64616082	4206008	464481	2339433	0.039411	877042879
T8M	1064127	53646079	3725473	245293	670915	0.014647	2180868161



Supplementary Figure 2 – **Depth of coverage per individual.** Median (dark gray), first (light gray) and third (black) quartile of the depth of coverage for the 4 Atlantic males (T1M – T4M), the 4 eastern Mediterranean males (T5M – T8M), the 6 western Mediterranean females (T1F – T8F), and the descendant of each family (T1C – T8C). The average coverage depth of each family is indicated on the right side of the bars.

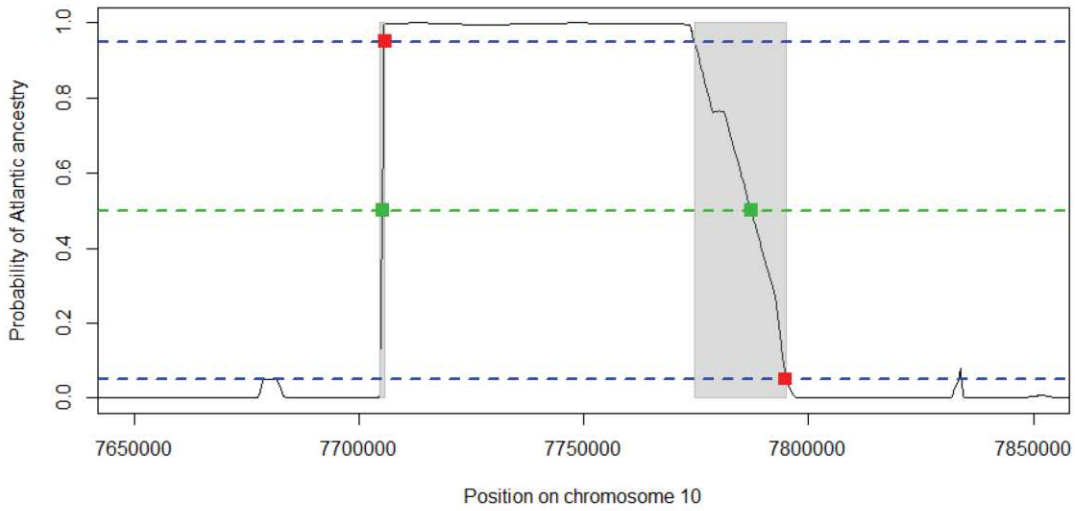
Supplementary Note 2: Analysis of spatial population structure

In order to determine the genetic position of the newly sequenced genomes with respect to the range-wide population structure of the European sea bass, we merged the new SNP dataset containing 2,628,725 SNPs with a RAD-derived SNP dataset containing 134,815 SNPs (mean read depth per SNP ranging from 10X to 150X, minor allele frequency (MAF) threshold = 0.01, genotyping rate threshold = 0.8). The RAD SNP dataset consists of 112 individuals from six locations: two in the Atlantic (Biarritz (France), N = 12; Mondego Estuary (Portugal), N = 26), two in the western Mediterranean (Palavas/Mauguio lagoon (France), N = 40, Annaba/Mellah lagoon (Algeria), N = 25) and two in the eastern Mediterranean (Syracuse (Italy), N = 1, Zarzis/El Biben lagoon (Tunisia), N = 8). Some individuals were already analyzed[1] and other were specifically added for this study. RAD SNP genotypes were obtained using the same reference mapping/genotyping pipeline as the one used for the whole genomes, but without using family priors. The two VCFs files derived from the whole genomes and the RAD-Sequencing data were merged together using VCFtools[3].

We then used the R package adegenet[4] to perform a Principal Component Analysis (PCA) of the combined dataset using only high-quality common variants. A total of 13,094 SNPs with a minor allele frequency > 0.10 and a genotyping rate greater than 0.9 were used for this analysis.

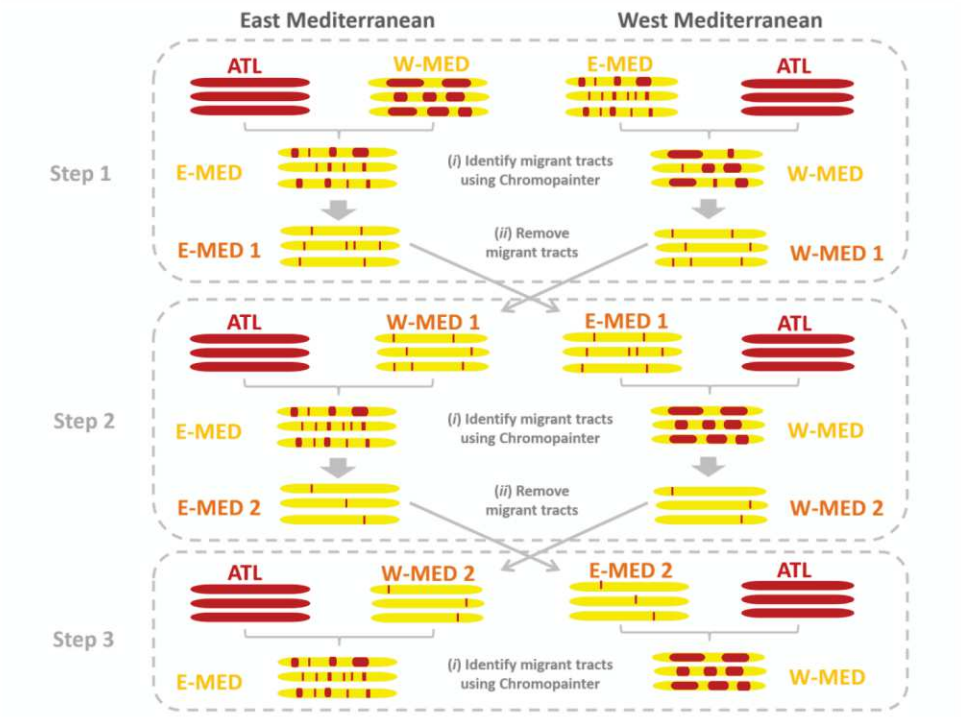
Supplementary Note 3: Detection of introgressed haplotypes and reconstruction of ancestral Mediterranean genomes

We used Chromopainter[5] with 150 iterations for the EM algorithm (expectation-maximization) to estimate the probability for each position along each haplotype to come from the Atlantic and Mediterranean population. Given that Chromopainter requires an estimate of the local recombination rate, we used the population recombination rate ($\rho = 4N_e r$) averaged between Atlantic and Mediterranean populations, already estimated in a previous study[1]. We then developed a method to analyze the ancestry probability profiles produced by Chromopainter to identify the beginning and the end of each introgressed tract (Supplementary Fig. 3).



Supplementary Figure 3 – **Probability profile of Atlantic ancestry for a Mediterranean sample haplotype along a 200kb region located on chromosome 10.** The probability of Atlantic ancestry was determined at each SNP by Chromopainter taking the local recombination rate into account. The squares represent the starting and ending positions of an introgressed tract of Atlantic origin before (red) and after (green) shifting the positions to the nearest point with a 0.5 probability (green dotted line) in the zone of uncertainty (grey rectangles). The blue dotted lines represent the probability thresholds used to consider a haplotype as truly Atlantic (0.95) or truly Mediterranean (0.05).

We reconstructed a non-introgressed Mediterranean population by removing introgressed Atlantic tracts to generate reference samples to be used in Chromopainter. We identified and removed introgressed Atlantic tracts within Mediterranean genomes using a three-step procedure (Supplementary Fig. 4). First, we used the Atlantic and E-MED populations as references to identify Atlantic haplotypes introgressed in the W-MED population, and reciprocally for the eastern Mediterranean population. We then removed the identified Atlantic tracts and replaced each of them by tracts identified as purely Mediterranean. The Mediterranean haplotypes used for replacement were obtained by determining the majority consensus sequence from all the sequences identified as purely Mediterranean at these positions. One consensus was generated for each Mediterranean population separately. Therefore, the gaps created by the removal of introgressed Atlantic tracts were filled with the most frequent allele found locally in the Mediterranean population considered. This step was repeated a second time using the Mediterranean populations reconstructed without introgressed tracts in the previous step as new references. This additional step aimed at removing residual tracts of Atlantic ancestry that could not be detected in the first step due to high rates of introgression in some regions of the genome in both eastern and western Mediterranean populations. After this step, only genomic regions that are completely swamped by Atlantic alleles should remain undetected, which is probably very rare. The two resulting Mediterranean populations obtained after these two steps should therefore correspond to the Mediterranean ancestral population before the beginning of gene flow. They were used in a third and last step as Mediterranean reference populations to identify migrant tracts (Supplementary Fig. 4).

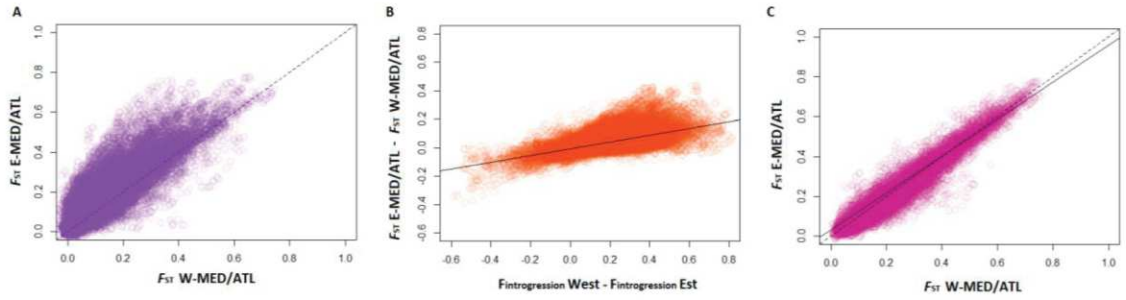


Supplementary Figure 4 – **Schematic overview of the three steps used to reconstruct the ancestral diversity of Mediterranean genomes and identify migrant tracts.** For each step, 3 chromosomes are represented per population for illustration. The red color is used for Atlantic DNA fragments and the yellow color for Mediterranean DNA fragments.

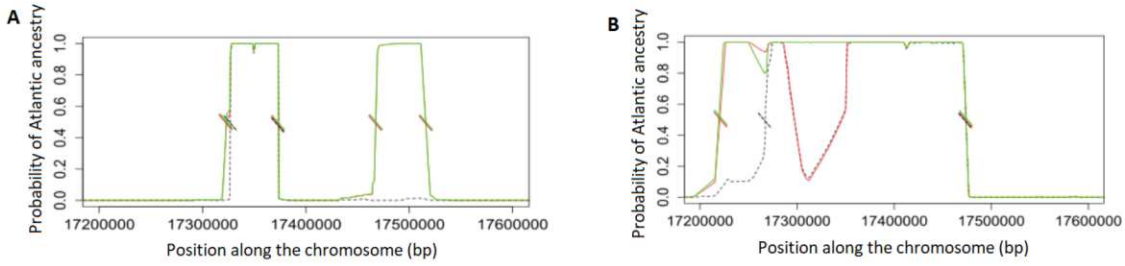
Supplementary Note 4: Effectiveness of ancestral reconstruction

We evaluated the effectiveness of our strategy to reconstruct ancestral Mediterranean genomes. Genome-wide differentiation measures showed that the E-MED population is genetically more differentiated from the Atlantic population than the W-MED population (Supplementary Fig. 5A). This is particularly true for genomic regions with intermediate F_{ST} values. If this difference is due to gene flow, then the W-MED population should be more strongly impacted by introgression than the E-MED population. Indeed, we found that the difference in F_{ST} measured between Atlantic and W-MED versus Atlantic and E-MED (*i.e.* $F_{ST\ (E\text{-}MED/ATL)}$ - $F_{ST\ (W\text{-}MED/ATL)}$) was positively correlated with the differential of introgression (*i.e.* $F_{Introgression\ (W\text{-}MED)}$ - $F_{Introgression\ (E\text{-}MED)}$) between W-MED and E-MED (Supplementary Fig. 5B). This result confirms that the genetic differences observed between the two Mediterranean populations are explained by an increased frequency of introgression within western compared to eastern Mediterranean genomes. As a corollary, we also show that our strategy to reconstruct ancestral Mediterranean genomes generates very similar results between W-MED and E-MED populations. Indeed, genetic differentiation between Atlantic and ancestral W-MED is highly positively correlated to genetic differentiation between Atlantic and ancestral E-MED, with a regression slope close to 1 confirming that reconstructed ancestral Mediterranean populations are equally differentiated from the Atlantic population (Supplementary Fig. 5C). This suggests that the method to reconstruct the ancestral state of Mediterranean genomes independently between W-MED and E-MED has efficiently reconstructed the same gene pool, supposed to reflect the genetic composition of the Mediterranean population before the beginning of gene flow. We note that this approach only treats the effect of introgression and does not address the coalescence of Mediterranean alleles to reconstruct ancestral genomes. It may also bias the frequency of Mediterranean alleles by replacing Atlantic tracts by the most frequent Mediterranean haplotype. However, this does not affect our approach since we mostly used the reconstructed ancestral genomes as references to improve the detection of introgressed tracts.

Using the reconstructed Mediterranean ancestral populations as references in Chromopainter improved the detection of migrant tracts within Mediterranean genomes in two different ways. First, the use of reconstructed reference populations allowed the detection of previously undetected Atlantic haplotypes (Supplementary Fig. 6A). Second, it also increased the precision of tract length estimation (Supplementary Fig. 6B). In most cases, the removal of Atlantic tracts in the Mediterranean reference populations increased the Atlantic ancestry probability in the genomic regions where introgression was strong, thus enabling a better detection and more precise measurement of migrant tracts. Finally, estimated ancestry probability profiles were very similar using the ancestral Mediterranean population reconstructed after one or two steps of introgressed tracts removal (Supplementary Fig. 6). This suggests that two steps were enough to efficiently detect and remove Atlantic tracts within Mediterranean genomes, and that adding a third step would not significantly improve the approach.



Supplementary Figure 5 – **Genome-wide correlations between different statistics averaged in 100 kb windows.** A. Correlation of genetic differentiation (F_{ST}) measured between Atlantic and western Mediterranean populations (F_{ST} W-MED/ATL) and between Atlantic and eastern Mediterranean populations (F_{ST} E-MED/ATL). The dotted line represents the equation y = x. B. Correlation between the difference in F_{ST} between the comparisons E-MED/ATL and W-MED/ATL and the differential of introgression between western and eastern Mediterranean populations. The black line is the linear regression based on the data. C. Correlation of genetic differentiation (F_{ST}) measured between Atlantic and ancestral western Mediterranean (F_{ST} W-MED/ATL) and between Atlantic and ancestral eastern Mediterranean (F_{ST} E-MED/ATL). The dotted line represents the equation y = x and the black line shows the linear regression based on the data.

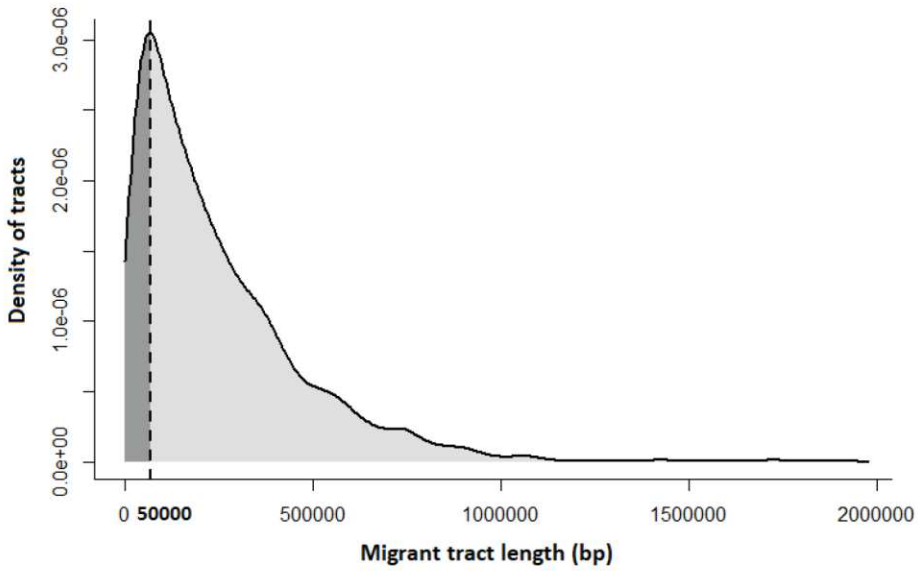


Supplementary Figure 6 – **Examples of Atlantic ancestry probability profiles inferred along chromosome 1A for one chromosome haplotype of a western Mediterranean individual**. Estimation of local ancestry was performed using the Atlantic population as reference together with either the contemporary E-MED population (black dotted lines), the reconstructed ancestral E-MED population after one (red lines) or two rounds of introgressed tracts removal (green lines). Oblique marks indicate the starting and ending positions of detected Atlantic tracts using each of the three different Mediterranean populations as reference (black, red, green). Using a reconstructed ancestral Mediterranean population allows to A. detect new migrant tracts and B. improve the estimation of migrant tract length.

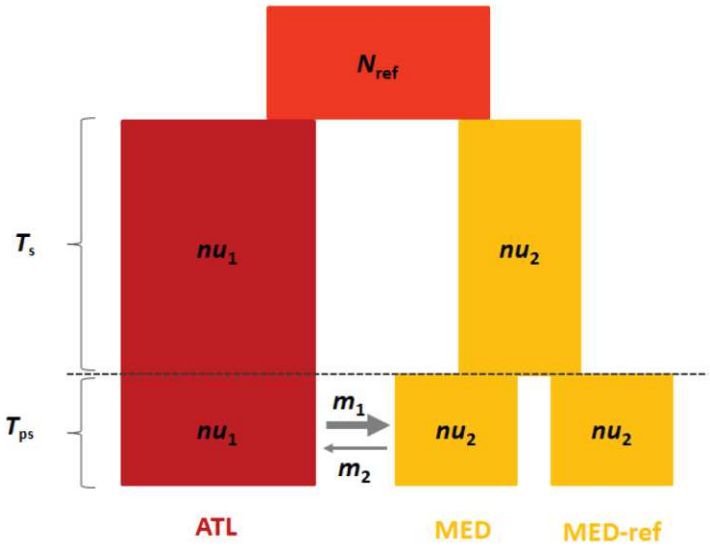
Supplementary Note 5: Analysis of migrant tract length distribution

The distribution of migrant tract length is informative about the timing and intensity of the genetic exchanges that have occurred between populations. After introgression into a recipient population, a migrant tract is progressively eroded over generations due to recombination with the introgressed genetic background[6,7]. The length of a migrant tract thus depends on the number of generations elapsed since introgression (t) and on the local recombination rate of the genomic region in which it is located (r). The average length of migrant tracts ($\overline{L}$) following a single pulse of admixture occurring t generations ago can be approximated by $\overline{L} = [\ 1 - f\ r\ t - 1\ ]^{-1}$, where f is the fraction of the population replaced by migrants[8]. We used the previous equation to estimate the date of introgression of the most abundant class of tracts found in low-recombining regions of the sea bass genome, which have an average length of 50 kb (Supplementary Fig. 7). Using $f = 0.31$ (the mean frequency of introgression estimated in W-MED individuals, see results), $r = 8.45\times10^{-9}$ M/pb (estimated in a previous study[1]) and assuming a generation time of 5 years[1], we calculated that these tracts have introgressed approximately 17,000 years BP. Since 85% of the introgressed tracts of Atlantic origin found in low-recombining regions are on average longer than 50kb, they probably introgressed the W-MED population less than 17,000 years ago. Reciprocally, only 15% of the tracts are on average shorter than 50 kb and have therefore possibly introgressed the W-MED population during an older contact episode. Therefore, the secondary contact model previously inferred without using linkage information[1] appears consistent with this result. In order to assess its goodness-of-fit more thoroughly, we compared the observed length distributions of migrant tracts with the ones obtained by simulation under the secondary contact model.

Coalescent simulations under the secondary contact model included a third, non-introgressed Mediterranean population, to use later as a reference in Chromopainter. To do so, we splitted the Mediterranean population into two replicates at the beginning of the secondary contact so that only one replicate is affected by gene flow from the Atlantic (Supplementary Fig. 8). Model parameter values corresponded to the ones inferred previously[1]. However, since this model included two different categories of loci experiencing different effective migration rates, we used the weighted average migration rate taking into account the relative fraction of the genome occupied by each category of loci. A simulation was performed independently for each chromosome using the local recombination rate previously inferred for each population[1]. Finally, we used Chromopainter to get the length distribution of migrant tracts for each of the two simulated populations.



Supplementary Figure 7 – **Distribution of the average length of Atlantic migrant tracts found in 100 kb windows located in low-recombining regions of W-MED genomes.** The vertical dotted line shows the most abundant class length represented by windows with an average tract length of 50 kb. About 15 % of the windows have a lower mean tract length (dark grey area) and 85% of the windows a higher mean tract length (light grey area). Regions are considered as low-recombining if $\rho \leq 10$, with ρ being the population-scaled recombination parameter previously estimated[4].



Supplementary Figure 8 – **Schematic representation of the model implemented to simulate the migrant tract length distribution**. An ancestral population of size N_{ref} splits into two derived populations of size nu_1 and nu_2, which evolve without exchanging genes during T_s generations. At the end of this allopatric divergence period, the Mediterranean population splits into two populations of size nu_2, and only one of them exchanges genes with the Atlantic population during T_{ps} generations. Migration from ATL to MED occurs at rate m_1, and at rate m_2 in the opposite direction. MED-ref was used as a reference population together with the Atlantic population for the detection of introgressed tracts in Chromopainter to detect introgressed tracts.

Supplementary Note 6: Testing waves of historical gene flow

We tested whether there have been several periods of allopatric isolation and secondary contact between the Atlantic and Mediterranean Sea bass lineages. To do so, we used a method[9] that exploits the information contained in a collection of pairwise sequence alignments by summarizing the length distribution of tracts of identity-by-state (IBS). An L-base IBS tract is defined as a segment of L contiguous identical base pairs between two consecutive SNPs. The distribution of IBS tracts shared between DNA sequences from different populations contains information about population divergence, paste gene flow and genetic diversity that existed at different time in the populations. It can therefore be used to jointly estimate the timing and magnitude of past admixture events, population divergence times and changes in effective population size. The method uses an approximate formula to predict the expected IBS tract length distribution within and between populations under different models of historical divergence with or without gene flow. The empirical and predicted distributions are then compared to maximize a composite likelihood function.

We developed a flexible model that accounts for a large diversity of demographic histories using nine parameters (Supplementary Fig. 3B). The model can represent three different categories of scenarios, (*i*) continuous migration, (*ii*) secondary contact and (*iii*) periodic pulses. The three categories of scenarios were then compared to each other for each value of $m \times n$ to identify the best model. Note that the periodic pulses scenario is only defined for $m \times n \in \{4, 6, 8, 9, 10\}$ (Supplementary Fig. 3C). Parameter bounds were set as follows: [0.1; 2] for the effective population sizes, [0.001; 0.5] for the admixture fractions, [0.001; 1] for the time parameters, except for the periodic pulses scenario in which the bounds of T_{diff} were set to [0.1; 1] to force divergence periods to be at least $0.1N_e$ generations long. The genetic diversity θ ($4N_e\mu$) and the population scaled recombination rate ρ ($4N_e r$) parameter were estimated from previous data[1], and set to $\theta = 0.001$ and $\rho = 0.001$ with $N_e = N_{ref} = 100{,}000$. Finally, the binning scheme was adjusted to better capture the signature of recent admixture, using 32 bins with b_0=100 and b_{i+1}=1.3×b_i. Therefore, IBS tracts shorter than 100 base pairs were not considered for likelihood optimization, but no upper threshold was set on the longest IBS tracts.

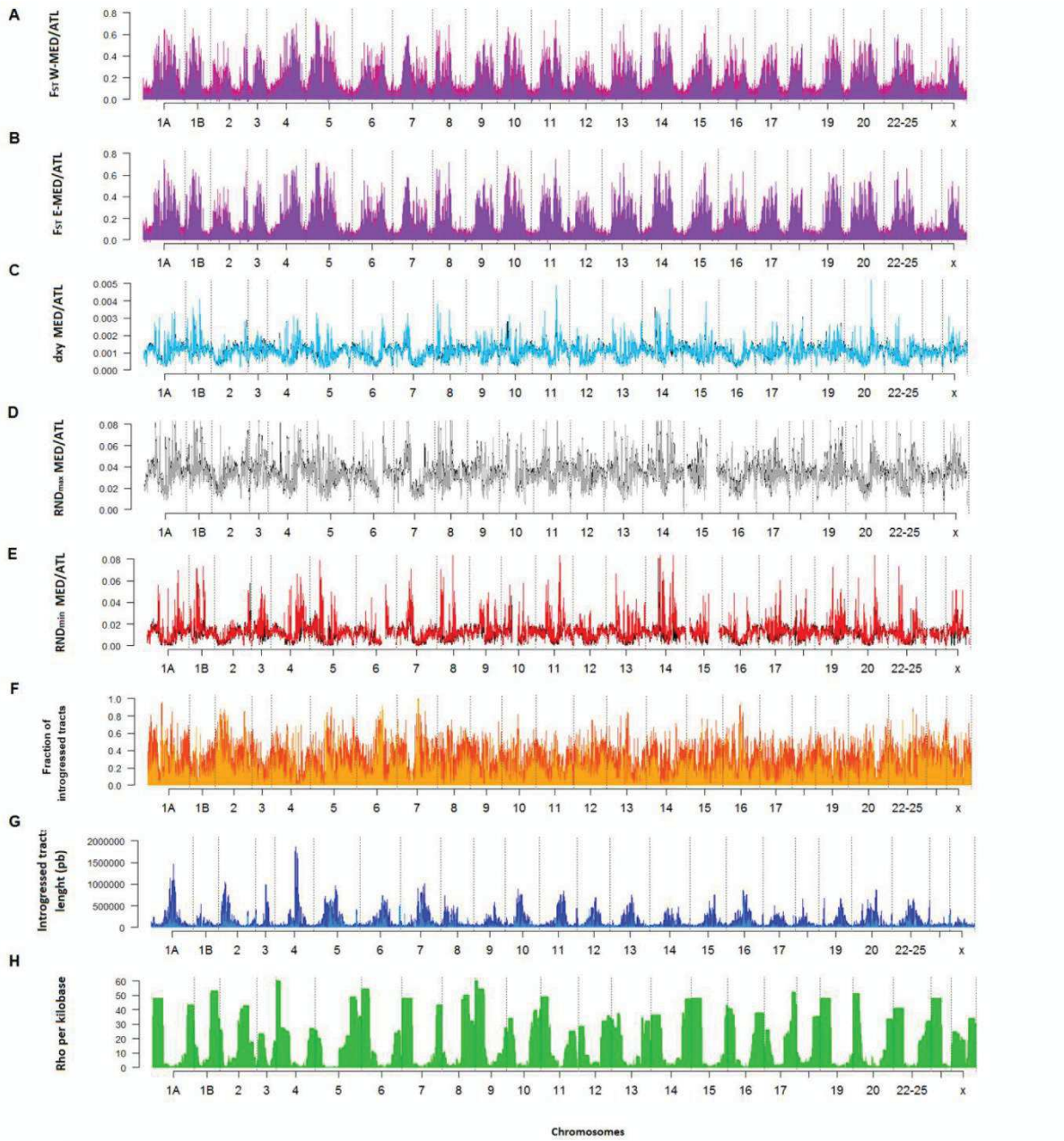
Supplementary Table 2 – **Results of model fitting for 27 combinations of *m* and *n* parameters.** The log(*Likelihood*) value is showed for the best run obtained over 20 optimizations for each (*m*:*n*) combination.

Number of pulses	1	2	3	4	5	6	7	8	9	10
Secondary contact	(1:1) -245443	(1:2) -126504	(1:3) -85028	(1:4) -73381	(1:5) -66147	(1:6) -61448	(1:7) -55760	(1:8) -55824	(1:9) -54024	(1:10) -52613
Continuous migration		(2:1) -133220	(3:1) -90439	(4:1) -76903	(5:1) -70269	(6:1) -65633	(7:1) -62301	(8:1) -59814	(9:1) -57892	(10:1) -56366
Periodic pulses				(2:2) -75199		(2:3) -66411 (3:2) -67636		(2:4) -61908 (4:2) -66536	(3:3) -64122	(2:5) -59102 (5:2) -67111
Best model	SC	SC	SC	SC	SC	SC	SC	SC	SC	SC

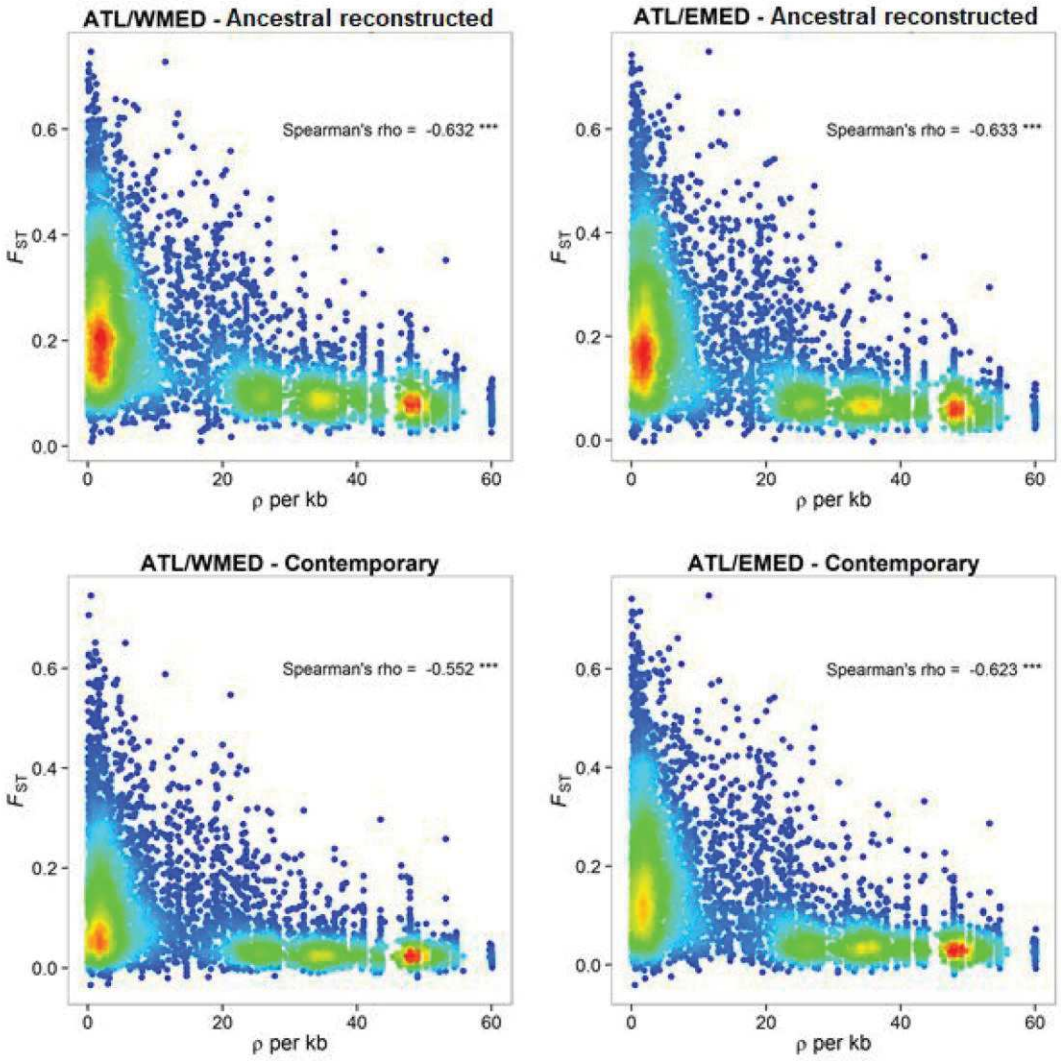
Supplementary Table 3 – **Results of model fitting for 27 combinations of *m* and *n* parameters.** Details of parameter values are provided for the best run obtained over 20 optimizations for each *m*:*n* combination. The best-fit model (1:10) is highlighted in red.

m	*n*	N_1	N_2	N	T_c	T_{diff}	f_1	f_2	*Likelihood*	Model
1	1	0.1408	0.3534	0.8956	0.242	0.001	0.5	0.0445	-245443	Sec. contact
1	2	0.119	0.4057	1.0668	0.158	0.2538	0.5	0.0372	-126504	Sec. contact
1	3	0.109	0.469	1.1813	0.2402	0.2057	0.5	0.001	-85028	Sec. contact
1	4	0.1621	0.429	1.2345	0.3189	0.1923	0.4118	0.032	-73381	Sec. contact
1	5	0.1883	0.4135	1.2682	0.3617	0.1837	0.3573	0.0403	-66147	Sec. contact
1	6	0.204	0.4061	1.2903	0.391	0.1746	0.3191	0.0421	-61448	Sec. contact
1	7	0.221	0.4024	1.3191	0.4824	0.1218	0.3038	0.0467	-55760	Sec. contact
1	8	0.2223	0.3988	1.3163	0.4301	0.1572	0.2659	0.0406	-55824	Sec. contact
1	9	0.2286	0.3967	1.3244	0.442	0.1492	0.2463	0.0392	-54024	Sec. contact
1	10	0.2332	0.3951	1.3306	0.4561	0.1417	0.2296	0.0377	-52613	Sec. contact
2	1	0.1315	0.3862	1.0426	0.1887	0.001	0.5	0.0468	-133220	Continuous
2	2	0.1284	0.4493	1.2087	0.1312	0.1	0.4645	0.0142	-75199	Periodic
2	3	0.1899	0.4081	1.295	0.1791	0.1	0.3454	0.0437	-66411	Periodic
2	4	0.2119	0.3993	1.3476	0.2061	0.1	0.2873	0.0448	-61908	Periodic
2	5	0.2247	0.3956	1.3769	0.2241	0.1	0.248	0.0425	-59102	Periodic
3	1	0.1107	0.4432	1.1433	0.1399	0.001	0.5	0.0104	-90439	Continuous
3	2	0.1898	0.4052	1.3222	0.0951	0.1	0.3501	0.0482	-67636	Periodic
3	3	0.2157	0.396	1.3854	0.1217	0.1	0.27	0.0465	-64122	Periodic
4	1	0.1166	0.4662	1.1956	0.1096	0.001	0.4677	0.001	-76903	Continuous
4	2	0.2075	0.3974	1.3831	0.0655	0.1	0.2943	0.0495	-66536	Periodic
5	1	0.1625	0.4264	1.2291	0.0973	0.001	0.3883	0.0293	-70269	Periodic
5	2	0.2109	0.3965	1.3981	0.0382	0.1	0.2502	0.044	-67111	Periodic
6	1	0.1825	0.4141	1.2514	0.0846	0.001	0.3447	0.0357	-65633	Continuous
7	1	0.1949	0.4078	1.2677	0.0744	0.001	0.3123	0.0375	-62301	Continuous
8	1	0.2044	0.4033	1.2807	0.0664	0.001	0.2863	0.0378	-59814	Continuous
9	1	0.2115	0.4002	1.2911	0.0598	0.001	0.2649	0.0373	-57892	Continuous
10	1	0.2172	0.398	1.2995	0.0545	0.001	0.2467	0.0364	-56366	Continuous

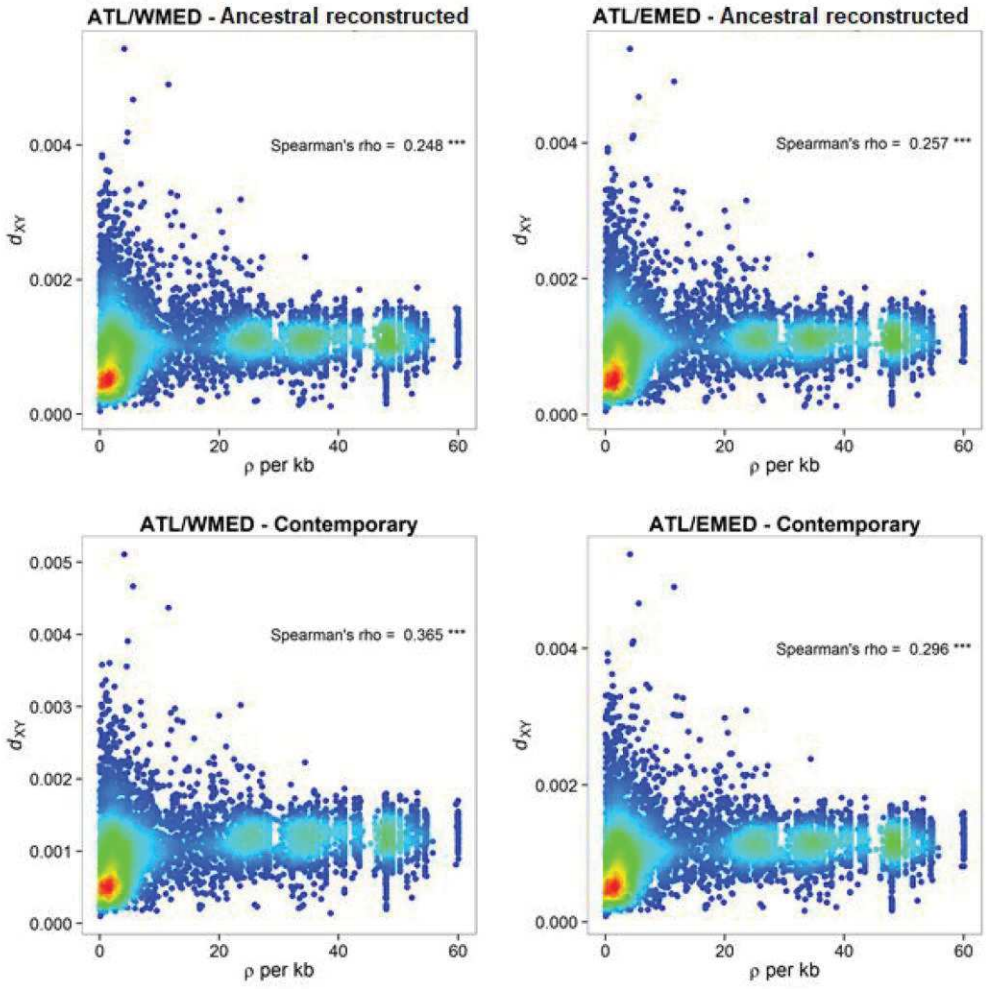
m : Number of periods of divergence-contact
n : Number of admixture pulse during contact
N_1 : Effective size of the Atlantic population in units of N_{ref}
N_2 : Effective size of the Mediterranean population un units of N_{ref}
N : Effective size of the ancestral population in units of N_{ref}
T_c : Duration of contact in units of N_{ref} generations
T_{diff} : Duration of divergence in units of N_{ref} generations
f_1 : Fraction of Atlantic migrating to Mediterranean population
f_2 : Fraction of Mediterranean migrating to Atlantic
generation time : 5 years



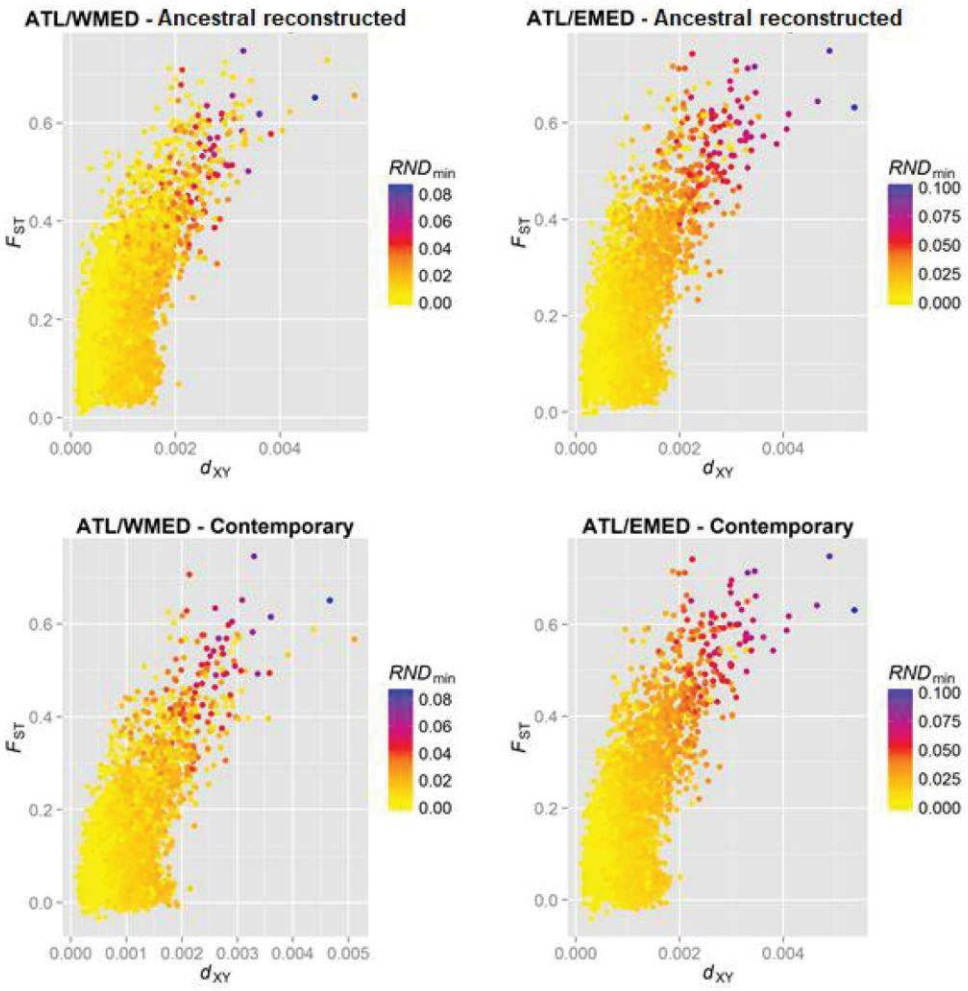
Supplementary Figure 9 – **Comparison of different population genetics statistics calculated in non-overlapping 100 kb windows along the sea bass genome.** Dashed vertical lines represent the limits between chromosomes. Chromosome x does not refer to a sexual chromosome. A. F_{ST} measured between the Atlantic and the contemporary (purple) or ancestral reconstructed (mauve) W-MED population. B. F_{ST} measured between the Atlantic and the present (purple) or ancestral reconstructed (mauve) E-MED population. C. d_{XY} calculated between the Atlantic and the W-MED (black) and E-MED (blue) populations. D. RND_{max} measured between the Atlantic and the W-MED (black) and E-MED (grey) populations. E. RND_{min} measured between the Atlantic and the W-MED (black) and E-MED (red) populations. F. Fraction of introgressed tracts in the W-MED (orange) and E-MED (yellow) populations. G. Average length of introgressed tracts in the W-MED (dark blue) and E-MED (light blue) populations. H. Population-scaled recombination rate (ρ=$4N_e r$ per kb) averaged between Atlantic and Mediterranean populations.



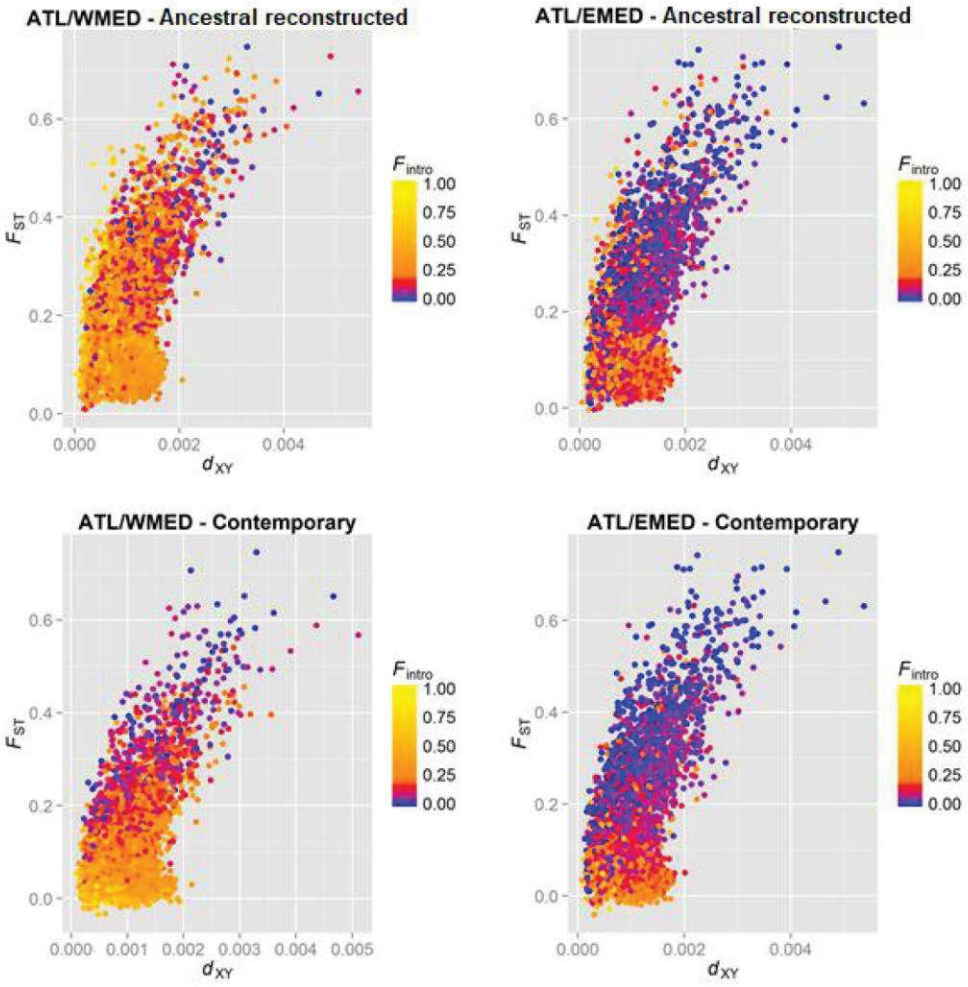
Supplementary Figure 10 – **Relationships between the population-scaled recombination rate (ρ=$4N_e r$ per kb) and genetic differentiation (F_{ST}) calculated in non-overlapping 100 kb windows.** The density of points appears in color scale from low (blue) to high (red) densities.



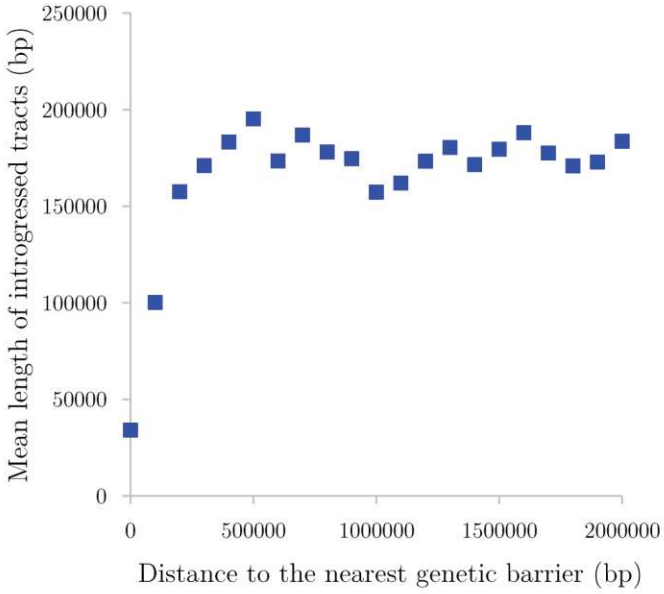
Supplementary Figure 11 – **Relationships between the population-scaled recombination rate (ρ=$4N_e r$ per kb) and nucleotide divergence (d_{XY}) calculated in non-overlapping 100 kb windows.** The density of points appears in color scale from low (blue) to high (red) densities.



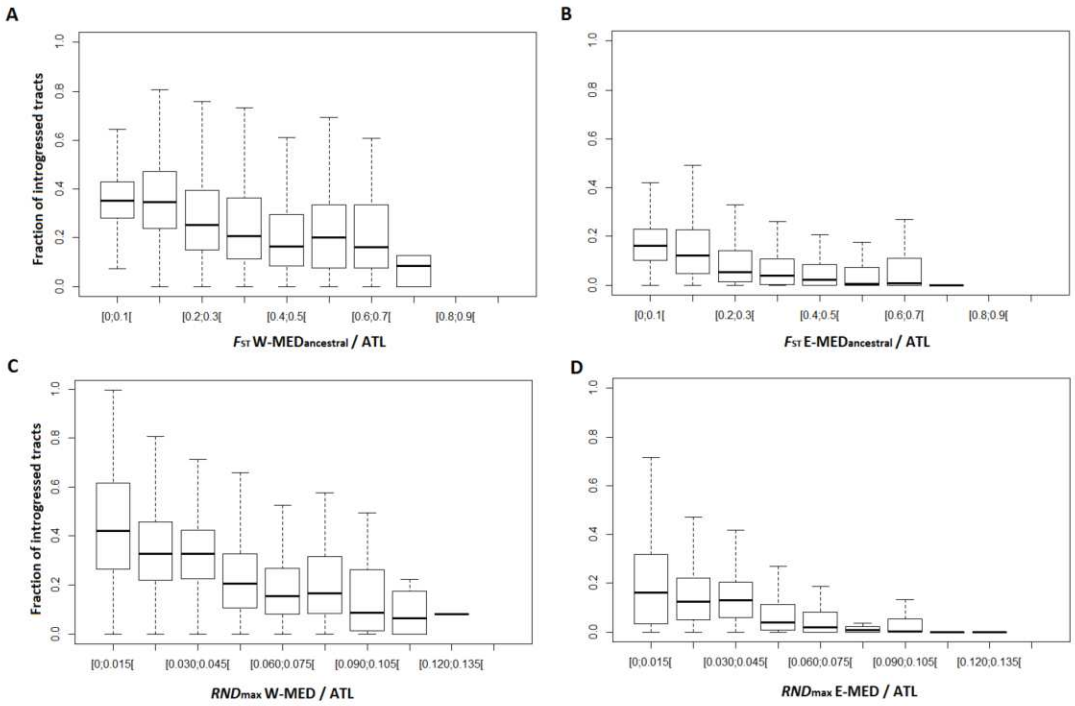
Supplementary Figure 12 – **Relationships between genetic differentiation (F_{ST}) and nucleotide divergence (d_{XY}) calculated in non-overlapping 100 kb windows.** The color scale indicates the value of RND_{min} in the corresponding window from low (yellow) to high (blue) values.



Supplementary Figure 13 – **Relationships between genetic differentiation (F_{ST}) and nucleotide divergence (d_{XY}) calculated in non-overlapping 100 kb windows.** The color scale indicates the frequency of introgression in the corresponding window from low (blue) to high (yellow) frequencies.



Supplementary Figure 14 – **Relationship between physical distance to the nearest genetic barrier and the mean length of introgressed tracts of Atlantic ancestry found in the W-MED population.** We used the 99th percentile of the distribution of RND_{min} values as a threshold to define outlier genomic regions (100 kb windows) that are highly resistant to introgression and therefore likely contain barrier loci. Each window was used as a reference position from which we calculated the mean length of Atlantic migrant tracts at increasing physical distances. Measures were finally combined and averaged across outlier regions.



Supplementary Figure 15 – **Fraction of introgressed tracts for different levels of ancestral genetic differentiation and allelic divergence measured with $RND_{\max}$.** A. Introgression as a function of ancestral F_{ST} measured between the Atlantic and the reconstructed ancestral W-MED or E-MED (B.) population. C. Introgression as a function of $RND_{\max}$ measured between the Atlantic and W-MED or E-MED (D.) population. Each box represents the lower and upper quartiles and the median of introgression frequency values. The negative correlation between the fraction of introgressed tracts and ancestral F_{ST} is not a methodological artifact due to the removal of introgressed tracts from contemporary genomes. This procedure would on the contrary tend to overestimate the ancestral F_{ST} in highly introgressed regions, where the reconstructed ancestral diversity of the Mediterranean population may be downwardly biased.

ANNEXE 2: Matériel supplémentaire de l'article:

The contribution of ancient admixture to reproductive isolation between European sea bass lineages.

Supplementary Materials

Duranton et al.

The contribution of ancient admixture to reproductive isolation between European sea bass lineages

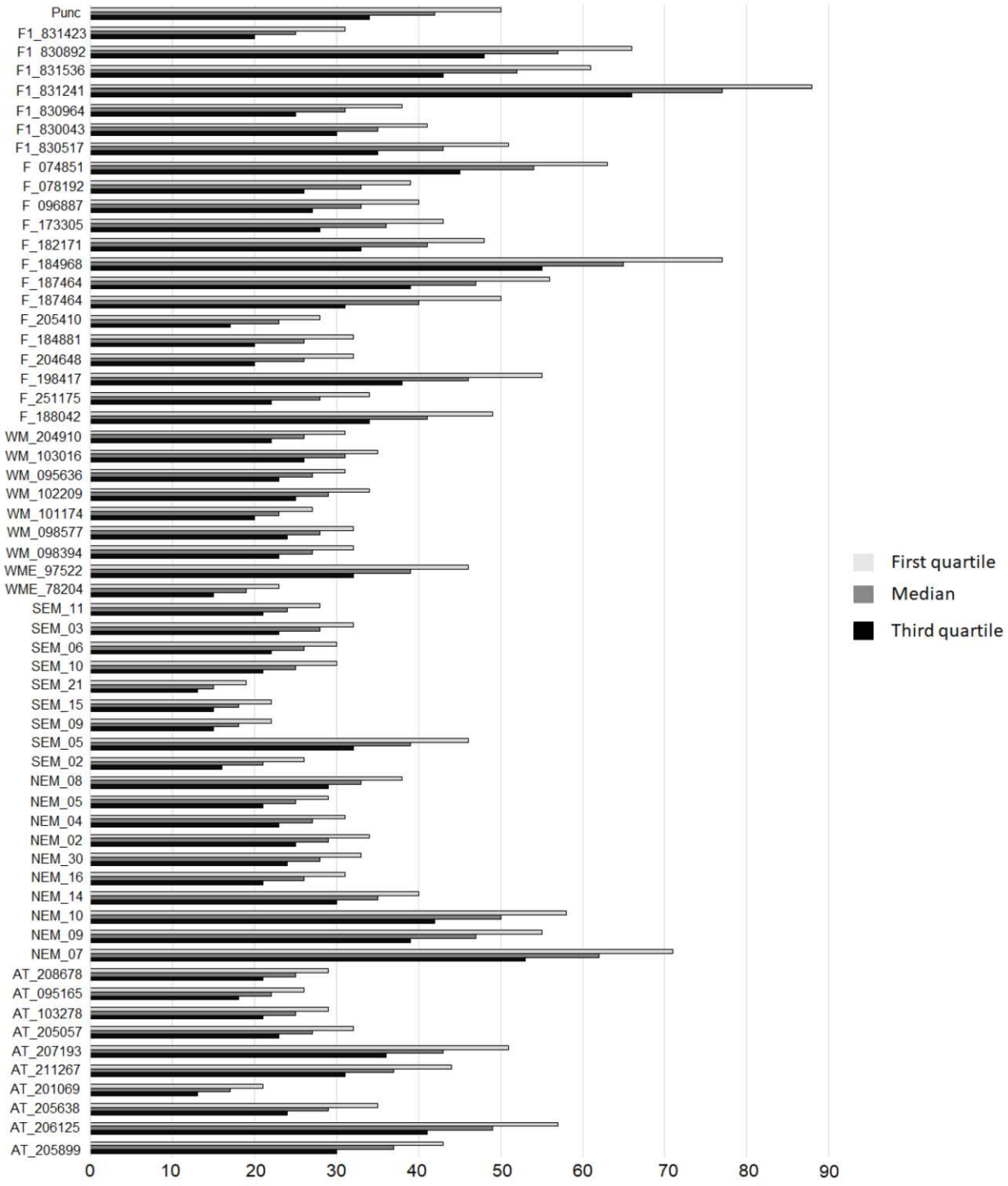
May 2019

1 - Whole-genome resequencing

Individual	Library kit	Run	Unpaired reads examined	Reads pairs examined	Unmapped reads	Read pair duplicates	Percent duplication	Estimated library size
AT_205899	PCR-free	2	710525	51574949	1779055	5243750	0.102331	338339229
		1	1599933	51480924	2756181	13676012	0.26773	89393199
AT_206125	PCR-free	2	1162366	66285637	2548540	5600109	0.085594	533205434
		1	2099730	70022633	3645908	17135878	0.246806	135168917
AT_205638	PCR-free	2	592020	42345534	1467942	5222111	0.123944	225073296
		1	949465	40386831	1834665	12051209	0.299717	60298353
AT_211069	PCR-free	2	340601	23707392	825243	2554580	0.108193	145663521
		1	503439	21859655	978983	5980255	0.274728	36599828
AT_211267	PCR-free	2	718656	50229040	1771632	4354008	0.087411	386708710
		1	1676865	52761066	2870845	12284403	0.234926	107357281
AT_207193	PCR-free	2	933708	59620496	2174352	6025969	0.101973	396946010
		1	1725667	62518162	3095545	16425410	0.26464	109956484
AT_205057	Nano	1	847273	65506340	2196319	3561652	0.055356	770163673
AT_103278	Nano	1	818074	58633597	2032366	3419072	0.059535	634069837
AT_095165	Nano	1	569451	50971720	1623871	2381592	0.047552	676857516
AT_208678	Nano	1	783941	59577971	2004383	3017378	0.051526	754343749
NEM_07	PCR-free	2	1698446	117368673	4077408	18711889	0.160839	399832566
		1	710006	52729283	1759686	9029507	0.172299	160945863
NEM_09	PCR-free	2	1050548	83890852	2741562	10541945	0.126624	405629701
		1	480858	41381960	1304016	8080455	0.195879	117595746
NEM_10	PCR-free	2	1316276	94951027	3241172	12858549	0.136506	424680397
		1	471446	38196025	1226054	6943165	0.182347	113906586
NEM_14	PCR-free	1	1627807	92941773	3483253	13813625	0.149824	408915958
NEM_16	PCR-free	2	593154	44861040	1513572	4786524	0.107339	258646208
		1	224528	19291930	613914	3058638	0.158891	69285554
NEM_30	Nano	1	863049	68382690	2267473	3819091	0.056962	788549675
NEM_02	Nano	1	926545	70986840	2365455	4528848	0.065048	712885179
NEM_04	Nano	1	743971	64687003	2096245	3486045	0.054803	763144409
NEM_05	Nano	1	759149	59477335	1973059	2899192	0.049731	779637467
NEM_08	Nano	1	1035595	82073920	2694179	5626675	0.069673	721645856
SEM_02	PCR-free	2	465634	34841686	1192898	4342291	0.125164	169755455
		1	214257	17618667	574709	3265401	0.18572	51785382
SEM_05	PCR-free	2	928763	69243971	2348789	8962764	0.130297	323138092
		1	405300	32921159	1070004	6303847	0.192022	92016456
SEM_09	PCR-free	1	631808	46758401	1598742	6363989	0.136711	219184320
SEM_15	PCR-free	1	661396	47118796	1624458	6575007	0.140163	218120903
SEM_21	PCR-free	1	570375	40958914	1403661	6581858	0.161225	158247672
SEM_10	Nano	1	687721	60932741	1962181	3344853	0.056067	668105655
SEM_06	Nano	1	747676	63988648	2077118	4195923	0.066478	588179698
SEM_03	Nano	1	838812	67601876	2244974	4340684	0.065364	639954765
SEM_11	Nano	1	622507	60102973	1868897	4606246	0.077472	461536800
WME_78204	PCR-free	2	809260	34759734	1527580	5125749	0.148367	131097459
		1	258012	13636123	536218	2139774	0.158096	47348614
WME_97522	PCR-free	1	1057906	74398643	2547664	11626900	0.157196	259519641
		2	334180	28357701	894368	4341746	0.153585	98230713

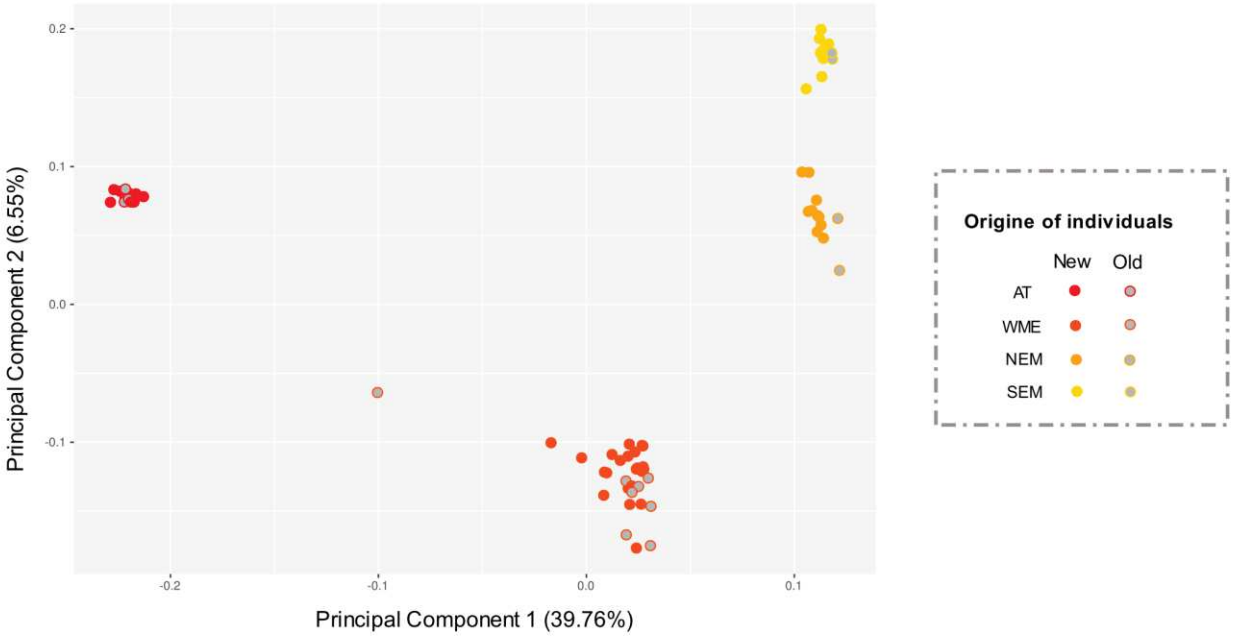
WM_098394	Nano	1	913210	66517786	2295822	4729605	0.072642	588581722
WM_098577	Nano	1	998106	67307402	2403470	4393065	0.066926	665052074
WM_101174	Nano	1	661295	55575994	1816347	2859239	0.052294	687069162
WM_102209	Nano	1	958027	70934697	2430433	4794683	0.068752	674026270
WM_095636	Nano	1	832945	65205207	2218543	4450311	0.069881	603852551
WM_103016	Nano	1	936302	74935945	2439552	5067344	0.068693	698973985
WM_204910	Nano	1	757779	63134426	2068975	3361863	0.054263	738990499
F_188042	PCR-free	2	852878	64569247	2180018	7934029	0.124367	320396420
		1	709483	51550989	1753305	14425349	0.281215	94143894
F_251175	PCR-free	2	672662	45797284	1605188	6306069	0.139438	196194463
		1	495515	31008533	1115349	8438385	0.274215	56574458
F_198417	PCR-free	2	853900	72155477	2365866	9631180	0.134809	339331281
		1	753758	61720283	2031438	18708110	0.304173	99301743
F_204648	PCR-free	2	514431	40272333	1338091	5120777	0.128122	193936148
		1	380574	30334980	989602	8009966	0.264841	59517624
F_184881	PCR-free	2	603672	42897682	1470562	5369542	0.127079	203711488
		1	467932	28829982	1044640	7264342	0.254274	57132036
F_204510	PCR-free	2	527829	40700253	1373179	5945218	0.147716	149284193
		1	324113	18235037	695701	2922069	0.163692	59486925
F_197773	PCR-free	2	3213575	64690627	4529189	7411979	0.13246	334487318
		1	4291010	52645926	5367162	14004442	0.291582	98076543
F_187464	PCR-free	2	983080	74571529	2552874	9194735	0.125058	370950906
		1	826343	59104941	2051287	15924558	0.270975	112752384
F_184968	PCR-free	2	1837939	110640723	4107239	15000697	0.138978	475982869
		1	1568769	77255006	3138653	20751473	0.272601	140131340
F_182171	PCR-free	2	846844	63186472	2156862	7172435	0.115273	344174477
		1	712446	48567533	1704558	12497390	0.259083	98293631
F_173305	PCR-free	2	904549	59066109	2142933	7660562	0.132085	268343212
		1	741286	40810732	1586048	10850386	0.268928	75798284
F_096887	PCR-free	2	753057	62259844	2029947	9814488	0.158953	211109766
		1	338562	26714113	876818	4406579	0.166428	84146648
F_078192	PCR-free	2	648072	46605554	1600276	4852019	0.105393	286878630
		1	1053911	43471190	2016367	10667848	0.247131	82272808
F_074851	PCR-free	2	1075079	79548876	2720599	8253526	0.105379	500993586
		1	1427481	74762818	3057529	18081154	0.243812	144112228
F1_830517	PCR-free	2	826604	65446108	2192368	8283817	0.127667	330242903
		1	720708	56592447	1879556	16747699	0.296774	94555459
F1_830043	PCR-free	1	1308835	96155133	3314643	14935653	0.156796	384114166
F1_830964	PCR-free	2	592785	46930900	1569853	6075256	0.130339	232370733
		1	483220	39816019	1295388	11354089	0.285826	70042740
F1_831241	PCR-free	2	1557618	114292058	3963642	13635367	0.120982	626917372
		1	2737677	111792696	5264401	32105291	0.289761	174771294
F1_831536	PCR-free	2	921352	79958678	2574800	9152029	0.115817	445062716
		1	1199337	68461264	2715665	18366601	0.26993	116091736
F1_830892	PCR-free	2	1151356	88028097	2966116	10762510	0.123891	451642355
		1	1022164	73024308	2507784	20468681	0.281729	130908069
F1_831423	PCR-free	2	429128	37300868	1214448	4647009	0.125227	188613285
		1	354593	30409425	978823	8038081	0.26486	58653591
Punc	PCR-free	2	1975030	70156271	3671546	8873276	0.132611	323908505
		1	1846933	44658812	2923019	11439503	0.266085	86609202

Supplementary Table 1 – Summary statistics of sequencing and mapping data for each individual. Individuals whose name is in bold are those involved in crossing.

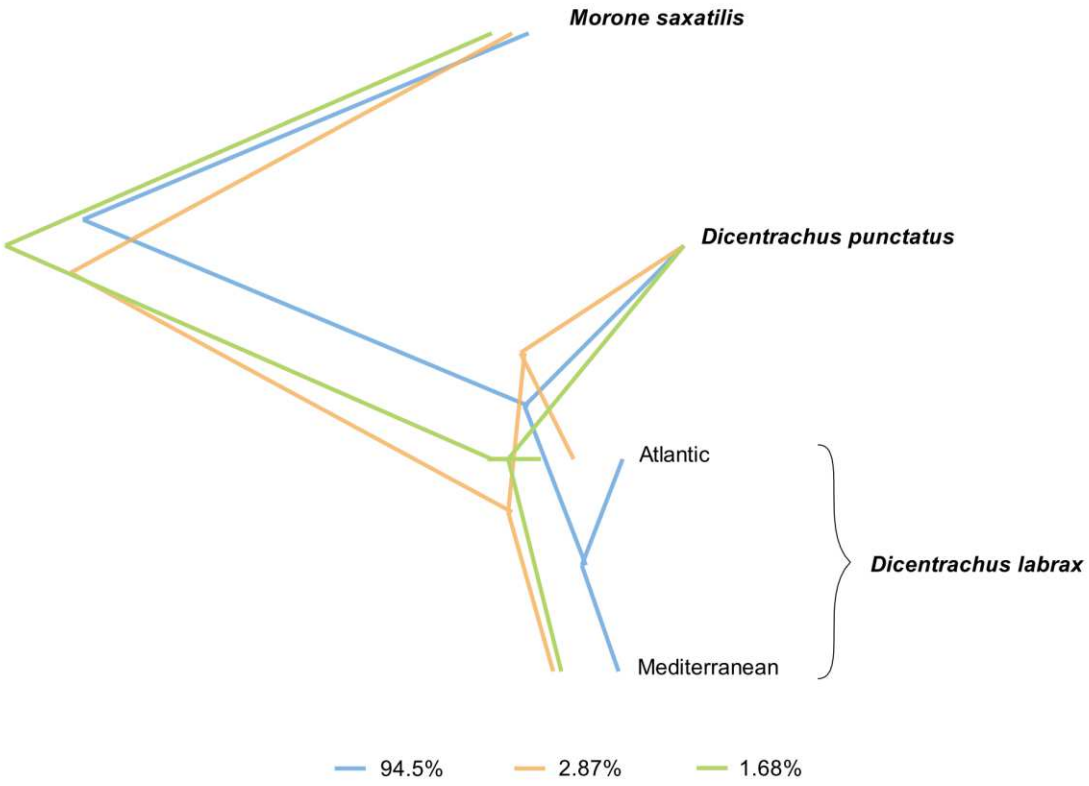


Supplementary Figure 1 - **Depth of coverage per individual**. Median (dark gray), first (light gray) and third (black) quartile of the depth of coverage for the 10 Atlantic males (AT), the 23 individuals from the western Mediterranean sea (14 females (F) and 9 males (WM/WME)), the 19 individuals from the eastern Mediterranean sea (9 males from the south (SEM) and 9 males from the north (NEM)), the 7 hybrids (F1) and the D. punctatus individual (Punc).

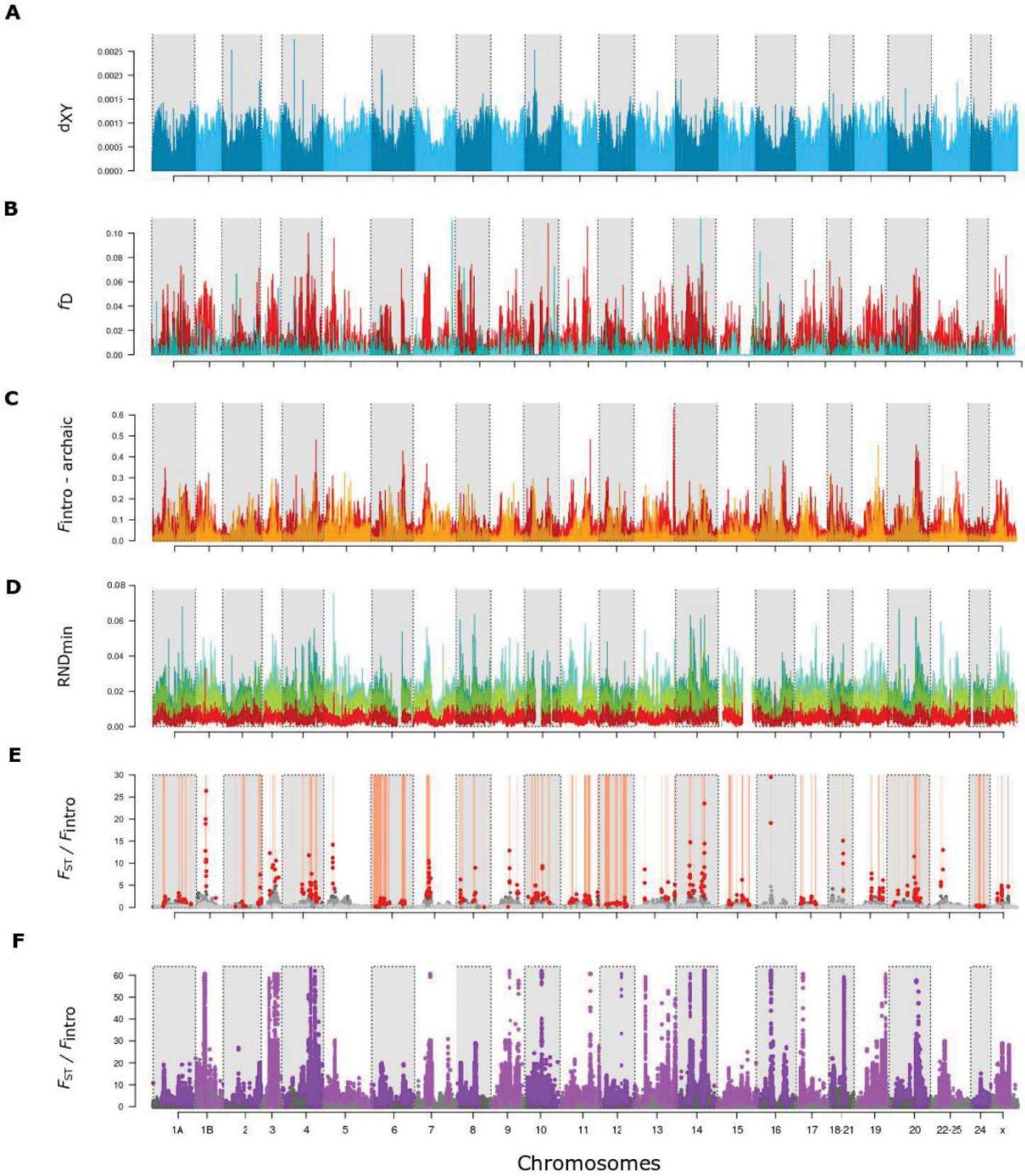
2 – Population structure and phylogenetic analyses



Supplementary Figure 2 **- Principal Component Analysis of the European sea bass population genetic structure**. The analysis was performed on the 52 newly sequenced genomes (colored circles) and the 16 from a previous data set (1) (gray circles with colorful outline). We used the R package adegenet (2) on 91,073 SNPs with a minor allele frequency > 0.4. Individuals originated from four different geographic locations the Atlantic ocean (red, AT), the west (orange, WME), the north-east (dark yellow, NEM) and the south east (light yellow, SEM) of the Mediterranean sea. The first PCA axis explains 39.76% of the total inertia and distinguish the Atlantic and Mediterranean populations while the second PCA axis explains 6.55% of the total inertia and reveals a structure within the Mediterranean population.

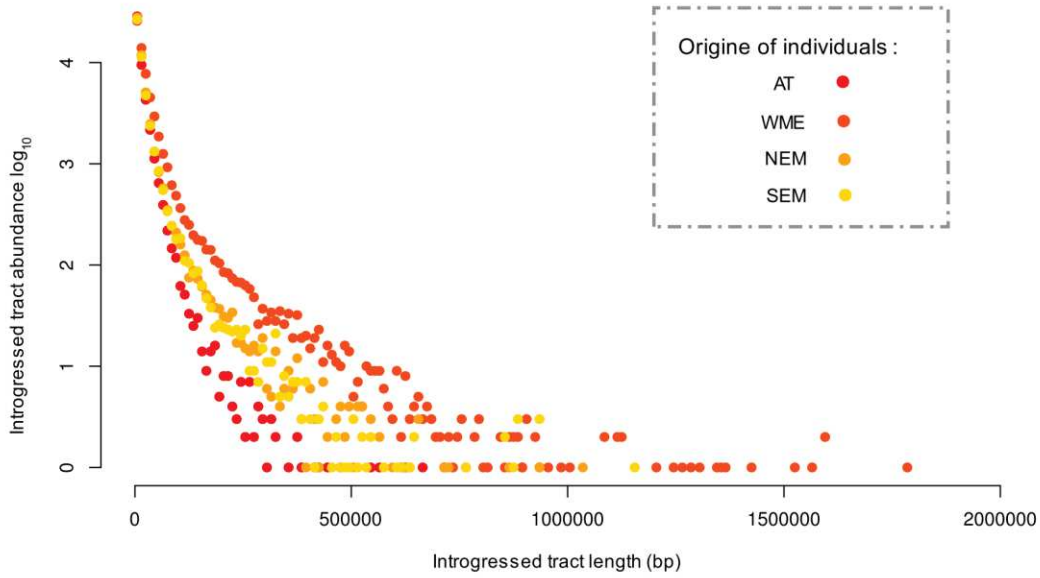


Supplementary Figure 3 –**Consensus trees of the 155,155 Maximum-likelihood trees inferred in 2kb windows along the genome between *M. saxatilis*, *D. punctatus* and Atlantic and Mediterranean *D. labrax* lineages**. There were four different topologies, the most frequent representing the species tree; 94.5% (blue), the second one grouping the Atlantic lineage with *D. punctatus*; 2.87 % (orange), a third one grouping *D. punctatus* with the Mediterranean lineage; 1.68 % (green) and a last one with unresolved relationship between *D. labrax* lineages and *D. punctatus*; 0.05% (not showed).



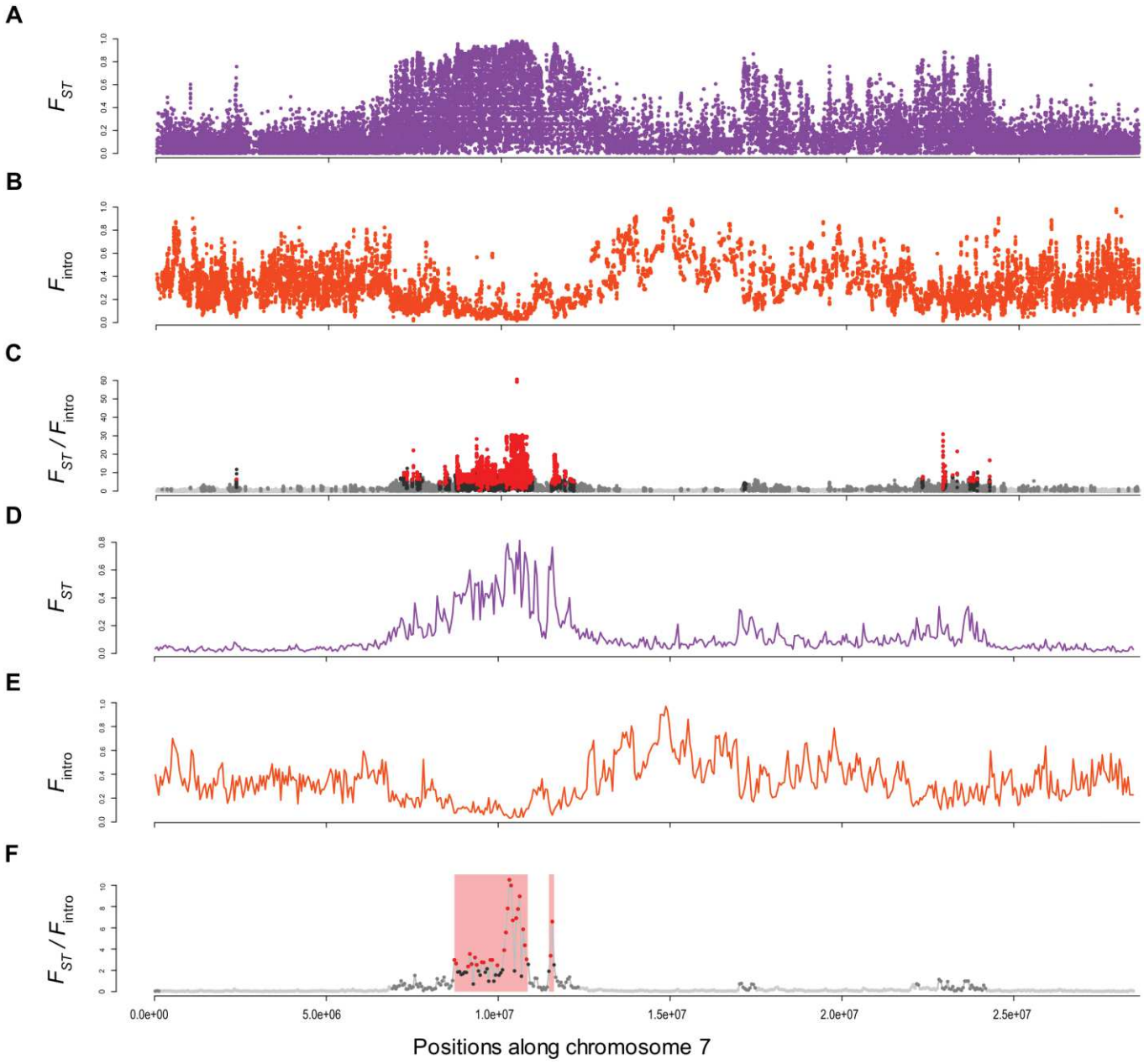
Supplementary Figure 4 – **Statistics measured in non-overlapping 50 kb windows along the genome. A.** d_{XY} measured between the Atlantic and Mediterranean (including eastern and western population) *D. labrax* lineage **B.** f_D statistic measured using in red (((MED, AT), PUN), SAX), in green (((AT, WEST), PUN), SAX) and in blue (((AT, EAST), PUN), SAX). **C.** Fraction of archaic introgressed tracts ($F_{archaic}$) in the eastern Mediterranean and Atlantic population of *D. labrax*. **D.** RND_{min} measured between *D. punctatus* and *D. labrax* Atlantic (red), western (green) and eastern (blue) Mediterranean populations. **E.** Ratio of F_{ST} and F_{intro} used to run the HMM approaches on 50 kb windows that rely on 3 states 1 (light grey), 2 (medium grey) and 3 (dark grey). Red points passed the control for false discovery. We defined island of reproductive isolation as continuous regions containing only red and dark grey points (red boxes). **F.** Ratio of F_{ST} and F_{intro} used to run the HMM approaches on SNPs, purple points are SNPs identified as involved in reproductive isolation that passed the control for false discovery.

4 - Detection of introgressed haplotypes



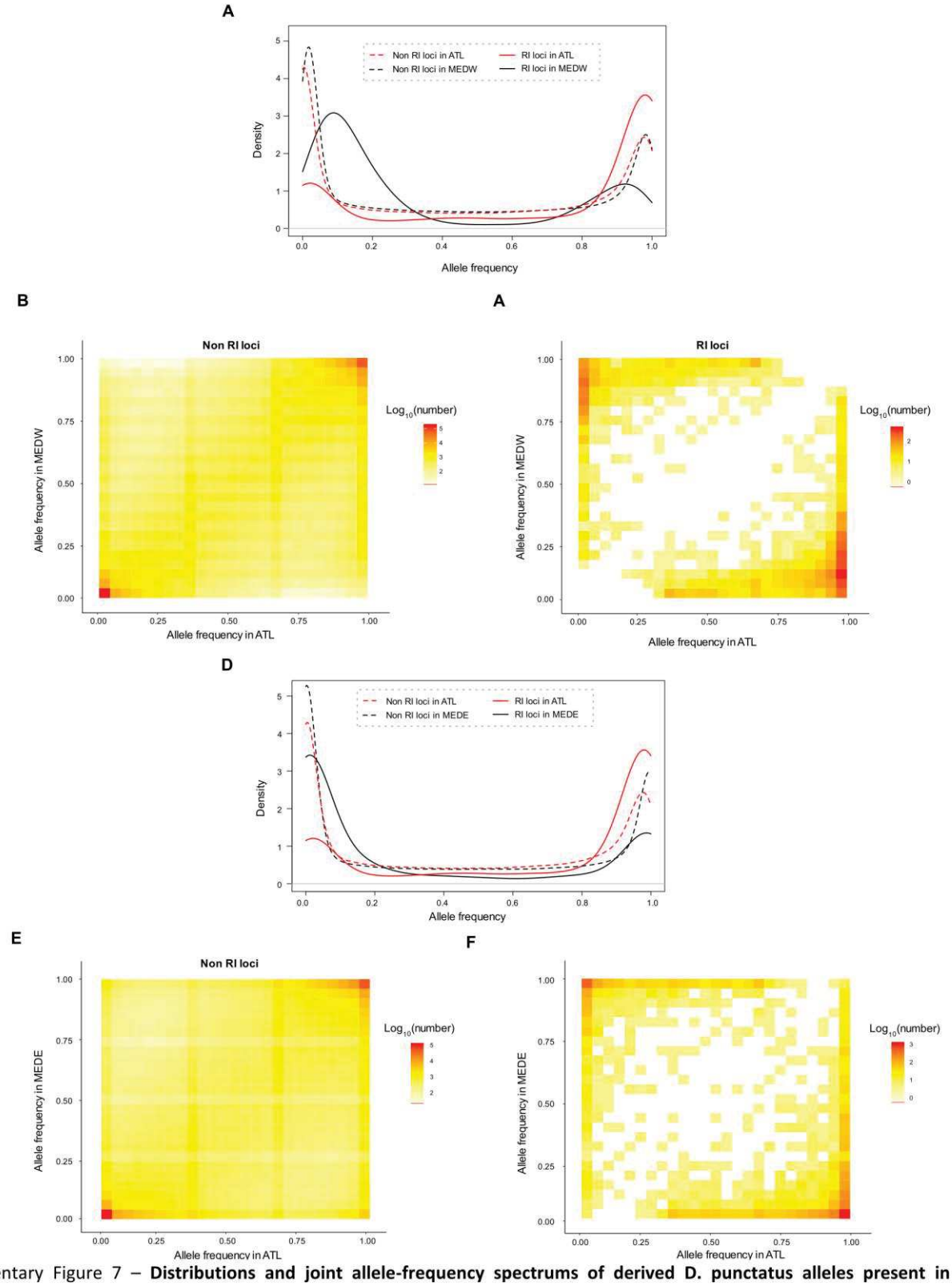
Supplementary Figure 5 – **Introgressed tract length distributions.** Length distributions of Mediterranean tracts introgressed into the Atlantic population (red) and of Atlantic tracts introgressed into the western (orange), north-eastern (dark yellow) and south-eastern (light yellow) Mediterranean populations. Distributions were generated over the whole genome using 11 individuals per population.

5 - Delineation of regions involved in reproductive isolation between *D. labrax* lineages



Supplementary Figure 6 – **Data and results for the SNPs and 50kb window based HMM approach to identify regions involved in reproductive isolation between the two lineage of *D. labrax* along chromosome 7. A.** F_{ST} measured between the Atlantic and western Mediterranean population of *D. labrax* for each SNPs and in every non-overlapping 50 kb windows (**D**). **B.** Fraction of Atlantic tracts introgressed in western Mediterranean genomes (F_{intro}) for each SNPs and in every non-overlapping 50 kb windows (**E**). **C.** Statistic analyzed by the HMM approaches (F_{ST} divided by F_{intro}) for each SNPs and every 50 kb non-overlapping window (**F**). Ratio of F_{ST} and F_{intro} used to run the HMM approaches that rely on 3 states that identify; neutral genomic regions (state 1, light grey), genomic regions under linked selection (state 2, medium grey) and genomic regions involved in reproductive isolation (state3, dark grey). Red points are those that passed the control for false discovery. For the window approach we defined island of RI as continuous regions containing only red and dark grey points (red boxes).

6 – Frequency of introgressed *D. punctatus* tracts



Supplementary Figure 7 – **Distributions and joint allele-frequency spectrums of derived D. punctatus alleles present in D. labrax**. Distribution of D. punctatus derived alleles frequency in AT (red) and WEST (black) (**A**) or East (**D**) *D. labrax* individuals for loci involved (solid line) or not (dashed lines) in reproductive isolation between the two *D. labrax* lineages. **B.** Joint allele-frequency spectrum of derived *D. punctatus* allele for the WEST (31 individuals) and AT (14 individuals) populations for 594,797 SNPs not involved and 7,372 SNPs involved (**C**) in reproductive isolation. **E.** Joint allele-frequency spectrum of derived *D. punctatus* allele for the EAST (23 individuals) and AT (14 individuals) populations for 594,454 SNPs not involved and 7,366 SNPs involved (**C**) in reproductive isolation.

7 – Estimation of the time since introgression between *D. punctatus* and *D. labrax*

Chromosome	*P*	*r*	*a*	T_{admix}
1A	0.062	$2.22e^{-8}$	0.063	94499
1B	0.044	$5.62e^{-8}$	0.060	66152
2	0.037	$4.55e^{-8}$	0.038	105730
3	0.048	$3.25e^{-8}$	0.053	139364
4	0.047	$4.68e^{-8}$	0.058	87390
5	0.047	$4.44e^{-8}$	0.056	93722
6	0.048	$4.38e^{-8}$	0.052	105456
7	0.055	$3.29e^{-8}$	0.051	161342
8	0.051	$4.72e^{-8}$	0.041	131003
9	0.041	$5.55e^{-8}$	0.042	86450
10	0.061	$4.69e^{-8}$	0.058	112733
11	0.046	$4.79e^{-8}$	0.060	79832
12	0.058	$4.39e^{-8}$	0.055	120243
13	0.045	$4.05e^{-8}$	0.058	95632
14	0.061	$5.33e^{-8}$	0.052	110692
15	0.048	$6.08e^{-8}$	0.049	81369
16	0.048	$5.08e^{-8}$	0.057	82769
17	0.048	$4.75e^{-8}$	0.044	114928
18-21	0.059	$3.75e^{-8}$	0.064	123309
19	0.038	$4.85e^{-8}$	0.054	72489
20	0.045	$4.30e^{-8}$	0.058	90803
22-25	0.057	$4.89e^{-8}$	0.057	102440
24	0.082	$9.38e^{-8}$	0.048	90731
X	0.066	$5.63e^{-8}$	0.064	90680

Supplementary Table 2 – values used to estimate T_{admix} for each chromosome.

8 – References

1. Duranton M, Allal F, Fraïsse C, Bierne N, Bonhomme F, Gagnaire P-A. The origin and remolding of genomic islands of differentiation in the European sea bass. Nature Communications. 28 juin 2018;9(1):2518.

2. Jombart T. adegenet: a R package for the multivariate analysis of genetic markers. Bioinformatics. 1 juin 2008;24(11):1403-5.

ANNEXE 3: Matériel supplémentaire de l'article:

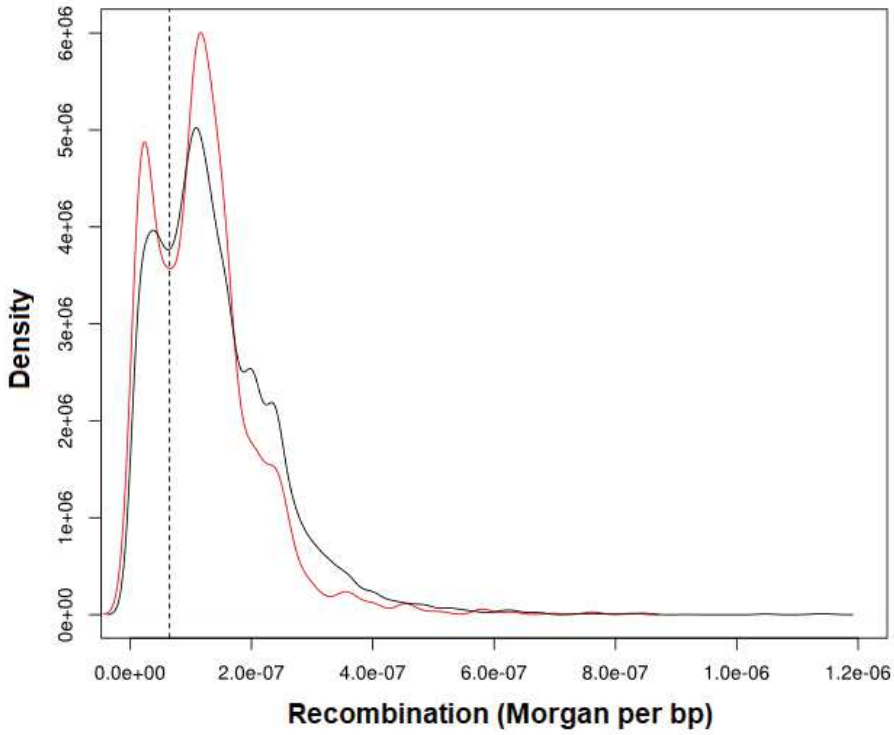
The relation between recombination, reproductive isolation and patterns of molecular evolution in the European sea bass genome.

Supplementary Material

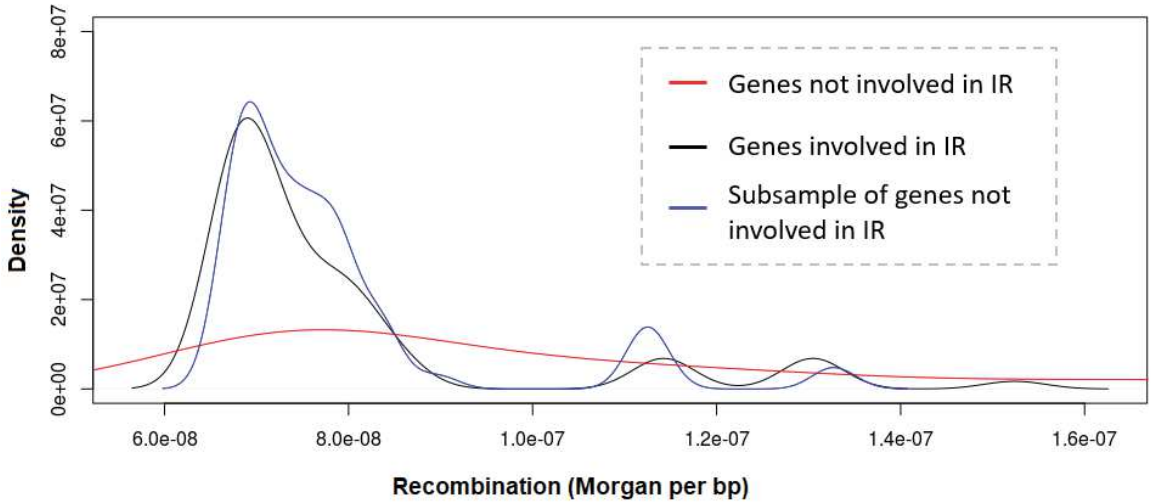
Duranton *et al.*

"The relation between recombination, reproductive isolation and patterns of molecular evolution in the European sea bass genome"

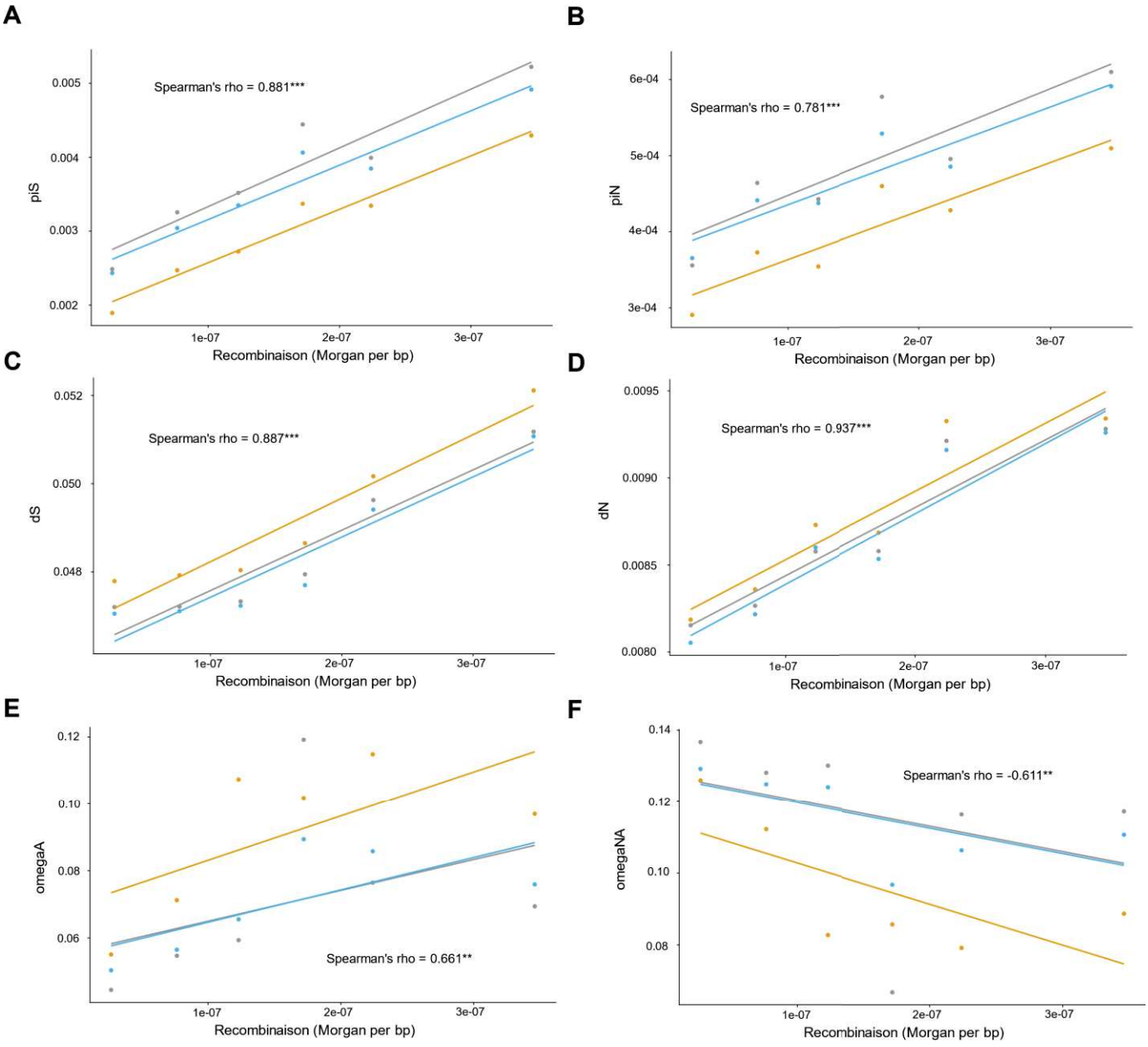
September 2019



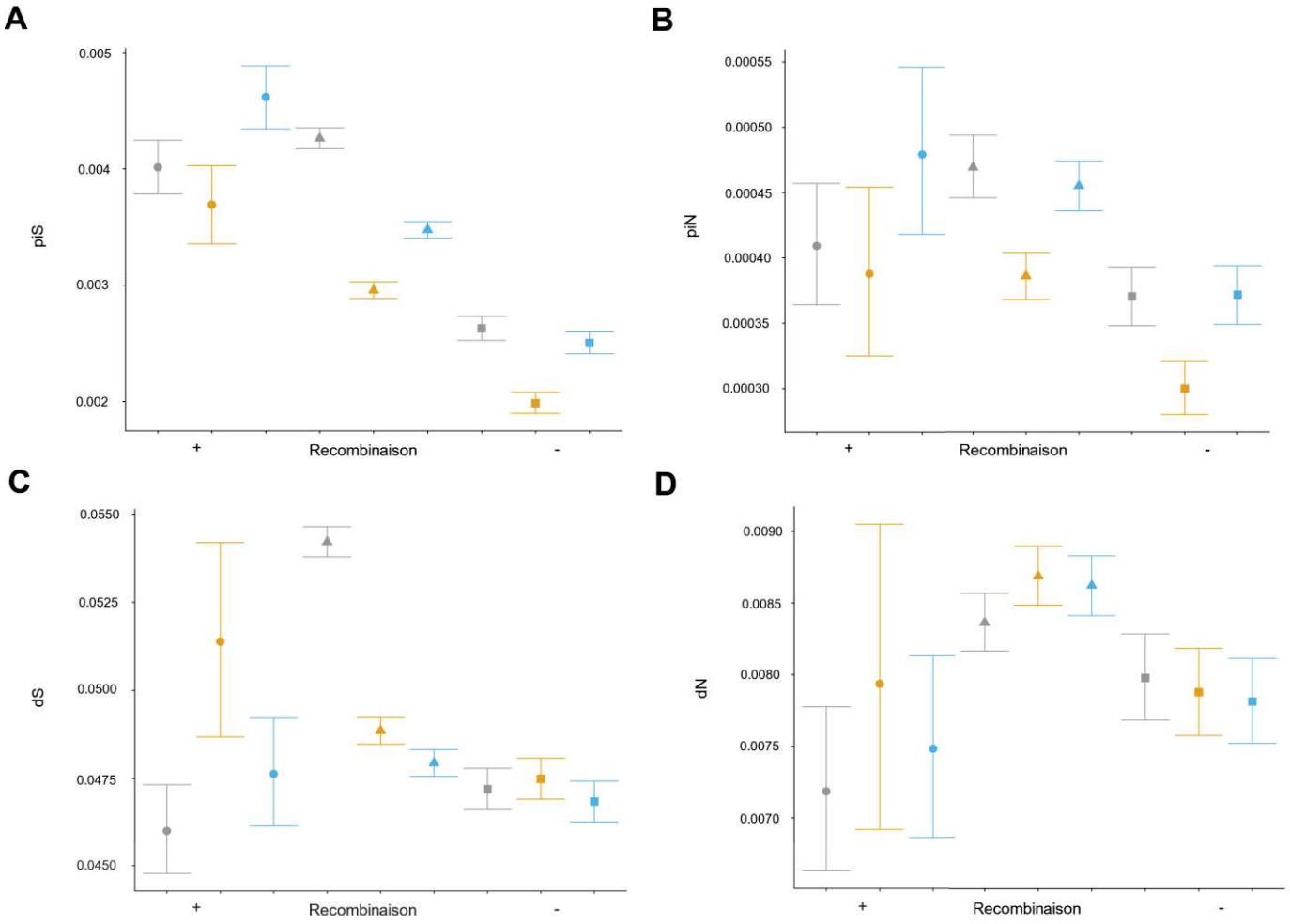
Supplementary Figure 1 – Distribution of recombination rate for gene involved (red) and not involved (black) in RI. The vertical dotted line represents the threshold we fixed to sort regions of low and high recombination.



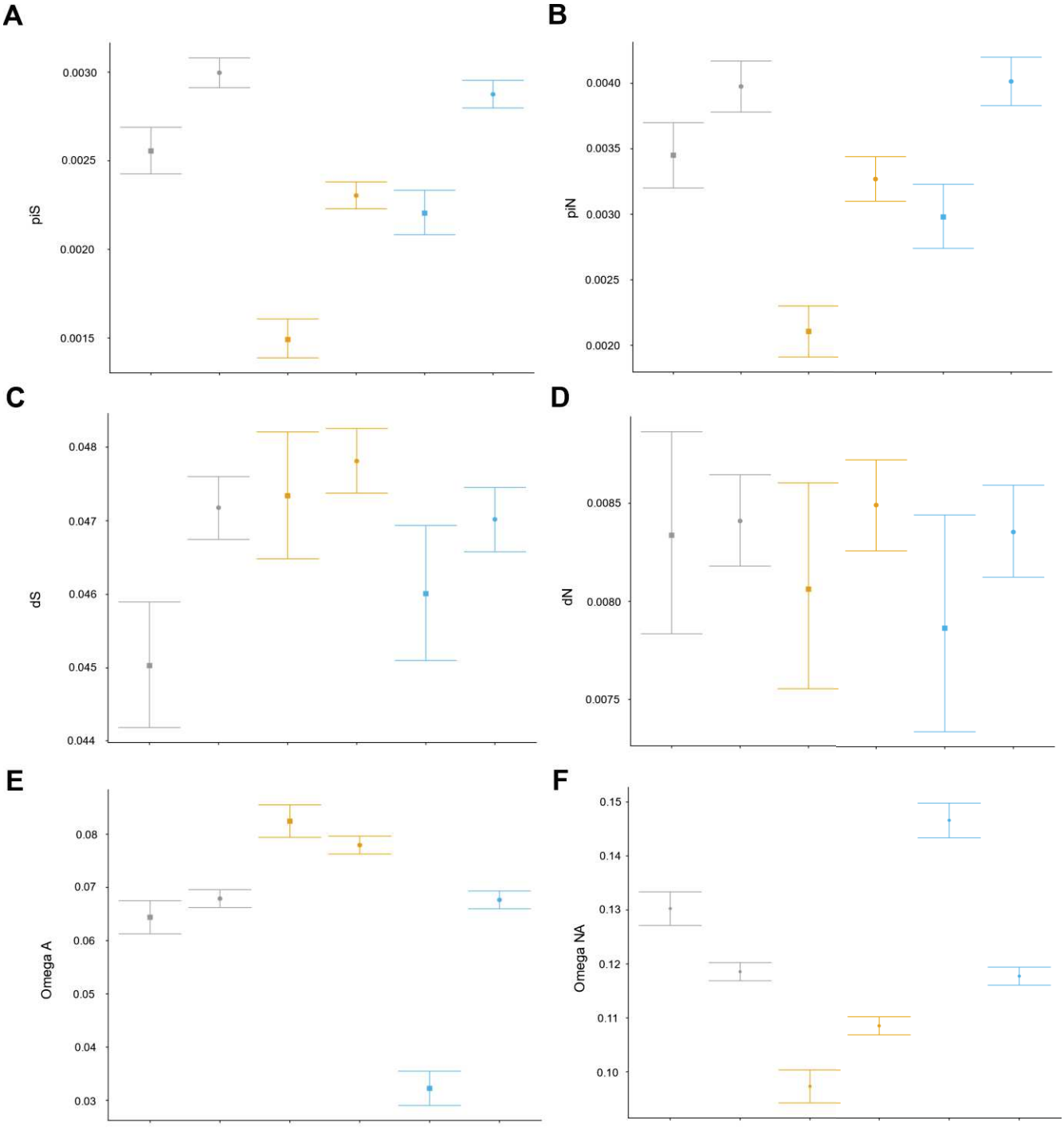
Supplementary Figure 2 – **Example for the chromosome LG1B of the subsampling of gene not involved in RI to reproduce the distribution of recombination rate of genes involved in RI**. Distribution of recombination rate for gene involved (black) and not involved in RI before (red) and after the subsampling (blue).



Supplementary Figure 3 – **Correlation between recombination rate and patterns of molecular evolution.** Measures were computed independently for the Atlantic (ATL; grey), western (MEDW; blue) and eastern-Mediterranean (MEDE; orange) populations. Stars indicate the level of significance of the p-value (** when p-value < 0.01 and *** when p-value < 0.001) for the linear correlation measured including all populations. **A.** piS, **B.** piN, **C.** dS, **D.** dN, **E.** ωa and **F.** ωna.



Supplementary Figure 4 – **Patterns of molecular evolution in recombination hotspots compared to the rest of the genome.** Measures were computed for recombination hotspots (circles), regions of high (triangles) and low (squares) recombination, independently for the Atlantic (ATL; grey), western (MEDW; blue) and eastern-Mediterranean (MEDE; orange) populations. Verticals bars represent the 95% confidence intervals. **A.** piS, **B.** piN, **C.** dS and **D.** dN.



Supplementary Figure 5 – **Comparison of molecular evolution patterns.** Measures were computed independently for the Atlantic (ATL; grey), western (MEDW; blue) and eastern-Mediterranean (MEDE; orange) populations for genes involved (squares) or not (circles) in RI presenting similar recombination rates. Verticals bars represent the 95% confidence intervals. **A.** piS, **B.** piN, **C.** dS, **D.** dN, **E.** ωa and **F.** ωna.

GO	Description	Frequency	Log_{10} p-value	Uniqueness	Dispensability
GO:0007155	cell adhesion	0.544 %	-3.7520	0.92	0.00
GO:0098609	cell-cell adhesion	0.251 %	-3.5591	0.92	0.92
GO:0007610	behavior	0.170 %	-3.2034	0.96	0.00
GO:0007611	learning or memory	0.047 %	-5.3990	0.81	0.00
GO:0007612	learning	0.029 %	-3.1232	0.81	0.95
GO:0007613	memory	0.022 %	-4.1124	0.82	0.94
GO:0009743	response to carbohydrate	0.041 %	-3.2644	0.89	0.00
GO:0022610	biological adhesion	0.550 %	-3.6364	0.96	0.00
GO:0034220	ion transmembrane transport	3.528 %	-8.1739	0.44	0.00
GO:0006820	anion transport	1.956 %	-5.1308	0.48	0.76
GO:0006812	cation transport	3.242 %	-5.7328	0.45	0.82
GO:0030001	metal ion transport	1.677 %	-5.7721	0.45	0.75
GO:0098662	inorganic cation transmembrane transport	1.858 %	-4.7545	0.42	0.83
GO:0098655	cation transmembrane transport	2.290 %	-5.1574	0.42	0.86
GO:0098660	inorganic ion transmembrane transport	2.317 %	-5.7305	0.44	0.78
GO:0015672	monovalent inorganic cation transport	1.824 %	-3.7825	0.45	0.82
GO:0050804	modulation of synaptic transmission	0.057 %	-4.2161	0.77	0.05
GO:0045216	cell-cell junction organization	0.051 %	-3.9393	0.90	0.12
GO:0050996	positive regulation of lipid catabolic process	0.007 %	-3.6840	0.81	0.15
GO:0046321	positive regulation of fatty acid oxidation	0.003 %	-3.1403	0.80	0.80
GO:0042391	regulation of membrane potential	0.135 %	-3.7959	0.84	0.18
GO:0048009	insulin-like growth factor receptor signaling pathway	0.007 %	-3.0017	0.79	0.29
GO:0019933	cAMP-mediated signaling	0.011 %	-3.1232	0.77	0.30
GO:0019935	cyclic-nucleotide-mediated signaling	0.013 %	-3.2526	0.77	0.79
GO:0034330	cell junction organization	0.056 %	-3.2823	0.90	0.32
GO:0007187	G-protein coupled receptor signaling pathway, coupled to cyclic nucleotide second messenger	0.046 %	-3.5482	0.77	0.33
GO:0015732	prostaglandin transport	0.002 %	-3.6144	0.63	0.40
GO:0007166	cell surface receptor signaling pathway	0.920 %	-3.7878	0.74	0.41
GO:0061072	iris morphogenesis	0.001 %	-3.2907	0.87	0.42
GO:0006811	ion transport	5.344 %	-8.6421	0.63	0.46
GO:0034762	regulation of transmembrane transport	0.202 %	-4.4724	0.50	0.46
GO:0032412	regulation of ion transmembrane transporter activity	0.035 %	-3.1931	0.42	0.96
GO:0032409	regulation of transporter activity	0.039 %	-3.3089	0.57	0.74
GO:2001257	regulation of cation channel activity	0.020 %	-3.9788	0.42	0.84
GO:0034765	regulation of ion transmembrane transport	0.197 %	-3.8268	0.38	0.84
GO:0022898	regulation of transmembrane transporter activity	0.036 %	-3.2472	0.51	0.99
GO:0051049	regulation of transport	0.529 %	-3.1451	0.51	0.85
GO:0055085	transmembrane transport	8.916 %	-7.9318	0.61	0.54
GO:0035725	sodium ion transmembrane transport	0.120 %	-4.9788	0.51	0.55

GO:2000969	positive regulation of alpha-amino-3-hydroxy-5-methyl-4-isoxazole propionate selective glutamate receptor activity	0.001 %	-3.2907	0.46	0.57
GO:0007188	adenylate cyclase-modulating G-protein coupled receptor signaling pathway	0.043 %	-3.1244	0.77	0.61
GO:0072511	divalent inorganic cation transport	0.393 %	-3.1701	0.51	0.62
GO:0006813	potassium ion transport	0.476 %	-3.9393	0.48	0.67
GO:0006814	sodium ion transport	0.305 %	-3.6021	0.50	0.72
GO:0071805	potassium ion transmembrane transport	0.331 %	-4.2984	0.44	0.73
GO:0071804	cellular potassium ion transport	0.331 %	-4.2984	0.46	0.93
GO:0050890	cognition	0.053 %	-4.4461	0.86	0.69
GO:0098656	anion transmembrane transport	1.012 %	-3.4034	0.46	0.70
GO:0015711	organic anion transport	1.192 %	-3.2000	0.48	0.85

Supplementary Table 1 – **Enrichment analysis at the level of biological process between genes involved or not in RI.**

GO	Description	Frequency	Log_{10} p-value	Uniqueness	dispensability
GO:0005215	transporter activity	8.494 %	-8.3288	0.92	0.00
GO:0015075	ion transmembrane transporter activity	3.726 %	-9.0458	0.10	0.00
GO:0022857	transmembrane transporter activity	5.870 %	-9.3990	0.18	0.74
GO:0022890	inorganic cation transmembrane transporter activity	1.904 %	-9.0246	0.10	0.81
GO:0022832	voltage-gated channel activity	0.220 %	-3.2882	0.15	0.88
GO:0008509	anion transmembrane transporter activity	1.160 %	-3.9788	0.14	0.76
GO:0005261	cation channel activity	0.299 %	-7.7520	0.10	0.92
GO:0022836	gated channel activity	0.424 %	-7.4498	0.13	0.94
GO:0022804	active transmembrane transporter activity	2.834 %	-5.2418	0.14	0.78
GO:0022838	substrate-specific channel activity	0.634 %	-7.5686	0.09	0.97
GO:0005244	voltage-gated ion channel activity	0.220 %	-3.2882	0.12	0.90
GO:0008324	cation transmembrane transporter activity	2.433 %	-9.3904	0.11	0.84
GO:0005216	ion channel activity	0.624 %	-7.7033	0.08	0.71
GO:0022839	ion gated channel activity	0.022 %	-7.5452	0.23	0.73
GO:0015077	monovalent inorganic cation transmembrane transporter activity	1.394 %	-5.7190	0.11	0.88
GO:0046873	metal ion transmembrane transporter activity	0.999 %	-9.6696	0.12	0.85
GO:0015267	channel activity	0.699 %	-6.6498	0.11	0.98
GO:0015318	inorganic solute uptake transmembrane transporter activity	0.004 %	-8.6326	0.34	0.44
GO:0015293	symporter activity	0.291 %	-3.7122	0.22	0.58
GO:0015294	solute:cation symporter activity	0.184 %	-3.6162	0.18	0.76
GO:0015081	sodium ion transmembrane transporter activity	0.220 %	-6.4802	0.18	0.63
GO:0015079	potassium ion transmembrane transporter activity	0.298 %	-4.2240	0.17	0.76
GO:0022803	passive transmembrane transporter activity	0.699 %	-6.5952	0.21	0.65
GO:0015291	secondary active transmembrane transporter activity	0.949 %	-4.2899	0.18	0.67

Supplementary Table 2 – **Enrichment analysis at the level of molecular functions between genes involved or not in RI.**

ANNEXE 4 : Matériel supplémentaire du Chapitre 4

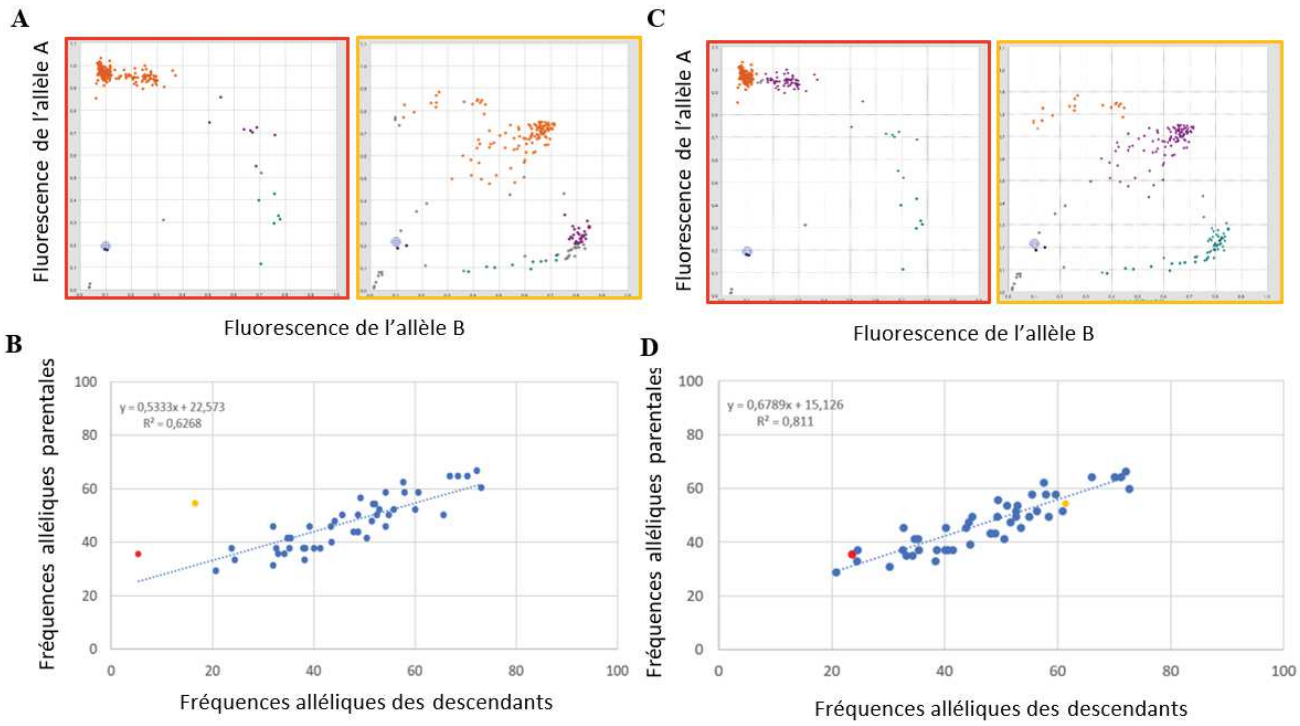


FIGURE ANNEXE 1 – Exemple de correction des données de génotypages sur 2 SNPs parmi les 48. Les panels du haut représentent les graphiques de fluorescence pour l'ensemble des individus sur deux SNPs différents (rouge et jaune) chaque point représente un individu ; et la couleur le génotype auquel il a été assigné AA (orange), AB (violet), BB (turquoise) ou gris s'il n'est pas assigné. Le cercle bleu correspond à la zone où doivent se trouver les contrôles négatifs (points noirs). Les panels du bas représentent la corrélation entre les fréquences alléliques des parents et des descendants sur l'ensemble des 48 SNPs. Les points rouge et jaune correspondent aux SNPs pour lesquels les graphiques de fluorescence sont représentés au-dessus **A.** Assignation automatique avec un seuil de distance de 80%. **B.** Corrélation avant la correction manuelle. **C.** Assignation des individus corrigée manuellement. **D.** Corrélation après correction.

Seuil de probabilité d'assignation	Nombre de descendants simulés assignés	Nombre de descendants assignés aux deux parents	Nombre de descendants assignés à un seul parent	Nombre d'erreurs
0,99	895	478	417	0
0,95	1142	630	512	0
0,75	1419	795	624	0
0,50	1607	871	736	0

TABLEAU SUPPLEMENTAIRE 1 – Résultats de l'étude de la puissance de COLONY2. Sur 2160 simulés nombre d'individus correctement assignés au total, à deux parents, à un parent et nombre d'erreurs d'assignation pour différents seuils de probabilité d'assignation fixés.

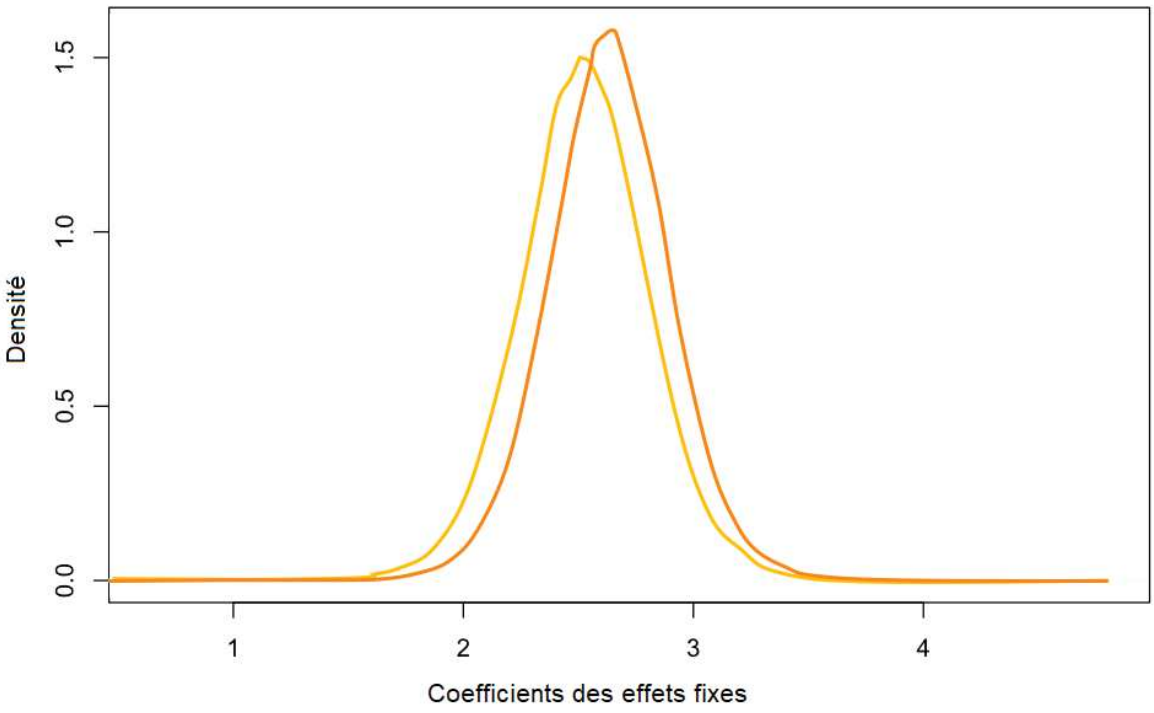


FIGURE ANNEXE 2 – Distribution des effets fixes estimés du type de croisement sur la survie à 1 an estimés par l'approche Bayésienne. La distribution des effets fixes associés aux croisements backcross-MED est représenté en orange et celle des méditerranéens en jaune.

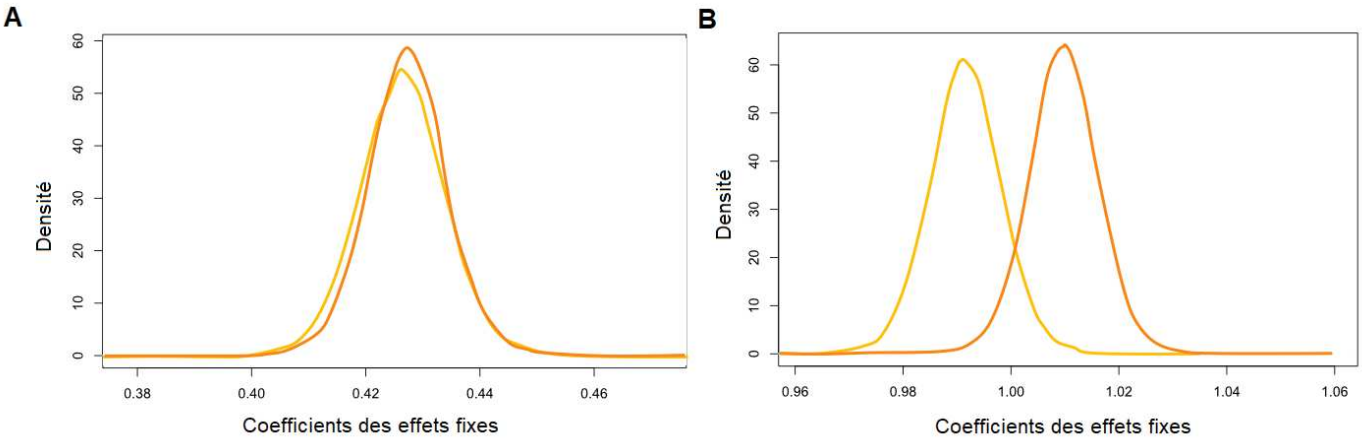


FIGURE ANNEXE 3 – Distribution des effets fixes estimés du type de croisement sur la croissance et la condition mesurées à 1 an estimés par l'approche Bayésienne. La distribution des effets fixes associés aux croisements backcross-MED est représenté en orange et celle des méditerranéens en jaune. **A.** croissance et **B.** condition.

Axe de l'ACP	P-value de l'effet fixe : « type de croisement »		Corrélation avec G		Corrélation avec K	
	Backcross-MED	MED	R^2	P-value	R^2	P-value
1	0,77	0,62	0,18	$0{,}30e^{-67}$	0,05	$0{,}41e^{-18}$
2	0,82	0,43	0,06	$0{,}17e^{-23}$	0,10	$0{,}28e^{-37}$
3	0,17	0,23	0,01	0,61	0,01	$0{,}78e^{-3}$
4	0,49	0,93	0,01	$0{,}50e^{-37}$	0,24	$0{,}50e^{-94}$
5	0,24	0,30	0,01	$0{,}21e^{-1}$	0,07	$0{,}38e^{-26}$
6	0,52	0,81	0,01	$0{,}68e^{-3}$	0,04	$0{,}58e^{-17}$
7	0,46	0,60	0,02	$0{,}45e^{-7}$	0,01	$0{,}70e^{-6}$

TABLEAU ANNEXE 3 – Association des variables type de croisement, croissance (G) et condition (K) aux 7 premiers axes de l'ACP grâce à des GLMM ajustés par approche bayésienne ou des corrélations linéaires.

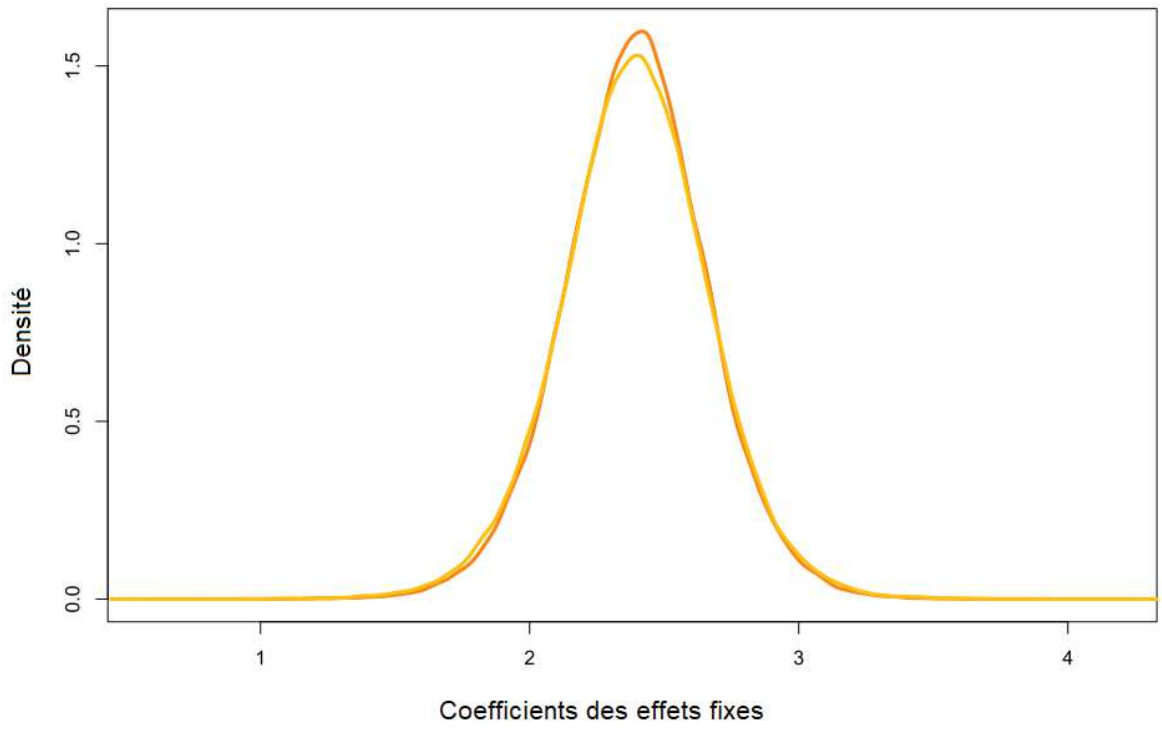


FIGURE ANNEXE 4 – Distribution des effets fixes estimés du type de croisement sur la survie à 2 an estimés par l'approche Bayésienne. La distribution des effets fixes associés aux croisements backcross-MED est représenté en orange et celle des méditerranéens en jaune.

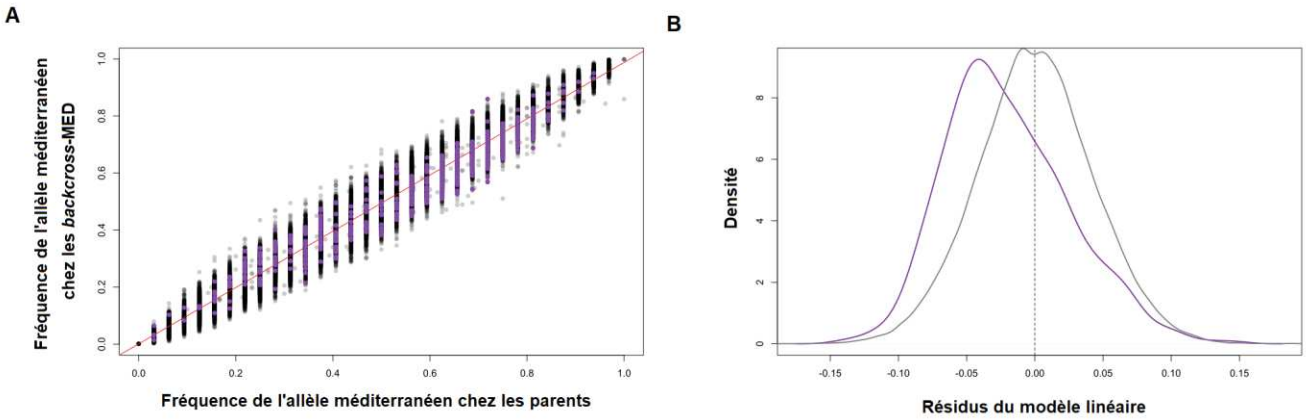


FIGURE ANNEXE 5 – Corrélation linéaire des fréquences alléliques parentales et des *backcross*-MED sur 49993 SNPs dont 762 impliqués dans l'isolement reproductif. A. Chaque point représente un SNPs impliqué (violet) ou non (gris) dans l'isolement reproductif. La droite rouge représente la régression linéaire entre les fréquences alléliques. **B.** Distribution des écarts au modèle pour les SNPs impliqués (violet) ou non (gris) dans l'isolement reproductif.

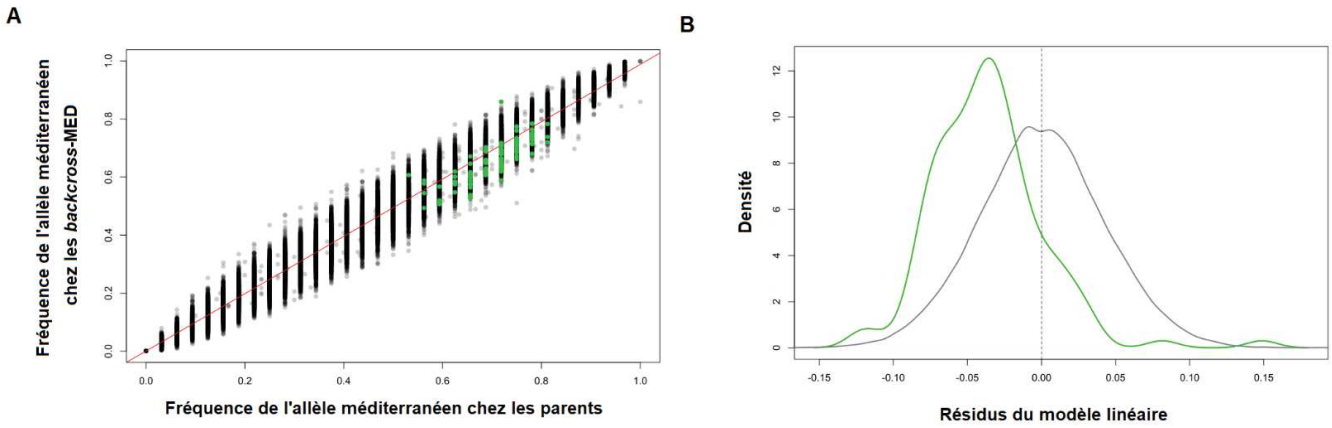


FIGURE ANNEXE 6 – Corrélation linéaire des fréquences alléliques parentales et des *backcross*-MED sur 49993 SNPs dont 127 impliqués dans l'isolement reproductif et différentiellement fixés entre les populations naturelles atlantiques et est-méditerranéennes. A. Chaque point représente un SNPs non impliqué (gris) ou impliqué dans l'IR et différentiellement fixé entre les populations naturelles atlantiques et est-méditerranéennes (vert). La droite rouge représente la régression linéaire entre les fréquences alléliques. **B.** Distribution des écarts au modèle pour les SNPs non impliqués (gris) ou impliqués dans l'IR et différentiellement fixé entre les populations naturelles atlantiques et est-méditerranéennes (vert).

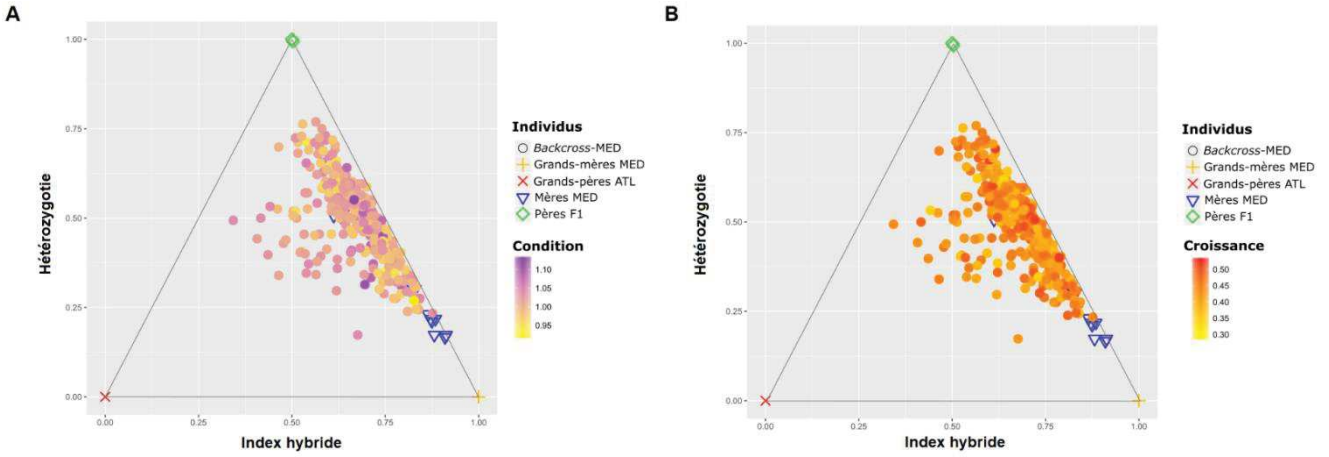


FIGURE ANNEXE 7 – Etude du lien génotype-phénotype chez 380 *backcross*-MED sur 156 SNPs impliqués dans l'isolement reproductif et différentiellement fixés entre les grands-parents atlantiques et méditerranéens des *backcross*-MED. Les *backcross*-MED (ronds) ainsi que leurs ascendants, leurs grand-mères et mères d'origine méditerranéenne (respectivement croix jaunes et triangles bleus), leurs grands-pères d'origine atlantique (croix rouges) et leurs pères hybrides de première génération (losanges verts) sont représentés dans un espace triangulaire indiquant en abscisse leur indice hybride (c'est-à-dire la proportion d'allèles d'origine méditerranéenne) et en ordonnés leur niveau d'hétérozygotie. La couleur des ronds indique **A.** la condition et **B.** la croissance des *backcross*-MED à 1 ans.

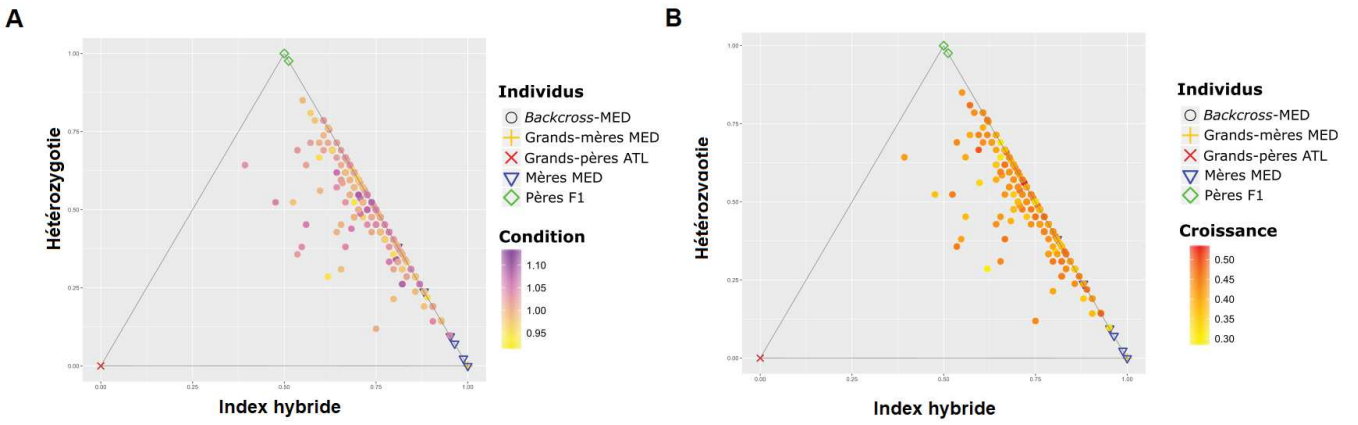


FIGURE ANNEXE 8 – Etude du lien génotype-phénotype chez 299 *backcross*-MED sur 42 SNPs impliqués dans l'isolement reproductif et présentant des fortes valeurs de F_{ST} entre les géniteurs atlantiques et méditerranéens. Les *backcross*-MED (ronds) ainsi que leurs ascendants, leurs grand-mères et mères d'origine méditerranéenne (respectivement croix jaunes et triangles bleus), leurs grands-pères d'origine atlantique (croix rouges) et leurs pères hybrides de première génération (losanges verts) sont représentés dans un espace triangulaire indiquant en abscisse leur indice hybride (c'est-à-dire la proportion d'allèles d'origine méditerranéenne) et en ordonnés leur niveau d'hétérozygotie. La couleur des ronds indique **A.** la condition et **B.** la croissance des *backcross*-MED à 2 ans.

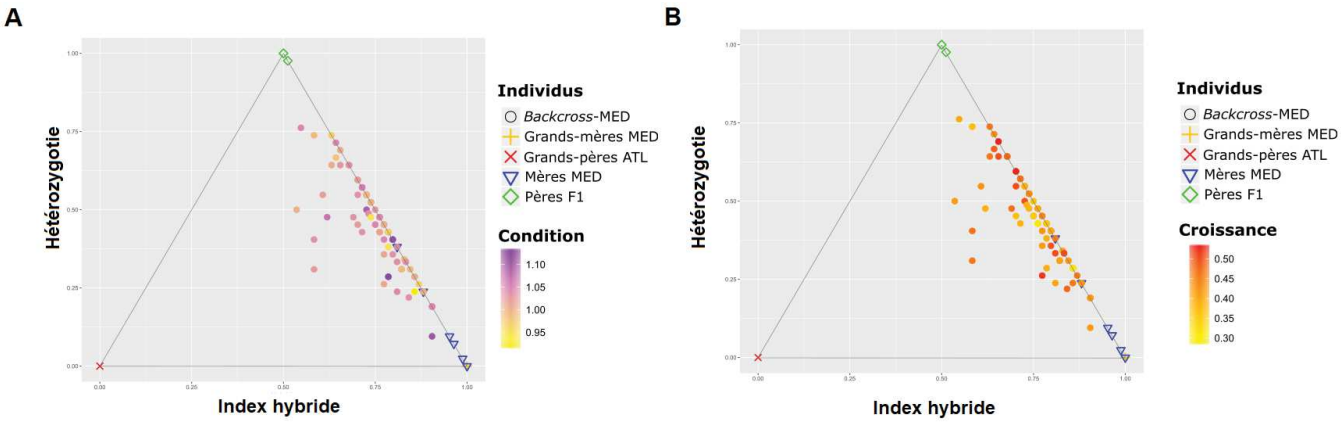


FIGURE ANNEXE 9 – Etude du lien génotype-phénotype chez 81 *backcross*-MED morts entre la première et deuxième année sur 42 SNPs impliqués dans l'isolement reproductif et présentant des fortes valeurs de F_{ST} entre les géniteurs atlantiques et méditerranéens. Les *backcross*-MED (ronds) ainsi que leurs ascendants, leurs grand-mères et mères d'origine méditerranéenne (respectivement croix jaunes et triangles bleus), leurs grands-pères d'origine atlantique (croix rouges) et leurs pères hybrides de première génération (losanges verts) sont représentés dans un espace triangulaire indiquant en abscisse leur indice hybride (c'est-à-dire la proportion d'allèles d'origine méditerranéenne) et en ordonnés leur niveau d'hétérozygotie. La couleur des ronds indique **A.** la condition et **B.** la croissance des *backcross*-MED à 1 ans.

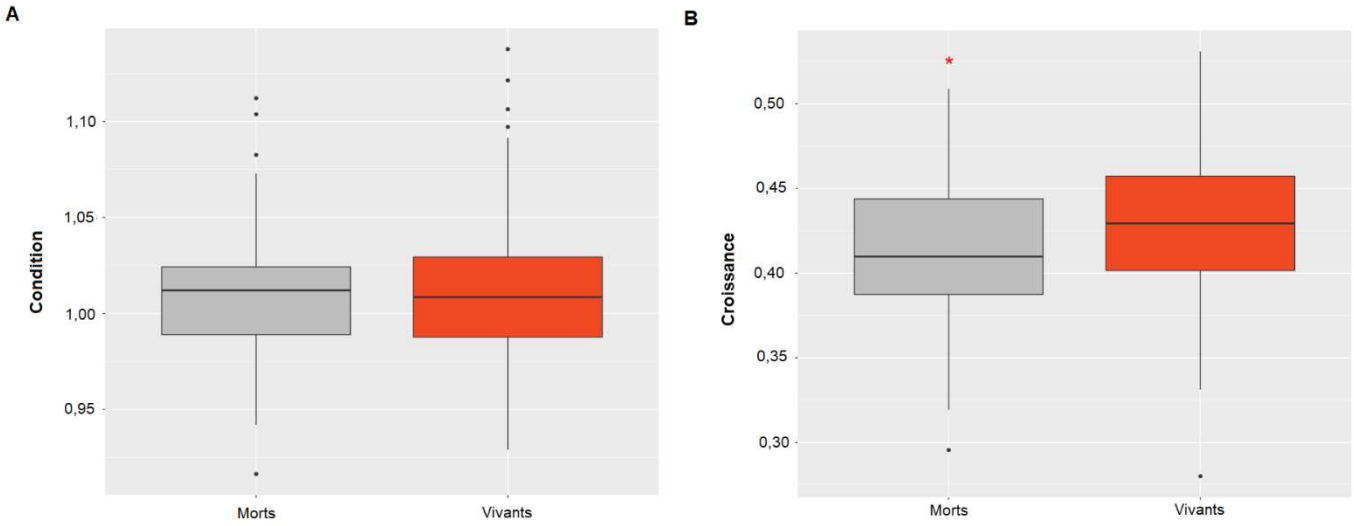


FIGURE ANNEXE 10 – Comparaison de la croissance et de la condition à 1 an des backcross-MED survivants et morts durant leurs deuxièmes années de vie. La couleur orange correspond aux individus survivants (299) et la grise aux morts (81). L'étoile rouge indique que la différence entre les deux est significative (p<0,5). **A.** Condition et **B.** croissance des *backcross*-MED à 1 ans.

ANNEXE 5: Importance de considérer la sélection en liaison dans les inférences démographiques, article:

Digest: Demographic inferences accounting for selection at linked sites.

doi:10.1111/evo.13

Digest: Demographic inferences accounting for selection at linked sites†

Alexis Simon[1,*] **and Maud Duranton**[1,2,*]

[1]*Institut des Sciences de l'Evolution-Montpellier, Université de Montpellier, CNRS-IRD-EPHE-UM, France*

[2]*E-mail: maud.duranton@umontpellier.fr*

Received April 18, 2018
Accepted May 7, 2018

Complex demography and selection at linked sites can generate spurious signatures of divergent selection. Unfortunately, many attempts at demographic inference consider overly simple models and neglect the effect of selection at linked sites. In this issue, Rougemont and Bernatchez (2018) applied an approximate Bayesian computation (ABC) framework that accounts for indirect selection to reveal a complex history of secondary contacts in Atlantic salmon (*Salmo salar*) that might explain a high rate of latitudinal clines in this species.

Identifying signatures of selection within genomes is a long-standing objective in evolutionary biology. This quest is complicated by factors such as species' demographic history, which has long been recognized for its potential to produce footprints that mimic selection. Secondary contacts, for instance, can generate clines in allele frequencies correlated with spatial variation in ecological factors that may be interpreted incorrectly as representing local adaptation to an environmental gradient (Fig. 1). More recently, selection at linked sites has received particular attention as another confounding mechanism that generates genomic regions of increased divergence between populations as expected in the presence of divergent selection (Fig. 1B). Furthermore, selection at linked sites seems to affect a sufficiently large fraction of the genome of many species and challenges the basic assumption that most loci evolve neutrally. Despite those issues, few studies have attempted to include the effects of selection at linked sites in demographic divergence models (Roux et al. 2016; Rougeux et al. 2017).

In this issue, Rougemont and Bernatchez (2018) present an inference of the demographic history of Atlantic salmon (*Salmo salar*) populations using a dataset of 5034 SNPs genotyped in 77 populations covering the whole species' range. They first reassessed the population structure and confirmed the existence of genetic subdivision among regions. They then used an approximate Bayesian computation (ABC) method to compare demographic scenarios with various histories of gene flow, including divergence with migration, ancient migration, and secondary contact.

ABC avoids the explicit computation of likelihood, allowing users to choose the most appropriate summary statistics to be combined for model selection. This approach allows users to evaluate more complex and flexible models compared to other inference methods. Furthermore, ABC can be coupled with machine learning, which is well adapted to high-dimensional data. Although selection was not explicitly simulated in this study, the implemented models build on the theoretical predictions that selection against introgression locally reduces the effective migration rate between populations (Barton and Bengtsson 1986), whereas selection at linked sites enhances the effect of drift by locally reducing the effective population size (Charlesworth et al. 1993). Even if selection at linked sites can be either positive (through selective sweeps) or negative (through background selection), these two mechanisms both reduce neutral nucleotide diversity over time.

Using this modeling framework, Rougemont and Bernatchez (2018) provide a new perspective on the evolutionary history of a widely studied system in which a complex demographic history

*These authors are cofirst authors.

†This article corresponds to Rougemont Q., L. Bernatchez. 2018 The demographic history of Atlantic Salmon (*Salmo salar*) across its distribution range reconstructed from Approximate Bayesian Computations. Evolution. https://doi.org/10.1111/evo.13486.

© 2018 The Author(s). *Evolution* © 2018 The Society for the Study of Evolution.

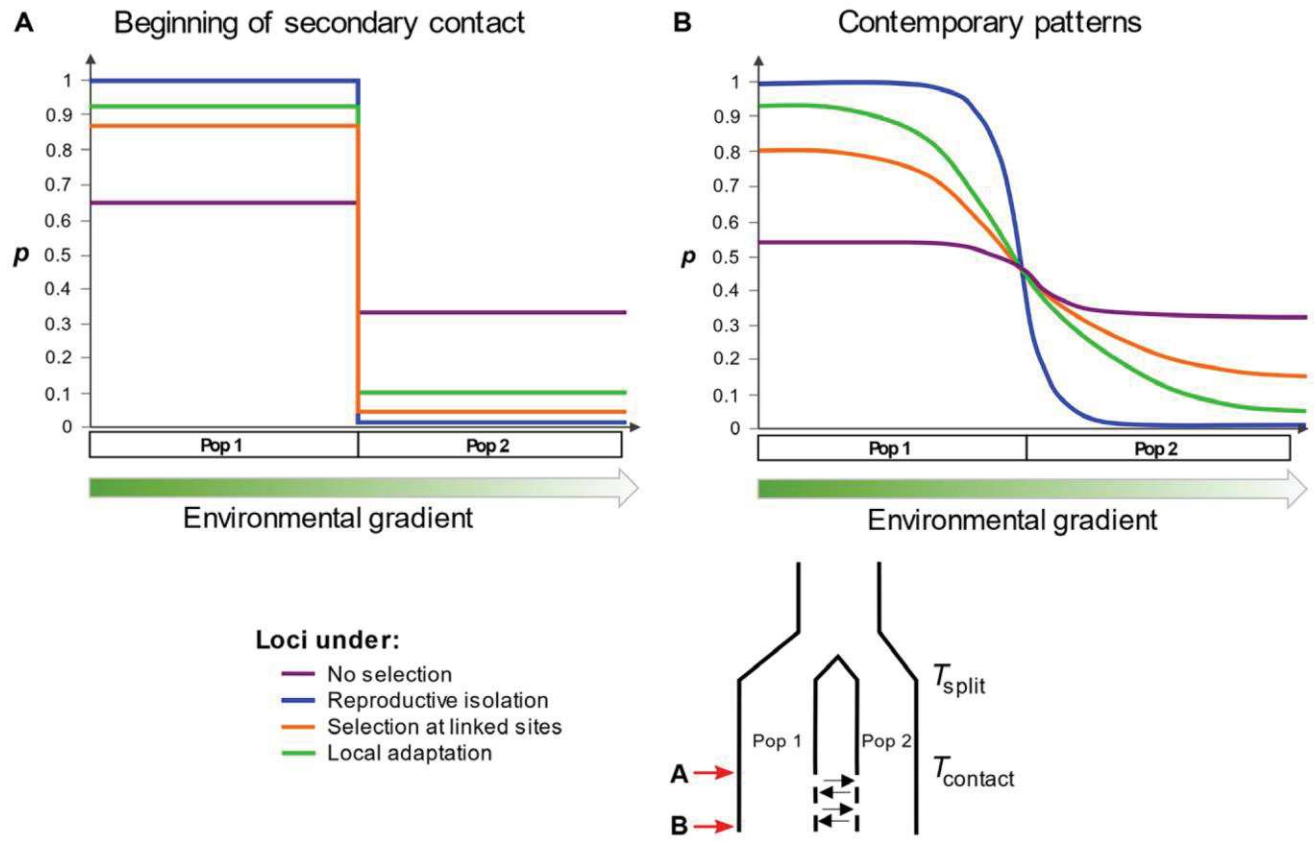


Figure 1. The fate of polymorphisms affected by different processes before and after admixture. Each color represents a different locus affected by a specific process during divergence (i.e., between T_{split} and $T_{contact}$) and secondary contact. Plots represent allele frequency clines across space, where populations 1 and 2 occupy different parts of an environmental gradient. The red arrows point to the time when cline patterns are observed. Since they are evolving under different processes, loci reach various levels of divergence and experience different levels of gene flow. (A) During divergence, differentiation builds up by either selection or genetic drift. Local adaptation (green) and reproductive isolation (blue) alleles rapidly increase in frequency in one population and decrease in the other. Loci under selection at linked sites (orange) diverge more rapidly than neutral ones (violet) because linked selection locally reduces the effective population size, which increases genetic drift. (B) During the contact, gene flow rapidly erodes the differentiation of the neutral and linked selected loci, but the cline of the second is steeper because it was more differentiated. Local adaptation loci experience reduced gene flow, and their clines are fixed by the environment. The reproductive isolation cline rapidly converged with the local adaptation cline. Clines of reproductive isolation and selection at linked sites can therefore be mistaken for local adaptation if these processes have not been taken into account when performing demographic inferences.

may confound the detection of local adaptation. A scenario of secondary contacts was favored both between and within American and European Atlantic salmon populations, providing strong evidence for a reticulated demographic history. This work emphasizes the importance of identifying a null model of demographic history for a given species that accounts for both population structure and selection at linked sites. Future studies seeking to identify signatures of divergent selection in natural populations will need to take into account variation in effective population size both in time and along the genome. Finally, this study illustrates the flexibility of ABC demographic inferences to account for dataset particularities such as the ascertainment bias in SNP chip arrays, confirming that this approach could be widely applied to many nonmodel species.

Future studies would benefit from going a step further toward modeling the whole picture of the demographic history by considering more populations simultaneously. This goal could be achieved within an ABC framework, because population structure and phylogeography of Atlantic salmon are now well known. However, for species for which there is no a priori knowledge about population history, considering a large number of populations (meaning an exponentially large number of alternative scenarios) can quickly become intractable in terms of computational time. One way forward could be to use non-ABC methods with

minimal assumptions (e.g., Minimal-Assumption Genomic Inference of Coalescence, Weissman and Hallatschek 2017; or Approximate Blockwise Likelihood Estimation, Reddy et al. 2017). Such methods would capture the key aspects of populations' history and provide the starting point for further refinements. Those methods also have the potential to incorporate the aforementioned confounding factors such as selection at linked sites. In conclusion, an important prerequisite before starting to uncover signatures of divergent selection is to identify a relevant null model of demographic history accounting for indirect selective effects.

LITERATURE CITED

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

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

Reddy, C. B., M. J. Hickerson, L. A. Frantz, and K. Lohse. 2017. Blockwise site frequency spectra for inferring complex population histories and recombination. bioRxiv, 077958. https://doi.org/10.1101/077958.

Rougemont, Q., and L. Bernatchez. 2018. The demographic history of Atlantic Salmon (*Salmo salar*) across its distribution range reconstructed from approximate Bayesian computations. Evolution. https://doi.org/10.1111/evo.13486.

Rougeux, C., L. Bernatchez, and P. A. Gagnaire. 2017. Modeling the multiple facets of speciation-with-gene-flow toward inferring the divergence history of lake whitefish species pairs (*Coregonus clupeaformis*). Genome Biol. Evol. 9:2057–2074.

Roux, C., C. Fraïsse, J. Romiguier, Y. Anciaux, N. Galtier, and N. Bierne. 2016. Shedding light on the grey zone of speciation along a continuum of genomic divergence. PLOS Biol. 14. https://doi.org/10.1371/journal.pbio.2000234.

Weissman, D. B., and O. Hallatschek. 2017. Minimal-assumption inference from population-genomic data. Elife 6:e24836.

Associate Editor: K. Moore
Handling Editor: M. Noor

SUBMIT A DIGEST

Digests are short (~500 word), news articles about selected original research included in the journal, written by students or postdocs. These digests are published online and linked to their corresponding original research articles. For instructions on Digests preparation and submission, please visit the following link: https://sites.duke.edu/evodigests/.

REMERCIEMENTS

Je vais commencer en remerciant les deux personnes sans qui cette thèse n'aurait pas été possible : François et Pierre-Alexandre. François, merci d'avoir été mon directeur de thèse, même si nos échanges n'ont pas forcément été très fréquents, tu as toujours été présent pour m'apporter ton soutient et eu ce regard bienveillant sur cette thèse. Merci pour toutes les discussions que nous avons pu avoir à deux comme à trois, et d'avoir su me rassurer sur mon avenir dans le domaine de la recherche. Merci aussi pour les petits cours privés de génétique des populations, je n'aurai jamais su que le F_{ST} est en réalité un indice de fixation sans toi =) .

Pierre-Alexandre, merci pour ton implication et ton enthousiasme permanent, à tes côtés chaque petit résultat semble être une découverte majeure, difficile de ne pas aimer ce que l'on fait. Merci pour ta présence et ton soutient, tu m'as progressivement laissé de plus en plus de liberté ce qui m'a permis de prendre confiance en moi. Même si c'est parfois difficile de te suivre vu le nombre d'idées que tu peux avoir à la seconde, tu as toujours su m'écouter et j'ai toujours senti que mes idées étaient considérées. Merci pour toutes les discussions que l'on a pu avoir quasiment jusqu'à la veille du rendu de ce manuscrit, tu as toujours pris le temps de m'expliquer les choses et j'en ai appris plus à tes côtés que tout ce que j'aurai pu lire pendant trois ans. Merci aussi pour ta patience à corriger mes fautes d'orthographes répétitives (Dicentrachus, une dernière pour la route ^^). Au final, même s'il y a eu des moments plus difficiles que d'autres et bien que j'ai finalement touché bien plus d'écailles que ce que tu m'avais promis, j'ai passé trois super années de thèse et c'est en grande partie grâce à toi.

Je souhaiterai également remercier les membres de mon jury de thèse, Carole Smadja, Tatiana Giraud, Laurence Després et Christophe Lemaire (Krostif, J'espère avoir bien repris le flambeau des études de génétique des populations du bar européen !). Merci d'avoir accepté de faire partie de mon jury et de vous être intéressé à mes travaux. Je remercie aussi les membres de mon comité de thèse Pierre Boursot, François Allal, Frédérique Viard, Nathalie Charbonnel, Benoît Nabholz et Nicolas Galtier pour vos conseils et les discussions qui m'ont permis d'avoir de nouvelles idées.

Etant donné que mes travaux sur le bar ont, en réalité, commencé et continué pendant ma première année de thèse à la station marine de Sète, je souhaiterai remercier toutes les personnes que j'ai croisé là-bas et qui ont contribué à la bonne ambiance dans la station. Merci à l'équipe de l'autre côté du couloir : Sophie, Cathy, Miriam, Diane, Joana, Noémie et des remerciements spéciaux pour Florent et tes petits messages de soutien venu de l'autre bout du monde. Merci aussi à Marianne, Florentine pour tes talents de monteuse vidéo et à Nicolas qui restera pour moi toujours associé à la station de Sète. Nico merci pour ta bienveillance, tes conseils et ta motivation pour les '*journal club'*. Enfin des remerciements particuliers pour Ahmed, c'était un plaisir de partager un bureau/laboratoire avec toi, tu as été mon binôme et mon soutien durant ces débuts de thèse, la vie à la station n'aurait pas été la même sans toi, même si je n'ai toujours pas compris pourquoi il te fallait 30 minutes pour aller chercher ta veste deux étages plus haut ^^.

Je voudrai également remercier les personnes de la station de Palavas qui se sont, de près ou de loin, occupées de la bonne santé des poissons issus de mes croisements expérimentaux, notamment Marie-Odile, François, Stéphane ; et à ceux qui m'ont aidé pour les longues chaînes de biométrie, comme Ronan et Sara. Un énorme merci à Alain, rien de tout ce qui a été fait sur les croisements expérimentaux n'aurait été possible sans toi. Merci

pour ton expertise et ta gentillesse envers moi qui n'avais jamais touché un poisson de ma vie. J'en profite pour remercier une nouvelle fois François Allal, cette thèse n'aurait pas pu se dérouler aussi bien sans toi.

Je souhaiterai également remercier toutes les personnes de l'ISEM, c'était un plaisir de pouvoir faire cette thèse au sein de ce laboratoire. Remerciements spéciaux pour les voisins du bâtiments 24, Rémy, Jimmy, Khalid, Fred, Erick, Nelly, Emilly, Stephen, Sophie, Corine et Véronique merci pour la bonne ambiance dans le couloir et les parties de mots croisés le midi, même si j'ai encore beaucoup de progrès à faire. Plus particulièrement merci à Cathy qui a toujours été disponible quand j'avais des problèmes de codes ou tout simplement besoin de parler.

Merci également à tous les doctorants que j'ai pu croiser pendant ces trois années de thèse et qui ont tous, chacun à leur manière, su m'apporter leur soutien. Un grand merci en premier à Alexis, on a commencé ensemble et on a fini aussi (presque) ensemble. Merci d'avoir été le calme dans mes tempêtes pendant quasiment quatre ans. Moi qui m'angoisse si facilement, c'était plutôt agréable d'avoir quelqu'un qui sait garder son calme juste derrière moi. J'en profite aussi pour remercier Alison, merci pour ta bonne humeur, pour avoir organisé la première (mythique) soirée plage, pour m'avoir toujours écouté quand j'avais besoin de me plaindre et pour l'organisation du voyage en Norvège. Un très gros merci à Maeva. Au final, on aura vraiment passé qu'une seule année ensemble mais remplie de bons moments. De Molène à Zurich, en passant par l'ESEB (à Groningen et Montpellier) et Amsterdam, je ne garde que de bons souvenirs avec un léger parfum de rhum arrangé ^^. Merci également à Manon et Marjolaine (l'une ne va pas sans l'autre). Merci pour tous les bons moments en dehors du labo, du block out aux soirées qui finissent bien évidement en dansant sur de la trans. Merci plus particulièrement à Marjo car je ne sais pas si j'aurai vraiment compris ce qu'est l'évolution moléculaire sans toi, et à Manon d'avoir été une oreille attentive et un soutien quand j'en avais tant besoin. Enfin, un grand Merci à Maurine qui a su écouter et comprendre toutes mes personnalités =) Merci d'être l'organisatrice officielle de ma journée de thèse, pour les nombreuses pauses café à discuter et pour me soutenir jusqu'au bout (tu feras aussi la bise à tes zouzs). Merci aussi à Pierre Barry, ton arrivé dans l'équipe aura apporté un brin de zénitude dans le bureau. Merci à Natalia pour la bonne humeur que tu as apporté lors de ton petit passage dans l'équipe, grâce à toi je pourrai dire que j'ai appris à parler anglais avec une espagnole. Merci aussi à Flavià et Manon qui, au travers de leurs stages, m'ont bien aidé à avancer dans ma thèse. Et enfin merci à tous les autres pour la bonne ambiance pendant les weekends d'intégration et le reste, Adrien, le grand et le petit Yoann, Clémentine, Rémi, Quentin, Maxime, Cécille, Myriam, Julien, Maeva, Camille et Alan pour les discussions plus ou moins scientifiques autour d'un verre.

Je voudrai finir en remerciant ma famille pour avoir été ce point de repère qui m'a permis de garder en tête que, malgré tout l'investissement que j'ai mis dans cette thèse, ce n'était qu'une partie et une étape de ma vie, et qu'aucune difficulté n'est insurmontable. Un merci tout particulier à Julie. Je n'en serais probablement pas là sans toi. Après tout, c'est principalement pour papoter avec toi sur les bancs de la fac que je me suis inscrite en première année de bio ! Merci pour tous ces bons souvenirs, des fous rires en salle de TP (parce que, oui, les paramécies bougent sous le microscope et que, bien sur que non, je n'ai pas vidé le bécher d'acide dans l'évier !) aux partiels passés en pyjama ;-). Plus

généralement merci d'être toujours là pour moi. A tous, merci d'être ceux que vous êtes et de croire en moi et merci aussi à ceux qui ne sont plus là et qui me manquent.

Un merci plus particulier à mes parents et à ma sœur. Merci à mes parents de m'avoir toujours encouragé dans mes études, même si j'ai souvent changé d'avis et ai eu l'air un peu perdue à ne pas savoir ce que je voulais faire. Merci Papa pour avoir réussi à me faire comprendre qu'avoir un travail qui nous plait rend la vie beaucoup plus facile. Je sais que je donne souvent l'impression de pas écouter tes conseils mais tu vois tu n'as pas toujours parlé dans le vent. Maman, merci d'être toujours présente tout en me laissant vivre et construire ma vie en liberté. Je sais que la solution à tous mes problèmes n'est jamais bien loin, puisqu'il me suffit de t'appeler pour la trouver. Merci Lisa pour m'avoir toujours soutenue. Même si c'est toi la plus petite tu as toujours été là pour me soutenir et rattraper (ou moins cacher) mes erreurs et je ne t'en remercierai jamais assez. A tous les trois, merci de m'avoir aidé à grandir et à devenir celle que je suis aujourd'hui.

Pour finir je souhaiterai remercier la personne sans qui je ne serais certainement pas en train d'écrire ces mots, Baptiste. Tout d'abord Merci pour ton soutient indéfectibles pendant ces trois années dans les bons et les mauvais moments, contrairement à moi tu n'as jamais douté que j'arriverais au bout de cette thèse ! Merci aussi de m'avoir écouté me plaindre de mes analyses bio-informatiques sans vraiment savoir de quoi je parlais et d'avoir toujours essayer de me comprendre quand je cherchais à te transmettre mon enthousiasme pour mes derniers résultats. Merci pour toutes les petites choses pour lesquelles tu m'as aidé pendant cette thèse. De la découpe des (2000) étiquettes pour poissons, aux nombreuses répétions d'orales où tu m'as écouté et bien sûr chronométré en passant par la correction de mes nombreuses fautes d'orthographe. Enfin plus généralement merci d'être la personne que tu es et d'avoir choisi de partager ta vie avec moi ! Merci de me comprendre et de m'accepter telle que je suis sans jamais avoir cherché à me changer. Tu es devenu au fil des années cette nouvelle partie de moi qui fait que je remarque ton absence bien plus que ta présence. Je suis pleinement consciente de la chance que j'ai de pouvoir me réveiller avec toi chaque jour et à quel point chaque moments passés ensemble sont précieux ! j'espère pouvoir en vivre encore pleins d'autres.

MERCI A TOUS

Résumé: La spéciation est un processus évolutif conduisant à la formation de nouvelles espèces grâce à l'établissement progressif de barrières d'isolement reproductif entre populations en cours de divergence. Comprendre de quoi sont constituées ces barrières, quelles sont les forces évolutives ayant permis leur mise en place et comment elles impactent la valeur sélective des individus hybrides sont des questions fondamentales en biologie évolutive. Cependant, répondre à ces questions en étudiant de vraies espèces qui n'interagissent plus au travers d'échanges génétiques peut s'avérer difficile, tant les barrières d'isolement reproductif sont nombreuses. C'est pourquoi nous avons choisi d'étudier dans cette thèse le bar européen (*Dicentrarchus labrax*), une espèce de poisson marin subdivisée en deux lignées évolutives en cours de spéciation, et représentées par les populations atlantique (bar) et méditerranéenne (loup). Afin de comprendre comment la divergence s'est établie et maintenue entre ces deux lignées au cours du temps, nous avons combiné plusieurs approches. Dans un premier temps, une étude de génomique des populations réalisée sur des individus sauvages nous a permis de préciser le contexte démographique dans lequel la divergence a eu lieu et d'identifier les mécanismes évolutifs ayant permis à la différenciation génétique de s'établir. Nous avons ainsi confirmé que les variations chromosomiques du taux de recombinaison influencent le maintien dans le génome d'allèles impliqués dans l'isolement reproductif. Les locus d'isolement occupent des régions où les niveaux de divergence nucléotidique sont particulièrement élevés. Nous avons pu relier ces excès de divergence à la présence d'allèles anciennement introgressés. En effet, des échanges génétiques semblent avoir eu lieu il y a près de 80000 ans entre la population de bar atlantique et une espèce proche, le bar moucheté (*Dicentrarchus punctatus*). Nos inférences montrent que ces échanges ont très probablement facilité la mise en place de l'isolement reproductif entre les deux lignées actuelles de bar européen. Dans un deuxième temps, nous avons étudié les signatures d'évolution moléculaire des gènes qui participent à l'isolement reproductif. Nous avons montré que ce sont principalement des gènes sous fortes contraintes évolutives subissant de fortes pressions de sélection purificatrice. Dans un troisième temps, nous avons utilisé des croisements expérimentaux afin de déterminer s'il existe une dépression d'hybridation chez les premières générations hybrides rencontrées en nature. Nous avons ainsi pu constater que la valeur sélective d'individus issus d'une première génération de rétrocroisement n'était pas inférieure à celle de leurs parents méditerranéens. Nos résultats montrent au contraire que les allèles d'origine atlantique sont favorisés chez ces hybrides au niveau des locus d'isolement, c'est-à-dire là même où ils sont contre-sélectionnée sur le long terme. La sélection agissant sur les haplotypes atlantiques introgressés suit donc une dynamique temporelle complexe. Ainsi, cette thèse aura influencé la vision classique du processus de spéciation allopatrique, généralement assimilé à une accumulation progressive d'une divergence génétique facilitée par l'absence de flux génique. Ici, nous avons montré que des échanges génétiques anciens avec une troisième lignée ont probablement accéléré l'émergence de l'isolement reproductif entre les deux lignées de bar. De plus, nous avons montré que l'isolement reproductif ne résulte pas d'une forte contre-sélection des hybrides de premières générations, mais d'une dynamique de sélection plus complexe des fragments introgressés, dont le sens s'inverse au fil des générations. Cette thèse montre que la sélection en liaison, en interaction avec les variations locales du taux de recombinaison, joue un rôle fondamental dans la mise en place et le maintien de l'isolement reproductif.

Mots clé : Génomique – Spéciation - Phasage haplotypique - Isolement reproductif - Contact secondaire - Evolution moléculaire.

Abstract: Speciation is the evolutionary process of species formation through the progressive establishment of reproductive isolation barriers between diverging populations. Understanding what kind of loci constitute these barriers, which evolutionary forces underlie their formation and how they impact the fitness of hybrids are fundamental questions in evolutionary biology. However, studying true biological species that do not interact through genetic exchanges do not help answering these questions as many reproductive isolation barriers potentially exist, making the identification of the initial barriers a difficult task. This is why we here focus our study of speciation on the European sea bass (*Dicentrarchus labrax*), a marine fish species subdivided into two incipient species, which are represented by the Atlantic and Mediterranean evolutionary lineages. In order to understand how divergence built up and maintained between these two lineages, we combined several complementary approaches. First, a population genomic study of wild individuals allowed us to specify the demographic context in which divergence took place and to identify the evolutionary mechanisms that allowed genetic differentiation to unfold at the genome level. We found that chromosomal variations in recombination rate have influenced the establishment of reproductive isolation. Furthermore, genomic regions involved in reproductive isolation showed particularly high levels of sequence divergence, that we related to the presence of anciently introgressed alleles. Our findings indicate that past genetic exchanges between *D. labrax* Atlantic lineage and a closely related species, the spotted sea bass (*Dicentrarchus punctatus*), have facilitated the establishment of reproductive isolation between the two extent *D. labrax* lineages. Secondly, we studied molecular evolution patterns of genes involved in reproductive isolation. We showed that these genes mainly display strong evolutionary constraints and thus undergo strong purifying selection. Thirdly, we used experimental crossings to determine if backcrossed individuals have a reduced fitness, which could be expected if there is hybrid depression. We observed that backcrossed individuals had the same fitness than non-backcross Mediterranean relatives. On the contrary, Atlantic alleles are even favored for these hybrids, and this occurs at the same loci where negative selection operates against Atlantic alleles over the long-term. These results thus reveal a complex temporal dynamic of selection on foreign Atlantic haplotypes. Overall, this work challenges the classical view of allopatric speciation which is generally thought as the progressive accumulation of barriers as the by-product of divergence between two populations, facilitated by the absence of gene flow. Here, we showed that on the contrary, genetic exchanges between *D. labrax* Atlantic population and a third lineage have probably accelerated the emergence of reproductive isolation between the two European sea bass lineages. Furthermore, reproductive isolation is generally assumed to rely on strong selection against first-generations hybrids. However, we showed that this is probably not the case for the European sea bass in which the dynamics of selection on foreign genetic material could be more complex, involving an inversion of selective effect over generations. This thesis shows that linked selection, in interaction with local recombination rate, plays a fundamental role in the establishment and maintenance of reproductive isolation.

Key words : Genomic – Speciation – Haplotype phasing – Reproductive isolation – Secondary contact – Molecular evolution.